nach0-pc: Multi-task Language Model with Molecular Point Cloud Encoder

Maksim Kuznetsov [1*], Airat Valiev [2], Alex Aliper [2], Daniil Polykovskiy [1], Elena Tutubalina [2], Rim Shayakhmetov [1], Zulfat Miftahutdinov [1]

[1]Insilico Medicine Canada Inc.,
[2]Insilico Medicine AI Ltd.

Abstract

Recent advancements have integrated Language Models (LMs) into a drug discovery pipeline. However, existing models mostly work with SMILES and SELFIES chemical string representations, which lack spatial features vital for drug discovery. Additionally, attempts to translate chemical 3D structures into text format encounter issues such as excessive length and insufficient atom connectivity information. To address these issues, we introduce nach0-pc, a model combining domain-specific encoder and textual representation to handle spatial arrangement of atoms effectively. Our approach utilizes a molecular point cloud encoder for concise and order-invariant structure representation. We introduce a novel pre-training scheme for molecular point clouds to distillate the knowledge from spatial molecular structures datasets. After fine-tuning within both single-task and multi-task frameworks, nach0-pc demonstrates performance comparable with other diffusion models in terms of generated samples quality across several established spatial molecular generation tasks. Notably, our model is a multi-task approach, in contrast to diffusion models being limited to single tasks. Additionally, it is capable of processing point cloud-related data, which language models are not capable of handling due to memory limitations. These lead to our model having reduced training and inference time while maintaining on par performance.

Introduction

Language Models (LMs) have shown exceptional natural language understanding (Devlin et al. 2019), and performance across diverse natural language tasks (Raffel et al. 2020; Lewis et al. 2020). LMs also show efficiency as conversational agents, engaging in meaningful dialogues (Brown et al. 2020).

Recent studies (Pei et al. 2023; Livne et al. 2024) have demonstrated the ability of LMs in processing specialized chemical languages, such as SMILES (Weininger 1988) and SELFIES (Krenn et al. 2020). These models exhibit proficiency in understanding and manipulating textual representations of chemical data, enabling their application across various tasks. For instance, LMs have been utilized for molecular property prediction (Ross et al. 2022), molecular generation (Edwards et al. 2022), and chemical reaction prediction (Lu and Zhang 2022).

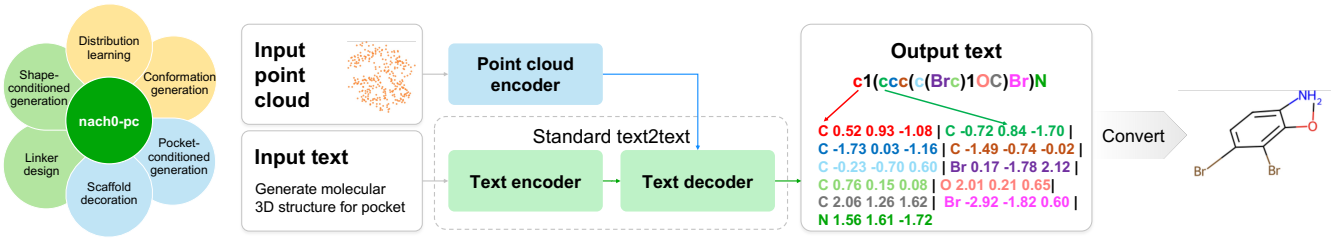
While SMILES and SELFIES capture chemical graph, they lack spatial features crucial for drug discovery methodologies involving precise atom arrangements and interactions. Recent studies (Flam-Shepherd and Aspuru-Guzik 2023) show LMs can generate meaningful 3D chemical structures in text formats like PDB, CIF, and XYZ, representing atom coordinates and features. However, this approach suffers from excessive length, requiring dozens of tokens per atom, becoming impractical for larger structures like proteins. Additionally, these formats lack information on atom connectivity, necessitating the use of external software tools like Open Babel(O'Boyle et al. 2011) to determine chemical bonds. Such external tools are sensitive to noise in atom positions, which can significantly alter reconstructed chemical graphs or cause molecular fragmentation.

In our study, we introduce nach0-pc, a Language Model designed for generative tasks involving 3D molecular structures. This model combines domain-specific encoder and textual representation of spatial atom arrangements, enabling effective handling of 3D molecular structures both as inputs and outputs. We leverage a molecular point cloud encoder to derive a concise, meaningful, and order-invariant representation of molecular and protein 3D structures. Furthermore, we propose a textual format for generating spatial molecular structures by initially generating SMILES representations, followed by specifying atom coordinates in accordance with the SMILES sequence order. This format allows LM determine molecular graph and eliminates the reliance on external software for reconstructing bonds.

The key contributions of our work are as follows:

1. A novel nach0-pc model that enhances standard encoder-decoder language models by integrating a specialized molecular point cloud encoder and tokens, incorporating unique features like point embedding calculation and position bias adjustments tailored for point cloud data.
2. A new pre-training approach that uses dropout on entire subfragments to train the model to predict missing parts of incomplete molecular point clouds, generated through fragment omitting and/or blurring strategies.
3. Extensive experiments demonstrating superior or comparable performance to LM baselines and state-of-the-art

*Corresponding author: kuznetsov@insilicomedicine.com



(a) Summary of tasks

(b) Overview of nach0-pc

Figure 1: Two diagrams of tasks and nach0-pc. (a) Three types of tasks are considered: text→text tasks in yellow; molecular point cloud + text→text in green; and molecular/protein point cloud + text→text in blue. (b) Every spatial molecular generation task we consider is cast as feeding our model text and/or molecular point cloud as input and training it to generate spatial molecular structures as output text.

diffusion approaches across six spatial molecular generation tasks.

Related Work

Language Models in Chemistry The sequential nature of molecules enables the use of Transformer models and masked language modeling pre-training methods (Lu and Zhang 2022). Recent advancements have introduced domain-specific LMs (Edwards et al. 2022; Pei et al. 2023), based on T5, designed specifically to incorporating both chemical and linguistic knowledge. MolT5 (Edwards et al. 2022) uses initial pre-training on a collection of molecule SMILES and texts, followed by single-task fine-tuning on downstream tasks. Another recent model, multi-domain nach0 (Livne et al. 2024), undergoes fine-tuning on a diverse set of 28 task-dataset pairs, employing instruction tuning in a multi-task fashion.

Spatial molecular structure generative models The majority of recently published spatial molecular structure generative models is based on the denoising diffusion probabilistic model (DDPM) paradigm (Ho, Jain, and Abbeel 2020). EDM (Hoogeboom et al. 2022) follows this methodology to solve **spatial molecular distribution learning**. A notable drawback of the diffusion approach for generating 3D molecules is its reliance on external software like OpenBabel to reconstruct molecular bonds from atom coordinates. This limitation has been addressed in several further works. The MolDiff model (Peng et al. 2023) integrates an additional bond predictor to guide diffusion and ensure bond consistency alongside 3D atom coordinates. GeoDiff (Xu et al. 2022), Torsional Diffusion (Jing et al. 2022) models can take molecular graph to perform **conformation generation** task. GeoMol(Ganea et al. 2021) is also a significant advancement in the field of cheminformatics and drug discovery. It is an end-to-end, non-autoregressive, and SE(3)-invariant machine learning approach to generate distributions of low-energy molecular 3D conformers. DiffLinker (Igashov et al. 2024) and LinkerNet (Guan et al. 2023a) models are a 3D equivariant diffusion model learned to generates the *linker* fragment between given disconnected molecular subfragments in **linker design** task. These models can find a stable linker and connect the fragments, resulting in a low-energy conformation for the entire molecule. DiffDec (Xie et al. 2024) model was proposed for **scaffold decoration** task and generates R-groups given a core of the molecule called *scaffold*. The ShapeMol (Chen et al. 2023) model utilizes a diffusion approach for **shape-conditioned generation**, where a reference ligand molecule is represented as a blurred spatial area approximating the molecular surface volume, and the model suggests new molecules with shapes closely resembling the reference. The **pocket-conditioned generation** task involves creating molecules that seamlessly fit within a designated pocket space, establishing favorable interactions to enhance binding affinity. D3FG(Lin et al. 2023), TargetDiff(Guan et al. 2023b), and DecompDiff (Guan et al. 2023c) are capable of designing novel 3D molecules from scratch that effectively bind to specified protein pockets.

Architecture

As shown in Fig. 1b, the architecture of the proposed nach0-pc model extends the standard encoder-decoder LM with a domain-specific molecular point cloud encoder and point cloud tokens in a plug-and-play manner. We use the T5 architecture (Raffel et al. 2020) as a base.

The method's plug-and-play nature allows one to choose whether to train a model from scratch or fine-tune a pre-trained LM model alongside a point cloud encoder for text and point cloud tasks. We initialize nach0-pc's LM component with nach0 (Livne et al. 2024), the state-of-the-art natural language and chemical LM. Ablation of LM components is presented in Supplementary material (Kuznetsov et al. 2024).

Input/Output Data Format

The model accepts textual input and *optionally* molecular point cloud input, while the output is textual. An input molecular point cloud consists of an unordered collection of points, denoted as $\{p_i\}$, where each point $p_i = (c_i, f_i)$ is described by Cartesian coordinates $c_i = (x_i, y_i, z_i)$ and an unordered set of tokens $f_i = \{f_i^j\}$ representing its features. For example, the features of a point corresponding to a ligand's atom might include its atom symbol and atom charge, such as $f_i = \{\texttt{'ligand'}, \texttt{'N'}, \texttt{'+'}\}$. The feature set

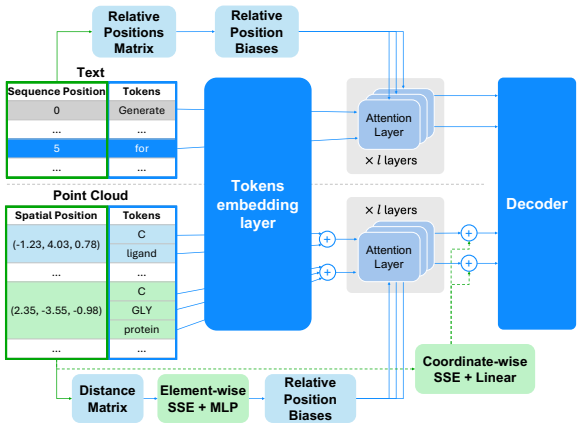


Figure 2: An overview of the encoder architecture adapted for point cloud data and standard text input. For point clouds, tokens represent features at specific spatial positions. Tokens are embedded via a token embedding layer, followed by summation pooling to optimize memory and processing efficiency. Scalar Sinusoidal Embeddings (SSE) integrate continuous spatial coordinates and relative pairwise distances.

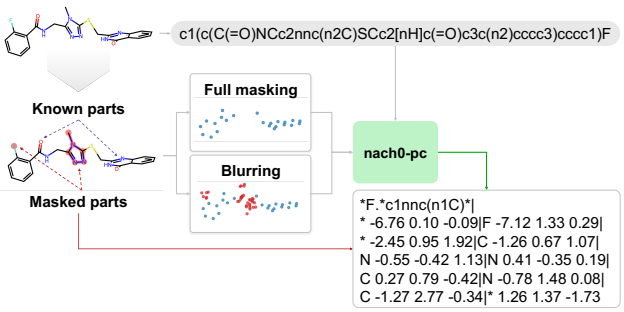


Figure 3: The pre-training scheme for 3D molecular structures datasets. The model learns to reconstruct blurred or masked molecular fragments. The model generates the missing parts during pre-training, including their SMILES representations, attachment points, and atom coordinates.

for a protein's atom point can be extended to encompass its atom name and the amino acid name it belongs, such as $f_i = \{$`'pocket', 'C', 'GLY', 'CA'`$\}$. All point features are represented as word tokens and can be utilized in textual input and output. For instance, atom symbols and charges are present in point features and SMILES tokens.

We represent a 3D molecule as text by combining SMILES and XYZ formats. Starting with SMILES to describe the molecular graph, we concatenate lines of XYZ format, describing each atom's symbol and positions in the same order they appear in SMILES. To limit the token count, we use two digits after the decimal point and tokenize each coordinate by splitting at the decimal point (`'-1.23'` $\rightarrow$ `['-1','.23']`), thus each coordinate can be described by two tokens. It takes only 12 tokens to describe one atom.

In our work we consider hydrogen-depleted molecular graphs for input point cloud and output SMILES+XYZ representation. To prepare the input point clouds, we center coordinates by deducting the average position from each point. To augment inputs and outputs, we apply the random rotations to the point clouds and make use of non-canonical SMILES representations for textual inputs/outputs.

For large ligand and pocket point clouds, we perform prioritized point downsampling - we keep ligand, C-alpha, and terminal atoms of protein amino acids and downsample other points. This procedure reduces the number of points to the fixed number and so the model fits memory constraints.

Point Cloud Encoder

The point cloud encoder utilizes the standard Transformers encoder architecture with several changes (see Fig. 2) inspired by the nature of point cloud data. The first key difference is the point embedding calculation. While word tokens are transformed using a token embedding layer, the same layer is applied to point tokens, followed by a summation pooling operation to aggregate point token embeddings. Representing spatial molecular data with point clouds uses far fewer embeddings than conventional text representation. For example, a molecular point cloud employs only a few dozens of embeddings whereas text-based representations require several hundreds of embeddings. It significantly decreases memory usage while increasing the speed of attention layers sensitive to input size.

The second notable distinction can be found in point spatial coordinates. Contrary to textual representation, which is a sequence of discrete token indices, every point's position in a point cloud is depicted using continuous Cartesian coordinates. First, we modify the relative position biases computation to embed pairwise distances instead of relative sequence positions. Also, we embed the coordinates of the point and sum them with point embeddings. We do it on the last step right before passing embeddings to the decoder, ensuring that the self-attention layers remain invariant to any point cloud translation or rotation.

To embed scalar continuous values of distances and coordinates, we are introducing Scalar Sinusoidal Embeddings (SSE). It takes inspiration from positional sinusoidal embeddings (Vaswani et al. 2017) and extends it to map continuous scalar values into a high-dimensional vector. We divide

Algorithm 1: Point Cloud Encoder

Input: Point features $f_i = \{f_i^j\}$ and pos $c_i = (x_i, y_i, z_i)$
Output: Point embeddings n_i^{out}

$n_i^0 = \sum_{f_i^j \in f_i} Emb(f_i^j)$ ▷ embed and aggregate feats
for $l = 1, 2, \ldots, L$ **do**
  $b_{hij}^l = MLP_h^l(SSE(||c_i - c_j||))$
  ▷ for each head h compute relative attention biases
  $n^l = SelfAttention^l(n^{l-1}, b^l)$ ▷ update embs
end for
$n_i^{out} = n_i^L + Lin_{xyz}(SSE(x_i), SSE(y_i), SSE(z_i))$
▷ embed coordinates

Group	Metrics	State-of-the-art			nach0-pc	
		MolDiff	OpenLLaMA	nach0	Multi-Task	Single-Task
Basic	Validity and Connectivity (↑)	**99.3%**	95.0%	98.6%	97.8%	99.0%
	Uniqueness (↑)	92.8%	**99.8%**	**99.8%**	99.2%	**99.8%**
	Novelty (↑)	97.7%	85.0%	**98.0%**	96.2%	97.1%
	Diversity (↑)	0.763	0.745	0.739	**0.781**	0.734
Druglikeness	QED (↑)	0.679	0.673	0.667	**0.771**	0.664
	SA (↑)	**0.875**	0.808	0.851	0.872	0.848
	Lipinski (↑)	4.981	4.972	4.938	**4.992**	4.938
3D substructures	JS. bond lengths (↓)	0.436	0.257	0.148	0.193	**0.142**
	JS. bond angles (↓)	0.181	0.136	0.096	0.100	**0.94**
	JS. dihedral angles (↓)	0.198	**0.110**	**0.110**	0.132	0.113
Bonds	JS. # bonds per atoms (↓)	0.121	**0.064**	0.111	0.233	0.094
	JS. freq. bond types (↓)	0.170	**0.031**	0.046	0.052	0.041
	JS. freq. bond pairs (↓)	0.153	**0.028**	0.037	0.040	0.033
	JS. freq. bond triplets (↓)	0.137	**0.034**	0.046	0.054	0.041
Rings	JS. # rings (↓)	0.079	**0.033**	0.56	0.270	0.036
	JS. # n-sized rings (↓)	0.102	**0.024**	0.026	0.058	**0.024**
	# Intersecting rings (↑)	8	8	8	**9**	8

Table 1: Spatial molecular distribution learning performance metrics on GEOM-DRUGS.

the input scalar s by wavelengths w_i, uniformly initialized on a logarithmic grid, and use the yielded sine and cosine function values to form the resulting embedding vector. The step-by-step description can be found in Alg. 1.

$$SSE_{2i}(s) = sin(s/w_i), SSE_{2i+1}(s) = cos(s/w_i)$$

Molecular Point Cloud Pre-training

Drawing inspiration from the token dropout pre-training objective of T5 (Raffel et al. 2020), we introduce a novel pre-training scheme for molecular point clouds (Fig. 3). In our approach, we feed the model with incomplete molecular point clouds and train the model to reconstruct the missing pieces. We employ the BRICS (Degen et al. 2008) algorithm to split the molecule into several fragments and randomly select a subset of these fragments with some predefined probability (ensuring that at least one is chosen) and exclude them from the molecule. The pre-training scheme is designed to be flexible, handling both datasets consisting only of spatial molecular structures and datasets with protein pockets and ligand pairs. In the latter, the ligand is masked, while the protein pocket remains unchanged in the input point cloud. Further, we form the input point cloud by (i) removing or (ii) obfuscating chosen fragments by omitting the features of a point (such as atomic symbol, charge, and valency), replicating the point multiple times, and introducing Gaussian noise to the coordinates of these copies.

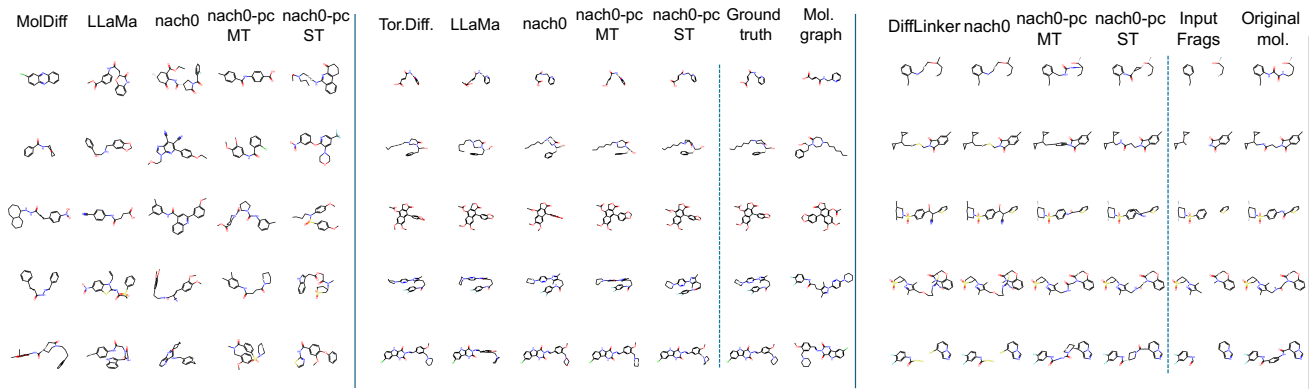
During pre-training, the model's task is to accurately recover the absent (or blurred) components, specifying their SMILES representations as well as atomic coordinates. Each recovered fragment should include a connecting point indicated by the symbol `'*'`. In cases where the model should reconstruct multiple missing fragments, the fragments are separated with the `'.'` token. We note that 3D pre-training enhances the model's performance on downstream tasks (see Impact of 3D pre-training in the Supplementary material (Kuznetsov et al. 2024)).

Experiments

We evaluate the quality of the nach0-pc model across several established spatial molecular generation tasks: (i) 3D molecular structures generation: spatial molecular distribution learning, conformation generation, (ii) molecular completion: linker design, scaffold decoration, (iii) shape-conditioned generation, (iv) pocket-conditioned generation. We highlight **the best** and the second best metrics values for better readability and provide sampled spatial molecular structures for visual inspection of results.

We compare our nach0-pc model with well-established baseline models proposed for each mentioned task. We also compare it with text-only language model - we train OpenLLaMA model (Geng and Liu 2023) with the same (350M) number of parameters as nach0-pc, along with nach0 (Livne et al. 2024) on spatial molecular distribution learning, conformation generation and linker design tasks, where input/output text and molecular condition in SMILES+XYZ format can be described with fewer than a thousand tokens. The rest of tasks either require large protein pocket condition or involve blurring that substantially inflate the size of the point cloud representation, making these tasks practically unfeasible in text-only paradigm.

We train our proposed nach0-pc model in both multi-task and single-task regimes to fairly evaluate it against both LM and non-LM models. Our work adopts small molecules ZINC (Irwin et al. 2020), MOSES (Polykovskiy et al. 2020), and GEOM-Drugs (Axelrod and Gómez-Bombarelli 2022) datasets, as well as the CrossDocked2020 (Francoeur et al. 2020) dataset, which includes pocket-ligand pairs. In cases when tasks use the same dataset, to avoid any potential data leakage, we use the same dataset split. We combine all stated datasets to both pretrain and finetune the nach0-pc model in multi-task regime. To address the discrepancy in dataset sizes, we normalize the training process by employing a batch-balancing technique that involves retrieving random object from uniformly sampled task. See Supple-



(a) Distribution learning (b) Conformation generation (c) Linker design

Figure 4: Samples for (a) spatial distribution learning, (b) conformation generation and (c) linker design tasks.

mentary material (Kuznetsov et al. 2024) for additional details on datasets, models, and metrics.

Model Parameters and Training Details

In our research, we primarily utilize a model based on the nach0 (Livne et al. 2024) architecture. Our experiments involve a `base` model variant, characterized by 12 layers, a hidden state of 768 dimensions, a feed-forward hidden state of 2048 dimensions, and 12 attention heads. We utilize the same parameters to build our point cloud encoder. The total model size has 370M parameters. The model was trained using two NVIDIA A6000 GPUs. The pre-training and fine-tuning stages were executed using the following hyperparameters: a batch size of 64 for both pre-training and fine-tuning, a learning rate set to 1e-4, a weight decay of 0.01, and a cosine schedule. Both the pre-training and fine-tuning stages lasted for 100000 steps.

Spatial Molecular Distribution Learning and Conformation Generation Tasks

This section evaluates the model's ability to generate structurally and physically plausible spatial molecular objects. The **spatial molecular distribution learning** task assesses whether the model can produce novel 3D molecular structures whose distribution is close to the ground truth. Following (Peng et al. 2023), we adopt high-quality GEOM-Drugs (Axelrod and Gómez-Bombarelli 2022) dataset, which offers gold-standard conformation ensembles generated using metadynamics in CREST (Pracht, Bohle, and Grimme 2020). We evaluate generated molecules from several perspectives, including basic (validity, novelty, etc) drug-likeness, and 3D substructures, bonds and rings distribution divergences.

In the **conformation generation**, we focus on generating plausible *conformations* given molecular graph. The *conformations* of a molecule are its energetically favorable 3D structures, each representing a local minimum on the potential energy surface. Similar to the previous task, we employ the GEOM (Axelrod and Gómez-Bombarelli 2022) dataset for this task. We follow (Jing et al. 2022) and employ the Average Minimum RMSD (AMR) and Coverage metrics. These metrics are evaluated from Recall and Precision view.

We utilize the same train/validation/test splits as in the conformation generation task from the Torsional Diffusion(Jing et al. 2022) paper and retrain baseline if they were trained on another split.

The results for two tasks are presented in Tables 1 and 2. Examples of generated structures are given on Figs. 4a, 4b. The proposed nach0-pc model shows the best or second best results for the majority of metrics on both tasks. As shown in Tab. 1, text-only OpenLLaMA shows slightly better results on bonds and rings distribution metrics. Our hypothesis, it is due to the fact the input of this task is the same all time, so decoder-only model better utilize parameters rather than encoder-decoder. On the conformation generation task, the only model outperforming nach0-pc on the whole set of metrics is Torsional Diffusion, which utilizes samples from external software RDKit (Landrum et al. 2023) as an initial generation point on the inference stage. Nevertheless, nach0-pc shows stronger results than all other purely neural baselines that produce molecular conformation from scratch.

Method	Recall				Precision			
	COV (↑)		AMR (↓)		COV (↑)		AMR (↓)	
	Avg	Med	Avg	Med	Avg	Med	Avg	Med
Single-Task baselines								
ETKDG	38.4	28.6	1.06	1.00	40.9	30.8	0.99	0.89
GeoMol	44.6	41.4	0.87	0.83	43.0	36.4	0.93	0.84
GeoDiff	42.1	37.8	0.83	0.81	24.9	14.5	1.14	1.09
Tor. Diff.	**72.7**	**80.0**	**0.58**	**0.56**	**55.2**	**56.9**	**0.78**	**0.73**
Multi-Task Text-Only LM baselines								
OpenLLaMA	10.2	0.8	1.48	1.43	19.7	1.7	1.31	1.28
nach0	54.0	54.55	0.74	0.73	30.7	20.7	1.06	1.06
nach0-pc								
Multi-Task	50.1	50.0	0.76	0.75	27.7	16.7	1.11	1.11
Single-Task	57.7	59.5	0.70	0.69	32.4	23.1	1.03	1.03

Table 2: Generated conformer ensembles quality on GEOM-DRUGS (baselines' metrics from (Jing et al. 2022)).

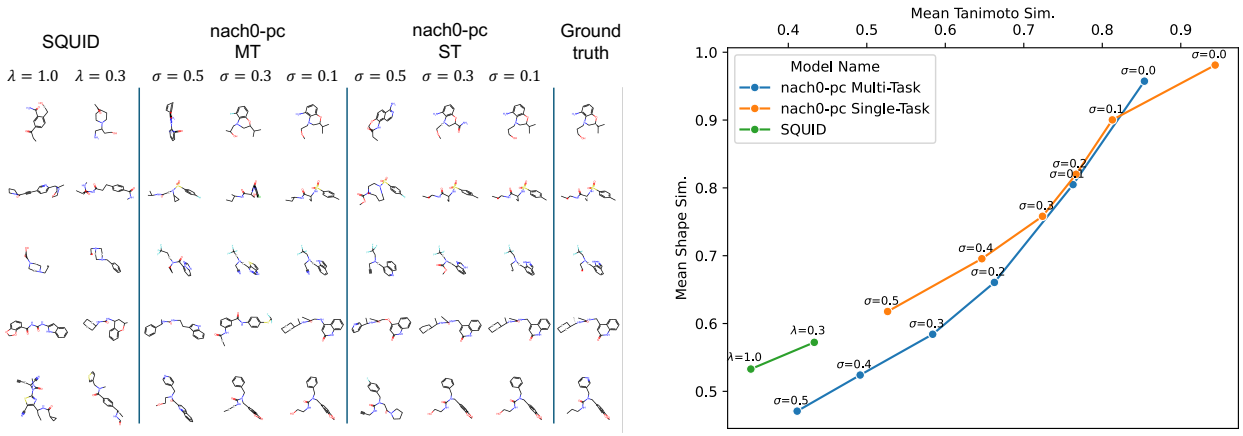


Figure 5: (left) Generated molecular structures and (right) structural/shape similarity trade-off for various noise parameters on shape-conditioned generation task.

Method	Val. (↑)	Uniq. (↑)	Filt. (↑)	RMSD (↓)	SC (↑)
State-of-the-art					
DeLinker	**98.3%**	**44.2%**	84.88%	5.48	0.49
3DLinker	71.5%	29.2%	83.72%	**0.11**	0.92
DiffLinker	93.8%	24.0%	86.26%	0.34	**0.93**
nach0	55.2%	19.9%	98.78%	1.41	0.85
nach0-pc					
Multi-Task	81.6%	27.6%	99.00%	1.28	0.86
Single-Task	89.7%	12.3%	**99.55%**	1.04	0.88

Table 3: Models performance evaluation for linker design task (metrics for baselines from (Igashov et al. 2024)).

Model	Validity (↑)	Uniqueness (↑)	QVina Score (↓)
Reference	-	-	-8.47
Scaffolds	-	-	-7.73
State-of-the-art			
Pocket2Mol	51.14%	44.27%	-8.11
FLAG	87.95%	**65.30%**	-7.62
DiffDec	98.00%	48.54%	**-8.25**
nach0-pc			
Multi-Task	97.63%	42.58%	-7.931
Single-Task	**99.17%**	18.12%	-8.123

Table 4: Scaffold decoration task metrics (metrics of baselines from (Xie et al. 2024)).

Shape-conditioned Generation

This section focuses on producing molecules spatially similar to the reference structure but structurally dissimilar. This is achieved by representing the reference molecule as a *shape* - an area where molecular atom nucleus and electron clouds are located. We represent molecular shape as a point cloud - we replicate each atom several times and add Gaussian noise with predefined standard deviation σ to atom positions while removing all point features completely. σ allows to balance between spatial and chemical similarity.

Following the methodology outlined in the SQUID paper (Adams and Coley 2023), we conduct training and test-stage sampling using the RDKit (Landrum et al. 2023) conformations computed for the MOSES dataset (Polykovskiy et al. 2020). We provide a comparison of nach0-pc with the SQUID (Adams and Coley 2023) model in Fig. 5. We alternate noise injection parameters, standard deviation σ for nach0-pc and prior interpolation coefficient λ for SQUID, to show available trade-offs between structural and shape similarity. nach0-pc provides a wider range of available trade-offs, covering a high structural to high shape similarity area. Moreover, nach0-pc produces more spatially similar objects for low structural similarity values than SQUID.

Linker Generation and Scaffold Decoration Tasks

These tasks assess the ability of models to complete disjoint or partially-defined molecular structures. In **linker design** task, models operate with several disconnected fragments and should produce small molecular structures that spatially and chemically connect the given fragments and complete into one chemical structure. The same as in the DiffLinker work (Igashov et al. 2024), we employ a subset of the ZINC dataset, comprising 250000 random molecules with conformations generated using RDKit (Landrum et al. 2023). This dataset also provides an input fragments/linker splits.

In the second task, **scaffold decoration**, the model takes the core part of the molecule called *scaffold* and complete side-chain specific motifs called R-groups. Usually, scaffold decoration is employed to enhance some molecular properties, for instance, binding affinity with a specific protein. Following DiffDec (Xie et al. 2024), we adopt Multi R-Group Decoration Task on CrossDocked (Francoeur et al. 2020) dataset, containing $100K$ ligand-protein pairs, where each ligand is split into a scaffold and R-groups.

Our nach0-pc takes molecular fragments/scaffold and protein pocket, if available, as an input point cloud and produces the linker/R-groups without repeating input atoms. The produced molecular substructures contain the attachment points described by a symbol `'*'` and coordinates.

Model	Valid. (↑)	Div. (↑)	Vina Dock (↓)		High Affinity (↑)
			Avg	Med	
Reference	100%	-	-7.45	-7.26	-
State-of-the-art					
AR	92.95%	0.70	-6.75	-6.62	37.9%
Pocket2Mol	**98.31%**	0.69	-7.15	-6.79	48.4%
TargetDiff	90.36%	**0.72**	**-7.80**	**-7.91**	**58.1%**
nach0-pc					
Multi-Task	91.78%	0.32	-6.52	-6.86	38.2%
Single-Task	89.82%	0.40	-6.50	-6.62	41.1%

Table 5: Pocket-conditioned generation performance metrics (metrics of baselines from (Guan et al. 2023b)).

We utilize these attachment points to combine input fragments with generated ones into a coherent molecule. The model can produce several R-groups in the scaffold decoration task by separating them with a symbol `'.'`. We compare nach0-pc with DeLinker (Imrie et al. 2020), 3DLinker (Huang et al. 2022), DiffLinker (Igashov et al. 2024) on the linker generation task, and benchmark against LibINVENT (Fialková et al. 2022), FLAG (Zhang et al. 2023), and DiffDec (Xie et al. 2024) for scaffold decoration. As shown in Tables 3 and 4 nach0-pc can complete input molecular point clouds with a high success rate, producing molecules (Fig. 4c) that pass 2D filters such as PAINS (Baell and Holloway 2010). Despite moderate structural diversity, nach0-pc produces spatially diverse molecules. Moreover, it enhances scaffold binding affinity, working on par with other state-of-the-art models in scaffold decoration.

Pocket-conditioned Generation

Finally, nach0-pc was trained to generate novel high-affinity structures for a given *protein pocket* condition. Similar to TargetDiff, we used the CrossDocked2020 dataset (Francoeur et al. 2020) and trained our model on 100000 high-affinity ligand-protein complexes. During the test stage, we randomly sampled 100 molecules for each protein pocket in the test set. One can find generated ligands visualisation in Fig. 6. We assess the validity, diversity and docking scores of generated structures and provide the comparison with AR(Luo et al. 2021), Pocket2Mol(Peng et al. 2022) and TargetDiff(Guan et al. 2023b) baseline models in Tab 5.

Our model shows high validity and diversity. While there is a significant gap between nach0-pc and the TargetDiff performance, it demonstrates comparable results to AR and Pocket2Mol based on docking scores and binding affinity.

Model Training Time and CO2 Impact

All experiments were conducted utilizing the CoreWeave infrastructure. Our model the training and evaluation were performed on an Nvidia A6000. The total training and evaluation time for our model was 164.5 hours, resulting in an estimated CO2 emission of 20.73 kgCO2eq. For the training and evaluation of MolDiff and EDM models, we utilized an Nvidia A4000. The models required 1020 GPU hours, leading to an estimated CO2 emission of 94.36 kgCO2eq. For more details see (Kuznetsov et al. 2024).

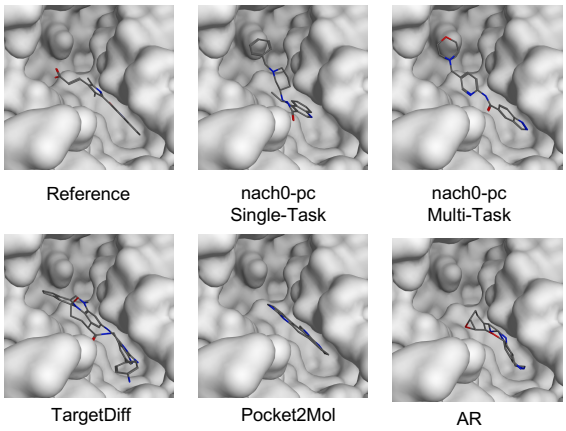


Figure 6: Generated molecular structures for pocket-conditioned generation (reference name `4iwq_A`) task.

Conclusion

We have introduced nach0-pc, a novel model adept at generating diverse and physically plausible molecular 3D structures. By combining a domain-specific point cloud encoder with an encoder-decoder language model along with combined SMILES+XYZ textual format and novel molecular point cloud pre-training technique, nach0-pc effectively addresses challenges associated with handling chemical 3D structures and a SMILES sequence. Through extensive fine-tuning within single-task and multi-task frameworks, nach0-pc exhibits comparable performance to various state-of-the-art diffusion and LM baseline models. As future work, it would be valuable to explore training on a broader range of NLP and Chemistry 2D/3D tasks in a multi-task fashion, including molecular properties prediction based on a spatial input and protein-related tasks. The integration of proposed ideas into decoder-only approaches remains to be explored.

Acknowledgments

We are grateful to Elizaveta Ekimova for her significant assistance in preparing the accompanying graphic material.

References

Adams, K.; and Coley, C. W. 2023. Equivariant Shape-Conditioned Generation of 3D Molecules for Ligand-Based Drug Design. In *The Eleventh International Conference on Learning Representations*.

Axelrod, S.; and Gómez-Bombarelli, R. 2022. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1): 185.

Baell, J. B.; and Holloway, G. A. 2010. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry*, 53(7): 2719–2740. PMID: 20131845.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell,

A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chen, Z.; Peng, B.; srinivasan parthasarathy; and Ning, X. 2023. Shape-conditioned 3D Molecule Generation via Equivariant Diffusion Models. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*.

Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; and Rarey, M. 2008. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem*, 3(10): 1503–1507.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Edwards, C.; Lai, T.; Ros, K.; Honke, G.; Cho, K.; and Ji, H. 2022. Translation between Molecules and Natural Language. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 375–413. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Fialková, V.; Zhao, J.; Papadopoulos, K.; Engkvist, O.; Bjerrum, E. J.; Kogej, T.; and Patronov, A. 2022. LibINVENT: Reaction-based Generative Scaffold Decoration for in Silico Library Design. *Journal of Chemical Information and Modeling*, 62(9): 2046–2063. PMID: 34460269.

Flam-Shepherd, D.; and Aspuru-Guzik, A. 2023. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as XYZ, CIF, and PDB files. arXiv:2305.05708.

Francoeur, P. G.; Masuda, T.; Sunseri, J.; Jia, A.; Iovanisci, R. B.; Snyder, I.; and Koes, D. R. 2020. Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *Journal of Chemical Information and Modeling*, 60(9): 4200–4215. PMID: 32865404.

Ganea, O.; Pattanaik, L.; Coley, C.; Barzilay, R.; Jensen, K.; Green, W.; and Jaakkola, T. 2021. GeoMol: Torsional Geometric Generation of Molecular 3D Conformer Ensembles. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 13757–13769. Curran Associates, Inc.

Geng, X.; and Liu, H. 2023. OpenLLaMA: An Open Reproduction of LLaMA.

Guan, J.; Peng, X.; Jiang, P.; Luo, Y.; Peng, J.; and Ma, J. 2023a. LinkerNet: Fragment Poses and Linker Co-Design with 3D Equivariant Diffusion. In Oh, A.; Neumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 77503–77519. Curran Associates, Inc.

Guan, J.; Qian, W. W.; Peng, X.; Su, Y.; Peng, J.; and Ma, J. 2023b. 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. In *The Eleventh International Conference on Learning Representations*.

Guan, J.; Zhou, X.; Yang, Y.; Bao, Y.; Peng, J.; Ma, J.; Liu, Q.; Wang, L.; and Gu, Q. 2023c. DecompDiff: Diffusion Models with Decomposed Priors for Structure-Based Drug Design. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 11827–11846. PMLR.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 6840–6851. Curran Associates, Inc.

Hoogeboom, E.; Satorras, V. G.; Vignac, C.; and Welling, M. 2022. Equivariant Diffusion for Molecule Generation in 3D. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 8867–8887. PMLR.

Huang, Y.; Peng, X.; Ma, J.; and Zhang, M. 2022. 3DLinker: An E(3) Equivariant Variational Autoencoder for Molecular Linker Design. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 9280–9294. PMLR.

Igashov, I.; Stärk, H.; Vignac, C.; Schneuing, A.; Satorras, V. G.; Frossard, P.; Welling, M.; Bronstein, M.; and Correia, B. 2024. Equivariant 3D-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*.

Imrie, F.; Bradley, A. R.; van der Schaar, M.; and Deane, C. M. 2020. Deep Generative Models for 3D Linker Design. *Journal of Chemical Information and Modeling*, 60(4): 1983–1995. PMID: 32195587.

Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J. W.; and Sayle, R. A. 2020. ZINC20 - A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.*, 60(12): 6065–6073.

Jing, B.; Corso, G.; Chang, J.; Barzilay, R.; and Jaakkola, T. 2022. Torsional Diffusion for Molecular Conformer Generation. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 24240–24253. Curran Associates, Inc.

Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; and Aspuru-Guzik, A. 2020. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4): 045024.

Kuznetsov, M.; Valiev, A.; Aliper, A.; Polykovskiy, D.; Tutubalina, E.; Shayakhmetov, R.; and Miftahutdinov, Z. 2024. nach0-pc: Multi-task Language Model with Molecular Point Cloud Encoder. *arXiv preprint arXiv:2410.09240*.

Landrum, G.; Tosco, P.; Kelley, B.; Ric; Cosgrove, D.; sriniker; gedeck; Vianello, R.; NadineSchneider; Kawashima, E.; Jones, G.; N, D.; Dalke, A.; Cole, B.; Swain, M.; Turk, S.; AlexanderSavelyev; Vaucher, A.; Wójcikowski, M.; Take, I.; Scalfani, V. F.; Probst, D.; Ujihara, K.; guillaume godin; Walker, R.; Lehtivarjo, J.; Pahl, A.; Berenger, F.; jasondbiggs; and strets123. 2023. rdkit/rdkit: 2023_09_3 (Q3 2023) Release.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Lin, H.; Huang, Y.; Zhang, O.; Liu, Y.; Wu, L.; Li, S.; Chen, Z.; and Li, S. Z. 2023. Functional-Group-Based Diffusion for Pocket-Specific Molecule Generation and Elaboration. In Oh, A.; Neumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 34603–34626. Curran Associates, Inc.

Livne, M.; Miftahutdinov, Z.; Tutubalina, E.; Kuznetsov, M.; Polykovskiy, D.; Brundyn, A.; Jhunjhunwala, A.; Costa, A.; Aliper, A.; Aspuru-Guzik, A.; and Zhavoronkov, A. 2024. nach0: multimodal natural and chemical languages foundation model. *Chem. Sci.*, 15: 8380–8389.

Lu, J.; and Zhang, Y. 2022. Unified Deep Learning Model for Multitask Reaction Predictions with Explanation. *Journal of Chemical Information and Modeling*, 62(6): 1376–1387. PMID: 35266390.

Luo, S.; Guan, J.; Ma, J.; and Peng, J. 2021. A 3D Generative Model for Structure-Based Drug Design. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 6229–6239. Curran Associates, Inc.

O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; and Hutchison, G. R. 2011. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1): 33.

Pei, Q.; Zhang, W.; Zhu, J.; Wu, K.; Gao, K.; Wu, L.; Xia, Y.; and Yan, R. 2023. BioT5: Enriching Cross-modal Integration in Biology with Chemical Knowledge and Natural Language Associations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1102–1123. Singapore: Association for Computational Linguistics.

Peng, X.; Guan, J.; Liu, Q.; and Ma, J. 2023. MolDiff: Addressing the Atom-Bond Inconsistency Problem in 3D Molecule Diffusion Generation. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 27611–27629. PMLR.

Peng, X.; Luo, S.; Guan, J.; Xie, Q.; Peng, J.; and Ma, J. 2022. Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 17644–17655. PMLR.

Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; and Zhavoronkov, A. 2020. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology*, 11.

Pracht, P.; Bohle, F.; and Grimme, S. 2020. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics*, 22(14): 7169–7192.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; and Das, P. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12): 1256–1264.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1): 31–36.

Xie, J.; Chen, S.; Lei, J.; and Yang, Y. 2024. DiffDec: Structure-Aware Scaffold Decoration with an End-to-End Diffusion Model. *Journal of Chemical Information and Modeling*, 64(7): 2554–2564. PMID: 38267393.

Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; and Tang, J. 2022. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In *International Conference on Learning Representations*.

Zhang, Z.; Min, Y.; Zheng, S.; and Liu, Q. 2023. Molecule Generation For Target Protein Binding with Structural Motifs. In *The Eleventh International Conference on Learning Representations*.