

---

# Building LEGO Using Deep Generative Models of Graphs

---

**Rylee Thompson**  
University of Guelph  
rylee@uoguelph.ca

**Elahe Ghalebi, Terrance DeVries and Graham W. Taylor**  
University of Guelph  
Vector Institute  
{eghalebi,terrance,gwtaylor@uoguelph.ca}

## Abstract

Generative models are now used to create a variety of high-quality digital artifacts. Yet their use in designing physical objects has received far less attention. In this paper, we advocate for the construction toy, LEGO, as a platform for developing generative models of sequential assembly. We develop a generative model based on graph-structured neural networks that can learn from human-built structures and produce visually compelling designs.

## 1 Introduction

Sequential assembly is the process of creating a desired form by connecting a series of geometric primitives. For example, furniture may be constructed from wooden segments, walls from individual bricks, and quilts from fabric patches. When the individual pieces are few and simple with a limited number of interlocking structures, this facilitates assembly by humans or robots [1]. Applications of automated assembly include modular packaging, or at a larger scale, pre-fabricated buildings.

Generative models can now create high-quality and diverse digital artifacts, particularly in the domain of images [2] and text [3]. Generative models are also being applied to physical design, such as computational chemistry and materials but they are less prevalent in architecture, landscape design, or manufacturing. Motivated by the opportunity of generative design of our physical environment, in this paper we explore the use of generative models to modular physical design.

Advances in generative models in areas such as images and text is in part due to the availability of data and ease of experimentation. Other areas of machine learning, such as reinforcement learning, have benefited from the availability of simulators. Simulators are conducive to experimentation supporting algorithm development and, at the same time, entertaining and familiar (c.f. the Arcade Learning Environment [4]). In the context of physical design, specifically the sequential assembly of modular structures, we believe that there is an analogue: the plastic construction toy, LEGO. Structures built from LEGO have a number of interesting properties: (i) they are complex enough to approach real-world design; ii) they are familiar and fun enough to attract interest from the generative models community; and iii) there is a huge amount of official and user-created structures available as a source of training data.

Learning to sequentially assemble LEGO structures based on human-generated examples is a sequential decision making problem. It can be approached through a number of formalisms familiar to the ML community, including reinforcement learning [5], imitation learning [6], evolutionary design [7, 8], and Bayesian optimization [9]. We are unaware of the use of deep generative models for LEGO. However, generative graph models (GGMs), an emerging sub-field of graph representation learning [10], are particularly well suited to this domain.

LEGO structures can be represented as a graph, with nodes representing bricks and edges representing connections between bricks. Nodes can hold information, such as brick type, orientation, and colour.

Edges can hold information such as how the bricks connect to one another. Graph-based representations of LEGO have been used as a search space for evolutionary design [7], and these should be learnable by GGMs such as Deep Generative Models of Graphs (DGMG) [11] or GraphRNN [12].

This paper explores GGMs for designing LEGO structures that are “human-like” in their build quality. It makes the following contributions:

- We make a case for the use of deep generative models in sequential assembly. We propose LEGO as a platform for experimentation that balances accessibility and realism.
- We propose a model based on [11] incorporating physical constraints designed for and evaluated on LEGO but general enough for other modular assembly applications.
- We perform an extensive evaluation of various metrics proposed in the generative modeling of images, using a novel permutation analysis.

## 2 Background

The problem we consider in this paper is the creation of novel “human-like” assemblies expressed as graphs. Here, we briefly review the works most relevant to our problem domain and methodology.

### 2.1 LEGO modelling

LEGO has received much attention within the computer graphics community. One of the most frequently studied LEGO-related problems is finding a constructable and stable layout of bricks for a target object, so-called “legolization” [13, 14, 15]. Legolization techniques [16] typically generate an assembly for a given 3D design. In the generative setting, we aim to create *novel* builds which may be conditioned on context, for example a textual description or object class, but not a specific target. This makes evaluation much more challenging because we are not measuring the discrepancy between a target build and the model’s output. Instead, we must assess the generative model based on the *quality* and *diversity* of its builds, as well as the *consistency* with any conditioning information.

Another community that has considered LEGO as a problem domain is evolutionary computing. These works are concerned with creative outputs that respect certain structural and aesthetic properties. Relevant to our work, several papers design new representations for LEGO that can be optimized. Devert et al. [8] introduce a representation of construction plans for LEGO-like structures that has several desirable properties such as re-usability and modularity. Peysakhov and Regli [7] utilize a graph representation and Genetic Algorithms to create LEGO structures.

In computer vision, Jones et al. [17] recover a 3D LEGO model of an assembly from a video of it being assembled. Like this work, we model LEGO designs as a sequential process. In contrast, we work with a more direct representation of structures and do not attempt to learn through vision. Most relevant to our work is [9], which applies Bayesian optimization to the sequential assembly of LEGO structures with generation conditioned on a high-level description or a specific training example. This is extremely similar to the problem we explore.

### 2.2 Generative graph models

Traditional graph generative models are based on random graph theory that formalizes a simple stochastic generation process and have well-understood mathematical properties. For example, Kronecker graphs [18] build up graphs from a small base generator graph by iteratively applying the Kronecker product. The resulting graphs are, by construction, very self-similar due to their simplicity and therefore only capture a few graph statistics such as degree distribution. These graph models remain highly-constrained in terms of the graph structures they can represent.

Recent graph generative models use neural networks to capture the distribution over random graphs. The quality of graph generative modeling depends on learning the distribution given a collection of relevant training graphs. A number of deep generative models are based on variational autoencoders (VAEs). For instance, the GraphVAE algorithm [19] uses a VAE to learn a matrix of edge probabilities between every possible pair in a graph. An extension of GraphVAE, the Regularized Graph VAE, uses validity constraints to regularize the output distribution of the decoder. Although these models

have greater capacity to learn structural information from data than the traditional models, capturing a set of specific global properties such as graph connectivity, and node compatibility is challenging.

A few models generate a graph sequentially by choosing nodes and edges step-by-step. For instance, Dai et al. [20] use recurrent neural networks (RNNs) to model graph formation, and to capture semantic validity, attribute grammars are applied on the output distribution. Deep Generative Models of Graphs (DGMG) [11] employs a RNN to make a sequence of decisions: whether to add a new node, which node to add, whether to add an edge, and which destination node to connect to the new node. DGMG assumes the probability of a graph is the sum over all possible node permutations. GraphRNN [12], uses a recurrent neural network to obtain the distribution over the  $i$ th row of the lower triangle of the adjacency matrix, conditioned on the previous rows. Nodes and edges are generated with a graph-level RNN which adds a new node and an edge-level RNN which generates a connection between this new node and the nodes of the graph.

Domain-specific methods such as molecular graphs have recently attracted a lot of attention in drug discovery and material science [21, 22, 23]. We believe that mixing domain-specific knowledge and machine learning in physical design will yield similar successes.

### 2.3 Evaluation of graph generative models

In general, evaluating generative models of graphs is challenging due to their complex structural dependencies [24]. Recent relevant graph generative models [12, 11] evaluate the quality of generation by considering a set of graph statistics: the degree distribution, clustering coefficient and number of occurrence of all orbits with  $n$  nodes. To estimate the similarity between generated graphs and the ground truth, the maximum mean discrepancy (MMD) is calculated for each of these statistics using the Wasserstein distance. MMD determines whether two sets of samples from the distribution  $p$  and  $q$  are derived from the same distribution. However, MMD works on fairly small graphs and for medium-sized graphs such as those we encounter in LEGO, the computation of MMD is very slow due to the Gaussian EMD kernel.

Another metric that measures the distance between generated and actual samples is the Fréchet Inception Distance (FID) [25]. The FID was originally introduced to measure the quality of generated image samples. FID along with other common GAN evaluation metrics use the pretrained image classifier Inception v3 [26] to obtain feature representations of images, which enables a more straightforward comparison between generated and reference distributions [27, 25, 28, 29, 30]. Liu et al. [31] adapt FID to the graph domain by replacing Inception v3 with a Graph Isomorphism Network (GIN) classifier [32]. [31] also introduces GIN classifier accuracy, which is simply the percentage of class conditional samples that can be successfully recognized by the GIN classifier.

## 3 Methodology

Here we present the Deep Generative Model of LEGO Graphs (DGMLG), a sequential generative model of LEGO structures. We first define a representation that allows us to convert a LEGO structure to a graph. Then, we define the graph generation process for LEGO and show how it is trained. Finally, we adapt a number of generative evaluation metrics to the graph setting.

### 3.1 LEGO structure graph representation

We utilize a graph representation very similar to the one proposed by Peysakhov et al. [7], with small changes to improve compatibility with our generative model. LEGO structures that contain standard symmetrical pieces such as bricks or plates can be represented by this graph representation, however structures with more complex pieces such as wheels or axles cannot. A LEGO structure that meets this criteria may be represented by a directed and labeled graph  $G$ , where the nodes represent LEGO bricks and the edges represent connections between bricks. In addition, node labels encode the orientation and size of each brick, and edge labels specify how two bricks are connected by encoding a two-element offset  $(x,y)$  between them. A directed edge from node  $u$  to node  $v$  indicates that brick  $u$  provides studs to brick  $v$  (i.e.  $v$  sits on  $u$ ). Although flexible, our graph representation does not ensure that all graphs are physically realizable as LEGO structures. It is possible to form a graph where any two bricks occupy the same physical space, or to overconstrain a brick to be in multiple

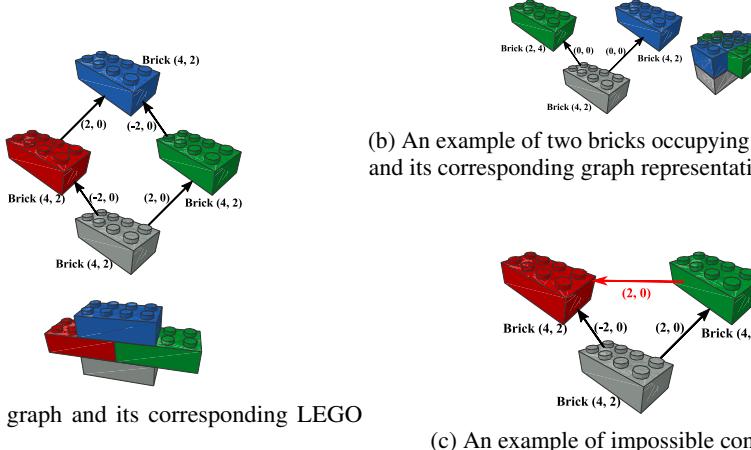


Figure 1: Valid and invalid LEGO builds and their graph representations.

locations at once. Invalid graphs are not physically realizable because of one or more of the discussed issues. An example of these issues along with a valid LEGO graph is shown in Figure 1.

An idiosyncrasy with our graph-based representation of LEGO structures is that some edges are redundant. An edge connecting  $u$  to  $v$  may imply that  $u$  is also connected to another node  $k$ . We refer to this as an “implied edge”, and an example can be found in Figure 1a; if any edge is removed the underlying LEGO structure remains unchanged, and the removed edge becomes implied. Generated graphs are considered to be valid regardless of any missing implied edges.

### 3.2 Sequential graph generation

We expand upon the DGMG graph generation process [11] to create a model that is capable of making decisions regarding typed and directed edges. Wherever possible we modelled our extensions after the sequence of actions a human might follow when creating a LEGO structure. DGMG uses a sequential process to generate nodes one at a time and connect them to the existing partial graph. At each iteration, it determines whether a new node of a particular type should be added, or the generation process should terminate. If a new node is added, DGMG then chooses whether to add an edge to this node or not. A node in the existing graph is then chosen as a destination for a newly added edge. This edge generation process is repeated until the decision is made to stop connecting new edges, in which case the process continues from the node generation step. The entire process is repeated until the node generation step makes the decision to terminate. DGMG uses graph-structured neural networks (graph nets) to complement structure building decisions by using message passing and graph propagation to yield node and edge representations [10].

**Graph propagation** For a graph  $G = (V, E)$ , we associate a node embedding vector  $\mathbf{h}_v \in \mathbb{R}^H$  for all  $v \in V$ , and an edge embedding vector  $\mathbf{s}_{u,v} \in \mathbb{R}^S$  for all  $e \in E$ . The set of all node embeddings in  $G$  is denoted by  $\mathbf{h}_V = \{h_1, h_2, \dots, h_{|V|}\}$ . The embeddings are initialized using corresponding node and edge types, and are used in the graph propagation process to aggregate information across the graph. As in [11], the function  $f_e$  that computes the message vector  $a_v$  from  $u$  to  $v$  is a fully-connected network, and the node-update function  $f_n$  is a gated recurrent unit (GRU) cell:

$$\mathbf{a}_v = \sum_{u:(u,v) \in E} f_e(\mathbf{h}_u, \mathbf{h}_v, \mathbf{s}_{u,v}) \quad \forall v \in V, \quad (1) \quad \mathbf{h}'_v = f_n(\mathbf{a}_v, \mathbf{h}_v) \quad \forall v \in V. \quad (2)$$

Li et al. [11] suggest using a different set of parameters for  $f_e$  and  $f_n$  for each round of graph propagation to increase model capacity, and we use this setting. The function  $\text{prop}^{(T)}(\mathbf{h}_V, G)$  denotes  $T$  rounds of graph propagation, and returns a set of updated node embeddings  $\mathbf{h}_V^{(T)}$ . This is equivalent to repeating Eq. 1 and Eq. 2  $T$  times. We set  $T$  to 2 throughout all experiments:

$$\mathbf{h}_V^{(T)} = \text{prop}^{(T)}(\mathbf{h}_V, G). \quad (3)$$

The new node vectors  $\mathbf{h}_V^{(T)}$  are carried through the decision modules below, making them recurrent across these decisions. The node vectors are also recurrent across the graph propagation steps.

To obtain graph embeddings, we first map the node representations to a higher dimension using a fully connected network  $f_m$ :  $\mathbf{h}_v^G = f_m(\mathbf{h}_v)$ . We then apply a gated sum over all nodes to obtain a single vector  $\mathbf{h}_G$ . The function  $g_m$  is a fully collected network which maps each node embedding to a single value, and determines  $\mathbf{g}_v^G$ , the importance of each node, for use in the gated sum:

$$\mathbf{h}_G = \sum_{v \in V} \mathbf{g}_v^G \odot \mathbf{h}_v^G, \quad (4) \quad \mathbf{g}_v^G = \sigma(g_m(\mathbf{h}_v)). \quad (5)$$

**Add node module** In this module, we produce the probability of adding a node of each type and the probability of terminating the process using an existing graph  $G$  and its corresponding node embeddings  $\mathbf{h}_V$ . We first use Eq. 3 to run  $T$  rounds of graph propagation to obtain updated node vectors  $\mathbf{h}_V^{(T)}$ , which are then used to create a graph representation vector as in Eq. 4. The graph embedding vector is then passed through a standard MLP  $f_{an}$  with softmax activation to obtain the probability associated with each possible action:

$$f_{addnode}(G) = \text{softmax}(f_{an}(\mathbf{h}_G)). \quad (6)$$

**Add edge module** In the add edge module, we take an existing graph  $G$ , a newly added node  $v$ , and compute probabilities for three possible outcomes: not adding an edge to  $v$ , adding an incoming edge to  $v$ , or adding an outgoing edge from  $v$ . These values are determined by passing the graph representation vector  $\mathbf{h}_G$  and the new node embedding  $\mathbf{h}_v$  through another MLP  $f_{ae}$  with softmax out:

$$f_{addedge}(G, v) = \text{softmax}(f_{ae}(\mathbf{h}_G, \mathbf{h}_v)). \quad (7)$$

**Choose destination module** This module computes a score  $x_u$  for every  $u \in V \setminus \{v\}$  using an MLP  $f_s$ , and normalizes the vector  $\mathbf{x}$  through a softmax to obtain the probability of connecting the new node  $v$  to  $u$  with direction determined by the add edge module:

$$x_u = f_s(\mathbf{h}_u^{(T)}, \mathbf{h}_v), \quad \forall u \in V \setminus v, \quad (8) \quad f_{dest}(G, v) = \text{softmax}(\mathbf{x}). \quad (9)$$

This module may handle typed or directed edges by making  $x_u$  a vector of scores the same size of edge types [11]. However, we separate these decisions to avoid combinatorial explosion in the number of outputs. The arguments to  $f_s$  in Eq. 8 are rearranged such that the source node and destination node for a newly added edge are always given to the MLP in the same order.

**Choose edge type module** This module determines the edge type of the newly added edge between nodes  $u$  and  $v$  by choosing the  $(x,y)$  offset between bricks  $u$  and  $v$ . We treat this decision as two independent events, and use two separate MLPs  $f_{ex}$  and  $f_{ey}$  to determine the  $x$  and  $y$  offset, respectively:

$$x = \text{softmax}(f_{ex}(\mathbf{h}_u^{(T)}, \mathbf{h}_v)), \quad (10) \quad y = \text{softmax}(f_{ey}(\mathbf{h}_u^{(T)}, \mathbf{h}_v)). \quad (11)$$

We experiment with two ways to pose this problem: i) as an ordinal regression problem where each output has a clear rank [33], or ii) by treating the offsets as categories, as in classification. The former is described in Appendix B, while the latter is shown in Eq. 10 and Eq. 11. In addition, this module introduces a new type of invalid graph; although two bricks may be connected in the graph, it is possible that we generate an offset such that it is physically impossible for them to connect.

In summary, to generate a graph for the LEGO graph representation: (1) choose whether to add a LEGO brick with a given size and orientation, or terminate the assembly process, (2) choose whether to attach this new brick into the LEGO structure, and if so whether it should connect on top of or underneath of a pre-existing brick, (3) if the newly added brick is to be connected, decide which brick it should connect to, otherwise go back to step (1), and (4) choose the relative offset of these bricks from one another, and restart from step (2). We perform class-conditioned generation by adding a “one-hot” class-conditioning vector  $\mathbf{c}$  to the input of each structure building MLP described above.

### 3.3 Training and evaluation

Given a set of training graphs, we train our model to maximize the joint log-likelihood  $\mathbb{E}_{P_{data}(G)} \log p(G)$  using categorical cross-entropy. For each LEGO structure, a graph generating sequence is created knowing that the ordering is analogous to an assembly that a human would make. The likelihood for each individual step is computed using the output modules described in § 3.2.

To evaluate the generated structures, we adapt several popular and thoroughly tested GAN evaluation metrics to the generative graph domain. Most are designed for images and make use of a pretrained

Inception v3 network as a fixed feature extractor. As in [31], we replace Inception v3 with a Graph Isomorphism Network (GIN) [32] to obtain feature representations of graphs<sup>1</sup>.

## 4 Experiments

We execute two main experiments. The first demonstrates the effectiveness and value of the § A metrics for generative LEGO assembly. The second evaluates our model quantitatively with these metrics, and qualitatively by visualizing generated LEGO structures. In both experiments we use the LEGO dataset from Kim et al. [9]. This dataset consists of 12 classes and a total of 360 LEGO structures built using 2x4 LEGO bricks, with each structure created by one of twelve human subjects. More information regarding the dataset is provided in Appendix C.

### 4.1 Permutation analysis

This experiment demonstrates the value of our proposed metrics in the evaluation of generative graph models. We begin by duplicating the LEGO dataset and designating the original as an unchanging reference. Throughout the experiment, we apply several stacking permutations to each graph in the copy. At each iteration every graph in this copy is randomly permuted, and the changes are carried over to the next iteration. These permutations cause the distribution of the permuted dataset to slowly drift from the reference distribution. We expect the drift to be detected by all evaluation metrics.

The permutations applied are simple: with equal probability, a randomly selected node and all associated edges are deleted from the graph, or a node with a randomly selected brick type is added to the graph. To ensure some similarity to the reference dataset at a very high level, node deletions that result in a disjoint graph are prohibited. To complete the addition of a node, we form a connection to a random pre-existing node, and assign a random edge direction and type under the constraint that the resultant graph must represent a valid LEGO structure. Once a valid edge is created, all implied edges are determined and subsequently added to the graph. The constraints ensure that the permuted dataset will resemble randomly assembled LEGO structures in both the LEGO and graphical representations, and allows for the results to be evaluated both qualitatively and quantitatively. We perform 500 accumulating permutations to each graph in the copy.

We can see from Figure 2 that all metrics included in the experiment capture the distribution shift relatively well. It is expected that each LEGO structure will reach a point where it appears that it was randomly assembled, and the metrics should then begin to asymptote. This is the case for all metrics with the exception of FD and KD; FD appears to asymptote, while KD is almost exponential, indicating a flaw may be present in the metric. GIN accuracy, density, and coverage decline extremely quickly and appear to be sensitive to any sort of distribution shift. With the exception of KD and recall, all metrics are fairly smooth with little noise. We include degree MMD from You et al. [12] in the experiment, and it is also capable of capturing the distribution shift.

### 4.2 Generation

Next we train our generative model on the same dataset considered in the permutation analysis. In addition, we employ the Bayesian optimization sequential LEGO assembly method from [9] as a baseline for our generative experiments. That method employs a voxel representation of LEGO structures, and restricts the decision space to prevent the creation of invalid assemblies. More information regarding the baseline is available in Appendix D.

First, we compare two methods for configuring the edge type module: using a thermometer encoding as motivated by ordinal regression [33] where the integer nature of the offsets is explicit, and as a softmax where each offset is treated as a category and thus maximally different. Surprisingly, the latter seems to provide a small increase to stability during training and provides slightly better results across the majority of metrics as shown in Figure 3.

We compare our model’s performance to the baseline using the GIN-based metrics proposed in Section A. We fix the reference dataset to be the entirety of the dataset, and generate 200 samples

---

<sup>1</sup>When we generalize a method beyond Inception v3, we will simplify its acronym. For example FID →FD. To avoid introducing another acronym, in § 4 we use the simplified acronym because our embeddings are unambiguously GIN.

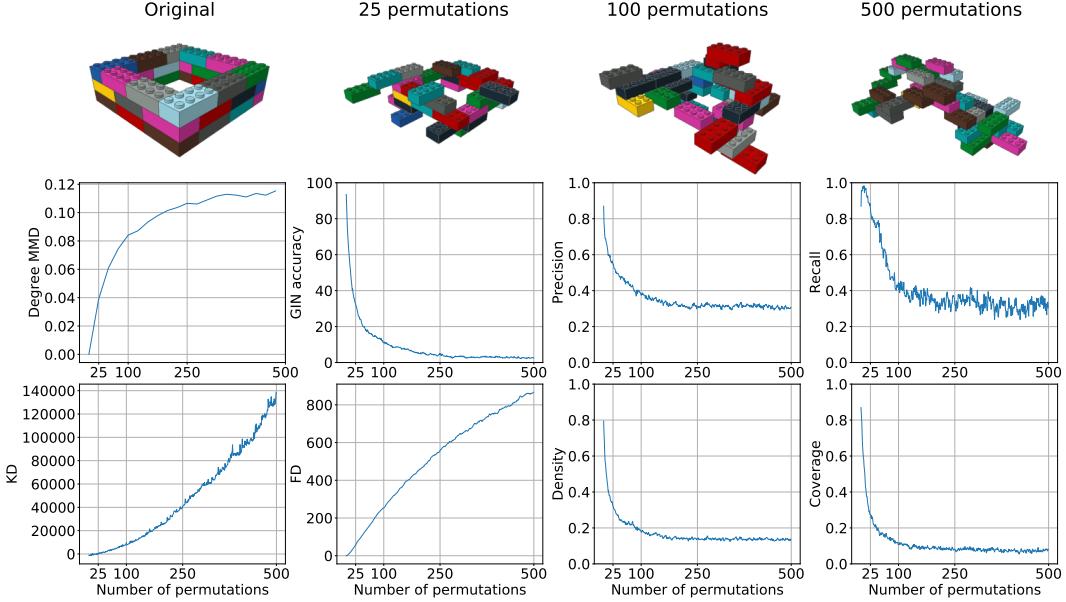


Figure 2: An example of a LEGO structure as it undergoes several permutations (top), and plots of various metrics as stacking permutations are applied (bottom). All GIN-based metrics are calculated at every iteration while degree MMD is calculated every 25 iterations due to its computational cost.

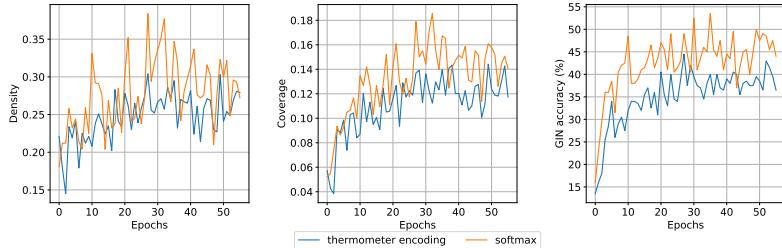


Figure 3: A comparison of density, coverage, and GIN accuracy metrics for two variants of DGMLG, which employ different methods for determining edge types.

with each method to prevent any discrepancy caused by differing sizes. We use the harmonic mean of density and coverage to identify the model with the highest performance through all epochs. We have found FD and KD to be relatively noisy with only 200 generated samples, and similar to the unified metrics, the harmonic mean of D&C provides a balance between sample quality and diversity.

We report all proposed metrics in Table 1 for two different versions of this model: the unrestricted variant (DGMLG), and one where we prevent any decisions that would result in an invalid structure (DGMLG-Re). Both versions of DGMLG perform significantly better than the baseline across most GIN-based metrics. An interesting result is that DGMLG outperforms DGMLG-Re. We expected that limiting the generated graphs to physically realizable LEGO structures would improve performance due to increased resemblance to the training set. This may be a result of the difference in average size of the generated graphs; DGMLG averages 52 nodes, DGMLG-Re averages 78 nodes, and the training set averages 56 nodes. It is unclear whether more training, or reducing the maximum size of generated graphs will improve the performance of DGMLG-Re relative to DGMLG.

We include visualizations of the LEGO structures generated by DGMLG, along with each sample’s nearest neighbour in the training set to show that the model is not simply generating copies of training examples in Figure 4. The generated structures are novel yet share common features with the corresponding class, indicating that DGMLG is capable of understanding and replicating patterns found in the dataset. For example, the generated structures in Figure 4a and Figure 4d extend patterns to create larger structures than what are found in the dataset. Also, the structure in Figure 4c shows

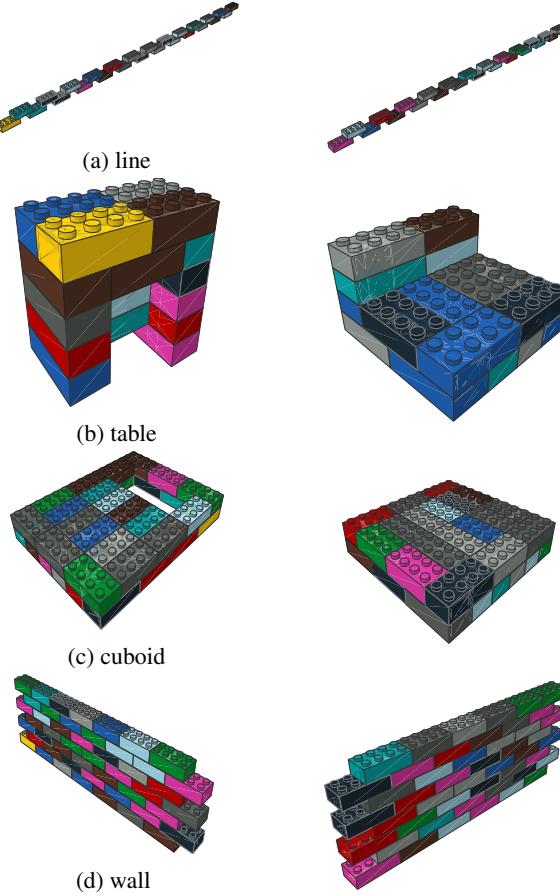


Figure 4: High quality LEGO structures generated by DGMLG and the class each generation was conditioned on (left), and the corresponding nearest neighbour by GIN graph embedding (right). The colour of each brick is randomly selected for aesthetics and not included anywhere in the graph generation process.

Table 1: Performance evaluation of DGMLG vs. the Bayesian optimization baseline over the proposed GIN-based metrics. The  $\uparrow/\downarrow$ symbols indicate that higher/lower is better, respectively. GIN Acc = GIN classifier accuracy, P = Precision, R = Recall, D = Density, C = Coverage.

	FD $\downarrow$	GIN Acc $\uparrow$	P $\uparrow$	R $\uparrow$	D $\uparrow$	C $\uparrow$	KD $\downarrow$	% novel $\uparrow$	% valid $\uparrow$
DGMLG	<b>150</b>	<b>60.5</b>	<b>0.62</b>	<b>0.92</b>	<b>0.48</b>	<b>0.23</b>	<b>2054</b>	89.5	25
DGMLG-Re	334	50.5	0.59	0.88	0.44	0.21	2.4e4	87	<b>100</b>
Bayesian optimization [9]	345	24.5	0.47	0.70	0.23	0.078	5297	<b>93.5</b>	<b>100</b>

that the model has learned that this particular class should be perfectly flat without any overhanging bricks. We estimate that these lie in about the top 5% of the generated structures based on our assessment of visual attractiveness.

## 5 Conclusion

We demonstrated that graph generative models can be readily used in sequential assembly of physical structures with visually satisfying results. We also showed the value of adapting several common evaluation metrics popularized by GANs to the generative graph domain. For future real-world physical design, it is vital to determine the stability of generated structures to ensure safety. We intend to pursue a larger-scale dataset with a wider variety of brick types and LEGO structures.

## References

- [1] Yinan Zhang and Devin Balkcom. “Interlocking Block Assembly”. In: *Algorithmic Foundations of Robotics XIII*. Springer International Publishing, 2020, pp. 709–726.
- [2] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.
- [3] Tom B Brown et al. “Language models are few-shot learners”. In: *arXiv preprint arXiv:2005.14165* (2020).
- [4] Marc G Bellemare et al. “The Arcade Learning Environment: An evaluation platform for general agents”. In: *J. Artif. Intell. Res.* 47 (2013), pp. 253–279.
- [5] Sai Krishna Gottipati et al. “Learning to Navigate in Synthetically Accessible Chemical Space Using Reinforcement Learning”. In: *Proceedings of the International Conference on Machine Learning* 1 (2020).
- [6] Y Mollard et al. “Robot programming from demonstration, feedback and transfer”. In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Sept. 2015, pp. 1825–1831.
- [7] Maxim Peysakhov and William C Regli. “Using assembly representations to enable evolutionary design of Lego structures”. In: *Artif. Intell. Eng. Des. Anal. Manuf.* 17.2 (May 2003), pp. 155–168.
- [8] Alexandre Devert, Nicolas Bredeche, and Marc Schoenauer. “Blindbuilder: A New Encoding to Evolve Lego-Like Structures”. In: *Genetic Programming*. Springer Berlin Heidelberg, 2006, pp. 61–72.
- [9] Jungtaek Kim et al. “Combinatorial 3D Shape Generation via Sequential Assembly”. In: (Apr. 2020). arXiv: 2004.07414 [cs.CV].
- [10] William L Hamilton, Rex Ying, and Jure Leskovec. “Representation Learning on Graphs: Methods and Applications”. In: *IEEE Data Engineering Bulletin* 40.3 (2017), pp. 52–74.
- [11] Yujia Li et al. “Learning Deep Generative Models of Graphs”. In: *International Conference on Learning Representations (ICLR) Workshop Track*. 2018.
- [12] Jiaxuan You et al. “GraphRNN: Generating Realistic Graphs with Deep Auto-regressive Models”. In: *International Conference on Machine Learning (ICML)*. 2018.
- [13] Romain Testuz, Yuliy Schwartzburg, and Mark Pauly. “Automatic Generation of Constructable Brick Sculptures”. In: *Eurographics 2013 - Short Papers*. Ed. by M - A Otaduy and O Sorkine. The Eurographics Association, 2013.
- [14] Sheng-Jie Luo et al. “Legolization: optimizing LEGO designs”. In: *ACM Trans. Graph.* 34.6 (Oct. 2015), pp. 1–12.
- [15] J Zhou, X Chen, and Y Xu. “Automatic Generation of Vivid LEGO Architectural Sculptures”. In: *Comput. Graph. Forum* 34 (Feb. 2019), 104:1.
- [16] Jae Woo Kim, Kyung Kyu Kang, and Ji Hyoung Lee. “Survey on Automated LEGO Assembly Construction”. en. In: *WSCG 2014: Poster Papers Proceedings: 22nd International Conference in Central European Computer Graphics, Visualization and Computer Vision in co-operation with EUROGRAPHICS Association*. 2014.
- [17] J Jones, G D Hager, and S Khudanpur. “Toward Computer Vision Systems That Understand Real-World Assembly Processes”. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2019, pp. 426–434.
- [18] Jure Leskovec et al. “Kronecker graphs: an approach to modeling networks.” In: *Journal of Machine Learning Research* 11.2 (2010).
- [19] Martin Simonovsky and Nikos Komodakis. “GraphVAE: Towards generation of small graphs using variational autoencoders”. In: *International Conference on Artificial Neural Networks*. Springer. 2018, pp. 412–422.
- [20] Hanjun Dai et al. “Syntax-directed variational autoencoder for structured data”. In: *arXiv preprint arXiv:1802.08786* (2018).
- [21] Nicola De Cao and Thomas Kipf. “MolGAN: An implicit generative model for small molecular graphs”. In: *arXiv preprint arXiv:1805.11973* (2018).
- [22] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. “Junction tree variational autoencoder for molecular graph generation”. In: *arXiv preprint arXiv:1802.04364* (2018).

- [23] Yibo Li, Liangren Zhang, and Zhenming Liu. “Multi-objective de novo drug design with conditional graph generative model”. In: *Journal of cheminformatics* 10.1 (2018), p. 33.
- [24] Lucas Theis, Aäron van den Oord, and Matthias Bethge. “A note on the evaluation of generative models”. In: *arXiv preprint arXiv:1511.01844* (2015).
- [25] Martin Heusel et al. “GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium”. In: *CoRR* abs/1706.08500 (2017). arXiv: 1706.08500. URL: <http://arxiv.org/abs/1706.08500>.
- [26] Christian Szegedy et al. “Rethinking the Inception Architecture for Computer Vision”. In: *CoRR* abs/1512.00567 (2015). arXiv: 1512.00567. URL: <http://arxiv.org/abs/1512.00567>.
- [27] Muhammad Ferjad Naeem et al. *Reliable Fidelity and Diversity Metrics for Generative Models*. 2020. arXiv: 2002.09797 [cs.CV].
- [28] Tim Salimans et al. “Improved Techniques for Training GANs”. In: *CoRR* abs/1606.03498 (2016). arXiv: 1606.03498. URL: <http://arxiv.org/abs/1606.03498>.
- [29] Tuomas Kynkänniemi et al. “Improved precision and recall metric for assessing generative models”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 3927–3936.
- [30] Mikołaj Bińkowski et al. *Demystifying MMD GANs*. 2018. arXiv: 1801.01401 [stat.ML].
- [31] Chia-Cheng Liu, Harris Chan, and Kevin Luk. “Auto-regressive Graph Generation Modeling with Improved Evaluation Methods”. In: *33rd Conference on Neural Information Processing Systems*. Vancouver, Canada, 2019. URL: <https://grlearning.github.io/papers/77.pdf>.
- [32] Keyulu Xu et al. “How Powerful are Graph Neural Networks?” In: *CoRR* abs/1810.00826 (2018). arXiv: 1810.00826. URL: <http://arxiv.org/abs/1810.00826>.
- [33] Jianlin Cheng. “A neural network approach to ordinal regression”. In: *CoRR* abs/0704.1028 (2007). arXiv: 0704.1028. URL: <http://arxiv.org/abs/0704.1028>.
- [34] Konstantin Shmelkov, Cordelia Schmid, and Karteeek Alahari. “How good is my GAN?” In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 213–229.

## Appendices

### A Evaluation metrics

In this section, we describe the specific metrics that we adapt to the graph setting and apply in § 4.

**GIN Accuracy** A simple measure of generation quality is to determine what percentage of class conditional samples can be successfully recognized by a pretrained classifier. If the generated samples share similar properties with the target class, then the labels predicted by the classifier should match the conditioning labels. The classification accuracy metric was introduced in the GAN literature as GAN-test [34], and has been adapted for evaluating graph generations through the use of a pretrained GIN classifier in [31]. GIN accuracy does not measure sample diversity. As such, a model that only produces a single realistic sample may still achieve a good score.

**Fréchet Distance (FD)** Fréchet Distance may be used to measure the distance between two probability distributions, and is commonly used for evaluating the quality of generative models. Generated samples and reference samples are embedded into some task relevant feature space (e.g. Inception v3 for images), and a multivariate Gaussian is fit to each set of features. The Fréchet Distance between the two distributions can then be calculated as  $D = \|\mu - \hat{\mu}\|_2^2 + \text{Tr}(\Sigma + \hat{\Sigma} - 2(\Sigma\hat{\Sigma})^{1/2})$ , where  $\mu$  and  $\Sigma$  are the mean and covariance of the reference distribution, and  $\hat{\mu}$  and  $\hat{\Sigma}$  are the mean and covariance of the generated distribution. FD jointly considers both generation realism and diversity, and as such is a good overall measure of model performance.

**Kernel Distance (KD)** Kernel Inception Distance [30] was introduced as an alternative to FID that brought with it several advantages: KID does not assume a parametric form of the distribution of the embedding space, it compares skewness in addition to the mean and variance that FID measures, and it is an unbiased estimator. KID can be measured by computing the squared maximum mean discrepancy (MMD) between distributions after a polynomial kernel has been applied.

**Precision and Recall (P&R)** Precision and Recall [29] were introduced to address FID’s coupling of diversity and quality by providing measures that evaluate each property separately. Precision measures generation realism, while Recall measures sample diversity. To measure P&R with graphs, all real and generated samples are first embedded into a GIN feature space. Manifolds are then constructed by extending a radius from each embedded sample in a set to its  $K^{th}$  nearest neighbour to form a hypersphere, with the union of all hyperspheres representing a manifold. Two manifolds are produced: one for real graphs, and one for generated graphs. Precision is defined as the percentage of generated graphs that fall within the manifold of real graphs, while Recall is defined as the percentage of real graphs which fall within the manifold of generated graphs. Though useful, Precision and Recall are susceptible to outliers [27].

**Density and Coverage (D&C)** Density and Coverage have recently been introduced as robust alternatives for Precision and Recall, respectively [27]. Density is calculated as the average number of real examples within whose manifold radius each generated sample falls into. Coverage is described as the percentage of real examples which have a generated sample fall within their manifold radius.

### B Thermometer encoding

We experiment with a thermometer encoding motivated by ordinal regression [33] in the edge type module, which accounts for the explicit ordering of each integer offset. In a typical classification problem the goal is to predict the probability that a data point  $x$  belongs to its class  $k$ , and the target vector is  $t = (0, 0, \dots, 1, \dots, 0, 0)$ , where  $t_k$  is one and all other elements are zero. This is easily done through the use of a softmax activation function. In the thermometer encoding, if a data point  $x$  belongs to class  $k$ , then it must also be classified into categories  $(1, 2, \dots, k - 1)$ . Thus, the target vector becomes  $t = (1, 1, \dots, 1, 0, \dots, 0)$  where  $t_i$  is one if  $i \leq k$ , and zero if  $i > k$ . We obtain a prediction  $O$  for this vector by independently applying the sigmoid activation function to each output as described by Cheng [33]. During generation, we begin with  $o_0$  and sample from consecutive elements of  $O$  until a zero is obtained.

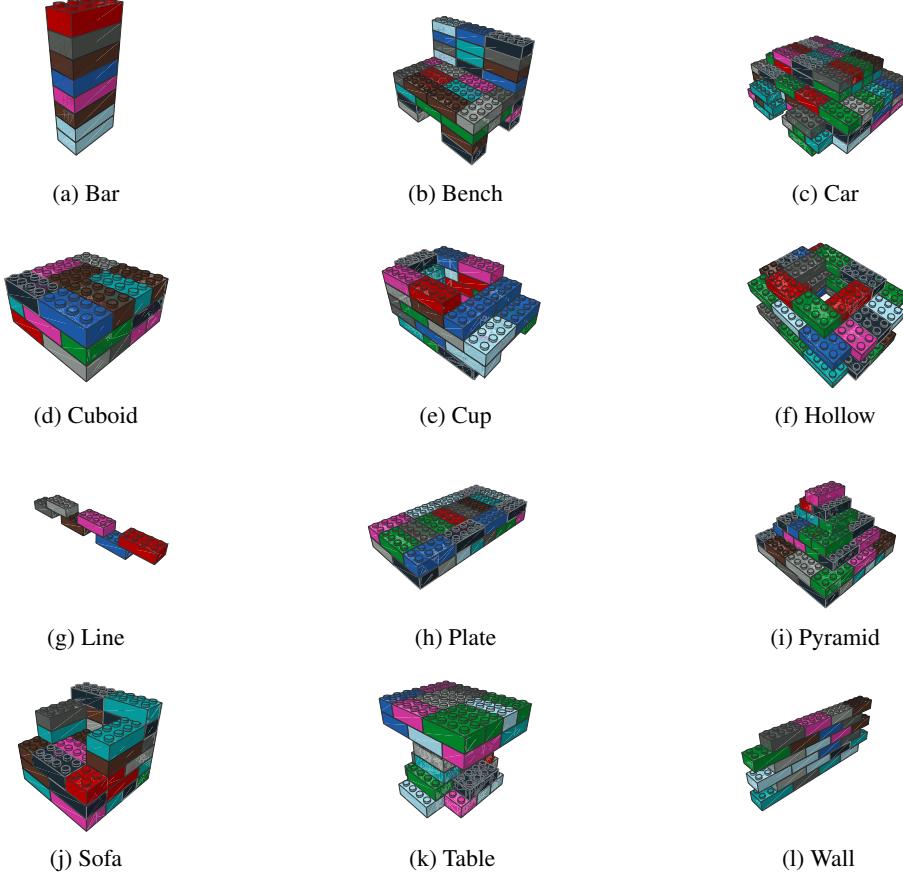


Figure 5: Representative examples of each class in the dataset

## C The LEGO dataset

The LEGO dataset consists of 12 classes and a total of 360 LEGO structures, each of which was created by one of twelve human subjects. Each class represents a relatively simple shape like “bench” or “pyramid”, and a representative example of each class in the dataset is shown in Figure 5. The dataset has an average size of 56 bricks/nodes and 80 edges, and records the sequence of actions followed to create each structure; it is perfect for our needs. In addition, our LEGO graph representation is sensitive to rotational shifts, meaning the dataset can be readily augmented by rotating the original structures by 90, 180, and 270 degrees. We leverage this sensitivity to train generative models with a larger capacity than what would typically be possible with the original dataset alone. An example of each class in the dataset is shown in Figure 5.

## D Bayesian optimization baseline

The Bayesian optimization baseline from Kim et al. [9] utilizes a voxel representation of LEGO structures, i.e. each structure could be represented by a 3D matrix of occupied and unoccupied cells or voxels. Given a partially assembled LEGO structure, the position of the next LEGO brick is posed as a score maximization problem, and Bayesian optimization is utilized to efficiently select positions to evaluate. The authors describe two different evaluation functions — one that uses an entire class of training examples, and one that only uses a single example. The latter is analogous to using Bayesian optimization to create a LEGO structure with shape guidance from a single training example, and we randomly select 200 samples from our dataset to recreate using this method as our baseline. Framing the problem as they have eliminates the possibility of creating an invalid LEGO structure. Note that this should be the most favourable setting for their model under all the metrics, with the exception of metrics that measure diversity (Recall and Coverage).