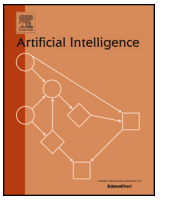
Artificial Intelligence 338 (2025) 104239

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



AI-driven transcriptome profile-guided hit molecule generation ☆

Chen Li *, Yoshihiro Yamanishi

Graduate School of Informatics, Nagoya University, Chikusa, Nagoya, 464-8601, Japan

ARTICLE INFO

Keywords:
Transcriptome profiles
Hit molecule generation
Dual VAEs

ABSTRACT

De novo generation of bioactive and drug-like hit molecules is a pivotal goal in computer-aided drug discovery. While artificial intelligence (AI) has proven adept at generating molecules with desired chemical properties, previous studies often overlook the influence of disease-specific cellular environments. This study introduces GxVAEs, a novel AI-driven deep generative model designed to produce hit molecules from transcriptome profiles using dual variational autoencoders (VAEs). The first VAE, ProfileVAE, extracts latent features from transcriptome profiles to guide the second VAE, MolVAE, in generating hit molecules. GxVAEs aim to bridge the gap between molecule generation and the biological context of disease, producing molecules that are biologically relevant within specific cellular environments or pathological conditions. Experimental results and case studies focused on hit molecule generation demonstrate that GxVAEs surpass current state-of-the-art methods, in terms of reproducibility of known ligands. This approach is expected to effectively find potential molecular structures with bioactivities across diverse disease contexts.

1. Introduction

Identifying hit molecules with the desired bioactivity and therapeutic effects within the vast expanse of chemical space is a major challenge in drug discovery [1]. Despite the theoretical existence of over 10^{60} organic molecules, only a tiny fraction exhibit drug-like properties. Traditionally, hit identification has depended on experimental techniques such as high-throughput screening (HTS) [2]. HTS is a valuable tool for finding bioactive hits [3], yet its success rate tends to be low, with many molecules proving inactive or less promising [4]. Additionally, these experimental methods are usually labor-intensive and time-consuming, often requiring significant resources and expertise to execute effectively. The entire drug development process can extend up to 12 years and require an investment exceeding 1.8 billion US dollars, encompassing various stages from initial screening and preclinical testing to multiple phases of clinical trials and regulatory approval [5]. This lengthy and costly process underscores the need for more efficient and cost-effective methods of hit identification and drug development. Despite extensive pre-market testing designed to ensure the efficacy and safety of drug candidates, more than 90% of these candidate molecules fail to reach the market [6]. This high failure rate is a testament to the inherent difficulties and high-risk nature of drug discovery and development. It highlights the critical need for innovation in methodologies and technologies for identifying and developing new therapeutic agents. Advances in computational drug discovery [7] and artificial intelligence (AI) [8] are poised to complement traditional experimental approaches, potentially enhancing the efficiency and success rates of drug discovery pipelines.

☆ This paper is an invited revision of a paper which first appeared at the 38th AAAI Conference on Artificial Intelligence (AAAI-24).

* Corresponding author.
E-mail addresses: li.chen.z2@a.mail.nagoya-u.ac.jp (C. Li), yamanishi@i.nagoya-u.ac.jp (Y. Yamanishi).

https://doi.org/10.1016/j.artint.2024.104239
Received 7 July 2024; Received in revised form 29 September 2024; Accepted 18 October 2024

Available online 22 October 2024

AI-driven *de novo* molecule generation represents a paradigm shift in computer-aided drug discovery. By integrating deep learning and computational chemistry, this approach generates entirely new molecules from scratch, without relying on existing compounds. It leverages a range of innovative strategies to design novel molecules with desirable chemical properties, offering a powerful tool for advancing drug discovery. Sophisticated deep generative models, such as generative adversarial networks (GANs) [9–11] and variational autoencoders (VAEs) [12–15], are central to this advancement. These models revolutionize the drug discovery pipeline by efficiently generating molecules with predefined bioactivity profiles. Trained on vast datasets comprising compounds with known bioactivities, these models excel at discerning intricate molecular patterns and relationships, allowing them to produce new molecules that closely mimic desired structural and functional characteristics. By generating molecules with desired optimized properties, AI-driven *de novo* approaches streamline the iterative process of drug design and discovery, significantly accelerating the exploration and identification of promising candidates for further development. This integration of AI into molecule generation not only enhances the efficiency of drug discovery efforts but also expands the scope of chemical diversity accessible for therapeutic innovation. As these technologies evolve, they hold promise for unlocking novel treatments that address unmet medical needs with unprecedented precision and efficacy.

However, current AI-driven approaches frequently neglect crucial biological aspects, especially the complex interactions among molecules within biological systems and their specific cellular contexts in disease states. This oversight undermines the capacity of AI-generated molecules to interact effectively with biological targets and achieve desired therapeutic effects. A primary goal of AI-driven drug discovery is to reconcile chemical feasibility with biological efficacy. This requires deeper insights into molecular interactions within biological environments, crucial for enhancing the utility and success rate of AI-generated molecules in clinical applications. Adopting a holistic approach presents new opportunities for discovering breakthrough therapies that not only possess optimal chemical properties but also demonstrate robust efficacy and safety profiles in complex biological settings. An emerging strategy to address these challenges involves generating hit molecules using omics data, encompassing genomics, epigenomics, and transcriptomes [16]. Transcriptome profiles offer a comprehensive molecular landscape describing how human cells respond to drug therapy and the underlying pathological processes of diseases [17]. For example, transcriptome profiles provide valuable insights into gene activities across various cellular contexts, including specific disease states [18]. Analyzing transcriptome profiles reveals dynamic changes such as gene expression and transcriptional variations driven by physiological conditions or disease progression, offering insights into the cellular context affecting therapeutic response and disease evolution.

Transcriptome profile-based approaches offer significant promise in drug discovery, yet they come with inherent challenges [19]. Firstly, the availability of comprehensive and high-quality transcriptome profiles for molecules remains limited. Additionally, interpreting and integrating transcriptome profiles pose substantial complexity and require advanced bioinformatics and statistical techniques to extract meaningful insights and identify relevant molecular signatures. Moreover, applying transcriptome profiles in the context of drug discovery demands a deep understanding of underlying biological and disease mechanisms. While some studies have leveraged transcriptome profiles to generate hit molecules [20,21], these efforts often lack direct associations between the source molecules and their corresponding transcriptome profiles. Consequently, the performance of generated hit molecules, such as their ability to interact effectively with target proteins, remains relatively modest and indicates room for improvement.

In this study, we introduce *GxVAEs*, an innovative approach for generating hit molecules from transcriptome profiles using dual VAEs: ProfileVAE and MolVAE. ProfileVAE acts as a feature extractor, utilizing transcriptome profile data to derive latent features that guide MolVAE in producing molecules. By integrating the dual VAEs, GxVAEs effectively bridge the gap between molecule generation and the intricate cellular environments of biological systems, thereby enhancing the biological relevance of the generated molecules tailored for specific diseases. This study builds upon our previous work [22], expanding on the methodology and highlighting key contributions:

- **Integration of cellular environments in molecule generation**: This study incorporates the influence of cellular environments on disease contexts during the molecule generation process.
- **Simple yet robust model architecture**: The streamlined dual VAE design produces hit molecules from transcriptome profiles.
- **Demonstrated superiority over state-of-the-art (SOTA) models**: GxVAEs exhibit superior performance compared to existing models designed for similar objectives, validating their effectiveness in advancing drug discovery in biomedicine.

The remainder of the study is organized as follows. Section 2 surveys existing approaches in molecule generation. Section 3 introduces GxVAEs, our proposed method for generating hit molecules from transcriptome profiles. Section 4 details comprehensive experiments and case studies aimed at validating the effectiveness of GxVAEs compared to SOTA models. Finally, Section 5 summarizes this study and outlines future research directions.

2. Related work

De novo molecule generation is instrumental in designing new molecules tailored to specific chemical properties, utilizing three primary data representations: 1-dimensional sequence-based representations, such as simplified molecular input line entry system (SMILES) strings [23], 2-dimensional (2D) molecular graphs [24], and 3-dimensional (3D) geometric structures [25]. These representations offer distinct advantages in capturing different aspects of molecular structures, from linear sequences suitable for database storage and modeling to detailed graphical and spatial arrangements essential for understanding molecular interactions and properties.

2.1. SMILES-based molecule generation

SMILES strings provide a straightforward linear representation of molecular structures, making them well-suited for efficient database storage and computational modeling. They are easy to manage in chemical databases and straightforward to process in molecular modeling applications. Deep generative models such as TransVAE [13] and GrammarVAE [14] utilize VAEs, whereas ScaffoldRNN [26], BIMODAL [27], and EarlGAN [28] employ recurrent neural networks (RNNs) for molecule generation from SMILES strings. To bolster robustness in molecule generation, TransORGAN [11] adopts a Transformer architecture within a GAN framework to produce valid SMILES strings. Similarly, TenGAN [29] and SpotGAN [30] apply a pure Transformer encoder and a reversal Transformer (first-decoder-then-encoder) GAN respectively, to train models capable of generating diverse SMILES representations. The increasing popularity of diffusion models has led to the emergence of SMILES-based approaches such as MDM [31] and TGM-DLM [32], which leverage diffusion processes for molecule generation and have gained significant attention in the research community. Moreover, alternative representations of SMILES, such as DeepSMILES [33] and GenSMILES [34], have been proposed to enhance molecular validity. These variants aim to improve upon the traditional SMILES format for better molecule generation capabilities. Additionally, SELFIES [35] presents another sequential representation that tends to produce overly long chains and large rings in molecules [36].

2.2. Graph-based molecule generation

Molecular graphs serve as detailed representations where atoms are nodes and chemical bonds are edges, providing specifics on atom and bond types (single, double, triple). JTVAE [15] employs a node tree methodology for constructing these graphs, integrating tree-structured scaffolds through graphical message-passing networks [37]. MolGAN [10] uses a reinforcement learning-guided GAN framework to generate molecular graphs with defined chemical properties, while ALMGIG [38], an extension of the bidirectional GAN model, employs adversarial cyclic consistency loss to learn molecular space distributions and generate new molecules. Similarly, geometric structure approaches often employ equivalent neural networks [39] to capture precise spatial arrangements and atom interactions, offering insights into molecular geometry and potential interactions with biological targets, thereby advancing applications in drug discovery and materials science.

2.3. Hit molecule generation

Previous studies in *de novo* molecule generation have primarily focused on optimizing the chemical properties of generated molecules based on their structural information. However, these approaches often overlooked the biological environment and cellular context of diseases [40], as well as the cell-specific activities of the molecules produced [41,42]. The evolution of omics data analysis, particularly transcriptome profiles in drug discovery [43,44], has spurred efforts to incorporate comprehensive biological response data into the development of therapeutic molecules tailored for specific diseases.

ExpressionGAN [20] and TRIOMPHE [21] stand as leading deep generative models for generating hit molecules interacting therapeutic target proteins without prior annotations of training molecules. ExpressionGAN integrates systems biology and molecular design to generate molecules capable of inducing desired transcriptome features directly from gene expression profiles. TRIOMPHE first calculates correlation coefficients to evaluate the strength of the relationship between target perturbation profiles (which reflect cellular responses to target proteins) and chemically induced profiles (which represent cellular responses to compound treatment). It then selects the molecules most closely aligned with the chemically induced profiles as source molecules for a VAE model. This model subsequently generates molecules designed to correspond with the target perturbation profiles.

Despite the potential of transcriptome profiles to tailor bioactive molecules for arbitrary target proteins within cellular contexts, ExpressionGAN encounters challenges such as producing hit molecules with low validity ($> 8.5\%$) and limited ability to replicate known ligands. Furthermore, understanding the transcriptional correlation between ligands and targets remains elusive. In contrast, TRIOMPHE relies exclusively on transcriptome profiles for correlation calculations, omitting a VAE in the molecule generation process. To address these limitations, this study introduces GxVAEs, an innovative approach designed to bridge the gap between the cellular environment and therapeutic molecule generation. By leveraging transcriptome profiles, GxVAEs facilitate the creation of biologically meaningful molecules tailored to specific disease contexts, thereby advancing precision in drug discovery. Unlike traditional *de novo* molecule generation, which often focus solely on optimizing chemical properties, GxVAEs integrate biological relevance into the molecule design process, aiming to produce therapeutically relevant compounds with a higher degree of accuracy.

3. GxVAEs

Fig. 1 presents an overview of GxVAEs, a dual VAE framework consisting of ProfileVAE and MolVAE, designed for generating hit molecules from transcriptome profiles. ProfileVAE serves as a crucial component by extracting essential features from high-dimensional transcriptome profiles, condensing them into low-dimensional feature vectors. These extracted features capture nuanced information related to biological responses, thereby guiding MolVAE in the generation of molecules. MolVAE, on the other hand, utilizes bidirectional gated recurrent units (GRUs) [45] conditioned on the feature vectors extracted by ProfileVAE. This conditioning ensures that the molecule generation process is influenced by the underlying biological context inferred from the transcriptome

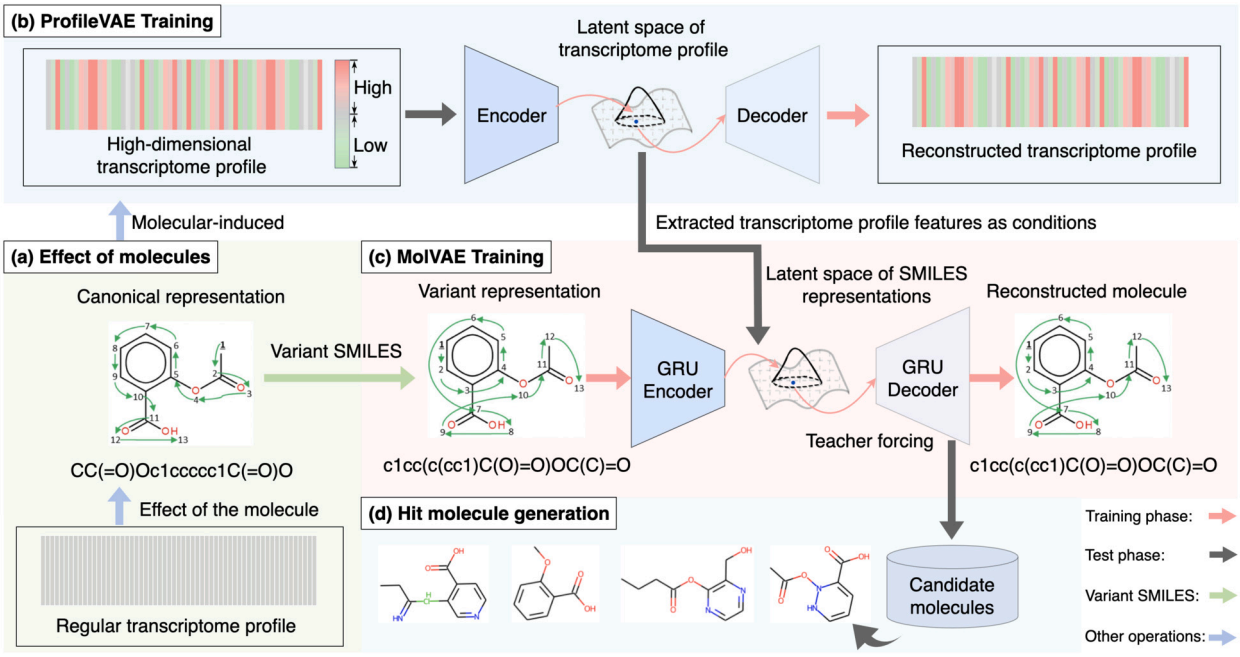


Fig. 1. Overview of GxVAEs. (a) Exposure a molecule to cells induces a specific transcriptome profile. This transcriptome profile, altered by the molecule "CC(=O)oc1ccccc1C(=O)O", manifests as fold changes in gene expression values. (b) The affected transcriptome profile is fed into the ProfileVAE encoder, where it is transformed into a latent space representation of transcriptome profiles. The resulting feature vector is then used by the ProfileVAE decoder to reconstruct the original transcriptome profile. (c) The extracted transcriptome profile features are combined with the variant SMILES representation of the molecule (e.g., "c1cc(c(cc1)C(O)=O)OC(C)=O"), serving as input to the bidirectional GRU encoder of MolVAE. This encoder extracts condition-constrained molecular features and maps them to the latent space representation of SMILES. Subsequently, the bidirectional GRU decoder reconstructs these features back into the original molecular structure. During training process, teacher forcing is employed to stabilize learning and expedite the convergence of MolVAE. (d) During inference (illustrated by black arrows), only ProfileVAE's encoder and MolVAE's decoder are utilized. An unknown transcriptome profile is input to ProfileVAE's encoder, with the extracted features then used by MolVAE's decoder to generate hit molecules exhibiting transcriptome profiles.

profiles. MolVAE generates molecules in the form of variant SMILES strings [29,30], which allow for flexibility in molecular representation. To optimize the training process, teacher forcing [46] is employed during the MolVAE training. This technique stabilizes training and accelerates convergence by guiding the model with correct sequences during initial phases, facilitating more efficient learning and better performance in generating molecules that align with desired biological and chemical criteria.

ProfileVAE. Let $\boldsymbol{X} = [x_1, x_2, \cdots, x_K]$ denote a transcriptome profile (e.g., gene expression profile) containing K genes, where x_k represents the k-th gene value. The objective of ProfileVAE is to model the marginal likelihood of the transcriptome profile through the following generative procedure:

$$\max_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{q_\theta(\boldsymbol{C}|\boldsymbol{X})}[\log p_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{C})], \tag{1}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operation, $\boldsymbol{C}$ represents the latent variable, and $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are the parameters of the ProfileVAE encoder and decoder, respectively. Additionally, $p_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{C})$ refers to the likelihood function, while $q_{\boldsymbol{\theta}}(\boldsymbol{C}|\boldsymbol{X})$ represents the posterior distribution. The loss function for the ProfileVAE is formulated as follows:

$$\begin{aligned}\mathcal{L}_G(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{X},\boldsymbol{C},\beta) &= \mathbb{E}_{q_\theta(\boldsymbol{C}|\boldsymbol{X})}[\log p_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{C})] \\ &\quad - \beta \cdot D_{KL}(q_{\boldsymbol{\theta}}(\boldsymbol{C}|\boldsymbol{X})||p_{\boldsymbol{\phi}}(\boldsymbol{C})),\end{aligned} \tag{2}$$

where β represents the weight of the Kullback–Leibler (KL) divergence D_{KL} [47]. During inference, the latent vector $\boldsymbol{C}$ for a given transcriptome profile is derived using the reparameterization trick with a unit normal distribution:

$$\boldsymbol{C} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{3}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the mean and standard deviation of the Gaussian distribution, respectively. This reparameterization technique ensures that the sampling process is differentiable, enabling efficient backpropagation and optimization of ProfileVAE's parameters during the training process.

MolVAE. Let $\boldsymbol{S} = [s_1, s_2, \cdots, s_T]$ represent a SMILES string, where s_t denotes the t-th atom. Assume $\boldsymbol{S}$ is reconstructed from the latent variable $\boldsymbol{Z}$ and conditioned on the extracted features through $\boldsymbol{S} \sim p_{\boldsymbol{\varphi}}(\boldsymbol{S}|\boldsymbol{Z})$, parameterized by $\boldsymbol{\varphi}$. The probability distribution can be expressed as:

$$p_{\boldsymbol{\varphi}}(\boldsymbol{S}|\boldsymbol{Z}) = \prod_{t=1}^{T} p_{\boldsymbol{\varphi}}(s_t|s_1, \boldsymbol{Z}, \boldsymbol{C}). \tag{4}$$

Here, the feature vectors $\boldsymbol{C}$ derived from the transcriptome profiles are utilized as conditioning inputs for MolVAE. The objective of MolVAE is to optimize the lower bound of the true log-marginal likelihood, formulated as:

$$\begin{aligned} \mathcal{L}_S(\boldsymbol{\varphi}, \boldsymbol{\psi}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}) = \mathbb{E}_{q_{\boldsymbol{\psi}}(\boldsymbol{Z}|\boldsymbol{S})}[\log p_{\boldsymbol{\varphi}}(\boldsymbol{S}|\boldsymbol{Z})] \\ - D_{KL}(q_{\boldsymbol{\psi}}(\boldsymbol{Z}|\boldsymbol{S})||p_{\boldsymbol{\varphi}}(\boldsymbol{Z})). \end{aligned} \tag{5}$$

Variant SMILES enables traversal of a molecular graph starting from different atoms, yielding diverse non-canonical SMILES representations of the same molecular structure. These variants are crucial for training models, ensuring the generation of diverse molecules during the training phase. An example illustrating variant SMILES is included in Appendix A. Furthermore, teacher forcing is employed during the training of MolVAE to enhance stability and accelerate convergence (refer to Appendix B for details).

Algorithm 1 Procedures for GxVAEs.

Data: Transcriptome profiles $\boldsymbol{X}$ and the intercorrelated SMILES strings $\boldsymbol{S}$

```
/* Training phase of the ProfileVAE */
for i ← 1 to g-epochs do
    Train ProfileVAE with Eq. (2).
    Update the parameters of θ and φ.
end for
Retain only the encoder of ProfileVAE and extract the features of transcriptome profiles with Eq. (3).
/* Training phase of the MolVAE */
for j ← 1 to s-epochs do
    Train MolVAE with Eq. (5).
    Update the parameters of φ.
end for
Retain only the decoder of MolVAE and generate SMILES strings with Eq. (4).
/* Inference phase of the GxVAEs */
Combine ProfileVAE's encoder with MolVAE's decoder for GxVAEs.
for n ← 1 to N do
    Extract features C from transcriptome profiles.
    Sample Gaussian noise Z.
    Generate hit molecules from transcriptome profiles.
end for
```

Algorithm 1 delineates the training strategies employed for GxVAEs. The process begins with the training of ProfileVAE, where the model is initially trained, and only the encoder component is retained after completion. This encoder is crucial for extracting features from transcriptome profiles. Subsequently, these features, accompanied by the corresponding SMILES strings, serve as inputs for training MolVAE. Following the completion of MolVAE training, the decoder component is exclusively retained for molecule generation purposes. In the final step, we integrate ProfileVAE's encoder with the MolVAE's decoder, allowing for the joint generation of hit molecules from transcriptome profiles.

4. Experiments

We conducted comprehensive experiments to investigate the following research questions (RQs):

- **RQ1**: To what extent does ProfileVAE capture features from transcriptome profiles?
- **RQ2**: Does MolVAE effectively capture internal molecular features, and how proficiently does it generate molecules?
- **RQ3**: Can GxVAEs generate potential candidate hit molecules?

4.1. Experimental setup

Datasets. Three distinct types of transcriptome profiles were employed to evaluate the effectiveness of GxVAEs:

- **Chemically induced profiles** were obtained from the LINCS database [48], which catalogs transcriptome value changes in response to diverse chemical compounds across 77 human cell lines. For instance, our analysis included data from the MCF7 cell line treated with a comprehensive set of 13,755 distinct molecules. This dataset provides insights into how different chemicals induce varied molecular responses within cellular contexts.

- **Target perturbation profiles** were obtained from the LINCS database, focusing on transcriptome alterations caused by targeted genetic interventions across the same 77 human cell lines. These interventions typically involve manipulating the expression levels of key proteins like RAC-α serine / threonine-protein kinase (AKT1), RAC-β serine / threonine-protein kinase (AKT2), Aurora B kinase (AURKB), cysteine synthase A (CTSK), epidermal growth factor receptor (EGFR), histone deacetylase 1 (HDAC1), mammalian target of rapamycin (MTOR), phosphatidylinositol 3-kinase catalytic subunit (PIK3CA), decapentaplegic homolog 3 (SMAD3), and tumor protein p53 (TP53), which play crucial roles in numerous cancers.
- **Disease-specific profiles** were retrieved from the crowd extracted expression of differential signatures (CREEDS) database [49]. These profiles aggregate transcriptome patterns from 14,804 genes associated with various diseases such as gastric cancer, atopic dermatitis, and Alzheimer's disease. By consolidating transcriptome profiles from multiple patients, we derived representative profiles that reflect the molecular signatures unique to each disease condition.

In this study, the chemically induced profiles dataset served as the training set for our GxVAEs model, providing essential data for the model to learn how to generate molecules based on cellular responses to chemical treatments. The target perturbation profile dataset was utilized to evaluate the GxVAEs model's ability to generate molecules tailored to specific protein targets involved in disease mechanisms, allowing us to assess how well the generated molecules align with the intended target perturbations. For the case study on therapeutic molecule generation, the disease-specific profile dataset was employed to determine how effectively GxVAEs could produce molecules suited to specific disease contexts, thereby enabling us to evaluate the model's performance in generating therapeutically relevant candidates.

Implementation details. ProfileVAE employed an encoder and decoder architecture comprising three feedforward layers with dimensions of 512, 256, and 128. The learning rate was set to 1e−4, and dropout with a probability of 0.2 was applied during training. MolVAE utilized an embedding size of 128 and featured three hidden layers sized at 256. The learning rate for MolVAE was 5e−4, with a dropout probability of 0.1. Additionally, a temperature parameter β of 1.0 was incorporated. The maximum length of generated SMILES strings was limited to 100 characters. Both ProfileVAE and MolVAE used latent vectors of dimensionality 64, with a consistent batch size across all experiments. ProfileVAE was trained over 2000 epochs, while MolVAE underwent 200 epochs of training. Computational experiments were performed using an NVIDIA Tesla T4 GPU with CUDA version of 11.4. A more detailed description of the experimental setup can be found in Appendix C. The source code used for the experiments is available on GitHub.[1]

4.2. Evaluation measures

We rigorously evaluated the quality and effectiveness of newly generated hit molecules using the following metrics:

- **Validity** determines the percentage of chemically valid molecules, verified using the RDKit tool [50]. Low validity suggests potential issues during the model training. The calculation is expressed as:

$$\text{Validity} = \frac{\text{Number of chemically valid molecules}}{\text{The total number of molecules generated}}.$$

- **Uniqueness** quantifies the proportion of unique molecules among those identified as valid. Low uniqueness indicates issues such as mode collapse. The calculation is performed as follows:

$$\text{Uniqueness} = \frac{\text{Number of non-repeated molecules}}{\text{Number of valid molecules}}.$$

- **Novelty** measures the proportion of unique molecules that are absent from the training set. Low novelty indicates an over-fitting problem. The calculation is shown below:

$$\text{Novelty} = \frac{\text{Number of novel molecules}}{\text{Number of unique molecules}}.$$

- **Quantitative estimate of drug-likeness (QED)** assesses whether a particular molecule is similar to known drugs. Usually, QED score is determined by assigning weights to eight molecular descriptors: molecular weight, octanol-water partition coefficient, number of hydrogen bond donors, number of hydrogen bond acceptors, molecular polar surface area, number of rotatable bonds, number of aromatic rings, and number of structural alarms [51]. The calculation is as follows:

$$\text{QED} = \exp\left(\frac{\sum_{i=1}^{8} W_i \ln d_i}{\sum_{i=1}^{8} W_i}\right),$$

where d_i represents the desirability function and W_i denotes the weight assigned to the i-th molecular descriptor. These weights are typically determined through chemical experiments. In this study, we calculated the QED score using the RDKit tool. A higher QED score indicates a higher likelihood that the molecule exhibits drug-like properties.

[1] Source code available at: https://github.com/naruto7283/GxVAEs.

- **Synthetic accessibility (SA)** score evaluates the ease of molecule synthesis. A higher SA score indicates less complex and more easily synthesizable compounds, calculated by

$$\mathrm{SA} = r_s - \sum_{i=1}^{5} p_i.$$

Here, r_s represents the "synthetic knowledge" derived from analyzing synthetic molecule features. It is the ratio of the summed contributions of all fragments to the number of fragments in the molecule [52]. The terms p_i (where i ranges from 1 to 5) denote ring complexity, stereo complexity, macrocycle penalty, size penalty, and bridge penalty, all calculated using the RDKit tool.

- **Lipo-hydro partition coefficient (logP)** score is defined as the ratio of a molecule's concentration in a mixture of two immiscible solvents at equilibrium [29], which can be expressed as

$$\mathrm{logP} = \log \frac{c_o}{c_w},$$

where c_o and c_w denote the substance's activity in the organic and water phases, respectively. A higher logP score represents increased lipophilicity of the molecule towards the organic phase.

- **Tanimoto coefficient** score was used to assess the structural similarity of generated molecules with known ligands of the corresponding target proteins, employing the ECFP4 fingerprint [53]. Molecules that are structurally similar to a ligand often exhibit a similar mode of action. The calculation is performed as follows:

$$\mathrm{Tanimoto} = \frac{N_{AB}}{N_A + N_B - N_{AB}},$$

where N_A and N_B represent the number of bits in the fingerprints of molecular structures A and B, respectively, and N_{AB} denotes the number of bits common to both fingerprints. In practice, we utilize "GetMorganfingerprintAsBitVect" and "BulkTanimotoSimilarity" from the RDKit tool to compute ECFP4 fingerprints and Tanimoto coefficients.

The statistics, including validity, uniqueness, and novelty, were utilized for evaluating the generated molecules. Chemical properties such as QED, SA, and logP scores were employed to assess quality of the generated molecules. Additionally, Tanimoto coefficients were used to measure the similarity of the generated molecules with known ligands of the corresponding target proteins. All measures were normalized to a 0-1 scale to ensure consistency and comparability across evaluations.

4.3. Baselines

The effectiveness of GxVAEs is evaluated by comparing it to two primary baseline models, which are described as follows:

- **ExpressionGAN [20]**: This previous SOTA model integrates systems biology with molecular design by employing a GAN)conditioned on transcriptomic data. This model autonomously designs molecules with a high probability of inducing a desired transcriptomic profile. By inputting a gene expression signature of the target state, the model is capable of generating hit molecules for specific targets without requiring prior annotations of the training compounds.
- **TRIOMPHE [21]**: This model uses transcriptome profiles to design new molecules targeting specific phenotypes and ligand-target interactions. It analyzes the correlation between chemically induced and genetically perturbed transcriptome profiles, and employs a VAE to generate molecules that match the desired profiles.

4.4. Evaluation of ProfileVAE (*RQ1*)

To demonstrate ProfileVAE's capability in extracting biological features from transcriptome profiles, we compared the distributions of input transcriptome profiles with their reconstructed counterparts. Fig. 2 depicts the transcriptome profile distributions induced by the molecules "$C_{40}H_{80}NO_8P$", "$C_{25}H_{39}N_3O_5$", "$C_{13}H_{16}N_2OS$", and "$C_6H_{15}N_3S$" (shown in yellow). These molecules have molecular weights of 734, 462, 248, and 161, with SMILES lengths of 63, 53, 36, and 15. The figures also display their corresponding reconstructed profiles (shown in green). The results illustrate that ProfileVAE successfully reconstructed the input transcriptome profiles across a variety of molecular weights and SMILES lengths. Fig. D.1 in the Appendix depicts the average distribution of all transcriptome profiles in the real set and their corresponding reconstructed profiles. The reconstructed profiles closely approximate those of the real set, showcasing ProfileVAE's capability to effectively extract features from transcriptome profiles and validate its effectiveness in reconstructing profiles based on these features.

Additionally, Fig. D.2 illustrates a positive linear relationship between SMILES lengths and molecular weights, indicating that longer SMILES strings generally correspond to heavier molecular weights. Figs. 3 and D.3 show the relationship between the mean square errors (MSE) of reconstructed transcriptome profiles and their respective SMILES lengths and molecular weights. The results reveal a significant negative correlation between MSE, which gauges the disparity between reconstructed and actual transcriptome profiles, and both SMILES lengths and molecular weights. This suggests that ProfileVAE effectively captures transcriptome profiles associated with molecules characterized by longer SMILES strings and molecular weights in the range of [400, 600]. Conversely, due to limited data availability for molecular weights below 200 and above 600 (about 300 and 100 samples, respectively), the MSE exhibits signs of instability or increases.

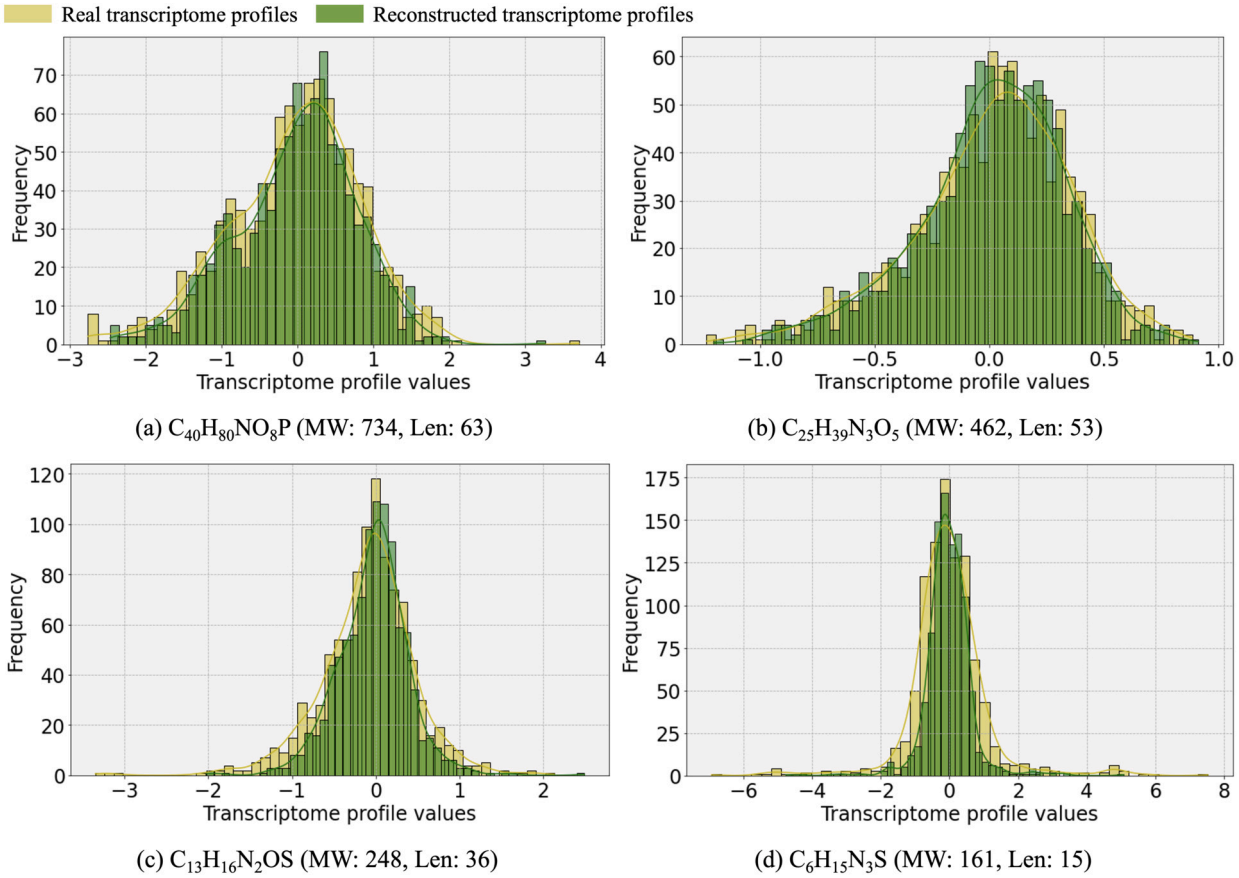


Fig. 2. Real and reconstructed distribution of transcriptome profiles generated by ProfileVAE. MW and Len represent molecular weight and SMILES length, respectively. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

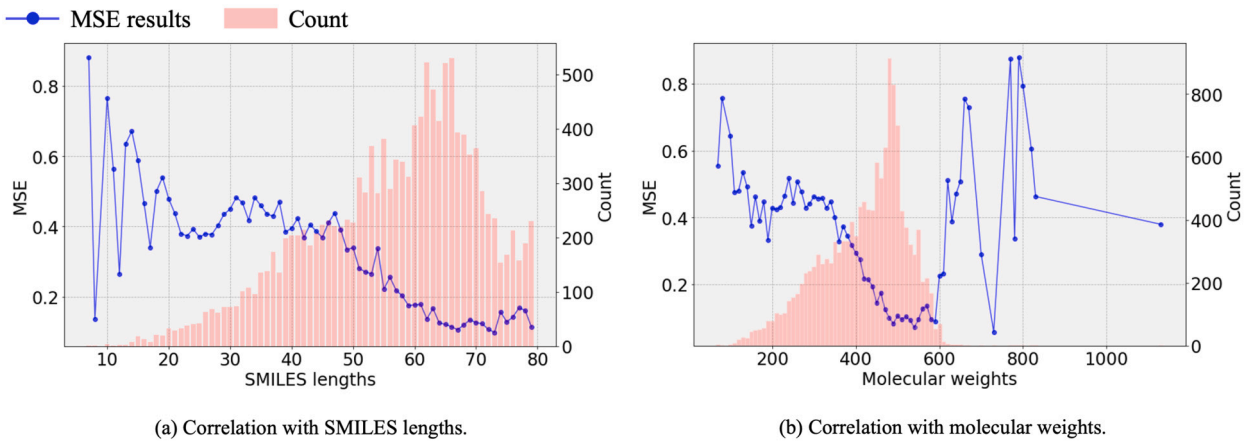


Fig. 3. Correlation between the reconstructed transcriptome profiles and SMILES lengths and molecular weights.

Furthermore, we performed singular value decomposition (SVD) and principal component analysis (PCA) on both the real transcriptome profiles and their corresponding reconstructed profiles. The 2D and 3D visualization plots are shown in Figs. 4, D.4, and D.5. The results illustrate that the reconstructed transcriptome profile data points exhibit significant overlap with the points from the real set, visually demonstrating the efficacy of ProfileVAE.

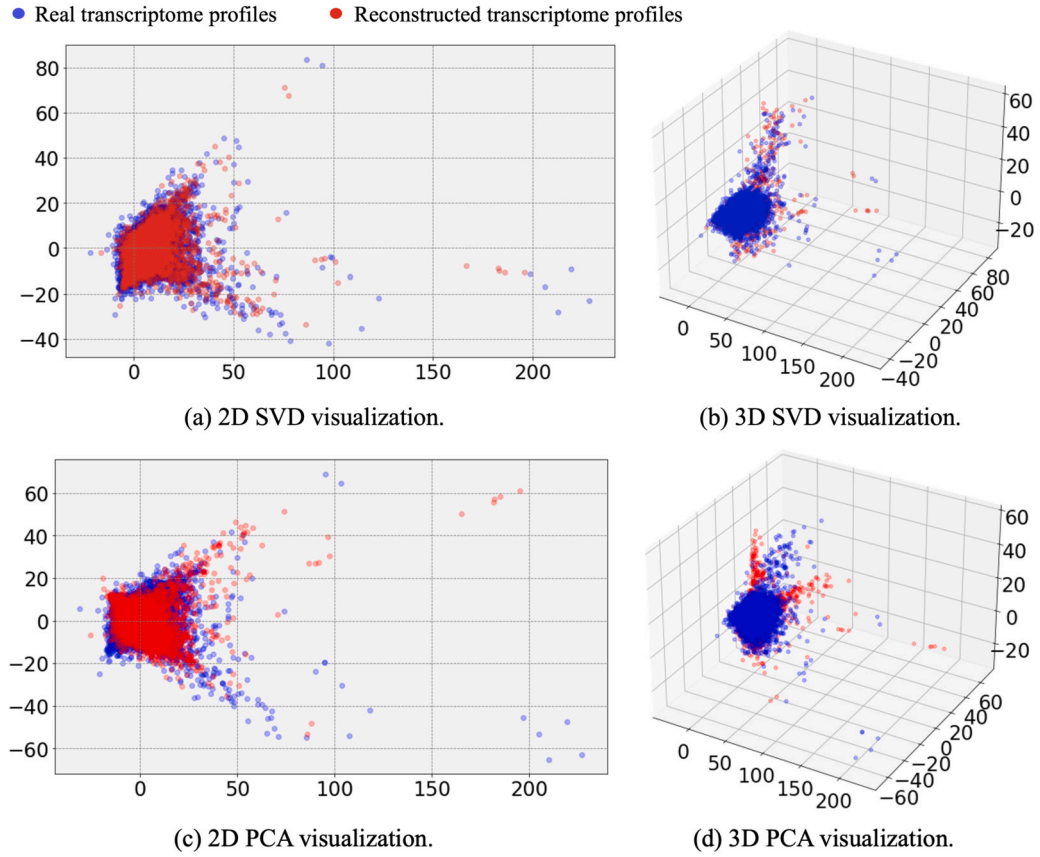


Fig. 4. Visualization plots of all real and reconstructed transcriptome profiles.

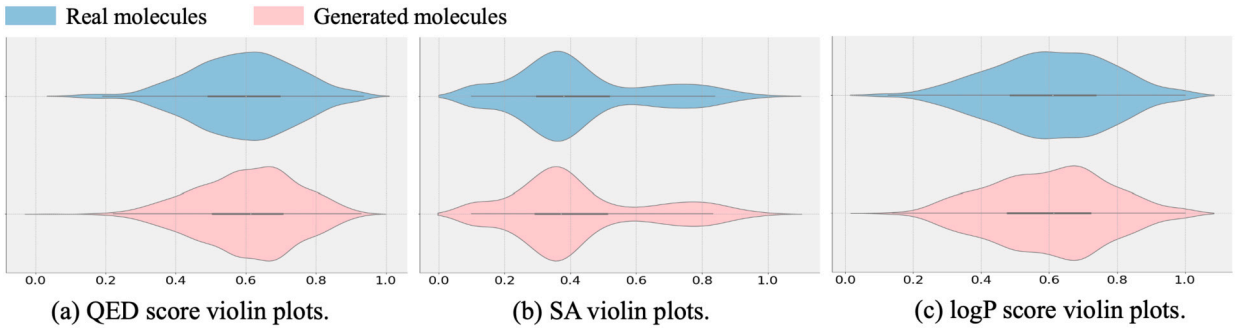


Fig. 5. Violin plots of property scores for molecules from the real set and MolVAE.

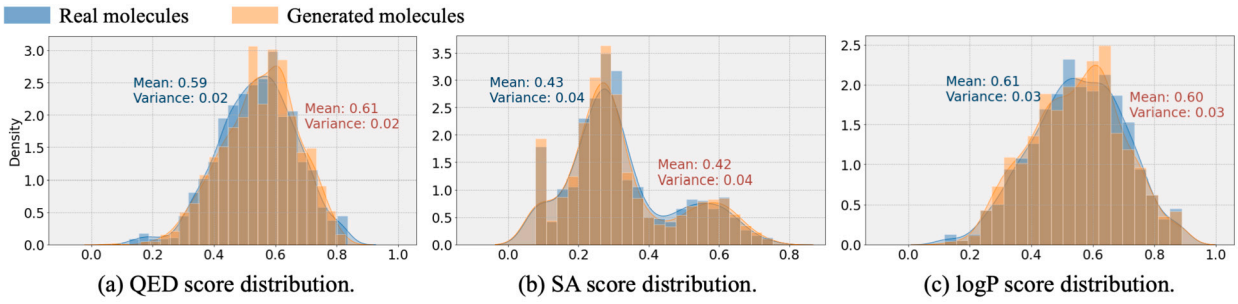


Fig. 6. Distributions of molecular property scores for molecules generated by MolVAE.

Table 1
Statistics for the dataset and molecules generated by MolVAE. AvgLen, MaxLen, and MinLen indicate the average, maximum, and minimum lengths of the SMILES strings.

Data	AvgLen	MaxLen	MinLen	MW
Real set	57	79	10	429
Generation	57	83	10	430

Table 2
Comparison of the top-k generated molecules.

Property	Data	Top-200	Top-400	Top-600	Top-800	Top-1000
QED	Real set	0.81	0.75	0.71	0.67	0.64
	Generation	0.81	0.75	0.72	0.68	0.65
SA	Real set	0.79	0.66	0.58	0.52	0.47
	Generation	0.80	0.67	0.58	0.52	0.47
logP	Real set	0.86	0.79	0.75	0.70	0.66
	Generation	0.86	0.79	0.74	0.70	0.65

Table 3
Comparative evaluation of TRIOMPHE baseline and GxVAEs. The values in gray cells indicate maximum values.

Target Protein	TRIOMPHE			GxVAEs		
	Validity	Uniqueness	Novelty	Validity	Uniqueness	Novelty
AKT1	20.9	49.3	100.0	88.0	88.6	100.0
AKT2	21.7	76.0	100.0	89.0	91.0	98.8
AURKB	20.3	74.9	100.0	89.0	93.3	98.8
CTSK	13.3	74.4	100.0	91.0	94.5	98.8
EGFR	13.2	75.8	99.0	89.0	93.3	100.0
HDAC1	25.5	52.2	99.2	78.0	96.2	97.3
MTOR	4.3	86.0	100.0	91.0	93.4	98.8
PIK3CA	10.9	97.2	100.0	92.0	93.5	97.7
SMAD3	26.6	67.7	100.0	86.0	91.9	98.7
TP53	12.2	77.9	100.0	85.0	96.5	98.8

4.5. Evaluation of MolVAE (**RQ2**)

Fig. E.1 in the Appendix displays the change curves for both the reconstruction loss and the ratio of valid molecules generated by MolVAE. It was observed that as the number of training epochs increased, the reconstruction loss decreased gradually, accompanied by a gradual increase in the validity of the generated molecules.

Fig. 5 presents violin plots comparing the QED, SA, and logP scores of molecules from the real set (in blue) and those generated by MolVAE (in red). The similarity in their distributions suggests that MolVAE generated valid molecules while maintaining consistent intrinsic chemical properties. To further analyze the structural characteristics of generated molecules, Fig. E.2 in the Appendix showcases the top 12 molecules based on the properties from both the real set and MolVAE. The molecules generated by MolVAE tend to exhibit structures and property scores similar to those in the real set.

Additionally, Table 1 provides statistics for both the real set and molecules generated by MolVAE. The average, maximum, and minimum lengths, as well as the average molecular weight of the generated SMILES strings, closely align with those of the real set. This consistency highlights MolVAE's effectiveness in learning data distributions of SMILES strings.

Fig. 6 and Table 2 depict the distributions of chemical property scores and the top-k generated molecules. The results demonstrate that MolVAE-generated molecules exhibit highly consistent distributions with the real set across QED, SA, and logP properties. This illustrates that MolVAE preserves intrinsic property features of SMILES strings during molecule generation.

4.6. Evaluation of GxVAEs (**RQ3**)

In biochemistry, ligand binding to a target protein alters its activity, triggering a cellular response. Here, we focused on generating ligand candidate molecules for ten target proteins by utilizing their transcriptome profiles. Knockdown transcriptome profiles were

Table 4
Comparison of the Tanimoto coefficients for the baselines and GxVAEs.

Target Protein	ExpressionGAN	TRIOMPHE	GxVAEs
AKT1	0.32	0.42	0.85
AKT2	0.29	0.35	0.43
AURKB	0.36	0.34	0.47
CTSK	0.31	0.29	0.38
EGFR	0.30	0.31	0.74
HDAC1	0.34	0.30	0.55
MTOR	0.39	0.69	0.52
PIK3CA	0.26	0.32	0.35
SMAD3	0.44	0.48	0.98
TP53	0.46	0.53	0.76

used to generate inhibitors for the proteins AKT1, AKT2, AURKB, CTSK, EGFR, HDAC1, MTOR, and PIK3CA, while overexpression transcriptome profiles were used to generate activators for SMAD3 and TP53.

Table 3 showcases the effectiveness of GxVAEs in generating hit molecules. Note that ExpressionGAN has a limited capacity to generate valid molecules, the comparison was made solely with the TRIOMPHE baseline. The results reveal that GxVAEs are three times more effective than TRIOMPHE in generating ligand-like molecules from the transcriptome profiles of the ten target proteins, achieving a validity rate of at least 78.0% for HDAC1. Moreover, the uniqueness of the molecules generated by GxVAEs surpasses that of TRIOMPHE, except for PIK3CA (93.5% versus 97.2%). The novelty of GxVAEs (97.7%) is also close to that of TRIOMPHE.

Table 4 compares the Tanimoto coefficients of GxVAEs with those of two SOTA models. For each target protein, we sampled 100 SMILES-like strings from its corresponding transcriptome profile. We then computed the Tanimoto coefficient between each generated sample and the known ligands of the target protein, and recorded the maximum Tanimoto coefficient in the table. A higher Tanimoto coefficient indicates a higher probability that a generated molecule effectively binds to a target protein. The findings indicate that GxVAEs produced candidate ligands with markedly higher Tanimoto coefficients across all ten target proteins compared to the two SOTA models. For example, the Tanimoto coefficient for AKT1 generated by GxVAEs was 2.7 and 2.0 times higher than those of the baselines.

To visually compare the known ligands with those generated by GxVAEs, we present their top 1 and top 5 molecular structures in Figs. 7 and F.1, respectively. The chemical structures of the hit molecules generated by GxVAEs seem to be similar to those of the known ligands. Overall, GxVAEs are able to produce ligand-like structures in generating hit molecules from transcriptome profiles, with the biological activity of the generated molecules significantly surpassing the SOTA baselines in terms of the reproducibility of known ligand structures.

4.7. Case study: therapeutic molecule generation

In disease states, complex combinations of gene abnormalities disrupt normal transcriptome patterns. As a case study, our aim was to generate therapeutic molecules from transcriptome profiles of real patients. Fig. 8 outlines the process facilitated by GxVAEs for therapeutic molecule generation. By aggregating transcriptome profiles from multiple patients with the same disease, we derived a disease-specific transcriptome profile that captures the transcriptome landscape of the disease. To identify potential therapeutic candidates, we inverted the disease-specific transcriptome profile (reversal profile). Molecules capable of inducing a transcriptome pattern similar to this reversal profile are considered potential therapeutics. Subsequently, these disease reversal profiles were input into GxVAEs to generate candidate therapeutic molecules. The generated molecules, which exhibit structural similarity to approved drugs for treating the disease, are anticipated to possess therapeutic efficacy. Tanimoto coefficients, reflecting structural similarity to approved drugs, were utilized to assess the drug-likeness.

We attempted to generate new therapeutic drug candidate molecules for gastric cancer, atopic dermatitis, and Alzheimer's disease. We made a comparison between GxVAEs and DRAGONET [54] using identical datasets of patient transcriptome profiles and molecule chemical structures to ensure a fair evaluation. Fig. 9 illustrates the therapeutic molecules generated by both methods, alongside their Tanimoto coefficients relative to approved drugs. Note that each approved drug was associated with 100 SMILES-like strings sampled from its corresponding disease reversal profile. For example, hydrocortisone (DB00741), a glucocorticoid used for treating atopic dermatitis, immune disorders, and allergies, achieved a Tanimoto coefficient of 1.0 with molecules generated by GxVAEs using patient-specific transcriptome profiles for atopic dermatitis. This result highlights GxVAEs' capability to accurately capture structural features akin to approved drugs for treating atopic dermatitis. Additionally, GxVAEs produced molecules designed for gastric cancer and Alzheimer's disease, and the generated molecules exhibited structural similarities to known approved drugs. This suggests that GxVAEs hold promise in generating therapeutic molecules with enhanced properties compared to DRAGONET.

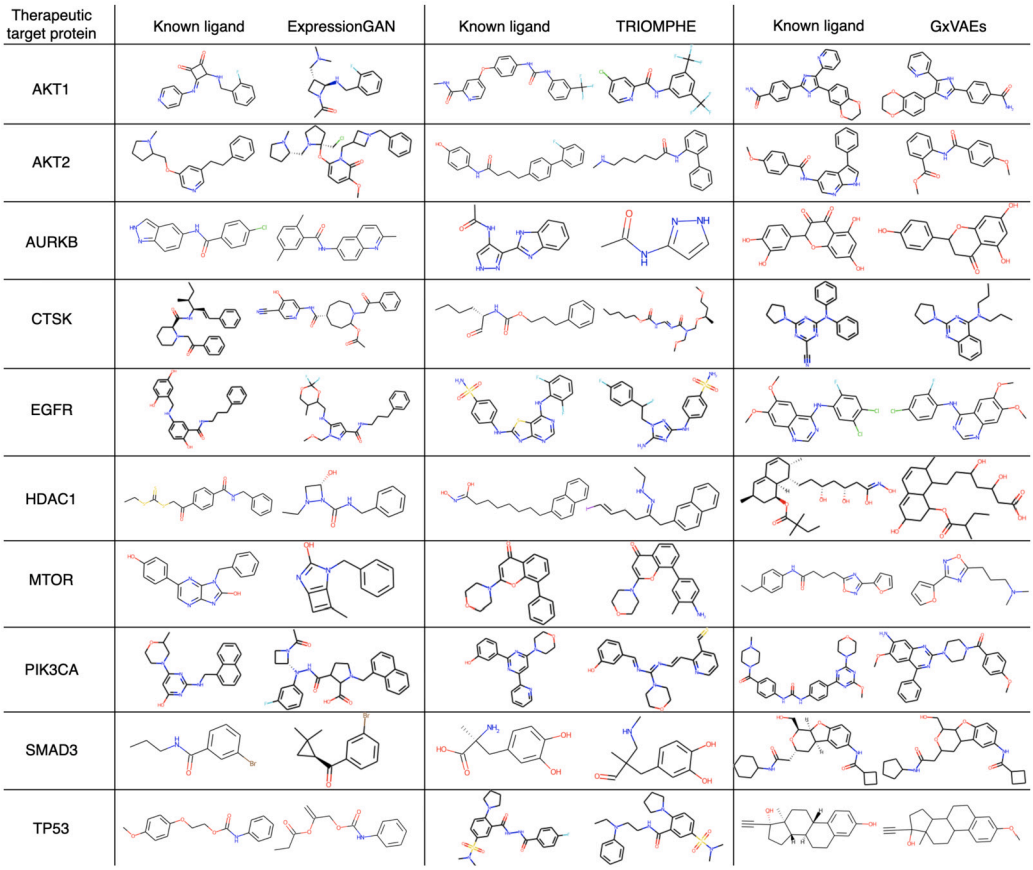


Fig. 7. Comparison of hit molecules generated by ExpressionGAN and TRIOMPHE baselines with those by GxVAEs, focusing specifically on molecules with the highest Tanimoto coefficients for corresponding known ligands.

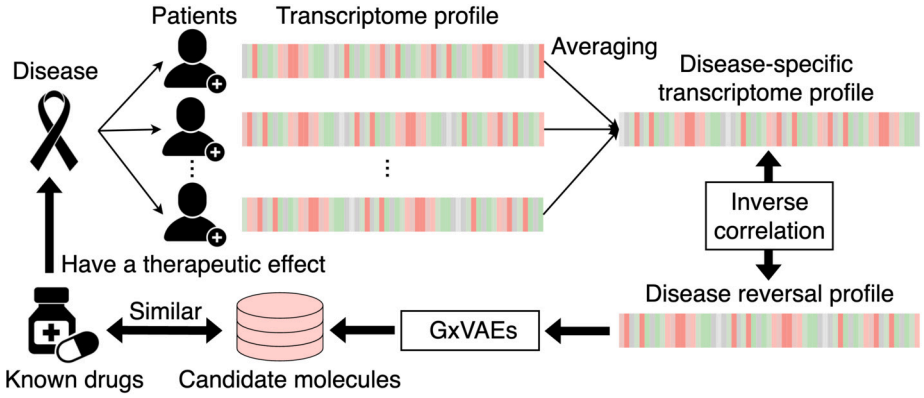


Fig. 8. GxVAEs for therapeutic molecule generation.

5. Conclusion

This study introduced GxVAEs, integrating two VAEs—ProfileVAE and MolVAE—to generate hit molecules from transcriptome profiles. ProfileVAE extracted features from transcriptome profiles, which MolVAE utilized as conditions to generate hit molecules. Experimental results demonstrated that GxVAEs surpassed current SOTA benchmarks, efficiently producing hit molecules from transcriptome profiles. Moreover, we showcased GxVAEs' capability to design molecular structures potentially effective in treating various diseases based on patient-specific disease profiles.

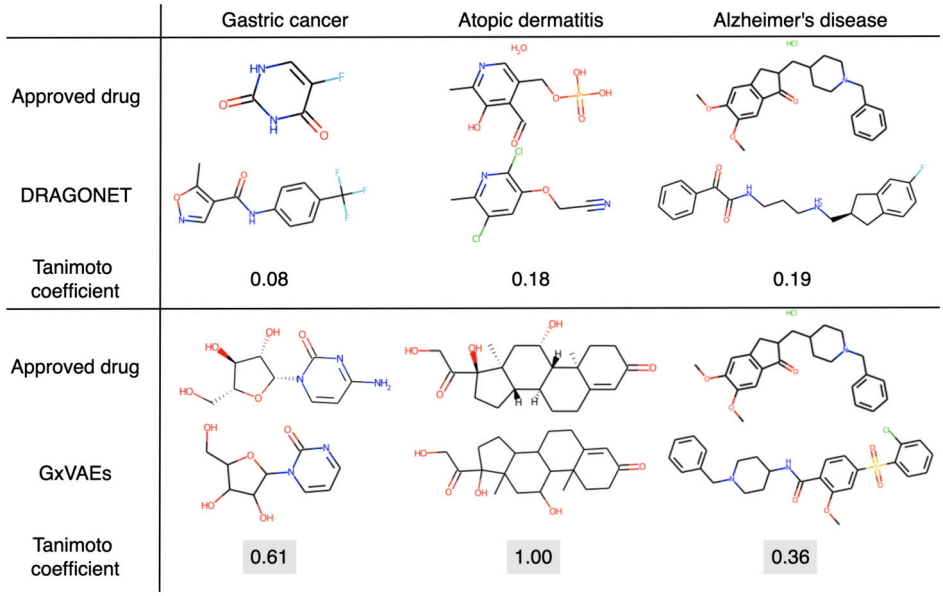


Fig. 9. Therapeutic molecule generation by DRAGONET baseline and GxVAEs for three diseases.

GxVAEs have two main limitations. First, the diversity of the generated molecules may be limited by the size of the latent space. If MolVAE uses a fixed latent vector, it could restrict the range of variability in the newly created molecules. Second, the model's performance can be affected by the quality and representativeness of the transcriptome profiles used during training, which may impact the relevance of the generated molecules to real-world disease contexts.

To address these issues in future work, we plan to improve the latent space representation and incorporate a broader range of diverse and comprehensive transcriptome profiles. Additionally, we aim to validate the efficacy of the candidate therapeutic molecules generated by GxVAEs in real disease contexts. By overcoming these limitations, we believe that GxVAEs can greatly enhance the process of generating hit molecules and offer valuable support to chemists and pharmacologists in their research endeavors.

CRediT authorship contribution statement

Chen Li: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Yoshihiro Yamanishi:** Writing – review & editing, Writing – original draft, Supervision, Software, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the International Research Fellow of Japan Society for the Promotion of Science (Postdoctoral Fellowships for Research in Japan [Standard]), AMED under Grant Number JP22nk0101111, and JSPS KAKENHI [grant numbers 20H05797, 21K18327].

Appendix A. Variant SMILES

Fig. A.1 demonstrates an example production of variant SMILES strings. Generally, a molecular graph has a canonical SMILES string (shown with the left figure). The numbers and arrows in the figure denote the traversal order of the atoms. Unlike in natural language processing, the canonical molecule has various SMILES representations (shown as the middle figure) according to different traversal orders (one example is shown in the right figure) called variant SMILES. However, the same molecular graph can be represented using these ten variant SMILES strings. Therefore, variant SMILES strings can be produced to improve the pretraining of the generator and prevent it from learning only a single semantic and syntactic feature.

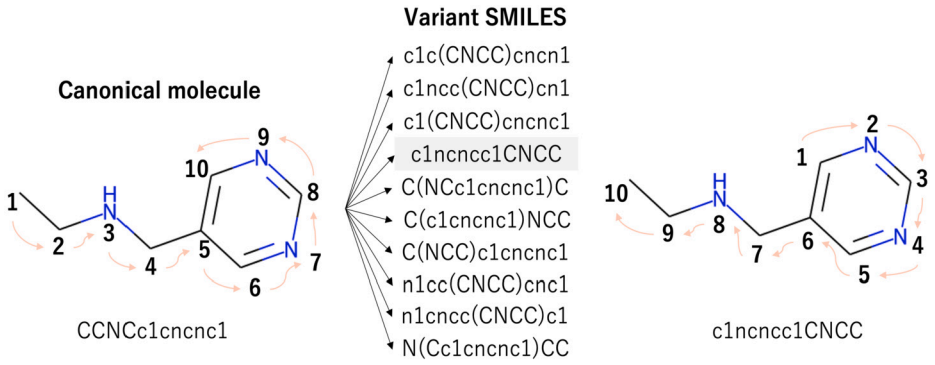


Fig. A.1. Example showing the production of variant SMILES strings.

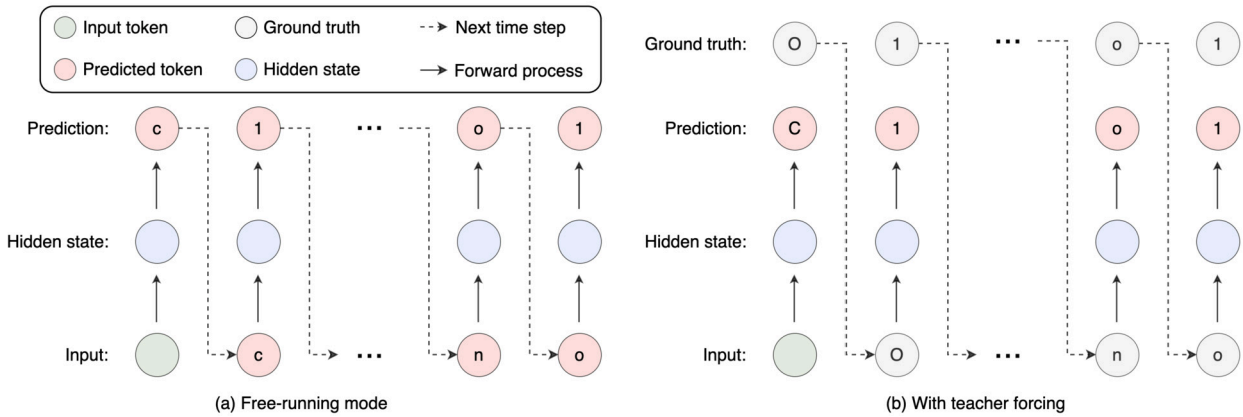


Fig. B.1. An illustration of the application of teacher forcing in a GRU model.

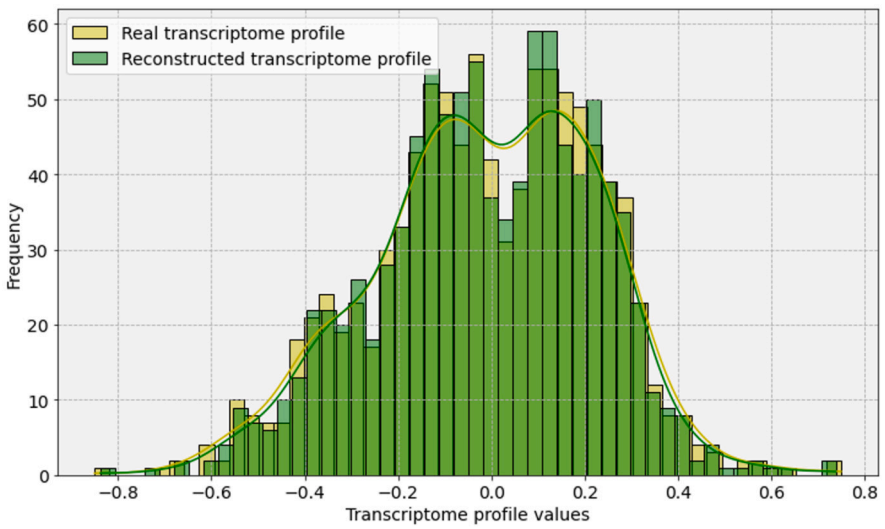


Fig. D.1. Average distribution of all transcriptome profiles exposed in the MCF7 cell.

Appendix B. Teacher forcing

Teacher forcing is commonly employed in the training of machine learning models, especially in the realm of sequence-to-sequence models like recurrent neural networks (RNNs) and their variants, such as long short-term memory networks (LSTMs) and gated recurrent units (GRUs). In GxVAEs, the primary purpose of employing teacher forcing is to expedite and stabilize the training process.

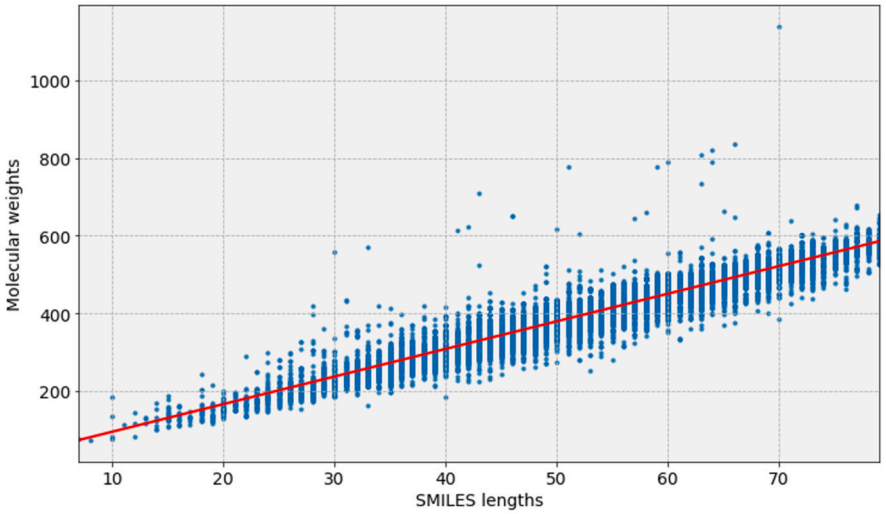


Fig. D.2. Relationship between molecular weights and SMILES lengths.

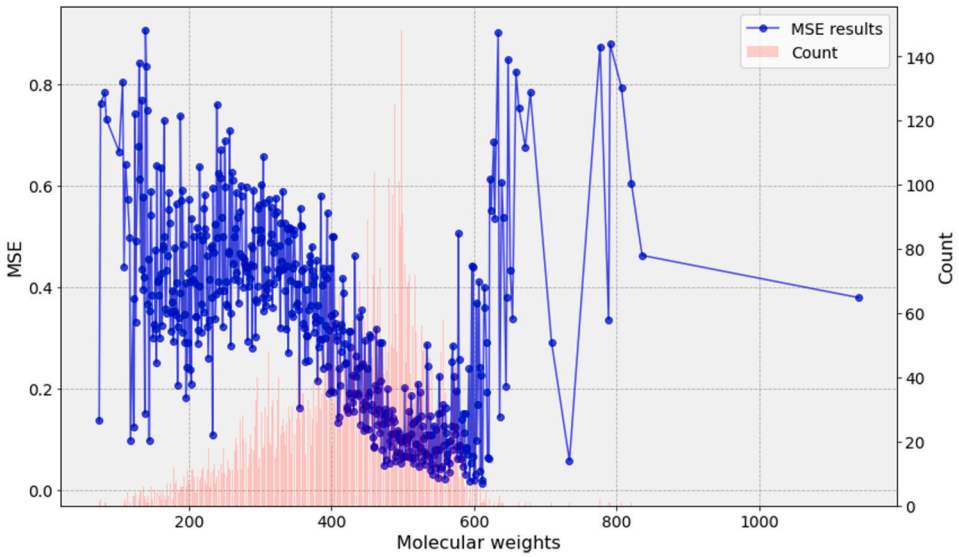
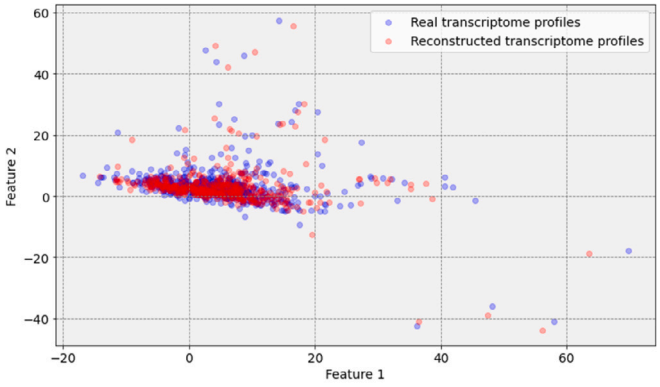


Fig. D.3. Relationship between reconstructed transcriptome profiles and molecular weights.

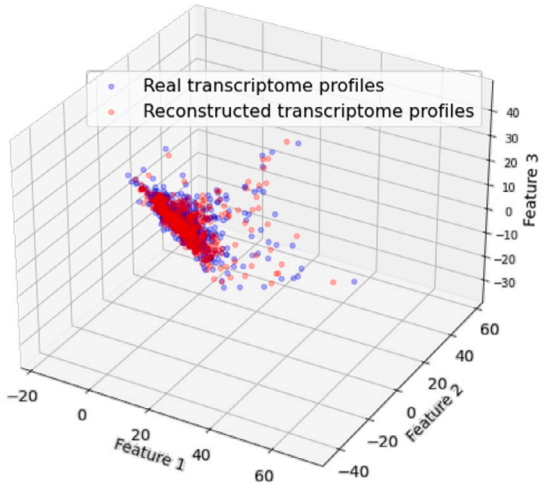
This technique involves using the ground truth target output from the training dataset as input for MolVAE during training, rather than relying solely on MolVAE's own generated output from the previous time step.

Fig. B.1 provides a visual representation of the application of teacher forcing in a GRU model. In the absence of the teacher forcing technique, commonly referred to as free-running mode (depicted in Fig. B.1 (a)), the predicted token at time step t is typically employed as the input for time step $t + 1$. However, this approach can lead to challenges when the prediction at time step t diverges from the label or ground truth at the same time step. Such discrepancies may propagate errors, resulting in subsequent predictions or generations deviating from the intended direction. Unlike the free-running mode, where the predicted token at time step t becomes the input for time step $t + 1$, teacher forcing introduces a distinctive approach, as illustrated in Fig. B.1 (b). In the context of teacher forcing, the GRU model is fed with the true or ground truth token at each time step during training, regardless of the model's own predictions. For instance, if the GRU model predicts an incorrect token at time step 1 (e.g., the atom "c"), differing from the ground truth (e.g., the atom "o"), teacher forcing uses "o" as the input for the second time step instead of the model's prediction.

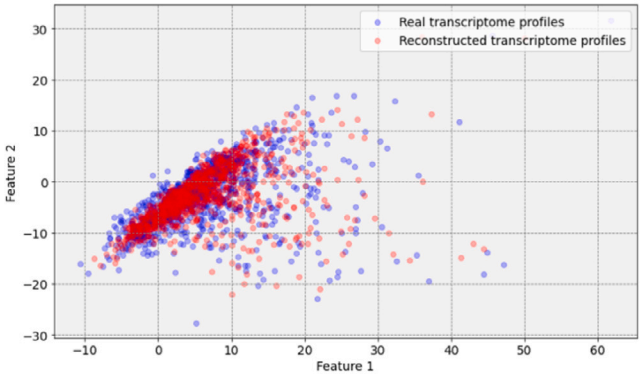
This technique aims to align the model's learning process more closely with the intended sequence, thereby enhancing stability and expediting convergence during training.



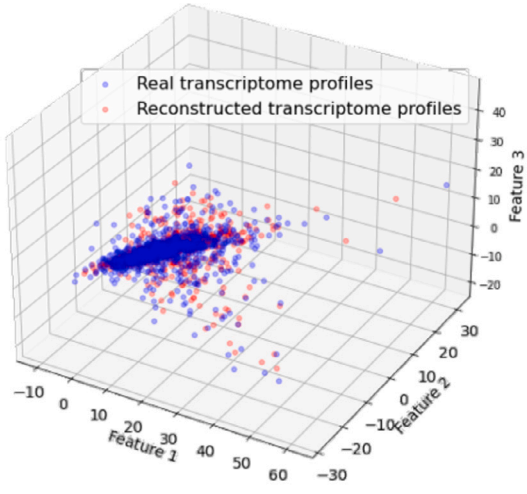
(a) 2D SVD visualization: 500 points.



(b) 3D SVD visualization: 500 points.



(c) 2D SVD visualization: 1000 points.



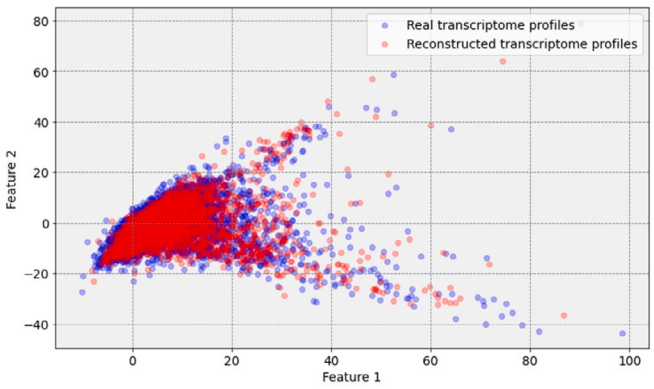
(d) 3D SVD visualization: 1000 points.

Fig. D.4. 2D and 3D SVD visualization plots of real and reconstructed transcriptome profiles.

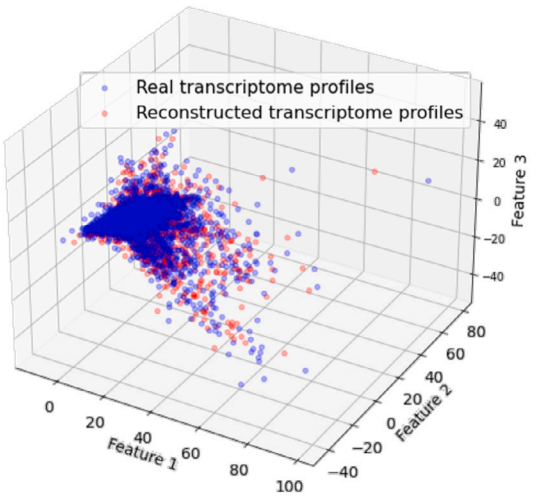
Appendix C. Experimental setup

In this section, we provide details on the computational resources and datasets used for training and evaluating the GxVAEs model.

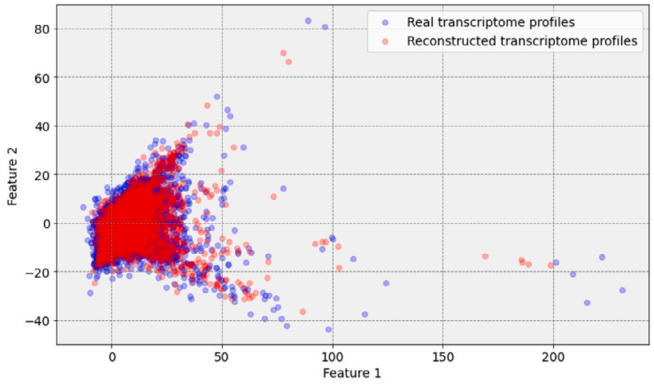
- **Computational Resources:** All experiments were conducted using an NVIDIA Tesla T4 GPU with CUDA version 11.4. The model was implemented in Python version 3.6.10.
- **Datasets:** The GxVAEs model was trained and validated using a dataset with 13,554 transcriptome profile samples for training and 200 samples for validation, each profile consisting of 978 dimensions.
- **Training Time:** The training of ProfileVAE required approximately 3 days to complete, while MolVAE was trained in about 6 hours.
- **Computational Efficiency:** The use of the NVIDIA Tesla T4 GPU enabled efficient processing of the large-scale dataset and accelerated the training process. CUDA 11.4 provided the necessary support for high-performance computations, ensuring that the training was completed within a reasonable time frame.



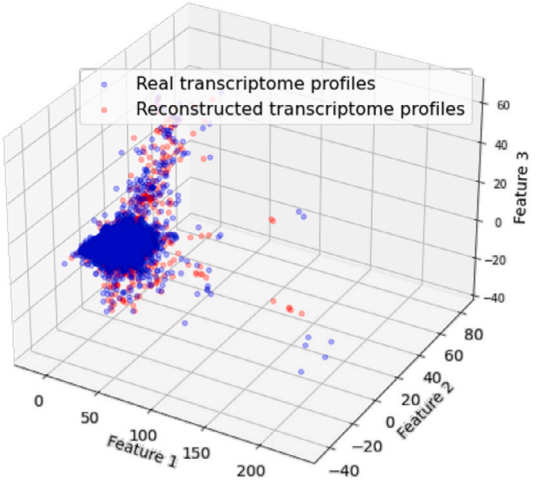
(e) 2D SVD visualization: 5000 points.



(f) 3D SVD visualization: 5000 points.



(g) 2D SVD: 10000 points.



(h) 3D SVD: 10000 points.

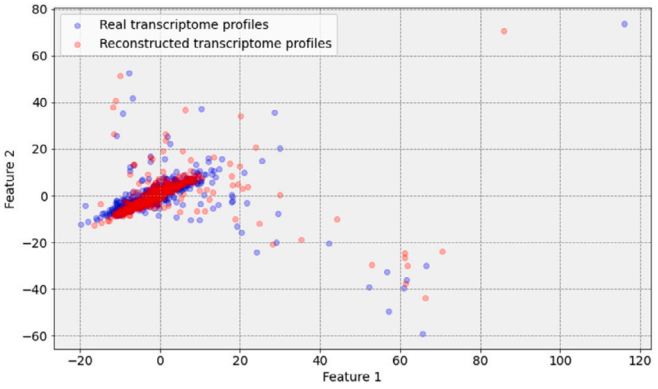
Fig. D.4. (*continued*)

Appendix D. Additional evaluation results of ProfileVAE

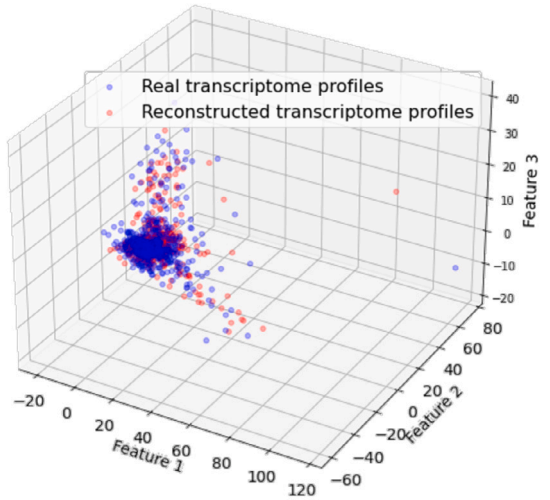
Fig. D.1 illustrates the average distribution of all transcriptome profiles exposed in the MCF7 cell, reconstructed by ProfileVAE. The average distribution of reconstructed transcriptome profiles closely overlaps with that of the real transcriptome profiles, thereby validating the effectiveness of ProfileVAE in feature extraction and reconstruction.

Fig. D.2 depicts the relationship between molecular weights and SMILES lengths. The data demonstrates a distinct positive linear relationship (highlighted by the red line), signifying that as the SMILES representation lengthens, the molecular weight of the molecule increases accordingly. This finding underscores the direct association between molecular complexity, as indicated by SMILES length, and molecular weight.

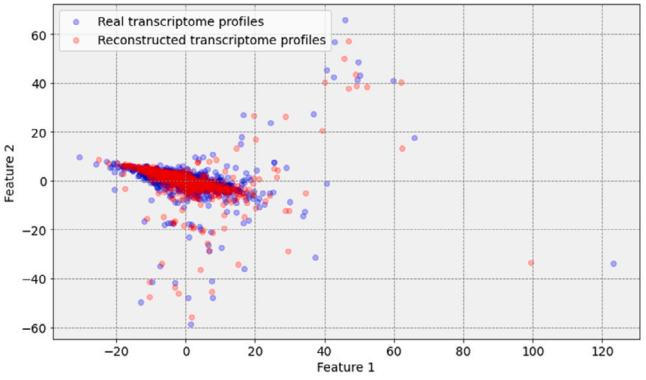
Figs. 3 (b) and D.3 depict the relationship between molecular weights, scaled by factors of 1 and 10, respectively, and the reconstructed transcriptome profiles. The results in Fig. D.3 detail that the MSE of transcriptome profiles exhibits an overall negative correlation with molecular weights within the range of [200, 600]. Due to the limited amount of data within the molecular weight



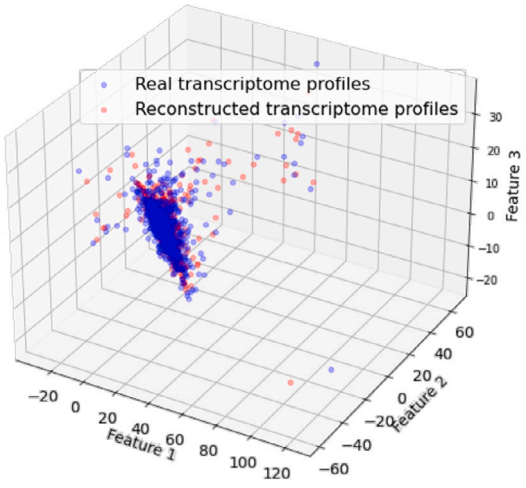
(a) 2D PCA visualization: 500 points.



(b) 3D PCA visualization: 500 points.



(c) 2D PCA visualization: 1000 points.



(d) 3D PCA visualization: 1000 points.

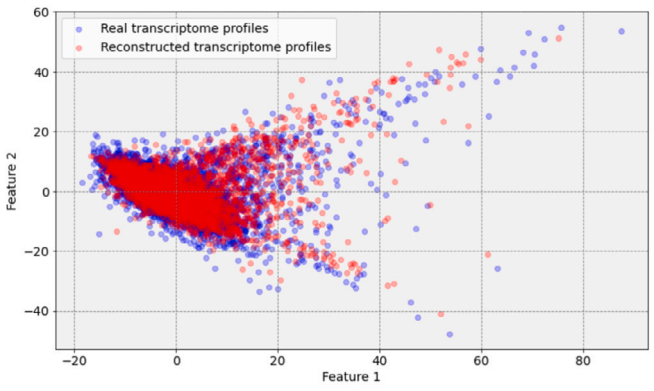
Fig. D.5. 2D and 3D PCA visualization plots of real and reconstructed transcriptome profiles.

ranges of [0, 100] and [600, 1000], approximately 300 and 100 instances, respectively, their MSE shows instability and tends to increase.

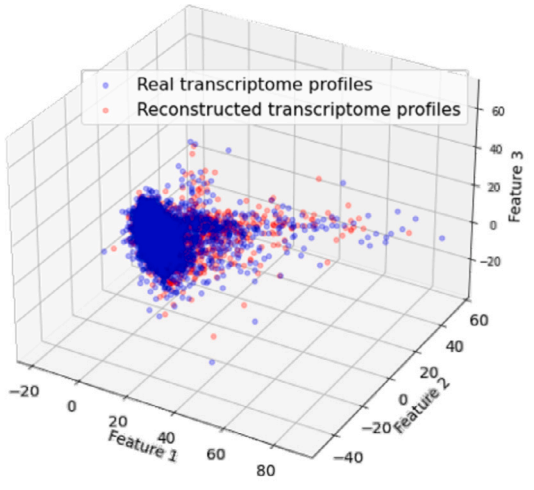
Figs. D.4 and D.5 show the SVD and PCA visualization plots of real and reconstructed transcriptome profiles, respectively. The results indicate that the 2D and 3D data points representing the dimension-reduced reconstructed transcriptome profiles closely overlap with those of the real transcriptome profiles, demonstrating the effectiveness of ProfileVAE.

Appendix E. Additional evaluation results of MolVAE

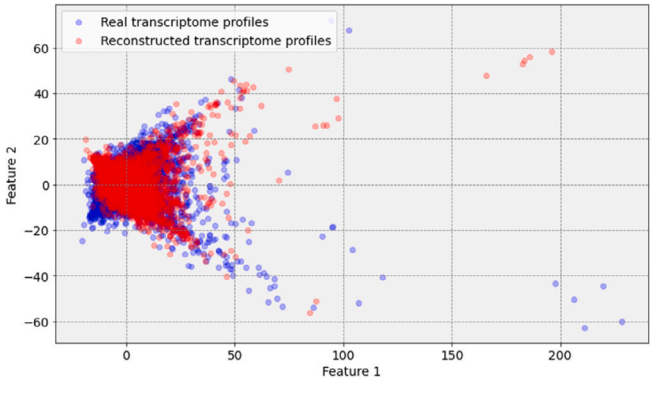
Fig. E.1 shows the reconstruction loss of the MolVAE and the ratio of valid molecules generated by MolVAE. The brown curve indicates the losses of MolVAE over the training epochs. The blue curve denotes the ratio of valid molecules generated by MolVAE with over training epochs. Note that the valid molecules are examined using the RDKit tool.



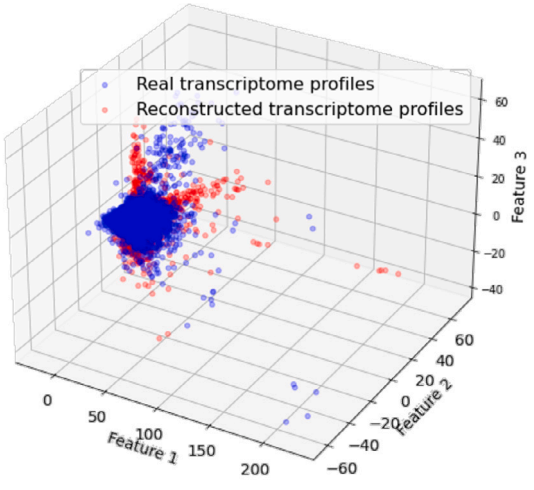
(e) 2D PCA visualization: 5000 points.



(f) 3D PCA visualization: 5000 points.



(g) 2D PCA: 10000 points.



(h) 3D PCA: 10000 points.

Fig. D.5. (*continued*)

Appendix F. Additional evaluation results of GxVAEs

For each target protein, we randomly selected 100 SMILES-like strings from its corresponding transcriptome profile. Subsequently, we calculated the Tanimoto coefficient between each generated sample and the known ligands of the target protein. The top 5 coefficients and their corresponding molecular structures are shown in Fig. F.1.

Data availability

Data will be made available on request.

References

[1] C.M. Dobson, et al., Chemical space and biology, Nature 432 (7019) (2004) 824–828.
[2] R.P. Hertzberg, A.J. Pope, High-throughput screening: new technology for the 21st century, Curr. Opin. Chem. Biol. 4 (4) (2000) 445–451.
[3] J.W. Scannell, J. Bosley, J.A. Hickman, G.R. Dawson, H. Truebel, G.S. Ferreira, D. Richards, J.M. Treherne, Predictive validity in drug discovery: what it is, why it matters and how to improve it, Nat. Rev. Drug Discov. 21 (12) (2022) 915–931.

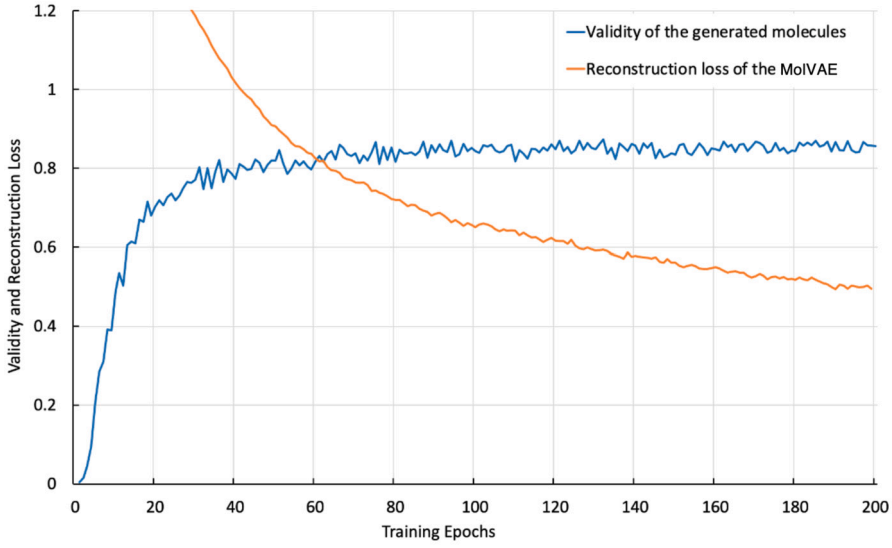
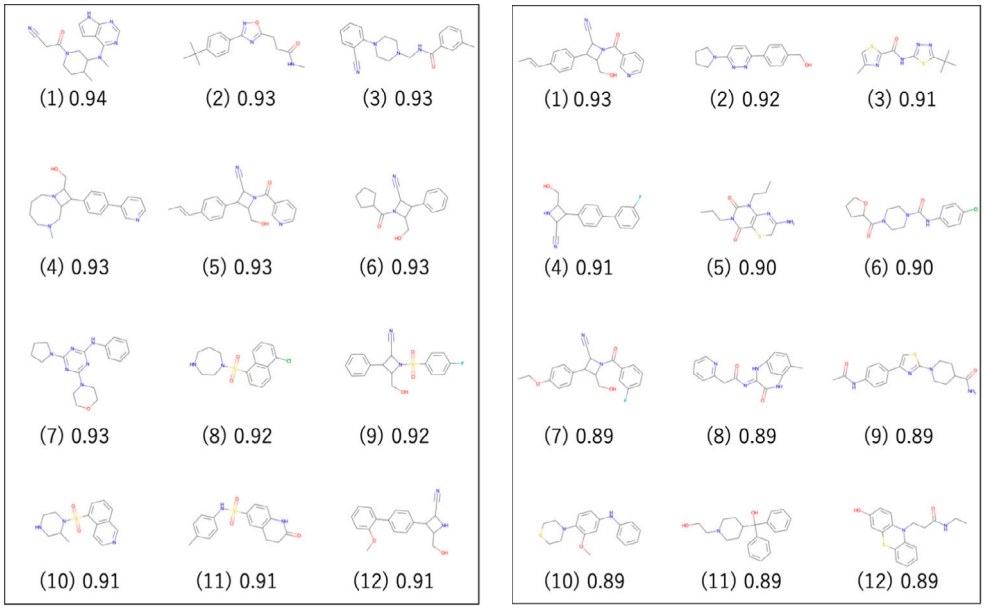


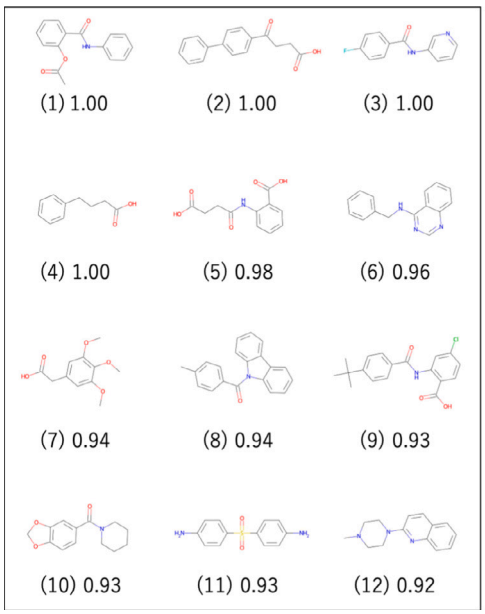
Fig. E.1. Change curves for the reconstruction loss of the real dataset and the ratio of valid molecules to all molecules generated by GxVAEs.



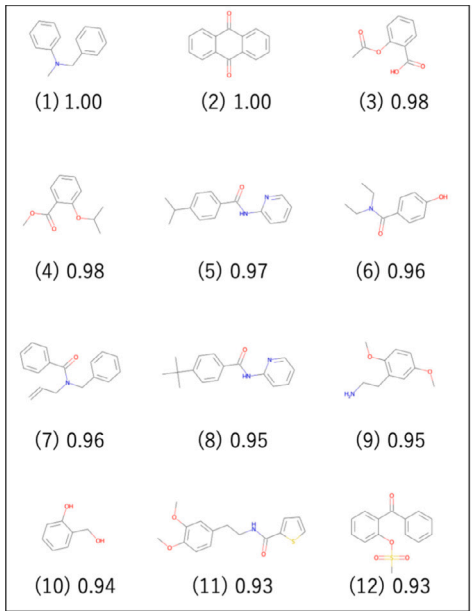
(a) Real molecules with QED scores. (b) Generated molecules with QED scores.

Fig. E.2. Top 12 molecules sorted by property scores from training and generated sets by MolVAE.

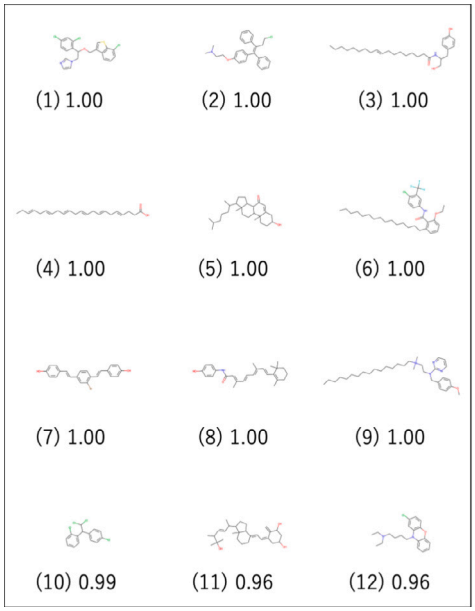
[4] A.Z. Rahman, C. Liu, H. Sturm, A.M. Hogan, R. Davis, P. Hu, S.T. Cardona, A machine learning model trained on a high-throughput antibacterial screen increases the hit rate of drug discovery, PLoS Comput. Biol. 18 (10) (2022) 1–22.
[5] S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham, D. Greyson, The cost of drug development: a systematic review, Health Policy 100 (1) (2011) 4–17.
[6] B. Shaker, S. Ahmad, J. Lee, C. Jung, D. Na, In silico methods and tools for drug discovery, Comput. Biol. Med. 137 (2021) 104851.
[7] A.V. Sadybekov, V. Katritch, Computational approaches streamlining drug discovery, Nature 616 (7958) (2023) 673–685.
[8] K.-K. Mak, Y.-H. Wong, M.R. Pichika, Artificial intelligence in drug discovery and development, in: Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays, 2023, pp. 1–38.
[9] G.L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P.L.C. Farias, A. Aspuru-Guzik, Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models, arXiv preprint, arXiv:1705.10843, 2017.
[10] N. De Cao, T. Kipf, MolGAN: an implicit generative model for small molecular graphs, ArXiv preprint, arXiv:1805.11973, 2018.
[11] C. Li, C. Yamanaka, K. Kaitoh, Y. Yamanishi, Transformer-based objective-reinforced generative adversarial network to generate desired molecules, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22), 2022, pp. 3884–3890.
[12] A.F. Oliveira, J.L. Da Silva, M.G. Quiles, Molecular property prediction and molecular design using a supervised grammar variational autoencoder, J. Chem. Inf. Model. 62 (4) (2022) 817–828.



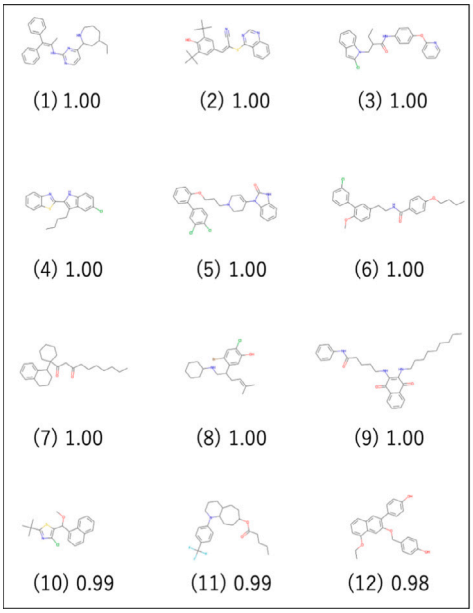
(c) Real molecules with SA scores.



(d) Generated molecules with SA scores.



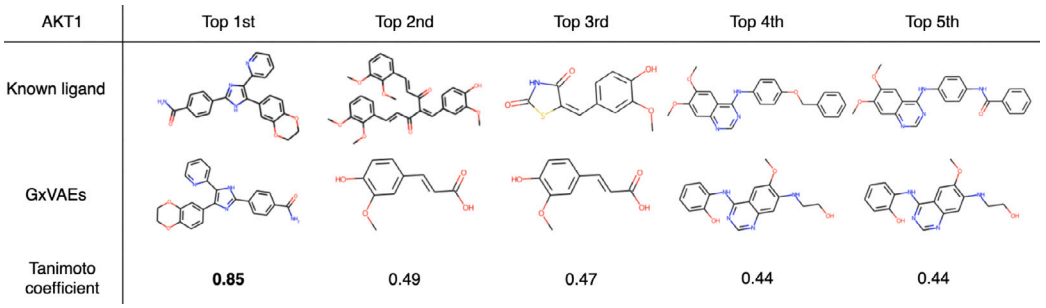
(e) Real molecules with logP scores.



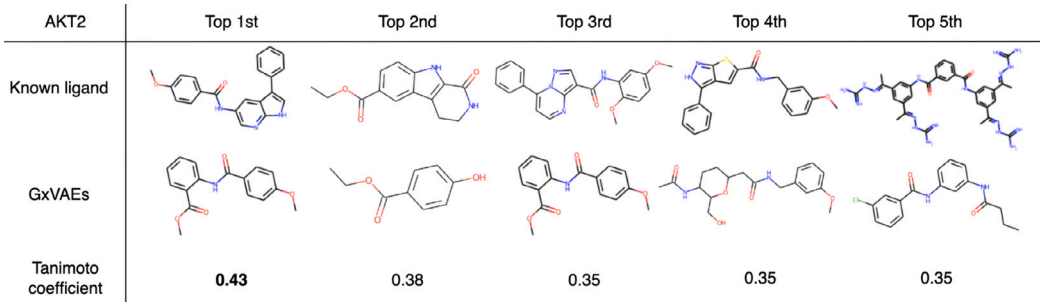
(f) Generated molecules with logP scores.

Fig. E.2. (*continued*)

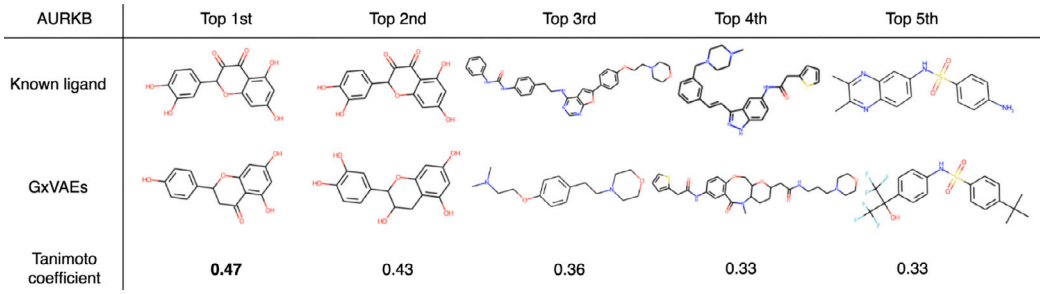
[13] O. Dollar, N. Joshi, D.A. Beck, J. Pfaendtner, Attention-based generative models for de novo molecular design, Chem. Sci. 12 (24) (2021) 8362–8372.
[14] M.J. Kusner, B. Paige, J.M. Hernández-Lobato, Grammar variational autoencoder, in: Proceedings of the International Conference on Machine Learning, PMLR, 2017, pp. 1945–1954.
[15] W. Jin, R. Barzilay, T. Jaakkola, Junction tree variational autoencoder for molecular graph generation, in: Proceedings of the International Conference on Machine Learning, PMLR, 2018, pp. 2323–2332.
[16] J. Born, M. Manica, A. Oskooei, J. Cadow, G. Markert, M.R. Martínez, PaccMannRL: de novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning, iScience 24 (4) (2021) 102269.
[17] J. Mun, G. Choi, B. Lim, A guide for bioinformaticians: omics-based drug discovery for precision oncology, Drug Discov. Today 25 (11) (2020) 1897–1904.
[18] M.H. Asyali, D. Colak, O. Demirkaya, M.S. Inan, Gene expression profile classification: a review, Curr. Bioinform. 1 (1) (2006) 55–73.
[19] M. Kang, E. Ko, T.B. Mersha, A roadmap for multi-omics data integration using deep learning, Brief. Bioinform. 23 (1) (2022) 1–16.
[20] O. Méndez-Lucio, B. Baillif, D.-A. Clevert, D. Rouquié, J. Wichard, De novo generation of hit-like molecules from gene expression signatures using artificial intelligence, Nat. Commun. 11 (1) (2020) 10.



(a) Top 5 Tanimoto coefficients and corresponding molecular structures for AKT1.



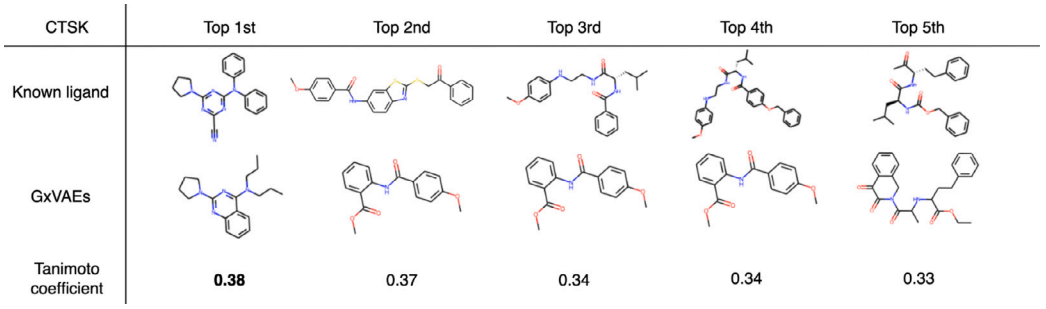
(b) Top 5 Tanimoto coefficients and corresponding molecular structures for AKT2.



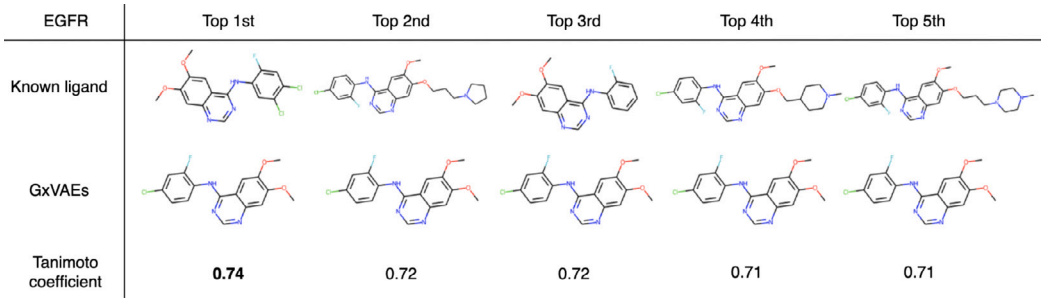
(c) Top 5 Tanimoto coefficients and corresponding molecular structures for AURKB.

Fig. F.1. Top 5 Tanimoto coefficients between known ligands and 100 generated molecules produced by GxVAEs.

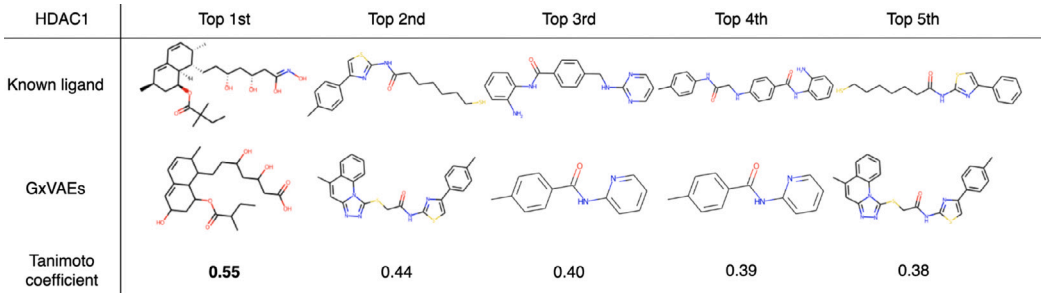
[21] K. Kaitoh, Y. Yamanishi, TRIOMPHE: transcriptome-based inference and generation of molecules with desired phenotypes by machine learning, J. Chem. Inf. Model. 61 (9) (2021) 4303–4320.
[22] C. Li, Y. Yamanishi, GxVAEs: two joint VAEs generate hit molecules from gene expression profiles, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 13455–13463.
[23] D. Weininger, SMILES: a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36.
[24] D.E. Manolopoulos, P.W. Fowler, Molecular graphs, point groups, and fullerenes, J. Chem. Phys. 96 (10) (1992) 7603–7614.
[25] O. Ganea, L. Pattanaik, C. Coley, R. Barzilay, K. Jensen, W. Green, T. Jaakkola, GeoMol: torsional geometric generation of molecular 3D conformer ensembles, Adv. Neural Inf. Process. Syst. 34 (2021) 13757–13769.
[26] J. Arús-Pous, A. Patronov, E.J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen, O. Engkvist, SMILES-based deep generative scaffold decorator for de-novo drug design, J. Cheminform. 12 (2020) 1–18.
[27] F. Grisoni, M. Moret, R. Lingwood, G. Schneider, Bidirectional molecule generation with recurrent neural networks, J. Chem. Inf. Model. 60 (3) (2020) 1175–1183.
[28] H. Tang, C. Li, S. Jiang, H. Yu, S. Kamei, Y. Yamanishi, Y. Morimoto, EarlGAN: an enhanced actor–critic reinforcement learning agent-driven gan for de novo drug design, Pattern Recognit. Lett. 175 (2023) 45–51.
[29] C. Li, Y. Yamanishi, TenGAN: pure transformer encoders make an efficient discrete GAN for de novo molecular generation, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, 2024, pp. 361–369.
[30] C. Li, Y. Yamanishi, SpotGAN: a reverse-transformer gan generates scaffold-constrained molecules with property optimization, in: ECML-PKDD, 2023, pp. 3884–3890.
[31] L. Huang, H. Zhang, T. Xu, K.-C. Wong, MDM: molecular diffusion model for 3D molecule generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 5105–5112.
[32] H. Gong, Q. Liu, S. Wu, L. Wang, Text-guided molecule generation with diffusion language model, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 109–117.
[33] N. O'Boyle, A. Dalke, DeepSMILES: an adaptation of smiles for use in machine-learning of chemical structures, ChemRxiv (2018).
[34] A.S. Bhadwal, K. Kumar, N. Kumar, GenSMILES: an enhanced validity conscious representation for inverse design of molecules, Knowl.-Based Syst. 268 (2023) 110429.



(d) Top 5 Tanimoto coefficients and corresponding molecular structures for CTSK.



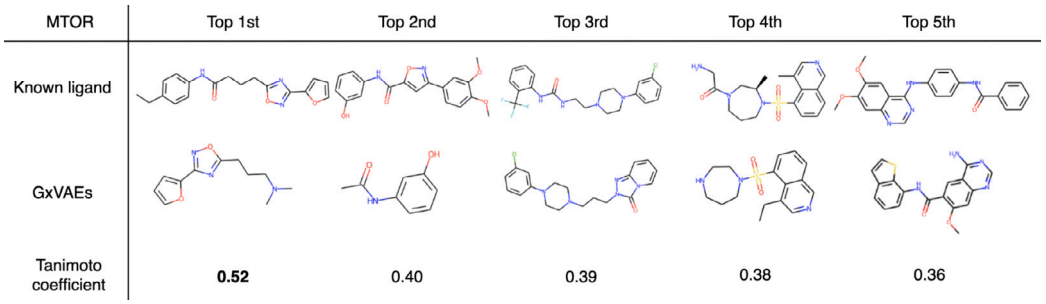
(e) Top 5 Tanimoto coefficients and corresponding molecular structures for EGFR.



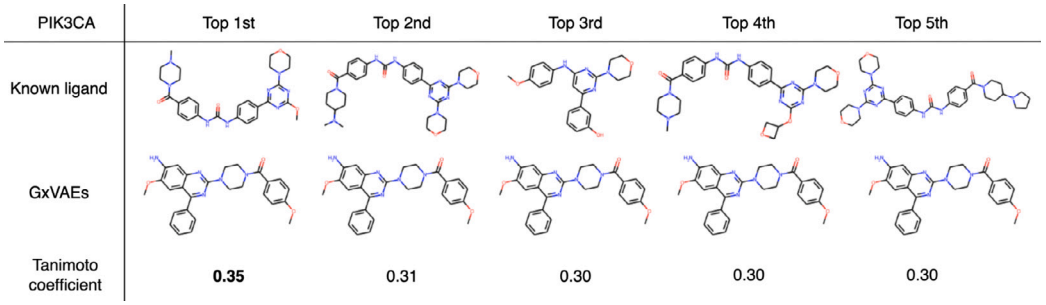
(f) Top 5 Tanimoto coefficients and corresponding molecular structures for HDAC1.

Fig. F.1. (*continued*)

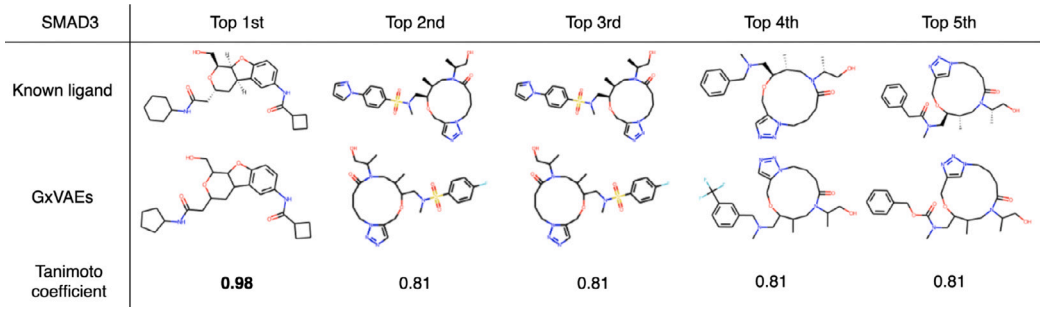
[35] M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N.C. Frey, P. Friederich, T. Gaudin, A.A. Gayle, K.M. Jablonka, et al., SELFIES and the future of molecular string representations, Patterns 3 (10) (2022).
[36] E. Noutahi, C. Gabellini, M. Craig, J.S. Lim, P. Tossou, Gotta be SAFE: a new framework for molecular design, Digit. Discov. 3 (4) (2024) 796–804.
[37] H. Dai, B. Dai, L. Song, Discriminative embeddings of latent variable models for structured data, in: Proceedings of the International Conference on Machine Learning, PMLR, 2016, pp. 2702–2711.
[38] S. Pölsterl, C. Wachinger, Adversarial learned molecular graph inference and generation, in: ECML-PKDD, 2021, pp. 173–189.
[39] V.G. Satorras, E. Hoogeboom, M. Welling, E (n) equivariant graph neural networks, in: Proceedings of the International Conference on Machine Learning, PMLR, 2021, pp. 9323–9332.
[40] J. Zhao, Q. Feng, W.-Q. Wei, Integration of omics and phenotypic data for precision medicine, in: Systems Medicine, Springer, 2022, pp. 19–35.
[41] S. Morganti, P. Tarantino, E. Ferraro, P. D'Amico, B.A. Duso, G. Curigliano, Next generation sequencing (NGS): a revolutionary technology in pharmacogenomics and personalized medicine in cancer, in: Translational Research and Onco-Omics Applications in the Era of Cancer Personal Genomics, 2019, pp. 9–30.
[42] R. Pereira, J. Oliveira, M. Sousa, Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics, J. Clin. Med. 9 (1) (2020) 132.
[43] B. Turanli, K. Karagoz, G. Gulfidan, R. Sinha, A. Mardinoglu, K.Y. Arga, A network-based cancer drug discovery: from integrated multi-omics approaches to precision medicine, Curr. Pharm. Des. 24 (32) (2018) 3778–3790.
[44] B. Chen, L. Garmire, D.F. Calvisi, M.-S. Chua, R.K. Kelley, X. Chen, Harnessing big 'omics' data and AI for drug discovery in hepatocellular carcinoma, Nat. Rev. Gastroenterol. Hepatol. 17 (4) (2020) 238–251.
[45] K. Cho, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.
[46] C. Yan, J. Yang, H. Ma, S. Wang, J. Huang, Molecule sequence generation with rebalanced variational autoencoder loss, J. Comput. Biol. 30 (1) (2023) 82–94.
[47] J.M. Joyce, Kullback-Leibler divergence, in: International Encyclopedia of Statistical Science, Springer, 2011, pp. 720–722.
[48] Q. Duan, C. Flynn, M. Niepel, M. Hafner, J.L. Muhlich, N.F. Fernandez, A.D. Rouillard, C.M. Tan, E.Y. Chen, T.R. Golub, et al., LINCS canvas browser: interactive web app to query, browse and interrogate LINCS l1000 gene expression signatures, Nucleic Acids Res. 42 (W1) (2014) W449–W460.
[49] Z. Wang, C.D. Monteiro, K.M. Jagodnik, N.F. Fernandez, G.W. Gundersen, A.D. Rouillard, S.L. Jenkins, A.S. Feldmann, K.S. Hu, M.G. McDermott, et al., Extraction and analysis of signatures from the gene expression Omnibus by the crowd, Nat. Commun. 7 (1) (2016) 12846.



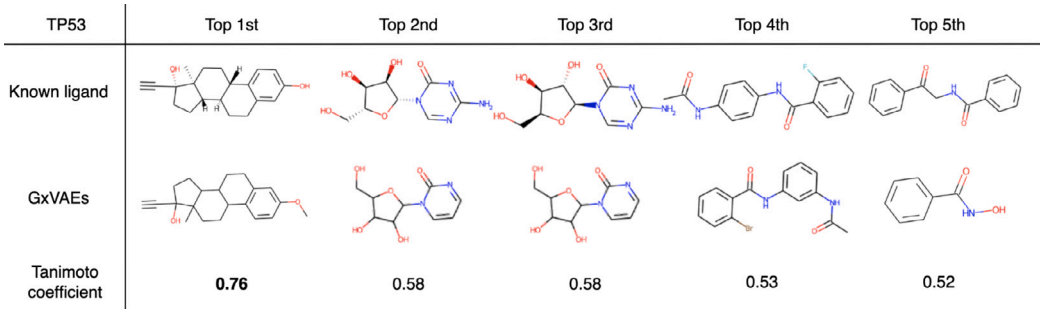
(g) Top 5 Tanimoto coefficients and corresponding molecular structures for MTOR.



(h) Top 5 Tanimoto coefficients and corresponding molecular structures for PIK3CA.



(i) Top 5 Tanimoto coefficients and corresponding molecular structures for SMAD3.



(j) Top 5 Tanimoto coefficients and corresponding molecular structures for TP53.

Fig. F.1. (*continued*)

[50] G. Landrum, Rdkit documentation, Release 1 (1–79) (2013) 4.
[51] Y. Kwon, J. Lee, MolFinder: an evolutionary algorithm for the global optimization of molecular properties and the extensive exploration of chemical space using smiles, J. Cheminform. 13 (2021) 1–14.
[52] P. Ertl, A. Schuffenhauer, Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, J. Cheminform. 1 (2009) 1–11.
[53] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (5) (2010) 742–754.
[54] C. Yamanaka, S. Uki, K. Kaitoh, M. Iwata, Y. Yamanishi, De novo drug design based on patient gene expression profiles via deep learning, Mol. Inform. (2023).