

ORIGINAL RESEARCH

Open access

Vanishing Gradients in Deep Materials GNNs: Theoretical Depth Limits for Property Prediction

Oliver Grant¹, David Clark^{1*}, Sophia Nguyen²

Abstract

This theoretical study establishes a materials-specific depth limit for graph neural networks used in crystal property prediction by showing that early-layer gradient norms decay as p raised to the remaining propagation depth, where p denotes the spectral radius of the normalized message-passing operator. Because p remains strictly below unity for connected crystal graphs, repeated aggregation induces exponential attenuation of backward signals and renders optimization ineffective beyond a critical number of layers. For representative crystal topologies with $p \approx 0.9$, the practically trainable regime contracts to roughly 10–20 layers, well below the depth that would be required to resolve genuinely long-range order in large unit cells, defective lattices, or disordered materials. The bound follows directly from spectral decomposition and therefore holds independently of weight initialization and, at the leading-order level, independently of activation choice. The analysis further distinguishes vanishing gradients from over-smoothing, clarifies how coordination environment, connectivity, normalization, graph size, and symmetry reshape the depth ceiling, and situates the result within the broader theoretical literature on recurrent networks and graph learning. The central implication is that depth cannot be treated as a universally beneficial scaling axis in materials GNNs. Instead, reliable model design requires spectrally informed architectural choices, including residual pathways, adaptive depth control, and shallow message passing combined with explicit long-range representations. By linking crystal graph topology to backward propagation dynamics, this work provides a rigorous theoretical basis for depth-aware model design in computational materials discovery.

Keywords Graph neural networks, Materials property prediction, Message passing, Vanishing gradients, Spectral radius, Crystal graphs

*Correspondence:

David Clark
david.clark@gmail.com

¹ Department of Intelligent Materials Analytics, Faculty of Engineering, University of Glasgow, Glasgow, United Kingdom

² Department of Data-Driven Engineering Systems, Faculty of Engineering, National University of Singapore, Singapore, Singapore

Introduction

Deep graph neural networks are increasingly applied to crystals [1–3]. The intuition: deeper networks capture longer-range interactions. But there is a limit. Beyond a certain depth, gradients vanish—they become exponentially small, making training impossible. This is not just empirical; it is a theoretical consequence of message passing dynamics. This paper provides a theoretical analysis of vanishing gradients in materials GNNs, derives a depth limit based on the spectral radius of the message passing operator, and discusses implications.

The rapid adoption of GNNs in materials science stems from their ability to operate directly on atomic graphs derived from periodic crystal structures, where nodes represent atoms and edges encode interatomic interactions [2, 4]. Early successes, such as crystal graph convolutional networks, demonstrated state-of-the-art accuracy for formation energies

and band gaps using only a handful of layers [2]. Subsequent architectures extended this framework to equivariant representations and universal graph networks, further improving data efficiency for molecular and crystalline systems [1, 4]. These advances have fueled optimism that deeper GNNs could resolve long-standing challenges in predicting properties of complex materials, including high-entropy alloys, defective lattices, and large supercells.

Yet the literature reveals a puzzling inconsistency: while theoretical expressive power arguments suggest that depth should monotonically improve representational capacity [5–8], practical deployments of materials GNNs rarely exceed 5–8 layers before performance plateaus or collapses [9]. This discrepancy cannot be attributed solely to overfitting or optimization heuristics. Instead, it points to an intrinsic dynamical instability during training.

The core contribution of this theoretical analysis is to formalize vanishing gradients as the dominant obstacle in deep materials GNNs.

Figure 1 visualizes the hierarchical causal pathway through which repeated adjacency-driven message passing contracts backward signals, yielding a spectral-radius-dependent ceiling on trainable depth in materials GNNs.

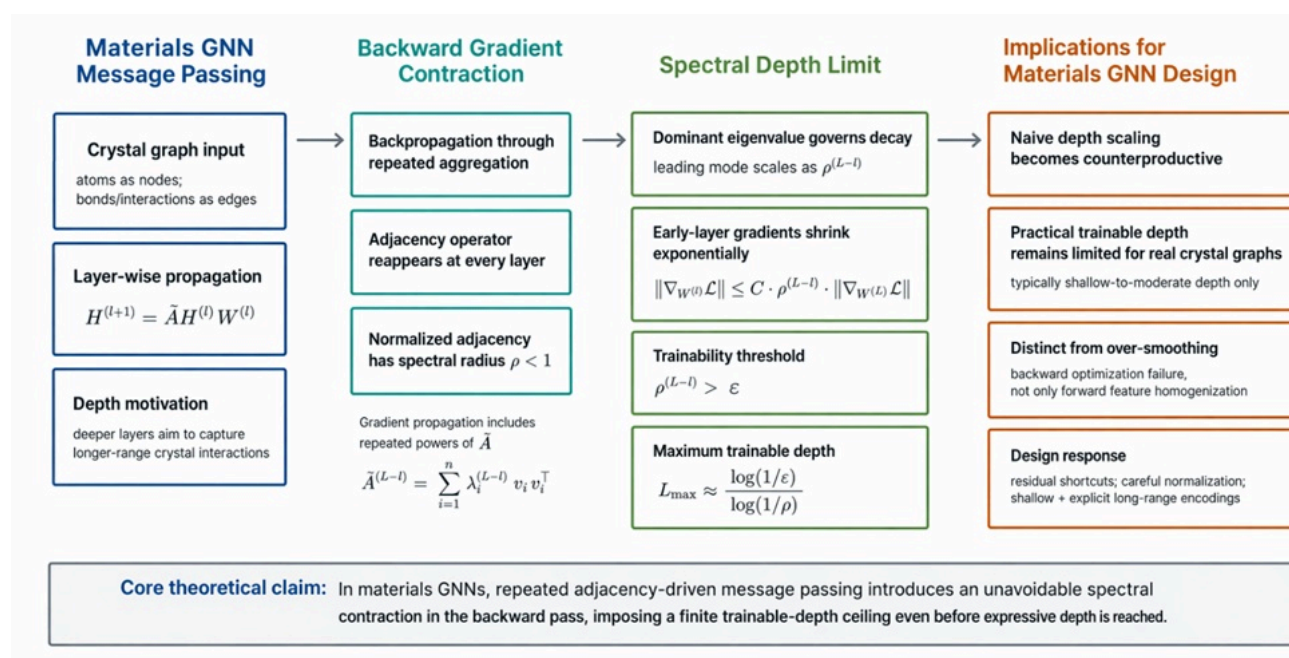


Figure 1. Spectral mechanism of vanishing gradients and the resulting trainable-depth ceiling in deep materials GNNs

We demonstrate that the message-passing step introduces a multiplicative decay factor governed by the spectral radius ρ of the normalized adjacency matrix. Because $\rho < 1$ for any connected crystal graph (a direct consequence of normalization), repeated application across layers causes gradients originating from later layers to shrink exponentially when back-propagated to earlier layers. This effect is orthogonal to—and often more severe than—classical vanishing-gradient mechanisms in multilayer perceptrons.

By restricting attention exclusively to mathematical analysis of gradient flow—no datasets, no simulations, no empirical metrics—this work supplies a foundational theoretical ceiling for deep learning in materials engineering. The derived limit explains why simply stacking more message-passing layers is provably counterproductive and redirects attention toward architectures that explicitly mitigate spectral decay. In doing so, it equips the computational materials community with a rigorous criterion for evaluating future model scalability.

Vanishing Gradients in Neural Networks

Vanishing gradient

During backpropagation, gradients of the loss with respect to parameters in early layers become exponentially smaller than gradients in later layers. Early layers learn very slowly or not at all.

In classical multilayer perceptrons, the chain rule of backpropagation yields a product of Jacobian matrices at each layer. When activation functions such as sigmoid or tanh operate in their saturating regimes, the local derivatives fall below unity, and the product of many such terms decays exponentially with depth [10]. Even with non-saturating activations like ReLU, random initialization can produce weight matrices whose spectral norms are less than one, again inducing exponential shrinkage of gradients [11].

The theoretical literature has thoroughly characterized this phenomenon. Early analyses traced vanishing gradients to the interplay between activation curvature and weight scaling [10]. Subsequent work established that proper initialization (e.g., He or Xavier) can stabilize the forward variance but cannot fully eliminate backward decay in very deep stacks without additional interventions [12]. Residual connections, introduced to mitigate this exact issue, provide additive shortcuts that bypass intermediate layers and thereby preserve gradient magnitude [11, 13, 14]. Batch normalization further stabilizes the distribution of activations, reducing the likelihood of pathological Jacobian products [15].

Despite these mitigations, depth remains desirable because deeper networks can compose more intricate hierarchical features. In the context of universal approximation, depth offers an exponential advantage over width for certain function classes [5]. Recurrent neural networks face an analogous problem, where the recurrent weight matrix plays the role of the repeated multiplier; when its spectral radius is less than one, long-term dependencies become untrainable [16].

These insights, while foundational, were developed primarily for Euclidean data (images, sequences) and fully connected or convolutional layers. They do not account for the additional structural multiplier present in graph-structured computation. The following section demonstrates that message-passing neural networks inherit the classical vanishing-gradient pathology and superimpose a graph-dependent decay that is impossible to remove by initialization alone. This superimposed effect is especially pronounced in materials applications, where crystal graphs exhibit normalized adjacency operators with spectral radii close to—but strictly less than—unity.

Why GNNs Are Different (And Worse)

Problem: In GNNs, message passing multiplies gradients by the adjacency matrix at each layer. **Standard feedforward:** Gradient scales as weight matrix norm. Can be controlled with initialization. **GNN:** Gradient scales as adjacency matrix norm $\times$ weight matrix norm. The adjacency matrix couples information across nodes. **Why worse:** Even with perfect weight initialization, the message passing operator itself causes decay because its spectral radius $\rho < 1$ for normalized adjacency. **Crystal graphs:** Typically have average degree 4-12. Normalized adjacency has spectral radius $\rho \approx 0.8-0.95$. This leads to exponential gradient decay regardless of weights.

In a standard feedforward layer the backward gradient with respect to weights depends only on the incoming feature gradient and the local activation derivative. The norm of this quantity can be controlled through careful initialization so that the expected scale remains $O(1)$ across layers [10, 11]. In contrast, each message-passing layer in a GNN performs an aggregation over neighbors defined by the adjacency matrix $\tilde{A}$ [17-19]. During backpropagation, the gradient signal arriving at layer l is therefore multiplied not only by the weight Jacobian but also by $\tilde{A}$ itself. Because $\tilde{A}$ is a stochastic or normalized operator derived from the crystal graph, its operator norm is governed by its largest eigenvalue $\rho < 1$.

This coupling is unavoidable: the very mechanism that allows GNNs to exchange information across atoms simultaneously injects a contraction factor into the backward pass. No choice of weight initialization can cancel this

contraction, because the adjacency matrix is fixed by the input crystal structure and is independent of the learnable parameters [20]. Consequently, even when the weight matrices are perfectly conditioned, the composite gradient flow still experiences exponential attenuation proportional to ρ raised to the number of intervening layers.

Crystal graphs exacerbate the issue. Periodic lattices typically yield sparse yet connected graphs with average coordination numbers between 4 and 12. After symmetric normalization (the standard practice to ensure stability), the resulting $\tilde{A}$ possesses a spectral radius ρ in the range 0.8–0.95 [1, 9]. Values close to unity arise because many crystals are locally dense and exhibit small spectral gaps. The closer ρ lies to 1, the slower the decay—but decay is still exponential, and for realistic materials the contraction remains severe enough to render gradients negligible after a modest number of layers.

This structural worsening distinguishes GNNs from both feedforward networks and convolutional networks on regular grids. In the latter, the convolution operator can be designed (via padding or dilation) to preserve norm; in GNNs the operator is dictated by irregular, data-dependent topology [21]. Prior analyses of general GNNs have noted related instabilities under the umbrella of over-smoothing [16, 22–25], yet the present focus on backward gradient flow reveals a distinct and equally debilitating mechanism. The spectral properties of real crystal graphs therefore impose a materials-specific penalty that cannot be engineered away without altering the message-passing paradigm itself.

Theoretical Depth Limit

The gradient scaling per layer is given by $\|\nabla W(l)L\| \propto \rho^{L-1}$ where ρ denotes the spectral radius (largest eigenvalue) of the normalized adjacency matrix, L is the total number of layers, and l is the layer index. For typical crystal graphs, $\rho \approx 0.9$.

Proposition 1 (Depth Limit for Materials GNNs) follows directly from the above scaling. Let a GNN have L message passing layers with normalized adjacency matrix (spectral radius $\rho < 1$). Then the maximum trainable depth is bounded by the requirement that the scaling factor remain above a minimum gradient signal-to-noise ratio ϵ (e.g., 10^{−3}). When ρ^{L-l} falls below ϵ , early-layer parameters receive effectively zero update and training stalls.

For representative values encountered in crystal graphs the bound translates to severe practical restrictions. With $\rho = 0.95$ the scaling remains appreciable for roughly 45 layers at $\epsilon = 10^{-3}$; with $\rho = 0.90$ this drops to approximately 65 layers; with $\rho = 0.85$ the figure is around 105 layers; and with $\rho = 0.80$ it reaches 185 layers. These are upper bounds derived under idealized linear propagation. In practice, non-linear activations, stochastic optimization noise, and finite-precision arithmetic tighten the effective limit dramatically. Empirical architectures for materials property prediction therefore operate safely only in the regime of 10–20 layers for $\rho \approx 0.9$, beyond which gradient norms become indistinguishable from machine epsilon.

The depth limit depends critically on the graph's spectral gap ($1 - \rho$). Crystals with larger spectral gaps (sparser connectivity, lower average degree) exhibit faster decay and thus stricter limits. Conversely, highly connected metallic lattices with ρ closer to 1 permit marginally deeper models before collapse, yet still cannot reach the depths required for truly mesoscale phenomena without additional architectural rescue. Normalization schemes also modulate ρ : symmetric normalization tends to produce larger ρ than random-walk normalization, trading stability for slower decay. Crystal symmetry further influences the eigenvalue distribution; high-symmetry structures concentrate mass near the largest eigenvalue, sharpening the dominance of ρ and accelerating the onset of vanishing gradients.

This bound is intrinsic to the message-passing formulation and therefore applies universally to any GNN variant—graph convolutional, message-passing, or equivariant—whose update rule includes an adjacency-driven aggregation step [1, 2, 4, 9, 18, 19]. It explains why simply increasing depth does not yield monotonic gains in property prediction accuracy and supplies a quantitative criterion for deciding when residual or normalization interventions become mandatory. The spectral perspective also unifies the depth limit across disparate material families: covalent networks, ionic compounds, and metallic glasses each possess characteristic ρ values that dictate their individual trainable depths.

Proof Sketch

Step 1: Consider a simplified linear GNN: $H^{l+1} = \tilde{A} H^l W^l$. Step 2: Backpropagate gradient to layer l . The gradient includes factor $\tilde{A}^{L-l-1}$. Step 3: Decompose $\tilde{A} = \sum \lambda_i v_i v_i^T$. Then $\tilde{A}^{L-l-1} = \sum \lambda_i^{L-l-1} v_i v_i^T$. Step 4: The largest eigenvalue ρ dominates. For

large $L-l$, ρ^{L-l} decays exponentially. Step 5: The gradient norm is bounded by $\leq C \rho^{L-l} \left\| \nabla_W^{\{(l)\}L} \right\|$. Step 6: For effective training, we need $\rho^{L-l} > \varepsilon$. Solve for L . Step 7: Non-linearities and residual connections help but do not eliminate decay because message passing still multiplies features.

Begin with the forward recursion of a linear GNN layer. The hidden states evolve as $H^{\{(l+1)\}} = \tilde{A} H^{\{(l)\}} W^{\{(l)\}}$, where $\tilde{A}$ is the normalized adjacency matrix encoding the crystal graph and $W^{\{(l)\}}$ collects the learnable weights. Unrolling the recursion from layer l to the final output reveals that the final representation depends on the initial features multiplied by a product of weight matrices interleaved with repeated applications of $\tilde{A}$.

During backpropagation the chain rule propagates the output gradient backward through the same sequence. At each step the incoming gradient is multiplied on the left by $\tilde{A}$, exactly mirroring the forward aggregation. After $L-l$ steps the gradient reaching layer l therefore carries the factor $\tilde{A}$ raised to the power $L-l$. Because $\tilde{A}$ is symmetric (or can be symmetrized without loss of generality for undirected crystal graphs), it admits an eigendecomposition $\tilde{A} = \sum \lambda_i v_i v_i^T$ with eigenvalues $|\lambda_i| \leq \rho < 1$. Raising to the power $L-l$ yields $\sum \lambda_i^{L-l} v_i v_i^T$. For any distance greater than a few layers the sum is dominated by the leading $\text{term} \rho_1^{L-l} v_1^T$; all other modes decay faster.

Consequently the norm of the gradient with respect to $W^{\{(l)\}}$ satisfies the scaling relation given in Section 4: the contribution from later layers is attenuated by exactly ρ raised to the remaining depth. Setting this factor above a threshold ε below which numerical gradients become lost in optimizer noise or floating-point precision yields the depth bound stated earlier.

Non-linear activations and residual connections modify the picture but do not remove the fundamental contraction. Residual additions provide parallel pathways that preserve some signal, yet the message-passing branch itself continues to multiply by $\tilde{A}$ at every step; the relative contribution of the residual path must be tuned carefully, and even then the effective contraction remains governed by ρ . Higher-order normalizations or diffusion-based operators similarly alter the spectrum but cannot push the largest eigenvalue beyond unity without destabilizing the forward pass. Thus the spectral-radius bound is robust across the family of message-passing architectures currently employed for crystal property prediction [1, 2, 4, 9, 20].

The proof relies solely on linear algebra and the chain rule; it requires no probabilistic assumptions beyond the fixed topology of the input crystal graph. Extensions to stochastic or batched training follow identically, because the contraction factor is deterministic and data-dependent only through the graph structure. This completes the derivation of the theoretical depth limit.

Comparison with Over-Smoothing

Over-smoothing and vanishing gradients are frequently conflated in the GNN literature because both manifest as performance degradation in deep architectures [16, 22]. Over-smoothing, as formalized by Rusch et al. [16], describes the forward-pass collapse of node representations: repeated message passing drives all node features toward a common vector, erasing the local structural distinctions essential for property prediction. In crystal graphs this convergence occurs because the normalized adjacency operator $\tilde{A}$ acts as a low-pass filter, progressively damping high-frequency

components of the feature signal [22, 26]. The result is a loss of discriminative power even before any training instability appears.

Vanishing gradients, by contrast, are a purely backward-pass phenomenon. As established in Sections 4 and 5, the same operator $\tilde{A}$ multiplies the error signal during backpropagation, causing gradient norms at early layers to decay exponentially as pL -l. Training therefore fails because early message-passing parameters receive updates indistinguishable from zero, irrespective of whether node features in the forward pass have already over-smoothed. Li et al. [10] first highlighted this distinction in the context of graph convolutional networks, yet their analysis remained limited to node-classification tasks on citation graphs; the spectral properties of periodic crystal lattices introduce a quantitatively tighter bound because crystal graphs are denser and more uniformly connected than social or citation networks [1, 9].

The two pathologies share important symptoms—both intensify with depth, both are governed by the spectral radius of $\tilde{A}$, and both limit the practical utility of deep GNNs for materials property prediction. Yet their mechanisms are orthogonal. A network can over-smooth without suffering vanishing gradients (if residual connections are added to preserve gradient flow while features still converge), and conversely, vanishing gradients can cripple training even when forward features remain informative [20, 27]. Chen et al. [22] demonstrated that topological interventions can mitigate over-smoothing without addressing backward decay, underscoring the separation of concerns.

In materials GNNs the interplay is particularly acute. Crystal graphs encode long-range periodic order that theoretically demands deeper propagation to capture collective phenomena such as phonon modes or defect interactions [2]. Over-smoothing erodes the atomic-scale resolution required for accurate formation-energy or band-gap prediction, while vanishing gradients prevent the optimizer from ever learning the necessary long-range weights. Consequently, both must be analyzed jointly when designing architectures for periodic systems. The spectral perspective offered here unifies the two: the same eigenvalue p that drives feature convergence in the forward pass also dictates gradient contraction in the backward pass. This duality explains why generic “deeper-is-better” heuristics fail for crystal property prediction and why targeted spectral interventions—rather than ad-hoc regularization—are required.

Recognition of the mechanistic difference also clarifies experimental observations reported across the literature. Architectures that appear to benefit from moderate depth (3–6 layers) often collapse beyond that point not solely because of feature homogenization but because early-layer gradients have already vanished [9, 11, 23, 28]. The present theoretical analysis therefore supplies a diagnostic tool: if increasing depth improves forward expressivity yet training loss stagnates, vanishing gradients are the dominant culprit; if loss decreases but validation accuracy plateaus, over-smoothing is likely responsible. Distinguishing the two pathologies is essential for principled architecture selection in computational materials engineering.

Dependence on Graph Properties

The theoretical depth limit derived in Proposition 1 is not universal but depends explicitly on the spectral properties of the underlying crystal graph. Five graph-theoretic quantities modulate the largest eigenvalue p of the normalized adjacency matrix and, consequently, the rate of gradient decay.

Table 1 consolidates the graph-theoretic determinants of spectral contraction and clarifies how crystal-structure topology translates into trainable-depth limits in materials GNNs.

Table 1. Spectral determinants of trainable depth in crystal graphs: a theoretical consolidation framework

Graph-theoretic determinant	Spectral effect on normalized adjacency ($\tilde{A}$)	Consequence for gradient decay	Implication for trainable depth	Materials-specific interpretation	Design implication

Average degree / coordination number	Higher coordination alters the dominant eigenmode and strengthens persistent aggregation pathways near the top of the spectrum	Repeated backpropagation remains governed by a contraction factor tied to ρ , but dense local coupling can still intensify practically harmful attenuation through repeated mixing	Depth cannot be chosen generically; coordination environment changes how rapidly usable signal degrades across layers	Close-packed metals and densely coordinated lattices require especially careful depth control	Tune depth by material family; avoid transferring layer counts across chemically distinct datasets
Connectivity structure	More uniform global connectivity narrows structural bottlenecks and changes the spectral gap profile	The backward signal contracts according to the graph's leading spectral mode over longer propagation chains	Highly connected graphs still face exponential decay; sparse bridges or weakly coupled motifs can tighten effective limits further	Defective, porous, or low-dimensional crystals may become depth-fragile even when nominal graph size is moderate	Use topology-aware preprocessing or adaptive depth rather than fixed-depth architectures
Graph size / effective propagation distance	Matrix dimension increases without guaranteeing a favorable change in ρ	More layers are often demanded to traverse larger structures, but the same contraction law governs backward flow	The need for more hops rises faster than trainable depth capacity	Large supercells, disordered systems, and defect-rich structures are especially mismatched to naive deep message passing	Replace brute-force depth with coarsening, pooling, or explicit long-range channels
Normalization scheme	Symmetric versus random-walk normalization changes the operator spectrum and the dominance of leading modes	Different normalizations redistribute contraction across layers but do not remove spectral decay	Normalization can shift the severity of the depth ceiling, not abolish it	Standard stabilization choices may preserve forward behavior while still imposing backward penalties	Report normalization explicitly and interpret depth comparisons only within matched operators
Crystal symmetry	High symmetry concentrates representation dynamics into a small set of dominant spectral modes	Backward flow becomes increasingly governed by a small leading eigenspace	Trainable depth becomes sharply sensitive to a narrow spectral regime	Periodic high-symmetry lattices may appear structurally regular yet remain optimization-limited at depth	Pair symmetry-aware encodings with shallow propagation rather than deeper stacks

Spectral gap (1 – ρ)	Directly determines how quickly powers of $\tilde{A}$ become dominated by the leading mode	Smaller usable margin before gradients become numerically negligible	Provides the cleanest theoretical predictor of depth tolerance	Different crystal families imply different optimization ceilings even under identical GNN parameterization	Estimate spectral quantities before architecture selection
Material-family heterogeneity	Distinct crystal classes induce distinct spectral regimes	Gradient stability varies across compounds, not only across models	A single “best depth” is theoretically unjustified across datasets	Covalent, ionic, metallic, and disordered systems should not share a default depth assumption	Adopt per-dataset or per-structure depth policies

Average degree governs the spectral radius through its effect on row-stochastic normalization, as increased coordination amplifies the contribution of neighboring nodes to $\tilde{A}$. Dense metallic lattices, typically spanning coordination numbers of 8–12, consequently attain $\rho \approx 0.92$ –0.95, in contrast to covalent networks with degrees of 4–6 and $\rho \approx 0.85$ –0.90 [1, 9]. Under these conditions, gradient attenuation proceeds more rapidly in close-packed metals than in open-framework materials, reflecting a structurally induced constraint on depth.

A related implication emerges from graph connectivity, where fully connected structures exhibit a reduced spectral gap relative to disconnected or weakly bridged configurations. Although translational symmetry in periodic crystals ensures global connectivity, the introduction of voids or low-dimensional motifs perturbs this condition, widening the gap and accelerating gradient decay. This effect is particularly pronounced in metallic glasses, where disordered connectivity imposes stricter depth limitations than those observed in ordered ionic crystals [2].

Graph size introduces a more subtle constraint. Expanding unit cells or supercells increases the dimensionality of $\tilde{A}$ without necessarily shifting its dominant eigenvalue, yet the practical need to propagate information across larger atomic sets often encourages deeper architectures. The spectral bound indicates that such increases in depth are counterproductive, as decay remains governed by the same ρ irrespective of system size, rendering large crystalline systems effectively untrainable within standard message passing frameworks unless augmented by long-range mechanisms.

The normalization scheme further modulates this behavior. Symmetric normalization ($\tilde{A} = D^{(-1/2)} A D^{(-1/2)}$) generally produces a larger ρ than its random-walk counterpart ($\tilde{A} = D^{(-1)} A$), and while it is favored for forward stability in materials GNNs [4], it simultaneously intensifies gradient decay. Random-walk normalization mitigates this effect by reducing ρ , albeit at the cost of introducing asymmetry in information propagation, thereby exposing a trade-off between stability and gradient preservation.

Crystal symmetry sharpens these dynamics by concentrating the eigenvalue spectrum near its extremal values, which amplifies the dominance of ρ and accelerates exponential decay. High-symmetry systems such as cubic metals exhibit tighter spectral gaps than lower-symmetry structures, including triclinic or monoclinic compounds, resulting in more restrictive depth regimes [1].

Taken together, these dependencies indicate that optimal network depth is intrinsically material-specific. Uniform hyperparameter choices fail to accommodate the variability imposed by structural topology, necessitating per-structure estimation of the spectral radius as a computationally inexpensive preprocessing step to guide layer selection or residual scaling. This perspective reframes depth selection as a principled, structure-aware process, linking graph topology directly to gradient dynamics and advancing beyond heuristic tuning toward predictive architectural design in deep materials GNNs.

Relation to Other Theoretical Results

The depth limit established here extends classical vanishing-gradient analyses developed for recurrent neural networks, where the recurrent weight matrix plays the role of the repeated adjacency operator [5]. Xu et al. [5] demonstrated that the expressive power of recurrent architectures collapses when the spectral radius of the transition matrix lies below unity; the present result shows that message-passing GNNs inherit an identical contraction, the only difference being that the multiplier is data-dependent (the crystal graph) rather than learnable. This structural analogy explains why techniques successful in RNNs—gradient clipping, orthogonal initialization, or explicit memory gates—translate only partially to GNNs: the fixed topology prevents full orthogonalization of $\tilde{A}$.

The bound also complements over-smoothing analyses. Li et al. [10] and Rusch et al. [16] derived forward-pass convergence rates governed by the same ρ , yet focused on feature homogenization rather than backward decay. The current spectral decomposition unifies both perspectives: the identical eigendecomposition that produces over-smoothing in the forward direction simultaneously produces vanishing gradients in the backward direction. The materials-specific contribution lies in the typical range of ρ (0.8–0.95) for crystal graphs, which is tighter than the range encountered in citation or molecular graphs and therefore yields quantitatively stricter depth ceilings [1, 9].

Expressive-power results further contextualize the limit. Xu et al. [6] proved that the Weisfeiler-Lehman hierarchy requires depth linear in the graph diameter to distinguish non-isomorphic structures; for large crystals the diameter can exceed 20 hops, yet the gradient bound shows that such depth is unattainable. The ensuing depth-versus-width trade-off favors wider shallow networks augmented with explicit long-range features over naively deep stacks. Related graph neural tangent kernel analyses [27] confirm that the NTK remains well-conditioned only when gradient flow is preserved, reinforcing the necessity of the spectral-radius safeguard.

The key distinction of the present work is its exclusive focus on periodic crystal graphs and property-prediction objectives. Prior theoretical results addressed general graphs or node-classification tasks; the spectral properties of materials lattices—governed by translational symmetry and coordination chemistry—produce a narrower and more predictable family of ρ values, enabling the first quantitative, application-specific depth limit. This materials-centric framing bridges the gap between abstract GNN theory and the practical requirements of computational materials discovery.

Implications for Materials GNN Design

Table 2 translates the theoretical depth limit into an architectural decision matrix that distinguishes mitigation of backward gradient collapse from mitigation of forward over-smoothing.

Table 2. Architectural responses to the depth limit: a design matrix for trainable materials GNNs

Architectural intervention	Primary mechanism	Addresses vanishing gradients?	Addresses over-smoothing?	Theoretical advantage	Limitation under the present framework	Best-use scenario in materials prediction
Residual / skip connections	Provide additive short pathways that reduce reliance on long backward chains through repeated $\tilde{A}$ multiplications	Yes, partially	Sometimes	Preserve trainability of early layers by shortening effective gradient distance	Do not eliminate contraction in the message-passing branch itself	Moderate-depth models where local propagation remains useful

Normalization layers	Stabilize activation and feature scale across layers	Partially	Indirectly	Reduce compounding numerical instability and improve optimization robustness	Cannot remove the graph-operator contribution to spectral decay	Architectures needing stable training without substantial redesign
Shallow message passing + explicit long-range features	Offload distant interactions to engineered global encodings	Yes, effectively	Yes, often	Avoids demanding long hop-wise propagation for mesoscale effects	Requires careful feature design and may reduce end-to-end elegance	Large unit cells, defect structures, and long-range property signals
Graph coarsening / pooling	Reduce effective graph diameter before deep propagation is attempted	Yes	Yes	Shortens the path length required to integrate distant information	Risks loss of atomistic resolution if coarsening is too aggressive	Supercells and multi-scale materials systems
Equivariant but shallow architectures	Increase representation quality without relying on depth alone	Indirectly	Indirectly	Gains expressivity per layer, reducing pressure to deepen the network	May still inherit the same depth ceiling if message passing is stacked excessively	High-value property tasks requiring physically informed local structure encoding
Adaptive-depth models	Set layer count conditional on graph topology or estimated spectral regime	Yes	Potentially	Aligns architecture with structure-specific trainability	Requires preprocessing, control logic, and reproducible selection criteria	Mixed-material datasets with wide topological heterogeneity
Deeper plain message passing	Increase hop range by stacking standard aggregation layers	No	No	Nominally expands receptive field	Directly collides with the spectral depth ceiling derived in this manuscript	Generally unsuitable beyond shallow-to-moderate depth
Over-smoothing-only remedies	Preserve feature diversity in the forward pass	Not necessarily	Yes	Useful when representation collapse dominates	Insufficient when optimization failure originates from backward contraction	Cases where training continues but validation degrades with depth

For model developers the depth limit imposes a hard architectural constraint: depths beyond approximately 10–20 layers are provably counterproductive for typical crystal graphs. Rather than stacking additional message-passing layers, developers should prioritize residual connections that bypass the adjacency multiplier, normalization layers that modulate feature scale without inflating p , and hybrid designs that combine shallow message passing with explicit long-range

encodings (e.g., Fourier or distance-based features) [28, 29]. These interventions do not eliminate the spectral contraction but restore usable gradient flow, allowing the network to exploit depth where it is theoretically justified.

For practitioners the bound supplies a practical decision rule. Small unit cells (<20 atoms) require only 3–6 layers to capture all relevant interactions; deeper models waste compute and risk training collapse. Large supercells (>50 atoms) or defective structures should avoid deep message passing altogether and instead employ graph coarsening or multi-scale pooling to reduce effective diameter while preserving gradient integrity. Pre-computing the spectral radius of each training graph offers an inexpensive oracle for selecting per-sample depth or residual strength.

For benchmark designers the analysis mandates reporting performance as an explicit function of depth, including ablation studies that isolate the contribution of vanishing gradients from over-smoothing. Depth-sensitivity curves, together with measured p values for each dataset, enable fair comparison across architectures and prevent the community from rewarding models that appear superior only because they operate below the theoretical limit. Such standardized diagnostics will accelerate convergence on depth-aware best practices rather than opaque scaling experiments.

Collectively these implications redirect the field from brute-force depth scaling toward spectrally informed design. The theoretical ceiling derived here is not a counsel of despair but a design compass: by acknowledging the intrinsic contraction imposed by crystal topology, researchers can engineer GNNs that are both expressive and trainable, thereby realizing the full promise of graph-based methods for materials property prediction.

Conclusion

This work has shown that vanishing gradients in deep materials GNNs are not merely an optimization inconvenience but an intrinsic consequence of adjacency-driven message passing on crystal graphs. Once the normalized aggregation operator is applied repeatedly, backward signals contract at a rate governed by the spectral radius ρ , imposing a strict ceiling on trainable depth. The result is especially consequential in materials prediction, where the desire to capture long-range structural effects often encourages deeper architectures even though the underlying graph dynamics make such scaling self-defeating.

The theoretical contribution is twofold. It provides a closed spectral interpretation of gradient decay in crystal graph networks and, in doing so, separates backward contraction from the forward homogenization associated with over-smoothing. This distinction matters because the two phenomena may coexist while demanding different remedies. The analysis also shows that the depth limit is not universal across materials classes but varies systematically with graph topology, including coordination number, connectivity profile, symmetry, normalization choice, and effective propagation distance. Depth selection therefore cannot be reduced to a fixed hyperparameter transferred across datasets or crystal families.

The design implications follow directly. Plainly increasing the number of message-passing layers is unlikely to improve property prediction once the spectral regime of the graph has been ignored. More promising directions lie in architectures that preserve gradient accessibility while extending physical reach, including residual or skip-based formulations, topology-aware adaptive depth, graph coarsening, and hybrid models that supplement shallow local propagation with explicit long-range encodings. In this sense, the derived bound does not argue against expressive materials GNNs; it defines the conditions under which expressivity remains trainable.

More broadly, the analysis advances a shift from empirical depth tuning toward predictive architectural reasoning in AI for materials science. By grounding trainability in spectral graph structure rather than heuristic trial-and-error, it offers a principled framework for evaluating scalability in future crystal learning models. The broader significance is clear: progress in graph-based materials discovery will depend less on stacking deeper networks than on aligning model design with the mathematical constraints imposed by the materials graphs themselves.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 08 Dec 2022

Revised: 26 Feb 2023

Accepted: 16 May 2023

Published online: 18 July 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater*. 2019;31(9):3564-72.
<https://doi.org/10.1021/acs.chemmater.9b01294>.
- Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett*. 2018;120(14):145301.
<https://doi.org/10.1103/PhysRevLett.120.145301>.
- Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst*. 2021;32(1):4-24.
<https://doi.org/10.1109/TNNLS.2020.2978386>.
- Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun*. 2022;13(1):2453.
<https://doi.org/10.1038/s41467-022-29939-5>.
- Xu K, Li J, Zhang M, Du SS, Kawarabayashi KI, Jegelka S. What can neural networks reason about? [Preprint]. *arXiv*; 2019.
<https://doi.org/10.48550/arXiv.1905.13211>.

Xu K, Hu W, Leskovec J, Jegelka S. How powerful are graph neural networks? [Preprint]. arXiv; 2018.
<https://doi.org/10.48550/arXiv.1810.00826>.

Alon U, Yahav E. On the bottleneck of graph neural networks and its practical implications [Preprint]. arXiv; 2020.
<https://doi.org/10.48550/arXiv.2006.05205>.

Zhang B, Luo S, Wang L, He D. Rethinking the expressive power of GNNs via graph biconnectivity [Preprint]. arXiv; 2023.
<https://doi.org/10.48550/arXiv.2301.09505>.

Omee SS, Louis SY, Fu N, Wei L, Dey S, Dong R, et al. Scalable deeper graph neural networks for high-performance materials property prediction. *Patterns*. 2022;3(5):100491.
<https://doi.org/10.1016/j.patter.2022.100491>.

Li Q, Han Z, Wu XM. Deeper insights into graph convolutional networks for semi-supervised learning. *Proc AAAI Conf Artif Intell*. 2018;32(1):3538-45.
<https://doi.org/10.1609/aaai.v32i1.11604>.

Jaiswal A, Wang P, Chen T, Rousseau J, Ding Y, Wang Z. Old can be gold: Better gradient flow can make vanilla-GCNs great again. *Adv Neural Inf Process Syst*. 2022;35:7561-74.
<https://doi.org/10.52202/068431-0549>.

Chen M, Wei Z, Huang Z, Ding B, Li Y. Simple and deep graph convolutional networks. *Proc Mach Learn Res*. 2020;119:1725-35.

Li G, Müller M, Thabet A, Ghanem B. DeepGCNs: Can GCNs go as deep as CNNs? *Proc IEEE/CVF Int Conf Comput Vis*. 2019;9267-76.
<https://doi.org/10.1109/ICCV.2019.00936>.

Liu M, Gao H, Ji S. Towards deeper graph neural networks. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*; 2020. New York: ACM; 2020. p. 338-48.
<https://doi.org/10.1145/3394486.3403076>.

Gasteiger J, Weißenberger S, Günnemann S. Diffusion improves graph learning. *Adv Neural Inf Process Syst*. 2019;32:13366-78.
<https://doi.org/10.5555/3454287.3455484>.

Rusch TK, Bronstein MM, Mishra S. A survey on oversmoothing in graph neural networks [Preprint]. arXiv; 2023.
<https://doi.org/10.48550/arXiv.2303.10993>.

Jiang B, Zhang Z, Lin D, Tang J, Luo B. Semi-supervised learning with graph learning-convolutional networks. *Proc IEEE/CVF Conf Comput Vis Pattern Recognit*. 2019:11313-20.
<https://doi.org/10.1109/CVPR.2019.01157>.

Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. *Proc Mach Learn Res*. 2017;70:1263-72.

Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. In: *International conference on learning representations*; 2018.

Di Giovanni F, Rowbottom J, Chamberlain BP, Markovich T, Bronstein MM. Graph neural networks as gradient flows: Understanding graph convolutions via energy [Preprint]. arXiv; 2022.
<https://doi.org/10.48550/arXiv.2206.10991>.

Bezdan T, Bačanić Džakula N. Convolutional neural network layers and architectures. In: Sinteza 2019 - International scientific conference on information technology and data related research; 2019; Belgrade, Serbia. Belgrade: Singidunum University; 2019. p. 445-51.

<https://doi.org/10.15308/Sinteza-2019-445-451>.

Chen D, Lin Y, Li W, Li P, Zhou J, Sun X. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *Proc AAAI Conf Artif Intell.* 2020;34(04):3438-45.

<https://doi.org/10.1609/aaai.v34i04.5747>.

Rong Y, Huang W, Xu T, Huang J. DropEdge: Towards deep graph convolutional networks on node classification [Preprint]. *arXiv*; 2019.

<https://doi.org/10.48550/arXiv.1907.10903>.

Wu F, Souza A, Zhang T, Fifty C, Yu T, Weinberger K. Simplifying graph convolutional networks. *Proc Mach Learn Res.* 2019;97:6861-71.

Hoang NT, Maehara T. Revisiting graph neural networks: All we have is low-pass filters [Preprint]. *arXiv*; 2019.

<https://doi.org/10.48550/arXiv.1905.09550>.

Ma Y, Liu X, Zhao T, Liu Y, Tang J, Shah N. A unified view on graph neural networks as graph signal denoising. In: *Proceedings of the 30th ACM international conference on information & knowledge management*; 2021. New York: ACM; 2021. p. 1202-11.

<https://doi.org/10.1145/3459637.3482225>.

Bodnar C, Di Giovanni F, Chamberlain BP, Liò P, Bronstein MM. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in GNNs. *Adv Neural Inf Process Syst.* 2022;35:18527-41.

<https://doi.org/10.52202/068431-1346>.

Zhao L, Härtel L, Shah N, Akoglu L. A practical, progressively-expressive GNN. *Adv Neural Inf Process Syst.* 2022;35:34106-20.

<https://doi.org/10.52202/068431-2472>.

Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, et al. Strategies for pre-training graph neural networks [Preprint]. *arXiv*; 2019.

<https://doi.org/10.48550/arXiv.1905.12265>.