Phenotypic Profile-Informed Generation of Drug-Like Molecules via Dual-Channel Variational Autoencoders

Hui Liu , **Shiye Tian** , **Xuejun Liu**[*]

College of Computer and Information Engineering, Nanjing Tech University, Nanjing, 211800, China.
{hliu, xjliu}@njtech.edu.cn,

Abstract

The de novo generation of drug-like molecules capable of inducing desirable phenotypic changes is receiving increasing attention. However, previous methods predominantly rely on expression profiles to guide molecule generation, but overlook the perturbative effect of the molecules on cellular contexts. To overcome this limitation, we propose SmilesGEN, a novel generative model based on variational autoencoder (VAE) architecture to generate molecules with potential therapeutic effects. SmilesGEN integrates a pre-trained drug VAE (SmilesNet) with an expression profile VAE (ProfileNet), jointly modeling the interplay between drug perturbations and transcriptional responses in a common latent space. Specifically, ProfileNet is imposed to reconstruct pre-treatment expression profiles when eliminating drug-induced perturbations in the latent space, while SmilesNet is informed by desired expression profiles to generate drug-like molecules. Our empirical experiments demonstrate that SmilesGEN outperforms current state-of-the-art models in generating molecules with higher degree of validity, uniqueness, novelty, as well as higher Tanimoto similarity to known ligands targeting the relevant proteins. Moreover, we evaluate SmilesGEN for scaffold-based molecule optimization and generation of therapeutic agents, and confirmed its superior performance in generating molecules with higher similarity to approved drugs. SmilesGEN establishes a robust framework that leverages gene signatures to generate drug-like molecules that hold promising potential to induce desirable cellular phenotypic changes. The source code and datasets are available at: https://github.com/hliulab/SmilesGEN.

1 Introduction

In the realm of drug discovery, target-based drug design (TDD) has emerged as the dominant strategy over the past few decades, driven by advancements in molecular biology and proteomics [Swinney and Anthony, 2011; Attwood et al., 2021]. TDD is centered on the identification and validation of specific molecular targets-typically proteins or RNAs-that play a critical role in disease processes [Munson et al., 2024]. Once a target is confirmed, compounds are designed or screened based on their ability to interact with the target in a manner that modulates its activity, thereby leading to therapeutic benefits [Meissner et al., 2022]. However, the process of molecular target validation is inherently intricate and resource-intensive, which result in a narrow focus on well-characterized targets [Vincent et al., 2022]. Additionally, TDD is vulnerable to off-target effects, where drugs interact with unintended targets, potentially causing unpredictable and sometimes severe side effects [Sadri, 2023].

In contrast, phenotypic drug discovery (PDD) adopts a different approach by concentrating on the observable effects of compounds on disease-relevant biological systems [Berg, 2021; Vincent et al., 2022; Moffat et al., 2017]. In fact, PDD has a long and established history, with many new drugs identified based on the observed therapeutic effects on disease phenotypes [Vincent et al., 2022]. A major advantage of PDD is that it does not require prior knowledge of specific drug targets, making it particularly advantageous in cases where the underlying molecular mechanism of a disease is not well understood [Moffat et al., 2017; Meissner et al., 2022]. However, the absence of a well-defined molecular target in phenotypic drug discovery poses significant challenges on refining the chemical structure of lead compounds to optimize their efficacy, safety, and drug-like properties. A promising strategy to overcome these challenges involves leveraging transcriptional response as indicators of phenotypic changes to guide the drug design and optimization process [Malandraki-Miller and Riley, 2021]. Expression profiles provide a detailed, high-resolution snapshot of how a compound influences cellular pathways and processes, offering a molecule-level understanding of the observed phenotypic outcomes. By comparing the expression profiles of treated and untreated cells, researchers can identify potential drug candidates that elicit desired phenotypic shifts. Therefore, expression profiles can serve as phenotypic signatures in the discovery process. Thereafter, the terms "expression profile" and "phenotypic profile" are used interchangeably in this study.

Several large-scale assays have been conducted to measure a wealth of drug-induced transcriptional responses *in*

vitro. For example, the L1000 platform [Subramanian *et al.*, 2017] offers a cost-effective, high-throughput method for generating expression profiles subjected to drug perturbations. Based on such data resource, deep generative models, such as generative adversarial networks (GANs) [Goodfellow *et al.*, 2014] and variational autoencoders (VAEs) [Kingma and Welling, 2013], have been proposed to design molecules with potential to elicit intended bioactivity [Malandraki-Miller and Riley, 2021]. However, existing approaches generate molecules without considering the cellular context relevant to specific disease. To guide drug discovery using phenotypic profile, several methods have been developed to utilize drug-induced expression profiles as conditional variables for molecular generation. These methods include ExpressionGAN [Méndez-Lucio *et al.*, 2020], TRIOMPHE [Kaitoh and Yamanishi, 2021], GxVAEs [Li and Yamanishi, 2024], and Gex2SGen [Das *et al.*, 2023]. Despite their advancement, these phenotype-based methods have two primary limitations. First, they rely exclusively on post-treatment expression profiles to guide drug generation, neglecting the change of expression profiles induced by drug perturbations. Second, they do not model the interplay between molecules and cellular environment, which is crucial for generating molecules with potential to induce desired phenotypic changes.

To address these issues, we propose SmilesGEN, a novel deep generative framework designed for *de novo* generation of drug-like molecules. SmilesGEN uniquely integrates drug-induced gene expression changes with phenotype-informed molecule generation within a unified framework. Unlike previous approaches, SmilesGEN uses dual-channel variational autoencoders to explicitly model the interplay between molecules and cellular environment. By mapping both molecules and expression profiles to a common latent space, the drug-induced phenotypic changes can be straightforwardly formulated. SmilesGEN is designed to satisfy two critical constraints. First, by mitigating the disturbative effects of molecules on post-treatment phenotypic profiles, SmilesGEN aims to restore the pre-treatment cellular state. Second, when conditioned on desired phenotypic profiles, SmilesGEN is engineered to generate molecules with potential to induce such phenotypic outcomes. Extensive experiments demonstrate that SmilesGEN outperforms current state-of-the-art methods, generating molecules with potential bioactivity and desirable drug-like properties. By establishing the link between molecule generation and cellular context, SmilesGEN effectively increases the likelihood of generating molecules that interact productively with disease-associated cellular machinery.

The main contributions of this study are summarized as below:

- **Phenotype-informed *de novo* molecule generation**: Unlike previous methods designed to generate molecules with proper physicochemical property, this study aims to generate molecules that potentially elicit desirable phenotypic outcomes.

- **Explicit formalization of drug perturbative effect on phenotypic change**: this study explicitly formalizes the phenotypic change to capture the net effect of the drug

perturbation, enabling to extract more informative features for molecular generation.

- **Superior performance over state-of-the-art (SOTA) methods**: Extensive experiments demonstrate that our method outperforms two baselines and five previous methods, and generates therapeutic molecules with chemical structures highly similar to those of known ligands.

2 Related Work

2.1 De Novo Drug Generation

Some generative models have been employed to design novel molecular structures, aiming to explore chemical space and generate molecules with both novelty and chemical validity. These approaches often utilize Simplified Molecular Input Line Entry System (SMILES) strings as input, which are well-suited for generative models like recurrent neural networks [Segler *et al.*, 2018; Merk *et al.*, 2018], VAEs [Gómez-Bombarelli *et al.*, 2018; Blaschke *et al.*, 2018; Zhang and Chen, 2022], and Transformers [Wang *et al.*, 2022; Dollar *et al.*, 2021]. Attention-based models leverage attention mechanisms to learn the "molecular grammar" from SMILES strings, enabling the generation of coherent molecular sequences. Despite the advantages of compact and standardized representation provided by SMILES, these methods often produce invalid molecular structures, as the stringed representation of SMILES does not always guarantee chemical validity. So, some methods employ Graph Neural Networks (GNNs) to represent and generate molecular structures [De Cao and Kipf, 2018; Li *et al.*, 2018; Jin *et al.*, 2020], in order to capture the chemical structure of molecules. In addition, reinforcement learning techniques have been incorporated into some methods to optimize the structures of generated molecules [Zhavoronkov *et al.*, 2019; Popova *et al.*, 2018].

2.2 Drug Generation from Expression Profile

With the advancement of RNA sequencing technology, the generation of large-scale multi-omics data has enabled researchers to integrate omics data into generative models for the design of de novo molecules. For instance, TRIOMPHE [Kaitoh and Yamanishi, 2021] employed a variational autoencoder framework conditioned on e n profiles to generate molecules. Gex2SGen [Das *et al.*, 2023] and GxVAEs [Li and Yamanishi, 2024] followed similar ideas that used the desired expression profiles as input to design drug-like molecules capable of eliciting phenotypic changes. Mendez-Lucio et al. [Méndez-Lucio *et al.*, 2020] further employed a generative adversarial network (GAN) in conjunction with expression profiles to automatically design molecules with a high probability to induce therapeutic effects. Although the integration of transcriptomic profiles into generative models marks a significant advancement, these methods overlook the interplay between drug molecules and cellular context.
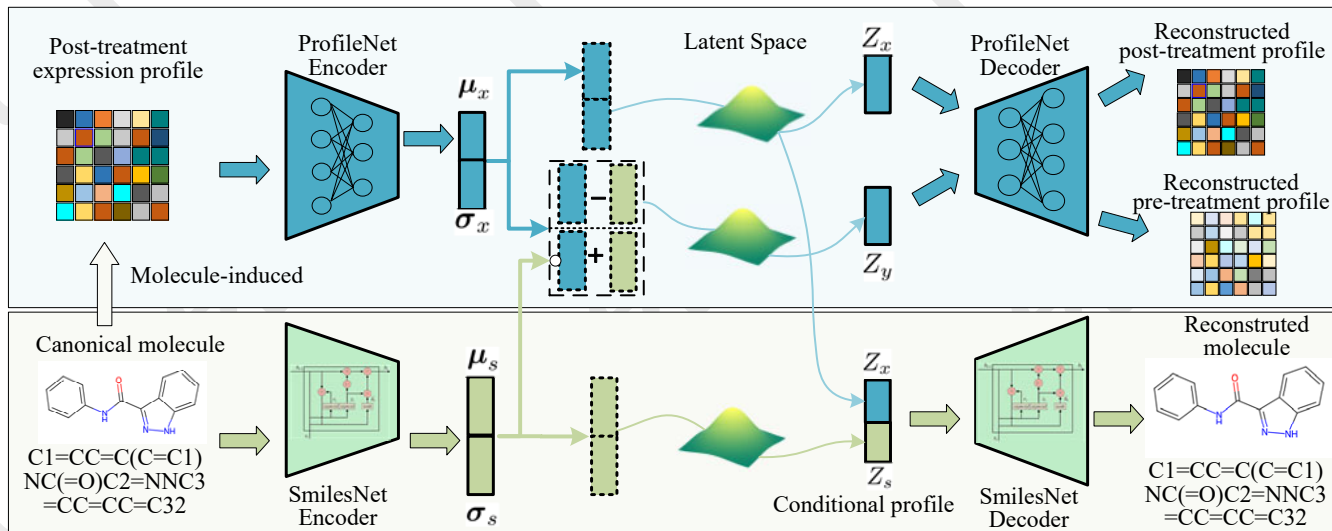
Figure 1: SmilesGEN framework comprises two interacting variational autoencoders, SmilesNet and ProfileNet, which work jointly to extract features from molecules and phenotypic profiles and reconstruct molecules and post-treatment signatures, as well as pre-treatment signatures.

3 Methods

3.1 SimlesGEN Framework

The proposed SmilesGEN framework comprises dual-channel VAEs: a molecule VAE (SmilesNet) and an expression profile VAE (ProfileNet). As shown in Figure 1, SmilesNet is a VAE [Chung *et al.*, 2014] that utilizes gated recurrent unit (GRU) backbone for the encoding and generation of molecules. SmilesNet is pretrained on a large-scale SMILES descriptors, enabling it to explore the extensive chemical space and generate meaningful molecular structures. ProfileNet is another VAE composed of multiple feed-forward layers to encode and decode expression profiles. A key feature of ProfileNet is to explicitly formalize the net effect of drug perturbations on cellular state in the latent space, by enforcing the constraint that the encoded post-treatment expression profile, when adjusted by the molecular representation, should restore the corresponding pre-treatment expression profile. Conceptually, this formalization simulates the therapeutic effects in a reverse direction that removal of drug treatment allows the cellular environment to revert to its unperturbed state. By modeling the dynamic interplay between molecules and cellular environments, our model enhances its capacity to learn informative and expressive features related to both molecule structures and their induced phenotypic changes. Once trained, the encoded expression profiles are fed into the SmilesNet decoder to generate molecules with a high likelihood of eliciting the desired phenotypic outcomes.

3.2 Drug Variational Autoencoder

The molecules are represented using SMILES strings, a compact and standardized notation for describing molecular structures. Let a molecule S be denoted by a string $S = [s_1, s_2, \cdots, s_n]$, in which each character corresponds to an atom or a chemical bond. Suppose the posterior distribution

learned by SmilesNet encoder is denoted as $q_\phi(Z_s|S)$ that is approximate to a normal distribution $\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s)$, where $\boldsymbol{\mu}_s$ and $\boldsymbol{\sigma}_s$ denote the mean and standard deviation. A latent variable Z_s is sampled from $q_\phi(Z_s|S)$ and then used to reconstruct the molecule S following a stochastic process $S \sim p_\varphi(S|Z)$ parameterized by φ:

$$p_\varphi(S|Z_s, Z_x) = \prod_{i=1}^{n} p_\varphi(s_i|s_{1:i-1}, Z_s, Z_x). \quad (1)$$

in which Z_x represents the conditional variable drawn from the probability distribution learned from drug-induced phenotypic profiles (see below). During the pre-training phase, Z_x in is substituted with Z_s. SmilesNet tries to maximize the lower bound of the true log-marginal likelihood as follows:

$$\begin{aligned}\mathcal{L}_S(\phi, \varphi) = &E_{Z_s \sim q_\phi(Z_s|S)}(\log p_\varphi(S|Z_s, Z_x)) \\ &- D_{KL}(q_\phi(Z_s|S)||p_\varphi(Z_s))\end{aligned} \quad (2)$$

where $E(\cdot)$ is an expectation operation, ϕ and φ denote the parameters of the SmilesNet encoder and decoder, respectively.

3.3 Profile Variational Autoencoder

Denote the post-treatment expression profile by $X = (x_1, x_2, \cdots, x_m)$, where each element represents the expression levels of a gene when subjected to a molecule. Similarly, denote by $Y = (y_1, y_2, \cdots, y_m)$ the pre-treatment expression profile of corresponding genes. Suppose the posterior distribution learned by ProfileNet encoder is denoted as $q_\theta(Z_x|X)$ that is approximate to a normal distribution $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x)$, where $\boldsymbol{\mu}_x$ and $\boldsymbol{\sigma}_x$ denote the mean and standard deviation. From the posterior distribution, a latent variable Z_x is sampled and taken as the conditional variable by ProfileNet decoder to reconstruct X following a generative process $X \sim p_\psi(X|Z_x)$. ProfileNet aims to maximize the

marginal likelihood to generate the expression profile from the generative process:

$$\mathcal{L}_X(\theta, \psi) = E_{Z_x \sim q_\theta(Z_x|X)}(\log p_\psi(X|Z_x)) \\ - D_{\text{KL}}(q_\theta(Z_x|X)||p_\psi(Z_x)) \quad (3)$$

where θ and ψ are actually the parameters of the ProfileNet encoder and decoder, respectively.

3.4 Modeling Drug-Induced Phenotypic Change

We aim to explicitly formalize the impact of drugs on cellular states within the latent space. Following the assumption of linear additivity that has been widely used in modeling cellular response to drug perturbations [Hetzel *et al.*, 2022; Huang and Liu, 2024], we assume that the linear subtraction of two random variables capture the data distribution of the pre-treatment expression profiles (control state), when mitigate the effect of drug perturbation from post-treatment expression profiles. As a result, we can draw samples from the resulting random variable to reconstruct the pre-treatment expression profile using the ProfileNet decoder. Denote the variable Z_y is approximated by the reparameterization trick:

$$Z_y = (\boldsymbol{\mu}_x - \boldsymbol{\mu}_s) + (\boldsymbol{\sigma}_x + \boldsymbol{\sigma}_s)\boldsymbol{\epsilon}, \quad (4)$$

in which $\boldsymbol{\epsilon}$ is sampled from a standard normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Following the same generative process $p_\psi(\cdot)$, we expect to reconstruct the pre-treatment expression profile Y when conditioned on Z_y. So, we define another loss function as below:

$$\mathcal{L}_Y(\theta, \psi) = E_{Z_y \sim \mathcal{N}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_s, \boldsymbol{\sigma}_x + \boldsymbol{\sigma}_s)}(\log p_\psi(Y|Z_y)) \\ - D_{\text{KL}}(q_\theta(Z_y|X)||p_\psi(Z_y)) \quad (5)$$

3.5 Full Objective

Taken together, we define the total loss function as below:

$$\mathcal{L}(\phi, \varphi, \theta, \psi) = \mathcal{L}_S(\phi, \varphi) + \mathcal{L}_X(\theta, \psi) + \mathcal{L}_Y(\theta, \psi) \quad (6)$$

The parameter are optimized by minimizing the total loss function. The training process of SmileGEN consists of two distinct stages. As outlined in Algorithm 1, the first stage involves pretraining SmilesNet on large-scale dataset of molecules. Once pretraining is completed, the encoder of SmilesNet is frozen. In subsequent training stage, the model takes as input the drug S, its perturbed expression profile X, and unperturbed expression profile Y to reconstruct them. The parameters are optimized by minimizing the objective loss function. After training, we retain the ProfileNet encoder and SmilesNet decoder to generate drug-like molecules from perturbed expression profiles.

In our practice, we employed the bidirectional GRUs as the backbone of SmilesNet encoder and decoder. Both the encoder and decoder were composed of three hidden layers of size 192. The ProfileNet encoder and decoder adopted multiple forward-feed layers with sizes 512, 256, and 192, respectively. During the pretraining stage, SmilesNet was optimized using the Adam optimizer with a learning rate of 5e-4 and a dropout rate of 0.1. The maximum length of the generated SMILES strings was restricted to 100 characters. In the subsequent training stage, both SmilesNet and ProfileNet were

---

Algorithm 1 Training Procedure of SmilesGEN.

Data: SMILES string $\mathbf{S}$, drug-perturbed profiles $\mathbf{X}$, unperturbed profiles $\mathbf{Y}$
1: /*Pre-training of SmilesNet*/
2: **for** $i \leftarrow 1\ to\ \mathbf{SEpoch}$ **do**
3:     Train SmilesNet to minimize Eq.(2).
4:     Update parameters φ and ϕ.
5: **end for**
6: Freeze SmilesNet encoder parameter ϕ.
7: /*Joint Training of SmilesNet and ProfileNet*/
8: **for** $j \leftarrow 1\ to\ \mathbf{GEpochs}$ **do**
9:     Train SmilesNet and ProfileNet to minimize Eq.(6).
10:     Update parameters θ, ψ, and φ.
11: **end for**

---

optimized using Adam optimizer with a reduced learning rate of 1e-4. The maximum epochs of the pre-training and training stages were set to 100 and 500, respectively. The batch size for both stages was set at 64. SmilesGEN was implemented using PyTorch 2.3.0. The training and all evaluation experiments were conducted on a CentOS Linux 8.2.2004 (Core) system, equipped with a GeForce RTX 4090 GPU and 128GB of memory.

4 Experiments

4.1 Datasets

For performance evaluation, we used the expression profiles induced by the genetic perturbation of 10 commonly target genes associated with the treatment of prevalent cancers. Among them, eight genes (AKT1, AKT2, AURKB, CTSK, EGFR, HDAC1, MTOR, and PIK3CA) were knockdown, and two genes (SMAD3 and TP53) were over-expressed by genetic perturbations. To evaluate the generated molecules, we obtained inhibitors and agonists that have been confirmed to target these proteins from the Drug Target Commons (DTC) database [Tang *et al.*, 2018]. We compute the performance metrics between these inhibitors/agonists and generated molecules.

We also utilized the expression profiles of cancer patients from The Cancer Genome Atlas (TCGA) repository [Weinstein *et al.*, 2013] to guide the generation of therapeutic molecules. Specifically, we extracted the disease-specific expression profiles of three types of cancers: triple-negative breast cancer (TNBC), colorectal cancer (CRC), and lung adenocarcinoma (LUAD). To evaluate the generated molecules, we retrieved a list of drugs approved for treatment of these cancers from DrugBank database [Wishart *et al.*, 2006]. The generated molecules were then assessed based on their structural similarity to the approved drugs, holding the rationale that structurally similar molecules are likely to exhibit therapeutic effects against the diseases.

4.2 Evaluation Metrics

For the generated molecules, we assessed their validity, uniqueness, novelty, and diversity. High validity and uniqueness are indicators of an effective molecule generation process, while high novelty indicates the model is not overfitting
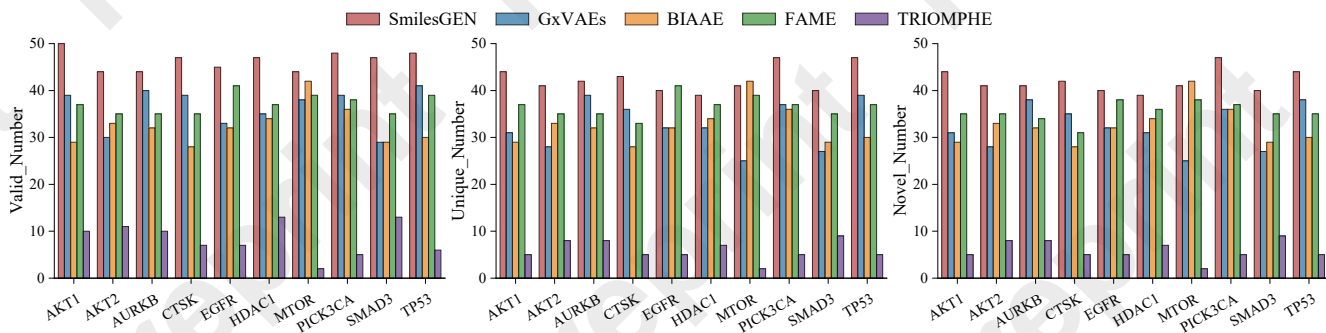
Figure 2: Performance comparison in terms of validity, uniqueness and novelty of generated molecules by SmilesGEN and four competing methods using the expression profiles induced by genetic perturbations.

Target Protein	Euclidean	Cosine	ExpressionGAN	TRIOMPHE	BIAAE	FAME	GxVAEs	SmilesGEN
AKT1	0.35	0.18	0.32	0.23	0.43	0.44	0.66	**0.98**
AKT2	0.36	0.33	0.29	0.50	0.42	0.55	0.38	**0.81**
AURKB	0.29	0.34	0.36	0.25	0.43	0.55	0.50	**0.63**
CTSK	0.18	0.29	0.31	0.32	0.37	0.43	0.41	**0.51**
EGFR	0.26	0.26	0.30	0.47	0.47	0.76	0.58	**0.81**
HDAC1	0.39	0.25	0.34	0.24	0.42	0.46	0.49	**0.56**
MTOR	0.33	0.19	0.39	0.30	0.49	0.42	0.50	**0.75**
PIK3CA	0.25	0.24	0.26	0.32	0.35	0.55	0.43	**0.75**
SMAD3	0.44	0.32	0.44	0.38	0.45	0.52	0.51	**0.76**
TP53	0.40	0.36	0.46	0.59	0.58	0.73	0.71	**1.00**

Table 1: Comparison of highest Tanimoto coefficients achieved the SmilesGEN and seven competing methods

to the training data. High diversity reflects the model's ability to generate molecules beyond the molecules in training set. To assess the drug-likeness of the generated molecules, we employed the quantitative estimate of drug-likeness (QED) score. Additionally, we evaluated the similarity between the generated molecules and known ligands of target proteins using Tanimoto similarity based on molecular fingerprints. This multifaceted assessment ensures the effectiveness and applicability of our model in generating molecules with potential therapeutic effect.

4.3 Evaluation on Benchmark dataset

To validate the model's capability to generate molecules based on desired expression profiles, we conducted experiments using the expression profile induced by genetic perturbations in the MCF7 cell line, a benchmark dataset that has been utilized in previous studies. Specifically, we used the knockdown-induced expression profiles of eight target genes (AKT1, AKT2, AURKB, CTSK, EGFR, HDAC1, MTOR, and PIK3CA) to generate inhibitors, and overexpression-induced profiles of two target genes (SMAD3 and TP53) to generate agonists. We first conducted a comparative analysis of our model against four existing methods in terms of validity, uniqueness, and novelty metrics. The competing methods include GxVAEs [Li and Yamanishi, 2024], BIAAE [Shayakhmetov *et al.*, 2020], FAME [Pham *et al.*, 2022] and TRIOMPHE [Kaitoh and Yamanishi, 2021]. As shown in Figure 2, the experimental results indicate that the molecules generated by our model achieved higher degree of validity,

uniqueness, and novelty than four competing methods.

Next, we evaluated the performance of SmilesGEN against seven existing models by assessing the Tanimoto coefficient between generated molecules and known ligands. To benchmark our method, we included two baseline models based on molecular similarity search (Euclidean and Cosine similarity), alongside five other competing methods: ExpressionGAN [Méndez-Lucio *et al.*, 2020], TRIOMPHE [Kaitoh and Yamanishi, 2021], BIAAE [Shayakhmetov *et al.*, 2020], FAME [Pham *et al.*, 2022], and GxVAEs [Li and Yamanishi, 2024]. For each target gene, 50 molecules were generated using each method, and the Tanimoto coefficients were computed between the generated molecules and the known ligands specific to that target gene. The highest Tanimoto coefficients for each method were reported, as shown in Table 1. Notably, SmilesGEN consistently achieved significantly higher Tanimoto coefficients compared to all other methods. Although the current state-of-the-art model, GxVAEs, outperformed the other competing methods, its performance remained notably inferior to that of SmilesGEN. Furthermore, Figure S1 showcases the molecular structures with the highest Tanimoto similarity to known ligands, as generated by SmilesGEN and five competing methods. Such visual comparisons further underscore SmilesGEN's superior ability to generate molecules with the highly similar structures to known ligands.

Moreover, we further evaluated the drug-likeness and diversity of the molecules generated by SmilesGEN. For each target gene, we generated 1,000 candidate molecules and an-

(a)   (b)
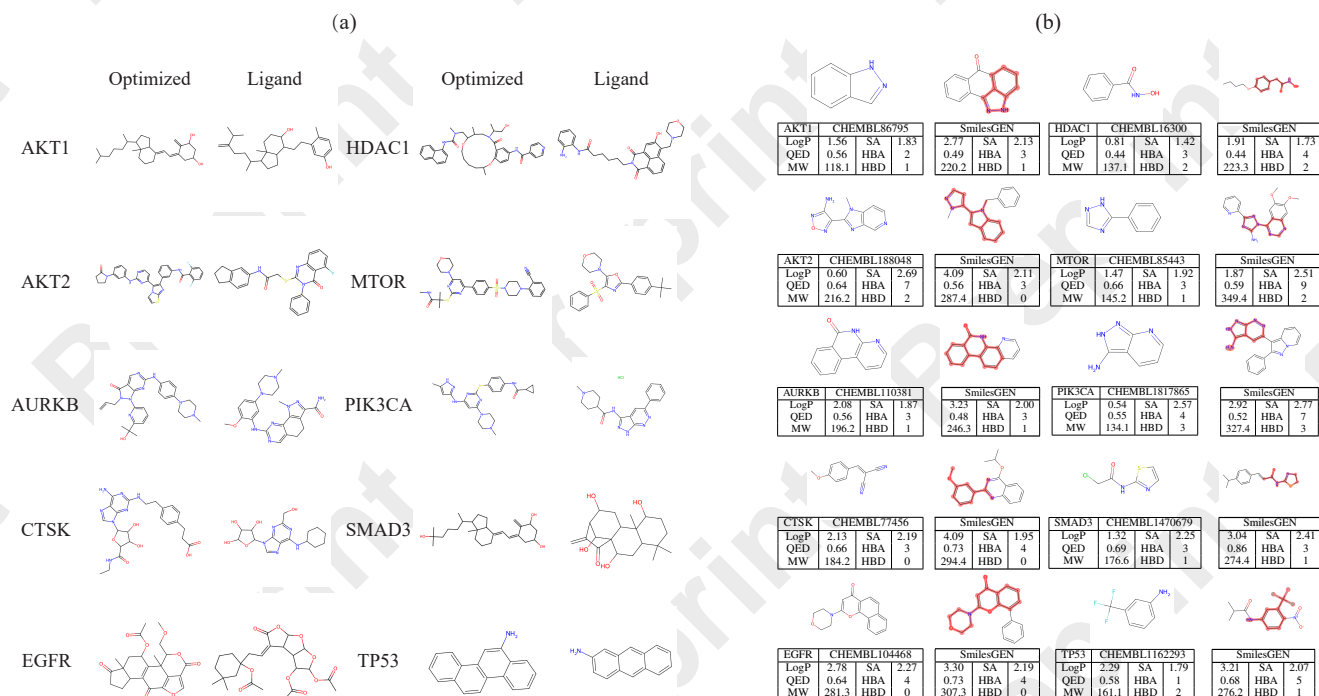

Figure 3: Performance evaluation for scaffold-based molecule optimization. (a) Optimized molecules based on benzene ring toward specific gene targets and their nearest known ligands. (b) Comparison between optimized molecules and corresponding short-chain ligands.

alyzed their physicochemical properties, including the quantitative estimate of drug-likeness (QED), logarithm of partition coefficient (LogP), molecular weight (MW), hydrogen bond acceptor (HBA), and hydrogen bond donor (HBD). The kernel density distributions of these physicochemical properties for both the generated molecules and the known ligands are presented in Figure S2. The results revealed that the QED distribution for our generated molecules exhibited a peak around 0.7, while 90% known ligands displayed peak QED values below 0.55, implying the generated molecules have higher degree of drug-likeness than known ligands. For other physicochemical metrics, the broader distribution range of the generated molecules compared to known ligands further confirmed the better diversity achieved by our method.

In addition, we applied Lipinski's Rule of Five (logP<10, molecular weight<500, hydrogen bond donors<5, hydrogen bond acceptors<10) to filter the 1,000 generated molecules, and then computed the Tanimoto similarities between the filtered molecules and known ligands. Table S1 showed the details of generated molecules with Tanimoto similarities exceeding 0.7. It is worth noting that among the refined molecules, more than 50% of the molecules generated towards desired expression profiles induced by seven gene perturbations (AKT1, AURKB, EGFR, MTOR, PIK3CA, SMAD3, TP53) exhibited high similarity with known ligands (Tanimoto≥0.7). We showcased the molecular structures of the generated molecules with highest Tanimoto similarity to known ligands (Figure S3). These findings provide strong

evidence of our method's capability to generate structurally diverse molecules with high drug-likeness.

4.4 Profile-Based Molecule Optimization

Molecule optimization is another important task in drug discovery, as it significantly contributes to improving therapeutic efficacy and pharmacokinetic properties. For proof-of-concept purpose, we evaluate our model in optimizing benzene rings—one of the most common scaffolds in drugs—towards active-like molecules, guided by the gene signature induced by genetic perturbations. Specifically, the SMILES of benzene rings and desired expression profiles were taken as input to generate molecular structures. Figure 3(a) presents the optimized molecules alongside their nearest known ligands. The results demonstrate that the generated molecules not only effectively retained the benzene ring scaffold, but also incorporated functional groups containing atoms such as oxygen (O), nitrogen (N), sulfur (S), and fluorine (F), showcasing the model's ability to enhance structural diversity and functionality.

Furthermore, we extended the optimization process to short-chain ligands associated with specific target genes. Figure 3(b) presented the short-chain ligands alongside their optimized counterparts, as well as their physicochemical properties, including LogP, QED, MW, SA, HBA, and HBD. We found that the optimized molecules incorporated additional benzene rings or functional groups, thereby expanding the molecular structures while maintaining or improving
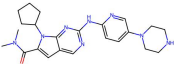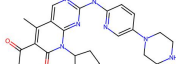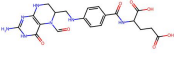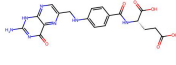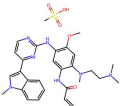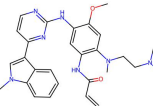
Target Protein	Approved Drug Name	Approved Molecule	Generated Molecule	Tanimoto
TNBC	Ribociclib			**0.56**
CRC	Cetuximab			**0.50**
LUAD	Alectinib			**0.98**

Table 2: Comparison between approved drugs and generated molecules for three types of cancers

their QED values relative to the original short-chain ligands. Meanwhile, other physicochemical properties were enhanced to varying degrees. These findings highlight the strong capability of our model in molecular optimization, offering robust support for drug design and development.

4.5 Therapeutic Molecule Generation

We further applied SmilesGEN to design therapeutic molecules for triple-negative breast cancer (TNBC), colorectal cancer (CRC), and lung adenocarcinoma (LUAD). These cancers are all characterized as malignant tumors, and a significant proportion of patients either do not respond to approved drugs or develop resistance over time. We obtained the expression profiles of patients with these cancers from The Cancer Genome Atlas (TCGA). Given the rationale that molecules capable of reversing disease-specific expression profiles might possess therapeutic potential [Tong *et al.*, 2023; Yamanaka *et al.*, 2023]. Therefore, we averaged the expression profiles across patients of the same cancer type and inverted the averaged expression profiles by multiplying the values by -1. These disease-specific reversal profiles were subsequently input into SmilesGEN to generate candidate therapeutic molecules.

For each type of cancer, we generated 1,000 candidate molecules. After refining the generated molecules using the Lipinski's Rule of Five, we obtained 783, 859, and 662 molecules for TNBC, CRC, and LUAD, respectively. Figure 4 displays the distribution of quantitative estimate of QED values of these molecules compared to approved drugs for treating these diseases. It is noteworthy that the mean QED values of the molecules generated for LUAD and TNBC exceeded 0.75, suggesting that our model is highly effective in generating drug-like molecules with therapeutic potential.

We further calculated the Tanimoto similarity between our generated molecules and known approved drugs. Table 2 highlights the molecules generated by our model that exhibit significant similarity to approved drugs for respective cancers. For example, Ribociclib, an approved drug for TNBC treatment [Li *et al.*, 2022], is a specific inhibitor of cyclin-dependent kinase 4/6 (CDK4/6), used in the treatment of hormone receptor-positive and human epidermal growth factor receptor 2-negative (HR+/HER2-) advanced or metastatic

breast cancer in postmenopausal women. The molecules generated by our model for these three cancers exhibited high Tanimoto similarity to these approved drugs. Notably, when generating therapeutic molecules for lung cancer using the reversal profile, the Tanimoto similarity between the generated molecules and Alectinib Hydrochloride reached 0.98. This high degree of similarity indicates that SmilesGEN effectively captures structural features similar to those of approved drugs, suggesting that our model can generate hit-like molecules with high therapeutic potential.
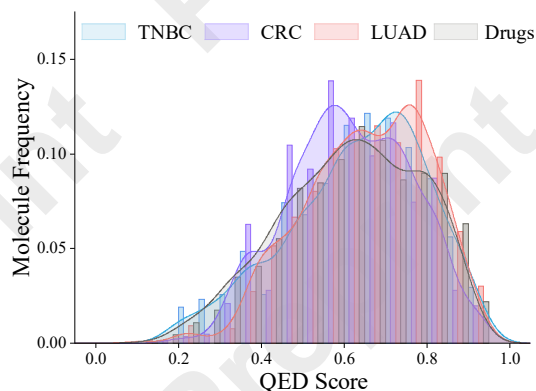

Figure 4: Distribution of QED values of generated molecules using the reversal profiles of three types of cancers.

5 Conclusion

In this paper, we introduce the SmilesGEN, a novel generative model designed to generate hit-like molecules based on desired phenotypic profiles. Our model explicitly models the interplay between molecules and cellular environments in the latent space, driving the model to learn informative and expressive features of molecules and expression profiles. Experimental results show that SmilesGEN outperforms current state-of-the-art models and can generate hit-like molecules with potential therapeutic effects. We believe that SmilesGEN represents an alternative approach to bridging chemistry and biology in the long and difficult road of drug discovery.

Acknowledgments

References

[Attwood *et al.*, 2021] Misty M Attwood, Doriano Fabbro, Aleksandr V Sokolov, Stefan Knapp, and Helgi B Schiöth. Trends in kinase drug discovery: targets, indications and inhibitor design. *Nature Reviews Drug Discovery*, 20(11):839–861, 2021.

[Berg, 2021] Ellen L Berg. The future of phenotypic drug discovery. *Cell chemical biology*, 28(3):424–430, 2021.

[Blaschke *et al.*, 2018] Thomas Blaschke, Marcus Olivecrona, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Application of generative autoencoder in de novo molecular design. *Molecular informatics*, 37(1-2):1700123, 2018.

[Chung *et al.*, 2014] Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, abs/1412.3555, 2014.

[Das *et al.*, 2023] Dibyajyoti Das, Broto Chakrabarty, Rajgopal Srinivasan, and Arijit Roy. Gex2sgen: designing drug-like molecules from desired gene expression signatures. *Journal of Chemical Information and Modeling*, 63(7):1882–1893, 2023.

[De Cao and Kipf, 2018] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[Dollar *et al.*, 2021] Orion Dollar, Nisarg Joshi, David AC Beck, and Jim Pfaendtner. Attention-based generative models for de novo molecular design. *Chemical Science*, 12(24):8362–8372, 2021.

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

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63:139 – 144, 2014.

[Hetzel *et al.*, 2022] Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Fabian Theis, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems*, 35:26711–26722, 2022.

[Huang and Liu, 2024] Wei Huang and Hui Liu. Predicting single-cell cellular responses to perturbations using cycle consistency learning. *Bioinformatics*, 40(Supplement_1):i462–i470, 2024.

[Jin *et al.*, 2020] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *International conference on machine learning*, pages 4839–4848. PMLR, 2020.

[Kaitoh and Yamanishi, 2021] Kazuma Kaitoh and Yoshihiro Yamanishi. Triomphe: transcriptome-based inference and generation of molecules with desired phenotypes by machine learning. *Journal of Chemical Information and Modeling*, 61(9):4303–4320, 2021.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv e-prints*, page arXiv:1312.6114, December 2013.

[Li and Yamanishi, 2024] Chen Li and Yoshihiro Yamanishi. Gxvaes: Two joint vaes generate hit molecules from gene expression profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13455–13463, 2024.

[Li *et al.*, 2018] Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia. Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03324*, 2018.

[Li *et al.*, 2022] Yun Li, Huajun Zhang, Yulia Merkher, Lin Chen, Na Liu, Sergey Leonov, and Yongheng Chen. Recent advances in therapeutic strategies for triple-negative breast cancer. *Journal of hematology & oncology*, 15(1):121, 2022.

[Malandraki-Miller and Riley, 2021] Sophia Malandraki-Miller and Paul R Riley. Use of artificial intelligence to enhance phenotypic drug discovery. *Drug Discovery Today*, 26(4):887–901, 2021.

[Meissner *et al.*, 2022] Felix Meissner, Jennifer Geddes-McAlister, Matthias Mann, and Marcus Bantscheff. The emerging role of mass spectrometry-based proteomics in drug discovery. *Nature Reviews Drug Discovery*, 21(9):637–654, 2022.

[Méndez-Lucio *et al.*, 2020] Oscar Méndez-Lucio, Benoit Baillif, Djork-Arné Clevert, David Rouquié, and Joerg Wichard. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature communications*, 11(1):10, 2020.

[Merk *et al.*, 2018] Daniel Merk, Lukas Friedrich, Francesca Grisoni, and Gisbert Schneider. De novo design of bioactive small molecules by artificial intelligence. *Molecular informatics*, 37(1-2):1700153, 2018.

[Moffat *et al.*, 2017] John G Moffat, Fabien Vincent, Jonathan A Lee, Jörg Eder, and Marco Prunotto. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature reviews Drug discovery*, 16(8):531–543, 2017.

[Munson *et al.*, 2024] Brenton P Munson, Michael Chen, Audrey Bogosian, Jason F Kreisberg, Katherine Licon, Ruben Abagyan, Brent M Kuenzi, and Trey Ideker. De novo generation of multi-target compounds using

deep generative chemistry. *Nature Communications*, 15(1):3636, 2024.

[Pham *et al.*, 2022] Thai-Hoang Pham, Lei Xie, and Ping Zhang. Fame: fragment-based conditional molecular generation for phenotypic drug discovery. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 720–728. SIAM, 2022.

[Popova *et al.*, 2018] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7):eaap7885, 2018.

[Sadri, 2023] Arash Sadri. Is target-based drug discovery efficient? discovery and "off-target" mechanisms of all drugs. *Journal of medicinal chemistry*, 66(18):12651–12677, 2023.

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

[Shayakhmetov *et al.*, 2020] Rim Shayakhmetov, Maksim Kuznetsov, Alexander Zhebrak, Artur Kadurin, Sergey Nikolenko, Alexander Aliper, and Daniil Polykovskiy. Molecular generation for desired transcriptome changes with adversarial autoencoders. *Frontiers in Pharmacology*, 11:269, 2020.

[Subramanian *et al.*, 2017] Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, and et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, 2017.

[Swinney and Anthony, 2011] David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507–519, 2011.

[Tang *et al.*, 2018] Jing Tang, Balaguru Ravikumar, Zaid Alam, Anni Rebane, Markus Vähä-Koskela, Gopal Peddinti, Arjan J van Adrichem, Janica Wakkinen, Alok Jaiswal, Ella Karjalainen, et al. Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. *Cell chemical biology*, 25(2):224–229, 2018.

[Tong *et al.*, 2023] Xiaochu Tong, Ning Qu, Xiangtai Kong, Shengkun Ni, Kun Wang, Lehan Zhang, Yiming Wen, Sulin Zhang, Xutong Li, and Mingyue Zheng. Transigen: Deep representation learning of chemical-induced transcriptional profile. *bioRxiv*, pages 2023–11, 2023.

[Vincent *et al.*, 2022] Fabien Vincent, Arsenio Nueda, Jonathan Lee, Monica Schenone, Marco Prunotto, and Mark Mercola. Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nature Reviews Drug Discovery*, 21(12):899–914, 2022.

[Wang *et al.*, 2022] Wenlu Wang, Ye Wang, Honggang Zhao, and Simone Sciabola. A transformer-based generative model for de novo molecular design. *arXiv preprint arXiv:2210.08749*, 2022.

[Weinstein *et al.*, 2013] John N Weinstein, Eric A Collison, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, and Houtan Noushmehr. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[Wishart *et al.*, 2006] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl_1):D668–D672, 2006.

[Yamanaka *et al.*, 2023] Chikashige Yamanaka, Shunya Uki, Kazuma Kaitoh, Michio Iwata, and Yoshihiro Yamanishi. De novo drug design based on patient gene expression profiles via deep learning. *Molecular Informatics*, 42(8-9):2300064, 2023.

[Zhang and Chen, 2022] Jie Zhang and Hongming Chen. De novo molecule design using molecular generative models constrained by ligand–protein interactions. *Journal of chemical information and modeling*, 62(14):3291–3306, 2022.

[Zhavoronkov *et al.*, 2019] Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.