Reinforcement Learning for Ising Models: Datasets and Benchmark

Levy Lin[1], **Zhiyuan Wang**[1], **Holden Mac Entee**[1], **Xingjian Zhao**[1], **and Xiao-Yang Liu**[1,2]
[1]Rensselaer Polytechnic Institute, [2]Columbia University,
{linl9, wangz60,macenh,zhaox8}@rpi.edu, XL2427@columbia.edu

Abstract

Searching for the ground state of Ising models remains an century-old unsolved problem, crucial for its analysis of physical systems [27, 29, 30] and abstraction to combinatorial optimization problems [28, 37]. Although, due to its huge discrete space and rough or glassy optimization landscape, heuristic methods are computationally infeasible at large scale. Reinforcement learning (RL) algorithms provide a promising alternative for obtaining high-quality suboptimal minima. However, there is no established dataset to benchmark RL methods on Ising problems. In this paper, we curated a comprehensive dataset of over 190,000 Ising instances and a state-of-the-art (SOTA) benchmark of Mixed Integer Program (MIP) solvers and RL methods. Through our dataset and benchmark, we promote interdisciplinary physics applications with the ML community and encourage physicists to apply the expertise of the ML community to their problems. Furthermore, we propose a novel transformer-based policy framework to tackle large scale Ising model problems. Our experiments demonstrate SOTA-level effectiveness and scalability with around 1% - 9.11% gap to industry level solvers on large scale Ising problems. Datasets and benchmarks are open source at link.

1 Introduction

Ising models, an over 100-year-old standing problem in physics [19], are a cornerstone in chemistry, statistical mechanics, machine learning, and physics due to their versatility as a fundamental model for physical and theoretical systems. For example, in the study of magnetic materials, Ising models are commonly used to analyze critical behavior [26] and phase transitions [2, 15, 11], thus assisting our understanding of complex phenomena such as Ising superconductors and quantum phase transitions [6]. Spin-Glasses, a disordered variant of the Ising model, represent a class of high-dimensional random functions, where the study of minima within Spin-Glasses plays an important role in high-dimensional statistics, optimization [28], and machine learning [36]. For these applications, an efficient exploration of the Ising model space is crucial. However, this task of searching for the ground state is NP-hard, since Ising models contain a huge non-convex discrete space with scaling expo-

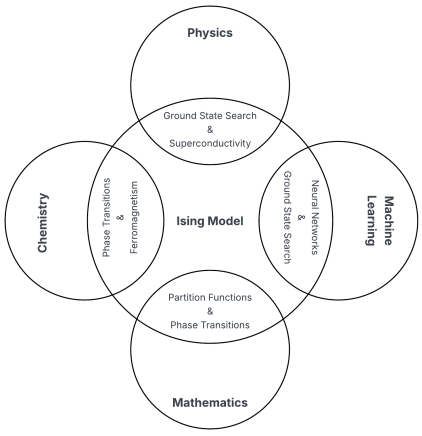

Figure 1: Interdisciplinary applications of Ising models.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

nential complexity [14, 20], rendering simulation and heuristic methods computationally infeasible in three dimensions or for vast two dimensional systems.

Although reinforcement learning frameworks or algorithms capable of obtaining high-quality minima have rapidly developed [17, 13, 32], the evaluation of these RL methods is difficult due to the lack of a standardized Ising model dataset and benchmark. In this paper, we curate a collective open-source dataset composed of high-interest Ising models, provide a baseline of MIP solvers, and a benchmark of SOTA RL methods. Our open-source baseline provides transparent and provable optimal, or near-optimal, solutions from industry-level solvers such as Gurobi and CPLEX. While our open-source benchmark of SOTA RL methods proposes an evaluation framework for measuring performance on Ising models. We hope to see progress made to our benchmark, expanding its utility in both ML and physics fields. Our dataset is available at link and our benchmark is on our website link.

Our aim is that the ML and physics communities will collaborate to maintain a benchmark for evaluating future improvements.

2 Problem Formulation

Ising models are formulated as a d-dimensional lattice, where each lattice is composed of N spins, $\boldsymbol{\sigma} \in \{-1, +1\}^N$, the interaction between spins, J_{ij}, and the external magnetic field, h_j. The Ising Hamiltonian (1) represents the energy of the system and is the objective when searching for the ground state (2), the spin configuration that minimizes the Hamiltonian.

$$H(\boldsymbol{\sigma}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_j h_j\sigma_j \tag{1}$$

$$\boldsymbol{\sigma}^* = \underset{\boldsymbol{\sigma}}{\operatorname{argmin}}\, H(\boldsymbol{\sigma}). \tag{2}$$

2.1 Markov Decision Process Modeling

We formulate the Ising optimization problems as Markov Decision Processes (MDPs).

- **State**. We define the state at timestep t as $s_t = [\boldsymbol{\sigma}_t, G]$, where $\boldsymbol{\sigma}_t \in \{-1, 1\}^N$ represents the current spin configuration and $G \in \mathbb{R}^{N\times N}$ is the weighted adjacency matrix of the input instance where $G_{ij} = J_{ij}$. The graph G remains static throughout the entire episode.
- **State Transition (Action)**. As there exists no natural 'action' for the Ising problem, a simple action formulation is to flip a fixed number of spins at each step. However, this introduces inductive bias and leads to the policy being more easily trapped in local minima. To avoid this, at each timestep t, we directly model the state transition $\delta(\boldsymbol{\sigma}_{t+1} \mid s_t)$, which defines the probability distribution over the next spin configuration $\boldsymbol{\sigma}_{t+1}$. We then sample from that distribution to obtain s_{t+1}.
- **Reward**. The reward is defined simply as the negative Hamiltonian:

$$r_t = -H(\boldsymbol{\sigma}_t)$$

- **Discount factor**. As typical in RL, we use a discount factor $\gamma \in (0, 1)$ to evaluate a discounted sum of rewards to encourage the policy to find the optimal solution faster.

A policy $\pi(\cdot \mid s_t)$ is a probability distribution over transitions in state s_t, which determines the probability of each spin being 1. The objective of the search agent is to learn a policy π that maximizes the cumulative discounted reward. We discuss how a policy gradient framework can be applied for this learning process in section 5.

3 Datasets for Ising Models

We curated a comprehensive open-source dataset of typical and prevalent Ising models used within the physics field to propose a standardized Ising model dataset, support benchmark evaluation of ML algorithms, and promote high-interest Ising model instances. All instances of Ising models in

our dataset are sourced from ML, RL, or Quantum-optimization papers with public datasets; references to each paper are shown in Table 1. In this section, we present how our dataset is organized, specifications on the types of Ising models, and their relevance to current research.

Since spins only interact with their nearest neighbor according to their dimensionality, we organized our dataset based on dimensionality, allowing for an intuitive distinction between different Ising models instances. The only exception is the ∞ dimension rank, which represents the set of Ising models where each spin interacts with every other spin regardless of spatial boundaries. Within each dimension, we provide an additional layer of distinction by labeling the type of the Ising model, distinguishing systems with unique spin alignments, interactions, or geometries. We label five main types of Ising models: Classic, Spin-Glass, Ferromagnetic, Anti-Ferromagnetic, and Synthetic. A Classic Ising model refers to a system with nearest-neighbor interactions and $J_{ij} \in \{-1, +1\}$ or $J_{ij} \in \{0, +1\}$. Spin-Glass Ising models refer to frustrated systems with random and disordered spin alignments and randomly distributed $J_{ij} \in \mathbb{R}$. Ferromagnetic and Anti-Ferromagnetic Ising instances represent magnetic systems with uniform and fixed spin alignment, where spins are either all aligned up or all down and $J_{ij} > 0$ or $J_{ij} < 0$. Synthetic Ising instances are similar to MaxCut instances, with an artificial complexity of extremely large or small J_{ij}.

Our dataset predominately features Spin-Glass instances and contains specific instances of high-interest such as Edwards-Anderson, Sherrington-Kirkpatrick, and Hopfield instances. Edwards-Anderson instances represent simple disordered, frustrated, or fragile systems and are studied for temperature chaos in a non-equilibrium context [3] and critical behavior across dimensions [35, 7]. The Sherrington-Kirkpatrick instances represent complex disordered and frustrated systems and are studied for their correlation to real spin-glass [41], non-equilibrium thermodynamics [1], and ground state [24]. Hopfield instances are being actively used to study neural networks [36] and memory retention in hardware [25].

The breakdown of our dataset is shown in Table 1.

Dimension	Type of Ising model	Instances	Spins	Couplings	Coupling Strength	Ref.
1D	Classic	75	32 – 128	31 – 127	0.00 – 1.00	[17]
	Spin Glass	4,000	3 – 10	6 – 55	-5.00 – 4.22	[23]
2D	Classic	100	16 – 64	28 – 120	-1.00 – 1.00	[21]
	Spin Glass	2,075	16 – 1,600	32 – 3,210	-5.09 – 4.66	[13, 17, 32]
	Ferromagnetic	90,000	1,600	83 – 1,246	-1.00	[31]
	Anti-Ferromagnetic	90,000	1,600	1874 – 3,037	1.00	[31]
	Synthetic	9	100 – 400	200 – 800	-294,541.00 – 375,001.00	[39]
3D	Spin Glass	666	24 – 8,000	38 – 40,279	-5.24 – 5.03	[13, 21, 42]
	Synthetic	9	125 – 343	375 – 1,029	-298,103.00 – 375,001.00	[39]
	Diamond	60	18 – 50	24 – 80	-1.00 – 1.00	[21]
4D	Spin Glass	1,450	81 – 4,096	324 – 16,384	-5.28 – 4.97	[13]
∞	Bi-clique	100	20 – 36	35 – 99	-2.00 – 0.89	[21]
	Spin Glass	5,440	3 – 900	6 – 319,600	-4.45 – 3.92	[12, 17, 23, 40]
	Synthetic	30	100 – 300	4,950 – 44,850	-246,443.00 – 280,065.00	[39]

Table 1: breakdown of collective dataset.

4 Benchmark Performance

In this section, we present our empirical results from the MIP baseline solvers and SOTA RL algorithms on Classic and Spin-Glass instances from our curated dataset.

4.1 Experimental Setup

Specifically, we select a set of simple 1D Classic Ising instances, to verify algorithmic correctness, and a set of challenging Spin-Glass instances to investigate the scalability performance of RL methods against increasing system size and dimensionality.

MIP solvers provide a provable upper bound, allowing a rigorous and simple evaluation of solution quality for RL algorithms. We utilize the following MIP solvers as baselines against our approximate RL benchmark algorithms.

- Gurobi [16]: a commercial MIP solver, which uses a branch-cut algorithm to search for the ground state of Ising models.

- IBM ILOG CPLEX [18]: a commercial MIP solver that uses a branch-cut algorithm to search for the ground state of Ising models.

We use the following SOTA ML benchmarking algorithms:

- REINFORCE: Our transformer-based policy framework integrates a graph transformer encoder to capture the Ising interaction structure and a seq2seq transformer policy that outputs full next-step spin configurations. Curriculum learning gradually increases the search horizon, while local search refinement help the policy escape local minima and converge quickly.
- Variational Neural Annealing [17]: A hybrid ML-Simulated annealing framework that utilizes auto-regressive RNNs to efficiently sample from the Boltzmann distribution and minimize variational free energy.
- MCPG [9]: An RL framework which combines Monte Carlo policy gradient methods with local search techniques to address binary combinatorial optimization problems, repurposed for the search for the ground state of Ising models.

For each Ising model, we evaluate performance by reporting the geometric average (3) of the model's percentage energy gap from the MIP solvers over 25 instances.

$$\Big(\prod_{i=1}^{n=25} \frac{GS_i - H(\boldsymbol{\sigma_i})}{GS_i} \Big)^{\frac{1}{n}}. \tag{3}$$

In addition, experimental details, code, and data availability are presented on our website located at link. To reproduce the results that we report, follow the tutorials outlined on the website to set up the solvers, download the datasets, and record the results.

4.2 Verification on BarabàsiAlbert (BA) Graphs

The BarabàsiAlbert (BA) graph is a scale-free network widely used in machine learning to evaluate Max-Cut performance due to its non-trivial degree distribution and community structure [4]. Following previous Max-Cut studies [9, 34], we evaluate our transformer-based REINFORCE policy on BA graphs of varying sizes from 100 to 1000 nodes.

Experimental setup. Each BA instance is generated with preferential attachment parameter $m = 4$. We compare our method against (i) the commercial solver Gurobi (mixed-integer quadratic programming) and (ii) the Monte Carlo Policy Gradient (MCPG) baseline. For each instance, the results are averaged over three random seeds, and we report the mean cut values.

Results. Table 2 shows that our REINFORCE policy achieves performance competitive with MCPG while consistently outperforming Gurobi on all scales. Furthermore, our REINFORCE algorithm beats MCPG by a margin of 0.03, in BA_500 of Table 2

BarabàsiAlbert Max-Cut Results				
Instance	Spins	REINFORCE	Gurobi	MCPG
BA_100	100	**283.70**	283.70	283.70
BA_200	200	**583.27**	583.27	583.27
BA_300	300	**880.43**	880.43	880.43
BA_400	400	**1179.70**	1179.33	1179.70
BA_500	500	**1479.70**	1475.67	1479.67
BA_600	600	1779.80	1772.83	**1780.07**
BA_700	700	2076.17	2067.50	**2077.33**
BA_800	800	2375.96	2361.57	**2377.80**
BA_900	900	2668.43	2655.93	**2675.03**
BA_1000	1000	2967.63	2952.20	**2974.57**

Table 2: Comparison of average Max-Cut values on BA graphs. Best results per instance are highlighted in bold.

4.3 Results for Spin Glass Instances

The results are summarized in Table 3. The Gurobi and CPLEX solvers achieve ground-state configurations of all instances and represent the baseline to which the RL algorithms are compared.

In 1D instances, VNA achieves spin configurations close to optimal, occasionally finding the true ground state. Although MCPG and REINFORCE do not find the ground state, they maintain competitive performance, verifying all three RL algorithms

In 2D instances, MCPG achieves energy values close to the reference optima in each instance with the geometric mean of energy gap percentages ranging from 0.000005% to 0.053214%. VNA and REINFORCE feature much worse performance with a geometric mean of energy gap percentage upper bound of 7.7% and 9.11% respectively.

In 3D instances, MCPG again achieves energy values near the reference optima in each instance. In contrast, REINFORCE achieves a result near optimum, with other instances ranging up to 5.679958%. Similarly, VNA achieves a competitive score for the first set of 3D instances and reaches a geometric mean upper bound of 5.137368%

In the infinite rank instances, REINFORCE achieves the best result of 0.504649% with VNA slightly behind at 0.519881% and MCPG performing the worst with a geometric mean of energy gap percentage of 0.976208%. Overall, while REINFORCE does not guarantee finding the global optimum, it maintains good approximation quality across different system sizes and dimensionality.

Dimension	Instance Type	Spins	MCPG	VNA	REINFORCE	Solvers	Ref.
1D	Classic	32	0.000003%	**0.00%**	0.000003%	0.00%	[17]
	Classic	64	0.000003%	**≈0.00%**	0.000084%	0.00%	[17]
	Classic	128	0.000003%	**≈0.00%**	0.000271%	0.00%	[17]
2D	Spin-Glass	100	**0.000005%**	4.647181%	0.114580%	0.00%	[32]
	Spin-Glass	144	**0.000003%**	6.815689%	2.288804%	0.00%	[32]
	Spin-Glass	196	**0.000002%**	7.444649%	3.614152%	0.00%	[32]
	Spin-Glass	256	**0.000004%**	7.733783%	4.479626%	0.00%	[32]
	EA	400	**0.000002%**	4.632651&	5.417051%	0.00%	[17]
	EA	625	**0.053214%**	5.529312%	8.689927%	0.00%	[13]
	EA	900	**0.001071%**	4.400705%	9.110858%	0.00%	[13]
3D	EA	64	**0.000003%**	0.031365%	0.000004%	0.00%	[13]
	EA	216	**0.000002%**	3.030916%	2.626845%	0.00%	[13]
	EA	512	**0.000002%**	5.137368%	5.679958%	0.00%	[13]
∞	SK	100	0.976208%	0.519881%	**0.504649%**	0.00%	[17]

Table 3: Combined results on Classic Ising and Spin-Glass instances. The lowest (best) energy per row is bolded.

4.4 Take-Away Message for Large-Scale Ising Models

It is important to note that on a large scale, RL-equipped methods achieve results that are significantly better than the solvers on the Max-Cut problem. This is likely due to the ability of RL to learn from the environment how to search the complex landscape of the Ising models more efficiently than classical heuristics. This observed superiority establishes RL as a powerful alternative in finding the ground state of Ising models and demonstrates the need for an interdisciplinary effort to propel progress in these fields. Although the same advantage is not currently observed on the Ising dataset, we believe that it can be achieved by further collaborative research of the physics and ML communities, and we believe that this dataset and benchmarks serve as a solid first step.

5 Conclusion and Future Work

In this paper, we have proposed an interdisciplinary physics-ML dataset to further efforts to find the ground state of the Ising model. Furthermore, we report a benchmark of state-of-the-art methods, including industry-level commercial solvers and machine learning approaches, to showcase the potential of RL methods. This project encourages the maintenance of reference curves to validate emerging methods in these fields. By maintaining an open-source dataset and benchmark, we hope to see more collaboration between the physics and ML fields.

We hope to see new or upgraded interdisciplinary RL methods to assist in finding the ground state of the Ising models. We plan to continue incorporating more data and methods into our benchmark, and extend the effectiveness of our algorithm to efficiently scale up and handle larger Ising systems.

References

[1] Miguel Aguilera, Masanao Igarashi, and Hideaki Shimazaki. Nonequilibrium thermodynamics of the asymmetric sherrington–kirkpatrick model. *Nature Communications*, 14(1):3685, 2023.

[2] Michael Aizenman, Hugo Duminil-Copin, and Vladas Sidoravicius. Random currents and continuity of ising model's spontaneous magnetization. *Communications in Mathematical Physics*, 334(2):719–742, Mar 2015.

[3] Marco Baity-Jesi, Enrico Calore, Andrés Cruz, Luis Antonio Fernandez, José Miguel Gil-Narvion, Isidoro Gonzalez-Adalid Pemartin, Antonio Gordillo-Guerrero, David Iñiguez, Andrea Maiorano, Enzo Marinari, Víctor Martin-Mayor, Javier Moreno-Gordo, Antonio Muñoz-Sudupe, Denis Navarro, Ilaria Paga, Giorgio Parisi, Sergio Perez-Gaviro, Federico Ricci-Tersenghi, Juan Jesús Ruiz-Lorenzo, Sebastiano Fabio Schifano, Beatriz Seoane, Alfonso Tarancon, Raffaele Tripiccione, and David Yllanes. Temperature chaos is present in off-equilibrium spin-glass dynamics. *Communications Physics*, 4(1):74, Apr 2021.

[4] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[5] Thomas Barrett, William Clements, Jakob Foerster, and Alex Lvovsky. Exploratory combinatorial optimization with reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:3243–3250, 2020.

[6] Massimo Bernaschi, Isidoro González-Adalid Pemartín, Víctor Martín-Mayor, and Giorgio Parisi. The quantum transition of the two-dimensional ising spin glass. *Nature*, 631(8022):749–754, Jul 2024.

[7] Stefan Boettcher. Physics of the edwardsanderson spin glass in dimensions d = 3, ,8 from heuristic ground state optimization. *Frontiers in Physics*, 12, Sep 2024.

[8] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

[9] Cheng Chen, Ruitao Chen, Tianyou Li, Ruichen Ao, and Zaiwen Wen. Monte carlo policy gradient method for binary optimization, 2023.

[10] Hanjun Dai, Elias B. Khalil, Yuyu Zhang, Bistra Dilkina, and Le Song. Learning combinatorial optimization algorithms over graphs. *arXiv preprint*, 2018.

[11] Hugo Duminil-Copin, Vladas Sidoravicius, and Vincent Tassion. Continuity of the phase transition for planar random-cluster and potts models with $1 \leq q \leq 4$. May 2015.

[12] Maxime Dupont, Bram Evert, Mark J. Hodson, Bhuvanesh Sundar, Stephen Jeffrey, Yuki Yamaguchi, Dennis Feng, Filip B. Maciejewski, Stuart Hadfield, M. Sohaib Alam, Zhihui Wang, Shon Grabbe, P. Aaron Lott, Eleanor G. Rieffel, Davide Venturelli, and Matthew J. Reagor. Quantum-enhanced greedy combinatorial optimization solver. *Science Advances*, 9(45):eadi0487, 2023.

[13] Changjun Fan, Mutian Shen, Zohar Nussinov, Zhong Liu, Yizhou Sun, and Yang-Yu Liu. Searching for spin glass ground states through deep reinforcement learning. *Nature Communications*, 14(1):725, 2023.

[14] Andreas Galanis, Leslie A. Goldberg, and Andres Herrera-Poyatos. The complexity of approximating the complex-valued ising model on bounded degree graphs. *SIAM Journal on Discrete Mathematics*, 36(3):2159–2204, 2022.

[15] J. González and T. Stauber. Ising superconductivity induced from spin-selective valley symmetry breaking in twisted trilayer graphene. *Nature Communications*, 14(1):2746, May 2023.

[16] Gurobi Optimization, LLC. *Gurobi Optimizer Reference Manual*, 2024.

[17] Mohamed Hibat-Allah, Estelle M. Inack, Roeland Wiersema, Roger G. Melko, and Juan Carrasquilla. Variational neural annealing. *Nature Machine Intelligence*, 3(11):952–961, 2021.

[18] IBM ILOG CPLEX. *V12.1: Users Manual for CPLEX*. International Business Machines Corporation, 2009.

[19] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, Feb 1925.

[20] Sorin Istrail. Statistical mechanics, three-dimensionality and np-completeness: I. universality of intracatability for the partition function of the ising model across non-planar surfaces (extended abstract). In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, STOC '00, pages 87–96, New York, NY, USA, 2000. Association for Computing Machinery.

[21] Andrew D. King, Alberto Nocera, Marek M. Rams, Jacek Dziarmaga, Roeland Wiersema, William Bernoudy, Jack Raymond, Nitin Kaushal, Niclas Heinsdorf, Richard Harris, Kelly Boothby, Fabio Altomare, Mohsen Asad, Andrew J. Berkley, Martin Boschnak, Kevin Chern, Holly Christiani, Samantha Cibere, Jake Connor, Martin H. Dehn, Rahul Deshpande, Sara Ejtemaee, Pau Farre, Kelsey Hamer, Emile Hoskinson, Shuiyuan Huang, Mark W. Johnson, Samuel Kortas, Eric Ladizinsky, Trevor Lanting, Tony Lai, Ryan Li, Allison J. R. MacDonald, Gaelen Marsden, Catherine C. McGeoch, Reza Molavi, Travis Oh, Richard Neufeld, Mana Norouzpour, Joel Pasvolsky, Patrick Poitras, Gabriel Poulin-Lamarre, Thomas Prescott, Mauricio Reis, Chris Rich, Mohammad Samani, Benjamin Sheldan, Anatoly Smirnov, Edward Sterpka, Berta Trullas Clavera, Nicholas Tsai, Mark Volkmann, Alexander M. Whiticar, Jed D. Whittaker, Warren Wilkinson, Jason Yao, T. J. Yi, Anders W. Sandvik, Gonzalo Alvarez, Roger G. Melko, Juan Carrasquilla, Marcel Franz, and Mohammad H. Amin. Beyond-classical computation in quantum simulation. *Science*, 388(6743):199–204, 2025.

[22] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[23] David Layden, Guglielmo Mazzola, Ryan V. Mishmash, Mario Motta, Pawel Wocjan, Jin-Sung Kim, and Sarah Sheldon. Quantum-enhanced markov chain monte carlo. *Nature*, 619(7969):282–287, 2023.

[24] Timothée Leleu, Farad Khoyratee, Timothée Levi, Ryan Hamerly, Takashi Kohno, and Kazuyuki Aihara. Scaling advantage of chaotic amplitude control for high-performance combinatorial optimization. *Communications Physics*, 4(1):266, 2021.

[25] Yongxiang Li, Shiqing Wang, Ke Yang, Yuchao Yang, and Zhong Sun. An emergent attractor network in a passive resistive switching circuit. *Nature Communications*, 15(1):7683, 2024.

[26] Tian Liang, S. M. Koohpayeh, J. W. Krizan, T. M. McQueen, R. J. Cava, and N. P. Ong. Heat capacity peak at the quantum critical point of the transverse ising magnet conb2o6. *Nature Communications*, 6(1):7611, Jul 2015.

[27] Zihua Liu, Erol Vatansever, Gerard T. Barkema, and Nikolaos G. Fytas. Critical dynamical behavior of the ising model. *Phys. Rev. E*, 108(3):034118, Sep 2023.

[28] Andrew Lucas. Ising formulations of many np problems. *Frontiers in Physics*, 2, 2014.

[29] Michael W. Macy, Boleslaw K. Szymanski, and Janusz A. Hoyst. The ising model celebrates a century of interdisciplinary contributions. *npj Complexity*, 1(1):10, Jul 2024.

[30] J. Majewski, H. Li, and J. Ott. The ising model in physics and statistical genetics. *Am. J. Hum. Genet.*, 69(4):853–862, Aug 2001.

[31] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G. R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 810:1–124, 2019.

[32] Kyle Mills, Pooya Ronagh, and Isaac Tamblyn. Finding the ground state of spin hamiltonians with reinforcement learning. *Nature Machine Intelligence*, 2(9):509–517, 2020.

[33] J. Oitmaa and D. D. Betts. The ground state of two quantum models of magnetism. *Canadian Journal of Physics*, 56(7):897–901, 1978.

[34] Tianle Pu, Changjun Fan, Mutian Shen, Yizhou Lu, Li Zeng, Zohar Nussinov, Chao Chen, and Zhong Liu. Transform then explore: a simple and effective technique for exploratory combinatorial optimization with reinforcement learning. *arXiv preprint*, 2024.

[35] F. Romá. Changing the universality class of the three-dimensional edwards-anderson spin-glass model by selective bond dilution. *Phys. Rev. B*, 103(6):064403, Feb 2021.

[36] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45(8):6056–6091, Apr 1992.

[37] Didier Sornette. Physics and financial economics (17762014): puzzles, ising and agent-based models. *Reports on Progress in Physics*, 77(6):062001, May 2014.

[38] Haoran Sun, Katayoon Goshvadi, Azade Nova, Dale Schuurmans, and Hanjun Dai. Revisiting sampling for combinatorial optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 32859–32874. PMLR, 2023.

[39] Angelika Wiegele and Frauke Liers. Biq mac library - binary quadratic and max cut library.

[40] Dian Wu, Lei Wang, and Pan Zhang. Solving statistical mechanics using variational autoregressive networks. *Phys. Rev. Lett.*, 122(8):080602, 2019.

[41] A. P. Young. Is the sherrington–kirkpatrick model relevant for real spin glasses? *Journal of Physics A: Mathematical and Theoretical*, 41(32):324016, 2008.

[42] Hao Zhang and Alex Kamenev. Computational complexity of three-dimensional ising spin glass: Lessons from d-wave annealer. *Phys. Rev. Res.*, 7(3):033098, 2025.

Appendix A Extended Background

Although there exist various methods to search for the ground state [33, 22], recent emerging methods are mostly ML-based because they can learn complex search strategies from data and can harness the power of modern hardware for more efficient optimization. We classify two categories of ML methods: ML ground-state and RL max-cut methods. ML ground-state methods refer to ML methods designed and applied to search for the ground state of Ising models [8, 17, 32]. RL max-cut methods refer to SOTA RL methods designed for max-cut, a problem mathematically similar to searching for the ground state [9, 38, 34, 5]. Despite the differences in applications, ML ground-state methods cannot scale or perform as well as RL max-cut methods because of their usage of outdated neural networks such as RNNs and reliance on handcrafted heuristics. On the other hand, RL max-cut methods achieve high-quality solutions at scale by either upgrading MCMC sampling, creating a probabilistic model to improve the parameterized policy distribution, or exploratory Q-learning frameworks.

Since RL max-cut methods are highly effective and max-cut is mathematically similar to searching for the ground state, we believe that RL max-cut methods can be efficient and scalable alternatives for finding the ground state of Ising models. By giving these SOTA tools to physics researchers, they gain access to high-quality solutions, better scalability, and accelerated progress in this field. However, the lack of Ising model datasets that accommodate ML applications suggests the need to curate a collective dataset of Ising model instances. Furthermore, there exists no benchmark for SOTA RL methods, hindering the evaluation and advancement of leading methods.

Appendix B Reinforcement Learning Methods

In this section, we discuss the current methods in Physics for finding the ground state of Ising models, and recent RL methods for solving combinatorial optimization problems. We also present how we apply a policy gradient based approach, with curriculum learning on spin masking to help explore the search space more efficiently and escape local minima to reach better solutions.

B.1 Challenges in Combinatorial Optimization

Combinatorial optimization (CO) problems aim to find a high-quality solution from a large discrete search space, where the number of possible solutions grows exponentially with the problem size.

Moreover, since combinatorial optimization problems are non-convex, the search space is filled with local minima. Heuristic methods, even hand-crafted by domain experts, often struggles to escape such local minima, and often fails to handle large-scale problems due to their linear scalability.

B.2 Methods in Physics Community

Traditional physical methods can be divided into two categories. Approximate methods include Simulated Annealing (SA) and Variational Monte Carlo (VMC), both of which rely on random sampling of the Boltzmann distribution. Therefore, as the dimension and system size increase, the sampling variance and mixing time increase sharply, resulting in unstable optimization and difficulty in scaling to large systems.

Exact methods such as ED obtain accurate eigenvalues and eigenstates by directly diagonalizing the Hamiltonian matrix in the full Hilbert space, but its dimension grows exponentially with the number of particles, causing memory and computational requirements to quickly get out of control and can only handle very small systems.

In summary, traditional physical heuristic methods face insurmountable bottlenecks when solving large-scale combinatorial optimization.

Given these bottlenecks, researchers are increasingly applying machine learning to combinatorial optimization, such as hybrid annealingML frameworks (VCA [17], COOL[32]) as well as pure ML strategies like DIRAC [13].

B.3 Methods in Reinforcement Learning Community

Machine learning enhanced physics methods retain physical heuristic information and have good interpretability, but are still limited by the bottlenecks of Boltzmann sampling and annealing processes, making the searching process difficult to expand and generalize. In contrast, pure reinforcement learning methods can learn solution transitions straight from the data, and directly drive the search through policies or value functions, which circumvents the limitation of heuristic-based transitions. In recent years, researchers have proposed various innovative reinforcement learning (RL) algorithms around combinatorial optimization problems. these pure reinforcement learning methods show significant advantages in large-scale combinatorial optimization problems:

- **Learned Policy**: The policies are learned through continuous interaction with the environment, adaptively balancing exploration and exploitation to avoid falling into local optima without relying on fixed heuristic-based transition rules.
- **Good Scalability**: RL agents can theocratically jump from one state to an arbitrary other state, so they can effectively handle large-scale problems and maintain high performance on large graphs.
- **Generalization ability**: Strategies learned on different training instances can usually capture common features of the problem; when faced with new instances or structures, RL agents can still provide high-quality solutions based on previously learned strategies.

Empiricaly, in recent years, reinforcement learning methods have achieved good results in combinatorial optimization tasks such as maximum independent set, TSP, graph partitioning, CVRP, etc., proving its high efficiency and wide applicability. In general, RL methods can be divided into three categories:

1. **Sampling methods**: such as instance-wise sampling (iSCO) [38], which gradually focuses on high quality areas by dynamic sampling and updating the distribution in the current solution neighborhood.
2. **Valuedriven methods**: such as ECO[5] and S2V [10], which use graph neural networks to approximate the value function $Q(s, a)$ and greedily select local changes, using value functions to guide the search path.
3. **Policynetwork methods**: such as MCPG [9], which directly use policy networks to generate complete or local solutions, and update the network parameters from end to end through policy gradients.

For the benchmark, we currently incldue ECO and MCPG because they demonstrated SOTA-level performance and are representative of value-driven and policy-network methods.

B.4 Policy-based Search Framework

Under the RL framework, we can model the objective as the negative Hamiltonian, so that the goal becomes training a policy to output the spin configuration with the minimal Hamiltonian.

$$J(\pi_\theta) = \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}[R(\tau)] = \mathbb{E}\left[\sum_{t=1}^{T}\gamma^t * H(\boldsymbol{\sigma}^t)\right]. \tag{4}$$

Policy gradient based reinforcement learning provides a natural and flexible approach to combinatorial optimization problems: the policy network learns to propose high-quality solutions by directly maximizing the expected reward (negative Hamiltonian for the Ising models). Since we model state transitions directly instead of actions, we modify the Policy Gradient theorem accordingly:

$$\nabla J(\pi_\theta) = \mathbb{E}\left[\sum_{t=1}^{T-1}\nabla_\theta \log(\pi_\theta(x_{t+1}|x_t)) \cdot \sum_{t=1}^{T-1}\gamma^t(-H(x_{t+1}))\right]. \tag{5}$$

$$L_{\text{actor}}(\theta) = -\mathbb{E}\left[\sum_{t=1}^{T-1}\log(\pi_\theta(x_{t+1}|x_t)) \cdot \sum_{t=1}^{T-1}\gamma^t(-H(x_{t+1}))\right]. \tag{6}$$

In addition to its own advantages, our purely policy-based approach also provides clear opportunities for combining with three other paradigms:

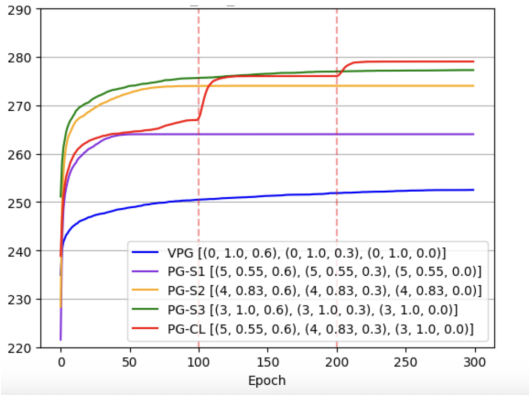

Figure 2: Training curves with different CL stages on BA_100_ID0.

Actor-Critic Inserting a critique module at the front end of the policy network to estimate the value of the current solution provides a more stable and lower-variance training signal for policy updates. The actor-critic architecture combines the flexibility of policy gradients with the reliability of value evaluation, which helps to improve convergence speed.

Filter Function Filter functions such as local search (LS), simulated annealing (SA), etc. are powerful tools to optimize the trajectory discovered by the policy network. After the policy network proposes an initial solution, these filtering functions can be used to efficiently further optimize the solution by focusing on local neighborhood. This step provides low-cost fine-grained optimization so that the policy can focus on a more global scale, ensuring that the proposed solution has high quality potential before finalizing.

Sampling Sampling methods, such as Markov chain Monte Carlo (MCMC) or random sampling, can be used to generate a diverse set of candidate solutions, which helps find a good initial solution and improves search efficiency. By generating a large number of candidate solutions, we increase the probability of discovering high-quality starting points for further exploration, increasing the likelihood of finding a global optimal solution.

Through this “policy-value-sampling" hybrid framework, reinforcement learning methods are very suitable for solving combinatorial optimization problems, and can also be combined with other related algorithmic ideas, showing strong performance potential and application flexibility.

B.5 Curriculum Learning

Our approach applies a random dynamic mask to the attention matrix of the graph structure and incorporates the local search steps and temperature scheduling into a multi-stage curriculum learning process. In the early stage of training, we mask most spins and maintain a high temperature to simplify the goal and help the model quickly escape local optima; as training progresses, these blocking probabilities, search steps, and temperature parameters are smoothly adjusted according to the preset plan, the mask range gradually expands, the search steps decrease, and the temperature gradually decreases, so that the model can integrate more global information and finely optimize structural features in the later stage. Compared with no CL, our approach shows a clear “staircase" improvement on the training curve , which significantly improves the overall training efficiency and final effect, as shown in Fig. 2.