Efficient and Reliable Deep Learning Methods and their Scientific Applications (25w5382)

Andrea Bertozzi (UCLA), Gitta Kutyniok, (LMU Munich), Stanley Osher (UCLA)
Bao Wang (University of Utah), Jack Xin (UC Irvine)

June 22 - 27, 2025

1 Overview of the Field

Deep learning (DL) brings unprecedented opportunities to solve challenging problems, ranging from image and language perception to scientific computing. However, most DL successes are based on heuristics rather than mathematical principles. Two prominent bottlenecks are: (1) lack of interpretability and performance guarantee; (2) excessive computational costs and dependence on massive high-quality data. Efficient deep learning (EDL) emerges as a branch of DL that aims to accelerate the prediction speed through a lightweight deep neural network (LWDNN) while maintaining the level of performance (accuracy) of its heavyweight counterpart. Though LWDNN is intended for AI applications in resource-limited environments, developing principled EDL methods to resolve the two bottlenecks mentioned above benefits computational science.

In this workshop, we invited leading experts to present cutting-edge mathematical and computational methods aligned with the themes of *principled and reliable low-dimensional and sparse approximations*, *efficient and trustworthy generative modeling*, *effective utilization of data priors*, and *in-context efficient learning*. The discussions also explored a wide range of scientific applications. A central focus of the workshop was on developing low-cost, lightweight models with interpretability supported by rigorous mathematical foundations, applicable across diverse tasks and domains.

The workshop delved into mathematics rooted in fundamental problems in cutting-edge AI. A major line of low-cost networks comes from low-precision approximations of weights and activations of neural networks (so-called quantization). If computing resources allow network re-training, mathematical issues of non-standard gradient descent arise in constructing and validating proxy gradients (a.k.a. straight through estimators) [44, 52]. Theoretical analysis of such a gradient driven learning algorithm to solve a challenging class of discrete high-dimensional nonconvex minimization problems (a.k.a Quantization Aware Training QAT) is emerging [130, 129, 73, 53]. When retraining is not an affordable option, as is often the case with LLMs, post training quantization (PTQ) [80, 76, 36] aims to extract a quantized model locally by minimizing a simplified surrogate loss at a significantly reduced algorithmic complexity than QAT. On the downside, PTQ suffers a heavier performance degradation than QAT, especially at the very low precision (binary bit) regime. Rigorous theoretical error analysis and backpropagation-free algorithms have been developed for PTQ based on linear algebra, statistics, and optimization ideas recently [135, 136, 134]. These efforts are essential for edge computing, as they greatly improve communication efficiency while safeguarding data privacy. In particular, these advances can significantly advances the field of federated learning [77, 113, 101, 46, 68, 100, 102, 47].

Efficient approximations of the vanilla softmax attention block [111] in transformers (as adopted in LLMs and computer vision) constitute another line of active research on reducing complexity to linear or almost lin-

ear growth in the regime of long input sequence length. Approximation ideas range from separable kernel function (linear attention [57] and improved variants [82, 114, 42, 34]), sparsity [58, 142, 141, 85], dissecting long- and short-range interactions [84], dilation [30], Fourier integral theorem [83] among others. A composite approximation consists of a local window attention and a global token mixing operation such as repeated window shifting [72], fast Fourier transform [106] and averaging [108]. Theoretical guarantees of local-global attention approximations await to be studied beyond those formulated as equivalence theorems in [106].

Reliability and safety of AI have caught the attention of the scientific community in recent years. AI safety research challenges broadly fall into three types [8]: 1) creating trustworthy AI systems (Development), 2) evaluating risks (Assessment), 3) monitoring and intervening after deployment (Control). Related issues are sustainability and energy efficiency. Deep mathematical understandings contribute to the forecasting and quantification tasks in the area. For example, rate distortion measures based on information theory helped minimize artifacts generated by AI [59], provably reliable AI-based communication systems helped comply with AI policies, and computability guarantees helped shape future energy-efficient AI computing [61].

The workshop explored *how the EDL tools developed around AI contribute to computational mathematics and scientific applications.*

One scientific application of broad interest is to advance and accelerate traditional computational methods of partial differential equations (PDE) based on EDL. Traditional methods are typically slow and costly in three or higher space dimensions when resolving solutions that are either in the low regularity class (non-smooth) or nearly singular (blow-up) in space or highly non-stationary (irregularly oscillatory) in time. In contrast, EDL methods are much faster at inference (prediction) once trained. An intriguing research topic at the workshop is on computing physically meaningful weak or multi-scale solutions through new formulations being both physical and adaptive to optimization techniques and neural networks approximations. New loss functions appeared recently for both the forward and inverse problems. Approaches for forward problems include score-based optimizations, implicit entropy/viscosity solution representations, jump condition satisfying and derivative-free approximations [128, 16, 143, 86, 19, 17, 120, 121, 139], going beyond physics informed formulations for differentiable (strong) solutions [56]. For both forward and inverse problems, interesting loss functions appeared in hybridized machine learning and classical methods [88, 33, 92]. In connection with efficient attention and transformers, much progress is seen in neural operator learning [75, 66] of PDEs with applications to inverse problems, mean field games and turbulent flows [78, 123, 65, 48, 37, 133, 55, 20]. An emerging area leveraging LLMs for scientific applications—including solving PDEs—is *in-context learning*, which has shown remarkable success in AI and scientific computing [15, 126, 27].

Another application is to design EDL methods to meet fundamental scientific complexities. Many physical problems—e.g. in material sciences and biophysics—exhibit two phenomena at the atomic level: the atomic structure exhibiting intricate *symmetries*, and *long-range interactions* (LRI) among distant atoms, causing the foremost computational challenges. Leveraging data symmetries in EDL enhances efficiency by exploiting inherent patterns or invariances in data, reducing computational complexity, and improving model generalization. For instance, symmetries such as translation, rotation, or scale invariance can be incorporated into neural architectures like convolutional neural networks (CNNs), which use weight sharing to reduce parameters while maintaining robustness to transformations. Techniques like group equivariant networks further generalize this concept by designing layers that are invariant to specific symmetry groups and leveraging steerable features, as discussed in [26, 109, 116, 7]. Additionally, data augmentation strategies exploit symmetries by generating transformed versions of training data, improving model robustness without additional labeled data [94]. These approaches collectively minimize redundant computations and enhance performance on tasks such as image classification and natural language processing. Learning LRI is another scientific challenge. Classical approaches, e.g. fast multipole method (FMM) and Ewald summation can compute LRI with sublinear complexity; however, there is no guaranteed EDL approach yet for learning LRI. Recurrent neural networks (RNNs) [87, 81], neural ODEs [24, 32, 125, 3], graph neural networks (GNNs) [40, 103, 6, 115, 5], and multiscale representations [117] have been tailored to learn LRI for scientific computing.

Recent advances in diffusion- [96, 45, 99] and flow-based [69, 1, 70] generative models have significantly expanded the frontier of scientific machine learning. These models offer powerful frameworks for learning complex data distributions, enabling high-fidelity generation and inference in domains ranging from molecular design [4, 104, 131] to climate modeling [64]. Diffusion models, inspired by stochastic processes, have

demonstrated remarkable performance in generating realistic samples by iteratively denoising data, while flow-based models provide exact likelihood estimation and invertible mappings, making them particularly suitable for tasks requiring precise control and interpretability. In scientific applications, these generative approaches are being leveraged to simulate physical systems, accelerate drug discovery, reconstruct high-resolution images in medical diagnostics, and solve inverse problems in physics and engineering. Their ability to integrate domain knowledge and scale to high-dimensional data makes them indispensable tools for modern computational science. Mathematically, fast score matching algorithms have been proposed to accelerate the training of diffusion models. However, generation process of diffusion- and flow-based generative models requires multiple expensive evaluations of a large neural networks, incurring massive computational cost. Consistency models [98, 97] enable single-step generation by mapping any noisy sample directly to data. Flow maps [38, 13] and mean flow models extend this capability by efficiently learning mappings between two distributions.

Developing fundamentally new ideas and approaches called for close collaborations among mathematicians, computational scientists, and domain experts. The workshop has been an excellent opportunity to bring together world-leading researchers and young talents from relevant disciplines. It featured both accomplished and rising scientists at the intersection of mathematics, machine learning, and applications; and created a lively forum for them to present recent developments and foster new interactions.

2 Recent Developments and Open Problems

Efficient and reliable deep learning brought up a cohort of fundamental mathematical and computational issues worth studying from the perspectives of low dimensional approximations in functional and probability spaces using various tools from numerical analysis and data science. We discuss recent developments and point out related open problems systematically below.

Efficient neural networks and accelerated inference. Quantization is a widely adopted low-precision discrete approximation of neural weights and activation functions to speed up network inference or prediction post-training. In quantization-aware training (QAT), there remain many open problems in understanding empirical algorithms such as binary-connect and straight-through estimators—two well-established quantization algorithms for EDL. Mathematical analysis in recent years has shed light on potential instabilities and oscillation behavior in minimizing population (infinite sample) loss of shallow networks [73], and a new notion of coarse gradient [129] to guide descent in solving high dimensional large-scale discontinuous optimization problems of broader interest. A significant recent theoretical progress in QAT is on dynamic behavior of coarse-gradient (straight-through estimator, a.k.a STE) based training algorithms solving high dimensional non-differentiable optimization problems with *finite sample loss functions* as objectives [53]. For a two-layer neural network with binary weights and activations, sample complexity bounds in terms of the data dimensionality guarantee the convergence of STE-based optimization to the global minimum. In the presence of label noise, the training sequence satisfies an intriguing recurrence property where the iterates repeatedly escape from and return to the optimal binary weights. The work leverages tools from 1-bit compressed sensing [14] and dynamical systems. A challenging open problem is to extend the theory to deep networks. Another interesting direction is the application of QAT as efficient PDE solvers [110].

Training DNNs is computationally expensive, e.g. training a large language model (LLM) costs millions of dollars. Moreover, prediction using a well-trained DNN is also very expensive. One of the major concerns of generative AI is that they are costly, e.g., a single search using ChatGPT is far more expensive than the traditional search engines. Post training quantization (PTQ) has been adopted as a low cost solution in LLMs. Two types of methods emerge from transformation techniques and rounding schemes. Scale invariance transforms (e.g. efficient Hadamard rotations [2]) redistribute quantization difficulty between weights and activations. Along the line of rounding schemes [80], recent works adopted principled discrete optimization using greedy, sequential rounding strategies to select quantized weights to minimize the layer-wise reconstruction error [76, 36, 135, 134]. The widely used OPTQ [36] is a sequential algorithm alternating between rounding the current weight and adjusting future un-quantized weights to reduce error (a.k.a. diffusion), yet without correcting quantization error in the activations or in quantizing previous layers. GPFQ [76] explicitly corrects quantization error in both the weights and activations in each rounding iteration, as

well as error from previous layers, but does not adjust the future un-quantized weights. Qronos [136] presented at the workshop subsumes and surpasses OPTQ and GPFQ, explicitly correcting quantization error in both the weights and activations of previous layers while diffusing error into future weights. It gave the first rigorous theoretical error bound for OPTQ and Qronos with stochastic rounding using convex ordering and high dimensional probability. A related data-driven post-training low-rank compression technique with recovery theorems is also developed. Earlier work on rounding (e.g. [80]) appeared too costly on LLMs due to gradient computation. A recent progress is made via coordinate descent and close-form solutions [134] free of back-propagation (gradient descent). Future work includes the extension of PTQ to multi-modal models (e.g. vision-language models), and combining per-layer and per-channel quantization strategies into a mix-precision quantization framework.

Efficient attention is a highly active area to reduce complexity of transformers, save energy consumption and accelerate their inference speeds on cloud and mobile platforms. Vanilla attention ([111], 2017) scales quadratically in the input sequence length, and various efforts have been made since to reduce complexity to linear scaling. LongNet [30] by Microsoft Research is a recent example on LLMs based on dilated attention (a type of sparse attention among others [142, 141]) that yields strong performance on long-sequence modeling and general language tasks, opening up new possibilities such as treating a whole corpus or even the entire Internet as a sequence. Variants of linear [42, 43, 34] and recursive (mamba) attentions [41] can also be competitive on certain data sets. On the other hand, linear complexity can be achieved by a composite attention structured as an integrated local and global approximation of full attention. An example is vision transformer Swin [72] where window attention (local) is shifted and globalized as network depth increases. Along this line, recent work investigated Fourier transform (FWin [107, 106]) and averaging (SEMA [108]) based globalization techniques. Fourier transform is a robust choice for moderately long time series without relying on enough network depths to globalize, while averaging is consistent with the asymptotic behavior of vanilla and linear attentions in the large sequence length limit (a.k.a. dispersion [112]), and scales better for large size images. For a fixed window size in the local attention, larger input image sizes (say from 224^2 to 1024^2) yield longer tokens after patch embeddings, causing all transformer networks to degrade in top-1 accuracies on ImageNet-1K. Without a retraining (finetuning), the inference accuracy of SEMA out-performs that of vision mamba [71] on larger size images. Future work remains to evaluate and further develop a SEMA type composite attention framework for large images in medical sciences and multi-modal data, as well as establish probabilistic error estimates where averaging is cast in a form of law of large numbers. For an analogue of averaging (aggregation) operation to promote long-range interaction on graphs, see [5].

Efficiency and fast inference issues arise in generative AI recently concerning diffusion- and flow-based generative models. The emerging research direction points to one/few step generation by imposing consistency in training loss [98], distillation [39, 35], flow map [12] or mean flow [38] matching. Interesting questions being discussed at the workshop include: 1) draw connections of one-step diffusion models with optimal transport and Schrödinger bridge that have been studied intensively in mathematics, 2) study mathematical foundation and error estimates of one-step diffusion models, 3) explore one-step diffusion models in scientific applications. Preliminary studies [139] suggest that one step diffusion models scale better than optimal transport maps of distributions as the underlying dimensions of data samples increase. Score-based diffusion models have been analyzed in terms of well-posedness and error bounds [22, 118, 79], as well as design space and fast algorithms [119, 91, 51, 89], providing tools along with optimization techniques over probability measure space [25] for making further progress. Interestingly, score functions in similar form appeared in the adjoint Monte Carlo method for Boltzmann equations [16, 128]. Likewise, a deep learning algorithm is developed for computing mean field control problems via forward-backward score dynamics [143]. Establishing partial differential equation-based frameworks to analyze the generation error dynamics is another emerging scientific area [50].

Connections between numerical analysis and efficient deep learning. The workshop revisited approximation properties of neural networks with ReLU activations to functions or solutions of elliptic PDEs and frequency contents of residuals under gradient descent minimization [95, 140]. The residuals tend to contain much more high frequency errors than under piecewise linear finite element approximations, suggesting that it is difficult for ReLU neural networks to approximate highly oscillatory or discontinuous functions [137]. Progress in approximating high frequency non-smooth functions and subsequent efficient learning is made by choosing sine or truncated sine activation functions and scaling of initial layer in a structured and balanced

multi-component and multi-layer network [138], a further development from [66].

Computing physically meaningful weak solutions to nonlinear PDEs is a timely topic in the workshop. In contrast to adopting PDE residuals to measure the errors of strong (smooth) solutions [56], there appears no simple and universal route. For first order Hamilton-Jacobi equations (HJE), the classical Lax-Oleinik formula provides a viscosity solution for convex Hamiltonians. In [86], an implicit representation extends the formula and engages a deep neural network for efficiently computing a variety of HJE's in high dimensions. Alternatively, enforcing entropy condition implicitly in a least squares type loss function incorporating divergence form or classical well-established discretizations appeared as a viable approach in training neural networks for efficiently computing shock waves in scalar conservation laws [18], interface and high frequency problems [19, 60], and HJE [33]. The minimum requirement of neural network based computation of weak (discontinuous) solutions calls for *physics-preserved* instead of physics-informed formulations. Yet another avenue to compute solutions in a distribution sense stems from probabilistic representations in similar spirit to scaling up to continuum mechanics from physical microscopic dynamics (stochastic differential equations). Martingale [17] and interacting particle ideas [120, 121, 139] led to spatial gradient free (mesh-free) approximations, promising for high dimensional reaction-diffusion-advection (second order in space) PDEs, and learning their physically parameterized solutions [62]. From a geometric view, parameterized Wasserstein flows [54, 124] offer another way to compute certain classes of high dimensional Fokker-Planck and Hamiltonian PDEs without spatial discretization.

Theoretical analysis of the effectiveness of leveraging in-context learning for solving PDEs have also been discussed in this workshop [126, 27]. In-context learning is a paradigm where LLMs or foundation models solve tasks by conditioning on a few examples provided in the input prompt—without updating model weights. When applied to PDEs, In context learning enables models to learn solution patterns from a small number of example input-output pairs (e.g., boundary conditions and corresponding solutions) and generalize to new, unseen PDE instances. In scientific computing, this approach is particularly valuable because: (1) It bypasses the need for retraining or fine-tuning on each new PDE. (2) It allows zero-shot or few-shot generalization to different geometries, boundary conditions, or PDE types. (3) It can be integrated with physics-informed priors or neural operators to improve accuracy and interpretability.

Integration of symmetry into neural network architecture design. The workshop discussed symmetry-integrating mechanism in neural network architecture design, ranging from both data representation [117] to predictive and generative AI [144]. Symmetry is omnipresent in both data science and physical sciences. For instance, scale invariance is a key factor in imaging sciences and intricate symmetries exhibit in crystalline structures. Crystal symmetries, representing invariances under certain transformations, such as rotations or reflections, influence the properties of the materials in unique ways. Integrating symmetry priors into deep learning is essential for enhancing physical fidelity and achieving data efficiency. Different kinds of symmetry-aware DNN architectures have been designed in the past several years, including invariant methods and equivariant methods using different types of steerable features. But it is still hard to say which kind of method is much more powerful; using higher-order steerable features requiring computing prohibitive Clebsch-Gordan tensor product but the accuracy gain is marginal. Existing empirical studies show that it is emerging to understand the role of invariant and covariant features and how architectures affect the models' performance. In this workshop, we brought experts in the scientific community to share and discuss latest advances in this scientific area. Meanwhile, our workshop has brought synergies between experts in different domains to explore symmetry-aware neural nets as scalable and efficient surrogates for scientific discovery and decision-making that are faithful to the underpinning physics for learning complex phenomena, e.g., plasma physics and fusion reactor design [128, 105, 92].

A challenging open scientific problem raised in this workshop is how to build multiscale symmetry-aware neural network for efficient deep learning. Integrating multiscale features is fundamental for the success of AI for Science, especially in computational molecular design. For instance, proteins are built from atoms forming amino acids (residues), which link into a linear polypeptide chain (primary structure). This chain folds into local secondary structures, like α-helices and β-sheets, each containing 2-40 contiguous residues and stabilized by hydrogen bonds. These progressively fold into higher-order structures, determining the protein's 3D geometry and functions. Integrating multiscale molecular features and complex 3D molecular geometry is essential for predictive modeling of proteins. Primary and secondary structures are among the most critical multiscale features, since (1) they underpin higher-order structures, and (2) identical primary

structures can even yield distinct secondary structures and properties. However, existing EDL methodologies struggle to overcome the challenges in integrating multiscale features and molecular geometry efficiently. Using a fully connected graph that incorporates all pairwise distances is common for capturing complete molecular geometry [29]; however, it becomes costly for proteins with thousands of residues.

Sample efficient deep learning and multi-modal/multi-fidelity deep learning. Training DNNs is data-hungry. Enforcing symmetry into DNN design can significantly improve data efficiency, which can be understood as a data augmentation scheme from a machine learning viewpoint [94]. Active learning (AL) is another remarkable way to quench a DNN's thirst for data. AL algorithms have access to a large but unlabeled dataset and sequentially select the most "informative" examples for labeling. At each iteration, an unlabeled example, e.g., the one that is most difficult to interpolate, is selected and labeled. The data interpolating function based on the previously labeled examples predicts a label for each unlabeled example. We then construct a new function that additionally interpolates each unlabeled example. However, the mathematical understanding of how symmetry and AL algorithms benefit sample efficiency is still lacking. Without such a theoretical foundation, we do not know the synergy between symmetry and AL. Even using AL, labeling scientific data is still expensive and challenging.

Another challenge for applying EDL to scientific computing lies in multi-modality and/or multi-fidelity of the scientific data: for a given problem, we can have image data, time series data, and even categorical data. Some of the data are obtained from expensive and accurate simulations while others are sensed from coarsed grained models. We have to apply an efficient data fusion mechanism to fuse these multi-modal and multi-fidelity data for scientific applications. It has been noticed that transformers and graph neural networks are two particularly appealing DNN models for deep learning with multi-modal and multi-fidelity data.

Reliable AI and scientific machine learning. Reliability and safety of LLMs have raised awareness of researchers worldwide, especially in medicine, healthcare, robotics and algorithmic decision making. There is an increased need of principled approaches to protect the society from harmful model outputs. One solution is to learn from past mistakes in cybersecurity, draw analogies with historical examples and develop lessons learned that can be applied to LLM safety [122]. Reliability also hinges on explainability, expressivity and generalization. A more expressive network than ReLU network is the so called spiking neural network [61], including spike response and leaky-integrate-and-fire models, which are more grounded in neural science. Their promising applications to reliable computing and numerical PDEs with theoretical guarantees are interesting to explore. Recently, global convergence theory is developed for neural network PDE models in scientific machine learning and recurrent neural networks on long sequential data [90, 63] using neural tangent kernel and distributional fixed point techniques.

Efficient and reliable deep learning methods have been applied to some of the most challenging scientific computing problems, including protein structure prediction, many-body problems, learning closure models, and simulation/generation of weak/discontinuous solutions of PDEs. Other promising scientific machine learning tasks for further study include (1) fast prediction of cancer cell growth, infectious disease spreading and wildfire burning; (2) inverse design of novel tailored crystal materials; (3) accelerating structure-based drug and artificial swarm designs; and (4) computational imaging.

3 Presentation Highlights

There were a total of 45 talks in our workshop, with 34 in-person talks and 11 virtual talks. Each workshop day was scheduled with topics related to our workshop themes, as summarized below.

- **Monday June 23:** *Morning Session:* This session highlighted innovative approaches at the intersection of data-driven modeling, optimization, and scientific machine learning. The first talk introduced an adjoint-based optimization framework for the Boltzmann equation, offering a powerful tool for kinetic theory and transport problems. The second presentation focused on data-driven strategies for constructing moment closures in radiation transport, addressing a key challenge in high-fidelity simulations. The third talk explored graph-based active learning techniques for hyperspectral unmixing in nearly blind settings, demonstrating the potential of AI in remote sensing and signal processing. The

final talk provided a broad overview of neural networks in scientific computing (SciML), outlining foundational concepts and posing critical questions that challenge current methodologies. Together, these talks underscored the growing role of AI and optimization in advancing computational science across diverse domains. *Afternoon Session:* This session brought together a compelling set of talks focused on advancing scientific machine learning through rigorous mathematical frameworks, efficient algorithms, and novel applications. The presentations covered a wide range of topics, including a new approach for solving Hamilton–Jacobi equations with applications to optimal transport, and deep learning methods grounded in classical convergent numerical schemes for differential equations. Talks also addressed optimal control in level-set front propagation models for complex flows and introduced homotopy training algorithms to improve learning efficiency and stability. Further contributions included efficient local-global attention approximations, unsupervised solution operator learning for mean-field games, and provable in-context learning of partial differential equations (PDEs). Collectively, these talks emphasized the integration of mathematical theory with scalable AI techniques to tackle high-dimensional, dynamic, and physically grounded problems in scientific computing. *Evening Session:* This session focused on cutting-edge approaches to solving inverse problems and learning PDE dynamics through data-driven and probabilistic methods. The first talk introduced DeepParticle, a novel framework that learns PDE dynamics by minimizing the Wasserstein distance on data generated from interacting particle methods, bridging physical modeling with deep learning. The second presentation explored neural inverse operators for solving PDE inverse problems, emphasizing flexibility and generalization across different problem settings. The final talk presented HJ-sampler, a Bayesian sampling method that leverages Hamilton–Jacobi PDEs and score-based generative models to address inverse problems in stochastic processes. Together, these talks highlighted the power of integrating physics-informed learning, generative modeling, and probabilistic inference to tackle complex scientific challenges.

- **Tuesday June 24:** *Morning Session:* This session explored the intersection of AI safety, uncertainty quantification, and advanced mathematical modeling for high-dimensional systems. The first talk addressed the growing concerns around AI risks and surveyed current approaches to ensuring AI safety, emphasizing the need for robust, transparent, and accountable systems. The second presentation introduced martingale deep learning methods for solving very high-dimensional quasi-linear PDEs and stochastic optimal control problems, offering a novel probabilistic framework for scientific computing. The final talk focused on active operator learning with predictive uncertainty quantification for PDEs, highlighting strategies for improving reliability and interpretability in data-driven models. Together, these talks underscored the importance of integrating safety, mathematical rigor, and uncertainty-aware learning in the development of next-generation AI tools for scientific applications. *Afternoon Session:* This session featured a diverse set of talks that advanced both the theoretical and practical frontiers of machine learning and scientific computing. Topics ranged from kernel-based methods for point cloud analysis and sparse radial basis function networks for solving nonlinear PDEs, to symbolic and large language model (LLM)-based approaches for modeling in the space of language. Several talks focused on foundational aspects of learning theory, including finite-sample analysis for binarized neural networks, optimization over probability measure spaces, and quantization and compression of neural networks with theoretical guarantees. The session also explored generative modeling through a Wasserstein bound for diffusion models under Gaussian tail assumptions, highlighting the importance of rigorous analysis in understanding model behavior. Together, these talks reflected a strong emphasis on mathematical rigor, computational efficiency, and the integration of structure and theory in modern AI systems. *Evening Session:* This session focused on advanced operator learning techniques and their applications to complex physical systems. The first talk introduced self-test loss functions for data-driven modeling of weak-form operators, offering a novel approach to improving model reliability and generalization. The second presentation proposed the nonlocal attention operator as a step toward developing foundation models for physical response prediction, emphasizing scalability and expressiveness. The final talk demonstrated the accurate fine-tuning of spatiotemporal Fourier neural operators for modeling turbulent flows, showcasing the potential of operator-based deep learning in capturing intricate dynamics in fluid systems. Together, these talks highlighted the growing sophistication of operator learning frameworks and their transformative potential in scientific machine learning.

- **Wednesday June 25:** *Morning Session:* This session delved into the theoretical underpinnings and generative modeling aspects of deep learning, with a focus on partial differential equations (PDEs) and diffusion-based methods. The first two talks provided rigorous mathematical insights into neural network approximation and convergence, including integral representations of Sobolev spaces using ReLU^k activations and error estimates for linearized networks, as well as convergence analysis of neural network methods for solving PDEs. The latter two talks shifted toward generative modeling, examining the generation accuracy of diffusion models and multimodal sampling via denoising annealing, and exploring the theoretical connections and discrepancies between diffusion processes and flow matching. Together, these talks highlighted the interplay between mathematical theory and generative modeling in advancing the reliability and interpretability of AI methods in scientific computing.

- **Thursday June 26:** *Morning session:* This session brought together a rich blend of theoretical and algorithmic insights aimed at deepening our understanding of neural networks and their applications in scientific computing. The first talk provided a comprehensive overview of the mathematical and computational foundations of neural networks, tracing their evolution from shallow to deep architectures and examining their learning dynamics. The second talk introduced a novel framework based on parameterized Wasserstein geometric flows, offering a fresh perspective on optimization and transport in high-dimensional spaces. The third presentation focused on a variational Bayesian approach for sequence model prediction and uncertainty quantification, highlighting the importance of probabilistic reasoning in time-dependent data. The final talk addressed the convergence properties of real-time recurrent learning (RTRL) algorithms for a class of recurrent neural networks, contributing to the theoretical understanding of training dynamics in sequential models. Collectively, these talks underscored the role of rigorous mathematical analysis in advancing the reliability, interpretability, and efficiency of modern AI systems. *Afternoon session:* This session showcased a diverse and forward-looking collection of talks centered on the development of reliable, structure-aware, and scientifically grounded AI methods. The presentations spanned a wide range of topics, including model-consistent strategies for PDE joint inversion, structure-preserving machine learning for dynamical systems and data-driven discovery, and generative modeling approaches such as stochastic interpolants for science and engineering. Several talks emphasized the integration of physical principles into AI frameworks, such as conservative neural network methods for nonlinear conservation laws and deep learning techniques for tensegrity structures and constitutive laws in materials science. The session concluded with a broader perspective on building reliable and sustainable AI systems, highlighting the importance of mathematical foundations in shaping the next generation of AI computing. Collectively, these talks underscored the growing synergy between machine learning, physical modeling, and scientific applications.

- **Friday June 27:** The open discussion of Friday morning was lively with different viewpoints expressed on EDL in scientific computing, and with EDL advances in scientific applications further illustrated.

4 Scientific Progress Made

Accelerating both the training and inference of deep neural networks (DNNs) is essential for the practical deployment of Efficient Deep Learning (EDL) in artificial intelligence and computational science applications. This workshop provided a valuable platform for exploring this challenge from multiple perspectives, bringing together insights from mathematical theory, algorithm design, and real-world implementation. Throughout the sessions, participants discussed a wide array of strategies aimed at enhancing the efficiency and scalability of EDL. These included techniques such as sparsification and low-rank modeling [74, 31, 93], which reduce computational complexity by exploiting the inherent structure of data and models. Knowledge distillation methods [67, 39, 35] were also examined for their ability to transfer information from large models to smaller, more efficient ones without significant loss in performance. Further discussions focused on quantization techniques [129, 76, 136, 134, 110], which enable low-precision computation to reduce memory and energy usage. Other promising directions included efficient attention mechanisms, optimal transport, one-step diffusion and flow matching, and implicit or probabilistic representations of weak solutions to high-dimensional partial differential equations (PDEs) [86, 120, 121, 19, 17, 33]. In addition, the workshop highlighted the potential of homotopy-based training algorithms, higher-order and symbolic methods [127, 51, 89, 9], and

sparse, geometry-aware data representations [117] to further advance the theoretical foundations and computational efficiency of EDL. These discussions not only deepened our understanding of the current landscape but also sparked new collaborations and research directions aimed at pushing the boundaries of what EDL can achieve.

A central theme of the workshop was the exploration of emerging research directions at the intersection of deep learning, generative AI, and applied mathematics. Participants engaged in in-depth discussions on how foundational understanding, efficient algorithm design, and rigorous mathematical analysis can collectively advance the capabilities of DNNs and generative models. These conversations spanned the entire pipeline—from model architecture and training strategies to inference techniques, error estimation, and convergence theory. The workshop also highlighted the growing impact of these developments on a wide range of scientific applications. In particular, we examined how EDL and related methods can be applied to inverse problems [88, 65], uncertainty quantification [78, 28], and remote sensing [21, 132, 11]. Further applications included complex systems such as fluid dynamics and fusion reactor design [128, 16, 92, 20, 90], as well as domains like atmospheric science [49, 107] and molecular modeling [10, 23]. These discussions underscored the importance of interdisciplinary collaboration in pushing the boundaries of what is possible with AI-driven scientific computing. By integrating theoretical insights with practical challenges, the workshop laid the groundwork for future breakthroughs in both methodology and application.

Looking ahead, the workshop participants expressed a strong commitment to maintaining ongoing communication and fostering collaboration on shared research interests. By staying connected and exchanging updates on their respective work, they aim to build a sustained and supportive research network that can accelerate progress in EDL and its applications. To further stimulate research and broaden the impact of EDL, participants also discussed the possibility of organizing future workshops and symposia at major venues such as SIAM and leading AI conferences. These events would serve as valuable platforms for disseminating new findings, attracting a wider community of researchers, and catalyzing interdisciplinary collaborations. Such continued engagement is expected to play a key role in shaping the future trajectory of EDL and its integration into scientific and technological innovation.

5 Outcome of the Meeting

The workshop featured a diverse lineup of speakers representing a broad spectrum of career stages and professional backgrounds. Participants included established professors, early-career postdoctoral researchers, and advanced graduate students, each bringing unique perspectives and expertise to the discussions. These speakers hailed from a wide array of institutions, including leading universities, government research laboratories, and industry organizations. Geographically, the workshop achieved impressive international representation, drawing contributors from North America, Europe, Asia, and the Middle East. In addition to its academic and professional diversity, the workshop placed a strong emphasis on promoting gender balance. Organizers made a concerted effort to ensure equitable representation of both male and female researchers across all levels of seniority. This commitment to inclusiveness not only enriched the scientific dialogue but also fostered a more welcoming and supportive environment for all participants, reinforcing the workshop's dedication to equity and diversity in the research community.

While the majority of the workshop sessions were conducted in person, each day included a dedicated online component to accommodate remote participants. To ensure smooth coordination between the physical and virtual audiences, a designated chairperson was assigned to oversee the online session. This individual played a crucial role in managing the technical aspects of the hybrid format, facilitating real-time communication, and ensuring that remote attendees could actively participate in discussions, ask questions, and contribute to the dialogue. This hybrid setup was thoughtfully designed to foster inclusivity and maximize engagement across all modes of attendance. By bridging the gap between in-person and virtual participants, the workshop created a dynamic and interactive environment that allowed for full participation regardless of physical location. The seamless integration of both formats not only expanded the reach of the event but also enriched the overall experience for everyone involved.

The communal atmosphere of the Banff International Research Station played a pivotal role in fostering meaningful interactions among participants. Shared meals and outdoor activities provided informal yet valuable opportunities for attendees to engage in open conversations, exchange ideas, and explore a wide

range of scientific topics beyond the formal sessions. These relaxed settings encouraged spontaneous brainstorming and the cross-pollination of research ideas across disciplines and career stages. A highlight of this collaborative spirit was the open discussion held on Friday morning, which proved to be particularly lively and intellectually stimulating. Participants expressed a variety of viewpoints, contributing to a rich and multifaceted dialogue. The session also showcased recent advances in the application of EDL, illustrating its growing impact across different scientific domains. This vibrant exchange of ideas exemplified the workshop's commitment to fostering an inclusive and forward-thinking research environment.

References

[1] M. Albergo, N. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[2] S. Ashkboos, A. Mohtashami, M. L. Croci, B. Li, P. Cameron, M. Jaggi, D. Alistarh, T. Hoefler, and J. Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. In *Neural Information Processing Systems*, 2024.

[3] J. Baker, E. Cherkaev, A. Narayan, and B. Wang. Learning proper orthogonal decomposition of complex dynamics using heavy-ball neural odes. *Journal of Scientific Computing*, 95(2):54, 2023.

[4] J. Baker, Y. Huang, S-H Wang, M. Pasini, A. Bertozzi, and B. Wang. Stabilized e(n)-equivariant graph neural networks-assisted generative models, 2024.

[5] J. Baker, Q. Wang, M. Berzins, T. Strohmer, and B. Wang. Monotone operator theory-inspired message passing for learning long-range interaction on graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 2233–2241, 2024.

[6] J. Baker, Q. Wang, C. Hauck, and B. Wang. Implicit graph neural networks: A monotone operator viewpoint. In *International Conference on Machine Learning*, pages 1521–1548. PMLR, 2023.

[7] J. Baker, S-H. Wang, T. De Fernex, and B. Wang. An explicit frame construction for normalizing 3D point clouds. In *Proceedings of the 41st International Conference on Machine Learning*, pages 2456–2473, 2024.

[8] Y. Bengio et al. The Singapore consensus on global AI safety research priorities building a trustworthy, reliable and secure ai ecosystem. *arXiv preprint arXiv:2506.20702*, May 2025.

[9] R. Bhatnagar, L. Liang, K. Patel, and H. Yang. From equations to insights: Unraveling symbolic structures in PDEs with LLMs. *arXiv preprint arXiv:2503.09986*, 2025.

[10] K. Bhattacharya, L. Cao, G. Stepaniants, A. Stuart, and M. Trautner. Learning memory and material dependent constitutive laws. *arXiv preprint arXiv:2502.05463*, 2025.

[11] G. Bhusal, Y. Lou, C. Garcia-Cardona, and E. Merkurjev. A general framework for group sparsity in hyperspectral unmixing using endmember bundles. *arXiv preprint arXiv:2505.14634*, 2025.

[12] N. Boffi, M. Albergo, and E. Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *arXiv preprint arXiv:2406.07507*, 2024.

[13] N. Boffi, M. Albergo, and E. Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. *Transactions Machine Learning Research*, 2025.

[14] P. Boufounos and R. Baraniuk. 1-bit compressive sensing. In *IEEE Annual Conference on Information Sciences and Systems*, pages 16–21, 2008.

[15] T. Brown and et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[16] R. Caflisch and Y. Yang. Adjoint Monte Carlo method. *Active Particles*, 4:461–505, 2024.

[17] W. Cai, S. Fang, W. Zhang, and T. Zhou. Martingale deep learning for very high dimensional quasi-linear partial differential equations and stochastic optimal controls. *arXiv preprint arXiv:2408.14395*, 2024.

[18] Z. Cai, J. Chen, and M. Liu. Least-squares neural network (lsnn) method for scalar nonlinear hyperbolic conservation laws: Discrete divergence operator. *Journal of Computational and Applied Mathematics*, 433:115298, 2023.

[19] Z. Cai, J. Choi, and M. Liu. Least-squares ReLU neural network (lsnn) method for linear advection-reaction equation: Discontinuity interface. *SIAM Journal on Scientific Computing*, 46(4):C448–C478, 2024.

[20] S. Cao, F. Brarda, R. Li, and Y. Xi. Spectral-refiner: Accurate fine-tuning of spatiotemporal Fourier neural operator for turbulent flows. *arXiv preprint arXiv:2405.17211*, 2024.

[21] B. Chen, Y. Lou, A. Bertozzi, and J. Chanussot. Graph-based active learning for nearly blind hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

[22] H. Chen, H. Lee, and J. Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023.

[23] M. Chen and J. Qin. Form-finding and physical property predictions of tensegrity structures using deep neural networks. In *Asilomar Conference on Signals, Systems, and Computers*, pages 1090–1094, 2024.

[24] R. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[25] S. Chen, Q. Li, O. Tse, and S. Wright. Accelerating optimization over the space of probability measures. *Journal of Machine Learning Research*, 26(31):1–40, 2025.

[26] T. Cohen and M. Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

[27] F. Cole, Y. Lu, T. Zhang, and R. O'Neill. Provable in-context learning of linear systems and linear elliptic PDEs with transformers. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024.

[28] D. Coscia, M. Welling, N. Demo, and G. Rozza. Barnn: A Bayesian autoregressive and recurrent neural network. *arXiv preprint arXiv:2501.18665*, 2025.

[29] J. Dauparas and et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[30] J. Ding, S. Ma, L. Dong, X. Zhang, S. Huang, W. Wang, N. Zheng, and F. Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023.

[31] T. Dinh, B. Wang, A. Bertozzi, S. Osher, and J. Xin. Sparsity meets robustness: Channel pruning for the Feynman-Kac formalism principled robust deep neural nets. In *International Conference on Machine Learning, Optimization, and Data Science*, 2020.

[32] E. Dupont, A. Doucet, and Y. Teh. Augmented neural odes. *Advances in neural information processing systems*, 32, 2019.

[33] C. Esteve-Yagüe, R. Tsai, and A. Massucco. Finite-difference least square methods for solving Hamilton-Jacobi equations using neural networks. *Journal of Computational Physics*, 524:113721, 2025.

[34] Q. Fan, H. Huang, and R. He. Breaking the low rank dilemma of linear attention. In *Computer Vision and Pattern Recognition (CVPR)*, 2025.

[35] K. Frans, D. Hafner, S. Levine, and P. Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, 2025.

[36] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. OPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

[37] Y. Gao, Q. Lang, and F. Lu. Self-test loss functions for learning weak-form operators and gradient flows. *arXiv preprint arXiv:2412.03506*, 2024.

[38] Z. Geng, M. Deng, X. Bai, J. Kolter, and K. He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

[39] Z. Geng, A. Pokle, and J. Kolter. One-step diffusion distillation via deep equilibrium models. In *Neural Information Processing Systems*, 2024.

[40] V. Gligorijević and et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

[41] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

[42] D. Han, X. Pan, Y. Han, S. Song, and G. Huang. Flatten transformer: Vision transformer using focused linear attention. In *International Conference on Computer Vision*, 2023.

[43] D. Han, Z. Wang, Z. Xia, Y. Han, Y. Pu, C. Ge, J. Song, S. Song, B. Zheng, and G. Huang. Demystify mamba in vision: A linear attention perspective. In *Neural Information Processing Systems*, 2024.

[44] G. Hinton. Neural networks for machine learning. Coursera, video lectures, 2012.

[45] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[46] Y. Hua, K. Miller, A. Bertozzi, C. Qian, and B. Wang. Efficient and reliable overlay networks for decentralized federated learning. *SIAM Journal on Applied Mathematics*, 82(4):1558–1586, 2022.

[47] Y. Hua, J. Pang, X. Zhang, Y. Liu, X. Shi, B. Wang, Y. Liu, and C. Qian. Towards practical overlay networks for decentralized federated learning. In *2024 IEEE 32nd International Conference on Network Protocols (ICNP)*, pages 1–11. IEEE, 2024.

[48] H. Huang and R. Lai. Unsupervised solution operator learning for mean-field games. *Journal of Computational Physics*, page 114057, 2025.

[49] S. Huang, Y. Wen, T. Adusumilli, K. Choudhary, and H. Yang. Parsing the language of expression: Enhancing symbolic regression with domain-aware symbolic priors. *arXiv preprint arXiv:2503.09592*, 2025.

[50] Y. Huang, T. Transue, S-H. Wang, W. Feldman, H. Zhang, and B. Wang. Improving flow matching by aligning flow divergence. In *Forty-second International Conference on Machine Learning*, 2025.

[51] Y. Huang, Q. Wang, A. Onwunta, and B. Wang. Efficient score matching with deep equilibrium layers. In *International Conference on Learning Representations*, 2024.

[52] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *Journal of Machine Learning Research*, 18(1):6869–6898, 2016.

[53] H. Jeong, J. Xin, and P. Yin. Beyond discreteness: Finite-sample analysis of straight-through estimator for quantization. *arXiv preprint arXiv:2505.18113*, 2025.

[54] Y. Jin, S. Liu, H. Wu, X. Ye, and H. Zhou. Parameterized Wasserstein gradient flow. *Journal of Computational Physics*, 524:113660, 2025.

[55] V. Karkaria, D. Lee, Y. Chen, Y. Yu, and W. Chen. Asno: An interpretable attention-based spatio-temporal neural operator for robust scientific machine learning. In *ICML Workshop on Reliable and Responsible Foundation Models*, 2025.

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

[57] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, 2020.

[58] N. Kitaev, L. Kaiser, and A. Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

[59] S. Kolek, D. Nguyen, R. Levie, J. Bruna, and G. Kutyniok. A rate-distortion framework for explaining black-box model decisions. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 91–115. Springer International Publishing, 2022.

[60] T. Kroells, E. Schmidt, C. Qiu, and J. Yan. Training data studies for the cell-average-based neural network method for linear hyperbolic and parabolic equations. *Beijing Journal of Pure and Applied Mathematics*, 2(1):147–181, 2025.

[61] G. Kutyniok. How can reliability of artificial intelligence be ensured? *Harvard Data Science Review*, 7, 2025.

[62] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider. A theoretical analysis of deep neural networks and parametric PDEs. *Constructive Approximation*, 55:73–125, 2022.

[63] S. Lam, J. Sirignano, and K. Spiliopoulos. Convergence analysis of real-time recurrent learning (rtrl) for a class of recurrent neural networks. *arXiv preprint arXiv:2501.08040*, 2025.

[64] L. Li, R. Carver, I. Lopez-Gomez, F. Sha, and J. Anderson. Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, 10(13):eadk4489, 2024.

[65] Q. Li, M. Oprea, L. Wang, and Y. Yang. Inverse problems over probability measure space. *arXiv preprint arXiv:2504.18999*, 2025.

[66] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

[67] Z. Li, B. Yang, P. Yin, Y. Qi, and J. Xin. Feature affinity assisted knowledge distillation and quantization of deep neural networks on label-free data. *IEEE Access*, 11:78042–78051, 2023.

[68] Z. Liang, B. Wang, Q. Gu, S. Osher, and Y. Yao. Differentially private federated learning with Laplacian smoothing. *Applied and Computational Harmonic Analysis*, 72:101660, 2024.

[69] Y. Lipman, R. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[70] X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

[71] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu. Vmamba: Visual state space model. In *Neural Information Processing Systems*, 2024.

[72] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[73] Z. Long, P. Yin, and J. Xin. Recurrence of optimum for training weight and activation quantized networks. *Applied and Computational Harmonic Analysis*, 62:41–65, 2023.

[74] C. Louizos, M. Welling, and D. Kingma. Learning sparse neural networks through ℓ_0 regularization. In *International Conference on Learning Representations*, 2018.

[75] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

[76] E. Lybrand and R. Saab. A greedy algorithm for quantizing neural networks. *Journal of Machine Learning Research*, 22(156):1–38, 2021.

[77] B. McMahan and et al. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[78] A. Mollaali, G. Zufferey, G. Constante-Flores, C. Moya, C. Li, M. Yue, and G. Lin. Conformalized prediction of post-fault voltage trajectories using pre-trained and finetuned attention-driven neural operators. *Neural Networks*, page 107809, 2025.

[79] C. Mooney, Z. Wang, J. Xin, and Y. Yu. Global well-posedness and convergence analysis of score-based generative models via sharp Lipschitz estimates. In *International Conference on Learning Representations*, 2025.

[80] M. Nagel, R. A. Amjad, M. v. Baalen, C. Louizos, and T. Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pages 7197–7206, 2020.

[81] T. Nguyen, R. Baraniuk, A. Bertozzi, S. Osher, and B. Wang. Momentumrnn: Integrating momentum into recurrent neural networks. *Advances in neural information processing systems*, 33:1924–1936, 2020.

[82] T. Nguyen, R. Baraniuk, R. Kirby, S. Osher, and B. Wang. Momentum transformer: Closing the performance gap between self-attention and its linearization. In *Mathematical and Scientific Machine Learning*, pages 189–204. PMLR, 2022.

[83] T. Nguyen, M. Pham, T. Nguyen, K. Nguyen, S. Osher, and N. Ho. Fourierformer: Transformer meets generalized Fourier integral theorem. In *Advances in Neural Information Processing Systems*, 2022.

[84] T. Nguyen, V. Suliafu, S. Osher, L. Chen, and B. Wang. Fmmformer: Efficient and flexible transformer via decomposed near-field and far-field attention. *Advances in neural information processing systems*, 34:29449–29463, 2021.

[85] M. Pagliardini, D. Paliotta, M. Jaggi, and F. Fleuret. Fast attention over long sequences with dynamic sparse flash attention. In *Advances in Neural Information Processing Systems*, volume 36, pages 59808–59831, 2023.

[86] Y. Park and S. Osher. Neural implicit solution formula for efficiently solving Hamilton-Jacobi equations. *arXiv preprint arXiv:2501.19351*, 2025.

[87] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr, 2013.

[88] K. Ren and L. Zhang. A model-consistent data-driven computational strategy for PDE joint inversion problems. *arXiv preprint arXiv:2210.09228*, 2022.

[89] Y. Ren, H. Chen, Y. Zhu, W. Guo, Y. Chen, G. Rotskoff, M. Tao, and L. Ying. Fast solvers for discrete diffusion models: Theory and applications of high-order algorithms. *arXiv preprint arXiv:2502.00234*, 2025.

[90] K. Riedl, J. Sirignano, and K. Spiliopoulos. Global convergence of adjoint-optimized neural PDEs. *arXiv preprint arXiv:2506.13633*, 2025.

[91] K. Rojas, Y. Zhu, S. Zhu, F. Ye, and M. Tao. Diffuse everything: Multimodal diffusion models on arbitrary state spaces. In *International Conference on Machine Learning*, 2025.

[92] S. Schotthöfer, M. Laiu, M. Frank, and C. Hauck. Structure-preserving neural networks for the regularized entropy-based closure of a linear, kinetic, radiative transport equation. *Journal of Computational Physics*, 533:113967, 2025.

[93] Z. Shao, K. Pieper, and X. Tian. Solving nonlinear PDEs with sparse radial basis function networks. *arXiv preprint arXiv:2505.07765*, 2025.

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

[95] J. Siegel and J. Xu. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *Foundations of Computational Mathematics*, 24(2):481–537, 2024.

[96] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

[97] Y. Song and P. Dhariwal. Improved techniques for training consistency models. In *The Twelfth International Conference on Learning Representations*, 2024.

[98] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

[99] Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[100] T. Sun, D. Li, and B. Wang. Adaptive random walk gradient descent for decentralized optimization. In *International conference on machine learning*, pages 20790–20809. PMLR, 2022.

[101] T. Sun, D. Li, and B. Wang. Decentralized federated averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4289–4301, 2022.

[102] T. Sun, D. Li, and B. Wang. On the decentralized stochastic gradient descent with markov chain sampling. *IEEE Transactions on Signal Processing*, 71:2895–2909, 2023.

[103] M. Thorpe, T. Nguyen, H. Xia, T. Strohmer, A. Bertozzi, S. Osher, and B. Wang. GRAND++: Graph neural diffusion with a source term. In *International Conference on Learning Representations*, 2022.

[104] J. Torge, C. Harris, S. Mathis, and P. Lio. DiffHopp: A graph diffusion model for novel drug design via scaffold hopping. *ICML Workshop on Comput. Biology*, 2023.

[105] H. Tran, M. Stoyanov, C. Hauck, and S. Smolentsev. Surrogate modeling for MHD flows in liquid metal fusion blankets: Initial assessment with analytical solutions. *Fusion Engineering and Design*, 218:115057, 2025.

[106] N. T. Tran and J. Xin. Fourier-mixed window attention for efficient and robust long sequence time-series forecasting. *Frontiers in Applied Math and Statistics*, 11, 2025.

[107] N.T. Tran, J. Xin, and G. Zhou. FWin transformer for dengue prediction under climate and ocean influence. In *International Conference on Machine Learning, Optimization, and Data Science*, volume 15509 of *Lecture Notes in Computer Science*, pages 160–175, 2025.

[108] N.T. Tran, F. Xue, S. Zhang, J. Lyu, Y. Zheng, Y-Y Qi, and J. Xin. SEMA: a scalable and efficient mamba like attention via token localization and averaging. *arXiv preprint arXiv:2506.08297*, 2025.

[109] T. Transue and B. Wang. Learning decentralized swarms using rotation equivariant graph neural networks. *arXiv preprint arXiv:2502.17612*, 2025.

[110] W. van den Dool, T. Blankevoort, M. Welling, and Y. Asano. Efficient neural PDE-solvers using quantization aware training. *arXiv preprint arXiv:2308.07350*, 2023.

[111] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[112] P. Velickovic, C. Perivolaropoulos, F. Barbero, and R. Pascanu. Softmax is not enough (for sharp out-of-distribution). In *International Conference on Machine Learning*, 2025.

[113] B. Wang, Q. Gu, M. Boedihardjo, L. Wang, F. Barekat, and S. Osher. DP-LSSGD: A stochastic optimization method to lift the utility in privacy-preserving ERM. In *Mathematical and Scientific Machine Learning*, pages 328–351. PMLR, 2020.

[114] B. Wang, H. Xia, T. Nguyen, and S. Osher. How does momentum benefit deep neural networks architecture design? a few case studies. *Research in the Mathematical Sciences*, 9(3):57, 2022.

[115] S-H Wang, J. Baker, C. Hauck, and B. Wang. Learning to control the smoothness of graph convolutional network features. *arXiv preprint arXiv:2410.14604*, 2024.

[116] S-H. Wang, Y-C. Hsu, J. Baker, A. Bertozzi, J. Xin, and B. Wang. Rethinking the benefits of steerable features in 3D equivariant graph neural networks. In *The Twelfth International Conference on Learning Representations*, 2024.

[117] S-H. Wang, Y. Huang, J. Baker, Y. Sun, Q. Tang, and B. Wang. A theoretically-principled sparse, connected, and rigid graph representation of molecules. In *International Conference on Learning Representations*, 2025.

[118] X. Wang and Z. Wang. Wasserstein bounds for generative diffusion models with Gaussian tail targets. *arXiv preprint arXiv:2412.11251*, 2024.

[119] Y. Wang, Y. He, and M. Tao. Evaluating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 37:19307–19352, 2024.

[120] Z. Wang, J. Xin, and Z. Zhang. Deepparticle: learning invariant measure by a deep neural network minimizing Wasserstein distance on data generated by an interacting particle method. *Journal of Computational Physics*, 464:111309, 2022.

[121] Z. Wang, J. Xin, and Z. Zhang. A deepparticle method for learning and generating aggregation patterns in multi-dimensional Keller-Segel chemotaxis systems. *Physica D: Nonlinear Phenomena*, 460:134082, 2024.

[122] D. Williams-King, L. Le, A. Oberman, and Y. Bengio. Can safety fine-tuning be more principled? lessons learned from cybersecurity. *arXiv preprint arXiv:2501.11183*, 2025.

[123] N. Winovich, M. Daneker, L. Lu, and G. Lin. Active operator learning with predictive uncertainty quantification for partial differential equations. *arXiv preprint arXiv:2503.03178*, 2025.

[124] H. Wu, S. Liu, X. Ye, and H. Zhou. Parameterized Wasserstein Hamiltonian flow. *SIAM Journal on Numerical Analysis*, 63(1):360–395, 2025.

[125] H. Xia, V. Suliafu, H. Ji, T. Nguyen, A. Bertozzi, S. Osher, and B. Wang. Heavy ball neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 34:18646–18659, 2021.

[126] L. Yang, S. Liu, T. Meng, and S. Osher. In-context operator learning with data prompts for differential equation problems. *Proceedings of the National Academy of Sciences*, 120(39):e2310142120, 2023.

[127] Y. Yang, Q. Chen, and W. Hao. Homotopy relaxation training algorithms for infinite-width two-layer relu neural networks. *Journal of Scientific Computing*, 102(2):40, 2025.

[128] Y. Yang, D. Silantyev, and R. Caflisch. Adjoint DSMC for nonlinear spatially-homogeneous Boltzmann equation with a general collision model. *Journal of Computational Physics*, 488:112247, 2023.

[129] P. Yin, J. Lyu, S. Zhang, S. Osher, Y. Qi, and J. Xin. Understanding straight-through estimator in training activation quantized neural nets. In *International Conf. on Learning Representations*, 2019.

[130] P. Yin, S. Zhang, J. Xin, and Y. Qi. Training ternary neural networks with exact proximal operator. *arXiv preprint arXiv:1612.06052*, 2016.

[131] K. Yoo, O. Oertell, Junhyun Lee, Sanghoon Lee, and Jaewoo Kang. TurboHopp: Accelerated molecule scaffold hopping with consistency models. *Advances in Neural Information Processing Systems*, 37:41157–41185, 2024.

[132] L. Yu, Y. Lou, and F. Chen. Uncertainty-aware graph-based hyperspectral image classification. In *International Conference on Learning Representations*, 2024.

[133] Y. Yu, N. Liu, F. Lu, T. Gao, S. Jafarzadeh, and S. Silling. Nonlocal attention operator: Materializing hidden knowledge towards interpretable physics discovery. *Advances in Neural Information Processing Systems*, 37:113797–113822, 2024.

[134] A. Zhang, Z. Yang, N. Wang, Y. Qi, J. Xin, X. Li, and P. Yin. COMQ: A backpropagation free algorithm for post-training quantization. *IEEE Access*, 13:99879–99891, 2025.

[135] J. Zhang, Y. Zhou, and R. Saab. Post-training quantization for neural networks with provable guarantees. *SIAM Journal on Mathematics of Data Science*, 5(2):373–399, 2023.

[136] S. Zhang, H. Zhang, I. Colbert, and R. Saab. Qronos: Correcting the past by shaping the future...in post-training quantization. *arXiv preprint arXiv:2505.11695*, 2025.

[137] S. Zhang, H. Zhao, Y. Zhong, and H. Zhou. Why shallow networks struggle with approximating and learning high frequency: A numerical study. *arXiv preprint arXiv:2306.17301*, 2023.

[138] S. Zhang, H. Zhao, Y. Zhong, and H. Zhou. Fourier multi-component and multi-layer neural networks: Unlocking high-frequency potential. *arXiv preprint arXiv:2502.18959*, 2025.

[139] T. Zhang, Z. Wang, J. Xin, and Z. Zhang. A bidirectional deepparticle method for efficiently solving low-dimensional transport map problems. *arXiv preprint arXiv:2504.11851*, 2025.

[140] H. Zhao and J. Xu. Convergence analysis and trajectory comparison of gradient descent for overparameterized deep linear networks. *Transactions on Machine Learning Research*, 2024.

[141] Y. Zheng, C. Hu, G. Lin, M. Yue, B. Wang, and J. Xin. Glassoformer: a query-sparse transformer for post-fault power grid voltage prediction. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3968–3972, 2022.

[142] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of AAAI*, 2021.

[143] M. Zhou, S. Osher, and W. Li. A deep learning algorithm for computing mean field control problems via forward-backward score dynamics. *Research in the Mathematical Sciences*, 12(3):42, 2025.

[144] W. Zhu, W. Khademi, E. Charalampidis, and P. Kevrekidis. Neural networks enforcing physical symmetries in nonlinear dynamical lattices: The case example of the Ablowitz–Ladik model. *Physica D: Nonlinear Phenomena*, 434:133264, 2022.