An efficient graph generative model for navigating ultra-large combinatorial synthesis libraries

Anonymous Author(s)
Affiliation
Address
email

Abstract

Virtual, make-on-demand chemical libraries have transformed early-stage drug discovery by unlocking vast, synthetically accessible regions of chemical space. Recent years have witnessed rapid growth in these libraries from millions to trillions of compounds, hiding undiscovered, potent hits for a variety of therapeutic targets. However, they are quickly approaching a size beyond that which permits explicit enumeration, presenting new challenges for virtual screening. To overcome these challenges, we propose the **C**ombinatorial **S**ynthesis **L**ibrary **V**ariational **A**uto-**E**ncoder (**CSLVAE**). The proposed generative model represents such libraries as a differentiable, hierarchically-organized database. Given a compound from the library, the molecular encoder constructs a query for retrieval, which is utilized by the molecular decoder to reconstruct the compound by first decoding its chemical reaction and subsequently decoding its reactants. Our design minimizes autoregression in the decoder, facilitating the generation of large, valid molecular graphs. Our method performs fast and parallel batch inference for ultra-large synthesis libraries, enabling a number of important applications in early-stage drug discovery. Compounds proposed by our method are guaranteed to be in the library, and thus synthetically and cost-effectively accessible. Importantly, CSLVAE can encode out-of-library compounds and search for in-library analogues. In experiments, we demonstrate the capabilities of the proposed method in the navigation of massive combinatorial synthesis libraries.

1 Introduction

Virtual high throughput screening (vHTS) [45] has gained significant traction in early-stage drug discovery, owing in no small part to make-on-demand chemical libraries utilizing a *combinatorial synthesis* construction. These libraries enable access to ultra-large swaths of chemical space from a considerably smaller set of chemically accessible building blocks that can be combined according to known synthesis routines. In recent years, these libraries have grown from millions, to billions, and now to trillions of compounds [18, 31, 42, 55]. For example, the Enamine REadily AccessibLe (REAL) libraries [16] leverage off-the-shelf molecular building blocks and parallel synthesis, permitting lead times on the order of a few weeks and ushering in an era of ever-decreasing latency between *in silico* and *in vitro* high throughput screening.

As a result of the combinatorial explosion these constructions enable, early-stage drug discovery has now "crossed the Rubicon" into the non-enumerative regime. This presents new challenges for *in silico* hit discovery and optimization, which rely on screening explicitly enumerated compounds. These methods are ill suited to the non-enumerative setting, scaling linearly with the number of compounds, which motivates our interest in designing scalable approaches for navigating such libraries.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

In this paper, we propose the **C**ombinatorial **S**ynthesis **L**ibrary **V**ariational **A**uto-**E**ncoder, or **CSLVAE** for short (pronounced like *c'est la vie*). CSLVAE is a graph-based generative model that exploits the structure of combinatorial synthesis libraries (CSLs) towards efficient navigation of the relevant chemical space. Our model learns a hierarchy of keys for accessing the library, and uses these keys to process queries for retrieval. The encoder processes a molecular graph and returns as output a query into the library, which the decoder retrieves through an efficient sequence of query-key comparisons that utilizes the hierarchical construction of CSLs, requiring minimal autoregression and admitting efficient parallelization.

The main contributions of this paper are as follows:

- We present a novel graph generative model which acts as a "neural database," providing random access to ultra-large, non-enumerable compound libraries. As such, our model is guaranteed to generate valid and cost-effectively accessible molecules.
- Our model overcomes challenges with long autoregressive chains in compound generation, improving scalability to large molecular graphs.
- Our model reduces the number of parameters ten-fold relative to comparable methods, and offers considerable improvements in computational complexity for searching through CSLs.

2 Related Work

Virtual high throughput screening and enumeration Often, the first step in a vHTS campaign is preparing a compound library for subsequent use [1, 13]. While highly optimized compound sampling and scoring techniques have been developed [14], these approaches nevertheless rely on an exhaustively accessible library. An exception is the virtual synthon hierarchical enumeration screening (V-SYNTHES) approach [41], which leverages the modular nature of parallel synthesis libraries. However, by design, V-SYNTHES does not permit query-based random access. On the other hand, SpaceMACS [43] and SpaceLight [3] can provide query-based access to modular libraries by decomposing the query into fragments, and matching those by similarity search to synthons in the library. In parallel to these efforts, machine learning has received significant attention in vHTS: for predicting activity scores given docked conformations [11, 40, 54], predicting activity scores given a ligand and protein separately (undocked) [36, 53], and in improving or altogether replacing classical molecular docking with machine learning approaches [37, 48, 49].

Deep learning approaches to molecular generation De novo drug design has been playing an increasingly important role in identifying novel chemical matter in drug discovery campaigns [7, 34, 47, 51]. The two dominant neural network-based paradigms for molecular generation are text-based and graph-based generative models. Early work in text-based generative models applied recurrent neural networks to SMILES strings [12, 44]. Although these methods have shown a great deal of promise and spurred interest in molecular generation within the ML community, they are not guaranteed to produce valid SMILES strings. Approaches utilizing grammar constraints of SMILES notation have been proposed to improve validity [8, 26]; separately the recently proposed SELFIES notation [25, 35] guarantees validity and has seen increased adoption as such. In both cases, however, there remain known drawbacks to modeling with such text-based representations of chemical matter (e.g., surjectivity, similar molecular structures having possibly large edit distances).

Graph-based generative models have received significant attention in recent years as an alternative to their text-based counterparts. The earliest of these models focused on generating graphs of a constant size in a single shot [46] or generating graphs of arbitrary size autoregressively, one atom or bond at a time [30, 39, 57]. These approaches also struggle to reliably produce chemically valid molecules—prompting the use of expensive but explicit validity checks during generation [33]—and encounter difficulties with large molecular graphs. In an effort to address both points, fragment-based approaches have been proposed and are growing in popularity [19, 21, 22, 24]. However, some issues persist due to general difficulties with autoregressive graph generation. Unlike text-based models, in which there is less ambiguity in the autoregression order (e.g., tokens are typically decoded in left-to-right order), graphs have no such canonical node order, which has presented challenges for graph autoencoders [30, 56]. Furthermore, both text- and graph-based generative models struggle with large molecular graphs due to the long autoregression chains required, which has been an active area of focus in the literature.

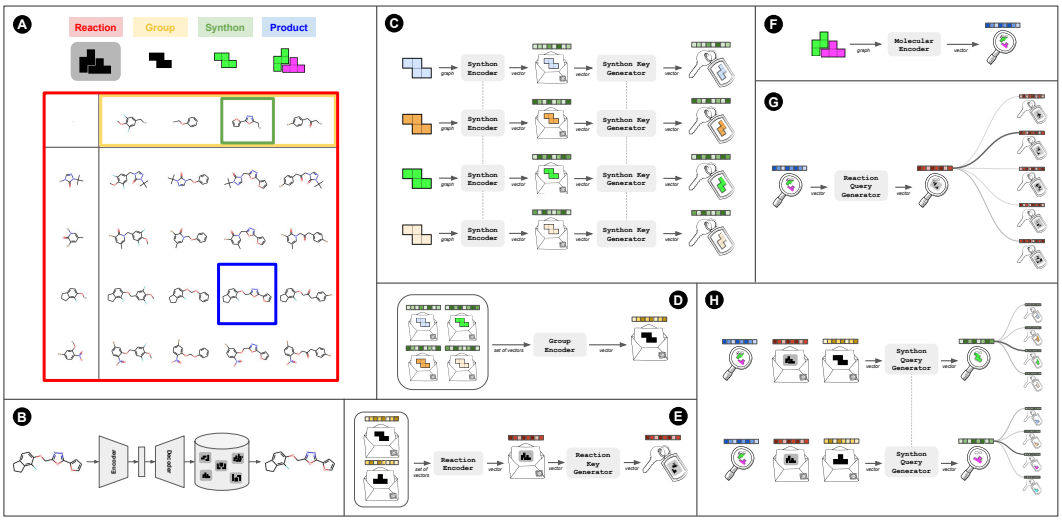

Figure 1: **Overview of our proposed method. Panel A:** Illustration of combinatorial synthesis libraries. **Panel B:** CSLVAE is an autoencoder which takes a molecular graph as input, produces a query into a library, and decodes a product from the library as output. **Panels C-E:** The library encoder of CSLVAE first encodes synthons, represented as molecular graphs, then learns a representation for the groups via a set neural network over the synthon representations, and finally learns a representation for the reactions via a set neural network over the group representations. Keys for the synthons and reactions are learned from their respective representations via MLPs. **Panel F:** The molecular encoder returns a query from an input molecular graph (in this case, a product from the library). **Panels G-H:** The molecular decoder processes the query to decode a product from the library, by first generating the reaction type and subsequently generating one synthon per group via query-key lookups.

Validity aside, there is the separate challenge of synthetic accessibility. Prior work has cast doubt on the synthetic feasibility of compounds proposed by many existing generative models [9]. Subsequent work has attempted to address these shortcomings by (i) including explicit penalties for synthetic inaccessibility via a scoring function [15], (ii) replacing fragments with those from known compounds [32, 38, 50], or (iii) inducing bias towards simple and known synthetic pathways [5, 6, 17].

3 Methodology

This section formalizes combinatorial synthesis libraries and our proposed model, CSLVAE. Figure 1 illustrates our approach.

3.1 Preliminaries

CSLs are composed of a set of multi-dimensional synthesis tables. Each synthesis table describes a multi-component *reaction*, denoted by $t \in T$. Following [42], we use the term "R-group," or *group* for short, to refer to a singular component in a reaction, and represent it by $r \in R$. We let $\psi : T \to \mathcal{P}(R)$ be the function which returns the set of groups $\psi(t) \subset R$ associated with a reaction t.

Each group is spanned by a possibly large number of molecular building blocks, called *synthons*, which can be utilized in the corresponding reaction. We denote a synthon by $s \in S$ and represent it with a molecular graph, $\mathcal{G}_s \in \mathscr{G}_S$. We note that a synthon can belong to multiple groups in R. For convenience, we use $\sigma : R \to \mathcal{P}(S)$ to represent the function which returns the set of synthons $\sigma(r) \subset S$ belonging to a particular group r.

A *product* $x \in X$ is a molecule that is synthesized according to a multi-component reaction with groups $\psi(t) = (r_t^{(1)}, \ldots, r_t^{(k_t)})$ and an associated synthon chain $u = (s_u^{(1)}, \ldots, s_u^{(k_t)})$, where $s_u^{(i)} \in \sigma(r_t^{(i)})$ for all $i = 1, \ldots, k_t$. We define f to be the synthesis rule which generates a compound x from a reaction and synthon chain pair (t, u), i.e., $x := f(t, u)$.

Hence, a CSL $\mathcal{D} = (S, R, T)$ is fully characterized by its synthons S, groups R, and reactions T, together with the synthesis rule f, reaction-to-groups mapping ψ, and group-to-synthons mapping σ.

Using the language of probability, one can describe a distribution over products induced by $\mathcal{D}$ via the following factorization:

$$p(x|\mathcal{D}) = \sum_{t \in T} \sum_{u \in U_t} p(t|\mathcal{D})\, p(u|t, \mathcal{D})\, p_f(x|t, u), \tag{1}$$

where $U_t = \sigma(r_t^{(1)}) \times \ldots \times \sigma(r_t^{(k_t)})$ is the set of all eligible synthon chains for reaction t. This factorization describes the generative process in which one first samples a reaction $t \sim p(t|\mathcal{D}) \propto |U_t|$, then samples a valid synthon chain $u \sim p(u|t, \mathcal{D}) = |U_t|^{-1}$ comprised of synthons from the respective groups in t, and joins these together via synthesis to form a product x (via the rule f).

As written, all valid (t, u) pairs in $\mathcal{D}$ are equally probable under p. Note that if every product in $\mathcal{D}$ can be reached according to just a single synthesis route, then $p(x|\mathcal{D})$ is a uniform distribution over the part of X accessible by $\mathcal{D}$.

3.2 Combinatorial Synthesis Library Variational Auto-Encoder (CSLVAE)

We consider the task of looking up a product x from a library $\mathcal{D}$, which amounts to finding the reaction t and synthon chain u satisfying $x = f(t, u)$. This can be cast as an inference problem seeking $p(t, u|x, \mathcal{D})$. A latent variable model for x gives rise to a variational formulation:

$$p(t, u|x, \mathcal{D}) = \int_{\mathbf{z}} p(t, u|\mathbf{z}, \mathcal{D})\, q(\mathbf{z}|x)\, \mathrm{d}\mathbf{z}, \tag{2}$$

with $p(\mathbf{z})$ denoting the prior distribution of the latent variable. One can further simplify the joint conditional distribution of t and u by first selecting the reaction t and then selecting the synthons $s_u^{(1)}, \ldots, s_u^{(k_t)}$ independently conditional on t:

$$p(t, u|\mathbf{z}, \mathcal{D}) = p(t|\mathbf{z}, \mathcal{D})\, p(u|\mathbf{z}, t, \mathcal{D}) \tag{3}$$

$$= p(t|\mathbf{z}, \mathcal{D}) \prod_{i=1}^{k_t} p(s_u^{(i)}|\mathbf{z}, t, r_t^{(i)}, \mathcal{D}). \tag{4}$$

This gives rise to a strategy in which one encodes a molecule x into the latent space, $\mathbf{z} \sim q(\mathbf{z}|x)$, and proceeds by first decoding the reaction type $t \sim p(t|\mathbf{z}, \mathcal{D})$ and then, conditional on the sampled reaction, decoding one synthon per group $s_u^{(i)} \sim p(s_u^{(i)}|\mathbf{z}, t, r_t^{(i)}, \mathcal{D})$ for $i = 1, \ldots, k_t$ to form the synthon chain $u \in U_t$.

The resulting latent variable model is the proposed CSLVAE. Panels B-H in Figure 1 give a step-by-step depiction of the method. In the forthcoming subsections, we will describe and formalize the three primary modules that comprise CSLVAE in detail: (i) the library encoder, (ii) the molecular encoder, and (iii) the molecular decoder. Module (ii) forms the basis for $q(\mathbf{z}|x)$, while modules (i) and (iii) form the basis for $p(t, u|\mathbf{z}, \mathcal{D})$.

3.2.1 Library encoder

A CSL $\mathcal{D} = (S, R, T)$ is organized hierarchically, with synthons S at the bottom of the hierarchy, groups R in the middle, and reactions T at the top. We propose a strategy for learning an associated hierarchy of representations that describe the library at these three levels of resolution in an end-to-end fashion. These representations can then be used in retrieving results from queries into the library. The library encoder is illustrated in panels C-E of Figure 1.

We start from the bottom of the hierarchy with the synthons S. To learn a representation for each synthon in a manner that is fully inductive, we use a graph neural network to parameterize the `SynthonEncoder` : $\mathscr{G}_S \to \mathbb{R}^{d_S}$, which applies a sequence of message passing steps followed by a readout to arrive at a d_S-dimensional representation for each synthon.

Moving up the hierarchy, we use a deep set neural network to represent the groups R by summarizing the representations for the synthons belonging to a particular group. Formally, the `GroupEncoder` :

$\mathbb{R}^{d_S} \times \cdots \times \mathbb{R}^{d_S} \to \mathbb{R}^{d_R}$ learns a d_R-dimensional representation for a given group from the set of d_S-dimensional representations of the constituent synthons. Following [58], the group encoder has the form $\texttt{GroupEncoder}(\{\mathbf{h}_s^S : \forall s \in \sigma(r)\}) = \rho\left(\bigoplus_{s\in\sigma(r)} \phi(\mathbf{h}_s^S)\right)$, where ϕ and ρ are parameterized by multi-layer perceptrons (MLPs) and $\bigoplus$ is a permutation-invariant aggregation operator. While other potentially more performant set-to-vector neural networks could be considered here, e.g., set transformers [29], an advantage of this simple construction is that it permits fast querying over library subsets at test time since $\phi(\mathbf{h}_s^S)$ can be cached for all $s \in S$, thereby reducing the required computations for making queries into partitions of the library. We utilize mean pooling as the aggregation operator to focus on characteristics of the distribution of synthons in a group (as opposed to sum pooling which generally expresses characteristics of the multiset), which improves performance when dealing with groups of varying cardinality.

Finally, to represent reactions, we use yet another deep set neural network as the $\texttt{ReactionEncoder} : \mathbb{R}^{d_R} \times \cdots \times \mathbb{R}^{d_R} \to \mathbb{R}^{d_T}$. This module takes as input a variable-sized set of group representations corresponding to the reactants in t and produces a d_T-dimensional representation for the reaction. We utilize sum pooling as the aggregation operator to learn multiset properties of the groups in a reaction.

Putting these together, the representations cascade as follows:

$$\mathbf{h}_s^S = \texttt{SynthonEncoder}(\mathcal{G}_s), \tag{5}$$

$$\mathbf{h}_r^R = \texttt{GroupEncoder}(\{\mathbf{h}_s^S : \forall s \in \sigma(r)\}), \tag{6}$$

$$\mathbf{h}_t^T = \texttt{ReactionEncoder}(\{\mathbf{h}_r^R : \forall r \in \psi(t)\}). \tag{7}$$

In (4), we considered a factorization of the likelihood such that, given a molecular representation $\mathbf{z}$, the molecular decoder proceeds by first decoding the reaction type and then decoding one synthon for each group separately conditional on the reaction type. Hence, we require a key vector for each reaction as well as for each synthon to compare with the associated query vectors (to be described). As such, we introduce a $\texttt{ReactionKeyGenerator} : \mathbb{R}^{d_T} \to \mathbb{R}^{k_T}$ which returns a key vector given a reaction representation, and similarly introduce a $\texttt{SynthonKeyGenerator} : \mathbb{R}^{d_S} \to \mathbb{R}^{k_S}$ to produce a key for each synthon. Each of these key generators can be parameterized by MLPs:

$$\mathbf{k}_t^T = \texttt{ReactionKeyGenerator}(\mathbf{h}_t^T), \tag{8}$$

$$\mathbf{k}_s^S = \texttt{SynthonKeyGenerator}(\mathbf{h}_s^S). \tag{9}$$

3.2.2 Molecular encoder

The $\texttt{MolecularEncoder} : \mathscr{G}_X \to \mathbb{R}^{d_X}$ takes as input a molecular graph $\mathcal{G}_x \in \mathscr{G}_X$ and returns a d_X-dimensional feature representation,

$$\mathbf{z} = \texttt{MolecularEncoder}(\mathcal{G}_x). \tag{10}$$

In our implementation, the $\texttt{MolecularEncoder}$ is a graph neural network with a variational linear layer stacked on top of the readout, which produces a sample $\mathbf{z} \sim q(\mathbf{z}|x)$ for a given input graph $\mathcal{G}_x$. As depicted in Figure 1, we interpret the representation $\mathbf{z}$ as a query induced by x into the library $\mathcal{D}$.

Note that the $\texttt{MolecularEncoder}$ can in principle take any valid molecular graph as input is therefore capable of producing queries for compounds that are *not* in the library $\mathcal{D}$. This can be useful for finding *analogues by catalog* – that is, compounds which can be purchased from a catalog and are chemical analogues of a query molecule. This use case will be investigated in a later section.

3.2.3 Molecular decoder

Given a query $\mathbf{z} = \texttt{MolecularEncoder}(\mathcal{G}_x)$ and a library $\mathcal{D}$, the decoder is tasked with retrieving the molecule from the library, i.e., identifying the reaction and synthon chain which yield the molecule as a product. We proceed by generating the reaction first. A $\texttt{ReactionQueryGenerator} : \mathbb{R}^{d_X} \to \mathbb{R}^{k_T}$ generates a query from the molecular representation to compare against the reaction keys:

$$\mathbf{q}^T = \texttt{ReactionQueryGenerator}(\mathbf{z}), \tag{11}$$

$$\left\{p(t_j|\mathbf{z},\mathcal{D})\right\}_{j=1}^{|T|} = \text{softmax}\left(\left\{\frac{\mathbf{q}^T \cdot \mathbf{k}_{t_j}^T}{\sqrt{k_T}}\right\}_{j=1}^{|T|}\right). \tag{12}$$

This defines a probability distribution over reaction types in T. We can sample according to this distribution to arrive at a reaction $t \sim p(t|\mathbf{z}, \mathcal{D})$.

Given the sampled reaction t, we know the required groups (via ψ) and furthermore know the synthons which are eligible for each group (via σ). To decode the synthon chain, we introduce a `SynthonQueryGenerator` : $\mathbb{R}^{d_X+d_T+d_R} \to \mathbb{R}^{k_S}$ which can be used to query synthons for each group $r_t^{(i)} \in \psi(t)$ as follows:

$$\mathbf{q}_{t,r}^S = \texttt{SynthonQueryGenerator}([\mathbf{z} \| \mathbf{h}_t^T \| \mathbf{h}_r^R]), \tag{13}$$

$$\left\{ p(s_j^{(i)} | \mathbf{z}, t, r_t^{(i)}, \mathcal{D}) \right\}_{j=1}^{|\sigma(r_t^{(i)})|} = \text{softmax} \left(\left\{ \frac{\mathbf{q}_{t,r_t^{(i)}}^S \cdot \mathbf{k}_{s_j^{(i)}}^S}{\sqrt{k_S}} \right\}_{j=1}^{|\sigma(r_t^{(i)})|} \right). \tag{14}$$

In our implementation, both the reaction and synthon query generators are parameterized by MLPs.

3.3 Training algorithm

In each training iteration, we sample a random subset from the library $\tilde{\mathcal{D}} \subset \mathcal{D}$ according to a distribution $p(\tilde{\mathcal{D}}|\mathcal{D})$ and use this subset in forming the synthon, group, and reaction representations and keys. In particular, we use a $p(\tilde{\mathcal{D}}|\mathcal{D})$ which first samples a subset of the reactions in $\mathcal{D}$ uniformly at random and then, for each reaction, samples an equal number of products also uniformly at random, including the resulting synthons in $\tilde{\mathcal{D}}$. We also utilize teacher forcing, feeding in the ground truth reaction when generating the synthon queries for the respective groups. Algorithm 1 in the supplementary materials describes the training procedure.[1]

Ex-post density estimation Given a trained generative model $p_\theta(x|\mathbf{z})$, we wish to sample products via $(x, \mathbf{z}) \sim p_\theta(x|\mathbf{z})\, p(\mathbf{z})$ (discarding $\mathbf{z}$).[2] However, this in general will not correspond well to a uniform distribution over the products in $\mathcal{D}$ due to the bias introduced by the batch sampling strategy outlined above.

Although this can be corrected with importance weighting during the training phase, we opt for a more practical approach by using an ex-post density estimation strategy [10]. We sample a large number of products from the target distribution $x \sim p(x|\mathcal{D})$ and encode these products via the molecular encoder $\mathbf{z} \sim q_\phi(\mathbf{z}|x)$. We then fit a density estimator to the aggregated samples, written $q_\lambda(\mathbf{z})$. In our experiments, we utilize a multivariate normal distribution for simplicity, but one could imagine using more expressive density estimators here (e.g., a mixture of multivariate normals).

Now, we can sample products via $(x, \mathbf{z}) \sim p_\theta(x|\mathbf{z})\, q_\lambda(\mathbf{z})$, which will more closely align with sampling from $p(x|\mathcal{D})$. This helps to correct for bias in the distribution over product space that is induced by the choice of batch sampling strategy.

3.4 Computational complexity, scalability, and efficiency

We now examine the computational complexity, scalability, and efficiency of the proposed method.

First, we note that the library $\mathcal{D}$ can be encoded with $O(|S| + |R| + |T|)$ complexity – the constant depends on the complexity of the synthon, group, and reaction encoders. Nonetheless, this is logarithmic in comparison to naively encoding each product in $\mathcal{D}$, which has $O(|\mathcal{D}|)$ complexity.

More noteworthy is the computational complexity of the molecular decoder. For clarity, let us consider a simplified $\mathcal{D}$ comprised of a single k-component reaction. Let M_i denote the number of synthons for group $i = 1, \ldots, k$. Naively, a nearest neighbor lookup in $\mathcal{D}$ requires $O(\prod_{i=1}^k M_i)$ complexity. CSLVAE, on the other hand, performs the lookup over synthons in each group independently, which has $O(\sum_{i=1}^k M_i)$ complexity – a logarithmic improvement. Hence, the proposed molecular decoder is highly suitable for ultra-large combinatorial libraries that are of interest in early-stage drug discovery.

[1] To keep notation consistent between Algorithm 1 and the equations it references, the input library to the algorithm is written as $\mathcal{D}^*$ (instead of $\mathcal{D}$) and the subsets are written as $\mathcal{D}$ (instead of $\tilde{\mathcal{D}}$).

[2] The parameter θ represents the parameters for the modules written in `typewriter` font.

Another advantage of CSLVAE's decoding strategy is that it relies only minimally on autoregression. In fact, the length of the autoregression is *exactly* two, irrespective of the size of the graph being generated. As such, CSLVAE gracefully scales to large and variable-sized molecular graphs that follow a combinatorial synthesis construction.

Lastly, we point out that our method is guaranteed to generate chemically valid—and perhaps more importantly, synthetically accessible—molecular graphs without performing explicit validity checks. This compares favorably with prior work, in which the validity of each candidate action is verified at each step of the autoregression, with invalid actions excluded from the choice set. Although cheminformatics libraries like RDKit [28] have efficient C++ implementations for these checks, they nonetheless increase runtime rather significantly. Further, in the absence of explicit validity checks, these models have been shown to generate invalid molecular graphs at a markedly higher rate [19].

4 Experiments

Data We demonstrate the capabilities of CSLVAE on the Enamine REAL library, which is comprised of 340K synthons and over one thousand reactions. The reactions in REAL range from two to four components and the number of synthons per R-group range from the single digits to tens of thousands. In total, the REAL library describes a chemical space of over 16 billion commercially available compounds.[3]

Training During training, we sample subsets of the library as follows. Of the roughly 1300 reaction types in the REAL database, we first sample 20 reactions independently, and subsequently sample 100 products per reaction, including the associated synthons in the library subset. These library subsets therefore describe roughly 300K-1.5M compounds each, which is significantly smaller than the complete library of 16B compounds. See Algorithm 1 (supplementary materials) for details.

Testing For test-time inference, we decode with respect to the full library of 16B compounds. This constitutes a test-time distribution shift relative to training, but we observed that CSLVAE generalizes remarkably well to the full library without modifications. For completeness, we include an analysis of the test-time distribution shift in the supplementary materials. In the forthcoming analyses, we share results when performing inference on the full library, as this is our primary objective.

4.1 Molecular reconstruction and generation

Table 1: Comparison of RationaleRL, JT-VAE, and CSLVAE (ours) on synthon-based generative modeling

	JT-VAE	RationaleRL	CSLVAE (ours)
# Parameters	4.7M	3.4M	380K
Validity	100.0%	100.0%	100.0%
Uniqueness	80.1%	96.3%	98.8%
Average likelihood	18.7%	62.3%	72.4%
In-library proportion	2.9%	50.9%	100.0%

We compare CSLVAE against two state-of-the-art molecular graph generative models: JT-VAE [20] and RationaleRL [22]. All three models were trained from scratch on the Enamine REAL library. Details on the experimental setup and architecture can be found in the supplementary materials.

In JT-VAE, molecular graphs are represented by junction trees over chemical fragments. Decoding proceeds by first generating the junction tree in a depth-first manner, placing a fragment in each node, and then subsequently orienting the fragments to match attachment points. RationaleRL, on the other hand, takes as input a starting chemical fragment or *rationale*. The decoder's objective is to complete the molecule in an autoregressive fashion (one graph edit per step). In our setting, we take a product from the library and remove all but one synthon, treating the resulting graph as the starting rationale. Thus, RationaleRL is tasked with generating the missing synthon.

[3] With permission from and attribution to Enamine Ltd., we release a subset of the library alongside our code for reproducing our experiments and to foster research in the machine learning community on combinatorial synthesis libraries.

Table 1 summarizes our key findings. First, we note that our implementation of CSLVAE has roughly 10x fewer parameters than the two alternatives considered, owing to the inductive nature of the library encoder. All three methods achieve 100% chemical validity, but CSLVAE achieves this result without explicit validity checks. The average likelihood is computed by taking the average of per-compound reconstruction likelihoods across a large number of products sampled from the library. This is a measure of how well the model is capable of reconstructing the full molecular graph (i.e., on average, how likely are we to reproduce the query molecule via the decoder) and can also loosely be interpreted as a measure of coverage/reachability (i.e., what proportion of the library are we able to faithfully cover). Finally, we highlight the challenges existing graph generative models face when applied to ultra-large CSLs, namely that they struggle to reliably generate in-library compounds. For JT-VAE, fewer than 1 in 34 compounds were found in REAL. Even though RationaleRL takes as input a product from REAL stripped of all but one synthon, it generated in-library completions only about half of the time (see Section A.3). In contrast, CSLVAE is guaranteed to stay in the library by design.

4.2 Latent space visualizations

Next, we qualitatively inspect the latent space learned by CSLVAE. In particular, we are interested in verifying whether the proposed model has learned a latent space which varies relatively smoothly over the covered chemical space (i.e., that small perturbations to the query induce only minor edits in the resulting molecular graph). We perform two kinds of checks: latent space interpolations and local neighborhood visualizations.

Panel A of Figure 2 contains an example of interpolations in the latent space. In particular, we interpolate the molecular queries for the starting compound (top left) and target compound (bottom right), in raster order. The molecules immediately adjacent to the starting and target compounds are the associated reconstructions. Products are decoded with respect to the full REAL library of 16B compounds. Below each molecule are its Tanimoto similarity[4] with the starting and target molecule, respectively. We observe that the interpolations traverse through regions of chemical space that gradually decrease (cf. increase) in similarity with respect to the starting (cf. target) compound.

Panel B of Figure 2 visualizes the latent space around a randomly sampled product from the REAL library. Following prior work [27], we form a random 2D plane in the high-dimensional latent space by sampling two random directions around the molecular query (center compound) and decoding the resulting products using the argmax decision rule. We observe that the latent space is smooth in the sense that molecules morph gradually, with only minor edits when the movement in latent space is small (e.g., one synthon at a time, modifications to smaller functional groups), and that the molecular scaffold is generally conserved locally.

4.3 Autoencoding as analogue search

Finally, we utilize CSLVAE to find analogues of a query compound in a large CSL. We present the model with two molecules: one which is in the library (Figure 3, left) and one which is not in the library (right). Given the molecular query, we generate conditionally random completions from the decoder. For the in-library example, we note that the model is able to successfully retrieve the query compound (row 2, column 1). Moreover, we find a number of instances in which the model returned compounds that have the same reaction type as the query molecule, but vary in one or two synthons, or returns the query molecule up to an addition/removal of a functional group. For the out-of-library example, we note that CSLVAE finds structurally similar compounds with high Tanimoto similarity, indicating the promise of this approach for fast analogue search in large (non-enumerable) libraries.

5 Closing remarks

We proposed CSLVAE, a new graph-based generative model for the navigation of combinatorial synthesis libraries. CSLVAE utilizes minimal autoregression, permitting efficient generation of large molecular graphs and improving scalability. Compounds generated by CSLVAE are chemically valid and cost-effectively accessible. CSLVAE is a neural database providing random access to non-enumerable libraries. In experiments, we demonstrate the capabilities of CSLVAE in modeling

[4]The Tanimoto similarity [2] is calculated as an intersection-over-union between a pair of bit vectors describing each molecule using a hashing-based technique called a *molecular fingerprint* [4].

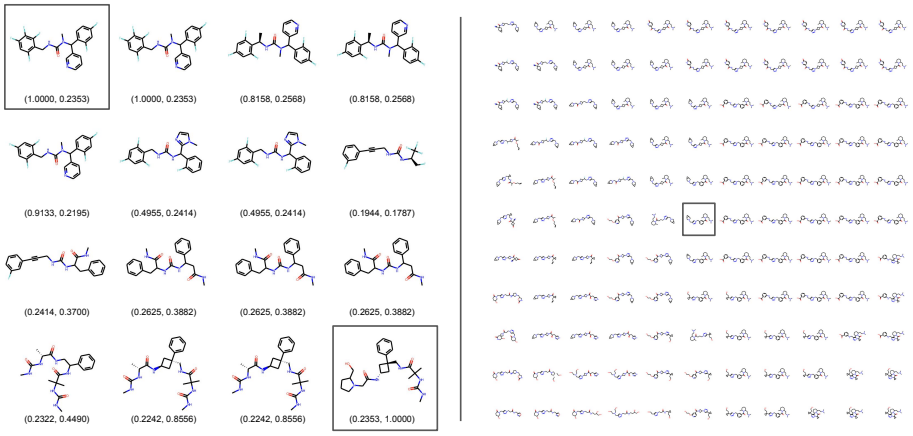

Figure 2: **Latent space visualizations. Panel A:** Moving left to right in raster order, we linearly interpolate from the starting compound to the target compound. The immediately adjacent molecules are reconstructions. Below each molecule are its Tanimoto similarities with the starting and target compound, respectively. **Panel B:** We sample two random directions in the latent space around a query compound and visualize the decoded molecules spaced evenly on the resulting 2D plane.

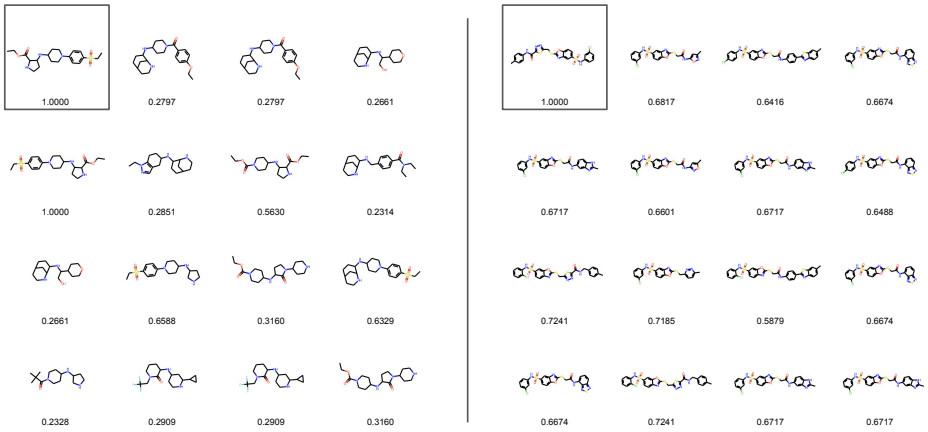

Figure 3: **Autoencoding as analogue search.** Each query molecule (top left) is autoencoded fifteen times. Below each molecule is its Tanimoto similarity with the query molecule. **Left:** Autoencoding a molecule contained in the library. **Right:** Autoencoding a molecule not contained in the library.

ultra-large and realistic make-on-demand libraries, paving a path towards more scalable strategies in the exploration of non-enumerable chemical libraries for early-stage drug discovery. Our method can be combined with established techniques for molecular optimization as a future direction.

Of course, our approach has its limitations; here, we highlight three. First, the synthon lookup in the decoder scales linearly with the number of synthons in a group, which can present challenges as libraries continue to add many synthons per group. This could be mitigated by using more scalable query-key designs, such as those utilizing locality-sensitive hashing [23], but is not considered here. Furthermore, CSLVAE may face difficulties in the presence of prominent group symmetry (e.g., polymers). The decoder would require modifications to break parity, but may not admit the same convenient parallelization. Lastly, softmax has limitations in mapping from real-valued potentials to choice probabilities due to its rigid substitution patterns [52]; sparse alternatives may be preferable, but are not considered in this paper. In future work, we intend to address these shortcomings and look to applications in virtual high throughput screening.

References

[1] Atanu Acharya, Rupesh Agarwal, Matthew B Baker, Jerome Baudry, Debsindhu Bhowmik, Swen Boehm, Kendall G Byler, SY Chen, Leighton Coates, Connor J Cooper, et al. Supercomputer-based ensemble docking drug discovery pipeline with application to covid-19. *Journal of chemical information and modeling*, 60(12):5832–5852, 2020.

[2] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

[3] Louis Bellmann, Patrick Penner, and Matthias Rarey. Topological similarity search in large combinatorial fragment spaces. *Journal of Chemical Information and Modeling*, 61(1):238–251, 2020.

[4] Andreas Bender and Robert C Glen. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2(22):3204–3218, 2004.

[5] John Bradshaw, Brooks Paige, Matt J. Kusner, Marwin H. S. Segler, and José Miguel Hernández-Lobato. A model to search for synthesizable molecules, 2019.

[6] John Bradshaw, Brooks Paige, Matt J. Kusner, Marwin H. S. Segler, and José Miguel Hernández-Lobato. Barking up the right tree: an approach to search over molecule synthesis dags, 2020.

[7] Hongming Chen. Can generative-model-based drug design become a new normal in drug discovery? *Journal of Medicinal Chemistry*, 65(1):100–102, December 2021.

[8] Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786*, 2018.

[9] Wenhao Gao and Connor W. Coley. The synthesizability of molecules proposed by generative models. *Journal of Chemical Information and Modeling*, 60(12):5714–5723, April 2020.

[10] Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019.

[11] Pawel Gniewek, Bradley Worley, Kate Stafford, Henry van den Bedem, and Brandon Anderson. Learning physics confers pose-sensitivity in structure-based virtual screening. *arXiv preprint arXiv:2110.15459*, 2021.

[12] 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.

[13] Christoph Gorgulla, Andras Boeszoermenyi, Zi-Fu Wang, Patrick D Fischer, Paul W Coote, Krishna M Padmanabha Das, Yehor S Malets, Dmytro S Radchenko, Yurii S Moroz, David A Scott, et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, 580(7805):663–668, 2020.

[14] David E Graff, Eugene I Shakhnovich, and Connor W Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chemical science*, 12(22):7866–7881, 2021.

[15] Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained bayesian optimization for automatic chemical design using variational autoencoders. *Chem. Sci.*, 11:577–586, 2020.

[16] Oleksandr O Grygorenko, Dmytro S Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E Gubina, and Yurii S Moroz. Generating multibillion chemical space of readily accessible screening compounds. *Iscience*, 23(11):101681, 2020.

[17] Julien Horwood and Emmanuel Noutahi. Molecular design in synthetically accessible chemical space via deep reinforcement learning. *ACS Omega*, 5(51):32984–32994, December 2020.

[18] John J. Irwin and Brian K. Shoichet. Docking screens for novel ligands conferring new biology. *Journal of Medicinal Chemistry*, 59(9):4103–4120, March 2016.

[19] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation, 2018.

[20] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

[21] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs, 2020.

[22] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In *International conference on machine learning*, pages 4849–4859. PMLR, 2020.

[23] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

[24] Xiangzhe Kong, Zhixing Tan, and Yang Liu. Graph piece: Efficiently generating high-quality molecular graphs with substructures, 2021.

[25] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[26] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International conference on machine learning*, pages 1945–1954. PMLR, 2017.

[27] Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder, 2017.

[28] Greg Landrum. Rdkit: Open-source cheminformatics, 2006.

[29] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[30] Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander Gaunt. Constrained graph variational autoencoders for molecule design. *Advances in neural information processing systems*, 31, 2018.

[31] Jiankun Lyu, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, Andrey A. Tolmachev, Brian K. Shoichet, Bryan L. Roth, and John J. Irwin. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, February 2019.

[32] 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, 2021.

[33] 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*, 2(2):025023, March 2021.

[34] Joshua Meyers, Benedek Fabian, and Nathan Brown. De novo molecular design and generative models. *Drug Discovery Today*, 26(11):2707–2715, November 2021.

[35] AkshatKumar Nigam, Robert Pollice, Mario Krenn, Gabriel dos Passos Gomes, and Alan Aspuru-Guzik. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (stoned) algorithm for molecules using selfies. *Chemical science*, 12(20):7079–7090, 2021.

[36] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[37] Joseph M Paggi, Julia A Belk, Scott A Hollingsworth, Nicolas Villanueva, Alexander S Powers, Mary J Clark, Augustine G Chemparathy, Jonathan E Tynan, Thomas K Lau, Roger K Sunahara, et al. Leveraging nonstructural data to predict structures and affinities of protein–ligand complexes. *Proceedings of the National Academy of Sciences*, 118(51), 2021.

[38] Pavel Polishchuk. Control of synthetic feasibility of compounds generated with CReM. *Journal of Chemical Information and Modeling*, 60(12):6074–6080, November 2020.

[39] Mariya Popova, Mykhailo Shvets, Junier Oliva, and Olexandr Isayev. Molecularrnn: Generating realistic molecular graphs with optimized properties, 2019.

[40] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, 57(4):942–957, 2017.

[41] Arman A Sadybekov, Anastasiia V Sadybekov, Yongfeng Liu, Christos Iliopoulos-Tsoutsouvas, Xi-Ping Huang, Julie Pickett, Blake Houser, Nilkanth Patel, Ngan K Tran, Fei Tong, et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature*, 601(7893):452–459, 2022.

[42] Arman A. Sadybekov, Anastasiia V. Sadybekov, Yongfeng Liu, Christos Iliopoulos-Tsoutsouvas, Xi-Ping Huang, Julie Pickett, Blake Houser, Nilkanth Patel, Ngan K. Tran, Fei Tong, Nikolai Zvonok, Manish K. Jain, Olena Savych, Dmytro S. Radchenko, Spyros P. Nikas, Nicos A. Petasis, Yurii S. Moroz, Bryan L. Roth, Alexandros Makriyannis, and Vsevolod Katritch. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature*, 601(7893):452–459, December 2021.

[43] Robert Schmidt, Raphael Klein, and Matthias Rarey. Maximum common substructure searching in combinatorial make-on-demand compound spaces. *Journal of Chemical Information and Modeling*, 2021.

[44] 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.

[45] Brian K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, December 2004.

[46] Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. In *International conference on artificial neural networks*, pages 412–422. Springer, 2018.

[47] Tiago Sousa, João Correia, Vítor Pereira, and Miguel Rocha. Generative deep learning for targeted compound design. *Journal of Chemical Information and Modeling*, 61(11):5343–5361, October 2021.

[48] Kate A Stafford, Brandon M Anderson, Jon Sorenson, and Henry van den Bedem. Atomnet poseranker: Enriching ligand pose quality for dynamic proteins in virtual high-throughput screens. *Journal of chemical information and modeling*, 62(5):1178–1189, 2022.

[49] Hannes Stärk, Octavian-Eugen Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. *arXiv preprint arXiv:2202.05146*, 2022.

[50] Kosuke Takeuchi, Ryo Kunimoto, and Jürgen Bajorath. R-group replacement database for medicinal chemistry. *Future Science OA*, 7(8), September 2021.

[51] Xiaochu Tong, Xiaohong Liu, Xiaoqin Tan, Xutong Li, Jiaxin Jiang, Zhaoping Xiong, Tingyang Xu, Hualiang Jiang, Nan Qiao, and Mingyue Zheng. Generative models for de novo drug design. *Journal of Medicinal Chemistry*, 64(19):14011–14027, September 2021.

[52] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

[53] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, 2019.

[54] Izhar Wallach, Michael Dzamba, and Abraham Heifets. Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*, 2015.

[55] Wendy A. Warr, Marc C. Nicklaus, Christos A. Nicolaou, and Matthias Rarey. Exploration of ultralarge compound collections for drug discovery. *Journal of Chemical Information and Modeling*, April 2022.

[56] Robin Winter, Frank Noé, and Djork-Arné Clevert. Permutation-invariant variational autoencoder for graph-level representation learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[57] Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation, 2018.

[58] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will release code, data, and instructions by the camera-ready deadline
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We include this information in the supplementary materials.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We include this information in the supplementary materials.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] We include this information in the supplementary materials.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] For the data that we intend to release alongside our code, we have received explicit permission from the vendor (Enamine) to share a subset of the full library for reproducibility experiments.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]