STRUCTURAL PRESERVATION AND THE LOGICAL EXPRESSIVENESS OF GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Bridges between graph neural networks (GNNs) and logical formalisms have been established by fixing architectural choices, such as the types of aggregation, combination, and activation functions. These choices define restricted classes of GNNs for which tight correspondences with logical formalisms can be obtained, by showing that logical formulae can be translated into equivalent GNNs and, conversely, that GNNs can be translated into equivalent formulae.

In this paper we take a semantic perspective by establishing the logical expressiveness of classes of GNN classifiers that are preserved under structural properties: embeddings (extensions), injective homomorphisms, and homomorphisms. We show that, for each such property, there exists a fragment of graded modal logic characterising the class of GNNs. In particular, preservation under embeddings, injective homomorphisms, and homomorphisms corresponds to existential graded modal logic, its existential-positive fragment, and existential-positive modal logic, respectively. These results characterise the expressiveness of broad classes of GNNs independently of specific architectural choices, but we also show that each of these classes admits a GNN architecture of the same expressiveness.

1 INTRODUCTION

Graph neural networks (GNNs) Gilmer et al. (2017) are models designed to operate directly on graphs of variable size while ensuring that model predictions are invariant under graph isomorphisms. In the aggregate-combine paradigm, each node maintains a vector representation that is iteratively updated by aggregating information from its neighbours. This design induces a strong structural bias: GNNs with L layers are invariant under isomorphisms and inherently local, in the sense that node representations depend only on the node's L-hop neighbourhood. These two properties bring GNNs close to logical formalisms, in particular modal logics, which are also local and isomorphism-invariant Barceló et al. (2020); Hauke & Wałęga (2026); Cuenca Grau et al. (2026). Establishing such connections is beneficial, as logic provides principled tools for analysing expressiveness and enabling symbolic methods for verification and explanation.

These connections have been studied extensively. In particular, bridges between GNNs and logic have been established by fixing architectural choices, such as aggregation, combination, and activation functions. These choices define restricted families of GNN classifiers for which precise correspondences with logical formalisms can be obtained. In particular, prior work has identified such families corresponding to extensions of first-order logic with counting terms Grohe (2024); Nunn et al. (2024) and Presburger quantifiers Benedikt et al. (2024), as well as to rule-based knowledge representation formalisms such as Datalog Tena Cucala et al. (2025); Tena Cucala et al. (2023).

In this paper, we complement such architecture-driven characterisations with a semantic framework based on structural properties that impose strictly stronger constraints on classifiers than invariance under graph isomorphisms, namely preservation under embeddings, injective homomorphisms, and homomorphisms. We show that for each such property, the corresponding class of GNN classifiers has the same expressiveness as a fragment of graded modal logic (Theorem 7, Corollary 8):

- The class of GNN classifiers preserved under **embeddings** has the same expressiveness as existential graded modal logic $\exists\mathcal{GML}$,
- The class of GNN classifiers preserved under **injective homomorphisms** has the same expressiveness as existential-positive graded modal logic $\exists^{+}\mathcal{GML}$, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- The class of GNN classifiers preserved under **homomorphisms** has the same expressiveness as existential-positive modal logic $\exists^{+}\mathcal{ML}$.

Our approach is inspired by finite model theory, where preservation theorems relate structural properties to logical definability Libkin (2004); Rosen (2002); Grädel et al. (2007); Rosen (1997); Abramsky & Reggio (2024). Classical results include the van Benthem–Rosen theorem Rosen (1997), which characterises modal logic as the bisimulation-invariant fragment of first-order logic, and Rossman's theorem Rossman (2008), which characterises existential-positive first-order logic as the fragment preserved under homomorphisms. We adapt this methodology to the setting of GNNs. Since existing preservation theorems do not directly apply in this setting, we develop new results tailored to graph structures induced by GNN computations.

The key intuition underlying our approach is that, in addition to being invariant under isomorphisms, GNNs are inherently local: in a GNN with L layers, the representation of a node depends only on its L-hop neighbourhood Barceló et al. (2020); Wang & Zhang (2022); Morris et al. (2019a). This locality allows us to reduce the analysis of GNNs to trees obtained via unravelling Barceló et al. (2020). Combined with preservation under structural relations, it yields the logical characterisations stated above.

The main technical challenge is to obtain a finite representation of such classes. To address this, we show a new result on well-quasi-orders (Theorem 4) which is of independent interest:

- Every class of trees of bounded height is **well-quasi-ordered** by the embedding relation.

Since L-unravellings have bounded height and well-quasi-orders admit only finitely many minimal elements, this yields a finite representation of each class via its minimal trees. At a high level, each minimal tree can be characterised by a logical formula, and the entire class is defined by a finite disjunction of such formulae. The fragment of modal logic is determined by the preservation relation.

In addition to the semantic characterisations, we show that each class of GNN classifiers defined via preservation properties can be captured by a specific architecture. In particular (Theorem 16):

- GNNs preserved under **injective homomorphisms** have the same expressiveness as monotonic GNNs (i.e., GNNs with non-decreasing aggregation and combination functions),
- GNNs preserved under **homomorphisms** have the same expressiveness as monotonic GNNs using MAX as aggregation,
- GNNs preserved under **embeddings** have the same expressiveness as monotonic GNNs with an additional layer that transforms node features independently of their neighbours.

Moreover, our results reveal an interesting property of GNN architectures considered in prior work, which we refer to as positive-weight GNNs, whose parameter matrices have non-negative entries. In particular, it follows that any GNN preserved under injective homomorphisms (and so, also any monotonic GNN) can be transformed into an equivalent positive-weight GNN. Thus, with respect to expressiveness, the syntactic restriction to non-negative weights is not restrictive within this class.

In summary, our contributions are as follows. We develop a framework for analysing the expressiveness of GNNs based on structural preservation properties. We establish preservation theorems linking these properties to definability in fragments of graded modal logic, supported by new well-quasi-ordering results ensuring finite representations. Finally, we provide architectural characterisations of the GNN classes defined by preservation properties.

2 Related Work

The expressiveness of GNNs has been studied from several perspectives. Their *discriminative power* measures their ability to distinguish graphs, and it is well known that message-passing GNNs are bounded by the Weisfeiler–Lehman (WL) graph isomorphism test. Thus, they cannot distinguish graphs that WL fails to separate Morris et al. (2019b); Xu et al. (2019). Work on the logical expressiveness of GNNs treats them as node classifiers that compute unary queries over graphs and establishes correspondences between GNN architectures and logical formalisms. In particular, GNN expressiveness is not captured by first-order logic alone and requires extensions with counting terms, Presburger quantifiers, or related mechanisms Grohe (2024); Benedikt et al. (2024); Nunn et al. (2024). Furthermore, connections to graded modal logic and related fragments have been

established Barceló et al. (2020). Recent work on bounded GNNs refines these correspondences by relating aggregation schemes with bounded multiplicities to two-variable fragments of first-order logic Cuenca Grau et al. (2026).

A complementary line of work studies connections between GNNs and rule-based knowledge representation formalisms motivated by logical reasoning and explainability. In particular, monotonic GNNs have been introduced to capture rule-based inference in Datalog Tena Cucala et al. (2025); Tena Cucala et al. (2023); Tena Cucala & Cuenca Grau (2024); Morris & Horrocks (2025); Morris et al. (2024). These models enforce structural preservation properties, such as preservation under homomorphisms or injective homomorphisms. These properties are enforced through architectural constraints, including non-negative weights, monotone activation functions, and restricted aggregation. While effective in practice, these approaches are inherently syntax-driven, as the desired preservation properties are enforced indirectly through design choices rather than characterised explicitly in semantic terms.

Our work is also related to preservation theorems in finite model theory, which characterise logical fragments in terms of structural preservation properties and, in some cases, invariance conditions. Classical results such as Łoś-Tarski, Lyndon, and Homomorphism Preservation theorems establish correspondences between fragments of first-order logic and preservation under embeddings, surjective homomorphisms, and homomorphisms, respectively Łoś & Suszko (1955); Lyndon (1959); Hodges (1997). Van Benthem's theorem (and its finite variant due to Rosen) characterises bisimulation-invariant first-order properties as those definable in modal logic Andréka et al. (1998); Rosen (1997). More recent work has extended preservation results to graded modal logic and related settings Abramsky & Reggio (2024). Our results on well-quasi-orders are related to recent work in a similar direction Lopez (2023), however, these techniques rely on additional structural assumptions, such as closure under substructures, that do not hold in our setting.

3 BACKGROUND

Graphs and Graph Relations We consider finite, undirected, simple, node-labelled graphs $G = (V, E, \lambda)$, where V is a finite set of nodes, E a set of undirected edges without self-loops, and $\lambda : V \to \{0,1\}^d$ assigns to each node a binary vector of fixed dimension d. Intuitively, d represents the number of possible labels (propositional features) and $\lambda_i(v) = 1$, for a label $i \in [d]$, indicates that node v has label i; the set $\{1, \ldots, n\}$ is represented as $[n]$. We write vectors in bold, e.g., $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, and use x_i to denote the i-th component of $\mathbf{x}$. For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we write $\mathbf{x} \leq \mathbf{y}$ if $x_i \leq y_i$ for all $i \in [d]$. This discrete labelling enables a direct connection with logical formalisms Barceló et al. (2020); Benedikt et al. (2024).

A *pointed graph* is a pair (G, v) consisting of a graph and a distinguished node. A *tree* is a connected acyclic graph. A *rooted tree* is a pointed graph (G, v) where G is a tree and v is its root. The *depth* of a node u in a pointed tree is the length of the unique path from the root to u, and the *height* of a tree G is the maximum depth of any node.

Let (G, w) and (H, v) be pointed graphs, with $G = (V, E, \lambda)$ and $H = (V', E', \lambda')$. We will consider the following definitions of mappings $f : V \to V'$, with $f(w) = v$. *Isomorphism:* f is a bijection, $\lambda(u) = \lambda'(f(u))$, and $\{u, z\} \in E$ iff $\{f(u), f(z)\} \in E'$. *Embedding:* f is an isomorphism from (G, w) onto the subgraph of H induced by $f(V)$. *Homomorphism:* f satisfies $\lambda(u) \leq \lambda'(f(u))$ for all $u \in V$, and $\{u, z\} \in E$ implies $\{f(u), f(z)\} \in E'$. If f is also injective, we call it an *injective homomorphism*.

Intuitively, these mappings capture different ways in which one graph can be embedded into another: isomorphisms preserve the entire structure, embeddings preserve a subgraph exactly, while homomorphisms allow labels to be extended and nodes to be merged (unless the mapping is injective). For examples of these relations see Figure 1. Clearly, every isomorphism is an embedding, every embedding is an injective homomorphism, and every injective homomorphism is a homomorphism.

Each notion above induces a corresponding binary relation $\preceq$ on pointed graphs, such that $(G, w) \preceq (H, v)$ if there exists a mapping of the corresponding type from (G, w) to (H, v).

Let $\mathcal{C}$ be a class of pointed graphs and $\preceq$ a binary relation as above. Class $\mathcal{C}$ is *preserved* under $\preceq$ if $(G, w) \in \mathcal{C}$ and $(G, w) \preceq (H, v)$ implies that $(H, v) \in \mathcal{C}$. Class $\mathcal{C}$ is *invariant* under $\preceq$ if $\mathcal{C}$

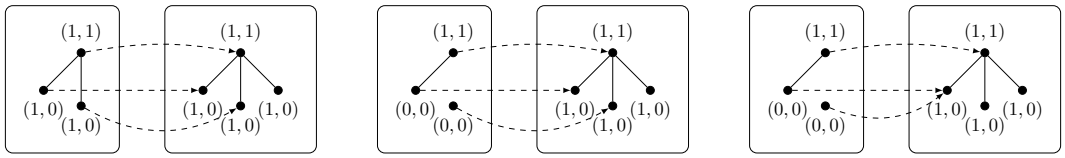


Figure 1: Embedding (a), injective homomorphism (not an embedding) (b), and a homomorphism which is not injective (c)

is preserved under both $\preceq$ and its inverse relation $\preceq^{-1}$. These preservation properties capture the robustness of classifiers under structural transformations of the input graph.

Graph Neural Networks We consider node classification with graph neural networks (GNNs) based on the standard *aggregate-combine* paradigm Benedikt et al. (2024); Barceló et al. (2020). GNNs iteratively update node representations by aggregating information from neighbours and combining it with the current representation.

Formally, a GNN layer is a pair $(\mathsf{agg}, \mathsf{comb})$, where agg is an *aggregation function* mapping multisets of vectors into vectors and comb is a *combination function* mapping pairs of vectors (or their concatenation) to vectors. Applying a layer to $G = (V, E, \lambda)$ produces a graph $G' = (V, E, \lambda')$ with the same nodes and edges, but with updated node features.

For each node $v \in V$, the updated label is defined by

$$\lambda'(v) = \mathsf{comb}\Big(\lambda(v), \mathsf{agg}(\{\!|\lambda(w)|\!\}_{w \in N_G(v)})\Big), \tag{1}$$

where $N_G(v) = \{w \mid \{v, w\} \in E\}$ denotes the set of neighbours of v, and $\{\!|\cdot|\!\}$ is a multiset. Aggregation and combination functions have fixed input and output dimensions and must be compatible; if agg maps vectors of dimension d to dimension d', then comb takes as input the concatenation of vectors of dimensions d and d'.

Often comb is a one-layer MLP with a componentwise activation σ (e.g., ReLU or truncated ReLU):

$$\lambda'(v) = \sigma\Big(\mathbf{b} + \lambda(v)\mathbf{A} + \mathsf{agg}\big(\{\!|\lambda(w) \mid w \in N_G(v)|\!\}\big)\mathbf{C}\Big), \tag{2}$$

where $\mathbf{b}$ is a bias vector, and $\mathbf{A}$ and $\mathbf{C}$ are real-valued matrices. Following previous work Barceló et al. (2020), we refer to such layers as *simple*. Common choices for agg include position-wise SUM, MAX, and MEAN although MEAN is not monotone with respect to the multiset order used later. Another aggregation is MAX-k-SUM Tena Cucala et al. (2023) which sums the k largest values per component.

A *GNN classifier* $\mathcal{N}$ of dimension d consists of L aggregate-combine layers followed by a classification function cls mapping vectors to truth values. In this paper, we consider only threshold-based classification functions satisfying $\mathsf{cls}(x) = \mathsf{true}$ if $x \geq t$ for some fixed threshold t and false otherwise. This restriction is without loss of generality for the expressiveness results proved in this paper. The output dimension of each layer matches the input dimension of the next. We write $\lambda^{(\ell)}(v)$ for the vector associated with node v after applying layer ℓ, with $\lambda^{(0)}(v) = \lambda(v)$ and $\lambda^{(L)}(v)$ the final representation. The output of $\mathcal{N}$ on (G, v) is the truth value $\mathcal{N}(G, v) = \mathsf{cls}\big(\lambda^{(L)}(v)\big)$.

Logical Expressiveness A *node classifier* is a function mapping pointed graphs to truth values true (accepted) or false (rejected). Two classifiers are *equivalent* if they accept exactly the same pointed graphs. We compare the expressiveness of families of classifiers arising from GNNs or logical formulae. We say that two families of classifiers have the same expressiveness if each classifier in one family has an equivalent classifier in the other, and vice versa.

4 A Logical Toolkit for Analysing Expressiveness

We introduce logical and combinatorial tools that underpin our analysis. Our approach is based on relating structural preservation properties to definability in fragments of graded modal logic.

4.1 Locality and Unravelling

GNNs are inherently local: after L layers, a node representation depends only on its L-hop neighbourhood. This locality can be captured by the notion of *unravelling*, which transforms a pointed graph (G, v) into a rooted tree $\mathsf{Unr}^L(G, v)$ whose nodes correspond to paths in G starting at v, of length at most L. The L-unravelling unfolds all paths of length up to L into a tree of height at most L. This construction mirrors the computation of a GNN, where each node aggregates information from its neighbourhood in a tree-shaped manner. An example of a pointed graph and its 2-unravelling is shown in Figure 2. Note that the unravelling unfolds cycles into distinct paths, thereby yielding a tree.

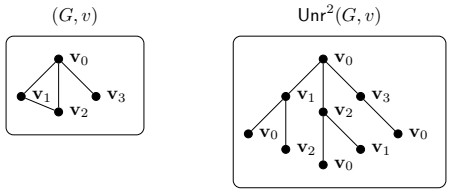


Figure 2: A pointed graph and its 2-unravelling

Importantly, GNNs with L layers are invariant under this transformation: evaluating a GNN with L layers on a graph yields the same result as evaluating it on its L-unravelling. This observation will be central to our analysis, as it allows us to reason about GNNs using tree-shaped structures, and has been exploited in prior work to relate GNNs to logical formalisms Barceló et al. (2020).

We now give the formal definition of unravelling.

Definition 1. *Let (G, v) be a pointed graph with $G = (V, E, \lambda)$, and let $L \in \mathbb{N}$. The* L-unravelling *of (G, v), denoted $\mathsf{Unr}^L(G, v)$, is the pointed graph $((V', E', \lambda'), v')$ defined as follows:*

- *V' contains a node for each sequence $(v, v_1, \ldots, v_\ell)$ of length $\ell \leq L$ such that $\{v_i, v_{i+1}\} \in E$ for all $i < \ell$;*
- *v' corresponds to the sequence (v) of length 0;*
- *E' has an edge between $(v, v_1, \ldots, v_{\ell-1})$ and $(v, v_1, \ldots, v_\ell)$ iff $\{v_{\ell-1}, v_\ell\} \in E$;*
- *λ' is defined by $\lambda'((v, v_1, \ldots, v_\ell)) = \lambda(v_\ell)$.*

Thus, $\mathsf{Unr}^L(G, v)$ is a tree of height at most L, whose nodes represent all paths of length at most L starting from v. This construction captures the locality of GNN computations: GNN classifiers are invariant under unravelling, as formalised next (see also (Barceló et al., 2020, Proposition C.4)).

Proposition 2. *Let $\mathcal{N}$ be an aggregate-combine GNN with L layers. Then, for every pointed graph (G, v) and every $K \geq L$, we have $\mathcal{N}(G, v) = \mathcal{N}(\mathsf{Unr}^K(G, v))$.*

Modal logics exhibit an analogous invariance property: the truth of a modal formula of depth L at a node depends only on its L-unravelling.

4.2 Modal Logic

Formulae of graded modal logic $\mathcal{GML}$ de Rijke (2000); Tobies (2001); Otto (2019) are defined inductively by the grammar $\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \Diamond^{\geq k}\varphi$, where p ranges over propositions p_i and $k \in \mathbb{N}$. The *depth* of φ is the maximum nesting depth of its modal operators. The *existential* fragment $\exists\mathcal{GML}$ restricts negation to occur in front of propositions, while the *existential-positive* fragment $\exists^+\mathcal{GML}$ is negation-free. The logic $\mathcal{ML}$ restricts the graded modal operators to $k = 1$, and the fragments $\exists\mathcal{ML}$ and $\exists^+\mathcal{ML}$ of $\mathcal{ML}$ are defined analogously.

We interpret formulae over pointed graphs[1] (G, v) with $G = (V, E, \lambda)$. We assume that each proposition p_i corresponds to the i-th component of the node feature vector.

[1]Modal formulae are typically interpreted over Kripke structures (directed graphs); in our setting, we work with undirected graphs by viewing the edge relation as symmetric.

The *satisfaction relation* $\models$ is defined inductively as follows:

$$
\begin{array}{lll}
(G,v) \models p_i & \text{iff} & \lambda(v)_i = 1 \\
(G,v) \models \neg\varphi & \text{iff} & (G,v) \not\models \varphi \\
(G,v) \models \varphi \wedge \psi & \text{iff} & (G,v) \models \varphi \text{ and } (G,v) \models \psi \\
(G,v) \models \varphi \vee \psi & \text{iff} & (G,v) \models \varphi \text{ or } (G,v) \models \psi \\
(G,v) \models \Diamond^{\geq k}\varphi & \text{iff} & \text{there exist at least } k \text{ neighbours } w \text{ of } v \text{ such that } (G,w) \models \varphi
\end{array}
$$

For example, the formula $p_1 \wedge \Diamond^{\geq 2} p_2$ holds in a pointed graph (G, v) if v satisfies p_1 (i.e., $\lambda(v)_1 = 1$) and it has at least two neighbours satisfying p_2 (i.e., their second component is 1). This formula belongs to $\exists\mathcal{GML}$ (and even to $\exists^{+}\mathcal{GML}$), but is not expressible in $\mathcal{ML}$.

It is well known that for each $\mathcal{GML}$ formula φ of depth L and each pointed graph (G, v), we have

$$(G,v) \models \varphi \quad \text{iff} \quad \mathsf{Unr}^{L}(G,v) \models \varphi.$$

That is, the class of pointed graphs satisfying φ is invariant under L-unravelling.

Given a modal logic $\mathcal{L}$, each formula $\varphi \in \mathcal{L}$ defines a node classifier by mapping a pointed graph (G, v) to true if $(G, v) \models \varphi$, and to false otherwise. We refer to such classifiers as $\mathcal{L}$-*classifiers*. A class $\mathcal{C}$ of pointed graphs is *definable* in a modal logic $\mathcal{L}$ if there exists a formula $\varphi \in \mathcal{L}$ such that $(G, v) \models \varphi$ if and only if $(G, v) \in \mathcal{C}$.

4.3 Well-quasi-orders

A key step in our approach is to ensure that only finitely many structures need to be considered when characterising a class of graphs. This finiteness property is guaranteed by well-quasi-orders (wqos).

A *quasi-order* (or *preorder*) $(A, \preceq)$ is a set A equipped with a binary relation $\preceq$ that is reflexive ($a \preceq a$ for all $a \in A$) and transitive ($a \preceq b$ and $b \preceq c$ implies $a \preceq c$). A quasi-order $(A, \preceq)$ is a *well-quasi-order* if every infinite sequence $a_1, a_2, \ldots$ in A contains indices $i < j$ such that $a_i \preceq a_j$. Intuitively, this rules out infinite antichains and ensures that every infinite sequence contains an increasing pair. In particular, any subset of a wqo has only finitely many $\preceq$-minimal elements (up to equivalence), which will be crucial for our results.

A central tool in our analysis is Higman's Lemma, which ensures that the wqo property is preserved when extended to sequences. Let A^* be the set of all finite sequences over A, and let $\preceq^*$ be the *subsequence order*, defined by $(a_1, \ldots, a_n) \preceq^* (b_1, \ldots, b_m)$ if there exists a strictly increasing map $f : \{1, \ldots, n\} \to \{1, \ldots, m\}$ such that $a_i \preceq b_{f(i)}$ for all $i \in \{1, \ldots, n\}$. For example, $(a_1, a_2) \preceq^* (b_1, b_2, b_3)$ holds if $a_1 \preceq b_1$ and $a_2 \preceq b_3$.

Lemma 3 (Higman's Lemma)**.** *If $(A, \preceq)$ is a wqo, then $(A^*, \preceq^*)$ is also a wqo.*

4.4 A New Well-Quasi-Ordering Result for Trees

We now apply the wqo framework to trees, in particular to those obtained via unravelling, although the results of this section apply more generally. We show that any (possibly infinite) class of trees obtained by L-unravelling has finitely many minimal trees with respect to embedding. This finiteness is crucial, as it allows us to represent an infinite class using only finitely many minimal elements. To establish this, we prove a stronger result: the embedding relation on any class of trees of bounded height is a wqo. Since L-unravellings have height at most L, and every wqo admits finitely many minimal elements, this result yields the desired finiteness.

This is the main technical result underlying our framework. Beyond its role in establishing our preservation theorems, it is of independent interest in the study of well-quasi-orders.

Intuitively, trees of bounded height can be decomposed into their immediate subtrees, which themselves have smaller height. This recursive structure allows one to reduce embeddings between trees to embeddings between sequences. Using Higman's Lemma we can ensure that such sequences form a well-quasi-order, yielding the result.

Theorem 4. *The embedding relation on any class of rooted trees, whose heights are uniformly bounded by a fixed constant, is a well-quasi-order.*

As a consequence, the same result holds for injective homomorphisms and homomorphisms, since these relations extend the embedding relation.

Corollary 5. *The injective homomorphism and homomorphism relations on any class of rooted trees of bounded height are well-quasi-orders.*

Any such class admits finitely many minimal trees, which provide a finite representation of the entire (infinite) class. This property underpins the preservation theorems developed in the next section. The bounded-height assumption is essential: without it, the wqo property fails.

Theorem 6. *The embedding, injective homomorphism, and homomorphism relations on classes of rooted trees of unbounded height are not, in general, well-quasi-orders.*

For example, consider the case of injective homomorphisms and the class $\mathcal{T}$ of rooted trees shown in Figure 3, where nodes are unlabelled. For any $n \neq m$, there is neither an injective homomorphism from (G_n, v_1) to (G_m, v_1), nor vice-versa. Intuitively, the position of the branching nodes depends on the height of the tree, making these structures incomparable under injective homomorphisms. Thus, $\mathcal{T}$ forms an infinite antichain with respect to injective homomorphisms, and so, it is not a wqo.

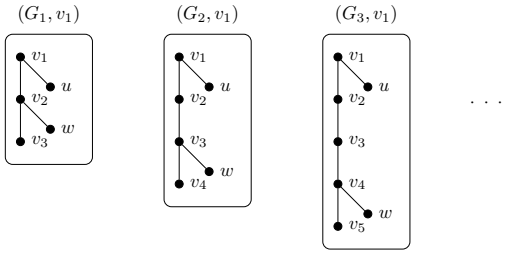


Figure 3: An infinite antichain of rooted trees with respect to the injective homomorphism

4.5 Preservation Theorems

We now combine the well-quasi-ordering results from the previous section with unravelling invariance to obtain preservation theorems.

At a high level, the idea is as follows. Let $\mathcal{C}$ be a class of pointed graphs invariant under L-unravelling and preserved under a structural relation $\preceq$ (such as embedding, injective homomorphism, or homomorphism). Then membership of a pointed graph in $\mathcal{C}$ is determined by its L-unravelling, which is a tree of bounded height. By the wqo results, such trees admit a finite set of $\preceq$-minimal elements. This finiteness allows us to represent $\mathcal{C}$ using a finite logical description: each minimal tree can be characterised by a formula, and the class $\mathcal{C}$ is defined by a finite disjunction of such formulae.

Each of the relations $\preceq$ corresponds to a fragment of modal logic: embeddings, injective homomorphisms, and homomorphisms correspond to $\exists\mathcal{GML}$, $\exists^+\mathcal{GML}$, and $\exists^+\mathcal{ML}$, respectively. We formalise this in the following theorem.

Theorem 7. *Let $\mathcal{C}$ be a class of pointed graphs, $L \in \mathbb{N}$, and let $\preceq$ be one of the following relations: embedding, injective homomorphism, or homomorphism. Then the following are equivalent:*

- *$\mathcal{C}$ is invariant under L-unravelling and preserved under $\preceq$*
- *$\mathcal{C}$ is definable by a formula of depth at most L in a corresponding logic: $\exists\mathcal{GML}$ if $\preceq$ is embedding, $\exists^+\mathcal{GML}$ if $\preceq$ is injective homomorphism, and $\exists^+\mathcal{ML}$ if $\preceq$ is homomorphism.*

Since L-layer GNN classifiers are invariant under L-unravelling, these preservation theorems yield a semantic characterisation of GNN expressiveness.

Corollary 8. *Let $\mathcal{N}$ be a GNN with L layers, let $\mathcal{C}$ be the class of pointed graphs accepted by $\mathcal{N}$, and let $\preceq$ be one of the relations: embedding, injective homomorphism, or homomorphism. Then the following are equivalent:*

- *$\mathcal{C}$ is preserved under $\preceq$*
- *$\mathcal{C}$ is definable by a formula of depth at most L in the corresponding fragment of graded modal logic: $\exists\mathcal{GML}$ if $\preceq$ is embedding, $\exists^+\mathcal{GML}$ if $\preceq$ is injective homomorphism, and $\exists^+\mathcal{ML}$ if $\preceq$ is homomorphism.*

5 Syntactic Characterisations of Structural Preservation Classes

We now complement the semantic characterisations with corresponding syntactic ones, showing that each preservation class identified in the previous section can be realised by a suitable class of GNN architectures. In particular, we show that GNN classifiers preserved under embeddings can be realised as monotonic GNNs preceded by a pointwise feature transformation. Removing this preprocessing step yields a class of monotonic GNNs corresponding to preservation under injective homomorphisms, which generalises previously proposed monotonic architectures. Further restricting the aggregation functions yields GNN classes corresponding to preservation under homomorphisms.

These architectures consist of two types of layers: monotonic layers which involve only non-decreasing combination and aggregation functions, and augmentation layers which transform initial node feature vectors into non-negative vectors independently of the graph structure.

Definition 9. *For multisets M and M' of vectors of the same dimension, we write $M \leq M'$ if there exists an injective mapping from the elements of M to the elements of M' such that $\mathbf{x} \leq f(\mathbf{x})$, for every $\mathbf{x} \in M$. A function f mapping vectors or multisets to vectors is* non-decreasing *if $x \leq y$ implies $f(x) \leq f(y)$ for all inputs x, y of the same type (e.g. vectors or multisets).*

Using this notion, we define monotonic and augmentation layers and the relevant GNN architectures.

Definition 10. *A GNN layer $(\mathsf{agg}, \mathsf{comb})$ is* monotonic *if both the aggregation* agg *and the combination* comb *functions are non-decreasing. An* augmentation layer *is a function $\eta : \mathbb{R}^d \to \mathbb{R}^{d'}_{\geq 0}$, applied pointwise to node feature vectors, where $\mathbb{R}_{\geq 0}$ are non-negative real numbers.*

Definition 11. *A* monotonic GNN *(MGNN) is a GNN whose layers are all monotonic. A* MAX-*MGNN is an MGNN using only* MAX *as aggregation. Then an* augmented (MAX-)MGNN *is an (*MAX-*)MGNN with an additional augmentation layer before monotonic layers.*

Monotonic GNNs have been studied previously in the context of explainability and logical reasoning, where architectural constraints are imposed to enforce structural preservation properties Tena Cucala et al. (2025); Tena Cucala et al. (2023). These models are designed to satisfy monotonicity conditions that ensure preservation under homomorphisms or injective homomorphisms. Our definitions subsume these approaches by isolating the underlying semantic property and expressing it in a modular, layer-wise manner. In particular, the class of monotonic GNNs defined above generalises the architectures proposed in prior work, which we recapitulate in the following definition.

Definition 12. *A* positive-weight GNN *is a GNN consisting of simple layers (see Equation* (2)*) in which all weight matrices have non-negative entries, the activation function is non-decreasing, and the aggregation functions are* SUM*,* MAX*, or* MAX-k-SUM*.*

These syntactic constraints ensure that all aggregation and combination functions are non-decreasing with respect to their inputs. Thus, the following can be observed.

Proposition 13. *Every positive-weight GNN is a monotonic GNN.*

The syntactic classes defined above satisfy the corresponding structural preservation properties.

Theorem 14. *Augmented MGNNs are preserved under embeddings, MGNNs are preserved under injective homomorphisms, and* MAX*-MGNNs are preserved under homomorphisms.*

These preservation results can be understood intuitively as follows. Monotonic layers ensure that increasing node features (with respect to the underlying order) or neighbourhood information cannot invalidate positive predictions. Injective homomorphisms preserve the structure of the graph without merging nodes, and therefore preserve such monotonic behaviour. In contrast, non-injective homomorphisms may merge nodes and reduce multiplicities, which can invalidate properties that depend on counting. Restricting aggregation to MAX eliminates this issue, as MAX is insensitive to multiplicities: merging nodes does not affect the maximum value. Thus, MAX-MGNN classifiers are preserved even under homomorphisms. Finally, augmentation layers allow for additional feature transformations applied pointwise, without changing the graph structure, thus preserving embeddings.

To obtain exact characterisations of GNNs classes, we start by establishing lower bounds. That is, we show that every formula in a relevant fragment can be realised by a GNN in the corresponding syntactic class.

Theorem 15. *The following statements hold:*

- *For each $\exists\mathcal{GML}$-classifier there is an equivalent augmented MGNN classifier,*
- *For each $\exists\mathcal{ML}$-classifier there is an equivalent augmented* MAX*-MGNN classifier,*
- *For each $\exists^{+}\mathcal{GML}$-classifier there is an equivalent MGNN classifier,*
- *For each $\exists^{+}\mathcal{ML}$-classifier there is an equivalent* MAX*-MGNN classifier.*

Furthermore, each GNN above can be assumed to be simple, with truncated ReLU $\sigma(\mathbf{x}) = \min(\max(0, \mathbf{x}), 1)$ as activation function.

To derive exact characterisations of the expressiveness of GNN classes we need to show matching upper bounds. For this, we use our preservation theorems together with the results in Section 4.5. Consequently we obtain the following tight bounds.

Theorem 16. *The following tight expressiveness results hold:*

- *Augmented MGNN classifiers have the same expressiveness as $\exists\mathcal{GML}$ classifiers,*
- *Augmented* MAX*-MGNN classifiers have the same expressiveness as $\exists\mathcal{ML}$ classifiers,*
- *MGNN classifiers have the same expressiveness as $\exists^{+}\mathcal{GML}$ classifiers,*
- MAX*-MGNN classifiers have the same expressiveness as $\exists^{+}\mathcal{ML}$ classifiers.*

These results provide concrete insights into the kinds of properties expressible by GNNs under different structural preservation constraints. At one end of the spectrum, augmented MGNNs, which correspond to preservation under embeddings, can express a wide range of local properties, including pointwise non-monotone properties such as "node v does not have label A". However, such models cannot express non-monotone properties that depend on the node's neighbourhood, such as exact counting ("exactly k neighbours of v satisfy property p"), parity conditions (e.g., "the number of neighbours of v is even"), or global properties such as connectivity.

Restricting to MGNNs, which correspond to preservation under injective homomorphisms, restricts expressiveness to monotone properties. These include existential and threshold-based conditions, such as "node v has at least k neighbours satisfying property p", or "there exists a path of length at most k from v to a node satisfying p".

At the most restrictive level, MAX-MGNNs, which correspond to preservation under homomorphisms, can express only purely existential properties, such as "node v has some neighbour satisfying property p". In particular, they cannot express threshold-based conditions, as MAX aggregation is insensitive to multiplicities.

Overall, these results reveal a fundamental trade-off: preservation under stronger graph transformations leads to simpler logical characterisations, but at the cost of reduced expressive power. In particular, preservation properties constrain GNN expressiveness to fragments of modal logic, ruling out global and counting-based properties that fall outside these fragments.

6 Conclusion

We have developed a semantic framework for analysing the expressiveness of GNNs based on structural preservation properties. By combining preservation theorems with new well-quasi-order results for trees, we established exact correspondences between preservation under structural relations, fragments of graded modal logic, and natural classes of GNN architectures. Our approach highlights a fundamental trade-off between structural invariance and expressive power: stronger preservation properties lead to simpler logical characterisations, but at the cost of reduced expressiveness. Our framework also extends naturally to directed graphs as the underlying logical results apply to Kripke models. The extension requires only minor adaptations to the GNN architecture, such as distinguishing between incoming and outgoing neighbourhoods.

AI use statement

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools very midldy, only for polishing language. We have reviewed all AI-assisted work and take the responsibility for the final content of this work, including text, claims, and proofs; including the contents of the technical appendix.

Ethics statement

Our paper is theoretical and it does not include any experiments, in particular those that could be harmful. Furthermore, we do no identify any other ethical issues in the paper.

Reproducibility statement

Research in this paper is theoretical; all the assumptions and results are precisely stated in the paper. Proof sketches are included in the main body of the paper whereas complete proofs are included in the appendix.

References

Samson Abramsky and Luca Reggio. Arboreal categories and equi-resource homomorphism preservation theorems. *Ann. Pure Appl. Log.*, 175(6):103423, 2024. doi: 10.1016/J.APAL.2024.103423. URL https://doi.org/10.1016/j.apal.2024.103423.

Hajnal Andréka, István Németi, and Johan van Benthem. Modal languages and bounded fragments of predicate logic. *J. Philos. Log.*, 27(3):217–274, 1998. doi: 10.1023/A:1004275029985. URL https://doi.org/10.1023/A:1004275029985.

Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider (eds.). *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003. ISBN 0-521-78176-0.

Pablo Barceló, Egor V. Kostylev, Mikaël Monet, Jorge Pérez, Juan L. Reutter, and Juan Pablo Silva. The logical expressiveness of graph neural networks. In *Proc. of ICLR*, 2020.

Michael Benedikt, Chia-Hsuan Lu, Boris Motik, and Tony Tan. Decidability of Graph Neural Networks via Logical Characterizations. In *Proc. of ICALP*, volume 297, pp. 127:1–127:20. Schloss Dagstuhl, 2024.

Bernardo Cuenca Grau, Eva Feng, and Przemyslaw Walega. The correspondence between bounded graph neural networks and fragments of first-order logic. In *Proc. of AAAI*, 2026.

Maarten de Rijke. A note on graded modal logic. *Stud Logica*, 64(2):271–283, 2000. doi: 10.1023/A:1005245900406. URL https://doi.org/10.1023/A:1005245900406.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proc. of ICML*, volume 70, pp. 1263–1272, 2017.

Erich Grädel, Phokion G. Kolaitis, Leonid Libkin, Maarten Marx, Joel Spencer, Moshe Y. Vardi, Yde Venema, and Scott Weinstein. *Finite Model Theory and Its Applications*. Texts in Theoretical Computer Science. An EATCS Series. Springer, 2007.

Martin Grohe. The descriptive complexity of graph neural networks. *TheoretiCS*, 3, 2024.

Stan P Hauke and Przemysław Andrzej Wałęga. Aggregate-combine-readout GNNs are more expressive than logic C2. In *Proc. of AAAI*, 2026.

Wilfrid Hodges. *A Shorter Model Theory*. Cambridge University Press, 1997. ISBN 978-0-521-58713-6.

Leonid Libkin. *Elements of Finite Model Theory*. Texts in Theoretical Computer Science. An EATCS Series. Springer, 2004.

Aliaume Lopez. *First Order Preservation Theorems in Finite Model Theory : Locality, Topology, and Limit Constructions. (Théorèmes de préservation pour la logique au premier ordre: localité, topologie et constructions limites)*. PhD thesis, University of Paris-Saclay, France, 2023. URL https://tel.archives-ouvertes.fr/tel-04534568.

Jerzy Łoś and Roman Suszko. On the extending of models (ii). *Fundamenta Mathematicae*, 42(2): 343–347, 1955.

Roger C Lyndon. Properties preserved under homomorphism. 1959.

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proc. of AAAI*, pp. 4602–4609. AAAI Press, 2019a.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proc. of AAAI*, volume 33, pp. 4602–4609, 2019b.

Matthew Morris and Ian Horrocks. Sound logical explanations for mean aggregation graph neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=7TY89cqLfE.

Matthew Morris, David J. Tena Cucala, Bernardo Cuenca Grau, and Ian Horrocks. Relational graph convolutional networks do not learn sound rules. In Pierre Marquis, Magdalena Ortiz, and Maurice Pagnucco (eds.), *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024*, 2024. doi: 10.24963/KR.2024/84. URL https://doi.org/10.24963/kr.2024/84.

Pierre Nunn, Marco Sälzer, François Schwarzentruber, and Nicolas Troquard. A logic for reasoning about aggregate-combine graph neural networks. In *Proc. of IJCAI*, pp. 3532–3540. ijcai.org, 2024.

Martin Otto. Graded modal logic and counting bisimulation. *CoRR*, abs/1910.00039, 2019. URL http://arxiv.org/abs/1910.00039.

Eric Rosen. Modal logic over finite structures. *J. Log. Lang. Inf.*, 6(4):427–439, 1997. doi: 10.1023/A:1008275906015. URL https://doi.org/10.1023/A:1008275906015.

Eric Rosen. Some aspects of model theory and finite structures. *Bull. Symb. Log.*, 8(3):380–403, 2002. doi: 10.2178/BSL/1182353894. URL https://doi.org/10.2178/bsl/1182353894.

Benjamin Rossman. Homomorphism preservation theorems. *J. ACM*, 55(3):15:1–15:53, 2008. doi: 10.1145/1379759.1379763. URL https://doi.org/10.1145/1379759.1379763.

David Tena Cucala, Bernardo Cuenca Grau, Boris Motik, and Egor V. Kostylev. On the correspondence between monotonic max-sum gnns and datalog. In Pierre Marquis, Tran Cao Son, and Gabriele Kern-Isberner (eds.), *Proc. of KR*, pp. 658–667, 2023.

David J. Tena Cucala and Bernardo Cuenca Grau. Bridging max graph neural networks and datalog with negation. In Pierre Marquis, Magdalena Ortiz, and Maurice Pagnucco (eds.), *Proc. of KR*, 2024.

David J Tena Cucala, Bernardo Cuenca Grau, Boris Motik, and Egor V Kostylev. Expressive power of monotonic graph neural networks via datalog. In *Handbook on Neurosymbolic AI and Knowledge Graphs*, pp. 780–807. IOS Press, 2025.

Stephan Tobies. PSPACE reasoning for graded modal logics. *J. Log. Comput.*, 11(1):85–106, 2001. doi: 10.1093/LOGCOM/11.1.85. URL https://doi.org/10.1093/logcom/11.1.85.

Xiyuan Wang and Muhan Zhang. How powerful are spectral graph neural networks. In *Proc. of ICML*, 2022.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *Proc. of ICLR*, 2019.

TECHNICAL APPENDIX

A PROOFS FOR SECTION 4.4

Theorem 4. *The embedding relation on any class of rooted trees, whose heights are uniformly bounded by a fixed constant, is a well-quasi-order.*

Proof. Let L be the uniform bound on the height of the trees under consideration, and let $\preceq$ be the embedding relation between pointed trees. Clearly, $\preceq$ is a quasi-order. We prove by induction on $\ell \leq L$ the following claim: for every class $\mathcal{T}$ of pointed trees of height at most ℓ, the quasi-order $(\mathcal{T}, \preceq)$ is a wqo.

Base case ($\ell = 0$). A tree of height 0 consists of a single root labelled by some vector in $\{0,1\}^d$. Since the dimension d of trees is finite, there are finitely many non-isomorphic trees of this form. Hence, for any class of pointed trees of height 0, the quasi-order induced by $\preceq$ is a wqo. Note that the finiteness of d is essential for this argument.

Inductive step. Let $\ell \geq 1$ and assume that the claim holds for $\ell - 1$. Let $\mathcal{T}$ be an arbitrary class of pointed trees of height at most ℓ. We show that $(\mathcal{T}, \preceq)$ is a wqo.

We begin by partitioning $\mathcal{T}$ into subclasses $\mathcal{T}^S$ indexed by $S \subseteq \{1, \ldots, d\}$, where $\mathcal{T}^S$ consists of all pointed trees in $\mathcal{T}$ whose root has exactly the labels in S. Since 2^d is finite and a finite disjoint union of wqos is also a wqo, it suffices to show that each $(\mathcal{T}^S, \preceq)$ is a wqo.

Fix $S \subseteq \{1, \ldots, d\}$. For each $(\mathfrak{M}, w)$ in $\mathcal{T}^S$, let $\tau_{\mathfrak{M},w} = (\tau_1, \ldots, \tau_n)$ be a finite sequence enumerating its immediate subtrees rooted at the children of w. As the children are unordered, we fix an arbitrary but canonical ordering on the subtrees to obtain a well-defined enumeration. The embedding relation $\preceq$ on pointed graphs induces the subsequence order relation $\preceq^*$ on their finite sequences, defined as in Higman's Lemma (see Lemma 3). To account for the arbitrary ordering of children, we consider the permutation-closed variant $\preceq^*_\pi$ defined by $\tau \preceq^*_\pi \tau'$ if and only if $\tau \preceq^* f(\tau')$ for some permutation f. We claim that for any $(\mathfrak{M}, w), (\mathfrak{N}, v) \in \mathcal{T}^S$, we have $(\mathfrak{M}, w) \preceq (\mathfrak{N}, v)$ if and only if $\tau_{\mathfrak{M},w} \preceq^*_\pi \tau_{\mathfrak{N},v}$. Indeed, an embedding from $(\mathfrak{M}, w)$ into $(\mathfrak{N}, v)$ must map each child of w injectively to a distinct child of v, such that the corresponding subtree embeds recursively. This yields $\tau_{\mathfrak{M},w} \preceq^*_\pi \tau_{\mathfrak{N},v}$. Conversely, if $\tau_{\mathfrak{M},w} \preceq^*_\pi \tau_{\mathfrak{N},v}$ then each subtree in $\tau_{\mathfrak{M},w}$ embeds into a distinct subtree in $\tau_{\mathfrak{N},v}$, yielding an embedding from $(\mathfrak{M}, w)$ into $(\mathfrak{N}, v)$.

Thus, to show that $(\mathcal{T}^S, \preceq)$ is a wqo, it suffices to show that $\big(\{\tau_{\mathfrak{M},w} \mid (\mathfrak{M}, w) \in \mathcal{T}^S\}, \preceq^*_\pi\big)$ is a wqo. To this end, let $\mathcal{I}$ be the set of all rooted trees of height at most $\ell - 1$. By the inductive hypothesis, $(\mathcal{I}, \preceq)$ is a wqo. By Higman's Lemma (Lemma 3), the set of finite sequences $\mathcal{I}^*$ ordered by the subsequence relation $\preceq^*$ is also a wqo. Since wqos are closed under taking subsets, and $(\tau_{\mathfrak{M},w} \mid (\mathfrak{M}, w) \in \mathcal{T}^S)$ is a subset of $\mathcal{I}^*$, it follows that $((\tau_{\mathfrak{M},w} \mid (\mathfrak{M}, w) \in \mathcal{T}^S), \preceq^*)$ is a wqo. Moreover, as $\preceq^*$ is a subrelation of $\preceq^*_\pi$ and $\preceq^*_\pi$ is a quasi-order, it follows that $\preceq^*_\pi$ is also a wqo on this set. This completes the inductive step and the whole proof. □

Theorem 6. *The embedding, injective homomorphism, and homomorphism relations on classes of rooted trees of unbounded height are not, in general, well-quasi-orders.*

Proof. We start with the case of injective homomorphisms, which also covers embeddings as a special case. Consider the class $\mathcal{T}$ of pointed trees that, for every $n \geq 1$, contains a rooted tree (G_n, v_1) with nodes $v_1, \ldots v_n, v_{n+1}, v_{n+2}, u, w$ which has edges $\{v_i < v_{i+1}\}$ for $1 \leq i \leq n+1$ as well as the edges $\{v_1, u\}$ and $\{v_{n+1}, w\}$, as depicted in Figure 3. All nodes are unlabelled (equivalently, they all have the same label).

In each (G_n, v_1) there are exactly two nodes at distance $n+1$ from the root, namely v_{n+2} and w. By contrast, if $m > n$, then (G_m, v_1) has a single node at distance $n+1$, namely v_{n+2}. Hence, for any $n \neq m$, there is neither an injective homomorphism from (G_n, v_1) to (G_m, v_1), nor vice-versa. Thus, $\mathcal{T}$ forms an infinite antichain with respect to injective homomorphisms and hence also with respect to embeddings.

We now turn to the case of homomorphisms. Consider the modification of $\mathcal{T}$ where, in each graph (G_n, v_1), nodes are labelled with vectors of length 4, where the labelling λ is defined such that

$\lambda(v)_1 = 1$ iff $v = v_i$ for some even i, $\lambda(v)_2 = 1$ iff $v = v_i$ for some odd i, $\lambda(v)_3 = 1$ iff $v = u$, and $\lambda(v)_4 = 1$ iff $v = w$.

Let $n < m$ and suppose, towards a contradiction, that f is a homomorphism from (G_n, v_1) to (G_m, v_1). We prove by induction on $i \leq n + 2$ that $f(v_i) = v_i$. The base case $i = 1$ is immediate. For the inductive step, assume that $f(v_i) = v_i$. Then, $f(v_{i+1})$ must be a successor of v_i in G_m satisfying the same labels as v_{i+1} in G_n. In G_m the only such successor is v_{i+1} itself, hence $f(v_{i+1}) = v_{i+1}$. In particular, $f(v_{n+1}) = v_{n+1}$. It follows that $f(w)$ is a successor of v_{n+1} in G_m with label 4. However, the only node in G_m with label 4 is w, which is a successor of v_{m+1} rather than v_{n+1}. This yields a contradiction. A symmetric argument shows that no homomorphism exists from (G_m, v_1) to (G_n, v_1). □

B Proofs for Section 4.5

In what follows we will prove the following theorem.

Theorem 7. *Let $\mathcal{C}$ be a class of pointed graphs, $L \in \mathbb{N}$, and let $\preceq$ be one of the following relations: embedding, injective homomorphism, or homomorphism. Then the following are equivalent:*

- *$\mathcal{C}$ is invariant under L-unravelling and preserved under $\preceq$*
- *$\mathcal{C}$ is definable by a formula of depth at most L in a corresponding logic: $\exists\mathcal{GML}$ if $\preceq$ is embedding, $\exists^+\mathcal{GML}$ if $\preceq$ is injective homomorphism, and $\exists^+\mathcal{ML}$ if $\preceq$ is homomorphism.*

The above theorem consists of three statements; we will prove them one by one in three subsections.

Since proofs exploit modal logic machinery, in the following subsections we will use notions and syntax used in this research area. In particular, instead of pointed graphs we will write about pointed (Kripke) models, and represent them as $\mathfrak{M} = (W, R, V)$, where W is a non-empty finite set of *worlds* (corresponding to nodes), $R \subseteq W \times W$ is the *accessibility relation* (corresponding to edges), and V is a *valuation function* mapping each propositional variable $p_i \in \mathsf{PROP}$ to a subset of W (corresponding to the node labelling, where PROP has as many labels as the dimension d of vectors).

B.1 Preservation Under Embeddings

Our aim is to show that any class of models preserved under embeddings and invariant under L-unravelling is definable by a $\exists\mathcal{GML}$-formula. Moreover, we show that such a formula can be chosen to have depth at most L. To this end, we introduce characteristic formulae in $\exists\mathcal{GML}$. These formulae are obtained as a modification of the standard characteristic formulae for graded modal logic from the literature Otto (2019); Cuenca Grau et al. (2026). The key difference is that our characteristic formulae in $\exists\mathcal{GML}$ do not contain negated modal operators.

Definition 17. *Let $(\mathfrak{M}, w)$ be a pointed model with $\mathfrak{M} = (W, R, V)$. For each $\ell \in \mathbb{N}$, we define a* characteristic $\exists\mathcal{GML}$ formula*, $\varphi^{\ell}_{\mathfrak{M},w}$ of depth ℓ inductively as follows. For $\ell = 0$, let*

$$\varphi^{0}_{\mathfrak{M},w} := \bigwedge\{p \mid (\mathfrak{M}, w) \models p\} \wedge \bigwedge\{\neg p \mid (\mathfrak{M}, w) \not\models p\}.$$

For $\ell + 1$, partition the R-successors of w into equivalence classes $C_1, \ldots, C_m$ according to the relation $\equiv^{\mathcal{GML}}_{\ell}$. For each C_j, choose an arbitrary representative $w_j \in C_j$ and define

$$\varphi^{\ell+1}_{\mathfrak{M},w} \;:=\; \varphi^{0}_{\mathfrak{M},w} \;\wedge\; \bigwedge_{j=1}^{m} \Diamond^{\geq |C_j|}\, \varphi^{\ell}_{\mathfrak{M},w_j}.$$

For full $\mathcal{GML}$, a characteristic formula of depth ℓ for a pointed model $(\mathfrak{M}, w)$ holds in $(\mathfrak{N}, v)$ if and only if there exists a graded ℓ-bisimulation between these two pointed models. Over finite models, this is equivalent to $(\mathfrak{M}, w) \equiv^{\mathcal{GML}}_{\ell} (\mathfrak{N}, v)$ Otto (2019). The characteristic defined above for $\exists\mathcal{GML}$ are strictly weaker, and these equivalences no longer hold. Instead, the $\exists\mathcal{GML}$ characteristic ensure a one-way structural property: the ℓ-unravelling of the model used to construct the characteristic formula embeds into the ℓ-unravelling of any model satisfying the formula. This property is captured by the following lemma, which plays a central role in the preservation theorem for embeddings.

Lemma 18. *Let $(\mathfrak{M}, w)$ and $(\mathfrak{N}, v)$ be pointed models and let $\ell \in \mathbb{N}$. We have $(\mathfrak{N}, v) \models \varphi^{\ell}_{\mathfrak{M},w}$ if and only if $\mathsf{Unr}^{\ell}(\mathfrak{M}, w)$ embeds into $\mathsf{Unr}^{\ell}(\mathfrak{N}, v)$.*

Proof. We prove the equivalence by induction on ℓ. Let $\preceq$ be the embedding relation between pointed models.

Base case ($\ell = 0$). Formula $\varphi^0_{\mathfrak{M},w}$ is a conjunction of all literals (propositions and their negations) satisfied at $(\mathfrak{M}, w)$. Thus, $(\mathfrak{N}, v) \models \varphi^0_{\mathfrak{M},w}$ iff $(\mathfrak{M}, w) \equiv_0^{\mathcal{GML}} (\mathfrak{N}, v)$. Since $\mathsf{Unr}^0(\mathfrak{M}, w)$ and $\mathsf{Unr}^0(\mathfrak{N}, v)$ consist single worlds, it holds iff $\mathsf{Unr}^0(\mathfrak{M}, w) \preceq \mathsf{Unr}^0(\mathfrak{N}, v)$.

Inductive step. Assume that the claim holds for $\ell \geq 0$. We prove that it also holds for $\ell + 1$. Recall that $\varphi^{\ell+1}_{\mathfrak{M},w} := \varphi^0_{\mathfrak{M},w} \wedge \bigwedge_{j=1}^{m} \Diamond^{\geq |C_j|} \varphi^{\ell}_{\mathfrak{M},w_j}$, where $C_1, \ldots, C_m$ are the equivalence classes of R-successors of w under $\equiv_\ell^{\mathcal{GML}}$, and $w_j \in C_j$ is an arbitrary representative. For each j, we also fix an enumeration $C_j = \{w_{j,1}, \ldots, w_{j,|C_j|}\}$.

($\Rightarrow$) Assume that $(\mathfrak{N}, v) \models \varphi^{\ell+1}_{\mathfrak{M},w}$. Then, $(\mathfrak{N}, v) \models \varphi^0_{\mathfrak{M},w}$, and hence $(\mathfrak{M}, w) \equiv_0^{\mathcal{GML}} (\mathfrak{N}, v)$. Moreover, for each j, there exist $|C_j|$ distinct successors $v_{j,1}, \ldots, v_{j,|C_j|}$ of v such that $(\mathfrak{N}, v_{j,k}) \models \varphi^{\ell}_{\mathfrak{M},w_j}$.

By the inductive hypothesis, for each such j and k, there exists an embedding $f_{j,k}$ from $\mathsf{Unr}^\ell(\mathfrak{M}, w_j)$ into $\mathsf{Unr}^\ell(\mathfrak{N}, v_{j,k})$. Since $w_{j,k} \equiv_\ell^{\mathcal{GML}} w_j$, the unravellings $\mathsf{Unr}^\ell(\mathfrak{M}, w_{j,k})$ and $\mathsf{Unr}^\ell(\mathfrak{M}, w_j)$ are isomorphic. Composing this isomorphism with $f_{j,k}$ yields an embedding of $\mathsf{Unr}^\ell(\mathfrak{M}, w_{j,k})$ into $\mathsf{Unr}^\ell(\mathfrak{N}, v_{j,k})$.

We now construct a global embedding f from $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ into $\mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$ as follows. Function f maps the root of $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ to the root of $\mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$. For each j and k, f coincides with the embedding constructed above on the subtree rooted at the node corresponding to $w_{j,k}$. Since the successors $v_{j,k}$ are pairwise distinct, the images of these embeddings are disjoint. It follows that f is an embedding of $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ into $\mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$.

($\Leftarrow$) Conversely, assume that f is an embedding of $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ into $\mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$. Then, f maps the root of $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ to the root of $\mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$, and therefore, $(\mathfrak{M}, w) \equiv_0^{\mathcal{GML}} (\mathfrak{N}, v)$. Thus, $(\mathfrak{N}, v) \models \varphi^0_{\mathfrak{M},w}$. Fix $j \in \{1, \ldots, m\}$. By the definition of unravelling, the root of $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ has exactly $|C_j|$ distinct nodes at depth 1 corresponding to the successors $w_{j,1}, \ldots, w_{j,|C_j|}$ of w. Since f is injective, it maps these nodes to $|C_j|$ distinct nodes at depth 1 in $\mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$. Each such node corresponds to a successor $v_{j,k}$ of v in $\mathfrak{N}$, and the subtree below this node is precisely $\mathsf{Unr}^\ell(\mathfrak{N}, v_{j,k})$. Restricting f to the corresponding subtrees yields embeddings, so $\mathsf{Unr}^\ell(\mathfrak{M}, w_{j,k}) \preceq \mathsf{Unr}^\ell(\mathfrak{N}, v_{j,k})$ for $k = 1, \ldots, |C_j|$. By the inductive hypothesis, this implies $(\mathfrak{N}, v_{j,k}) \models \varphi^{\ell}_{\mathfrak{M},w_{j,k}}$. Since each $w_{j,k} \in C_j$, we obtain that $\varphi^{\ell}_{\mathfrak{M},w_{j,k}}$ is identical to $\varphi^{\ell}_{\mathfrak{M},w_j}$. Moreover, the worlds $v_{j,k}$ are pairwise distinct successors of v. Hence, $(\mathfrak{N}, v) \models \Diamond^{\geq |C_j|} \varphi^{\ell}_{\mathfrak{M},w_j}$, as required. □

Equipped with Lemma 18 and Theorem 4, we are now ready to prove the preservation theorem for classes of models preserved under embeddings and invariant under L-unravelling. As we show below, these are exactly the classes definable by $\exists\mathcal{GML}$- of modal depth at most L.

Theorem 19. *The following are equivalent for any class $\mathcal{C}$ of pointed models and any $L \in \mathbb{N}$:*

- *$\mathcal{C}$ is invariant under L-unravelling and preserved under embeddings,*
- *$\mathcal{C}$ is definable by an $\exists\mathcal{GML}$ formula of depth at most L.*

Proof. ($\Rightarrow$) Let $\mathcal{T} = \{\mathsf{Unr}^L(\mathfrak{M}, w) \mid (\mathfrak{M}, w) \in \mathcal{C}\}$ and let $\preceq$ be the embedding relation. Since $\mathcal{C}$ is invariant under L-unravelling, we have that $\mathcal{T} \subseteq \mathcal{C}$. By Theorem 4, $(\mathcal{T}, \preceq)$ is a wqo and therefore admits only finitely many $\preceq$-minimal elements up to isomorphism. Let $\mathcal{T}_{\mathsf{min}}$contain exactly one representative for each such minimal isomorphism type. For each $(\mathfrak{T}, v) \in \mathcal{T}_{\mathsf{min}}$, we construct the $\exists\mathcal{GML}$ characteristic formula $\varphi^L_{\mathfrak{T},v}$ and define $\varphi_{\mathcal{C}} = \bigvee_{(\mathfrak{T},v)\in\mathcal{T}_{\mathsf{min}}} \varphi^L_{\mathfrak{T},v}$. Since the disjunction is finite and each $\varphi^L_{\mathfrak{T},v}$ is an $\exists\mathcal{GML}$-formula, $\varphi_{\mathcal{C}}$ is a (finite) $\exists\mathcal{GML}$-formula. We show that $\varphi_{\mathcal{C}}$ defines $\mathcal{C}$.

Let $(\mathfrak{M}, w) \in \mathcal{C}$. Then $\mathsf{Unr}^L(\mathfrak{M}, w) \in \mathcal{T}$. Since $(\mathcal{T}, \preceq)$ is a wqo, there exists a $\preceq$-minimal element $(\mathfrak{T}, v) \in \mathcal{T}_{\mathsf{min}}$ such that $(\mathfrak{T}, v) \preceq \mathsf{Unr}^L(\mathfrak{M}, w)$. Because $(\mathfrak{T}, v)$ is tree-shaped of height at most L, we have $\mathsf{Unr}^L(\mathfrak{T}, v) = (\mathfrak{T}, v)$. Hence, $\mathsf{Unr}^L(\mathfrak{T}, v) \preceq \mathsf{Unr}^L(\mathfrak{M}, w)$ and by Lemma 18 we obtain $(\mathfrak{M}, w) \models \varphi^L_{\mathfrak{T},v}$. It follows that $(\mathfrak{M}, w) \models \varphi_{\mathcal{C}}$.

Conversely, suppose that $(\mathfrak{M}, w) \models \varphi_{\mathcal{C}}$. Then, $(\mathfrak{M}, w) \models \varphi^L_{\mathfrak{T},v}$ for some $(\mathfrak{T}, v) \in \mathcal{T}_{\mathsf{min}}$. By Lemma 18, this implies $\mathsf{Unr}^L(\mathfrak{T}, v) \preceq \mathsf{Unr}^L(\mathfrak{M}, w)$. Since $(\mathfrak{T}, v)$ is tree-shaped of height at most L, we again have $\mathsf{Unr}^L(\mathfrak{T}, v) = (\mathfrak{T}, v)$, and hence $(\mathfrak{T}, v) \preceq \mathsf{Unr}^L(\mathfrak{M}, w)$. By construction, $\mathcal{T}_{\mathsf{min}} \subseteq \mathcal{T} \subseteq \mathcal{C}$, so $(\mathfrak{T}, v) \in \mathcal{C}$. Since $\mathcal{C}$ is preserved under embeddings, we have $\mathsf{Unr}^L(\mathfrak{M}, w) \in \mathcal{C}$. Finally, invariance of $\mathcal{C}$ under L-unravelling yields $(\mathfrak{M}, w) \in \mathcal{C}$.

($\Leftarrow$) Suppose that $\mathcal{C}$ is definable by an $\exists\mathcal{GML}$-formula φ of depth at most L. Then $\mathcal{C}$ is invariant under L-unravellings: for every pointed model, $(\mathfrak{M}, w) \models \varphi$ if and only if $\mathsf{Unr}^L(\mathfrak{M}, w) \models \varphi$ since the truth of φ depends only on the unravelling up to depth L. We argue that $\mathcal{C}$ is also preserved under embeddings. Let $(\mathfrak{M}, w) \models \varphi$ and suppose that there exists an embedding witnessing $(\mathfrak{M}, w) \preceq (\mathfrak{N}, v)$. Since embeddings preserve propositional labels and map successors injectively, a routine induction on the structure of φ shows that $(\mathfrak{N}, v) \models \varphi$. In particular, for each subformula of the form $\Diamond^{\geq k}\psi$ occurring in φ, the existence of k successors satisfying ψ in $(\mathfrak{M}, w)$ implies the existence of k successors satisfying ψ in $(\mathfrak{N}, v)$. Thus, $\mathcal{C}$ is preserved under embeddings. □

B.2 Preservation Under Injective Homomorphisms

We next characterise unravelling-invariant classes that are preserved under injective homomorphisms. Recall that every embedding is an injective homomorphism, but not conversely. This suggests that the logical language defining classes preserved under injective homomorphisms should be strictly weaker than the one required for preservation under embeddings. By Theorem 19, the latter language is $\exists\mathcal{GML}$. As we show next, the appropriate language in the case of injective homomorphisms is $\exists^+\mathcal{GML}$, the existential-positive fragment of $\mathcal{GML}$.

We first define characteristic $\exists^+\mathcal{GML}$-, obtained from the characteristic $\exists\mathcal{GML}$ by deleting all negative literals.

Definition 20. *Let $(\mathfrak{M}, w)$ be a pointed model with $\mathfrak{M} = (W, R, V)$. The* characteristic $\exists^+\mathcal{GML}$-formula $\psi^\ell_{\mathfrak{M},w}$ *of depth ℓ is defined as the formula obtained from the characteristic $\exists\mathcal{GML}$-formula $\varphi^\ell_{\mathfrak{M},w}$ (see Definition 17) by removing all negative literals $\neg p$.*

Note that $\psi^\ell_{\mathfrak{M},w}$ is defined in the same way as $\varphi^\ell_{\mathfrak{M},w}$, except that the conjunction $\bigwedge\{\neg p \mid (\mathfrak{M}, w) \not\models p\}$ is omitted. As a result, $\psi^\ell_{\mathfrak{M},w}$ contains no negations and is therefore an $\exists^+\mathcal{GML}$-formula.

While $\varphi^\ell_{\mathfrak{M},w}$ guarantees the existence of an embedding between ℓ-unravellings of models, the formula $\psi^\ell_{\mathfrak{M},w}$ guarantees the existence of an injective homomorphism between these unravellings. This is captured by the following lemma.

Lemma 21. *Let $(\mathfrak{M}, w)$ and $(\mathfrak{N}, v)$ be pointed models and let $\ell \in \mathbb{N}$. Then, $(\mathfrak{N}, v) \models \psi^\ell_{\mathfrak{M},w}$ if and only if there exists an injective homomorphism from $\mathsf{Unr}^\ell(\mathfrak{M}, w)$ to $\mathsf{Unr}^\ell(\mathfrak{N}, v)$.*

Proof. The proof is by induction on ℓ and closely parallels the proof of Lemma 18; we highlight only the points where injective homomorphisms replace embeddings. Throughout this proof, let $\preceq$ be the injective homomorphism relation.

Base case ($\ell = 0$). The formula $\psi^0_{\mathfrak{M},w}$ is the conjunction of all propositional symbols satisfied at w. Thus, $(\mathfrak{N}, v) \models \psi^0_{\mathfrak{M},w}$ if and only if $(\mathfrak{N}, v)$ satisfies at least the same propositions as $(\mathfrak{M}, w)$. Since $\mathsf{Unr}^0(\mathfrak{M}, w)$ and $\mathsf{Unr}^0(\mathfrak{N}, v)$ each consist of a single world, this holds if and only if $\mathsf{Unr}^0(\mathfrak{M}, w) \preceq \mathsf{Unr}^0(\mathfrak{N}, v)$.

Inductive step. Assume that the claim holds for some $\ell \geq 0$, and consider $\ell + 1$. We use the same notation as in the proof of Lemma 18.

(⇒) Assume that $(\mathfrak{N}, v) \models \psi^{\ell+1}_{\mathfrak{M},w}$. Then $(\mathfrak{N}, v) \models \psi^{0}_{\mathfrak{M},w}$, and for each equivalence class C_j, there exist $|C_j|$ distinct successors $v_{j,1}, \ldots, v_{j,|C_j|}$ of v such that $(\mathfrak{N}, v_{j,k}) \models \psi^{\ell}_{\mathfrak{M},w_j}$. By the inductive hypothesis, for each j, k, there exists an injective homomorphism and hence $\mathsf{Unr}^{\ell}(\mathfrak{M}, w_j) \preceq \mathsf{Unr}^{\ell}(\mathfrak{N}, v_{j,k})$. As in the proof of Lemma 18, these homomorphisms can be combined together into a single injective homomorphism from $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ to $\mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$ by mapping the roots and using the homomorphisms above on the corresponding subtrees. Since the targets lie in pairwise disjoint subtrees rooted at distinct depth-1 nodes, the resulting homomorphism is injective. Thus, $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w) \preceq \mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$.

(⇐) Conversely, suppose that there exists an injective homomorphism f from $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ to $\mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$. Then f maps the root to the root, which implies $(\mathfrak{N}, v) \models \psi^{0}_{\mathfrak{M},w}$. Fix $j \in \{1, \ldots, m\}$. Injectivity of f forces the $|C_j|$ distinct depth-1 nodes below the root of $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ to be mapped to $|\ C_j\ |$ distinct nodes of depth 1 of $\mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$. Each such node corresponds to a successor $v_{j,k}$ of v in $\mathfrak{N}$. Restricting f to the corresponding subtrees yields injective homomorphisms from $\mathsf{Unr}^{\ell}(\mathfrak{M}, w_{j,k})$ to $\mathsf{Unr}^{\ell}(\mathfrak{N}, v_{j,k})$. By the induction hypothesis we obtain $(\mathfrak{N}, v_{j,k}) \models \psi^{\ell}_{\mathfrak{M},w_j}$ for all k. Since the worlds $v_{j,k}$ are pairwise distinct successors of v, we conclude that $(\mathfrak{N}, v) \models \Diamond^{\geq |C_j|} \psi^{\ell}_{\mathfrak{M},w_j}$. As this holds for each j, $(\mathfrak{N}, v) \models \psi^{\ell+1}_{\mathfrak{M},w}$. □

Using the results above on characteristic $\exists\mathcal{GML}$-, we can now establish the following preservation theorem for unravelling-invariant classes preserved under injective homomorphisms.

Theorem 22. *The following are equivalent for any class $\mathcal{C}$ of pointed models and any $L \in \mathbb{N}$:*

- *$\mathcal{C}$ is preserved under injective homomorphisms and invariant under L-unravelling,*
- *$\mathcal{C}$ is definable by an $\exists^{+}\mathcal{GML}$-formula of depth at most L.*

Proof. The proof follows the same strategy as that of Theorem 19, but now exploits Lemma 21 instead of Lemma 18.

(⇒) Let $\mathcal{T} = \{\, \mathsf{Unr}^{L}(\mathfrak{M}, w) \mid (\mathfrak{M}, w) \in \mathcal{C} \,\}$ and let $\preceq$ be the injective homomorphism relation. Since $\mathcal{C}$ is invariant under L-unravellings, we have $\mathcal{T} \subseteq \mathcal{C}$. By Corollary 5, $(\mathcal{T}, \preceq)$ is a wqo and hence admits only finitely many $\preceq$-minimal elements up to isomorphism. Let $\mathcal{T}_{\mathsf{min}}$contain one representative for each such minimal isomorphism type. For each $(\mathfrak{T}, v) \in \mathcal{T}_{\mathsf{min}}$, we construct the $\exists^{+}\mathcal{GML}$ characteristic formula $\psi^{L}_{\mathfrak{T},v}$ and define $\psi_{\mathcal{C}} = \bigvee_{(\mathfrak{T},v)\in\mathcal{T}_{\mathsf{min}}} \psi^{L}_{\mathfrak{T},v}$. The argument showing that $\psi_{\mathcal{C}}$ defines $\mathcal{C}$ is identical to the corresponding one in the proof of Theorem 19, with Lemma 21 used in place of Lemma 18.

(⇐) Suppose that $\mathcal{C}$ is definable by an $\exists^{+}\mathcal{GML}$-formula φ of depth at most L. Then $\mathcal{C}$ is invariant under L-unravellings: for every pointed model $(\mathfrak{M}, w)$, $(\mathfrak{M}, w) \models \varphi$ if and only if $\mathsf{Unr}^{L}(\mathfrak{M}, w) \models \varphi$. We now show that $\mathcal{C}$ is preserved under injective homomorphisms. Let $(\mathfrak{M}, w) \models \varphi$ and let f be an injective homomorphism witnessing $(\mathfrak{M}, w) \preceq (\mathfrak{N}, v)$. Since injective homomorphisms preserve propositional labels monotonically and map successors injectively, a routine induction on the structure of φ shows that $(\mathfrak{N}, v) \models \varphi$, exactly as in the proof of Theorem 19. □

It is worth noting that logics corresponding to both $\mathcal{GML}$ and $\exists^{+}\mathcal{GML}$ have been studied extensively in the setting of description logics in knowledge representation and reasoning. In particular, the description logic corresponding to $\mathcal{GML}$ is $\mathcal{ALCQ}$ Baader et al. (2003), while the logic corresponding to $\exists^{+}\mathcal{GML}$ is $\mathcal{ELUQ}$ Morris & Horrocks (2025).

B.3 Preservation Under Homomorphisms

We now turn to classes preserved under homomorphisms. We show that if such a class is invariant under unravelling, then it is definable by existential-positive modal logic ($\exists^{+}\mathcal{ML}$). Our proof strategy follows the same general pattern as in the cases of embeddings and injective homomorphisms. In particular, we define an appropriate form of characteristic $\exists^{+}\mathcal{ML}$-.

Definition 23. *Let $(\mathfrak{M}, w)$ be a pointed model with $\mathfrak{M} = (W, R, V)$, the* characteristic $\exists^+\mathcal{ML}$-formula $\chi^{\ell}_{\mathfrak{M},w}$ *of depth ℓ is defined as the formula obtained from the characteristic $\exists\mathcal{GML}$-formula $\varphi^{\ell}_{\mathfrak{M},w}$ (see Definition 17) by removing all negative literals $\neg p$ and replacing each graded modal operator $\Diamond^{\geq k}$ with the basic modal operator $\Diamond$.*

Characteristic $\exists^+\mathcal{ML}$- are related to homomorphisms between unravellings by the following lemma.

Lemma 24. *Let $(\mathfrak{M}, w)$ and $(\mathfrak{N}, v)$ be pointed models and let $\ell \in \mathbb{N}$. Then, $(\mathfrak{N}, v) \models \chi^{\ell}_{\mathfrak{M},w}$ if and only if $\mathsf{Unr}^{\ell}(\mathfrak{M}, w) \preceq \mathsf{Unr}^{\ell}(\mathfrak{N}, v)$, where $\preceq$ denotes the homomorphism relation.*

Proof. The proof is by induction on ℓ.

Base case ($\ell = 0$). The formula $\chi^{0}_{\mathfrak{M},w}$ is a conjunction of propositional symbols satisfied at w. Thus, $(\mathfrak{N}, v) \models \chi^{0}_{\mathfrak{M},w}$ if and only if $(\mathfrak{N}, v)$ satisfies at least the same propositional symbols as $(\mathfrak{M}, w)$. Since $\mathsf{Unr}^0(\mathfrak{M}, w)$ and $\mathsf{Unr}^0(\mathfrak{N}, v)$ each consist of a single world, this holds if and only if $\mathsf{Unr}^0(\mathfrak{M}, w) \preceq \mathsf{Unr}^0(\mathfrak{N}, v)$.

Inductive step. Assume that the claim holds for some $\ell \geq 0$, and consider $\ell + 1$. We use the same notation as in the proof of Lemma 18.

($\Rightarrow$) Let $(\mathfrak{N}, v) \models \chi^{\ell+1}_{\mathfrak{M},w}$. Then, for each equivalence class C_j of successors of w in $\mathfrak{M}$, there exists at least one successor v_j of v in $\mathfrak{N}$ such that $(\mathfrak{N}, v_j) \models \chi^{\ell}_{\mathfrak{M},w_j}$. By the inductive hypothesis, for each j, $\mathsf{Unr}^{\ell}(\mathfrak{M}, w_j) \preceq \mathsf{Unr}^{\ell}(\mathfrak{N}, v_j)$. The unravelling $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ consists of the root together with subtrees rooted at depth-1 nodes corresponding to the successors of w. Since homomorphisms need not be injective, all depth-1 nodes corresponding to successors in the same class C_j may be mapped to the same depth-1 node corresponding to v_j. Mapping the root to the root and using the homomorphisms above on the corresponding subtrees yields a homomorphism from $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ to $\mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$.

($\Leftarrow$) Let f be a homomorphism from $\mathsf{Unr}^{\ell+1}(\mathfrak{M}, w)$ to $\mathsf{Unr}^{\ell+1}(\mathfrak{N}, v)$. Then, f maps the root to the root, implying $(\mathfrak{N}, v) \models \chi^{0}_{\mathfrak{M},w}$. Fix an equivalence class C_j. Every depth-1 node corresponding to a successor w_j of w must be mapped by f to some depth-1 node corresponding to a successor v_j of v. Restricting f to the corresponding subtree yields a homomorphism $\mathsf{Unr}^{\ell}(\mathfrak{M}, w_j) \to \mathsf{Unr}^{\ell}(\mathfrak{N}, v_j)$ and by the inductive hypothesis we obtain $(\mathfrak{N}, v_j) \models \chi^{\ell}_{\mathfrak{M},w_j}$. Hence, $(\mathfrak{N}, v) \models \Diamond\chi^{\ell}_{\mathfrak{M},w_j}$ for each j and therefore $(\mathfrak{N}, v) \models \chi^{\ell+1}_{\mathfrak{M},w}$. □

The preceding result allows us to establish the homomorphism preservation theorem for unravelling-invariant classes.

Theorem 25. *The following are equivalent for any class $\mathcal{C}$ of pointed models and any $L \in \mathbb{N}$:*

- *$\mathcal{C}$ is preserved under homomorphisms and invariant under L-unravelling,*
- *$\mathcal{C}$ is definable by an $\exists^+\mathcal{ML}$-formula of depth at most L.*

Proof. The proof follows the same general strategy as that of Theorem 19, but now uses Lemma 24 in place of Lemma 18.

($\Rightarrow$) Let $\mathcal{T} = \{\, \mathsf{Unr}^L(\mathfrak{M}, w) \mid (\mathfrak{M}, w) \in \mathcal{C} \,\}$. Since $\mathcal{C}$ is invariant under L-unravelling, we have $\mathcal{T} \subseteq \mathcal{C}$. Let $\mathcal{T}_{\mathsf{min}} \subseteq \mathcal{T}$ be a set of representatives of the homomorphism-minimal elements of $\mathcal{T}$. By Corollary 5, $\mathcal{T}_{\mathsf{min}}$ is finite (alternatively, finiteness also follows from the fact that there are only finitely many $\mathcal{ML}$-inequivalent of depth at most L). For each $(\mathfrak{T}, v) \in \mathcal{T}_{\mathsf{min}}$, let $\chi^{L}_{\mathfrak{T},v}$ be its characteristic $\exists^+\mathcal{ML}$-formula and define $\chi_{\mathcal{C}} = \bigvee_{(\mathfrak{T},v)\in\mathcal{T}_{\mathsf{min}}} \chi^{L}_{\mathfrak{T},v}$, which is a finite $\exists^+\mathcal{ML}$-formula of depth at most L.

By construction, $\chi_{\mathcal{C}}$ holds in every model in $\mathcal{T}_{\mathsf{min}}$ and by Lemma 24 it therefore holds in all models in $\mathcal{T}$. Since $\mathcal{C}$ is invariant under L-unravelling, it follows that $\chi_{\mathcal{C}}$ defines $\mathcal{C}$.

(⇐) Suppose that $\mathcal{C}$ is definable by an $\exists^+\mathcal{ML}$-formula φ of depth at most L. Since φ is positive, the standard translation yields an existential-positive first-order logic formula. Such are preserved under homomorphisms Rossman (2008). Moreover, since φ has depth at most L, it is invariant under L-unravelling. □

C Proofs for Section 5

Theorem 14. *Augmented MGNNs are preserved under embeddings, MGNNs are preserved under injective homomorphisms, and* MAX*-MGNNs are preserved under homomorphisms.*

Proof. We will show each of the three claims separately. We start from MGNNs, then consider MAX-MGNNs, and finally show the result for augmented MGNNs.

Claim 1. MGNNs are preserved under injective homomorphisms.

Let $h\colon (G, v) \to (G', v')$ be an injective homomorphism between pointed graphs, and let $\mathcal{N}$ be an L layer MGNN such that $\mathcal{N}(G, v) = 1$. We will show that $\mathcal{N}(G', v') = 1$.

Since $\mathcal{N}$ is monotonic, its classification function is non-decreasing, so it suffices to show that $\lambda^{(L)}(v) \leq \lambda^{(L)}(v')$. To this end, we prove by induction on $\ell \leq L$ that $\lambda^{(\ell)}(w) \leq \lambda^{(\ell)}(h(w))$, for every node w in G.

For $\ell = 0$, we observe that h preserves node labels, so $\lambda^{(0)}(w) \leq \lambda^{(0)}(h(w))$. For the inductive step, assume the claim holds for $\ell < L$. Since h is an injective homomorphism, it induces an injection from the multiset of neighbours of w in G to the multiset of neighbours of $h(w)$ in G'. By the induction hypothesis, this yields $\{\!|\lambda^{(\ell)}(u) \mid u \in N_G(w)|\!\} \leq \{\!|\lambda^{(\ell)}(u') \mid u' \in N_{G'}(h(w))|\!\}$. By monotonicity of the aggregation and combination functions, we conclude that $\lambda^{(\ell+1)}(w) \leq \lambda^{(\ell+1)}(h(w))$.

Claim 2. MAX-MGNNs are preserved under homomorphisms.

The proof follows the same general structure as in the case of MGNNs. Let $h\colon (G, v) \to (G', v')$ be a homomorphism between pointed graphs, and let $\mathcal{N}$ be an L layer MAX-MGNN with $\mathcal{N}(G, v) = 1$. We need to show that $\mathcal{N}(G', v') = 1$. As before it suffices to prove by induction on $\ell \leq L$ that $\lambda^{(\ell)}(w) \leq \lambda^{(\ell)}(h(w))$, for each node w in G.

For $\ell = 0$ the inequality holds h, as a homomorphism, preserves node labels. For the inductive step, assume the claim holds for $\ell < L$. Since h is a homomorphism, each neighbour u of w in G is mapped to a neighbour $h(u)$ of $h(w)$ in G'. By the induction hypothesis, $\lambda^{(\ell)}(u) \leq \lambda^{(\ell)}(h(u))$ and hence $\mathrm{MAX}(\{\!|\lambda^{(\ell)}(u) \mid u \in N_G(w)|\!\}) \leq \mathrm{MAX}(\{\!|\lambda^{(\ell)}(u') \mid u' \in N_{G'}(h(w))|\!\})$. By monotonicity of the combination function, we conclude that $\lambda^{(\ell+1)}(w) \leq \lambda^{(\ell+1)}(h(w))$.

Claim 3. Augmented MGNNs are preserved under embeddings.

Let $h\colon (G, v) \to (G', v')$ be an embedding between pointed graphs, and let $\mathcal{N}$ be an L-layer augmented MGNN such that $\mathcal{N}(G, v) = 1$. We will show that $\mathcal{N}(G', v') = 1$. To this end, it again suffices to prove by induction on $\ell \leq L$ that $\lambda^{(\ell)}(w) \leq \lambda'^{(\ell)}(h(w))$, for every node w in G.

This time, however, layer 1 is different from later layers, so we will consider it separately. For $\ell = 0$, since h is an embedding, it preserves node labels, and so, $\lambda^{(0)}(w) = \lambda'^{(0)}(h(w))$. For $\ell = 1$, we have that the first layer is the augmented layer. Since it is applied pointwise, it follows that $\lambda^{(1)}(w) = \lambda'^{(1)}(h(w))$.

For $\ell \geq 1$, the ℓth layer is monotonic. Moreover $\lambda^{(1)}(w)$, for any node w, contains only non-negative numbers. Therefore we can use the same argument as we used to show that MGNNs are preserved under injective homomorphisms. Thus, $\lambda^{(\ell)}(w) \leq \lambda'^{(\ell)}(h(w))$. □

Moreover, we can observe that MGNNs are not preserved under homomorphisms. To this end, let $G = (V, E, \lambda)$ with $V = \{v, u, w\}$, $E = \{\{v, u\}, \{v, w\}\}$, $\lambda(v) = \lambda(u) = \lambda(w) = 1$, and let $G' = (V', E', \lambda')$ with $V' = \{v', u'\}$, $E = \{\{v', u'\}\}$, and $\lambda'(v) = \lambda'(u) = 1$. Define $\mathcal{N}$ with a single layer, aggregation function SUM, parameters $\mathbf{b} = 0$ and $\mathbf{A} = \mathbf{C} = 1$, and classification function $\mathsf{cls}(x) = 1$ iff $x \geq 3$. Then, $\mathcal{N}(G, v) = 1$, but $\mathcal{N}(G', v') = 0$. However, the function f

defined by $f(v) = v'$, $f(u) = u'$ and $f(w) = u'$ is a (non-injective) homomorphism from (G, v) to (G', v').

Theorem 15. *The following statements hold:*

- *For each $\exists\mathcal{GML}$-classifier there is an equivalent augmented MGNN classifier,*
- *For each $\exists\mathcal{ML}$-classifier there is an equivalent augmented* MAX*-MGNN classifier,*
- *For each $\exists^+\mathcal{GML}$-classifier there is an equivalent MGNN classifier,*
- *For each $\exists^+\mathcal{ML}$-classifier there is an equivalent* MAX*-MGNN classifier.*

Furthermore, each GNN above can be assumed to be simple, with truncated ReLU $\sigma(\mathbf{x}) = \min(\max(0, \mathbf{x}), 1)$ as activation function.

Proof. The statement consists of four claims, that we will prove one by one.

Claim 1. For each $\exists\mathcal{GML}$-classifier there is an equivalent augmented MGNN.

Fix $\varphi \in \exists\mathcal{GML}$ over propositions $\mathsf{PROP}(\varphi)$ whose subformulae are $\mathsf{sub}(\varphi) = (\varphi_1, \ldots, \varphi_L)$, where $k \leq \ell$ whenever φ_k is a subformula of φ_ℓ. We construct an equivalent augmented MGNN $\mathcal{N}_\varphi$ with layers $0, \ldots, L + 1$.

The first layer ($\ell = 0$) is an augmentation layer which extends node features so that, for each proposition $p \in \mathsf{PROP}(\varphi)$, both p and its negation $\neg p$ are explicitly represented. Concretely, if the original node labels are vectors in $\{0, 1\}^d$, then the preprocessing layer maps each node v to a vector $\lambda^{(0)}(v)$ that contains, for each $p \in \mathsf{PROP}(\varphi)$, a coordinate for p with the same value as in the initial vector and a coordinate for $\neg p$, where the latter is computed as 1 minus the value of the coordinate for p. Thus, the augmentation layer doubles the dimension of the input vector and the vectors it computes contain only non-negative entries. Note that the augmentation layer can be implemented as a pointwise affine transformation.

The remaining L layers correspond to the subformulae of φ and are specified below. The classification function cls is defined by $\mathsf{cls}(\mathbf{x}) = \mathsf{true}$ iff the last component of $\mathbf{x}$ is 1.

Each layer $1, \ldots, L + 1$ is simple (as Equation (2)):

$$\lambda'(v) = \sigma\Big(\mathbf{b} + \lambda(v)\mathbf{A} + \mathsf{agg}\big(\{\!\!\{\lambda(w) \mid w \in N_G(v)\}\!\!\}\big)\mathbf{C}\Big), \tag{3}$$

with activation function σ being the componentwise truncated ReLU $\sigma(\mathbf{x}) = \min(\max(0, \mathbf{x}), 1)$ and the aggregation function agg being MAX-n-SUM, where n is the largest integer occurring in a graded modalities of φ.

The entries of $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{L\times L}$ and $\mathbf{b} \in \mathbb{R}^L$ are defined as follows:

1. if φ_ℓ is a proposition or the negation of a proposition, then $C_{\ell\ell} = 1$;
2. if $\varphi_\ell = \varphi_j \wedge \varphi_k$, then $C_{j\ell} = C_{k\ell} = 1$ and $b_\ell = -1$;
3. if $\varphi_\ell = \varphi_j \vee \varphi_k$, then $C_{j\ell} = C_{k\ell} = 1$ and $b_\ell = 0$;
4. if $\varphi_\ell = \Diamond^{\geq c}\varphi_k$, then $A_{k\ell} = 1$ and $b_\ell = -c + 1$.

All other entries are set to 0.

Observe that all entries of $\mathbf{A}$ and $\mathbf{C}$ are non-negative, so $\mathcal{N}_\varphi$ is a positive-weight GNN with additional augmented layer. Hence, by Proposition 13, $\mathcal{N}_\varphi$ is an augmented MGNN, as required.

We now show correctness of the construction. Consider the application of $\mathcal{N}_\varphi$ to a graph $G = (V, E, \lambda)$. We show that for each subformula φ_ℓ, each $i \in \{\ell, \ldots, L\}$, and each node $v \in V$,

$$(G, v) \models \varphi_\ell \quad \text{iff} \quad \lambda^{(i)}(v)_\ell = 1.$$

This implies that $\mathcal{N}_\varphi(G, v) = \mathsf{true}$ iff $(G, v) \models \varphi$. The proof is by induction on the structure of φ_ℓ.

- If φ_ℓ is a proposition then, $\lambda(v)^{(0)}_\ell = 1$ if $(G, v) \models \varphi_\ell$, and otherwise $\lambda(v)^{(0)}_\ell = 0$. Moreover, since φ_ℓ is a proposition, all aggregate-combine layers have $C_{\ell\ell} = 1$, $b_\ell = 0$, and $A_{k\ell} = 0$ for each k, so $\mathsf{comb}(\mathbf{x}, \mathbf{y})_\ell = \mathbf{x}_\ell$. Thus $\lambda(v)^{(0)}_\ell = \lambda(v)^{(i)}_\ell$ for any i.

- If φ_ℓ is the negation of a proposition, then its value is computed by the augmented layer. In particular, $\lambda(v)_\ell^{(0)} = 1$ if $(G, v) \models \varphi_\ell$, and otherwise $\lambda(v)_\ell^{(0)} = 0$. As in the propositional case, the construction ensures that this value is preserved across layers, so $\lambda(v)_\ell^{(0)} = \lambda(v)_\ell^{(i)}$ for all $i \geq \ell$.

- If $\varphi_\ell = \varphi_j \wedge \varphi_k$, then by construction, we have $C_{j\ell} = C_{k\ell} = 1$, $b_\ell = -1$, and $A_{m\ell} = 0$ for each m. Then by Equation (2), we have

$$\lambda(v)_\ell^{(i)} = \sigma(\lambda(v)_j^{(i-1)} + \lambda(v)_k^{(i-1)} - 1).$$

  By the inductive hypothesis, $\lambda(v)_j^{(i-1)} = 1$ iff $(G, v) \models \varphi_j$, and similarly for φ_k. Hence, $\lambda(v)_\ell^{(i)} = 1$ iff $(G, v) \models \varphi_j \wedge \varphi_k$, and otherwise $\lambda(v)_\ell^{(i)} = 0$ for all $i \geq \ell$.

- If $\varphi_\ell = \varphi_j \vee \varphi_k$, then by construction, we have $C_{j\ell} = C_{k\ell} = 1$, $b_\ell = 0$, and $A_{m\ell} = 0$ for each m. Then by Equation (2), we have

$$\lambda(v)_\ell^{(i)} = \sigma(\lambda(v)_j^{(i-1)} + \lambda(v)_k^{(i-1)}).$$

  By the inductive hypothesis, $\lambda(v)_j^{(i-1)} = 1$ iff $(G, v) \models \varphi_j$, and similarly for φ_k. Hence, $\lambda(v)_\ell^{(i)} = 1$ iff $(G, v) \models \varphi_j \vee \varphi_k$, and otherwise $\lambda(v)_\ell^{(i)} = 0$ for all $i \geq \ell$.

- If $\varphi_\ell = \Diamond^{\geq c}\varphi_k$, then by construction, we have $A_{k\ell} = 1$, $b_\ell = -c + 1$, and $C_{m\ell} = 0$ for each m. Since agg is the MAX-n-SUM, by Equation (2), we have

$$\lambda(v)_\ell^{(i)} = \sigma\left(\min\Big(n, \sum_{w \in N_G(v)} \lambda(w)_k^{(i-1)}\Big) - c + 1\right).$$

  By the inductive hypothesis, $\lambda(w)_k^{(i-1)} = 1$ iff $(G, w) \models \varphi_k$. Hence, $\lambda(v)_\ell^{(i)} = 1$ iff $(G, v) \models \Diamond^{\geq c}\varphi_k$, and otherwise $\lambda(v)_\ell^{(i)} = 0$.

Claim 2. For each $\exists\mathcal{ML}$-classifier there is an equivalent augmented MAX-MGNN.

Recall that $\mathcal{ML}$ is a fragment of $\mathcal{GML}$ where all graded modalities have index 1. Then, the MAX-n-SUM function in our construction is MAX-1-SUM, which is simply the component-wise MAX function. Hence $\mathcal{N}_\varphi$ becomes an augmented MAX-MGNN, as required.

Finally we consider two last claims together:

Claim 3. For each $\exists^+\mathcal{GML}$-classifier there is an equivalent MGNN.

Claim 4. For each $\exists^+\mathcal{ML}$-classifier there is an equivalent MAX-MGNN.

In both cases formulae are negation-free, so the augmented layer in the construction of $\mathcal{N}_\varphi$ can be omitted. Hence our construction gives rise to an MGNN or MAX-MGNN, depending on the case. □

Theorem 16. *The following tight expressiveness results hold:*

- *Augmented MGNN classifiers have the same expressiveness as $\exists\mathcal{GML}$ classifiers,*
- *Augmented* MAX-*MGNN classifiers have the same expressiveness as $\exists\mathcal{ML}$ classifiers,*
- *MGNN classifiers have the same expressiveness as $\exists^+\mathcal{GML}$ classifiers,*
- MAX-*MGNN classifiers have the same expressiveness as $\exists^+\mathcal{ML}$ classifiers.*

Proof. Construction of a required type GNN from formulae from a given modal logic follows directly from Theorem 15.

For the opposite direction we need to show that for a GNN $\mathcal{N}$ of a given type there exists an equivalent formula φ from a particular modal logic. Let $\mathcal{C}$ be the class of pointed graphs accepted by $\mathcal{N}$. By Theorem 14, $\mathcal{C}$ is preserved under: embeddings if $\mathcal{N}$ is an augmented MGNN, injective homomorphisms if $\mathcal{N}$ is an MGNN, and under homomorphisms if $\mathcal{N}$ is a MAX-MGNN. Thus, by Barceló et al. (2020) there exists a formula φ in the required fragment of modal logic.

It remains to consider the case when $\mathcal{N}$ is an augmented MAX-MGNN, and show that it is equivalent to some $\exists\mathcal{ML}$ formula. We observe that every MAX-MGNN can be seen as a bounded AC-GNN with set-based aggregation, denoted as $\mathrm{GNN}_{\mathsf{s}}^{\mathsf{AC}}$, as defined in Cuenca Grau et al. (2026). It has been shown that $\mathrm{GNN}_{\mathsf{s}}^{\mathsf{AC}}$s have the same expressive power as $\mathcal{ML}$ Cuenca Grau et al. (2026), so we obtain that $\mathcal{N}$ is equivalent to some $\mathcal{ML}$ formula. Furthermore, recall that augmented MGNNs are preserved under embeddings (Theorem 14) so $\mathcal{N}$, as a special case of an augmented MGNN is also preserved under embeddings. Consequently, it allows us to apply a well-known Rosen's preservation theorem for $\mathcal{ML}$, stating that an $\mathcal{ML}$ formula is preserved under embeddings if and only if it is equivalent to an $\exists\mathcal{ML}$ formula (Rosen, 1997, Proposition 6). As a result we obtain that $\mathcal{N}$ is equivalent to an $\exists\mathcal{ML}$ formula. □