Long-Range Ising Model: A Benchmark for Long-Range Capabilities in Graph Learning

Joël Mathys[1] [⋆], Henrik Christiansen[2], Federico Errica[2], and Francesco Alesiani[2]

[1] ETH Zürich {jmathys}@ethz.ch
[2] NEC Laboratories Europe {firstname.lastname}@neclab.eu

Abstract. Accurately modeling long-range dependencies in graph structured data is critical for many real-world applications. However, properly incorporating long-range interactions beyond the nodes' immediate neighborhood remains an open challenge for graph machine learning models. Existing benchmarks for evaluating long-range capabilities cannot guarantee that their tasks actually depend on long-range information or are artificial. Therefore, claiming long-range modeling improvements based on empirical performance on those datasets remains a fragile and weak form of evidence. We introduce the Long-Range Ising Model (LRIM) Graph Benchmark, a physics-grounded benchmark based on the well-studied Ising model whose ground truth theoretically depends on long-range dependencies. Our benchmark consists of multiple datasets that scale from 256 to 65k nodes per graph and provide controllable long-range dependencies through multiple tunable parameters, allowing precise control over the hardness and "long-rangedness" of tasks. We provide model-agnostic evidence showing that local information is insufficient, further validating the design choices of our benchmark. This is ongoing, new research to provide a framework towards principled and provable long-range capability evaluation for graph machine learning.

Keywords: Graph Benchmark, Long-range interactions, Graph Neural Networks

1 Introduction

Since the early days of deep learning on graphs [45, 33, 23, 43, 32] researchers have tried to automatically learn a task-defined mapping from graph-structured data to some output. At the very heart of the most popular architecture, namely Graph Neural Networks (GNNs), is the idea that repeated aggregation of local information expands the "receptive field" of each node [3] in a way similar to convolutional neural networks for images [28]. Such an expansion is crucial, for instance, in tasks where the true mapping requires a non-local processing of information among nodes in the graph. In this case, researchers typically talk about *capturing long-range dependencies.*

[⋆] Work conducted while the author was an intern at NEC Laboratories Europe.

Long-range dependencies or interactions are an important component of physics and chemistry, manifesting, for example, in quantum systems [15], protein folding [24] or astronomy [11]. An example from biology is mRNA splicing, a fundamental part of the gene expression process: splicing is inhibited if we "disable" long-range dependencies between distant regions of mRNA [41]. The protein binding mechanism, whose understanding is crucial for the development of vaccines, also depends on long-range interactions between proteins [19].

As of today, most deep learning models on graphs struggle to solve long-range tasks. For GNNs, the reason is tied to their efficient but limited message-passing scheme: expanding the receptive field rapidly increases the amount of information that every node has to process and store, generating a *computational bottleneck* [2]. When addressing the long-range limitations of existing models, it should be natural to benchmark novel methods on tasks that provably depend on long-range information while being relevant for real-world applications. Unfortunately, while popular existing benchmarks focus on real-world data [17], they cannot **guarantee** that *i)* the task to solve hinges on long-range dependencies and *ii)* that there are no shortcuts models can find while learning from the data [47]. In both cases, the reason is that the task definition is *unknown*, which is almost always the case in machine learning. Researchers have also tried to formalize a heuristic notion of "long-rangedness" [5], but *a priori* defining a task that can be tuned and theoretically analyzed has many advantages, for instance interpreting and inspecting the learned models that approximate the task function.

This work presents a first attempt at a *provable* long-range benchmark based on the well-studied and fundamental Ising model with power-law long-range node (called *spin*) interactions. The Ising model was originally introduced in statistical physics to study magnetic materials [39] and fundamental properties related to phase transitions [49, 46, 51]. Over the decades, the influence of the model has spread to aid understanding and analyzing complex phenomena far beyond the original intention, with applications in protein folding [10], percolation [4], the theory of disordered systems [38], or social systems such as stock markets [16] to name only a few. The $d = 2$ LRIM we use in this work models a ferromagnetic material by placing binary variables called spins on a grid lattice. The problem we want to solve is the prediction of the energy change $\varDelta E_i$ when one component is updated, as often required for Markov Chain Monte Carlo simulations of such systems. The tunable parameters of the Ising model control the dependency of each spin's energy on distant spins, thereby allowing us to control the impact of long-range dependencies; in other words, *we can easily control the "long-rangedness" of the task.*

In what follows, we describe the dataset creation process, highlighting important design choices such as the physics-based pseudo-critical temperature and the hardness of the task. We validate the hardness of the task by measuring the error of "partial oracle" functions that have access to a limited neighborhood for each node, showing that the error increases for specific choices of parameters controlling the impact of long-range dependencies. Then, we train classical GNNs and provide their performance as a baseline starting point for future evaluations, together with the details of the replicable evaluation process.

2 Related Work

The literature on long-range benchmarks for graph-structured data is rather limited. The most popular one is perhaps the Long Range Graph Benchmark (LRGB) [17], which proposes image segmentation tasks – adapted to graphs – together with peptides' function classification and property regression tasks. These tasks are considered to require long-range interactions, especially the image-based tasks where graph transformers [37] perform much better than classical GNNs, but there are no strong guarantees that they are. Moreover, such empirical claims are prone to reevaluations [48, 6]. On the other hand, synthetic tasks such as predicting eccentricity, shortest path distance, and diameter on randomly generated topologies [14] are provably long-range, but their real-world impact is unclear. Recently, node eccentricity has been computed on large real road networks, however, eccentricity was approximated by a 16-hop radius for computational reasons [30]. The recent dataset of [30] uses real-world data but considers an artificial task. In this work, we propose a benchmark that relies on simulated data whose task is connected to real-world applications and fully controllable. Hopefully, this allows us to have a greater impact while being able to carefully control whether models learn the true objective function.

Lastly, we mention that heterophilic datasets have often been believed to require long-range capabilities of GNNs, but this viewpoint was recently criticized by [2] by providing clear counterexamples that the task might induce heterophily regardless of the nature of the problem.

3 Background

Spin Models The description of systems with many interacting components is ubiquitous in the natural and social sciences. Using domain-specific modeling approaches, these systems can be investigated on a case-by-case basis using convoluted system prototypes that are often difficult to understand and interpret. Spin models, such as the Ising model, have proven powerful in describing relevant features of real systems while retaining simplicity. Based on simple microscopic interaction laws, they show rich emergent behavior and non-trivial phase-transition and spin-spin correlations.

Formally, the system is defined by the graph topology and the interaction law between the so-called spins $\mathbf{s}_i$, referred to as the spin Hamiltonian (without external magnetic field).

$$\mathcal{H}(\{\mathbf{s}_i\}) = -\frac{1}{2}\sum_{ij} J_{ij}\mathbf{s}_i\mathbf{s}_j, \tag{1}$$

where $\mathbf{s}_i \in \mathbb{R}^n$ are the spin variables described as unit vectors of dimension n, i.e., for $O(n = 1)$ one has binary spins $s_i = \pm 1$. The interaction potential

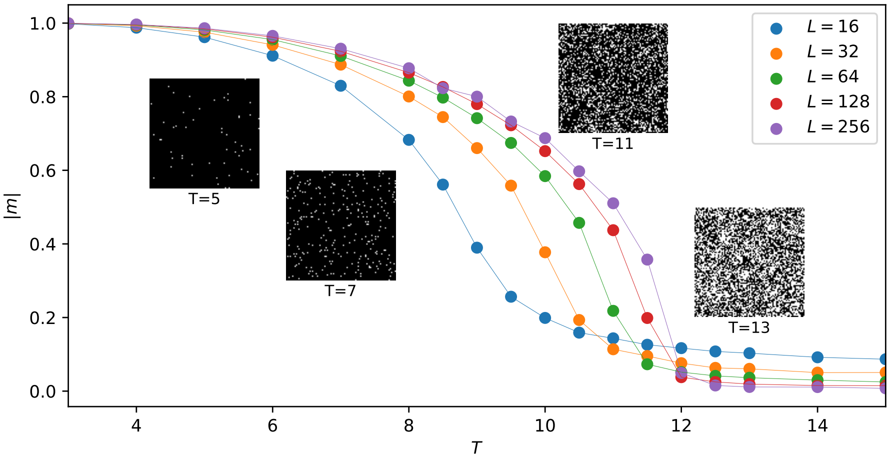

Fig. 1: Absolute magnetization $|M|$ versus temperature T for the LRIM with $\sigma = 0.6$ for different system sizes L. The snapshots show respective configurations from simulations for $L = 256$.

J_{ij} describes the graph connectivity, and depending on the choice one can interpolate from the fully-connected mean-field model to nearest-neighbor models where spins only interact within a short-ranged neighborhood. We here consider this model in contact with a thermal environment, that is, a canonical setting in which the microscopic configurations occur according to the Boltzmann distribution.

$$P(\{\mathbf{s}_i\}) = \exp\left(-\frac{\mathcal{H}(\{\mathbf{s}_i\})}{k_b T}\right), \tag{2}$$

where the Boltzmann constant is set to unity ($k_b = 1$) and T is the temperature.

The ferromagnetic model with $J_{ij} > 0$ for all edges with the nodes placed on a regular graph shows three distinct phases depending on T: *i*) disordered, *ii*) critical, or *iii*) ordered. The phases are characterized by distinct behavior of the connected correlation function

$$G_c(\mathbf{r}) = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle, \tag{3}$$

where $\langle \ldots \rangle$ symbolized expectations under the Boltzmann distribution of Eq. 2. One has

1. An ordered phase for $T < T_c$: At low temperatures, the system orders and one has spontaneous magnetization $m \neq 0$. $G_c(\mathbf{r})$ decays exponentially with distance $r = |\mathbf{r}_i - \mathbf{r}_j|$ as

$$G_c(\mathbf{r}) \sim e^{-r/\xi}, \quad r \to \infty. \tag{4}$$

Here, ξ is the correlation length and $\mathbf{r}_i$ is the position of spin s_i.

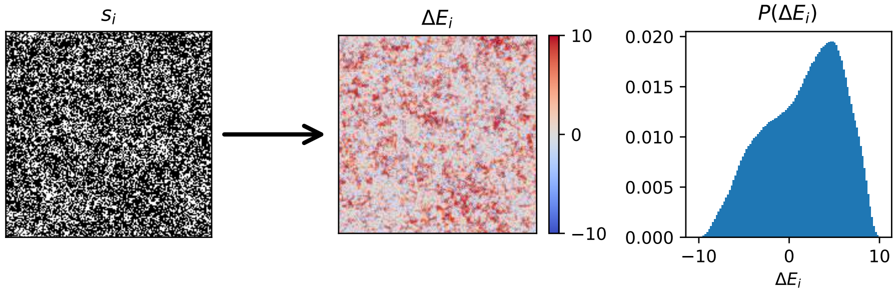

Fig. 2: Visualization of the mapping from spin configuration s_i to the corresponding ΔE_i including the resulting histogram of predictions. The presented snapshots were generated for $L = 256$ and $\sigma = 0.6$ at the pseudo critical point.

2. Critical phase at $T = T_c$: At the critical point, fluctuations are correlated across all distances, and no length scale dominates. Self-similar structures emerge, and the spin clusters become fractal. Furthermore, $G_c(\mathbf{r})$ decays algebraically
$$G_c(\mathbf{r}) \sim \frac{1}{r^\eta}, \quad r \to \infty, \tag{5}$$
with the critical exponent η.
3. Disordered phase with $T > T_c$: Above T_c, the system is disordered and in a paramagnetic state. The correlation function decays again exponentially. As the correlations vanish rapidly at large distances, there is no long-range order in this phase.

If one approaches the critical temperature from either side, the correlation length ξ diverges as
$$\xi \sim |T - T_c|^{-\nu}, \quad T \to T_c, \tag{6}$$
with the critical exponent ν. This means, that spins are correlated over large distances in a nontrivial way, posing an ideal way to construct appropriate configurations for our benchmark.

The critical exponents η and ν are universal, i.e., they do not depend on microscopic details of systems, and are instead determined by dimensionality d, symmetry of order parameters n, conservation laws, and range of interactions.

Long-Range Ising Model For the LRIM we consider in this work, we place the spins on a regular grid in $d = 2$ spatial dimensions. The spins interaction with power-law potential
$$J_{ij} = \frac{1}{|\mathbf{r}_j - \mathbf{r}_i|^{d+\sigma}}, \tag{7}$$
where σ controls the long-rangedness of the interactions. The value of σ not only controls the long-rangedness of the interaction, but also changes the equilibrium and nonequilibrium behavior of the model. This is reflected in different values

for the critical exponents, for example η for the correlation function and ν for the correlation length:

1. For $\sigma < 1$, one has mean-field behavior with the corresponding critical exponents [25].
2. For $1 < \sigma < \sigma_\times$, the critical exponents depend on σ.
3. For $\sigma_\times < \sigma$, one is in the short-range (nearest-neighbor) universality class.

The value of $\sigma_\times$ is discussed in the literature [40, 1, 44], with $\sigma_\times = 2$ [20] or $\sigma_\times = 1.75$ [42] being the most likely candidates.

For nonequilibrium simulations investigating phase-ordering kinetics and related properties [8], the nonequilibrium exponents show non-trivial dependence on σ, with $\sigma = 1$ being the point of interest in most cases for $d = 2$ [7, 13, 12, 34, 35].

Simulation To simulate the LRIM, a Markov Chain Monte Carlo simulation can be set up by proposing a random single spin flip $\{\ldots, s_i, \ldots\} \rightarrow \{\ldots, -s_i, \ldots\}$, and accepting the proposal according to the Metropolis criterion with probability

$$p = \min\left(1, \exp\left(-2\Delta E_i / k_B T\right)\right), \text{ with } \Delta E_i = s_i \sum_j s_j J_{ij}. \tag{8}$$

Although there exist cluster algorithms that decorrelate quickly in equilibrium [31, 22, 21], and recent advances for single spin flip simulations [36] that avoid exact calculations of ΔE_i by exploiting the way Monte Carlo simulations are constructed, the calculation of ΔE_i is at the heart of many other models and tasks where certain properties cannot be exploited [26].

To obtain our samples from the target distribution for the dataset, we implement the single cluster variant for the LRIM as presented in [21]. We make sure to equilibrate the simulation before measuring, and write out data only after sufficient decorrelation from the previous sample.

4 LRIM Graph Benchmark

In this section, we outline how exactly we make use of the previously discussed physical spin model to construct our LRIM Graph Benchmark. The main goal is to directly translate the system into an appropriate graph-based task formulation while preserving its simplicity and controllable mechanisms for long-range interactions. We focus on the $d = 2$ LRIM on a grid lattice and want to predict ΔE_i energies present throughout the system. Each LRIM instance gives us a graph G with $L \times L$ nodes that are arranged in a 2D periodic grid. Note that the topology is shared among all instances and that each node is connected to its 4 nearest neighbors. Moreover, each node has a single feature, representing the physical spin $\{-1, +1\}$. We formulate the energy prediction as a node regression task, where each node v has to predict its energy change $\Delta E \in \mathbb{R}$. A visual illustration of how the graph task is constructed is shown in Figure 3.

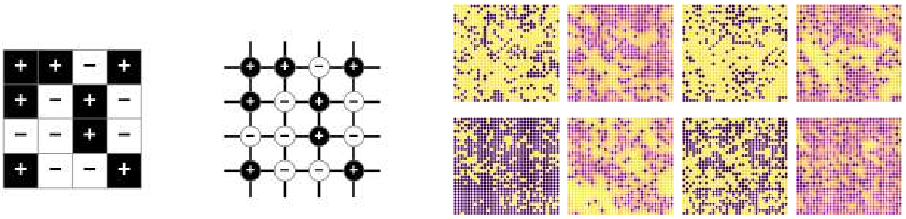

Fig. 3: On the left how an Ising configuration is represented as an attributed graph in the LRIM Graph Benchmark. On the right a depiction of four configurations and their associated ΔE from the LRIM-M dataset.

Table 1: Dataset specifications for the LRIM-S, LRIM-M, LRIM-L, LRIM-XL, and LRIM-XXL versions of the LRIM benchmark.

Property	LRIM-S	LRIM-M	LRIM-L	LRIM-XL	LRIM-XXL
Number of Nodes	256	1‘024	4‘096	16‘384	65‘536
Number of Edges	512	2‘048	8‘192	32‘768	131‘072
Diameter	16	32	64	128	256
Avg. Shortest Path	8.03	16.01	32.01	64.00	128.00
Avg. Effective Resistance	0.49	0.60	0.71	0.82	0.93
Node Degree	4	4	4	4	4
σ easy	1.6	1.6	1.6	1.6	1.6
σ hard	0.6	0.6	0.6	0.6	0.6
Number of Graphs	1‘000	1‘000	1‘000	1‘000	1‘000
Node Features	1	1	1	1	1
Edge Features	–	–	–	–	–
Task	Node Prediction	Node Prediction	Node Prediction	Node Prediction	Node Prediction
Prediction Target	$\Delta E \in \mathbb{R}$	$\Delta E \in \mathbb{R}$	$\Delta E \in \mathbb{R}$	$\Delta E \in \mathbb{R}$	$\Delta E \in \mathbb{R}$
Performance Metric	logMSE	logMSE	logMSE	logMSE	logMSE

We want to precisely evaluate the long-range capabilities with the LRIM benchmark and therefore provide datasets across different scales and difficulty levels. The benchmark comprises five dataset sizes from LRIM-S (256 nodes) to LRIM-XXL (65,536 nodes). Each dataset variant contains 1,000 distinct graph instances. To vary the long-range interaction strength, we generate two variants for each system size L (graph size). The "hard" variant uses $\sigma = 0.6$, creating stronger long-range dependencies that require models to aggregate information from more distant nodes. The "easy" variant sets $\sigma = 1.6$. Data generation follows the outlined Monte Carlo sampling protocol. For each system size L and σ value, we first determine the appropriate pseudo-critical temperature $T_c(\sigma, L)$ where the system exhibits longest correlation lengths, thereby creating the most interesting configurations. We then sample datapoints at ten temperatures equally spaced between $0.95T_c$ and $1.05T_c$, capturing the critical region where long-range correlations are present. Each configuration is sampled from a simulation which is first equilibrated followed by a decorrelation phase between two subsequent samples to ensure statistical independence between them. Then, we split the dataset into 80/10/10 for training, validation, and testing. Complete graph and dataset statistics for all variants are provided in Table 1.

We want to highlight the computational efficiency aspect when evaluating methods for their long-range capabilities, as improvement often comes at significant computational cost or techniques. To ensure rigorous and fair comparisons, we ask that **methods report their runtime complexity for their computational budget** (e.g. $\mathcal{O}(L \cdot E)$ for standard MPNNs with L layers and E edges) **and any precomputation costs**, including creation of additional structure and feature preprocessing. We put little restriction on what can be used on the benchmark on purpose, encouraging novel methods. However, all modifications and their computational overhead must be transparently documented.

5 Evaluation

5.1 Long-Range Analysis

In this section, we aim to demonstrate why our proposed LRIM benchmark is suitable for testing long-range interactions. Because the dataset consists of synthetic simulation data, we have complete knowledge about the underlying generation process as well as full control over the fundamental parameters. This allows us to study the intrinsic properties and determine whether long-range interactions are required independently of any specific proposed model baseline. We consider three different perspectives: how the simulation accuracy degrades when restricted to local neighborhoods, what limitations WL imposes on the realized dataset, and theoretical error bounds on the worst case. Together, these analyses provide strong evidence that our benchmark captures tasks that require long-range reasoning capabilities.

First, we analyze how prediction accuracy degrades when information is restricted to local neighborhoods only. We construct an "oracle" predictor that has access to the r-hop neighborhoods of each target node and predicts ΔE based on the correct contributions within the r-hop neighborhood. Although this is not a strict upper bound on achievable performance, as a model could pick up correlated information beyond the r-hop neighborhood in the data, we expect this to be close to the best possible predictor.

We vary the parameter r from 1 to the diameter of the graph on a selection of different datasets, illustrated in Figure 4. These datasets differ in their chosen $\sigma \in \{0.6, 1.6\}$, where a smaller σ corresponds to stronger long-range interactions, and system sizes $L \in \{16 \times 16, 32 \times 32\}$. These results show that task difficulty can be precisely controlled by both the parameter σ and the size of the system L. We observe that lower σ values consistently require larger neighborhoods to achieve the same prediction accuracy, since the $\sigma = 0.6$ curves lie above the $\sigma = 1.6$ curves across system sizes, indicative of stronger long-range interactions. Furthermore, for the same σ larger system sizes increase task difficulty, supported by the $L = 32$ curves lying above the $L = 16$ curves. The prediction error decays smoothly as the neighborhood size increases from local to global, showing that

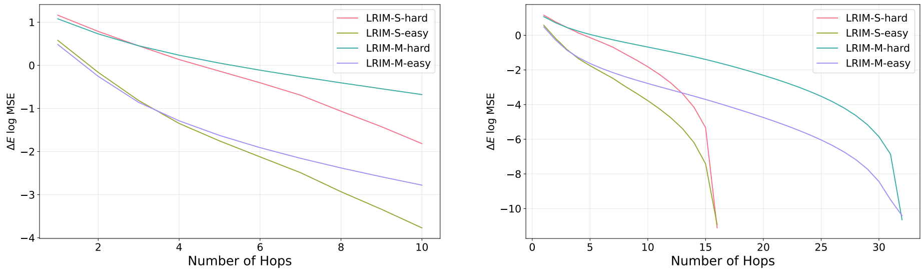

Fig. 4: LogMSE of oracle predictions restricted to the r-hop neighborhood. Results show that the task difficulty can be controlled and increased by both smaller σ values and larger system sizes as they require larger neighborhoods to achieve the same prediction accuracy. Therefore, to achieve low prediction error requires to account for long-range interactions across substantial fractions of the entire graph. The right side depicts the behavior over the full system size, whereas on the left only the first 10 hops are shown.

incorporating information from more distant nodes consistently improves the prediction accuracy. Crucially, this analysis reveals that achieving high accuracy requires considering interactions across substantial fractions of the entire graph.

Next, we discuss how easy it is for MPNNs to fit the instantiated datasets. The Weissfeiler-Leman (WL) [29] test provides a theoretical framework to understand the expressivity limitations of message-passing neural networks. Since standard MPNNs cannot distinguish between nodes that have identical WL labels, nodes within the same WL equivalence class must predict the same outputs. We compute 1-WL labels for all nodes in our datasets up to depth k, creating equivalence classes of nodes that are indistinguishable to k-layer MPNNs. For each equivalence class, we measure the range of ΔE values between the nodes. Figure 5 shows the maximum range among all equivalent classes depending on the size of the considered neighborhood. As seen in Figure 5, the maximum range is not negligible initially, but decreases as the neighborhood increases. This exemplifies two important insights: First, there are nodes with similar neighborhoods that have very different prediction targets in our datasets. This is desirable as it requires information beyond the immediate neighbors to distinguish these cases. As a consequence, there is an inherent drive towards an increased receptive field, which is also necessary to capture long-range dependencies, in order to uniquely shatter the equivalence classes. Second, the curve drops off faster than the analysis of the oracle predictor. That is, because of finite data, there exists a potential pitfall to approximate the true prediction with fewer than the minimum required number of layers. However, it is crucial to note that this analysis does not provide conclusions about how well models can generalize beyond the training data. In fact, we expect the number of layers to be necessary to closely follow the oracle predictor. However, we should be aware of this discrepancy between the number of layers required for (over)fitting and generalizing.

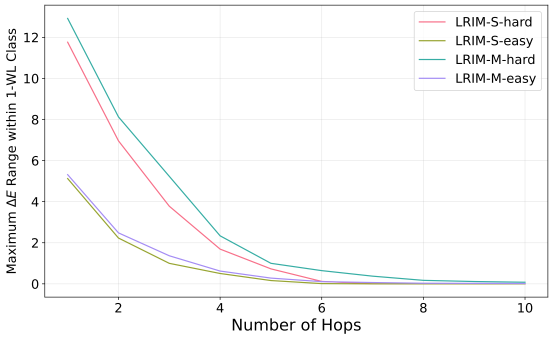

Fig. 5: Maximum range of ΔE values among nodes sharing the same 1-WL label, plotted as a function of number of hops. Results show that WL equivalence classes initially contain nodes with very different prediction targets, requiring larger receptive fields to distinguish them. The faster decay compared to the oracle predictor (Figure 4) indicates potential overfitting with finite data, though generalization likely requires distances closer to the oracle.

Finally, we provide a theoretical argument that underscores the fundamental necessity of long-range information for the best performance of the task. We establish a lower bound on the worst-case error that any method restricted to local neighborhoods must exhibit by not considering the rest of the graph.

Lemma 1. *Let f_θ be a model that predicts ΔE for a given node v on an instance X with a periodic grid graph G of size $N = n \times n$ with diameter $D = n$. If f_θ only considers the spins of nodes within radius $r \ll D$, then there exists a configuration X' where $f_\theta(X)_v = f_\theta(X')_v$ but $|Y'_v - f_\theta(X')_v| \geq n^{-\sigma}$.*

The idea is that we can create several instances that share the same local neighborhood, but the spins that were not taken into account can create a range of possible ΔE values. Therefore, no predictor that only considers the local information can have a maximum error significantly below that range.

5.2 Empirical Evaluation

To see how current graph learning approaches perform on our proposed benchmark, we evaluate three common architectures GIN [50], GCN [27], and GatedGCN [9] on the hard variants of our LRIM datasets. Moreover, we add an MLP baseline, which predicts only based on the individual node feature. Due to computational constraints, we focus our evaluation on the S - XL dataset sizes, omitting the computationally intensive XXL variant. We report the mean logMSE in Table 2 with standard deviations over 3 runs. Complete hyperparameter configurations for each architecture are provided in the Appendix. Performance consistently worsens as the size of the graph increases from LRIM-S to LRIM-XL, confirming the increased difficulty of larger datasets. All three

Table 2: We report the log MSE of different baselines on the hard LRIM benchmark variant across different dataset sizes, including computational complexity analysis. The number of edges E corresponds to $4N$ in our datasets.

	Preprocessing	Computation	LRIM-S-hard	LRIM-M-hard	LRIM-L-hard	LRIM-XL-hard
MLP	-	$\mathcal{O}(N)$	1.363 ± 0.000	1.353 ± 0.001	1.339 ± 0.001	1.326 ± 0.001
GIN	-	$\mathcal{O}(L \cdot E)$	0.630 ± 0.005	0.662 ± 0.003	0.705 ± 0.007	0.714 ± 0.005
GCN	-	$\mathcal{O}(L \cdot E)$	0.596 ± 0.001	0.641 ± 0.002	0.682 ± 0.001	0.693 ± 0.001
GatedGCN	-	$\mathcal{O}(L \cdot E)$	0.582 ± 0.003	0.621 ± 0.001	0.654 ± 0.001	0.667 ± 0.001

MPNNs achieve comparable performance with a slight edge by the GatedGCN architecture. This suggests that the challenge might not be due to the specific architecture, but in the challenges and limitations of local message-passing when used for long-range tasks.

To better understand how performance relates to the size of the receptive field, we perform an ablation study of the number of message-passing layers using the GIN architecture. Figure 6 shows the performance from 2 to 20 layers. Note that the dataset has a diameter of 16, therefore, we would expect improvement up to that point. We observe that performance consistently improves with increased depth. However, it plateaus around 8 to 12 layers, well before the diameter of 16. However, there is a significant gap between the empirically achieved performance of these message-passing based models and the oracle predictor. This might be due in part to the limited size of the current dataset version, although preliminary investigations with more data yielded similar results. However, it hints at known phenomena, such as computational bottlenecks, and requires further investigation. However, such a steep drop off is surprising and is exactly the kind of insight we hope to uncover with the LRIM Graph Benchmark as a new valuable tool towards developing more capable long-range techniques for graph learning.

6 Limitations

Our proposed LRIM benchmark is fundamentally a synthetic dataset with a well-understood mechanism, and as such it has limited direct real-world applicability. Its primary purpose and main advantage is to provide an understandable and controllable framework to assess long-range capabilities rather than to solve an open real-world problem. As such, we do not intend LRIM as a replacement for real-world benchmarks, but rather as a complementary tool for advancing the study of long-range interactions for the domain of graph learning. Furthermore, our current benchmark is limited to regular lattice structures, which may not capture the diverse topological patterns encountered in general graphs. Future extensions could incorporate other structured graph types, but this would require careful consideration of how to properly obtain appropriate simulated data and the accompanied long-range analysis. Finally, empirical performance on LRIM

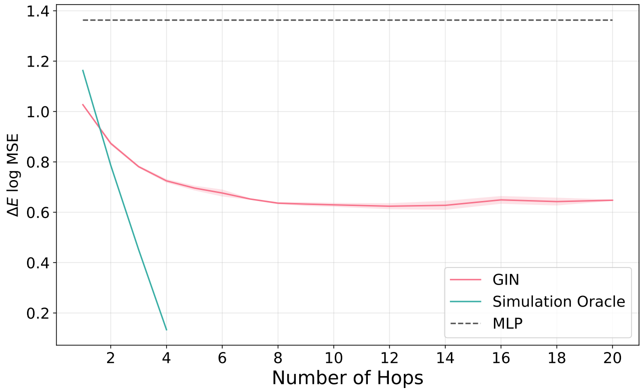

Fig. 6: LogMSE performance of GIN architecture depending on number of message-passing layers on LRIM-S-hard dataset. For each number of rounds a new model is trained and results are averaged over three seeds. Performance improves with increased depth but plateaus around 8-12 layers, well before reaching the graph diameter. Noteably, there is a significant performance gap between GIN performance and the simulation oracle.

may not translate directly to real-world tasks, as practical applications involve additional complexities beyond pure modeling of long-range interactions. The benchmark value lies in its ability to isolate and systematically evaluate this single but crucial capability as one component of their overall effectiveness.

7 Conclusion

Using current graph learning benchmarks, it is difficult to properly isolate and assess the ability to capture long-range dependencies. Performance is usually argued through empirical evidence, although it is often unclear to what extent tasks depend on or require long-range information. We introduce the Long-Range Ising Model (LRIM) Graph Benchmark, a physics-grounded framework based on the Ising model that provides controllable and provable long-range dependencies. This allows us to precisely control and vary the hardness of the task across datasets that scale from 256 to 65k nodes. In addition, we provide model-agnostic evidence through which LRIM tasks genuinely require long-range reasoning, with oracle prediction degrading when information is restricted to local neighborhoods. Our empirical evaluation reveals large gaps between current methods and oracle performance, highlighting fundamental limitations in existing graph learning approaches when confronted with provably long-range tasks. This benchmark establishes a foundation for developing, properly evaluating, and advance our understanding of what architectural innovations are needed to tackle long-range dependency modeling in graph-structured data.

References

1. Angelini, M.C., Parisi, G., Ricci-Tersenghi, F.: Relations between short-range and long-range Ising models. Physical Review E **89**, 062120 (2014)
2. Arnaiz-Rodriguez, A., Errica, F.: Oversmoothing, “oversquashing", heterophily, long-range, and more: Demystifying common beliefs in graph machine learning. arXiv:2505.15547 (2025)
3. Bacciu, D., Errica, F., Micheli, A., Podda, M.: A gentle introduction to deep learning for graphs. Neural Networks **129** (2020)
4. Balogh, J., Bollobás, B., Duminil-Copin, H., Morris, R.: The sharp threshold for bootstrap percolation in all dimensions. Transactions of the American Mathematical Society **364**, 2667 (2012)
5. Bamberger, J., Gutteridge, B., le Roux, S., Bronstein, M., Dong, X.: On measuring long-range interactions in graph neural networks. In: Proceedings of the 42nd International Conference on Machine Learning (ICML) (2025)
6. Bechler-Speicher, M., Finkelshtein, B., Frasca, F., Müller, L., Tönshoff, J., Siraudin, A., Zaverkin, V., Bronstein, M.M., Niepert, M., Perozzi, B., Galkin, M., Morris, C.: Position: Graph learning will lose relevance due to poor benchmarks (2025), https://arxiv.org/abs/2502.14546
7. Bray, A.J., Rutenberg, A.D.: Growth laws for phase ordering. Phys. Rev. E **49**, R27 (1994)
8. Bray, A.J.: Theory of phase-ordering kinetics. Advances in Physics **43**(3), 357–459 (1994)
9. Bresson, X., Laurent, T.: Residual gated graph ConvNets. arXiv:1711.07553 (2018)
10. Bryngelson, J.D., Wolynes, P.G.: Spin glasses and the statistical mechanics of protein folding. Proceedings of the National Academy of Sciences **84**, 7524 (1987)
11. Carroll, B.W., Ostlie, D.A.: An introduction to modern astrophysics. Cambridge University Press (2017)
12. Christiansen, H., Majumder, S., Henkel, M., Janke, W.: Aging in the long-range Ising model. Phys. Rev. Lett. **125**, 180601 (2020)
13. Christiansen, H., Majumder, S., Janke, W.: Phase ordering kinetics of the long-range Ising model. Phys. Rev. E **99**, 011301 (2019)
14. Corso, G., Cavalleri, L., Beaini, D., Liò, P., Veličković, P.: Principal neighbourhood aggregation for graph nets. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS) (2020)
15. Defenu, N., Donner, T., Macrì, T., Pagano, G., Ruffo, S., Trombettoni, A.: Long-range interacting quantum systems. Reviews of Modern Physics **95**, 035002 (2023)
16. Durlauf, S.N.: How can statistical mechanics contribute to social science? Proceedings of the National Academy of Sciences **96**, 10582 (1999)
17. Dwivedi, V.P., Rampášek, L., Galkin, M., Parviz, A., Wolf, G., Luu, A.T., Beaini, D.: Long range graph benchmark. In: Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS) (2022)
18. Ellens, W., Spieksma, F., Van Mieghem, P., Jamakovic, A., Kooij, R.: Effective graph resistance. Linear Algebra and its Applications **435**(10), 2491–2506 (2011), special Issue in Honor of Dragos Cvetkovic
19. Ferber, M., Zoete, V., Michielin, O.: T-cell receptors binding orientation over peptide/MHC class I is driven by long-range interactions. PloS one **7**, e51943 (2012)
20. Fisher, M.E., Ma, S.k., Nickel, B.: Critical exponents for long-range interactions. Physical Review Letters **29**, 917 (1972)

21. Flores-Sola, E., Weigel, M., Kenna, R., Berche, B.: Cluster Monte Carlo and dynamical scaling for long-range interactions. The European Physical Journal Special Topics **226**, 581 (2017)
22. Fukui, K., Todo, S.: Order-n cluster monte carlo method for spin systems with long-range interactions. Journal of Computational Physics **228**, 2629 (2009)
23. Gori, M., Monfardini, G., Scarselli, F.: A new model for learning in graph domains. Proceedings of the International Joint Conference on Neural Networks (IJCNN) (2005)
24. Gromiha, M.M., Selvaraj, S.: Importance of long-range interactions in protein folding. Biophysical chemistry **77**, 49 (1999)
25. Kadanoff, L.P.: More is the same; phase transitions and mean field theories. Journal of Statistical Physics **137**, 777 (2009)
26. Katzgraber, H.G., Young, A.P.: Monte Carlo studies of the one-dimensional Ising spin glass with power-law interactions. Physical Review B **67**, 134410 (2003)
27. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR) (2017)
28. LeCun, Y., Bengio, Y., others: Convolutional networks for images, speech, and time series. The Handbook of Brain Theory and Neural Networks **3361**, 1995 (1995)
29. Leman, A., Weisfeiler, B.: A reduction of a graph to a canonical form and an algebra arising during this reduction. Nauchno-Technicheskaya Informatsiya **2**(9), 12–16 (1968)
30. Liang, H., Sáez de Ocáriz Borde, H., Sripathmanathan, B., Bronstein, M., Dong, X.: Towards quantifying long-range interactions in graph machine learning: a large graph dataset and a measurement. arXiv:2503.09008 (2025)
31. Luijten, E., Blöte, H.W.: Monte Carlo method for spin models with long-range interactions. International Journal of Modern Physics C **6**, 359 (1995)
32. Micheli, A.: Neural network for graphs: A contextual constructive approach. IEEE Transactions on Neural Networks **20** (2009)
33. Micheli, A., Sestito, A.: A new neural network model for contextual processing of graphs. In: Proceedings of the Italian Workshop on Neural Networks (WIRN) (2005)
34. Müller, F., Christiansen, H., Janke, W.: Phase-separation kinetics in the two-dimensional long-range ising model. Physical Review Letters **129**, 240601 (2022)
35. Müller, F., Christiansen, H., Janke, W.: Nonuniversality of aging during phase separation of the two-dimensional long-range ising model. Physical Review Letters **133**, 237102 (2024)
36. Müller, F., Christiansen, H., Schnabel, S., Janke, W.: Fast, hierarchical, and adaptive algorithm for Metropolis Monte Carlo simulations of long-range interacting systems. Physical Review X **13**, 031006 (2023)
37. Müller, L., Galkin, M., Morris, C., Rampášek, L.: Attending to graph transformers. Transactions on Machine Learning Research (2024)
38. Parisi, G.: Nobel lecture: Multiple equilibria. Reviews of Modern Physics **95**, 030501 (2023)
39. Peierls, R.: On Ising's model of ferromagnetism. Mathematical Proceedings of the Cambridge Philosophical Society **32**, 477 (1936)
40. Picco, M.: Critical behavior of the Ising model with long range interactions. arXiv:1207.1018 (2012)
41. Rüegsegger, U., Leber, J.H., Walter, P.: Block of hac1 mrna translation by long-range base pairing is released by cytoplasmic splicing upon induction of the unfolded protein response. Cell **107**, 103 (2001)

42. Sak, J.: Recursion relations and fixed points for ferromagnets with long-range interactions. Physical Review B **8**, 281 (1973)

43. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Transactions on Neural Networks **20** (2009)

44. Shiratani, S., Todo, S.: Stochastic parameter optimization analysis of dynamical quantum critical phenomena in the long-range transverse-field Ising chain. Physical Review E **110**, 064106 (2024)

45. Sperduti, A., Starita, A.: Supervised neural networks for the classification of structures. IEEE Transactions on Neural Networks **8** (1997)

46. Stanley, H.E.: Introduction to phase transitions and critical phenomena. Oxford University Press (1987)

47. Tönshoff, J., Ritzert, M., Rosenbluth, E., Grohe, M.: Where did the gap go? reassessing the long-range graph benchmark. In: The 2nd Learning on Graphs Conference (LoG) (2023)

48. Tönshoff, J., Ritzert, M., Rosenbluth, E., Grohe, M.: Where did the gap go? reassessing the long-range graph benchmark (2023), https://arxiv.org/abs/2309.00367

49. Wilson, K.G.: The renormalization group and critical phenomena. Reviews of Modern Physics **55**, 583 (1983)

50. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? International Conference on Learning Representations (2019)

51. Yeomans, J.M.: Statistical mechanics of phase transitions. Clarendon Press (1992)

A Dataset

The Diameter D of a graph G as the maximum shortest hop distance between any two nodes in the graph. For a periodic grid graph of size N, the distances can be calculated using the manhattan distance and the diameter is $\sqrt{N}$.

$$D = \max_{u,v \in V(G)} d(u,v)$$

The average shortest path of a graph is the average shortest hop distance between two nodes in the graph. For a periodic grid graph of size $N = n \cdot n$ it is given as $\frac{n^3}{2(n^2-1)}$.

$$
\begin{aligned}
SP &= \frac{1}{N(N-1)} \sum_{u \in V(G)} \sum_{v \in V(G), v \neq u} d(u,v) \\
&= \frac{1}{N(N-1)} N \sum_{v \in V(G)} d(u, v_0) && \text{topology of all nodes is the same} \\
&= \frac{1}{(N-1)} \sum_{v \in V(G)} \min\{u_x, n-u_x\} + \min\{u_y, n-u_y\} && \text{manhatten distance} \\
&= \frac{1}{(N-1)} 2 \sum_{v \in V(G)} \min\{u_x, n-u_x\} && \text{symmetry and linearity of coordinates} \\
&= \frac{1}{(N-1)} 2n \sum_{i=1}^{n} \min\{i, n-i\} && \text{repeated summation of each row} \\
&= \frac{1}{(N-1)} 2n \left(\left(2 \sum_{i=1}^{\frac{n}{2}} i \right) - \frac{n}{2} \right) \\
&= \frac{1}{(N-1)} 2n \left(\frac{n}{2} (\frac{n}{2} + 1) - \frac{n}{2} \right) \\
&= \frac{n^3}{2(n^2-1)}
\end{aligned}
$$

The effective resistance was calculated using the networkx implementation following the Kirchhoff index[18] and normalized by the number of edges.

B Evaluation

Lemma 2. *Let f_θ be a model that predicts ΔE for a given node v on an instance X with a periodic grid graph G of size $N = n \times n$ with diameter $D = n$. If f_θ only considers the spins of nodes within radius $r \ll D$, then there exists a configuration X' where $f_\theta(X)_v = f_\theta(X')_v$ but $|Y'_v - f_\theta(X')_v| \geq n^{-\sigma}$.*

Proof. We construct two candidate instances X'_1, X'_2, which have the exact same spins as X within radius r and are all -1, respectively $+1$ outside of that. The error of any prediction will then be at least $\frac{1}{2}|Y'_{1,v} - Y'_{2,v}|$.

$$\begin{aligned}
Y'_{1,v} - Y'_{2,v} &= \sum_{u \in G} x'_{1,u} x'_{1,v} d(u,v)^{-(2+\sigma)} - \sum_{u \in G} x'_{2,u} x'_{2,v} d(u,v)^{-(2+\sigma)} && \text{def. of } \mathit{\Delta E} \\
&= \sum_{u \in G, d(u,v) > r} x'_{1,u} x'_{1,v} d(u,v)^{-(2+\sigma)} - \sum_{u \in G, d(u,v) > r} x'_{2,u} x'_{2,v} d(u,v)^{-(2+\sigma)} \\
&= \sum_{u \in G, d(u,v) > r} (x'_{1,u} - x'_{2,u}) x'_{1,v} d(u,v)^{-(2+\sigma)} \\
&= 2x'_{1,v} \sum_{u \in G, d(u,v) > r} d(u,v)^{-(2+\sigma)} \\
\frac{1}{2}|Y'_{1,v} - Y'_{2,v}| &= \sum_{u \in G, d(u,v) > r} d(u,v)^{-(2+\sigma)} \\
&\geq (N - r^2) \frac{1}{n}^{(2+\sigma)} \geq n^{-\sigma} - \frac{r^2}{n^{2+\sigma}} \\
&\geq n^{-\sigma} && r \ll n
\end{aligned}$$

C Empirical Evaluation

Model selection and hyperparameter optimization are performed exclusively on LRIM-S, with the best configurations then applied to larger datasets. All models were trained using the MSE loss for 500 epochs using the AdamW optimizer with weight decay of 1e-5 using a cosine scheduler with 5 epochs warmup and gradient clipping (l2 norm of 1). Moreover, the MPNNs use a residual connection between layers, a linear encoder as well as a two layer MLP for readout.

Table 3: Hyperparameter configurations for the best performing GNN architectures. Bold values indicate the optimal setting for each architecture.

Hyperparameter	gatedgcnconv	gcnconv	ginconv
base_lr	{0.0001, **0.001**}	{0.0001, **0.001**}	{0.0001, **0.001**}
batch_size	{**64**, 128}	{**64**, 128}	{**64**, 128}
batchnorm	{**False**, True}	{False, **True**}	{**False**, True}
dropout	{0.0, 0.1, **0.2**}	{0.0, **0.1**, 0.2}	{0.0, **0.1**, 0.2}
dim_inner	{**32**, 64, 128}	{32, 64, **128**}	{**32**, 64, 128}
layers_mp	{5, **10**}	{5, **10**}	{5, **10**}