MAGNET: MOTIF-AGNOSTIC GENERATION OF MOLECULES FROM SHAPES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in machine learning for molecules exhibit great potential for facilitating drug discovery from *in silico* predictions. Most models for molecule generation rely on the decomposition of molecules into frequently occurring substructures (motifs), from which they generate novel compounds. While motif representations greatly aid in learning molecular distributions, such methods struggle to represent substructures beyond their known motif set. To alleviate this issue and increase flexibility across datasets, we propose MAGNet, a graph-based model that generates abstract shapes before allocating atom and bond types. To this end, we introduce a novel factorisation of the molecules' data distribution that accounts for the molecules' global context and facilitates learning adequate assignments of atoms and bonds onto shapes. Despite the added complexity of shape abstractions, MAGNet outperforms most other graph-based approaches on standard benchmarks. Importantly, we demonstrate that MAGNet's improved expressivity leads to molecules with more topologically distinct structures and, at the same time, diverse atom and bond assignments.

1 INTRODUCTION

The role of machine learning (ML) models in generating novel compounds has grown significantly, finding applications in fields like drug discovery, material science, and chemistry (Bian & Xie, 2021; Butler et al., 2018; Choudhary et al., 2022; Hetzel et al., 2022; Moret et al., 2023). These models offer a promising avenue for efficiently navigating the vast chemical space and generating unique molecules with specific properties (Hoffman et al., 2022; Zhou et al., 2019). A key ingredient contributing to the success of these models is their ability to encode molecules in a meaningful way, often employing graph neural networks (GNNs) to capture the structural intricacies (Gilmer et al., 2017; Mercado et al., 2021; Shi et al., 2021). Moreover, the inclusion of molecular fragments, known as motifs, significantly influences the generation process by enabling the model to explicitly encode complex structures such as cycles (Sommer et al., 2023). This contrasts with the gradual formation of ring-like structures from individual atoms, forming chains until both ends unite.

In the context of fragment-based models, researchers have adopted various techniques to construct fragment vocabularies, which can be categorised into chemically inspired and data-driven approaches. For example, both JT-VAE (Jin et al., 2018) and MoLeR (Maziarz et al., 2022) adhere to a heuristic strategy that dissects molecules into predefined structures, primarily consisting of ring systems and acyclic linkers. However, the diverse appearances of molecular structures result in various challenges concerning the vocabulary. While the large fragments in JT-VAE do not generalise well to larger datasets, the number of fragments becomes an issue for MoLeR. Such challenges are not unique to heuristic fragmentation methods but also extend to data-driven approaches like PS-VAE (Kong et al., 2022) and MiCaM (Geng et al., 2022). These approaches can set the number of chosen fragments but often resort to representing cyclic structures by combinations of chains. MiCaM, in particular, takes an approach that additionally incorporates attachment points for each fragment, resulting in a vocabulary that is "connection-aware", increasing its size by a significant margin. This leads to a situation where the included fragments often fall short to comprehensively represent the full spectrum of molecules present in the datasets (Sommer et al., 2023). Consequently, a generative model must refer to individual atoms in order to generate uncommon structures, a demanding task as the infrequent structures are often also more complex.

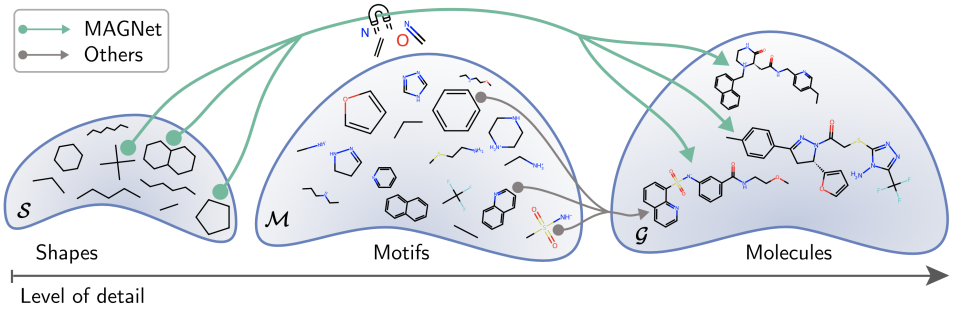

Figure 1: Current methods use motifs to compose new molecules. MAGNet advances this perspective and abstracts molecular subgraphs to untyped shapes. Being agnostic to atom and bond features in the beginning of the generation process, MAGNet generates molecules with great structural variety.

Our work addresses this issue by abstracting fragments to shapes, representing only the binary adjacencies, and using this abstraction for vocabulary creation. Since the same shape can lead to multiple distinct atom and bond representations, this effectively reduces the required vocabulary size to capture a dataset's molecules, while still encoding any hard-to-learn structural elements. Simultaneously, we propose an effective generation procedure in the form of our MAGNet model, which learns to generate molecules from abstract shape sets. The shape abstraction enables us sample the *entire* molecule context at each stage and construct the molecular graph in a hierarchical fashion. This hierarchy progresses from shapes and their connectivity to the atom level, where shape representations are generated, ultimately leading to the full definition of the molecule. Notably, our model is the first to freely learn the distribution over shape representations, enabling it to sample a greater variety of atom and bond attributes compared to fragment-based approaches.

While we identify limitations in some of the benchmarks for *effectively* evaluating whether complex structures have been captured, we provide alternative analyses to offer a more comprehensive understanding of the generative behaviour of our and other approaches. Overall, our results highlight MAGNet's ability to reliably reconstruct uncommon shapes, including large cycles, surpassing other methods that face challenges in this area. Additionally, MAGNet excels in sampling a wider range of shape representations, closely aligning with the original data distribution. We also demonstrate MAGNet's advantageous properties, including its fidelity to the presented latent code—a departure from existing methods. Finally, our model enables simultaneous conditioning on multiple scaffolds, underscoring its versatility for diverse applications in molecular generation.

2 Modelling Molecules from Shapes

We will first detail the factorisation of the data distribution of molecular graphs, denoted as $\mathbb{P}(\mathcal{G})$. Following this, we will introduce our novel approach to fragmenting molecules into shapes and present the corresponding MAGNet model.

Factorising the data distribution $\mathbb{P}(\mathcal{G})$ A molecular graph $\mathcal{G}$ is defined by its structure together with its node and edge features, describing atoms and bonds, respectively. In this work, we consider a factorisation of $\mathbb{P}(\mathcal{G})$ that builds the molecular graph $\mathcal{G}$ from its shape representation, i.e.

$$\mathbb{P}(\mathcal{G}) = \mathbb{P}(\mathcal{G} \mid \mathcal{G}_{\mathcal{S}})\, \mathbb{P}(\mathcal{G}_{\mathcal{S}}) \,,$$

where $\mathcal{G}$ refers to the normal molecular graph and $\mathcal{G}_{\mathcal{S}}$ to its coarser shape graph. To factorise $\mathbb{P}(\mathcal{G}_{\mathcal{S}})$ further, we consider a fragmentation into a multiset of shapes $\mathcal{S}$ and their typed connectivity $A \in \{0, \mathrm{C}, \mathrm{N}, \dots\}^{|\mathcal{S}|\times|\mathcal{S}|}$. Each shape $S \in \mathcal{S}$ can be classified as one of three categories, namely (i) rings, (ii) junctions, and (iii) chains. A shape S only holds *structural information* about its $s = |S|$ nodes, meaning we can represent it as a binary matrix, i.e. $S \in \{0,1\}^{s\times s}$. The shape connectivity A, on the other hand, is typed and encodes whether two shapes share an atom, signified by $A \neq 0$, and also the respective atom type of the join, e.g. C or N. Note that A is a shape-level representations which does not hold any information about the exact position of the connection between shapes.

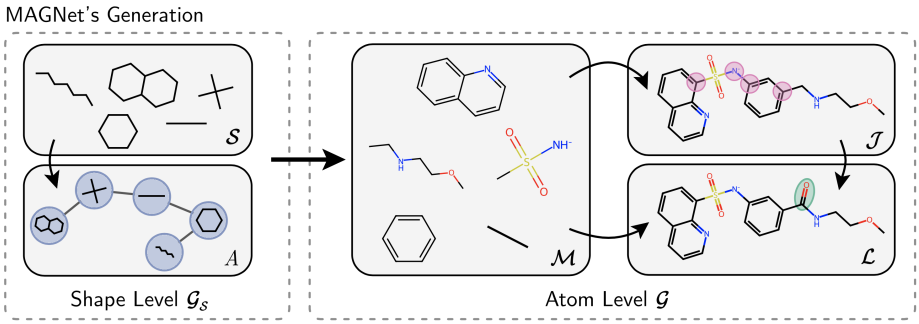

Figure 2: On the shape-level, MAGNet predicts the shape multiset $\mathcal{S}$ and its connectivity A. Progressing to the atom-level, $\mathcal{G}_S$ informs the generation of shape representations $\mathcal{M}$. To fully define the molecular graph $\mathcal{G}$, the joining positions $\mathcal{J}$ and the leaf atoms $\mathcal{L}$ are predicted.

Further, since all cyclic structures within the molecular graph are considered individual shapes, A always describes a tree on the shape-level.

Moving forward to the atom-level representation, a shape S can be equipped with node and edges features, corresponding to atom and bond types, respectively. We consider this representation of S to constitute a typed subgraph M, or fragment, of the input graph $\mathcal{G}$, and the associated multiset within the molecule $\mathcal{M}$. We note that a shape S can map to multiple distinct M as only the binary adjacencies will be shared between them; we will later on make use of context knowledge to select a suitable M.

The shape representations $\mathcal{M}$ and the shape connectivity A do not fully describe $\mathcal{G}$ yet as they do not specify the exact atomic positions where the shapes have to be joined. For this, we define the join set $\mathcal{J}$ which can be understood as the set of join nodes j that are contained in two motifs: $\mathcal{J} = \{j \mid j \in M_k,\ j \in M_l,\ A_{kl} \neq 0,\ S_k, S_l \in \mathcal{S}\}$. Finally, the full molecule is defined by attaching leaf atoms $\mathcal{L}$, which are those atoms with degree $d_i = 1$ and with neighbour $j \in \mathcal{N}_i$ with $d_j = 3$.

In conclusion, $\mathbb{P}(\mathcal{G})$ is factorised into the shape representations $\mathcal{M}$, the join set $\mathcal{J}$, and the leaf set $\mathcal{L}$:

$$\begin{aligned}\mathbb{P}(\mathcal{G}) &= \mathbb{P}(\mathcal{L}, \mathcal{J}, \mathcal{M}) \\ &= \mathbb{P}(\mathcal{L} \mid \mathcal{M}, \mathcal{J})\mathbb{P}(\mathcal{J} \mid \mathcal{M}, A)\,\mathbb{P}(\mathcal{M} \mid \mathcal{G}_\mathcal{S})\,\mathbb{P}(\mathcal{G}_\mathcal{S})\,,\end{aligned}$$

where we use $\mathcal{G}_\mathcal{S}$ to factorise the distribution. Note that $\mathcal{J}$ is *conditionally independent* of $\mathcal{S}$ given the motifs $\mathcal{M}$. The same applies to $\mathcal{L}$ being conditionally independent of A given $\mathcal{J}$.

During learning, we start with the coarser shape-level, that is $\mathcal{S}$ and A, and then proceed to the atom level for $\mathcal{M}$ and $\mathcal{J}$. Finally, we learn the distribution of leaves $\mathcal{L}$. Note that unlike sequential generation methods, this factorisation considers the entire molecule context at all times with only the level of detail increasing.

2.1 Fragmentation into shapes

MAGNet's fragmentation scheme aims to break down a given molecule into clear structural elements. Initially, we remove all leaf atoms $\mathcal{L}$ across the graph, following the approach outlined in previous works (Jin et al., 2020; Maziarz et al., 2022). Leaves are those nodes that have degree $d_i = 1$ and whose neighbouring node $j \in \mathcal{N}_i$ fulfills $d_{\mathcal{N}_i} = 3$. This step helps to divide the molecule $\mathcal{G}$ into cyclic and acyclic parts. Importantly, instead of modelling the connection between two fragments M_i and M_j with a connecting bond, we represent it by a shared atom, referred to as the join atom $\nu \in \mathcal{J}$. One benefit of this approach is that fragments are not truncated and, for example, two cycles which are connected by only one join atom are identified as two cycles instead of a single cycle that is connected to a chain at both ends.

Moreover, in order to reduce the number of required shapes as much as possible, we further decompose the resulting acyclic fragments. To this end, we introduce "junctions" which are defined by a center node with degree three or four present in an acyclic structure. The junction then contains the center nodes as well as its neighbours. This additional refinement reduces the number of fragments by a

factor of three, allowing to classify them into rings, junctions, and chains. On the ZINC dataset (Irwin et al., 2020), consisting of 249,456 compounds, our fragmentation results in 7371 typed subgraphs. When compared to the decomposition BBB procedure (Jin et al., 2020; Maziarz et al., 2022), our approach reduces complexity by collapsing acyclic structures into distinct shapes, leading to a reduction in vocabulary size by more than a half. Additionally, in contrast to data-driven methods like those outlined in Kong et al. (2022) and Geng et al. (2022), our decomposition method maintains structural integrity through a top-down approach.

Abstracting these typed subgraphs further to shapes ultimately results in 347 distinct shapes, where—in the best case—up to almost 800 fragments are consolidated into a single shape token. While the generative model is required to map a single shape to all its various representations, this fragmentation also enables us to model smoother transitions between shape representations. For instance, consider the molecules "C1NNCC1" and "C1NCNC1", which only differ in the relative positioning among atoms. Using shapes, the model can account for this difference without selecting potentially dissimilar tokens from a large vocabulary. We discuss the potential of our approach for transferability across datasets in Section 4.4 and Appendix C.

2.2 MAGNET'S GENERATION PROCESS

MAGNet is designed to represent the hierarchy of the factorisation into shape- and atom-level from Section 2, cf. Fig. 2. The model is trained as a VAE model (Kingma et al., 2015), where the latent vectors z are trained to encode meaningful semantics about the input data. That is, given a vector z from the latent space, MAGNet's generation process first works on the shape level to predict $\mathcal{G}_\mathcal{S}$, defined by the multiset $\mathcal{S}$ its connectivity A, before going to the atom level which is defined by the fragments $\mathcal{M}$, joins $\mathcal{J}$, and leaves $\mathcal{L}$.

Shape-Level On the shape-level, MAGNet first generates the **shape multiset** $\mathcal{S}$—the same shape can occur multiple times in one molecule—from the latent representation z. More specifically, we learn $\mathbb{P}(\mathcal{S} \mid z)$ by conditioning the generation on the latent code z and the intermediate representation of the shapes by a transformer model. On the sorted shape set, the network is optimised via Cross-Entropy (CE) loss, denoted by $L_\mathcal{S}$.

Given the shape multiset $\mathcal{S}$, MAGNet infers the **shape connectivity** A between shapes $S_i, S_j \in \mathcal{S}$. Formally, we learn $\mathbb{P}(A \mid \mathcal{S}, z) = \prod_{i,j=1}^{n} \mathbb{P}(A_{ij} = t \mid S, z)$ where $t \in \{0, \mathrm{C}, \mathrm{N}, \dots\}$ not only encodes the existence (or absence) of a shape connection but also its atom type. Further, we assume the individual connections, A_{ij} and A_{lk}, to be independent given the shape multiset $\mathcal{S}$ and the latent representation z. MAGNet implements this connectivity module using an MLP, optimised with CE loss L_A. We compute the loss on the shape level as $L_{\mathcal{G}_\mathcal{S}} = L_\mathcal{S} + L_A$. It is worth noting that the same shape can have different atom representations depending on its positions in the molecular graph, highlighting the importance of predicting A for generating atom-level shape representations.

Atom-Level Leveraging the shape set $\mathcal{S}$ and connectivity A, which together define a molecule's shape-level representation, MAGNet transitions to the atom-level by discerning appropriate node and edge attributes for each shape, referred to as **atom and bond types** $\mathcal{M}$. This step is pivotal in MAGNet's generation process, granting it independence from a pre-established fragment set and thus enhancing its flexibility. To model the shape representation $\mathbb{P}(M_i \mid \mathcal{S}, A, z)$ of shape S_i, the atoms are generated using a transformer model that constructs its memory by encoding the shape graph and incorporating embeddings tailored to each individual shape (Shi et al., 2021). Unlike the shape set, the sizes of the shapes are fixed, enabling a structured representation of each position within the shape.

Subsequently, the resulting atom embeddings are leveraged to determine the corresponding **bond types** M^b between connected nodes. This bond determination is achieved through the utilisation of an MLP, which effectively captures the necessary relationships for assigning correct bond types. Note that conditioning on A ensures that M_k includes all atoms required for connectivity also on the atom-level, i.e. the atom allocations for M have to respect all join types defined by the shape connectivity A: $A_{kl} \in \bigcup_j M_k^a \cap M_l^a$, where a signifies exclusive consideration of atoms. We optimise both atoms and bonds with a CE loss and denote the combined loss by $L_\mathcal{M}$.

Next, to establish connectivity on the atom level within the molecular graph, MAGNet proceeds to identify the **join nodes** $\mathcal{J}$. The join nodes' types are already determined by the connectivity A. MAGNet accomplishes this by predicting the specific atom positions p_a that need to be combined, collectively forming the join set $\mathcal{J}$. The likelihood of merging nodes i and j in shape representations M_k and M_l is represented by the merge probability $J_{ij}^{(k,l)} = \mathbb{P}(p_i \equiv p_j \mid \mathcal{M}, A, z)$, which constitutes the join matrix $J^{(k,l)} \in [0,1]^{V_{S_k} \times V_{S_l}}$. It is modelled and optimised using an MLP and CE loss $L_{\mathcal{J}}$.

Finally, predicting the **leaves** $\mathcal{L}$ involves two key aspects: determining the correct atom type for each leaf and establishing its connection to the molecule's atom representation, denoted as $\mathcal{C}$ for core molecule. To learn $\mathbb{P}(\mathcal{L}_S \mid \mathcal{C}, z)$, we assess each node position within the shape representation M_S. Similar to the approach employed in modelling the shape representation $P(M_i|S, A, z)$ for shape S_i, MAGNet utilises a transformer model to generate atom representations, optimised with CE loss $L_{\mathcal{L}}$. This model constructs its memory by encoding the present atom graph and incorporating embeddings tailored to each unique shape representation. The final atom loss is defined by $L_{\mathcal{G}} = L_{\mathcal{M}} + L_{\mathcal{J}} + L_{\mathcal{L}}$.

2.3 The MAGNet encoder

MAGNet's encoder aims to learn the approximate posterior $\mathbb{Q}(z \mid \mathcal{G})$. At its core, the encoder leverages a graph transformer Shi et al. (2021) for generating node embeddings of the molecular graph. Since MAGNet generates molecules in a coarse to fine-grained fashion, it is beneficial to encode information about the decomposition of the molecules. To achieve this, we use an additional GNN to capture the coarse connectivity within the shape graph. The embeddings of the molecular and the shape graph are computed by aggregating over the individual atom and shape nodes, respectively. In addition to these aggregations, we exclusively aggregate joins and leaves to represent the other essential components of the graph. The representation of the individual components—molecular graph, shapes, joins, and leaves—are concatenated and mapped to the latent space by an MLP, constituting the graph embedding $z_{\mathcal{G}}$. More details about the chosen node features as well as technical specifications of the encoder can be found in Appendix D.

Taken together, we optimise MAGNet according to the VAE setting, maximising the ELBO:

$$\begin{aligned} L &= \mathbb{E}_{z\sim\mathbb{Q}}\big[\mathbb{P}\big(\mathcal{G} \mid z\big)\big] + \beta D_{\mathrm{KL}}\big(\mathbb{Q}(z \mid \mathcal{G}) \mid P\big) \qquad \text{with} \quad P \sim \mathcal{N}(0, \mathbb{1}) \\ &= L_{\mathcal{G}_S} + L_{\mathcal{G}} + \beta D_{\mathrm{KL}}\,, \end{aligned}$$

where the KL-divergence D_{KL} serves to regularise the posterior $\mathbb{Q}(z \mid \mathcal{G})$ towards similarity with the Normal prior P in the latent space, weighted by β. However, we have observed that this regularisation alone is inadequate for achieving a smoothly structured latent space. To address this, we apply a Normalizing Flow to the latent space, aligning it more effectively with the prior (Tong et al., 2023).

3 Related Work

Molecule generation Existing generative models can be broadly divided into two categories: (1) string-based models, relying on string representations like SMILES or SELFIES (Adilov, 2021; Fang et al., 2023; Flam-Shepherd et al., 2022; Grisoni, 2023; Gómez-Bombarelli et al., 2018; Segler et al., 2018), which do not leverage structural information, and (2) graph-based models, which are inherently centered around molecular graphs. Graph-based approaches involve models that represent molecular graphs (1) primarily at the atom level and (2) predominantly through fragments. Regarding the generation process, they can be further divided into sequential methods (Ahn et al., 2021; Assouel et al., 2018; Bengio et al., 2021; Jin et al., 2020; Kajino, 2019; Khemchandani et al., 2020; Li et al., 2018; Lim et al., 2020; Liu et al., 2018; Luo et al., 2021; Mercado et al., 2021; Popova et al., 2019; Shi et al., 2020; Shirzad et al., 2022; Yang et al., 2021; You et al., 2019), building molecules per fragment while conditioning on a partial molecule, and all-at-once (AAO) approaches (Bresson & Laurent, 2019; De Cao & Kipf, 2018; Flam-Shepherd et al., 2020; Kong et al., 2022; Liu et al., 2021; Ma et al., 2018; Samanta et al., 2019; Simonovsky & Komodakis, 2018; Zang & Wang, 2020) that create each aspect of the molecular graph in a single step.

Fragmentation and shape representation Various techniques are available for constructing fragment vocabularies, with a distinction between chemically-inspired and data-driven approaches. For example, both HierVAE (Jin et al., 2020) and MoLeR (Maziarz et al., 2022) adopt a heuristic

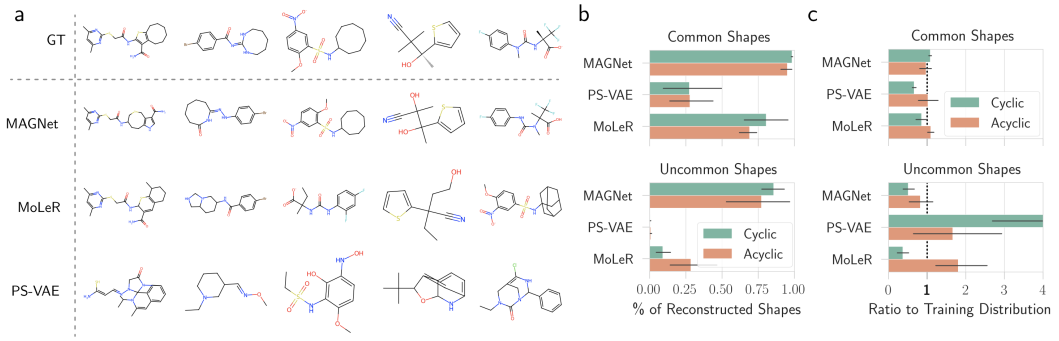

Figure 3: (**a**) Reconstruction of molecules that include large cycles or complex junctions. Relying on individual atoms to build these structures is not sufficient. Only MAGNet is able to reliably decode its latent code $z_{\mathcal{G}}$. (**b**) Percentage of reconstructed shapes. MAGNet substantially improves in reconstructing both common and—more importantly—uncommon shapes. (**c**) Comparison of sampled shapes to shape occurrences in the training distribution. A ratio of 1 is optimal.

strategy known as "breaking bridge bonds" to decompose molecules into rings and remainder fragments, emphasising chemically valid substructures. In a similar vein, JT-VAE (Jin et al., 2018) employs fragmentation guided by the construction of junction trees. In contrast, PS-VAE (Kong et al., 2022) and MiCaM (Geng et al., 2022) take a data-driven bottom-up approach, creating fragments by merging smaller components, starting from single atoms. MiCaM even integrates attachment points, resulting in a larger, "connection-aware" vocabulary.

MAGNet is a graph-based model that employs a unique approach by generating each hierarchy level in a single step, similar to existing AAO approaches. It positions itself between the traditional categories of single-atom and fragment-based models by utilising shapes as building blocks and subsequently generating appropriate atom and bond attributes. Our new fragmentation approach is designed to achieve concise shape representations, expanding its capacity to encode diverse structures within the same vocabulary size.

4 Experiments

We evaluate MAGNet's performance across several dimensions of the generation process. First, we investigate the reconstruction and sampling of shapes $\mathcal{S}$, as the fundamental component of MAGNet's factorisation, and assess how faithfully our model represents the diverse structural characteristics found in molecules. We continue to evaluate the generative performance using established benchmarks. Next, we analyse MAGNet's ability to determine suitable atom and bond allocations $\mathcal{M}$, highlighting its distinctive approach compared to baseline models. Moreover, we demonstrate how MAGNet's factorisation and shape vocabulary facilitate its ability to generalise effectively across different datasets in a zero-shot manner. Finally, we explore the possibilities enabled by our factorisation, such as conditioning on various levels of the generation process.

4.1 Using shapes to represent structural variety of molecules

Reconstructing large cycles Our first experiment provides qualitative insights into how accurately shapes are decoded, $\mathbb{P}(\mathcal{S} \mid z_{\mathcal{G}})$. For this experiment, we employ two baselines: PS-VAE (Kong et al., 2022) and MoLeR (Maziarz et al., 2022), the models with the best performance within their respective category on the GuacaMol Benchmark (see Sec. 4.2). We assess the decoder's performance in reconstructing molecules from the test set, which includes uncommon shapes like large rings or complex junctions. Our observations reveal that the baseline models have difficulty in constructing complex shapes, as illustrated in Figure 3a. This limitation is likely attributed to the absence of such shapes in their top-k vocabularies. Consequently, these models face the challenge of constructing shapes such as large rings from individual atoms. In contrast, our proposed model, MAGNet, operates with a moderately-sized shape vocabulary that includes complex shapes, enabling it to generate

Table 1: GuacaMol and MOSES Benchmark. We report mean and standard deviation using 5 random seeds and highlight the best overall graph-based method as well as **the best within each category.**

		GuacaMol		MOSES			
		FCD (↑)	KL (↑)	IntDiv (↑)	logP (↓)	SA (↓)	QED (↓)
SM.	CharVAE	0.17 ± 0.08	0.78 ± 0.04	**0.88 ± 0.01**	0.87 ± 0.14	0.48 ± 0.13	0.06 ± 0.03
	SM.-LSTM	**0.93 ± 0.00**	**1.00 ± 0.00**	0.87 ± 0.00	**0.12 ± 0.01**	**0.04 ± 0.02**	**0.00 ± 0.00**
Sequential	GraphAF	0.05 ± 0.00	0.67 ± 0.01	**0.93 ± 0.00**	0.41 ± 0.02	0.88 ± 0.10	0.22 ± 0.01
	HierVAE	0.53 ± 0.14	0.92 ± 0.01	0.87 ± 0.01	0.36 ± 0.17	0.20 ± 0.14	0.03 ± 0.00
	MiCaM	0.63 ± 0.02	0.94 ± 0.00	0.87 ± 0.00	0.20 ± 0.05	0.51 ± 0.03	0.08 ± 0.00
	JTVAE	0.75 ± 0.00	0.94 ± 0.00	0.86 ± 0.00	0.28 ± 0.03	0.34 ± 0.01	**0.01 ± 0.00**
	MoLeR	**0.80 ± 0.01**	**0.98 ± 0.00**	0.87 ± 0.00	**0.13 ± 0.02**	**0.06 ± 0.01**	**0.01 ± 0.01**
AAO	PSVAE	0.28 ± 0.01	0.83 ± 0.00	**0.89 ± 0.00**	0.34 ± 0.02	1.18 ± 0.05	0.05 ± 0.00
	MAGNet	**0.76 ± 0.00**	**0.95 ± 0.00**	0.88 ± 0.00	**0.22 ± 0.01**	**0.12 ± 0.01**	**0.01 ± 0.00**

molecules that closely adhere to the latent code and the corresponding ground truth molecules. We quantify this result through the displacement of latent codes in Appendix B.

MAGNet reliably decodes shapes Building on our analysis of large cycles and uncommon junctions, we extend our investigation to assess how effectively different models can reconstruct the shape set $\mathcal{S}$ in a general context. Given our focus on $\mathbb{P}(\mathcal{S} \mid z)$, we can disregard the shape connectivity A and representations $\mathcal{M}$. As illustrated in Fig. 3b, our findings demonstrate that MAGNet consistently outperforms both MoLeR and PS-VAE. Notably, our results provide additional support for the hypothesis that the other methods do not effectively learn the concept of a shape, relying primarily on the information encoded in their vocabulary. As a result, they may struggle to decode a common shape S when the signal originally comes from an uncommon representation M.

MAGNet matches the distribution of shapes more accurately To further check that uncommon shapes are also sampled, we analyse the shape set of generated molecules. If the other models are able to represent scaffolds that are not included in their vocabulary, they should be able to reflect the reference distribution of shapes. Fig. 3a shows that this is *not* the case in practice. For this evaluation, we decompose sampled molecules into their shapes. We then measure the models' over- and undersampling behaviour based on the ratio r_{S_i}:

$$r_{S_i} = \frac{c_s(S_i)}{\sum_k c_s(S_k)} \times \frac{\sum_k c_t(S_k)}{c_t(S_i)} \,,$$

where c_t and c_s refer to the count function applied to the training set and sampled sets, respectively. On common shapes, i.e. those that occur in more than 10% of the molecules, all evaluated models are able to match the ratio of the ZINC distribution. For uncommon shapes, however, both MoLeR and PS-VAE fail: while PS-VAE heavily oversamples both ring-like structures and chain-like structures, MoLeR oversamples chain-like structures and undersamples ring-like structures. MAGNet matches the reference distribution best across categories and we conclude that the proposed abstraction to shapes is also beneficial for generation.

4.2 Generative performance evaluated on common benchmarks

Employing two standard benchmarks for de-novo molecule generation, we establish MAGNet's competitive generative performance. The GuacaMol benchmark provides a framework for assessing the ability of a generative model to sample in accordance with the distribution of a molecular dataset (Brown et al., 2019). Next to evaluating the uniqueness and novelty of sampled molecules, the benchmark also computes distributional distances to the reference, i.e. the KL-divergence and Fréchet distance (FCD). We use the MOSES benchmark (Polykovskiy et al., 2020) to report measures for the internal diversity (IntDiv) of generated molecules as well as chemical properties such as synthetic accessability (SA), the octanol-water partition coefficient (logP), and the viability for drugs (QED).

Baselines for these benchmarks additionally include JTVAE (Jin et al., 2018), HierVAE (Jin et al., 2020), and MiCaM (Geng et al., 2022) as sequential, fragment-based methods. For those model with a variable vocabulary, we set the size to 350. We also include GraphAF (Shi et al., 2020) as a purely atom-based model. While the focus of this work lies on graph-based molecule generation,

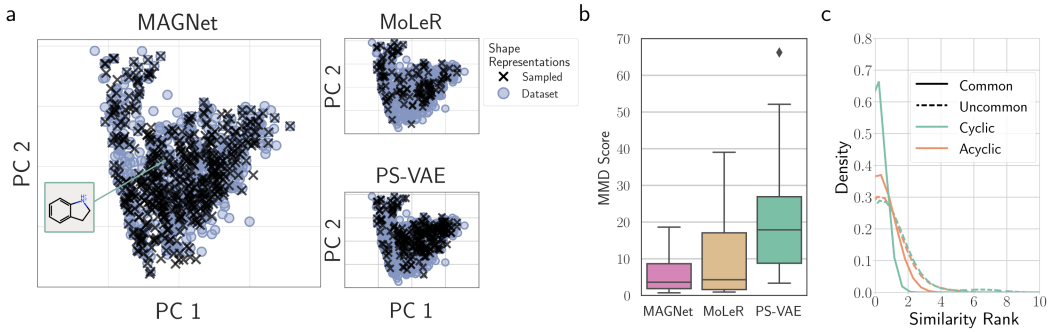

Figure 4: (**a**) Example of generated fragments by MAGNet and baseline methods. (**b**) MMD computation to quantify similarity between generated and ground truth shape representations. (**c**) Rank comparison between predicted fragments and their original counterparts.

we furthermore add the SMILES-based baseline of the GuacaMol benchmark SMILES-LSTM (SM-LSTM) (Segler et al., 2018) as well as the SMILES-based VAE CharVAE (Gómez-Bombarelli et al., 2018). Importantly, SM-LSTM does not have a latent space and can thus not perform targeted decoding. For all baselines, we use the hyperparameters specified in their respective works. We specify MAGNet's hyperparameter configuration in Appendix E.

MAGNet is the best AAO model on standard benchmarks The benchmark is conducted on 10^4 latent codes sampled from the prior distribution, $z \sim P$, and decoded into valid molecules. Our results for both benchmarks on the ZINC dataset are depicted in Table 1. We do not report Novelty and Uniqueness, as almost all evaluated models achieve 100% on these metrics. Solely GraphAF and HierVAE achieve 91% and 96% Uniqueness and Novelty, respectively. For baselines like SM-LSTM and CharVAE, which are not able to achieve 100% Validity, we sample until we obtain 10^4 valid molecules. While MoLeR sets the state of the art on both FCD and KL, MAGNet overall performs competitively, outperforming all other graph-based baselines. This supports the proposed factorisation in Section 2 while also challenging the common perception that methods for molecule generation must rely on motif vocabularies to obtain good generative performance.

Moreover, despite the FCD being one of the most important metrics for molecular distribution learning, we find that it fails to provide insights about the structural diversity of the molecules that are generated. Upon further investigation, we evaluate the benchmark on a subset of 10^4 molecules from the training data. The subset was filtered to include only the 10 most common shapes found in the dataset, resulting in an FCD score of 0.89. This observation offers an explanation for why models like MoLeR can achieve state-of-the-art FCD scores, despite not accurately capturing the distribution of uncommon shapes, as demonstrated in Section 4.1. This underscores that our evaluation of the structural diversity of molecule is orthogonal to these benchmarks, providing valuable insights into the tails of the molecular distribution.

4.3 Generation of shape representations $\mathcal{M}$

Having established MAGNet's ability to utilise its shape vocabulary to reliably decode a molecule's structure and sample diversely, we further evaluate MAGNet's atom and bond allocation to shapes.

MAGNet's shape representations are superior to fixed fragments The larger a given shape, the more the combinatorial aspect starts to dominate: with a size-limited vocabulary, it is challenging to reflect the diversity of a shape's realisations during decoding. This is shown in Fig. 4a, which provides a qualitative view on shape representations. We extract shape representations of a given shape from the molecules sampled in Table 1 and plot the two principal components of their fingerprints. Only for this shape, there are 791 representations in ZINC. Both PS-VAE as well as MoLeR are not able to cover the distribution fully, even though the shape appears commonly in the dataset. PS-VAE's and MoLeR's factorisation and vocabulary show lacking ability to cover all shape representations sufficiently. MAGNet, by contrast, covers all parts of the distribution, even outliers. Fig. 3b shows the MMD quantification of the results in Fig. 3a, confirming that MAGNet is able to best cover the

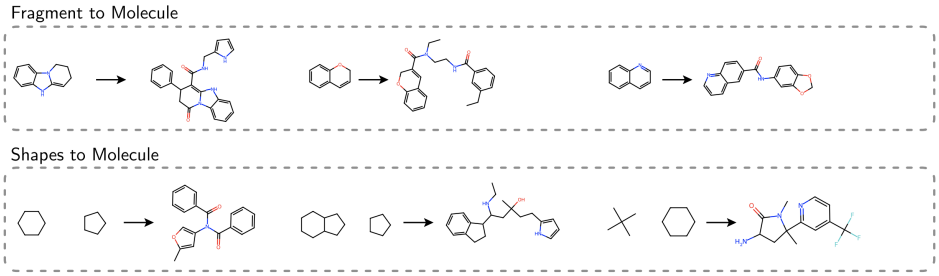

Figure 5: Examples of conditional molecule generation with MAGNet. The generation is conditioned on (*top*) a complete fragment, including atoms and edges, and (*bottom*) two distinct shapes.

entire distribution of shape representations. Being able to *reliably* decode a large variety of molecular scaffolds is especially important for downstream tasks such a molecule optimisation.

Allocation of atom and bonds to shapes Extending the sampling analysis in Fig. 3b, we quantify the process of turning an abstract shape into a chemically valid substructure in Fig. 4c. For each shape in the ZINC dataset, we compute the similarities between the set of all predicted and ground truth allocations. Given a ground truth assignment and a successful shape decoding, we measure how the decoded allocation ranks compared to known allocations. In the majority of cases, MAGNet achieves rank 0 or 1 in the shape allocation, with uncommon rings being the most challenging to decode.

4.4 Application across datasets and conditional sampling

Having analysed the generative performance of the MAGNet model and the benefit of the proposed shape fragmention, we continue to investigate how well the shape abstraction derived from the ZINC data translates to other datasets. After this, we showcase how one can use MAGNet's whole context generation for the generation of linkers and scaffold constrained generation.

Shape abstractions translate well across datasets To examine the flexibility of our shape fragmentation, we evaluate its transferability to unseen datasets with distinct molecular distributions through zero-shot generalization in Appendix C. We use the datasets QM9 (Wu et al., 2018), GuacaMol (Brown et al., 2019), CheMBL (Mendez et al., 2019) and L1000 (Subramanian et al., 2017). Note that we do not finetune any model on the unseen datasets and only use the vocabulary extracted from ZINC. MAGNet is able to achieve the highest similarity scores across all datasets, underscoring the flexibility of the fragmentation and MAGNet's expressive power across the space of drug-like molecules.

MAGNet efficiently generates molecules conditioned on shapes and scaffolds In the context of potential downstream applications, we investigate novel scaffold conditioning methods made possible by MAGNet's factorisation. Besides the latent space interpolation in Appendix A, Figure 5 illustrates that MAGNet is capable to condition not only on a single scaffold but also on multiple scaffolds, even when they are not directly connected within the resulting molecule. This poses a significant challenge for models like MoLeR, which rely on extending connected subgraphs for scaffold conditioning. Moreover, MAGNet enables conditioning solely based on shapes, enabling the free generation of atoms and bonds—a form of conditioning that was previously not possible.

5 Conclusion

We present MAGNet, a generative model for molecules that relies on a novel factorisation to disentangle structure from features, thus leading to a general abstraction in the space of molecules. MAGNet exhibits stronger performance at representing structures than existing models while also showing favourable results in generative tasks. While we argue that a global context like the one adopted in MAGNet is important for shape representations, modifications thereof can also be promising for sequential models. Finally, our proposed abstraction to shapes lends itself to *general* applications in graph generative models beyond the molecular world.

Reproducibility Statement To ensure reproducibility we make available the code for MAGNet. Additional documentation is provided in Appendix E, including all hyperparameters and training specifications necessary to reproduce the results discussed in this work. Both training and inference work on readily-available hardware, and detailed computational requirements are outlined in Appendix E. All data used for the experimental evaluation is publicly available.

Ethics Statement We commit to full transparency by making our research code publicly available. Together with openly accessible datasets, this facilitates widespread utilisation of MAGNet. However, this comes with potential risks of misuse inherent to the field of drug discovery. The capacity to generate molecules can go beyond benevolent drug discovery and can inadvertently lead to the creation of hazardous compounds or substances with unforeseen consequences. These risks emphasise the necessity for responsible use and oversight in the application of our methodology. However, our work also holds the potential to advance drug discovery efforts, potentially aiding in the identification of new pharmaceutical compounds.

REFERENCES

Sanjar Adilov. Neural Language Modeling for Molecule Generation, 2021.

Sungsoo Ahn, Binghong Chen, Tianzhe Wang, and Le Song. Spanning tree-based graph generation for molecules. In *International Conference on Learning Representations*, 2021.

Rim Assouel, Mohamed Ahmed, Marwin H. Segler, Amir Saffari, and Yoshua Bengio. DEFactor: Differentiable Edge Factorization-based Probabilistic Graph Generation. Technical report, arXiv, 2018.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation. Technical report, Neural Information Processing Systems, 2021.

Yuemin Bian and Xiang-Qun Xie. Generative chemistry: drug discovery with deep learning generative models. *Journal of Molecular Modeling*, 2021. ISSN 0948-5023. doi: 10.1007/s00894-021-04674-8.

Xavier Bresson and Thomas Laurent. A Two Step Graph Convolutional Decoder for Molecule Generation. Technical report, Machine Learning and the Physical Sciences Workshop, 2019.

Nathan Brown, Marco Fiscato, Marwin H.S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling*, 2019. ISSN 1549-9596, 1549-960X. doi: 10.1021/acs.jcim.8b00839.

Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0337-2.

Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon J. L. Billinge, Elizabeth Holm, Shyue Ping Ong, and Chris Wolverton. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 2022. ISSN 2057-3960. doi: 10.1038/s41524-022-00734-6.

Nicola De Cao and Thomas Kipf. MolGAN: An implicit generative model for small molecular graphs. Technical report, arXiv, 2018.

Yin Fang, Ningyu Zhang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Molecular Language Model as Multi-task Generator, 2023.

Daniel Flam-Shepherd, Tony Wu, and Alan Aspuru-Guzik. Graph Deconvolutional Generation. Technical report, arXiv, 2020.

Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. Language models can learn complex molecular distributions. *Nature Communications*, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30839-x.

Zijie Geng, Shufang Xie, Yingce Xia, Lijun Wu, Tao Qin, Jie Wang, Yongdong Zhang, Feng Wu, and Tie-Yan Liu. De novo molecular generation via connection-aware motif mining. In *The Eleventh International Conference on Learning Representations*, 2022.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Francesca Grisoni. Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology*, 2023. ISSN 0959-440X. doi: 10.1016/j.sbi.2023.102527.

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 2018. ISSN 2374-7943, 2374-7951. doi: 10.1021/acscentsci.7b00572.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, 2022.

Samuel C. Hoffman, Vijil Chenthamarakshan, Kahini Wadhawan, Pin-Yu Chen, and Payel Das. Optimizing molecules using efficient queries from property evaluations. *Nature Machine Intelligence*, 2022. ISSN 2522-5839. doi: 10.1038/s42256-021-00422-y.

John J. Irwin, Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical Information and Modeling*, 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c00675.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction Tree Variational Autoencoder for Molecular Graph Generation. Technical report, International Conference on Machine Learning, 2018.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical Generation of Molecular Graphs using Structural Motifs. Technical report, International Conference on Machine Learning, 2020.

Hiroshi Kajino. Molecular Hypergraph Grammar with Its Application to Molecular Optimization. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019.

Yash Khemchandani, Stephen O'Hagan, Soumitra Samanta, Neil Swainston, Timothy J. Roberts, Danushka Bollegala, and Douglas B. Kell. DeepGraphMolGen, a multi-objective, computational strategy for generating molecules with desirable properties: a graph convolution and reinforcement learning approach. *Journal of cheminformatics*, 2020.

Durk P Kingma, Tim Salimans, and Max Welling. Variational Dropout and the Local Reparameterization Trick. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015.

Xiangzhe Kong, Wenbing Huang, Zhixing Tan, and Yang Liu. Molecule Generation by Principal Subgraph Mining and Assembling, 2022.

G. Landrum. RDKit, 2010.

Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia. Learning Deep Generative Models of Graphs. Technical report, arXiv, 2018.

Jaechang Lim, Sang-Yeon Hwang, Seungsu Kim, Seokhyun Moon, and Woo Youn Kim. Scaffold based molecular design using graph generative model. *Chemical Science*, 2020. ISSN 2041-6520, 2041-6539. doi: 10.1039/C9SC04503A.

Meng Liu, Keqiang Yan, Bora Oztekin, and Shuiwang Ji. GraphEBM: Molecular Graph Generation with Energy-Based Models. Technical report, arXiv, 2021.

Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander L. Gaunt. Constrained Graph Variational Autoencoders for Molecule Design. Technical report, Advances in Neural Information Processing Systems, 2018.

Youzhi Luo, Keqiang Yan, and Shuiwang Ji. GraphDF: A Discrete Flow Model for Molecular Graph Generation. Technical report, International Conference on Machine Learning, 2021.

Tengfei Ma, Jie Chen, and Cao Xiao. Constrained Generation of Semantically Valid Graphs via Regularizing Variational Autoencoders. Technical report, Advances in Neural Information Processing Systems, 2018.

Krzysztof Maziarz, Henry Jackson-Flux, Pashmina Cameron, Finton Sirockin, Nadine Schneider, Nikolaus Stiefl, Marwin Segler, and Marc Brockschmidt. Learning to Extend Molecular Scaffolds with Structural Motifs. Technical report, arXiv, 2022.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019.

Rocío Mercado, Tobias Rastemo, Edvard Lindelöf, Günter Klambauer, Ola Engkvist, Hongming Chen, and Esben Jannik Bjerrum. Graph Networks for Molecular Design. *Machine Learning: Science and Technology*, 2021.

Michael Moret, Irene Pachon Angona, Leandro Cotos, Shen Yan, Kenneth Atz, Cyrill Brunner, Martin Baumgartner, Francesca Grisoni, and Gisbert Schneider. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nature Communications*, 2023. ISSN 2041-1723. doi: 10.1038/s41467-022-35692-6.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models, 2020.

Mariya Popova, Mykhailo Shvets, Junier Oliva, and Olexandr Isayev. MolecularRNN: Generating realistic molecular graphs with optimized properties. Technical report, arXiv, 2019.

Bidisha Samanta, Abir De, Gourhari Jana, Pratim Kumar Chattaraj, Niloy Ganguly, and Manuel Gomez-Rodriguez. NeVAE: A Deep Generative Model for Molecular Graphs, 2019.

Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1): 120–131, 2018.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation. Technical report, arXiv, 2020.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification, 2021.

Hamed Shirzad, Hossein Hajimirsadeghi, Amir H. Abdi, and Greg Mori. TD-GEN: Graph Generation Using Tree Decomposition. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

Martin Simonovsky and Nikos Komodakis. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders, 2018.

Johanna Sommer, Leon Hetzel, David Lüdke, Fabian J. Theis, and Stephan Günnemann. The Power of Motifs as Inductive Bias for Learning Molecular Distributions. 2023.

Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint 2302.00482*, 2023.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A Benchmark for Molecular Machine Learning, 2018.

Soojung Yang, Doyeong Hwang, Seul Lee, Seongok Ryu, and Sung Ju Hwang. Hit and Lead Discovery with Explorative RL and Fragment-based Molecule Generation. *Advances in Neural Information Processing Systems*, 2021.

Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. Technical report, arXiv, 2019.

Chengxi Zang and Fei Wang. MoFlow: An Invertible Flow Model for Generating Molecular Graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020. doi: 10.1145/3394486.3403104.

Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N. Zare, and Patrick Riley. Optimization of Molecules via Deep Reinforcement Learning. *Scientific Reports*, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-47148-x.

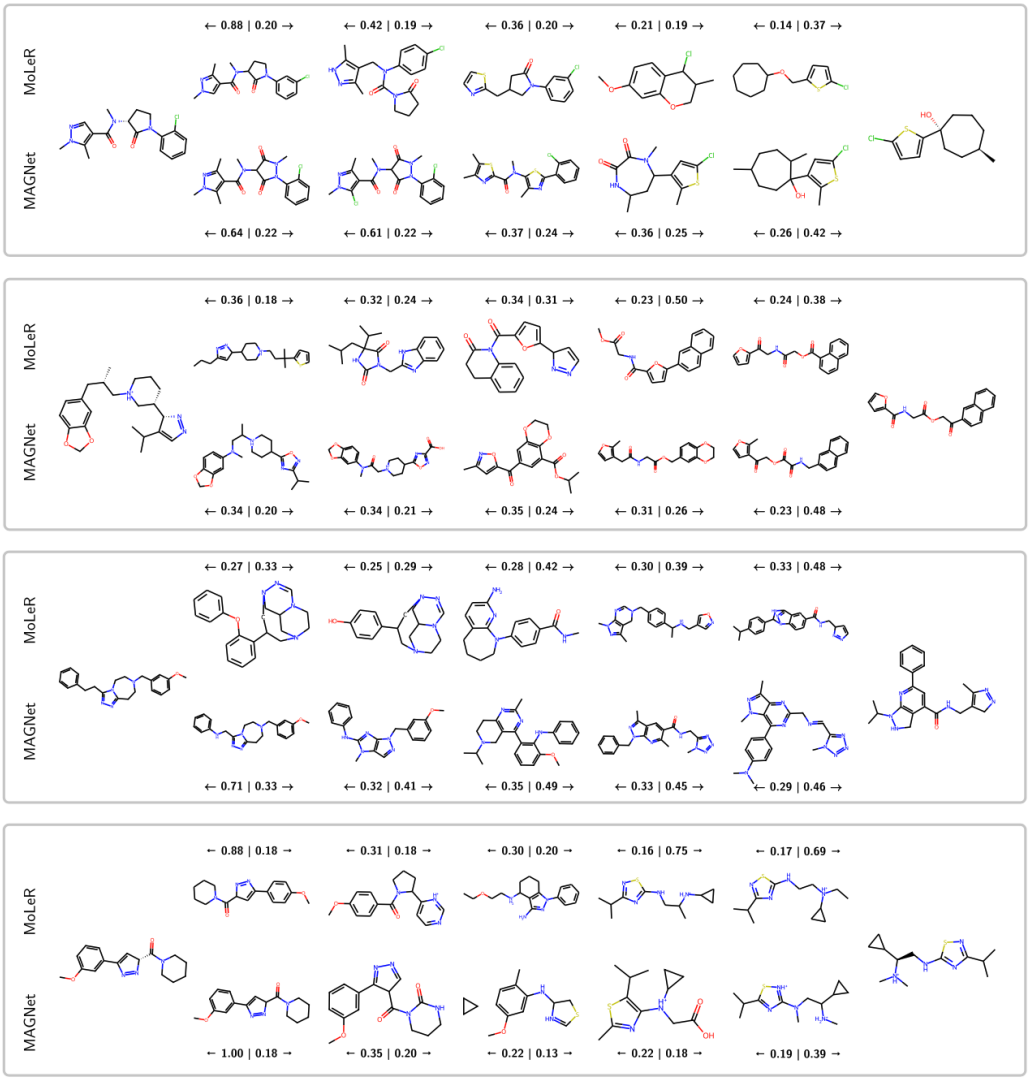

Figure 6: We provide four interpolation examples for MAGNet and MoLeR. The input molecules (left and right) are shared between the two models. We report the Tanimoto similarity as a rough estimate for the interpolation's goodness.

A Interpolation

Extending on Fig. 5, we additionally provide examples for latent space interpolation in Fig. 6. During interpolation, MAGNet stays faithful to the shapes present in the input molecules. The last row shows a failure case of MAGNet: it identifies a shape multiset that can not be fully connected to a molecule.

B Displacement of Latent Codes

To quantify the discrepancy between input and reconstructed molecule visible in Fig. 3a, we measure the displacement of latent codes. That is, we obtain the latent representation for the input molecule, decode this latent representation into the output molecule and then obtain the latent representation for the output molecule. This verifies what can be observed qualitatively in Fig. 3a–the evaluated baselines can not reliably decode complex shapes.

C TRANSFERABILITY OF SHAPES

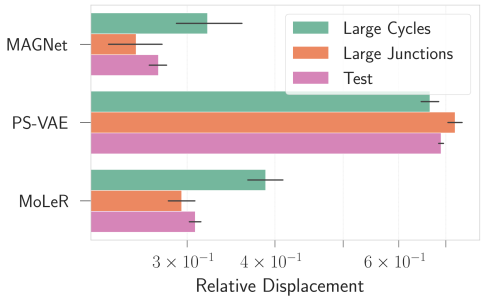

Figure 7: Displacement between latent representation of the input vs. the decoded output.

We calculate the Tanimoto similarity in the reconstruction setting for a variety of datasets, Fig. 8. For all evaluated datasets, MAGNet achieves the best similarity scores between molecules, highlighting the transferability of shapes across various distributions.

We compute the tanimoto scores only for those molecules that can be represented via the shapes that were extracted from the ZINC dataset. For the QM9 dataset, MAGNet can represent roughly 75% of the molecules in the dataset. This is due to unseen shapes which make up around 11% out of the total number of 289,966 shapes. For GuacaMol, MAGNet can represent around 97% of the molecules in the dataset. Out of the 9,562,028 shapes in GuacaMol, only 0.5% are missing from the shape vocabulary extracted from the ZINC dataset. We consider a more flexible decomposition of shapes that translates even better across datasets important future work.

D DETAILS ENCODER

We build the node features that are processed in MAGNet's encoder from different attributes, see Table 2. We include the atom type ('atom_id_dim'), its charge ('atom_charge_dim'), as well as its multiplicity value ('atom_multiplicity_dim'). We proceed accordingly for the shape level and include the shape id ('shape_id_dim'), its multiplicity ('shape_multiplicity_dim'), as well chemical features ('motif_feat_dim') computed through RDKit (Landrum, 2010). Since the latter are not learned during training, the features are mapped to the specified dimensionality by a linear map.

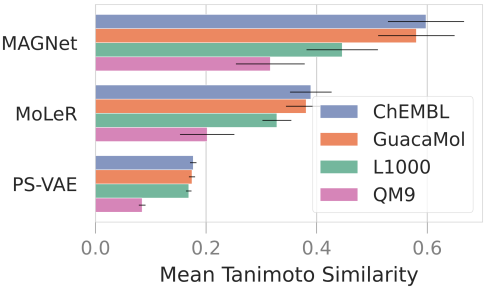

Figure 8: Tanimoto similarity for zero-shot reconstruction on unseen datasets.

After processing the resulting node features through the graph transformer (Shi et al., 2021) with 'num_layers_enc'-many layers, they are aggregated in different ways and mapped to specified dimensions as defined by 'enc_<>_dim' for the atoms, shapes, joins, and leaves, respectively. On top, the shape embeddings are additionally processed with the same transformer architecture ('num_layers_shape_enc') to inform the embedding about the shape-level connectivity. We then concatenate the resulting graph-level embeddings and further combine them with global molecule features, again computed via RDKit and then mapped to the required dimension ('enc_global_dim'), before mapping them to the latent space via the latent module which has 'num_layers_latent'-many layers.

E MAGNET: HYPERPARAMETERS AND TRAINING

Training MAGNet for one epoch takes around 30 minutes on a single 'NVIDIA GeForce GTX 1080 Ti'. We trained MAGNet for 30 epochs and fitted the latent normalizing flow post-hoc for 5000 epochs in total and conducted a random hyperparameter sweep including the learning rate, beta annealing scheme, and the number of layers for the encoder and latent module. The MAGNet model reported in the main text has 12.6 M parameters and its configuration is depicted in Table 2. In its current version, MAGNet processes roughly 70 molecules per second during training and samples about 8 molecules per second during inference.

Table 2: Parameter configuration of the best MAGNet runs.

	Parameter	Value
Train	batch_size	64
	flow_batch_size	1024
	lr	3.07×10^{-4}
	lr_sch_decay	0.9801
	flow_lr	1×10^{-3}
	flow_lr_sch_decay	0.99
	flow_patience	13
	gradclip	3
Model	latent_dim	100
dim_config	enc_atom_dim	25
	enc_shapes_dim	25
	enc_joins_dim	25
	enc_leaves_dim	25
	enc_global_dim	25
	atom_id_dim	25
	atom_charge_dim	10
	atom_multiplicity_dim	10
	shape_id_dim	35
	shape_multiplicity_dim	10
	motif_feat_dim	50
	shape_hidden	256
	shape_gnn_dim	128
	motif_seq_pos_dim	15
	leaf_hidden	256
	latent_flow_hidden	512

	Parameter	Value
Model	node_aggregation	sum
	num_layers_latent	2
	num_layers_enc	2
	num_layers_shape_enc	4
	num_layers_hgraph	3
loss_weights	joins	1
	leaves	1
	motifs	1
	hypergraph	1
beta_annealing	max	0.01
	init	0
	step	0.0005
	every	2500
	start	2000