Learnable Commutative Monoids for
Graph Neural Networks

Euan Ong
University of Cambridge
ellyro2@cam.ac.uk

Petar Veličković
DeepMind / University of Cambridge
petarv@deepmind.com

Abstract
Graph neural networks (GNNs) have been shown to be highly sensitive to the choice of aggregation function. While summing over a node’s neighbours can approximate any permutation-invariant function over discrete inputs, Cohen-Karlik et al. [2020] proved there are set-aggregation problems for which summing cannot generalise to unbounded inputs, proposing recurrent neural networks regularised towards permutation-invariance as a more expressive aggregator. We show that these results carry over to the graph domain: GNNs equipped with recurrent aggregators are competitive with state-of-the-art permutation-invariant aggregators, on both synthetic benchmarks and real-world problems. However, despite the benefits of recurrent aggregators, their $O(V)$ depth makes them both difficult to parallelise and harder to train on large graphs. Inspired by the observation that a well-behaved aggregator for a GNN is a commutative monoid over its latent space, we propose a framework for constructing learnable, commutative, associative binary operators. And with this, we construct an aggregator of $O(\log V)$ depth, yielding exponential improvements for both parallelism and dependency length while achieving performance competitive with recurrent aggregators. Based on our empirical observations, our proposed learnable commutative monoid (LCM) aggregator represents a favourable tradeoff between efficient and expressive aggregators.

1 Introduction
When dealing with irregularly structured data [Bronstein et al., 2021], neural networks typically need to process data of arbitrary sizes. In such scenarios, the heart of the network is arguably its aggregation function—a function that reduces a collection of neighbour feature vectors into a single vector. Indeed, graph neural networks (GNNs) have been shown empirically to be highly sensitive to the choice of aggregator [Veličković et al., 2019, Richter and Wattenhofer, 2020], with a wide range of aggregators (e.g. sum, max and mean) and their combinations [Corso et al., 2020] in common use. In this paper, we offer a new perspective for studying aggregators, with clear theoretical and practical implications. It can be said that the true objective of choosing an aggregator is to make it as simple as possible (i.e. to minimise the sample complexity required) for the parameters of the GNNs to exploit that aggregator in a way that makes it easier to solve the learning problem. Specifically, we study this in the context of learning to align the GNN’s aggregator to a desirable target aggregation function (as defined in [Xu et al., 2019a]). It is already a known fact that higher alignment implies reduced sample complexity [Xu et al., 2019a], and in the context of algorithmic reasoning, it is well-known that a neural network will be better at learning to imitate an algorithm if its aggregator matches that of the algorithm it is trying to imitate [Veličković et al., 2019, Xu et al., 2020]. However, beyond the realm of learning a task with a concrete aggregator, many real-world problems offer more challenging settings, wherein the optimal aggregator to learn is not clear—but unlikely to be a trivial fixed aggregator. To formalise this notion, while preserving the useful assumption of permutation invariance, we leverage commutative monoids as a formalism for both the aggregators supported by GNNs and the (potentially unknown) target aggregators one would wish to align to.
This formalism allows us to derive several relevant results, including the fact that using any fixed commutative monoid $F$ (e.g. sum or max) as an aggregator would compel the GNN to learn a commutative monoid homomorphism from $F$ to the target commutative monoid, purely from data. We hypothesise that this is often difficult to do robustly, and verify our hypothesis by demonstrating several instances (both synthetic and real-world) where fixed aggregators (including combinations of them [Corso et al., 2020]) fail to generalise.

Our perspective, inspired by the functional programming motif of folds (or catamorphisms) over arbitrary data structures, leads us to consider flexible and learnable aggregation functions, which can more easily fit a wide range of commutative monoids directly, without needing to learn such a homomorphism. The most popular such aggregator has previously been the RNN (i.e. ‘a fold over a list’) – used, for instance, in GraphSAGE [Hamilton et al., 2017]. The reason for RNNs’ expressive power is simple: their usage of a hidden recurrent state allows them to break away from the constraints of commutative monoids and aggregate inputs more flexibly. However, while empirically powerful, the sequential structure of RNN aggregators leads to clear shortcomings in efficiency: if an RNN had learnt to aggregate $n$ neighbours under a commutative monoid operation $\oplus$, it would do so with a depth that is linear in $n$, as $(((\ldots(x_1 \oplus x_2) \oplus x_3) \oplus \ldots) \oplus x_{n-1}) \oplus x_n$.

But, by folding over a binary tree instead of a list (in other words, rearranging the order of operations to a balanced binary tree $\ldots((x_1 \oplus x_2) \oplus (x_3 \oplus x_4)) \oplus \ldots \oplus (x_{n-1} \oplus x_n) \ldots)$), we derive an aggregator that achieves a favourable trade-off between flexibility and efficiency, empirically retaining most of the performance of RNNs while having a depth that is logarithmic in $n$. We also demonstrate how such layers can be effectively constrained and regularised to respect the commutative monoid axioms (essentially creating a learnable commutative monoid), leading to further gains in robustness.

## 2 Motivation

Before exploring GNN aggregators, we first review the structure of a GNN. For a graph $G = (V, E)$ whose nodes $u$ have one-hop neighbourhoods $N_u = \{v \in V \mid (v, u) \in E\}$ and features $x_u$, a message-passing GNN over $G$ is defined by Bronstein et al. [2021] as

$$h_u = \phi \left( \bigoplus_{v \in N_u} \psi(x_u, x_v) \right)$$

for $\psi$ the message function, $\phi$ the readout function and $\bigoplus$ a permutation-invariant aggregation function.

This GNN ‘template’ can be instantiated in many ways, with different choices of $\phi$, $\psi$ and $\bigoplus$ yielding popular architectures such as GCNs [Kipf and Welling, 2017] and GATs [Veličković et al., 2018].

### 2.1 To learn a complex aggregator is to learn a commutative monoid homomorphism

So we’ve seen that, in order to define a GNN, we must define a permutation-invariant aggregator $\bigoplus$ over its messages. But how can we characterise a permutation-invariant aggregator in general?

In abstract algebra (and in functional programming), a permutation-invariant aggregator over a set can be described as (maps into and out of) a commutative monoid. A commutative monoid $(M, \bigoplus, e_\bigoplus)$ is a set $M$ equipped with a commutative, associative binary operator $\bigoplus : M \times M \rightarrow M$ and an identity element $e_\bigoplus \in M$ – in other words, an instance of the following Haskell typeclass, satisfying the identities to the right for all $x$, $y$, $z$:

```haskell
class CommutativeMonoid a = x <> e == e
  where e :: a
        x <> y == y <> x
        x <> (y <> z) == (x <> y) <> z
```

Intuitively, commutative monoids over a set $M$ are ‘operations you can use to reduce a multiset, whose members are in $M$, to a single value’. These include GNN aggregators, like sum-aggregation $(\mathbb{R}^n, +, 0)$ and max-aggregation $(\mathbb{R}^n, \max, 0)$. Indeed, Dudzik and Veličković [2022] observe that, for the aggregation function $\bigoplus$ of a GNN to be well-behaved (in the sense of respecting the axioms of the multiset monad), it must form a commutative monoid $(S, \bigoplus, e_\bigoplus)$ over some subspace $S$ of $\mathbb{R}^n$.

The vast majority of GNNs choose a fixed permutation-invariant function $\bigoplus$ (or fixed combinations of them [Corso et al., 2020]). While some research [Pellegrini et al., 2020, Li et al., 2020] has explored
aggregation functions with learnable parameters, these functions are only very weakly parameterised, and give us limited additional expressivity.

For problems where we can anticipate the kind of aggregation function we might need, this approach works well: indeed, choosing a commutative monoid that aligns values, or representations encoding many different types of data) for which these monoids may not always be the most natural choice for the aggregation we’re trying to learn. So in such cases, ψ and ϕ must take on some of the work of mapping our representations into and out of a space where \( \oplus \)-aggregation makes sense.

Formally, suppose we use a GNN equipped with a fixed commutative monoid aggregator \((F, \oplus, e_{\oplus})\), on a problem for which the ‘true’ aggregation we want to perform is the commutative monoid \((M, *, e_{\bullet})\) over the GNN’s latent space. What would it take for our GNN to perform \( M \)-aggregation?

**Proposition 1.** Let \((M, *, e_{\bullet})\) and \((F, \oplus, e_{\oplus})\) be commutative monoids. Then for functions \( g : M \to F \) and \( h : F \to M \),

\[
\forall x \in X \quad \ast x = h \left( \bigoplus_{x \in X} g(x) \right)
\]

for all finite multisets \( X \) of \( M \), if and only if \( h \) is both a left inverse of \( g \) and a surjective monoid homomorphism from \( \langle g(M) \rangle \subseteq F^1 \) to \( M \).

Now, given Proposition 1 above (proven in Appendix A), suppose we had a trained GNN, parameterised by \( \phi : \mathbb{R}^k \times F \to \mathbb{R}^k \) and \( \psi : \mathbb{R}^k \times \mathbb{R}^k \to F \), with a fixed \( F \)-aggregator. Suppose this GNN has learned to imitate the \( M \)-aggregation commutative monoid. We will model this property as there existing functions \( \phi' : \mathbb{R}^k \times M \to \mathbb{R}^k \) and \( \psi' : \mathbb{R}^k \times \mathbb{R}^k \to M \), \( g : M \to F \) and \( h : F \to M \) such that

- \( \phi(x_u, m_{X(u)}) = \phi'(x_u, h(m_{X(u)})) \)
- \( \psi(x_u, x_v) = g(\psi'(x_u, x_v)) \)

and \( \ast x = h(\bigoplus_{x \in X} g(x)) \) for all finite multisets \( X \) of \( M \).

(Observe that this implies the following:

\[
\phi \left( x_u, \bigoplus_{v \in N_u} \psi(x_u, x_v) \right) = \phi'(x_u, h \left( \bigoplus_{v \in N_u} g(\psi'(x_u, x_v)) \right)) = \phi'(x_u, \ast (\bigoplus_{v \in N_u} \psi'(x_u, x_v)))
\]

for all nodes \( u, v \) in graphs \( G \).)

Hence \( h \) is a surjective monoid homomorphism from \( \langle g(M) \rangle \) to \( M \) (i.e. \( M \) is a subquotient of \( F \)).

So at a high level, for a GNN with aggregator \( F \) to imitate an aggregator \( M \), it must learn a function that can decompose into a surjective monoid homomorphism from a submonoid of \( F \) to \( M \).

### 2.2 Limitations on expressivity and generalisation for constructed aggregators

Given this result, what are the implications for prior and present work?

As has been seen in [Veličković et al., 2019, Sanchez-Gonzalez et al., 2020], it’s clear that if our fixed commutative monoid \( F \) is aligned with a target monoid \( M \) for the problem we want to solve – intuitively, ‘if the homomorphism doesn’t have to do much work’ – then we can easily learn to imitate \( M \). Indeed, if the target homomorphism is linear, and we have appropriate training set coverage, then by [Xu et al., 2020] it may well generalise out-of-distribution – a result that holds (to an extent) in the case of learning to imitate path-finding algorithms such as Bellman-Ford [Veličković et al., 2019].

But there are many cases where \( M \) is more complex, and there is no commonly-used fixed aggregator \( F \) for which we can simply apply a linear homomorphism to get from \( F \) to \( M \). One such example is

\[ \langle g(M) \rangle \] denotes the submonoid of \( F \) generated by \( g(M) \).
the problem of finding the 2nd-minimum element in a set. Here, the desired monoid \( M \) is as follows:

(Snippet 2)

```haskell
type M = (Int, Int)
instance CommutativeMonoid M where
  secondMinimum :: [Int] -> Int
  secondMinimum = secondMinimum
  where
    secondMinimum (a1, a2) = if a2 > a1 then a2 else a1
    secondMinimum [x] = x

In principle, there exists an \( F \) from which it is possible to construct a homomorphism to \( M \): by [Zaheer et al., 2017] and [Xu et al., 2019a], for any \( (M, *, e_* \rangle \) with \( M \subseteq \mathbb{Q}^n \), there exists a surjective monoid homomorphism \( h \) from \( (\mathbb{R}^n, +, 0) \) to \( (M, *, e_*) \). But Wagstaff et al. [2019] show that this guarantee may require an \( h \) that is highly discontinuous, and therefore not only hard to learn in-distribution\(^2\), but fully misaligned with the assumptions of the universal approximation theorem. Further, as \( \text{dom}(h) = \langle g(M) \rangle \), we are not learning a function whose domain is a bounded set, so we have little hope of generalising out-of-distribution. Indeed, we demonstrate in Section 3.1 that all common fixed aggregators fail to learn the 2nd-minimum problem, both in and out of distribution.

Similarly, Cohen-Karlik et al. [2020] show that sum-aggregators as implemented in [Zaheer et al., 2017] (i.e. maps into and out of the \((\mathbb{R}^n, +, 0)\) commutative monoid) require \( \Omega(\log 2^n) \) neurons to learn the parity function over sets of size \( n \). Intuitively, the crux of their proof is that the homomorphism the aggregator would have to learn from \((\mathbb{R}^n, +, 0)\) to the parity monoid is a periodic function with unbounded domain. Similar arguments hold for all aggregation tasks involving modular counting.

### 2.3 Fully learnable recurrent aggregators and their limitations

We will now take a step back from homomorphisms, and try to discover a more flexible aggregator. An emerging narrative within deep learning is that of representations as types [Olah, 2015]. If we view the construction of neural networks as the construction of differentiable, parameterised pure functional programs, many of the design patterns commonly used in deep learning correspond to higher-order functions commonly used in functional programming (FP). This paradigm has proven valuable in recent times, embodied by deep learning frameworks such as JAX [Bradbury et al., 2018].

In FP, a simple way to aggregate a multiset of elements is to represent them as a list and fold over it:³

(Snippet 3)

```haskell
fold :: (a -> b) -> b -> [a] -> b
fold f z [] = z
fold f z (x:xs) = f x (fold f z xs)
```

And in some sense, a recurrent neural network (RNN) is simply a fold over a list, parameterised by a learnable accumulator \( f \) and a learnable initialisation element \( z \):⁴ (Snippet 4)

```haskell
rnn :: Learnable
    ([Vec R h1] -> Vec R h2)
initialState :: Learnable (Vec R h2)
rnn = fold rnnCell
```

Hence a natural way to construct a learnable aggregator over multisets could be to use an RNN—\( a \) ‘learnable fold’— and to somehow ensure it is permutation-invariant.

Indeed, this approach has been used for permutation-invariant set aggregation, with Murphy et al. [2019] enforcing permutation-invariance by design by taking the average of an RNN applied

---

²Suppose \( f : X \rightarrow Y \) is a model trained to learn \( h : X \rightarrow Y \) given a training set \( \{(x_i, y_i)\}_{i=1}^n \subseteq D \) for \( y_i = h(x_i) \) and \( D \) the support of the training distribution. Now, for some loss function \( L : Y \times Y \rightarrow \mathbb{R} \), we say that \( f \) has learned \( h \) in-distribution if \( \mathbb{E}_{x \sim D}[L(f(x), h(x))] \) is small, and that \( f \) has learned \( h \) out-of-distribution if \( \mathbb{E}_{x \sim P}[L(f(x), h(x))] \) is small for distributions \( P \) over \( X\setminus D \).

³Note that \( a \rightarrow b \rightarrow b \) is an equivalent way (via currying) of specifying a function \( a \times b \rightarrow b \).

⁴Note that an RNN can also be viewed as a map to the carrier set of the monoid of endofunctions (i.e. functions from a set to itself—in this case, from \( b \) to \( b \)) under composition: see Appendix B for details.
to all permutations of its input, and Cohen-Karlik et al. [2020] regularising RNNs $f$ towards permutation-invariance by adding a pairwise regularisation term $L_{\text{swap}}(x_1, x_2) = \frac{1}{2} (f(f(x_1, x_2) - f(f(x_2, x_1)))^2$ (which we motivate through the lens of commutative monoids in Appendix B).

Recurrent aggregators have also occasionally seen use in GNNs [Hamilton et al., 2017, Xu et al., 2018], but they are scarcely used despite their competitive performance. We assume RNNs likely remain unpopular as a GNN aggregator due to their depth. Indeed, observe that an $N$-layer GNN equipped with a recurrent aggregator has (worst-case) depth $O(VN)$. By contrast, the same GNN equipped with a fixed aggregator has (worst-case) depth $O(N)$. And as many graphs on which we want to deploy GNNs can have upwards of 100,000 nodes [Hu et al., 2020], the same problems of efficiency and maximum dependency length observed by Vaswani et al. [2017] when using RNNs for sequence transduction also hold when using RNNs for graph message aggregation.

### 2.4 A compromise: fully learnable commutative monoids

So, if recurrent aggregators are too deep, is there any way to get a fully learnable aggregator? We’ve considered the fixed-aggregator approach, where we learn maps into and out of the carrier set of a pre-determined commutative monoid. We’ve considered the recurrent-aggregator approach, where we represent multisets as lists and implement aggregation as a learnable fold over lists. But another way to represent multisets in FP is as a balanced binary tree, over which aggregation is implemented as a fold parameterised by a commutative monoid. So what if we implemented aggregation as a learnable fold over a balanced binary tree? Or in other words, what if, instead of learning maps into and out of some commutative monoid, we simply learn the commutative monoid itself?

Let’s make precise what exactly we mean by ‘learning a commutative monoid’ for use in a GNN. Recall that a commutative monoid $(M, \oplus, e_M)$ is defined by its carrier set $M$, its binary operation $\oplus$ and its identity element $e_M$. So given some learnable commutative, associative binary operator $\oplus$ (written $\text{binOp} :: \text{Learnable} \ (\text{Vec} \ R \ h \rightarrow \text{Vec} \ R \ h \rightarrow \text{Vec} \ R \ h)$), and some learnable identity element $e_\oplus$ (written $\text{identity} :: \text{Learnable} \ (\text{Vec} \ R \ h)$), we can define a learnable commutative monoid over some learned embedding space (in other words, a subset of $\mathbb{R}^h$): (Snippet 5)

```haskell
instance CommutativeMonoid HiddenState where
  e = identity; <> = binOp
```

Thus, our aggregation function can be specified simply, as $\oplus_x x$, or

```haskell
aggregate :: Learnable ([HiddenState] -> HiddenState)
aggregate = reduce (<>)
```

Note that, here, the carrier set is implicit – when used in a GNN, we expect the message function (i.e. the producer of the elements to be aggregated) to learn a ‘return type’ representation whose members are elements of this implicit carrier set, and similarly for the ‘input type’ of the readout function.

Now, why do we care about this at all? Indeed, if we implement reduce as a fold, we’re no better off than if we just used a recurrent aggregator. But consider the computation graph (or rather, computation binary tree) of such an aggregation $x_1 \oplus (x_2 \oplus (x_3 \oplus x_4))$. By Tamar’s theorem [Tamari, 1962], the associativity of $\oplus$ means that the result of evaluating this computation tree is invariant under rotations of nodes in the tree. Therefore, in order to minimise the depth of the computation, we can rewrite our reduction as a balanced binary tree: $(x_1 \oplus x_2) \oplus (x_3 \oplus x_4)$ (see Appendix D).

And by doing so, for $V$ elements to aggregate, we obtain a network with $O(V)$ applications of $\oplus$ and $O(\log V)$ depth – an exponential improvement over our $O(V)$-depth recurrent aggregators.

### 2.5 Commutative, associative binary operators for learnable commutative monoids

So, given a commutative, associative binary operator, we can get our learnable commutative monoid with $O(\log V)$ depth. But how do we construct such an operator in the first place? As with permutation-invariant RNNs, we have two options: either we construct an operator that strongly enforces the axioms of commutativity and associativity by construction, or we construct some arbitrary binary operator and weakly enforce the axioms through regularisation.

---

$^3$Alternatively, we can see this, as in Appendix B, as learning maps into and out of the carrier set of the monoid of endofunctions.
Learnable Commutative Monoids for Graph Neural Networks

**Strong enforcement.** While some research has been conducted into learning algebraic structures with strongly enforced axioms [Abe et al., 2021, Martires, 2021], these approaches reduce to learning maps into and out of a fixed aggregator. We observe that, while we can strongly enforce commutativity in any binary operator \( f(x, y) \) by symmetrising it to \( g(x, y) = \frac{f(x, y) + f(y, x)}{2} \), we found no such construction for associativity which doesn’t sacrifice expressivity.

So given this, and given the importance of gating [Tallec and Ollivier, 2018] in neural networks applied over long time horizons, we can construct a simple strongly commutative binary aggregator (Binary-GRU) by symmetrising a GRU [Cho et al., 2014]: (Snippet 6)

```haskell
binaryGRU :: Learnable (Vec R h -> Vec R h -> Vec R h)
binaryGRU v1 v2 = do
g <- new gruCell (InputDim h) (HiddenDim h)
return (g v1 v2 + g v2 v1) / 2
```

**Weak enforcement.** Alternatively, just as we saw with recurrent aggregators in Section 2.3, for a learnable binary operator \( \oplus : \mathbb{R}^n \rightarrow \mathbb{R}^n \rightarrow \mathbb{R}^n \) we could weakly enforce commutativity and associativity through regularisation losses \( L_{\text{comm}}(x, y) = \lambda_{\text{comm}}(|| (x \oplus y) - (y \oplus x) ||^2 \) and \( L_{\text{assoc}}(x, y, z) = \lambda_{\text{assoc}}(|| (x \oplus (y \oplus z)) - ((x \oplus y) \oplus z) ||^2 \) (for implementation details, see Appendix E).

Now, by applying \( L_{\text{assoc}} \) to Binary-GRU, we obtain a strongly commutative, weakly associative binary operator (Binary-GRU-Assoc).

3 Assessing the utility of learnable commutative monoids

Now, we’ve seen three types of aggregator: fixed aggregators, recurrent aggregators and learnable commutative monoids. In order to explore their trade-offs in terms of expressivity, generalisation and efficiency, we conduct a range of experiments comparing the performance of

- state-of-the-art fixed aggregators (such as sum-aggregation [Zaheer et al., 2017], max-aggregation [Veličković et al., 2019] and PNA [Corso et al., 2020]),
- recurrent aggregators (specifically GRUs [Cho et al., 2014]), and
- learnable commutative monoid (LCM) aggregators (using the Binary-GRU and Binary-GRU-Assoc learnable operators as described in Section 2.5)

on the following synthetic and real-world problems:

2nd-minimum. We test fixed aggregators, recurrent aggregators and learnable commutative monoids on the problem of finding the second-smallest element in a set of binary-encoded integers. As observed in Section 2.2, this task is a synthetic aggregation problem with an ‘unusual’ commutative monoid, in that it doesn’t align well with common fixed aggregators. Therefore, we expect this task to be a standard problem for which learnable aggregators would outperform any commonly-used fixed aggregator, especially out-of-distribution.

PNA synthetic benchmark. We then proceed to test the in-distribution performance of our aggregators on the synthetic dataset presented in [Corso et al., 2020]. This dataset consists of aggregator-heavy, classical graph problems that are mostly aligned with the aggregators used to construct PNA. Thus, we expect PNA (and the relevant fixed aggregators) to perform strongly here, potentially even out-of-distribution. But while our learnable aggregators don’t necessarily have the inductive bias to approximate these monoids well over an unbounded domain, we expect them to perform competitively at learning the relevant monoids in-distribution.

PNA real-world benchmark. Finally, we test our aggregators on the real-world dataset presented in [Corso et al., 2020], consisting of chemical (ZINC and MolHIV) and computer vision (CIFAR10 and MNIST) datasets from the GNN benchmarks of Dwivedi et al. [2020] and Hu et al. [2020]. In contrast to the algorithmic tasks in the synthetic benchmark, we expect these real-world problems to contain ‘unusual’ target monoids: for both molecular and computer vision problems, it is likely that

---

6Note that we can instantiate this operator with different values of the regularisation parameter \( \lambda_{\text{assoc}} \) (hereafter referred to as \( \lambda \)) by which we scale the associativity loss.
our GNN will learn complex representations whose most natural monoid is not the image of a simple homomorphism from any common fixed aggregator. Therefore, we expect fully learnable aggregators (GRU and LCMs) to outperform fixed aggregators on this benchmark.

Training details for all experiments are provided in Appendix F. Notably, for all uses of learnable aggregators, we randomly shuffle each batch of sequences before feeding it to the aggregator as a form of regularisation through data augmentation.

3.1 2nd-minimum

For this experiment, we compared fixed (sum, max, PNA), recurrent (GRU) and LCM (Binary-GRU) aggregators on the synthetic 2nd-minimum set aggregation problem. In order to evaluate the effects of regularisation towards algebraic axioms on the performance of LCM aggregators, we also tested Binary-GRU-Assoc, sweeping over values of the regularisation parameter \( \lambda \) from \( 10^{10} \) to \( 10^{-7} \).

3.1.1 Experimental details

For training data, we used 65,536 multisets of integers \( \sim U(0, 255) \) of size \( \sim U(1, 16) \). For validation data, we used 1,024 multisets of integers \( \sim U(0, 255) \) of size 32. For evaluation data, we used 1,024 multisets of integers \( \sim U(0, 255) \) of size \( l \), for \( l \in [1, 200] \). We used a standard multiset-aggregation architecture \( f(X) := \sigma(\psi(\bigoplus_{x \in X} \phi(x))) \) for \( \bigoplus \) the aggregator being tested, and \( \phi \) and \( \psi \) MLPs. \( f \) takes as input a vector of 8-bit binary-encoded integers (as in [Yan et al., 2020]), and returns a binary-encoded integer in \([0, 1]^8\). The full architecture (with details on integer embedding) is outlined in Appendix C.

3.1.2 Results and discussion

Summary. Recall that this problem was chosen for its comparatively unusual commutative monoid, which we do not expect aligns well with fixed aggregators. Indeed, we confirm this hypothesis: we see in Figure 1 that fixed aggregators fail to learn 2nd-minimum in-distribution, that recurrent aggregators learn 2nd-minimum near-perfectly in-distribution, generalising well out-of-distribution, and that LCM aggregators learn 2nd-minimum near-perfectly in-distribution and are competitive with recurrent aggregators out-of-distribution, while achieving an exponential speedup over recurrent aggregators on large sets. Furthermore, we observe that regularising towards algebraic axioms improves the performance of LCM aggregators both in and out of distribution.

![Figure 1](image_url)

Figure 1: Generalisation performance for fixed (max, sum, PNA), recurrent (GRU) and LCM (Binary-GRU) aggregators, along with the best-performing regularised LCM aggregator (Binary-GRU-Assoc with \( \lambda = 10^{10} \)). The shaded region is bounded above and below by the maximum and minimum values across all runs. The vertical purple line denotes the maximum set size present in training data (16); the vertical blue lines denote powers of 2 (from \( 2^1 \) to \( 2^7 \)). For detailed results, see Appendix G.
In-distribution performance. Examining Figure 1, observe that only the fully-learnable aggregators – GRU, Binary-GRU and Binary-GRU-Assoc – managed to learn 2nd-minimum near-perfectly in-distribution, with the next best performing aggregator being PNA.8

Out-of-distribution performance (without regularisation). Observe that, out-of-distribution, all learnable aggregators generalise near-perfectly up to size 32 (twice the size of the input). Beyond this point, while the performance of the recurrent aggregator decays slowly (reaching 0.912 ± 0.017 at size 200), the performance of the LCM quickly drops (reaching 0.287 ± 0.068 at size 200). Despite this, both learnable aggregators consistently outperform the fixed aggregators out-of-distribution. Furthermore, out of the fixed aggregators, we see that the sum-aggregator’s performance plateaus extremely quickly, a result we may attribute to the domain of the learned homomorphism from the sum-aggregator being an unbounded set (see Section 2.2).

Efficiency. As hypothesised in Section 2.4, we see (in Appendix G, Figure 3) that LCMs are indeed exponentially faster than RNNs for large sets: for \( n = 20 \), Binary-GRU-Assoc takes 48.2 ± 0.4 seconds per epoch, and GRU takes 46.6 ± 0.5 seconds per epoch, while for \( n = 200 \), Binary-GRU-Assoc takes 79.4 ± 0.5 seconds per epoch, and GRU takes 397.2 ± 1.3 seconds per epoch.

Regularisation towards associativity. We show the results from the best-performing regularised LCM aggregator (\( \lambda = 10^0 \)) in Figure 1 and Table 2. Although the unregularised Binary-GRU performs better than all fixed aggregators, observe that the regularised Binary-GRU-Assoc outperforms its unregularised sibling both in and out of distribution, and achieves generalisation performance competitive with GRU. Furthermore, observe that the sudden performance drops experienced by Binary-GRU when the size of the set reaches a power of two (i.e. when the depth of the aggregation tree increases) are noticeably dampened for Binary-GRU-Assoc, suggesting that regularisation towards associativity helps prevent overfitting to a particular maximum aggregation tree height. For interest, we present the full results of the regularisation parameter sweep in Figure 4 in Appendix G.

3.2 PNA synthetic benchmark

For this experiment, we trained recurrent (GRU) and LCM (Binary-GRU, Binary-GRU-Assoc) aggregators on the synthetic benchmark from [Corso et al., 2020], comparing against the fixed-aggregator baselines presented there (for GATs [Velicković et al., 2018], GCNs [Kipf and Welling, 2017], GINs [Xu et al., 2019b] and MPNNs [Gilmer et al., 2017] with sum and max aggregators).

3.2.1 Experimental details

In the PNA paper [Corso et al., 2020], experiments testing fixed aggregators (sum, max, PNA) are conducted on a custom GNN architecture centred around an MPNN layer with dimension 16, split into four towers each with hidden dimension 4. As we hypothesise that the low dimensionality of these towers could harm the expressivity of learnable aggregators, we test our learnable aggregators both in MPNNs of hidden dimension 16, with four towers of hidden dimension 16, and in MPNNs of hidden dimension 128, with one tower of hidden dimension 128.

3.2.2 Results and discussion

Summary. Recall that this dataset consists of aggregator-heavy classical graph problems9 that are mostly aligned with the aggregators used to construct PNA. So, as expected, we see in Table 1 that PNA outperforms all other aggregators tested on the dataset in-distribution. But observe that, on these problems, our asymptotically more efficient LCMs are competitive with and sometimes beat GRUs – and indeed, on the node-based problems in the dataset, our LCMs are as strong as PNA.

In Appendix G, we observe the surprising result that LCMs are more stable than PNA out-of-distribution (OOD), and that regularising LCMs towards associativity improves OOD performance at the cost of impairing performance in-distribution. We also discuss the effects of increasing dimensionality on fixed aggregator performance, through the lens of commutative monoid homomorphisms.

8Note that, out of the fixed aggregators, PNA was the only one to achieve near-perfect accuracy on the training dataset, with a maximum training accuracy of around 0.997.

9three node-based algorithmic tasks (single-source shortest paths, eccentricity and computing the Laplacian of node feature vectors) and three graph-based algorithmic tasks (connectedness, diameter and spectral radius)
Learnable Commutative Monoids for Graph Neural Networks

In-distribution performance. Observe in Table 1 that, while PNA beats all other aggregators tested, our learnable aggregators perform competitively in-distribution, with all learnable aggregators (GRU, Binary-GRU and Binary-GRU-Assoc) beating all single-aggregator (i.e. non-PNA) architectures. Interestingly, our Binary-GRUs perform better than the corresponding GRUs: perhaps their inductive bias towards commutativity helps us learn in-distribution.

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg score</th>
<th>Node tasks</th>
<th>Graph tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SSSP</td>
<td>Ecc</td>
</tr>
<tr>
<td>GCN</td>
<td>-2.05</td>
<td>-2.16</td>
<td>-1.89</td>
</tr>
<tr>
<td>GAT</td>
<td>-2.26</td>
<td>-2.34</td>
<td>-2.09</td>
</tr>
<tr>
<td>GIN</td>
<td>-1.99</td>
<td>-2.00</td>
<td>-1.90</td>
</tr>
<tr>
<td>MPNN (sum)</td>
<td>-2.50</td>
<td>-2.33</td>
<td>-2.26</td>
</tr>
<tr>
<td>MPNN (max)</td>
<td>-2.53</td>
<td>-2.36</td>
<td>-2.16</td>
</tr>
<tr>
<td>PNA-128</td>
<td>-3.09</td>
<td>-2.94</td>
<td>-2.88</td>
</tr>
<tr>
<td>GRU</td>
<td>-2.91</td>
<td>-2.84</td>
<td>-2.71</td>
</tr>
<tr>
<td>Binary-GRU</td>
<td>-3.00</td>
<td>-2.85</td>
<td>-2.77</td>
</tr>
</tbody>
</table>

Table 1: Mean $\log_{10}(MSE)$ on the PNA test dataset

Per-task performance. We present the per-task performance of all 128-dimensional aggregators (together with fixed-aggregator baselines) in Table 1. Observe that, in fact, Binary-GRU-Assoc outperforms Binary-GRU in all tasks apart from the graph Laplacian. Furthermore, while learnable aggregators do not perform as strongly as fixed aggregators on whole-graph tasks, they perform as well as or better than fixed aggregators for node-based tasks. This may be because the benchmark implementation for whole-graph tasks uses a sum-aggregator over the readout values: it is likely difficult to learn a homomorphism from the sum aggregator to the complex latent-space monoid learned by the LCM, and perhaps fixed aggregators provide an inductive bias towards learning representations for which it is easier to map to and from the sum-aggregation monoid.

3.3 PNA real-world benchmark

For this experiment, we trained recurrent (GRU) and LCM (Binary-GRU) aggregators on the real-world benchmark from Corso et al. [2020], containing two molecular graph property prediction datasets (ZINC and MolHIV) and two superpixel graph classification datasets (CIFAR10 and MNIST). Note that, due to limitations on compute resources, we were not able to perform a regularisation parameter sweep to test Binary-GRU-Assoc. The GNN architecture used here is identical to that in [Corso et al., 2020], except that, for learnable aggregators, all MPNN towers have the same dimensionality as the MPNN itself (i.e. we do not divide the towers).

3.3.1 Results and discussion

Summary. Recall that the real-world benchmark has complex problems that do not necessarily align with common fixed aggregators. We observe in Figure 2 that, while PNA in general outperforms all other aggregators on property prediction problems over small molecular graphs, the more expressive GRU substantially outperforms PNA for the (more discrete) task of image classification. Also, note that the (asymptotically efficient) Binary-GRU LCM provides a good trade-off between these two aggregators, being the second-best aggregator for all but two problems. Finally, we see that learnable aggregators appear particularly powerful on problems involving graphs with edge features.

Molecular datasets. Observe that PNA is the strongest aggregator over both the ZINC dataset without edge features and the HIV dataset – indeed, due to the continuous nature of the properties we want to estimate in these datasets, it seems likely that the ‘natural’ monoids for aggregation over graphs in these datasets would align well with fixed aggregators.

Image datasets. By contrast, we observe that GRU-aggregators are the strongest when testing on image data, likely as their expressivity lets them easily learn a complex, perhaps more discrete aggregation function. And while Binary-GRU does not do quite as well as GRU here, in all but one case it outperforms PNA on this problem.

Edge features. Finally, observe that, if we add edge features to ZINC, GRU outperforms PNA – and comparing results on the CIFAR-10 dataset with and without edge features, the average accuracy
<table>
<thead>
<tr>
<th>Model</th>
<th>Zinc (MAE)</th>
<th>CIFAR10 (Acc)</th>
<th>MNIST (MAE)</th>
<th>MolHIV (%ROC-AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No edge features</td>
<td>std</td>
<td>No edge features</td>
<td>std</td>
</tr>
<tr>
<td>Devvedi et al et al</td>
<td>MLP</td>
<td>0.710</td>
<td>0.001</td>
<td>58.01</td>
</tr>
<tr>
<td></td>
<td>GIN</td>
<td>0.469</td>
<td>0.002</td>
<td>54.46</td>
</tr>
<tr>
<td></td>
<td>DiffPoll</td>
<td>0.466</td>
<td>0.006</td>
<td>57.59</td>
</tr>
<tr>
<td></td>
<td>GAT</td>
<td>0.463</td>
<td>0.002</td>
<td>65.46</td>
</tr>
<tr>
<td></td>
<td>Monet</td>
<td>0.407</td>
<td>0.007</td>
<td>53.42</td>
</tr>
<tr>
<td></td>
<td>GatedGCN</td>
<td>0.423</td>
<td>0.008</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>MPNN (sum)</td>
<td>0.381</td>
<td>0.005</td>
<td>0.288</td>
</tr>
<tr>
<td></td>
<td>MPNN (max)</td>
<td>0.468</td>
<td>0.002</td>
<td>0.328</td>
</tr>
<tr>
<td></td>
<td>PNA (no scalers)</td>
<td>0.413</td>
<td>0.006</td>
<td>0.247</td>
</tr>
<tr>
<td></td>
<td>PNA</td>
<td>0.320</td>
<td>0.032</td>
<td>0.188</td>
</tr>
<tr>
<td>Ours</td>
<td>Binary-GRU</td>
<td>0.340</td>
<td>0.003</td>
<td>0.175</td>
</tr>
<tr>
<td></td>
<td>GRU</td>
<td>0.342</td>
<td>0.004</td>
<td>0.175</td>
</tr>
</tbody>
</table>

**Figure 2:** Results of learnable aggregators on the PNA real-world dataset, in comparison with those analysed by Corso et al. [2020]. Best results in bold-face, second-best in underline.

improvement for fixed aggregators when adding edge features is 0.34%, whereas the equivalent improvement for learnable aggregators is 2.33%. Learnable aggregators may be particularly strong on tasks with edge features, as making full use of them tends to require the learning of a more complex aggregation function.

### 4 Conclusions

In this work we have conducted a thorough study of aggregation functions within graph neural networks (GNNs), demonstrating both theoretically and empirically that many tasks of practical interest rely on a nontrivial integration of neighbourhoods (i.e. a nontrivial commutative monoid). This motivates the use of fully-learnable aggregation functions, but prior proposals based on RNNs had several shortcomings in terms of efficiency. Accordingly, we propose learnable commutative monoid (LCM) aggregators, which trade off the flexibility of RNNs with efficiency of fixed aggregators, producing a simple, yet empirically powerful, GNN aggregator with only $O(\log V)$ depth.

**Implications for GNN practitioners.** Based on our results, we present some suggestions to those using GNNs in practice:

- **When choosing a fixed aggregator** $F$ for a GNN architecture, consider the type of aggregation your problem is likely to involve – if it can be framed as a commutative monoid $M$, is it likely that a homomorphism can be learned from $F$ to $M$?

- **For graph problems manipulating discrete data**, or problems for which the aggregation required doesn’t align with existing fixed aggregators, learnable aggregators may improve performance (especially out-of-distribution).

- **When choosing a learnable aggregator**, for problems over small graphs, recurrent aggregators will likely perform well – but if they prove too slow, you may wish to try a learnable commutative monoid aggregator.

- **And if your learnable aggregator is overfitting**, perhaps try regularising it towards the relevant axioms (e.g. invariance under pairwise swaps for recurrent aggregators, commutativity and associativity for learnable commutative monoids).
Author Contributions

The research idea of exploring learnable commutative monoid (LCM) aggregators was originated by Euan Ong and steered by Petar Veličković. The experimental pipeline was developed and experiments were conducted by Euan, with oversight, mentorship and management from Petar. The formal analysis of fixed aggregators and its framing in terms of functional programming were originated by Euan, with advice from Petar. Both authors contributed to writing the paper and responding to reviewer feedback.

Acknowledgements

We would like to thank Pietro Liò and Malcolm Scott for generously providing access to compute resources at short notice, without which this work would not be possible. We also thank Andrew Dudzik and Karl Tuyls for reviewing the paper prior to submission, and all our anonymous reviewers for their careful feedback.

References


Andrew Dudzik and Petar Veličković. Graph Neural Networks are Dynamic Programmers. pages 1–9, 2022. URL http://arxiv.org/abs/2203.15544. 2


A Proof of Proposition 1

Proposition 1. Let \((M, *, e_x)\) and \((F, \oplus, e_\oplus)\) be commutative monoids. Then for functions \(g : M \to F\) and \(h : F \to M\),

\[
\forall x \in X \quad x = h \left( \bigoplus_{x \in X} g(x) \right)
\]

for all finite multisets \(X\) of \(M\), if and only if \(h\) is both a left inverse of \(g\) and a surjective monoid homomorphism from \(\langle g(M) \rangle \subseteq F^{10}\) to \(M\).

Proof. We proceed by cases.

\((\to)\) Suppose \(\forall x \in X \quad x = h(\bigoplus_{x \in X} g(x))\) for all finite multisets \(X\) of \(M\).

When \(X = \{x\}\), have that \(h(g(x)) = x\) trivially, so \(h\) must be a left inverse of \(g\) (and is therefore surjective).

Now for \(x, y \in \langle g(M) \rangle\), we want to show that \(h(x \oplus y) = h(x) \ast h(y)\) and that \(h(e_\oplus) = e_x\).

To show the former, observe that \(x = \bigoplus_{a \in A} g(a)\) and \(y = \bigoplus_{b \in B} g(b)\) for some finite multisets \(A, B\) of \(M\).

Now have that

\[
\begin{align*}
h(x \oplus y) &= h \left( \bigoplus_{a \in A} g(a) \right) \oplus \left( \bigoplus_{b \in B} g(b) \right) \\
&= h \left( \bigoplus_{x \in A \cup B} g(x) \right) \\
&= \ast \quad x \\
&= \left( \ast \quad a \right) \ast \left( \ast \quad b \right) \\
&= h \left( \bigoplus_{a \in A} g(a) \right) \ast h \left( \bigoplus_{b \in B} g(b) \right) \\
&= h(x) \ast h(y)
\end{align*}
\]

as desired.

To show the latter, observe that \(h(e_\oplus) \ast h(f) = h(e_\oplus \oplus f) = h(f)\) for all \(f \in F\). As \(h\) is surjective, we have that \(h(F) = M\), so \(h(e_\oplus) \ast m = m \ast h(e_\oplus) = m\) for all \(m \in M\), and \(h(e_\oplus) = e_x\).

\(^{10}\langle g(M) \rangle\) denotes the submonoid of \(F\) generated by \(g(M)\).
(←) Suppose \( h \) is a left inverse of \( g \) and a surjective monoid homomorphism from \( \langle g(M) \rangle \) to \( M \). Then

\[
\begin{align*}
    h \left( \bigoplus_{x \in X} g(x) \right) &= h \left( \bigoplus_{i=1}^{n} g(x_i) \right) \\
                                   &= h \left( f(x_1) \bigoplus_{i=2}^{n} g(x_i) \right) \\
                                   &= h(g(x_1)) \ast h \left( \bigoplus_{i=2}^{n} g(x_i) \right) \\
                                   &= x_1 \ast h \left( \bigoplus_{i=2}^{n} g(x_i) \right) \\
                                   &= \ldots \\
                                   &= \bigast_{i=1}^{n} x_i \\
                                   &= \bigast_{x \in X} x
\end{align*}
\]

as desired.

B Motivating the conditions for permutation-invariance in RNNs

An alternative way to motivate the regularisation loss of Cohen-Karlik et al. [2020], through the lens of monoids, is to frame the recurrent aggregator as a monoid, and identify the conditions required for this monoid to be commutative.

Keeping in mind that ‘RNNs are just learnable folds’, we notice that endofunctions form a monoid under composition:

\[
\text{instance Monoid (a -> a) where} \\
e = \text{id} \\
<> = (.)
\]

and observing that, for instance,

\[
\begin{align*}
    \text{fold f z [x1, x2, x3]} &= f \ x1 \ (f \ x2 \ (f \ x3 \ z)) \\
    &= (f \ x1 \ . \ f \ x2 \ . \ f \ x3) \ z \\
    &= (\$ \ z) \ (f \ x1 \ . \ f \ x2 \ . \ f \ x3) \\
    &= (\$ \ z) \ (\text{reduce } (.) \ (\text{map } f \ [x1; x2; x3]))
\end{align*}
\]

we can rewrite fold as an aggregation over the composition monoid:

\[
\begin{align*}
    \text{fold :: (a -> b -> b) -> b -> [a] -> b} \\
    \text{fold f z = dec . reduce (.) . map enc} \\
    \text{where} \\
    \text{enc x = f x} \\
    \text{dec f = f z}
\end{align*}
\]

Now, applying this to our recurrent aggregator, we have

\[
\begin{align*}
    \text{rnn :: Learnable ([Vec R h1] -> Vec R h2)} \\
    \text{rnn = dec . reduce (.) . map enc} \\
    \text{where} \\
    \text{enc x = rnnCell x} \\
    \text{dec f = f initialState}
\end{align*}
\]
Observe that, for \( \text{rnn} \), the carrier set of the composition (sub)monoid consists of functions \( \text{rnnCell} \) for inputs \( x \) to the aggregation function. So, in order to enforce that this monoid is commutative, we must simply ensure that

\[
f \mathbin{<>} g = g \mathbin{<>} f
\]

\[
= (\text{rnnCell} \ x1) \cdot (\text{rnnCell} \ x2) = (\text{rnnCell} \ x2) \cdot (\text{rnnCell} \ x1)
\]

\[
= \text{rnnCell} \ x1 \ (\text{rnnCell} \ x2 \ h) = \text{rnnCell} \ x2 \ (\text{rnnCell} \ x1 \ h)
\]

for all inputs \( x1, x2 \) and all hidden states \( h \).

**C Architecture used for 2\textsuperscript{nd}-minimum benchmark**

We present Haskell pseudocode for the architecture used in the 2\textsuperscript{nd}-minimum benchmark below.

\[
h = 128
\]

\[
\text{ofMlp} :: \text{Learnable} \ (\text{Vec R h} \rightarrow \text{Vec R h})
\]

\[
\text{ofMlp} = \text{do}
\]

\[
\text{dense} \leftarrow \text{new ofLinearLayer} \ (\text{In h}) \ (\text{Out h})
\]

\[
\text{return} \ \text{gelu} \ . \ \text{dense}
\]

\[
\text{intEmbedding} :: \text{Learnable} \ (\text{Vec Bool 8} \rightarrow \text{Vec R h})
\]

\[
\text{intEmbedding} = \text{toLearnable} \ \lambda \int \rightarrow \text{do}
\]

\[
\text{one_vecs} \leftarrow \text{newList} \ (\text{Length 8}) \ (\text{Of (learnableParameter (Dim h))})
\]

\[
\text{zero_vecs} \leftarrow \text{newList} \ (\text{Length 8}) \ (\text{Of (learnableParameter (Dim h))})
\]

\[
\text{return}
\]

\[
[ \text{one}^* \ + \ \text{zero}^* \ (1 - i)
\]

\[
| \ (i, \ \text{one}, \ \text{zero}) \leftarrow \text{zip3 intVecs zeroVecs]}
\]

\[
\text{enc} :: \text{Learnable} \ (\text{Vec Bool 8} \rightarrow \text{Vec R h})
\]

\[
\text{enc} = \text{do}
\]

\[
\text{mlp} \leftarrow \text{new ofMlp}
\]

\[
\text{return} \ \text{mlp} \ . \ \text{intEmbedding}
\]

\[
\text{agg} :: \text{Learnable} \ (\text{[Vec R h]} \rightarrow \text{Vec R h})
\]

\[
\text{-- Implementation-dependent}
\]

\[
\text{dec} :: \text{Learnable} \ (\text{Vec R h} \rightarrow \text{Vec R 8})
\]

\[
\text{dec} = \text{do}
\]

\[
\text{mlp} \leftarrow \text{new ofMlp}
\]

\[
\text{dense} \leftarrow \text{new ofLinearLayer} \ (\text{In h}) \ (\text{Out h})
\]

\[
\text{return} \ \text{sigmoid} \ . \ \text{dense} \ . \ \text{mlp}
\]

\[
\text{net} :: \text{Learnable} \ (\text{[Vec Bool 8]} \rightarrow \text{Vec R 8})
\]

\[
\text{net} = \text{dec} \ . \ \text{agg} \ . \ \text{map} \ \text{enc}
\]

**D Implementing binary tree aggregation for learnable commutative monoids**

More precisely, given a learnable commutative monoid operator \( <> \) and a function \( \text{toBalancedTree} \) which takes a list of elements and returns a balanced Tree whose leaves contain these elements, we aggregate in the following way:

\[
\text{data Tree a = Lf a | Nd Tree Tree}
\]

\[
\text{toListBalancedTree :: [a] -> Tree a}
\]

\[
\text{fold :: (a -> a -> a) -> Tree a -> a}
\]

\[
\text{fold f = \case}
\]

\[
\text{Nd l r} \rightarrow (f (\text{fold f l}) (\text{fold f r}))
\]

\[
\text{Lf m} \rightarrow m
\]
Learnable Commutative Monoids for Graph Neural Networks

aggregate :: Learnable ([LearnableMonoid] -> LearnableMonoid)
aggregate = fold (<>). toBalancedTree

E Implementing regularisation losses for learnable commutative monoids

Observe that, for any learnable binary operator

(<> :: Learnable (Vec R h -> Vec R h -> Vec R h)

aggregating over a tree of messages (of type Tree (Vec R h)), we can construct regularisation losses that penalise the operator for violating commutativity and associativity each time it is applied:

-- Computes getLossesAtNode at every node in the tree,
-- returning a list of the results.
accumLosses :: ((Tree (Vec R h)) -> [R]) -> (Tree (Vec R h)) -> [R]
accumLosses getLossesAtNode = \case
  Nd a b ->
    getLossesAtNode (Nd a b) :
    (accumLosses getLossesAtNode a ++ accumLosses getLossesAtNode b)
  _ -> 

commLoss :: (Tree (Vec R h)) -> R
commLoss = mean . accumLosses getLossesAtNode
  where getLossesAtNode = \case
    Nd a b -> [(a <> b) - (b <> a)]**2
    _ -> []

assocLoss :: (Tree (Vec R h)) -> R
assocLoss = mean . accumLosses getLossesAtNode
  where
    loss a b c = |((a <> b) <> c) - (a <> (b <> c))|**2
    getLossesAtNode = \case
      Nd (Nd a b) (Nd c d) ->
        [loss (aggregate a) (aggregate b) (aggregate c),
         loss (aggregate b) (aggregate c) (aggregate d)]
      Nd (Nd a b) (Lf c) ->
        [loss (aggregate a) (aggregate b) c]
      _ -> []

aggregateWithLoss :: Learnable ([LearnableMonoid] -> LearnableMonoid)
aggregateWithLoss xs = aggregate tree
  with extraLosses = [commLoss tree, assocLoss tree]
  where tree = toBalancedTree xs

F Training details for experiments

On every experiment, for each model, we performed 3 training runs with different seeds; for each run we used a validation set to choose the highest-performing checkpoint for evaluation.

2nd-minimum. We trained each aggregator with the Adam optimiser for 1,000 epochs, with batch size 32 and learning rate $1e^{-4}$.

PNA synthetic benchmark. We trained each aggregator for 1,000 epochs. To ensure convergence, 16-dimensional models were trained with a learning rate of $10^{-3}$ as in Corso et al. [2020], and 128-dimensional models were trained with a learning rate of $10^{-4}$. All other hyperparameters were as in Corso et al. [2020].

PNA real-world benchmark. All hyperparameters (including training time) are as in [Corso et al., 2020].
G Detailed results for the 2\textsuperscript{nd}-minimum benchmark

We present more detailed results for the 2\textsuperscript{nd}-minimum benchmark below:

- Table 2 contains in-distribution and out-of-distribution results for all aggregators tested.
- Figure 3 presents network efficiency against set size for all aggregators tested.
- Figure 4 presents the full results of the regularisation parameter sweep for Binary-GRU-Assoc.

As a side note, when training the non-regularised Binary-GRU aggregators, we observed that while associativity regularisation loss increased initially, it started decreasing as the GNN’s training accuracy began to plateau. This potentially hints at the model’s learning trajectory: one might hypothesise that the point at which the loss decreases is the point at which the model shifts from memorisation to learning a parsimonious algorithm that generalises.

<table>
<thead>
<tr>
<th>Type</th>
<th>Aggregator</th>
<th>ID accuracy $n \in [1, 16]$</th>
<th>OOD accuracy $n = 32$</th>
<th>OOD accuracy $n = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recurrent</td>
<td>GRU</td>
<td>0.996 ± 0.001</td>
<td>0.998 ± 0.001</td>
<td>0.912 ± 0.017</td>
</tr>
<tr>
<td>LCM</td>
<td>Binary-GRU-Assoc</td>
<td>0.997 ± 0.002</td>
<td>0.997 ± 0.002</td>
<td>0.822 ± 0.064</td>
</tr>
<tr>
<td>LCM</td>
<td>Binary-GRU</td>
<td>0.997 ± 0.001</td>
<td>0.992 ± 0.005</td>
<td>0.443 ± 0.122</td>
</tr>
<tr>
<td>Fixed</td>
<td>PNA</td>
<td>0.961 ± 0.003</td>
<td>0.794 ± 0.012</td>
<td>0.110 ± 0.027</td>
</tr>
<tr>
<td>Fixed</td>
<td>Max</td>
<td>0.901 ± 0.007</td>
<td>0.723 ± 0.025</td>
<td>0.069 ± 0.039</td>
</tr>
<tr>
<td>Fixed</td>
<td>Sum</td>
<td>0.845 ± 0.010</td>
<td>0.261 ± 0.020</td>
<td>0.045 ± 0.011</td>
</tr>
</tbody>
</table>

Table 2: Accuracy (the fraction of multisets at each size for which the 2\textsuperscript{nd}-minimum is correctly identified) for fixed, recurrent and LCM aggregators, along with the best-performing regularised LCM aggregator (Binary-GRU-Assoc with $\lambda = 10^0$).

![Figure 3: Efficiency (mean time per epoch on a GPU, over 5 epochs) for fixed (max, sum, PNA), recurrent (GRU), LCM (Binary-GRU) and regularised LCM (Binary-GRU-Assoc) aggregators. The shaded region is bounded above and below by the maximum and minimum values across all runs.](image)

H Detailed results and discussion for the PNA synthetic benchmark

Out-of-distribution performance. We present the out-of-distribution performance of our aggregators in Figure 5. Note that the MPNN (max) curve corresponds to the second-best aggregator tested out-of-distribution in [Corso et al., 2020], after PNA – this curve stops at graphs of sizes between 45 and 50 as this is the maximum graph size on which the aggregator was tested in the paper.

Observe that all learnable aggregators generalise as well as, or better than, the max-aggregator. Notably, while the Binary-GRU-Assoc aggregator underperforms in-distribution compared to Binary-GRU, it beats Binary-GRU out-of-distribution and performs competitively with GRU: indeed, the regularisation towards associativity has improved performance out-of-distribution at the cost of a slight decrease in performance in-distribution.
Learnable Commutative Monoids for Graph Neural Networks

Figure 4: Mean generalisation performance for fixed, recurrent and LCM aggregators, sweeping across regularisation rate $\lambda$ for Binary-GRU-Assoc.

Figure 5: Mean generalisation performance (multi-task log$_{10}$ of the ratio between the MSE loss for the GNN and the MSE loss for the baseline) for fixed, recurrent and LCM aggregators on the PNA multi-task benchmark.

Notice also that all learnable aggregators are more stable than PNA for very large graphs – in fact, the 128-dimensional PNA explodes for graph sizes above 75.

**Dimensionality and overfitting.** Finally, we take a look at the effects of high dimensionality on the performance of various aggregators.

For learnable aggregators, increasing dimensionality seems to help performance. We demonstrated that, if learnable aggregators operate over a latent space with a high enough dimension, they can beat individual fixed aggregators on tasks the fixed aggregators should be aligned to, and can even be competitive with PNA. Informal testing showed that the performance of learnable aggregators drops substantially if the dimensionality of these aggregators is reduced.

By contrast, for fixed aggregators, increasing dimensionality seems to harm performance: Corso et al. [2020] found that “even when [models with fixed aggregators] are given 30% more parameters than the [model using] PNA, they are qualitatively less capable of capturing the graph structure”. (And for this reason, we did not test models with fixed aggregators in the 128-dimensional setting.)

For PNA, the story is slightly more complex: while the 16-dimensional PNA performs well in-distribution (and, to some extent, out-of-distribution), this improvement in performance is small, especially when compared to PNA’s standard deviation. And notably, unlike the 16-dimensional PNA, the 128-dimensional PNA explodes out-of-distribution.

So it seems that, when increasing the dimensionality of the aggregator, fixed aggregators may have more of a tendency to overfit.

One possible hypothesis for this phenomenon comes from observing that, by Section 2.2,
in cases where the problem we’re attempting to solve aligns with the fixed aggregator we want to use, we can often learn a simple homomorphism from the fixed aggregator to our latent space, and

while homomorphisms from fixed aggregators are expressive enough in principle to model any commutative monoid, the required homomorphism is complex and doesn’t generalise out-of-distribution.

Note that, even for choices of fixed aggregator where some tasks align with the underlying monoid, the aggregator still doesn’t align perfectly with the combined ‘multitask benchmark monoid’ that we would need to learn to imitate in order to perform all tasks simultaneously. So, if we have the dimensionality to do so, our fixed aggregator may try to combine the existing monoids to approximate this multitask monoid in-distribution, in a way that does not generalise. In other words, it may be easier to get better performance by learning a very complex homomorphism from our fixed aggregator that works well in-distribution but struggles to extrapolate, than by learning a simple homomorphism from the fixed aggregator that ‘mostly works’.

Under this hypothesis, low-dimensional feature spaces provide an inductive bias towards learning simple homomorphisms that generalise out-of-distribution.
I  Reference table for code snippets

Throughout this work, we present code snippets in Haskell [Marlow et al., 2010], a statically typed, purely functional programming language.

As the fundamental idea behind this work – using algebraic structures as a means of abstraction in software development – was popularised by Haskell and its surrounding community, we observe that the ideas presented in this paper are most concisely stated through the lens of Haskell.

Furthermore, in the spirit of Olah [2015], we observe that there is a very close correspondence between the construction of neural networks and the construction of purely functional programs: indeed, we believe that strongly typed, purely functional languages like Haskell offer great potential for safe, succinct specification and training of neural networks.

For those unfamiliar with Haskell, we present the Haskell snippets featured in the main body of this work, alongside roughly equivalent implementations in Python.

<table>
<thead>
<tr>
<th>Haskell</th>
<th>Python</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>class CommutativeMonoid a =</code></td>
<td><code>class CommutativeMonoid(Protocol, Generic[A]):</code></td>
</tr>
<tr>
<td><code>    e :: a</code></td>
<td><code>    def id() -&gt; A:</code></td>
</tr>
<tr>
<td><code>    &lt;&gt; :: a -&gt; a -&gt; a</code></td>
<td><code>    ...</code></td>
</tr>
<tr>
<td>`{</td>
<td><code>@staticmethod</code></td>
</tr>
<tr>
<td></td>
<td>where commutative monoids M satisfy</td>
</tr>
<tr>
<td></td>
<td>x &lt;&gt; e == e</td>
</tr>
<tr>
<td></td>
<td>x &lt;&gt; y == y &lt;&gt; x</td>
</tr>
<tr>
<td></td>
<td>x &lt;&gt; (y &lt;&gt; z) == (x &lt;&gt; y) &lt;&gt; z</td>
</tr>
<tr>
<td></td>
<td>}</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>```</td>
<td><code>    &quot;&quot;&quot;</code></td>
</tr>
<tr>
<td></td>
<td>where commutative monoids M satisfy</td>
</tr>
<tr>
<td></td>
<td>M.plus(x, M.id()) == x</td>
</tr>
<tr>
<td></td>
<td>M.plus(x, y) == M.plus(y, x)</td>
</tr>
<tr>
<td></td>
<td>M.plus(x, M.plus(y, z)) == M.plus(M.plus(x, y), z)</td>
</tr>
</tbody>
</table>
```                                                                       |

Snippet 1: Defining the interface for commutative monoids. In Haskell, we do this by specifying a typeclass, such that a commutative monoid over some type \( T \) is defined by giving an instance of the typeclass for type \( T \). In Python, we do this by defining an abstract class, such that a commutative monoid over some type \( T \) is defined by specifying a child class of `CommutativeMonoid[T]`. 
Learnable Commutative Monoids for Graph Neural Networks

### Snippet 2: Defining the 2nd-minimum commutative monoid.

**Haskell**

```haskell
type M = (Int, Int)
instance CommutativeMonoid M where
  e = (infinity, infinity)
  (a1, a2) <> (b1, b2) = (c1, c2)
    where c1:c2:_= =
          sort [a1, a2, b1, b2]

secondMinimum :: [Int] -> Int
secondMinimum = dec . agg . map enc
    where enc x = (x, infinity)
          agg = reduce (<>)
          dec (_, x2) = x2
```

**Python**

```python
class SecondMinCM(CommutativeMonoid[Tuple[int, int]]):
    @staticmethod
    def plus(a: Tuple[int, int], b: Tuple[int, int]) -> Tuple[int, int]:
        c1, c2 = sorted([*a, *b])[:2]
        return (c1, c2)
    @staticmethod
    def id() -> Tuple[int, int]:
        return (INFINITY, INFINITY)

def secondMinimum(xs: List[int]) -> int:
    encoded = [(x, INFINITY) for x in xs]
    (_, x2) = SecondMinimumCM.reduce(encoded)
    return x2
```

### Snippet 3: Implementing a polymorphic fold over lists.

**Haskell**

```haskell
fold :: (a -> b -> b) -> b -> [a] -> b
fold f z [] = z
fold f z (x:xs) = f x (fold f z xs)
```

**Python**

```python
def fold(f: Callable[[B, A], B], z: B, xs: List[A]):
    accumulator = z
    for x in xs:
        accumulator = f(accumulator, x)
    return accumulator
```

**Snippet 4: Implementing an RNN as a fold over lists.** Note that, as mentioned in Snippet 3, the RNN as implemented in Haskell will consume its list of input states ‘in reverse’.

**Haskell**

```haskell
rnnCell :: Learnable
  (Vec R h1 -> Vec R h2 -> Vec R h2)
initialState :: Learnable (Vec R h2)

rnn :: Learnable
  (Vec R h1) -> Vec R h2
rnn = fold rnnCell initialState
```

**Python**

```python
rnnCell: Callable[
    [HiddenState, InputState], HiddenState
]
initialState: HiddenState

def rnn(inputs: List[InputState]) -> HiddenState:
    return fold(rnnCell, initialState, inputs)
```
### Snippet 5: Defining a learnable commutative monoid over hidden states. We assume we have access to a learnable binary operation \( \text{binOp} \in (\mathbb{R}^h \times \mathbb{R}^h) \to \mathbb{R}^h \) and a learnable vector \( \text{identity} \in \mathbb{R}^h \).

```haskell
binOp :: Learnable (Vec R h -> Vec R h -> Vec R h)
identity :: Learnable (Vec R h)

type HiddenState = Vec R h
instance (CommutativeMonoid HiddenState) where
e = identity; <> = binOp

aggregate :: Learnable ([HiddenState] -> HiddenState)
aggregate = reduce (<>)
```

```python
class LearnableCommutativeMonoid(
    CommutativeMonoid[HiddenState]
):
    @staticmethod
    def plus(a: HiddenState, b: HiddenState) -> HiddenState:
        return binOp(a, b)
    @staticmethod
    def id() -> HiddenState:
        return identity

def aggregate(xs: List[HiddenState]) -> HiddenState:
    return LearnableCommutativeMonoid.reduce(xs)
```

### Snippet 6: Defining the Binary-GRU operator. In Haskell, we present this via a (hypothetical) monadic API for defining neural networks; in Python, we define a class in the style of TensorFlow / PyTorch modules.

```haskell
binaryGRU :: Learnable (Vec R h -> Vec R h -> Vec R h)
binaryGRU v1 v2 = do
g <- new (gruCell (InputDim h) (HiddenDim h))
return (g v1 v2 + g v2 v1) / 2
```

```python
class BinaryGRU:
    def __init__(self, h):
        self.gruCell: Callable[[HiddenState, HiddenState], HiddenState] = GRUCell()

def __call__(self, x: HiddenState, y: HiddenState):
    return (self.gruCell(x, y) + self.gruCell(y, x)) / 2
```