Toward Data-Efficient Intelligence: From Few-Shot Learning to Agentic Systems

Yaqing Wang
Beijing Institute of Mathematical Sciences and Applications
wangyaqing@bimsa.cn

Abstract

Modern artificial intelligence has made rapid progress by scaling data, models, and computation. Yet scale alone does not solve one basic problem. Many intelligent systems must learn, adapt, and act when direct experience is scarce, costly, noisy, or changing. A scientist may have only a few labeled molecules. A recommender system may see only sparse interactions for a new user or item. A language-model agent may need to align with a user's preference from a short interaction history. These settings differ, but they share the same question: how can AI extract more learning signal from less experience? My research studies this question through data-efficient generalization. This article summarizes a research path from few-shot learning to meta-learning, in-context learning, and data-efficient agentic systems. The central theme is that limited supervision can be made useful by the right priors, the right adaptation mechanism, and the right way to reuse experience. I also discuss two regimes where data efficiency is not only desirable but necessary: scientific scarcity and efficient data utilization.

1 Introduction

Modern AI is often described as a story of scale. Larger datasets, larger models, and larger compute budgets have led to stronger systems. This story is real. It explains much of the progress in vision, language, recommendation, and scientific modeling. But it is incomplete.

Many important settings are not data-rich at the point where a decision must be made. A new molecular property may have only a few measured examples. A new item in a recommender system may have little interaction history. A user may reveal preferences slowly and indirectly. An autonomous agent may receive only sparse feedback from tool use or the environment. In all these cases, the learner cannot rely only on fitting the target data. It must use prior knowledge, structure, and related experience.

This question has guided my research. I study *data-efficient generalization*: how learning systems can generalize, adapt, and make decisions under bounded supervision. The bounded resource may be labeled examples, expert annotations, demonstrations, reasoning traces, user feedback, tool-use records, memory, or real-world trials. The surface forms differ. The core difficulty is the same. Direct evidence for the target task is too small, so the learner must decide what old knowledge should transfer and what new information should change the model.

This article follows my research path through four stages[1]. The first stage is few-shot learning, which gives a clean formulation of the problem. It asks how a system can generalize from only a few labeled examples. Our survey organized this field by asking how prior knowledge enters data, models, and algorithms [Wang *et al.*, 2020b]. The second stage is meta-learning. It learns across tasks so that adaptation to a new task becomes faster and more stable. The third stage is in-context learning. Large language models can infer task behavior from examples in the prompt, without parameter updates [Brown *et al.*, 2020]. Reasoning prompts further show that intermediate traces can act as useful test-time experience [Wei *et al.*, 2022]. This changes where adaptation happens. The fourth stage is agentic learning. When AI systems act through tools, memory, and interaction, data efficiency expands beyond labels [Yao *et al.*, 2023]. Our recent survey on data-efficient agentic learning [Wang *et al.*, 2026] develops this view for the agentic era. It treats demonstrations, feedback, reasoning traces, tool-use records, memory, and trajectories as forms of agentic data. It asks how agents can acquire, structure, reuse, and learn from limited experience. In this stage, data-efficient generalization becomes a question of how agents improve under bounded supervision and costly interaction.

The main message is simple. Data-efficient intelligence is not the opposite of scale. It is a complement to scale. Scaling gives expressive models. Data efficiency tells us how to use priors, structures, and interactions so that these models can adapt when experience is limited.

2 Few-Shot Learning as a Core Problem

Few-shot learning is a natural starting point. It makes the small-data challenge explicit. A learner receives only a few labeled examples for the target task, but still needs to make

[1]This article further develops the perspective in my AAAI New Faculty Highlights short abstract [Wang, 2026].

reliable predictions. This setting appears in image recognition, drug discovery, query understanding, acoustic gesture recognition, and many other domains. It also connects to a broader goal: learning more like humans, who often grasp new concepts from limited experience.

Our survey on few-shot learning argued that the central issue is not merely the small number of samples [Wang *et al.*, 2020b]. The central issue is that empirical risk minimization becomes unreliable. With few examples, a large hypothesis space can fit the observations but generalize poorly. A small hypothesis space may generalize better but fail to represent the target function. This creates a tradeoff between approximation and estimation. Prior knowledge is needed to make the search feasible.

This view leads to a simple taxonomy. Prior knowledge can enter through data, models, or algorithms. Data-level methods enrich the limited supervised experience. They use unlabeled data, weak supervision, generated samples, or external knowledge. Model-level methods restrict or structure the hypothesis space. They use architectures, regularizers, domain knowledge, or pretrained representations. Algorithm-level methods change how the learner searches for a solution. Meta-learning is a central example. It learns an adaptation strategy from related tasks.

This taxonomy shaped much of my later work. In early structured-learning problems, we studied how shared components can be reused while allowing sample-dependent adaptation, and how robust losses can handle unknown noise. Examples include scalable and generalized convolutional sparse coding and sample-dependent dictionaries [Wang *et al.*, 2018b; Wang *et al.*, 2018a; Wang *et al.*, 2020a]. In low-rank matrix and tensor learning, we used structural assumptions and efficient nonconvex optimization to recover information from partial observations [Wang *et al.*, 2021c; Yao *et al.*, 2022]. Similar ideas also appear in generative time-series forecasting, where diffusion, denoising, and disentanglement help learn from short and noisy series [Li *et al.*, 2022]. These works were not always framed as few-shot learning. But they follow the same principle. When direct observations are limited or costly, useful structure can reduce the burden on data.

Few-shot learning therefore remains more than a benchmark family. It is a core learning problem. It asks what should be reused, what should be adapted, and how a model should decide which prior is relevant to a new task.

3 Learning to Adapt

Meta-learning addresses few-shot learning across tasks. Instead of learning each task from scratch, a meta-learner observes many tasks during training. It learns how to adapt from a small support set and is evaluated on query examples from the same task. At test time, the same learned adaptation strategy is applied to unseen tasks.

A useful way to view meta-learning is through three families. Optimization-based methods learn initializations or update rules for fast adaptation [Finn *et al.*, 2017]. Metric-based methods learn how to compare support and query examples [Vinyals *et al.*, 2016]. Amortization-based methods map a support set directly to task-specific parameters or modulations [Requeima *et al.*, 2019]. These families differ in form, but they share the same purpose. They convert experience across tasks into a prior for a new task.

Our work on molecular property prediction illustrates why this matters. Drug discovery often faces severe label scarcity. Wet-lab experiments are expensive, risky, and slow. Moreover, the same molecule may behave differently for different properties. In Property-Aware Relation networks (PAR), we used a property-aware encoder and a query-dependent relation graph to adapt molecular representations to the target property [Wang *et al.*, 2021a; Yao *et al.*, 2024]. The relation graph helps information propagate among molecules that are similar with respect to the current property. The meta-learning procedure separates generic molecular knowledge from property-specific adaptation.

Later, PACIA further focused on parameter-efficient adaptation for few-shot molecular property prediction [Wu *et al.*, 2024]. Instead of updating many model parameters for each task, it uses a GNN adapter to generate a small set of adaptive parameters. The adapter modulates message passing in both the encoder and predictor. This reduces overfitting and makes adaptation more efficient. These methods follow a common design rule: adapt the part that should change, and protect the part that should remain shared.

Task identity itself can also be useful. In episodic meta-training, each batch contains tasks. This gives a natural source of supervision. ConML uses this signal to construct a contrastive meta-objective [Wu *et al.*, 2025a]. Models learned from different subsets of the same task should be aligned. Models learned from different tasks should be distinguishable. This encourages a meta-learner to be stable within a task and discriminative across tasks. It is a simple idea, but it shows an important point: data efficiency can improve when we use signals already present in the learning process.

In-context learning (ICL) changes the form of adaptation. A large language model receives task examples in the prompt and predicts the output for a query. No parameter update is required. At first sight, this looks different from meta-learning. Yet the connection is strong. ICL can be viewed as learning to learn inside the forward pass of a transformer [Wu *et al.*, 2025b]. The model uses the context as a task-specific dataset. The prompt becomes the support set.

Our theoretical work studies this connection by comparing ICL models with classical meta-learners [Wu *et al.*, 2025b]. We show that transformer-based ICL can represent gradient-based, metric-based, and amortization-based meta-learning algorithms. This suggests that ICL has a broad hypothesis space over learning algorithms. It can learn data-dependent adaptation strategies rather than merely imitate a fixed classical learner. This helps explain why ICL can be powerful. It also reveals a risk. The learned algorithm is shaped by pretraining data, so its generalization is distribution-sensitive.

This view also motivates practical methods. If ICL is a data-dependent learning process, then examples in the context are not just demonstrations. They are the data from which the model infers a task. Their selection matters. GraphIC retrieves in-context examples for multi-step reasoning by modeling reasoning processes as thought graphs, rather than rely-

ing only on text similarity [Fu *et al.*, 2026]. RD-MCSA uses classification rationales and adaptive demonstration selection for multi-class sentiment analysis [Xie *et al.*, 2025]. Both works follow the same principle. To make ICL data-efficient, we should improve the quality of the limited context.

4 Learning from Interaction

As AI systems become agents, the meaning of data efficiency expands. An agent does not only classify examples. It perceives, reasons, calls tools, receives feedback, stores memory, and acts again. The available data are no longer only labeled pairs. They include demonstrations, trajectories, reasoning traces, tool outputs, preferences, reflections, and failures. We call these signals agentic experience [Wang *et al.*, 2026].

This setting raises new questions. Which experiences should be stored? Which should be retrieved? How should an agent distinguish stable preferences from temporary intents? How can it improve without asking for too much feedback? These questions are central because interaction is costly. Tool calls may fail. Human feedback is limited. Real-world trials may be risky.

AdaPA-Agent studies this issue for personalized LLM agents [Nie *et al.*, 2025]. Users often have multiple preferences at the same time. The contents of these preferences may be stable, but their strengths change with goals and context. AdaPA-Agent estimates dynamic preference strengths from existing user-agent interaction, without requiring extra explicit feedback. It then uses adaptive preference arithmetic to guide generation by combining next-token distributions conditioned on different preferences. The goal is to personalize from minimal interaction while retaining controllability.

Another direction is communication among language models. Many agent systems use natural-language messages between modules or agents. This is convenient, but it is discrete, lossy, and hard to optimize end to end. LMNet treats pretrained language models as reusable nodes in a network and lets them communicate through dense vectors [Wu *et al.*, 2026a]. Trainable edge modules carry information between stripped language-model nodes. This makes internal communication differentiable and learnable from final task supervision. The broader idea is that agentic systems should not only prompt better. They should organize, reuse, and optimize limited experience more efficiently.

Agentic learning brings the few-shot question back in a new form. The support set is no longer a fixed set of labeled examples. It is a stream of interaction. The prior is no longer only a pretrained model. It also includes memory, tools, preferences, workflows, and communication structure. The challenge is to decide what to reuse and what to update at each step.

5 Where Data Efficiency Matters

Data-efficient learning matters most when direct supervision is hard to obtain or hard to use. Two regimes have been central in my work: scientific scarcity, where data are expensive to obtain, and efficient data utilization, where data may be abundant globally but sparse, noisy, or dynamic for the target case.

Scientific scarcity is direct. In AI for science, high-quality labels often require experiments, simulations, or expert validation. In drug discovery, molecular property labels are scarce. PAR and PACIA address this by making molecular representations property-aware and adaptation parameter-efficient [Wang *et al.*, 2021a; Wu *et al.*, 2024]. KnowDDI studies drug-drug interaction prediction, where known interactions are rare and biomedical knowledge graphs provide useful prior information [Wang *et al.*, 2024b]. It learns knowledge subgraphs that improve prediction and also provide explaining paths. The goal is not only to improve accuracy, but also to make the prediction interpretable enough to support biomedical use.

Scientific scarcity also appears in physical simulation. Neural PDE solvers can approximate solution operators, but high-fidelity trajectories are expensive. DGNet uses Green's function theory as a structural prior for spatiotemporal PDE learning [Tan *et al.*, 2026]. By embedding the superposition principle into a graph-based discrete formulation, the model reduces the need to relearn physical structure from data. This helps data-efficient learning and supports generalization to unseen source terms. The lesson is general. If theory already tells us part of the structure, the model should not spend scarce data rediscovering it.

Personalization faces a different form of scarcity. Data may be abundant globally, but sparse locally. A recommender system may have many interactions overall, but a new user, item, scenario, or moment may still be data-poor. Cold-start recommendation captures this problem. ColdNAS searches for modulation structures for user cold-start recommendation [Wu *et al.*, 2023; Wu *et al.*, 2026b]. EmerG learns item-specific feature interactions for cold-start CTR prediction [Wang *et al.*, 2024a]. PERSCEN models user-specific interaction patterns and scenario-aware preferences for multi-scenario matching [Du *et al.*, 2025]. These methods ask which global patterns transfer and how local sparse signals should modulate the model.

Search and short text understanding show the same pattern. Medical search queries are short, noisy, and domain-specific. MEDIC uses medical knowledge graphs, syntactic information, pretrained language models, and co-click weak supervision to recognize fine-grained query intents with few labels [Wang *et al.*, 2022]. SHINE and SimpleSTC use graph learning to enrich short text representations and propagate sparse label information [Wang *et al.*, 2021b; Zheng *et al.*, 2022]. Robust heterogeneous graph learning follows a related idea: MDCL structures training with multi-faceted curricula to handle noisy features, edges, and labels [Wong *et al.*, 2026]. The key is not to add complexity for its own sake. The key is to provide the missing context that short texts and sparse labels cannot provide alone.

Across these applications, the common structure is clear. The target situation is locally data-poor. Prior knowledge is available but must be selected and adapted. The method must be statistically efficient and often computationally efficient. Data-efficient intelligence therefore connects learning theory, model design, and deployment constraints.

6 Unifying View and Outlook

The examples above look diverse. Some use molecular graphs. Some use language models. Some use knowledge graphs, feature graphs, PDE operators, or user histories. Yet they share a simple pattern. A learner faces a new target with too little direct evidence. It must reuse old knowledge, but not reuse it blindly.

This pattern leads to three design questions. First, where does the prior come from? The source of the prior should match the reason that data are scarce. Second, how is the prior used? A relation graph, an adapter, a contrastive objective, a retrieval rule, or a dense communication edge can all be a data-efficient mechanism if it reduces the search burden. This is also close to knowledge-aware parsimony: structure should reduce unnecessary learning burden rather than only increase model size [Yao *et al.*, 2025]. Third, what is the adaptation budget? Some settings allow gradient updates. Some allow only a prompt, a few milliseconds, or implicit feedback. The method must fit this budget.

These observations shape my future research. One direction is data-efficient AI for science. Scientific problems often contain formulas, operators, symmetries, conservation laws, and graph structures. Models should use this structure as part of their design. They should learn unknown components from data while respecting known principles. This direction is especially natural at BIMSA, where mathematical structure and machine learning can be studied together.

A second direction is data-efficient agents with human priors. Future agents must adapt to users, tasks, and environments from sparse interaction. They need memory, retrieval, preference modeling, tool use, and communication mechanisms that are reliable rather than ad hoc. They should ask for feedback only when needed and use past experience responsibly. This is important for personalization, but also for trust.

A third direction is a unified theory and evaluation protocol for data-efficient generalization. Few-shot learning, meta-learning, in-context learning, and agentic learning are often studied separately. But they share the same structure: limited direct evidence, prior knowledge, and adaptation under budget. A useful theory should explain when a prior reduces estimation error, when it introduces harmful bias, and how different signals should be combined. A useful evaluation should also count annotation cost, feedback burden, inference cost, latency, safety risk, and the cost of wrong adaptation, especially for interactive agents and GUI-agent benchmarks [Chen *et al.*, 2026].

The long-term goal is an AI system that behaves less like a static predictor and more like a reliable learner. It should know what it has seen, recognize what it has not seen, reuse prior knowledge carefully, and improve from each new piece of experience. This is the path from few-shot learning to data-efficient intelligence.

Acknowledgments

This work is sponsored by Beijing Nova Program.

References

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[Chen *et al.*, 2026] Yihong Chen, Shuai Wang, Yaqing Wang, and Quanming Yao. A survey on benchmarks of LLM-based GUI agents. *Transactions on Machine Learning Research (TMLR)*, 2026.

[Du *et al.*, 2025] Haotong Du, Yaqing Wang, Fei Xiong, Lei Shao, Ming Liu, Hao Gu, Quanming Yao, and Zhen Wang. PERSCEN: Learning personalized interaction pattern and scenario preference for multi-scenario matching. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

[Fu *et al.*, 2026] Jiale Fu, Yaqing Wang, Simeng Han, Jiaming Fan, and Xu Yang. GraphIC: A graph-based in-context example retrieval model for multi-step reasoning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[Li *et al.*, 2022] Yan Li, Xinjiang Lu, Yaqing Wang, and Dejing Dou. Generative time series forecasting with diffusion, denoise, and disentanglement. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[Nie *et al.*, 2025] Hongyi Nie, Yaqing Wang, Mingyang Zhou, Feiyang Pan, Quanming Yao, and Zhen Wang. Adaptive preference arithmetic: Modeling dynamic preference strengths for LLM agent personalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[Requeima *et al.*, 2019] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[Tan *et al.*, 2026] Yingjie Tan, Quanming Yao, and Yaqing Wang. DGNet: Discrete green networks for data-efficient learning of spatiotemporal PDEs. In *International Conference on Learning Representations (ICLR)*, 2026.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[Wang *et al.*, 2018a] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Online convolutional sparse coding with sample-dependent dictionary. In *International Conference on Machine Learning (ICML)*, 2018.

[Wang *et al.*, 2018b] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Scalable online convolutional

sparse coding. *IEEE Transactions on Image Processing (TIP)*, 27(10):4850–4859, 2018.

[Wang *et al.*, 2020a] Yaqing Wang, James T. Kwok, and Lionel M. Ni. Generalized convolutional sparse coding with unknown noise. *IEEE Transactions on Image Processing (TIP)*, 29:5386–5395, 2020.

[Wang *et al.*, 2020b] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[Wang *et al.*, 2021a] Yaqing Wang, Abulikemu Abuduweili, Quanming Yao, and Dejing Dou. Property-aware relation networks for few-shot molecular property prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[Wang *et al.*, 2021b] Yaqing Wang, Song Wang, Quanming Yao, and Dejing Dou. Hierarchical heterogeneous graph representation learning for short text classification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[Wang *et al.*, 2021c] Yaqing Wang, Quanming Yao, and James T. Kwok. A scalable, adaptive and sound nonconvex regularizer for low-rank matrix learning. In *The Web Conference (TheWebConf)*, 2021.

[Wang *et al.*, 2022] Yaqing Wang, Song Wang, Yanyan Li, and Dejing Dou. Recognizing medical search query intent by few-shot learning. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2022.

[Wang *et al.*, 2024a] Yaqing Wang, Hongming Piao, Daxiang Dong, Quanming Yao, and Jingbo Zhou. Warming up cold-start CTR prediction by learning item-specific feature interactions. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

[Wang *et al.*, 2024b] Yaqing Wang, Zaifei Yang, and Quanming Yao. Accurate and interpretable drug-drug interaction prediction enabled by knowledge subgraph learning. *Communications Medicine (Commun Med)*, 4:59, 2024.

[Wang *et al.*, 2026] Yaqing Wang, Zhenlin Luo, Peiyao Zhao, Yunfeng Cai, and Quanming Yao. Beyond scaling: A survey of data-efficient learning for LLM agents. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2026.

[Wang, 2026] Yaqing Wang. From few-shot learning to data-efficient intelligence. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[Wong *et al.*, 2026] Zhen Hao Wong, Hansi Yang, Quanming Yao, and Yaqing Wang. Robust heterogeneous network representation learning by multifaceted curriculum training. *Neural Networks*, 196:108438, 2026.

[Wu *et al.*, 2023] Shiguang Wu, Yaqing Wang, Qinghe Jing, Daxiang Dong, Dejing Dou, and Quanming Yao. ColdNAS: Search to modulate for user cold-start recommendation. In *The Web Conference (TheWebConf)*, 2023.

[Wu *et al.*, 2024] Shiguang Wu, Yaqing Wang, and Quanming Yao. PACIA: Parameter-efficient adapter for few-shot molecular property prediction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[Wu *et al.*, 2025a] Shiguang Wu, Yaqing Wang, Yatao Bian, and Quanming Yao. Learning to learn with contrastive meta-objective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[Wu *et al.*, 2025b] Shiguang Wu, Yaqing Wang, and Quanming Yao. Why in-context learning models are good few-shot learners? In *International Conference on Learning Representations (ICLR)*, 2025.

[Wu *et al.*, 2026a] Shiguang Wu, Yaqing Wang, and Quanming Yao. Language model networks: Supervision-efficient learning through dense communication. In *International Conference on Machine Learning (ICML)*, 2026.

[Wu *et al.*, 2026b] Shiguang Wu, Yaqing Wang, and Quanming Yao. Searching to modulate for cold-start recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 48(5):5710–5724, 2026.

[Xie *et al.*, 2025] Haihua Xie, Yinzhu Cheng, Yaqing Wang, Miao He, and Mingming Sun. RD-MCSA: A multi-class sentiment analysis approach integrating in-context classification rationales and demonstrations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[Yao *et al.*, 2022] Quanming Yao, Yaqing Wang, Bo Han, and James T. Kwok. Low-rank tensor learning with nonconvex overlapped nuclear norm regularization. *Journal of Machine Learning Research (JMLR)*, 23(111):1–60, 2022.

[Yao *et al.*, 2023] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[Yao *et al.*, 2024] Quanming Yao, Zhenqian Shen, Yaqing Wang, and Dejing Dou. Property-aware relation networks for few-shot molecular property prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(8):5413–5429, 2024.

[Yao *et al.*, 2025] Quanming Yao, Yongqi Zhang, Yaqing Wang, Nan Yin, James T. Kwok, and Qiang Yang. Beyond scaleup: Knowledge-aware parsimony learning from deep networks. *AI Magazine*, 2025.

[Zheng *et al.*, 2022] Kaixin Zheng, Yaqing Wang, Quanming Yao, and Dejing Dou. Simplified graph learning for inductive short text classification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.