Repurformer: Transformers for Repurposing-Aware Molecule Generation

Changhun Lee
UNIST
South Korea
changhun@unist.ac.kr

Gyumin Lee
Korea University
South Korea
optimizt@korea.ac.kr

Abstract

Generating as diverse molecules as possible with desired properties is crucial for drug discovery research, which invokes many approaches based on deep generative models today. Despite recent advancements in these models, particularly in variational autoencoders (VAEs), generative adversarial networks (GANs), Transformers, and diffusion models, a significant challenge known as *the sample bias problem* remains. This problem occurs when generated molecules targeting the same protein tend to be structurally similar, reducing the diversity of generation. To address this, we propose leveraging multi-hop relationships among proteins and compounds. Our model, Repurformer, integrates bi-directional pretraining with Fast Fourier Transform (FFT) and low-pass filtering (LPF) to capture complex interactions and generate diverse molecules. A series of experiments on BindingDB dataset confirm that Repurformer successfully creates substitutes for anchor compounds that resemble positive compounds, increasing diversity between the anchor and generated compounds.

1 Introduction

The design of valid and novel molecules with desired biological properties, known as *de novo* molecule generation, is vital to modern drug discovery. Recent advancements in deep generative models, particularly variational autoencoders (VAEs) (Kingma and Welling, 2022), generative adversarial networks (GANs) (Goodfellow et al., 2014), Transformers (Vaswani et al., 2017), and diffusion models (Ho et al., 2020), have significantly enhanced our ability to generate chemically valid and novel molecules. However, these models need to be further refined to generate molecules that interact with specific target proteins.

Target-specific molecule generation addresses this challenge by producing drug-like molecules that are more likely to bind with specific target proteins (Grechishnikova, 2021; Qian et al., 2022; Tan et al., 2022). Nonetheless, there remains a significant issue known as *the sample bias problem*, where reliance on existing protein-compound pairs results in the generation of structurally similar molecules. This phenomenon limits the diversity of generated molecules and hinders the discovery of novel compounds.

To address this, we propose leveraging multi-hop relationships among proteins and compounds to expand the generative space and increase the diversity of the generated molecules. Our method introduces the concept of repurposing-aware molecule generation, designed to identify and utilize latent multi-hop relations within the protein-compound interaction network.

In this paper, we present Repurformer, a novel model that integrates bi-directional pretraining and advanced signal processing techniques to overcome the limitations of existing models. Repurformer captures complex relationships between proteins and compounds by pretraining encoders in both protein-to-compound and compound-to-protein directions and applying Fast Fourier Transform (FFT) with low-pass filtering (LPF) to the latent space. This approach allows the model to distinguish the different scales of interactions. By focusing on low-frequency components, which correspond to the longer propagation through the multi-hop protein-compound interaction network, Repurformer generates as diverse compounds as possible with desired properties. In summary, the contributions of our work are threefold:

- We introduce a framework for repurposing-aware molecule generation to address the *sample bias problem* by leveraging multi-hop relations between proteins and compounds.
- We develop Repurformer, a model that integrates a bi-directional pretraining and an FFT-based approach to capture and utilize latent

Proceedings of the 1st Workshop on Language + Molecules (L+M 2024), pages 117–128

multi-hop relations in an end-to-end manner.

- We demonstrate that Repurformer successfully generates valid and diverse molecules, creating substitutes for anchor compounds that resemble positive compounds.

2 Preliminaries

2.1 *De novo* Molecule Generation

De novo molecule generation is the process of exploring vast chemical space and producing novel molecules with desired biological properties. With the rapid advancement of artificial intelligence, recent deep generative models have been widely used in molecule generation tasks.

For example, several VAE variants have been introduced thanks to its manipulable latent space, such as charVAE (Gómez-Bombarelli et al., 2018), SD-VAE (Dai et al., 2018), and JT-VAE (Jin et al., 2018). GAN has been adopted due to their capability to generate new molecules that are highly similar in structure to existing ones, including ORGAN (Guimaraes et al., 2018), ORGANIC (Sanchez-Lengeling et al., 2017), and MolCycleGAN (Maziarka et al., 2020). More recently, Transformers and diffusion models have been utilized, based on their success in language modeling and image generation, respectively, such as MolGPT (Bagal et al., 2022), MDM (Huang et al., 2023), and GeoLDM (Xu et al., 2023).

2.2 Target-specific Molecule Generation

In drug discovery, identifying drug-target interactions (DTI) is crucial for understanding the bioactivity and therapeutic effects of drugs for specific diseases. Although the deep generative models have proved useful in generating novel and chemically valid molecules, further screening is necessary to evaluate their potential to bind with specific protein targets. Building on this notion, several researchers have developed target-specific molecule generation models to produce novel, drug-like molecules that are highly likely to interact with specific target proteins, including Transformer-based generation (Grechishnikova, 2021), AlphaDrug (Qian et al., 2022), SiamFlow (Tan et al., 2022) and POLYGON (Munson et al., 2024).

2.3 Repurposing-Aware Molecule Generation

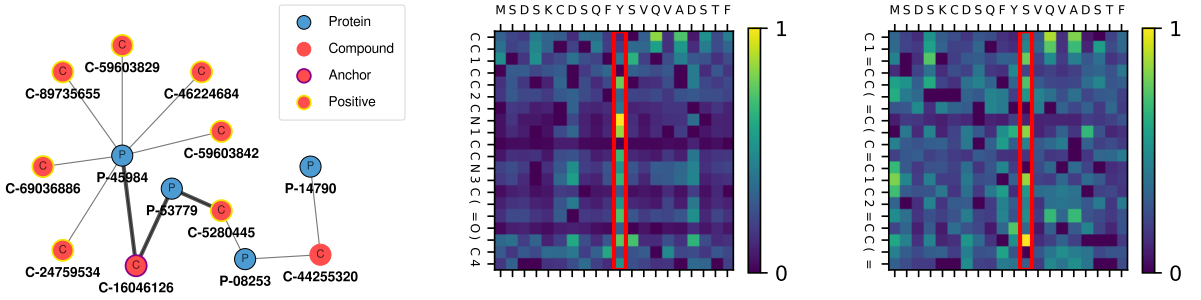
Drug repurposing is a strategy that identifies new therapeutic uses for approved drugs beyond their original indications. This approach offers significant advantages over developing entirely new drug, such as lower failure risk and development costs. The concept of drug repurposing can be defined as multi-hop relationships in the protein-compound interaction network, which is not directly connected but can be accessed through intermediaries. In chemical spaces, proteins and compounds have many-to-many relationships based on their structural coordination. This leads to the assumption that if a compound can reach a specific protein through another compound that shares a common protein (*i.e.,* in the multi-hop relationship), there is potential for repurposing the focal compound.

The repurposability in chemical spaces can introduce a new paradigm for molecule generation, by serving as a key to expanding the generative space and increasing molecular diversity. Previous approaches for target-specific molecule generation tend to generate structurally similar molecules for a specific protein target due to their dependence on known protein-compound interactions. While incorporating randomness in the generation process can contribute to molecular diversity, it may neglect structural coordination with targets, possibly resulting in a trade-off between diversity and binding affinity. In this context, leveraging latent repurposability within the multi-hop relationships among proteins and compounds can provide a reasonable boundary for molecule generation, broadening the generative space and enhancing molecular diversity without sacrificing their drug potency.

3 Problem Statement

The discovery of new compounds often relies on existing protein-compound pairs. This results in that the compounds targeting the same protein exhibit similar structures. In other words, the generative space of models tends to be bounded in limited regions, reducing the diversity of the generation. We refer to this as a *sample bias problem*.

To address this problem, we leverage multi-hop relations among proteins and compounds. Specifically, given a pair of protein p and compound c that are known to interact, we assume that the compound relates to p within a 3-hop relation, *i.e.,* a positive compound c^+, has a potential interaction with p. Definitions from 3.1 to 3.3 describe the key concepts of our approach, and Figure 1 visually represents the rationale. Note that both protein and compound are represented by



(a) Protein-Compound Graph (b) C-16046126 & P-45984 Attention (c) C-5280445 & P-45984 Attention

Figure 1: (a) illustrates a many-to-many relationship between proteins and compounds. The bold lines indicate potential repurposing flows by which, given an anchor compound's target protein p (P-45984), a positive compound c^+ (C-5280445) can be considered to replace the anchor compound $\acute{c}$ (C-16046126). Red boxes in (b) and (c) represent the parts of p (P-45984) to which $\acute{c}$ (C-16046126) and c^+ (C-5280445) attend, respectively. It is noteworthy that attending regions are right next to each other, implying c^+ may have a potential repurposability to p.

amino-acid and SMILES sequences, respectively: $p = [p_1, \cdots, p_{T_p}]$ and $c = [c_1, \cdots, c_{T_c}]$ with T_p and T_c being the fixed length of each sequence.

Definition 3.1 (Protein-Compound Graph)**.** *The relations between proteins and compounds can be represented as a bipartite graph $\mathcal{G}(\mathcal{P} \cup \mathcal{C}, \mathcal{E})$, where $\mathcal{P}$ and $\mathcal{C}$ denote the sets of protein and compound nodes, respectively. Specifically, $p^{(i)} \in \mathcal{P}$ represents the i-th protein and $c^{(j)} \in \mathcal{C}$ represents the j-th compound, for $i = 1, \cdots, M$ and $j = 1, \cdots, N$.*

Definition 3.2 (Protein-Compound Pair)**.** *A pair of nodes in $\mathcal{G}$ is represented by an edge $e_{ij} = \{(p^{(i)}, c^{(j)}) | p^{(i)} \in \mathcal{P}, c^{(j)} \in \mathcal{C}\} \in \mathcal{E}$. The presence of an edge e_{ij} indicates a link between the i-th protein and the j-th compound, such that $e_{ij} = 1$ if they are linked and $e_{ij} = 0$ otherwise.*

Definition 3.3 (Anchor/Positive Compounds)**.** *Given a target protein $p^{(i)}$, a compound $c^{(j)}$ is defined as an anchor compound $\acute{c}$ if $e_{ij} = 1$. For another protein $p^{(k)}$ where $e_{kj} = 1$, any compound $c^{(l)}$ $(l \neq j)$ that satisfies $e_{kl} = 1$ is regarded as a positive compound c^+ for the target protein $p^{(i)}$.*

4 Repurformer

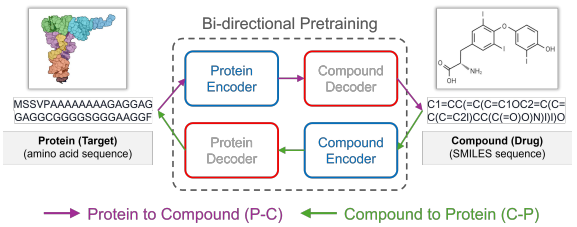
In this section, we propose Repurformer, a novel method designed to address the sample bias problem by leveraging multi-hop relations among proteins and compounds. Figure 2 illustrates how Repurformer seamlessly integrates the concepts of drug discovery and repurposing.

Bi-directional Pretraining To capture the many-to-many relationships between proteins and compounds, we employed bi-directional pretraining for the protein and compound encoders. Specifically, we built two Transformers with identical encoder-decoder structures but opposite training directions: one was trained in the protein-to-compound direction, and the other in the compound-to-protein direction (see Figure 2a). By doing so, we expect the protein encoder $f_p(c|p)$ and the compound encoder $f_c(p|c)$ to extract latent relations, z^p and z^c, that encompass both cases where proteins and compounds are the head and tail of an edge, and vice versa, *i.e.*, $f_p : c|p \rightarrow z^p$ and $f_c : p|c \rightarrow z^c$. For example, given a pair of $p^{(2)}$ and $c^{(1)}$ as shown in Figure 2b, z^p and z^c will represent the edges from $p^{(2)}$ to $c^{(1)}$ (*i.e.*, $p^{(2)} \rightarrow c^{(1)}$) and from $c^{(1)}$ to $p^{(1)}$ (*i.e.*, $c^{(1)} \rightarrow p^{(1)}$), respectively.

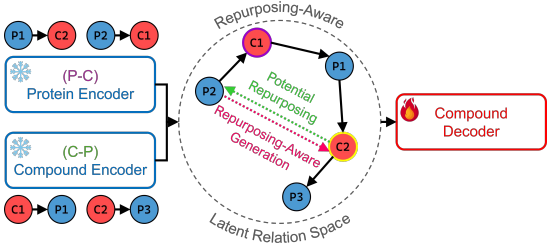
Transformer with Bi-directional Encoders The pretrained bi-encoders are then used as feature extractors; they are frozen and followed by a new compound decoder. The compound decoder $\pi(\cdot)$, parameterized by θ, receives a sum of the encoding vectors $h = z^p + z^c$ and a positive compound c^+ as inputs:

$$\hat{c}^+_{t+1} = \pi_\theta(\cdot | c^+_{1:t}, h_t) \quad \text{where} \quad h_t = z^p_t + z^c_t \,.$$

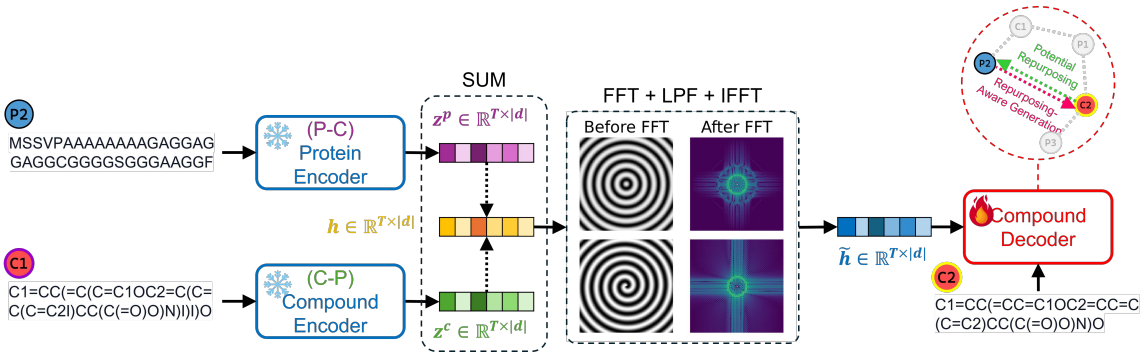
Here, $h_t \in \mathbb{R}^{|d|}$ represents a $|d|$-dimensional latent vector of 2-hop relation, *e.g.*, $p^{(2)} \overset{\text{1-hop}}{\rightarrow} c^{(1)}$ $(= \acute{c})$ $\overset{\text{2-hop}}{\rightarrow} p^{(1)}$ (see Figure 2b), from a t-th token perspective. Accordingly, feeding the compound decoder with a positive compound as a label enables it to learn potential repurposing relationships that emerge from an additional third-hop edge, *e.g.*, $\cdots \overset{\text{2-hop}}{\rightarrow} p^{(1)} \overset{\text{3-hop}}{\rightarrow} c^{(2)}$ $(= c^+)$. Putting it all to-



(a) Bi-directioanl Compound-Protein Relations Embedding



(b) Repurposing-Aware Molecule Generation



(c) Model Architecture

Figure 2: Overview of Repurformer

gether, the loss function is defined as follows:

$$\ln \pi_\theta(c^+|p,c) = \ln \prod_{t=1}^{T_c} \pi_\theta(c^+_{t+1}|c^+_{1:t},p,\acute{c}) = \sum_{t=1}^{T_c} \ln \pi_\theta(c^+_{t+1}|c^+_{1:t},p,\acute{c})$$

Fast Fourier Transform (FFT) The Fourier transform decomposes a function into its constituent frequencies using complex exponentials (sinusoids) as basis functions (Heckbert, 1995; Lee-Thorp et al., 2021). Given a sequence $\{x_1, \cdots, x_T\}$, the *discrete Fourier Transform* (DFT) is defined by the formula:

$$X_k = \sum_{t=0}^{T-1} x_t e^{-\frac{2\pi i}{T} tk}, \quad 0 \le k \le T-1$$

where X_k is the k-th frequency component, x_t is the t-th time-domain signal, and i is the imaginary unit. Calculating the DFT directly has a complexity of $O(T^2)$, which can be inefficient for large datasets. To address this, the *Fast Fourier Transform* (FFT) algorithm was proposed, reducing the complexity to $O(T \log T)$ (Cooley and Tukey, 1965; Brigham, 1988). In this study, we apply the FFT to $h \in \mathbb{R}^{T\times|d|}$ to construct eigenvectors along which the 2-hop propagation occurs. To be specific, the 2D DFT is utilized: one 1D DFT along the sequence dimension, $\mathcal{F}_{\text{seq}}$, and another 1D DFT along the feature dimension, $\mathcal{F}_{\text{dim}}$, keeping real-valued parts only as in Lee-Thorp et al. (2021):

$$H = \Re(\mathcal{F}_{\text{seq}}(\mathcal{F}_{\text{dim}}(h))) \in \mathbb{R}^{T\times|d|}.$$

Note that T is set to the length of a longer sequence; if $T_p > T_c$, then T is set as T_p and vice versa.

Low-Pass Filter (LPF) The Fourier-transformed features H comprise low frequencies that represent a globally smoothed signal and high frequencies that indicate a locally normalized signal. This separation of frequency components allows for distinct interpretations at different scales. For example, Tamkin et al. (2020) applied the discrete cosine transform (DCT) (Rao and Yip, 2014), which is closely related to the DFT, to separate latent information at different scales. They found that low frequencies capture topic-scale context while high frequencies capture word-scale context.

In our setting, a scale can be understood as the number of hops. Specifically, the lower frequency implies a longer propagation through multi-hop relations while the higher one implies a shorter propagation within a single-hop relation. From the repurposing perspective, we need to leverage the longer propagation so that only the multi-hop relations are considered. To achieve this, we can apply the *low-pass filtering* (Pollack, 1948; Costen et al., 1996), which removes the frequency components above a certain cutoff parameter α by setting

$H_{k,d} \leftarrow 0$ for all $k, d > \alpha$. This filtering can be easily implemented using a binary mask:

$$H_{\text{LPF}} = H \odot M$$

where $M = \{m_{t,d} | m_{t,d} \in \{0, 1\}, 1 \leq t \leq T, 1 \leq d \leq |d|\}$ is an one-hot matrix, with $m_{t,d} = 1$ for low-frequency components and $m_{t,d} = 0$ otherwise. Lastly, we transformed H_{LPF} back to the features of an original domain using the inverse FFT (IFFT), before passing it to the compound decoder:

$$\tilde{h} = \mathcal{F}_{\text{dim}}^{-1}(\mathcal{F}_{\text{seq}}^{-1}(H_{\text{LPF}})) \in \mathbb{R}^{T \times |d|} .$$

Implementation Details The structure of the Repurformer is essentially identical to that of pretrained transformers. It consists of encoder and decoder networks, each linearly stacked with 4 layers of 256 dimensions, with each layer divided into 4 heads of 64 dimensions. To tokenize the protein and compound sequences, we utilized existing vocabularies from previous works—the protein vocabulary from Rao et al. (2019) and the compound vocabulary from Honda et al. (2019). For training, we set the number of epochs, batch size, and learning rate to 20, 64, and 5e-05, respectively.

5 Experiments

Experiment Setup We collected data from BindingDB (Gilson et al., 2016) which contains over 2.8 million measured binding affinities of interactions between proteins and drug-like molecules. The collected dataset was then preprocessed to filter out missing values, duplicates, and proteins and compounds with excessively long or short sequences. In particular, given the many-to-many nature of protein-compound relationships, we selected compounds that interact with a reasonable number of individual proteins between 10 and 100, to enable our model to learn various compound structures reacting with different proteins. The resulting dataset comprised 60,719 protein-compound pairs derived from 3,006 proteins and 7,803 compounds. We split this dataset into train and test datasets with 8:2 ratio, ensuring that the proteins interacting with each compound did not overlap between the two sets. Our model was then trained on protein-compound pairs from the train set, representing proteins with amino acid sequences and compounds with canonicalized SMILES strings. We tokenized individual characters from amino acid sequences and SMILES strings, resulting in vocabularies containing 30 characters for proteins and 46 characters for compounds.

Evaluation Metrics To thoroughly assess the effectiveness and reliability of Repurformer, we employed several evaluation metrics, focusing on the generative performance of the model and physicochemical properties and drug-likeness of the molecules it generated. In terms of generative performance, we applied widely accepted metrics for sequence generation tasks: BLEU (Papineni et al., 2002), GLEU (Wu et al., 2016), and F1 score of ROUGE (Lin, 2004). In particular, we used 1- and 2-gram units as the evaluation basis for these generative metrics. We utilized physicochemical properties, specifically molecular weights and log of octanol-water partition coefficients (LogP) (Wildman and Crippen, 1999), to assess the feasibility of molecular structures as drugs. Furthermore, we used other widely used drug-likeness metrics, such as QED (Bickerton et al., 2012), SA (Ertl and Schuffenhauer, 2009), and NP (Ertl et al., 2008), to evaluate the potential effectiveness of the generated molecules as drug-like compounds.

Configurations This study aims to analyze whether the configuration of Repurformer is effective. Given that the distinguishing configuration of Repurformer is the application of FFT with LPF in the embedding space, we conducted comparative experiments with different configuration options:

- SUM Only: This is the baseline configuration. It directly passes h to the compound decoder.
- +FFT: This configuration transforms h to H but does not revert it to $\tilde{h}$.
- +MLP: This configuration adds a single fully-connected layer that mixes the values of h feature-wise.
- +FFT+MLP: This configuration mixes the frequencies of H.
- +FFT+MLP+IFFT w/ auxiliary losses: This configuration mixes the frequencies of H and reverts the mixed H to $\tilde{h}$. Note that L1, L2, and Frobenius norm are added as auxiliary losses to minimize the distance between the MLP output and $\tilde{h}$.

6 Results

Main Results To evaluate Repurformer, we conducted a comparative analysis of 11 configurations, focusing on generative performance, physicochemical properties, and drug-likeness.

		1-gram						2-gram					
		BLEU		GLEU		ROUGE		BLEU		GLEU		ROUGE	
		anc $\acute{c}$	pos c^+	anc $\acute{c}$	pos c^+	anc $\acute{c}$	pos c^+	anc $\acute{c}$	pos c^+	anc $\acute{c}$	pos c^+	anc $\acute{c}$	pos c^+
Baseline (SUM Only)		0.615	0.664	0.618	0.668	0.381	0.398	0.534	0.580	0.543	0.589	0.120	0.127
+FFT		0.155	0.164	0.179	0.188	0.060	0.054	0.098	0.104	0.126	0.132	0.024	0.016
+MLP		0.646	0.692	0.651	0.700	0.399	0.422	0.564	0.604	0.575	0.618	0.133	0.139
+FFT+MLP		0.281	0.289	0.306	0.318	0.011	0.012	0.144	0.156	0.198	0.209	0.004	0.004
+FFT+MLP+IFFT (w/ L1 Loss)		0.583	0.636	0.585	0.640	0.366	0.388	0.511	0.556	0.518	0.565	0.113	0.113
+FFT+MLP+IFFT (w/ L2 Loss)		0.623	0.672	0.627	0.679	0.382	0.398	0.543	0.588	0.553	0.600	0.118	0.123
+FFT+MLP+IFFT (w/ Frobenius Loss)		0.629	0.670	0.635	0.679	0.359	0.367	0.544	0.579	0.556	0.594	0.101	0.104
+FFT+LPF+IFFT (Ours)	α=2	0.620	0.660	0.626	0.670	0.331	0.348	0.512	0.548	0.528	0.567	0.102	0.112
	α=4	0.662	0.690	0.670	0.703	0.385	0.400	0.571	0.598	0.585	0.616	0.147	0.149
	α=6	0.583	0.630	0.587	0.635	0.386	0.406	0.513	0.553	0.521	0.563	0.142	0.150
	α=8	0.606	0.663	0.610	0.667	0.390	0.416	0.532	0.582	0.541	0.591	0.134	0.144

Table 1: Evaluation of Generative Performance. The numbers represent the average (n-gram-based) syntactic similarity of the generated compounds $\hat{c}^+$, which target specific proteins p, to both the anchor compounds $\acute{c}$ and the positive compounds c^+. Note that α is a cutoff parameter.

		MW $[0, \infty]$	LogP $[-\infty, \infty]$
Baseline (SUM Only)		588.506	4.870
+FFT		*N/A*	*N/A*
+MLP		537.230	4.317
+FFT+MLP		533.193	11.559
+FFT+MLP+IFFT (w/ L1)		631.012	4.655
+FFT+MLP+IFFT (w/ L2)		554.490	4.892
+FFT+MLP+IFFT (w/ Frobenius)		572.882	6.092
+FFT+LPF+IFFT (Ours)	α=2	475.473	6.005
	α=4	479.357	3.888
	α=6	584.083	7.442
	α=8	566.942	5.529

Table 2: Evaluation of Physicochemical Properties. The numbers in the MW and LogP columns represent average molecular weights and octanol-water partition coefficients, respectively. By Lipinski's Rule of Five (Lipinski et al., 2012), compounds with MW ≤ 500 and LogP ≤ 5 have good absorption and permeation.

		QED $[0, 1]$	SA $[1, 10]$	NP $[-5, 5]$
Baseline (SUM Only)		0.320	4.033	-0.629
+FFT		*N/A*	*N/A*	*N/A*
+MLP		0.332	3.479	-0.796
+FFT+MLP		0.164	2.086	0.154
+FFT+MLP+IFFT (w/ L1)		0.227	4.209	-0.564
+FFT+MLP+IFFT (w/ L2)		0.355	3.696	-0.659
+FFT+MLP+IFFT (w/ Frobenius)		0.250	4.046	-0.368
+FFT+LPF+IFFT(Ours)	α=2	0.468	4.289	0.072
	α=4	0.598	2.696	-0.682
	α=6	0.254	3.067	-0.679
	α=8	0.352	3.404	-0.984

Table 3: Evaluation of Drug-Likeness. The numbers represent how likely the generated compounds are to be effective drugs. Note that QED, SA, and NP represent a compound's drug-likeness, synthetic accessibility, and natural product-likeness.

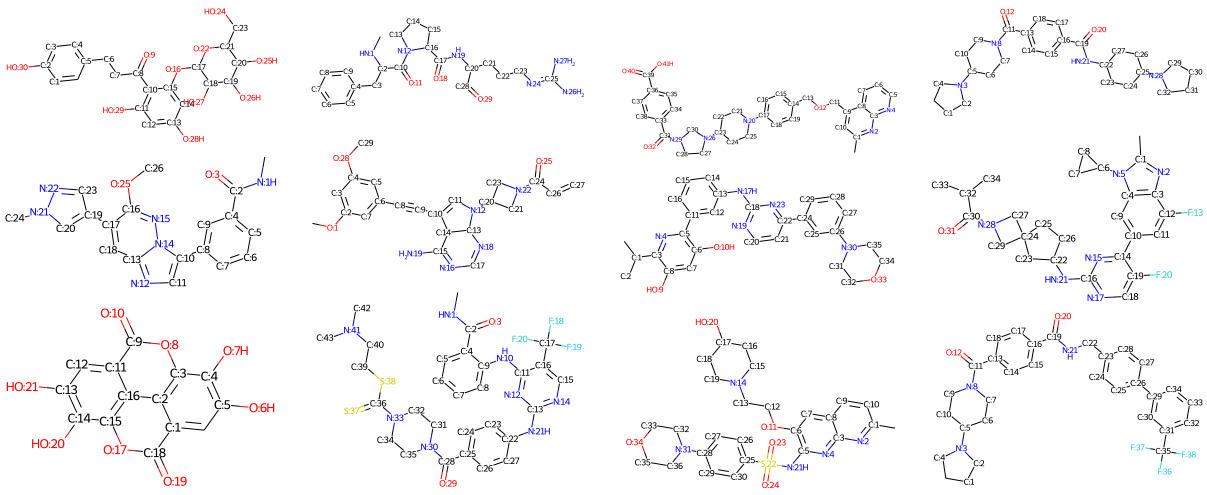
Table 1 shows the similarity of the generated compounds $\hat{c}^+$ to both the anchor $\acute{c}$ and positive c^+ compounds, calculated using BLEU, GLEU, and ROUGE scores. The results indicate that the "+MLP" and "Repurformer with $\alpha = 4$" exhibit remarkable performance compared to other configurations. Notably, the Repurformer ($\alpha = 4$) generated compounds with higher structural similarity to the anchor compounds than those generated by the +MLP configuration. This suggests that Repurformer successfully generates compounds that are potentially repurposable to the target proteins.

In Tables 2 and 3, we can compare the molecular properties of the generated compounds from different configurations. Table 2 shows that Repurformer with $\alpha = 4$ generates compounds that are the most physicochemically desirable. On the other hand, Table 3 shows that the Repurformer with $\alpha = 4$, with $\alpha = 2$, and "+FFT+MLP" configurations had comparative advantages in QED, SA, and NP, respectively. Given that QED is generally considered the most important metric for measuring drug similarity and efficacy, we can emphasize that Repurformer ($\alpha = 4$) excels in generating compounds with the highest potential for effective drug discovery. Figure 3 compares the generation results of the '+MLP' and 'Repurformer ($\alpha = 4$)' configurations.

Performance Comparison To assess the effectiveness of Repurformer as a target-specific molecule generation model, we compared its performance with the existing protein-specific generative approaches as external baseline models, including Transformer-based model (Grechishnikova, 2021) and AlphaDrug (Qian et al., 2022). Transformer-based model utilized the vanilla Transformer architecture (Vaswani et al., 2017) to generate compounds based on target proteins. This model viewed the target-specific molecule genera-

	1-gram				2-gram				Physicochemical Properties		Drug-Likeness		
	BLEU		GLEU		BLEU		GLEU						
	anc $\acute{c}$	pos c^+	anc $\acute{c}$	pos c^+	anc $\acute{c}$	pos c^+	anc $\acute{c}$	pos c^+	MW	LogP	QED	SA	NP
Repurformer (α=4)	0.662	0.690	0.670	0.703	0.571	0.598	0.585	0.616	479.357	3.888	0.598	2.696	-0.682
Transformer	0.541	0.495	0.599	0.572	0.476	0.440	0.533	0.513	9651.296	187.126	0.119	9.037	-0.128
AlphaDrug	0.638	0.652	0.665	0.685	0.555	0.567	0.585	0.603	389.616	2.947	0.507	2.685	-0.842

Table 4: Evaluation of Comparative Performance. Parts of evaluation metrics in terms of generative performance, physicochemical properties, and drug-likeness are used to compare the performance of Repurformer with the existing target-specific molecule generative models, such as Transformer and AlphaDrug.



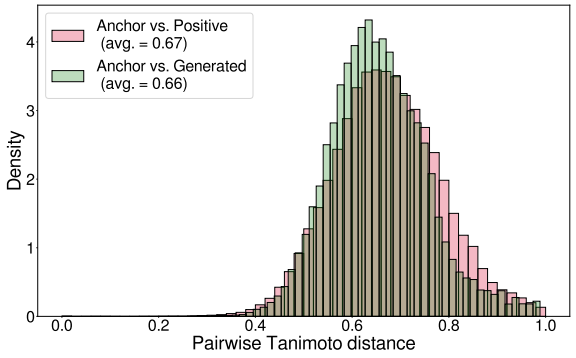
(a) Anchor $\acute{c}$ (b) Positive c^+ (c) Generated $\hat{c}^+$ (+MLP) (d) Generated $\hat{c}^+$ ($\alpha = 4$)

Figure 3: Comparison of 2D Molecule Drawings. From left to right, the drawings represent the anchor $\acute{c}$, positive c^+, and generated compounds $\hat{c}^+$, respectively. $\hat{c}^+$ is expected to interact with the target protein to which $\acute{c}$ interacts.

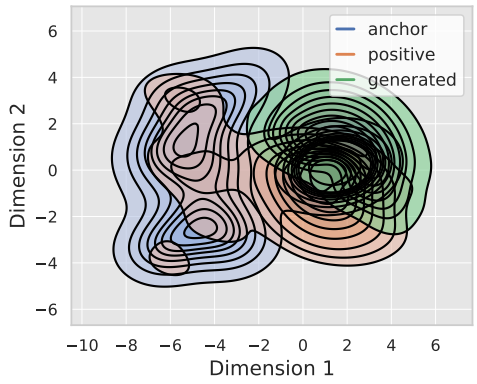
tion as a translational task, converting amino acid sequence into SMILES strings. AlphaDrug modified the vanilla Transformer by introducing skip-connections between its encoders and decoders, facilitating the joint embedding of target proteins and molecules. In addition, it employed a Monte Carlo tree search algorithm for the conditioned generation of novel molecules based on specific target proteins.

To ensure a fair comparison, we trained the external baseline models using our dataset using the same experiment setup and evaluation metrics as for Repurformer. Table 4 presents the performance comparison between our best configuration (Repurformer with α = 4) and the external baseline models. The comparison results demonstrate that the Repurformer (α = 4) outperformed the existing approaches on most evaluation metrics. In particular, our model generated compounds with high structural similarity to both the anchor and positive compounds than those generated by the external baseline models. This suggests that Repurformer can generate not only realistic but diverse compounds with methodological considerations for drug repurposability. Regarding drug-likeness, our model achieved the highest performance only on QED. Although the Transformer-based model excelled in SA and NP, the feasibility of its generated compounds is questionable due to its exceptionally high scores in physicochemical properties, which indicate the compounds might not be suitable as medicines. This is further validated by the evaluation of compound validity, as illustrated in Figure 8 in the Appendix. The compounds generated by the Transformer-based model were significantly less valid compared to those generated by Repurformer.

Mitigation of Sample Bias Figure 4a shows that the distance distribution of the generated compounds to the anchor compounds is similar to that of the positive compounds. We calculated the distance over the fingerprint domain to consider the patterns of molecular substructure. The result implies that the generated and positive compounds have different substructures from the anchor compounds to the same extent. However, Figure 4a compares *"the relative distances"* of the generated and positive compounds to the anchor compounds



(a) Distribution of Pairwise Tanimoto Distances: Generated vs. Anchor and Positive vs. Anchor. The distance was defined over the molecular fingerprints domain.



(b) 2D Gaussian KDE plots for the anchor, positive, and generated compounds. Before employing KDE, each compound was converted to an embedding vector using successive applications of Word2Vec and t-SNE embeddings.

Figure 4: (a) illustrates the distance distribution from the molecular fingerprint perspective. (b) describes the estimated two-dimensional Gaussian distribution of anchor, positive, and generated compounds.

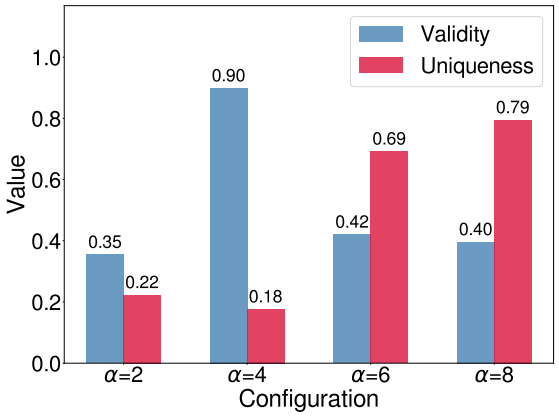


Figure 5: Validity-Uniqueness Trade-off at different values of α. Note that validity represents the quality of generated samples, while uniqueness represents the diversity of generated samples.

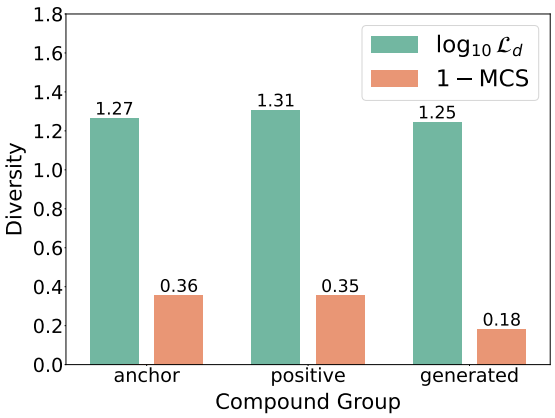


Figure 6: Internal Diversity per Compound Group. $\log_{10}\mathcal{L}_d$ measures the syntactic difference of SMILES strings, while MCS distance (1-MCS) measures the semantic dissimilarity at the atomic level.

"at the substructure level," making it difficult to directly compare the absolute distances to each other at the holistic level.

Figure 4b visualizes the overlapping representations among the anchor, positive, and generated compounds, *"directly comparing their absolute distances at the holistic level."* To do this, we extracted SMILES word embeddings (*e.g.,* C, N, F, =, +, [,], etc.) using Word2Vec (Mikolov et al., 2013) and defined the holistic representation of each molecule as the summation of these word embeddings. We then projected the holistic representation of each molecule into a 2-dimensional space by t-SNE (Van der Maaten and Hinton, 2008). Since t-SNE embeddings preserve pairwise similarities of high-dimensional data as neighboring points in a low-dimension, it allows for direct comparison of absolute distances between samples. Finally, we applied Gaussian kernel density estimation (KDE) to visualize the distribution of t-SNE embeddings.

The results from Figures 4a and 4b indicate that the generated compounds are more similar to the positive compounds than to the anchor compounds, both relatively and absolutely, and at substructural and holistic levels. This suggests that Repurformer successfully addressed the sample bias problem, creating substitutes for anchor compounds that resemble positive compounds.

Existence of Mode Collapse Mode collapse refers to a phenomenon where the generative model creates high-quality samples at the expense of in-distribution diversity (Adiga et al., 2018). In this

section, we demonstrate that Repurformer suffers from mode collapse and thus the *"internal"* diversity of generated compounds is relatively lower than anchor and positive compounds.

Figure 5 illustrates the negative relationship between the validity and uniqueness of the generated compounds by different values of α. Validity represents the ratio of samples that can be depicted as 2D molecular drawings by RDKit (*i.e.,* the quality of generation), while uniqueness represents the ratio of non-duplicated samples (*i.e.,* the diversity of generation). As α increases, we observe that uniqueness increases but validity decreases. This is an expected outcome given that the low-frequency signals represent global structure whereas the high-frequency signals represent local structure. For example, low-frequency signals (*i.e.,* lower α) focus on the most fundamental structures, increasing the validity of generated compounds but reducing their uniqueness. Conversely, high-frequency signals (*i.e.,* higher α) focus on local details, increasing the uniqueness of generated compounds but reducing their structural validity. In short, Figure 5 demonstrates that Repurformer may be susceptible to the validity-uniqueness trade-off, *i.e.,* mode collapse, and thus α must be carefully selected.

Figure 6 describes the *"internal"* diversity of valid compounds generated by Repurformer with $\alpha = 4$, along with anchor and positive compounds that share the same target proteins with the generated ones. Following Pereira et al. (2021), we evaluated the internal diversity within each compound group using two metrics: Levenshtein distance ($\mathcal{L}_d$) and maximum common substructure (MCS). The Levenshtein distance (Levenshtein et al., 1966), also known as edit distance, measures the difference between two SMILES strings by calculating the minimum number of insertions, deletions, and replacements needed to make the strings identical. On the other hand, the MCS (Cao et al., 2008) measures the ratio of the number of atoms in the maximum common substructure of two compounds to their total number of atoms. Since the MCS represents a similarity score normalized between 0 and 1, the MCS distance can be obtained by 1-MCS, which captures the atom-level dissimilarities of compounds. The diversity of each compound group was computed by averaging the pairwise distances of all compounds. The result indicates that the internal diversity of the generated compounds is lower than that of the anchor and positive compounds, suggesting mode collapse in Repurformer.

Note that the existence of mode collapse does not contradict the mitigation of sample bias. Mode collapse refers to less internal diversity among generated compounds, while mitigating sample bias involves creating substitutes for anchor compounds that resemble positive ones, thus increasing diversity between the anchor and generated compounds.

7 Limitations

This study has some limitations. First, due to inconsistencies between the tokens in our dataset and those we borrowed from previous research, some generated outputs contained <UNK> tokens, which had to be excluded. Second, the study lacks experiments on binding affinity, which are necessary to evaluate how strongly the generated compounds bind to proteins. These limitations must be addressed in future research.

8 Concluding Remarks

In this study, we introduced Repurformer, a novel model designed to address the sample bias problem in *de novo* molecule generation by leveraging multi-hop relationships. Repurformer integrates bidirectional pretraining with Fast Fourier Transform and low-pass filtering, to capture complex interactions between proteins and compounds. This approach focuses on low-frequency components, corresponding to longer propagation through multi-hop protein-compound interactions. The results show that Repurformer successfully generates valid and diverse molecules.

Building on these positive results, there are several promising directions for future improvement. Enhancing the backbone architecture by incorporating advanced models like diffusion or graph neural networks and techniques such as contrastive learning could further improve Repurformer's ability to capture multi-hop protein-compound interactions. The results from Figures 5 and 6 also suggest promising directions to improve Repurformer such as leveraging reinforcement learning to maximize diversity rewards or introducing Wasserstein loss to address mode collapse. Additionally, while our current experiments have shown the potential of Repurformer, it is critical to validate its applicability in real-world scenarios. Therefore, we need to verify the performance of Repurformer on existing drug repurposing cases. Considering these aspects will strengthen the practical implications and utilities of Repurformer.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00275796).

References

Sudarshan Adiga, Mohamed Adel Attia, Wei-Ting Chang, and Ravi Tandon. 2018. On the tradeoff between mode collapse and sample quality in generative adversarial networks. In *2018 IEEE global conference on signal and information processing (GlobalSIP)*, pages 1184–1188. IEEE.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. 2022. MolGPT: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. 2012. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98.

E Oran Brigham. 1988. *The fast Fourier transform and its applications*. Prentice-Hall, Inc.

Yiqun Cao, Tao Jiang, and Thomas Girke. 2008. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics*, 24(13):i366–i374.

James W Cooley and John W Tukey. 1965. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301.

Nicholas P Costen, Denis M Parker, and Ian Craw. 1996. Effects of high-pass and low-pass spatial filtering on face identification. *Perception & psychophysics*, 58:602–612.

Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. 2018. Syntax-directed variational autoencoder for molecule generation. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Peter Ertl, Silvio Roggo, and Ansgar Schuffenhauer. 2008. Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries. *Journal of Chemical Information and Modeling*, 48(1):68–74.

Peter Ertl and Ansgar Schuffenhauer. 2009. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8.

Michael K. Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. 2016. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1):D1045–D1053.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Daria Grechishnikova. 2021. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Scientific Reports*, 11(1):321.

Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. 2018. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. *arXiv preprint*. ArXiv:1705.10843 [cs, stat].

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. 2018. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276.

Paul Heckbert. 1995. Fourier transforms and the fast fourier transform (fft) algorithm. *Computer Graphics*, 2(1995):15–463.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Preprint*, arXiv:2006.11239.

Shion Honda, Shoi Shi, and Hiroki R Ueda. 2019. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*.

Lei Huang, Hengtong Zhang, Tingyang Xu, and Ka-Chun Wong. 2023. MDM: Molecular diffusion model for 3D molecule generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5105–5112.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2018. Junction tree variational autoencoder for molecular graph generation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2323–2332. PMLR. ISSN: 2640-3498.

Diederik P. Kingma and Max Welling. 2022. Auto-encoding variational bayes. *Preprint*, arxiv:1312.6114.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. 2012. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 64:4–17.

Łukasz Maziarka, Agnieszka Pocha, Jan Kaczmarczyk, Krzysztof Rataj, Tomasz Danel, and Michał Warchoł. 2020. Mol-CycleGAN: a generative model for molecular optimization. *Journal of Cheminformatics*, 12(1):2.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Brenton P. Munson, Michael Chen, Audrey Bogosian, Jason F. Kreisberg, Katherine Licon, Ruben Abagyan, Brent M. Kuenzi, and Trey Ideker. 2024. De novo generation of multi-target compounds using deep generative chemistry. *Nature Communications*, 15(1):3636.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tiago Pereira, Maryam Abbasi, Bernardete Ribeiro, and Joel P Arrais. 2021. Diversity oriented deep reinforcement learning for targeted molecule generation. *Journal of cheminformatics*, 13(1):21.

Irwin Pollack. 1948. Effects of high pass and low pass filtering on the intelligibility of speech in noise. *The Journal of the Acoustical Society of America*, 20(3):259–266.

Hao Qian, Cheng Lin, Dengwei Zhao, Shikui Tu, and Lei Xu. 2022. AlphaDrug: Protein target specific de novo molecular generation. *PNAS Nexus*, 1(4):pgac227.

K Ramamohan Rao and Ping Yip. 2014. *Discrete cosine transform: algorithms, advantages, applications*. Academic press.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. 2019. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32.

Benjamin Sanchez-Lengeling, Carlos Outeiral, and Gabriel L Guimaraes. 2017. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC).

Alex Tamkin, Dan Jurafsky, and Noah Goodman. 2020. Language through a prism: A spectral approach for multiscale language representations. *Advances in Neural Information Processing Systems*, 33:5492–5504.

Cheng Tan, Zhangyang Gao, and Stan Z. Li. 2022. Target-aware molecular graph generation. *arXiv preprint*. ArXiv:2202.04829 [cs].

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Scott A. Wildman and Gordon M. Crippen. 1999. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arxiv:1609.08144.

Minkai Xu, Alexander S. Powers, Ron O. Dror, Stefano Ermon, and Jure Leskovec. 2023. Geometric latent diffusion models for 3D molecule generation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 38592–38610. PMLR. ISSN: 2640-3498.

A Supplementary Materials

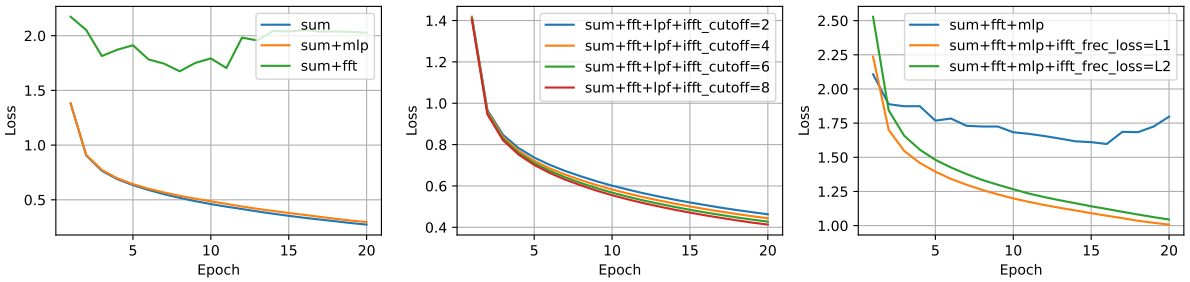


Figure 7: Comparison of Training Performance with Different Configurations. When embedding vectors from protein and compound encoders are mapped to the frequency domain using Fourier Transform (FFT), training performance does not improve unless they are transformed back to the original domain with an inverse Fourier Transform (iFFT). This indicates that applying FFT in the latent space leads to alignment issues between the encoders and the compound decoder.

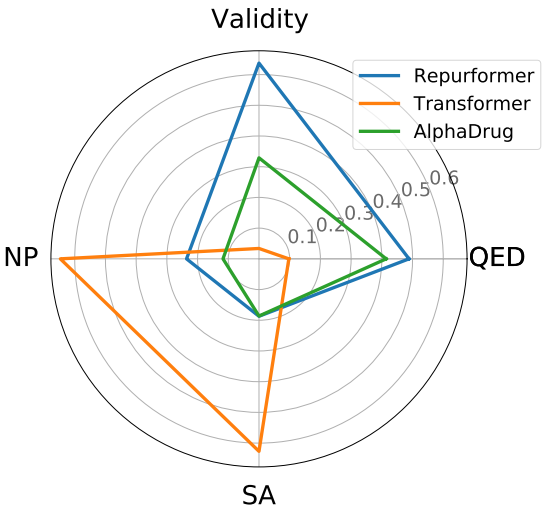


Figure 8: Comparison of Validity and Drug-Likeness Metrics. Validity, QED, SA, and NP scores were normalized to the same scale. Although the Transformer-based model (Grechishnikova, 2021) showed relatively higher SA and NP scores, its validity is extremely low. This indicates that the compounds generated by the Transformer-based model are not of sufficient quality to be considered as drug candidates.