GxVAEs: Two Joint VAEs Generate Hit Molecules from Gene Expression Profiles

Chen Li and Yoshihiro Yamanishi

Graduate School of Informatics, Nagoya University, Chikusa, Nagoya, 464-8601, Japan
li.chen.z2@a.mail.nagoya-u.ac.jp, yamanishi@i.nagoya-u.ac.jp

Abstract

The *de novo* generation of hit-like molecules that show bioactivity and drug-likeness is an important task in computer-aided drug discovery. Although artificial intelligence can generate molecules with desired chemical properties, most previous studies have ignored the influence of disease-related cellular environments. This study proposes a novel deep generative model called GxVAEs to generate hit-like molecules from **g**ene e**x**pression profiles by leveraging two joint variational autoencoders (**VAEs**). The first VAE, ProfileVAE, extracts latent features from gene expression profiles. The extracted features serve as the conditions that guide the second VAE, which is called MolVAE, in generating hit-like molecules. GxVAEs bridge the gap between molecular generation and the cellular environment in a biological system, and produce molecules that are biologically meaningful in the context of specific diseases. Experiments and case studies on the generation of therapeutic molecules show that GxVAEs outperforms current state-of-the-art baselines and yield hit-like molecules with potential bioactivity and drug-like properties. We were able to successfully generate the potential molecular structures with therapeutic effects for various diseases from patients' disease profiles.

Introduction

Hit identification, which involves discovering hit molecules with the desired bioactivity and therapeutic effects within an infinite chemical space, is a critical challenge in drug discovery (Dobson et al. 2004). Over 10^{60} organic molecules theoretically exist, but only a limited number of these molecules would have drug-like properties. Traditionally, hit identification campaigns have been conducted using experimental approaches such as high-throughput screening (HTS) (Hertzberg and Pope 2000). HTS provides a valuable methodology for identifying hits with desired bioactivity (Scannell et al. 2022). However, the hit rate observed in HTS campaigns tends to be relatively low, which results in many molecules that are inactive or that hold less promise being screened out (Rahman et al. 2022). Additionally, experimental approaches to drug discovery typically rely on labor-intensive and time-consuming processes. Unfortunately, the entire process of drug development takes up to 12 years and requires over 1.8 billion US dollars in capital investment. Despite rigorous pre-market drug testing, the failure rate of candidate molecules has exceeded 90% (Shaker et al. 2021).

De novo molecular generation using artificial intelligence (AI) techniques has emerged as a promising approach for computer-aided drug discovery. The power of deep learning and computational chemistry enables the generation of new molecules with desired bioactivity for specific therapeutic targets. Deep generative models, such as generative adversarial networks (GANs) (Guimaraes et al. 2017; De Cao and Kipf 2018; Li et al. 2022) and variational autoencoders (VAEs) (Oliveira, Da Silva, and Quiles 2022; Dollar et al. 2021; Kusner, Paige, and Hernández-Lobato 2017; Jin, Barzilay, and Jaakkola 2018) accelerate the drug discovery process by employing computational models to generate molecules having a specific bioactivity. Such models learn the underlying patterns and relationships within molecular structures by training on datasets of known molecules with the desired properties. By capturing the essential features of molecules, deep generative models can produce new molecules with similar structures and related properties. Most AI-driven molecular generation models focus on the generation of new molecules with improved chemical properties of interest. However, these approaches generate molecules without considering the biological environment and cellular context of a particular disease.

Generating hit-like molecules from omics data addresses these limitations by considering the broader biological context to generate molecules that are more likely to interact effectively with disease-related cellular machinery (Born et al. 2021). Omics data, such as genomics, epigenomics, and transcriptomics, provide a comprehensive molecular landscape for describing the human cellular response to drug therapy and the pathological history of patients with disease (Mun, Choi, and Lim 2020). Omics data that represents drug activity are therefore an essential resource for modern drug development. For example, gene expression profiles provide valuable insights into the activity levels of genes in different cellular contexts, including in specific diseases (Asyali et al. 2006). Gene expression profiles capture dynamic changes in gene expression that occur in response to various physiological and pathological conditions. Analysis of gene expression profiles provides insight into the cellular environment and its impact on disease progression and therapeutic response.

While omics-based approaches to drug discovery offer promising opportunities, it is essential to understand their limitations (Kang, Ko, and Mersha 2022). First, the availability of comprehensive and high-quality omics data for molecules is limited. Moreover, the interpretation and integration of omics data is complex and challenging. Analysis of omics data usually requires advanced bioinformatics and statistical techniques to extract meaningful insights and identify relevant molecular signatures. Interpretation of omics data in the context of drug discovery also requires a deep understanding of the underlying biological and disease mechanisms. Thus far, few previous studies have exploited omics data, especially gene expression profiles, to generate hit-like molecules (Méndez-Lucio et al. 2020; Kaitoh and Yamanishi 2021). However, the source molecules of these models are not associated with the corresponding gene expression profiles, and the performance of the generated hit-like molecules (e.g., association with target proteins) is relatively low and still shows potential for improvement.

In this study, we proposed GxVAEs to generate hit-like molecules from **g**ene e**x**pression profiles using two joint **VAEs** (i.e., ProfileVAE and MolVAE). First, ProfileVAE functions as a feature extractor to extract latent features from gene expression profiles. The extracted features are then used as conditions to guide MolVAE in generating hit-like molecules. GxVAEs bridges the gap between molecular generation and the cellular environment of a biological system, making the produced molecules more biologically meaningful in the context of specific diseases. The main contributions of this study are summarized as follows:

- **Hit-like molecular generation that considered the cellular environment**: This study considers the influence of the cellular environment on specific diseases during the molecular generation process.
- **A simple but effective model architecture**: The combination of two VAEs successfully generated hit-like molecules from gene expression profiles.
- **Superior performance over state-of-the-art (SOTA) models**: Experiments and case studies demonstrate that the proposed GxVAEs outperform the current SOTA models for the same objectives.

Related Work

De Novo Molecular Generation

De novo molecular generation aims to produce new molecules with specific chemical properties. Generally, the data structures used for *de novo* molecular generation are molecular graphs (Manolopoulos and Fowler 1992) and simplified molecular input line entry system (SMILES) strings (Weininger 1988). Molecular graphs contain more structural information about molecules than SMILES strings. The atoms and chemical bonds in the molecular graph are represented as nodes and edges, respectively. Each node contains information about the atom type. The edges indicate the connectivity between atoms and specify the type of bond, such as a single, double, or triple. A SMILES string is a sequential representation of a molecule and is more concise than a molecular graph. By representing molecular structures in a linear sequence, SMILES strings allow a compact and standardized representation of molecules. SMILES strings are easy to store in chemical databases and to process in molecular modeling. TransVAE (Dollar et al. 2021) and GrammarVAE (Kusner, Paige, and Hernández-Lobato 2017) are two deep generative models that produce molecules from SMILES strings using VAEs. JTVAE (Jin, Barzilay, and Jaakkola 2018) uses a node tree to generate molecules from a molecular graph. It first creates a tree-structured scaffold on a substructure and then combines the tree-structured scaffold into a molecule using a graphical message-passing network (Dai, Dai, and Song 2016). MolGAN (De Cao and Kipf 2018) and TransORGAN (Li et al. 2022) are discrete GANs that use reinforcement learning to generate molecular graphs and SMILES strings, respectively.

Hit-Like Molecular Generation

All the aforementioned studies used only chemical structure information and improved the relevant chemical properties of the generated molecules. However, these approaches do not consider the biological environment and the cellular context of a particular disease (Zhao, Feng, and Wei 2022) nor cell-specific activities of the produced molecules (Morganti et al. 2019; Pereira, Oliveira, and Sousa 2020). With the development of omics data analysis, especially gene expression profiles used in drug discovery (Turanli et al. 2018; Chen et al. 2020), chemists have begun to use comprehensive biological response information to generate therapeutic molecules for the treatment of specific diseases.

ExpressionGAN (Méndez-Lucio et al. 2020) and TRIOMPHE (Kaitoh and Yamanishi 2021) are the SOTA deep generative models that are most relevant to our study on generating hit-like molecules for target proteins with no prior annotation of the target training molecules. ExpressionGAN is a GAN model that bridges systems biology and molecular design to produce molecules with a high probability of inducing the desired transcriptome features automatically from gene expression profiles. TRIOMPHE first calculates the correlation of ligand-target interactions between the chemically induced transcriptome profiles of cellular responses to molecular treatment and the transcriptome profiles of gene perturbations that reflect cellular responses to gene knockdown or gene overexpression of target proteins. Next, a VAE was employed to generate a new molecule with the desired transcriptome profile. Finally, molecules showing bioactivity against the target protein were automatically designed using the transcriptome profile of interest.

Gene expression profiles can be employed to generate bioactive molecules against arbitrary target proteins by considering cellular context. However, ExpressionGAN generates hit-like molecules with low validity ($> 8.5\%$), and its ability to reproduce known ligands is limited. Moreover, the transcriptional correlation between ligands and targets is unclear. Additionally, TRIOMPHE only utilizes gene expression profiles in correlation calculations, whose VAE is not involved in molecular generation. To address these issues, GxVAEs were proposed in this study. By bridging the gap between the cellular environment and the generation of ther-

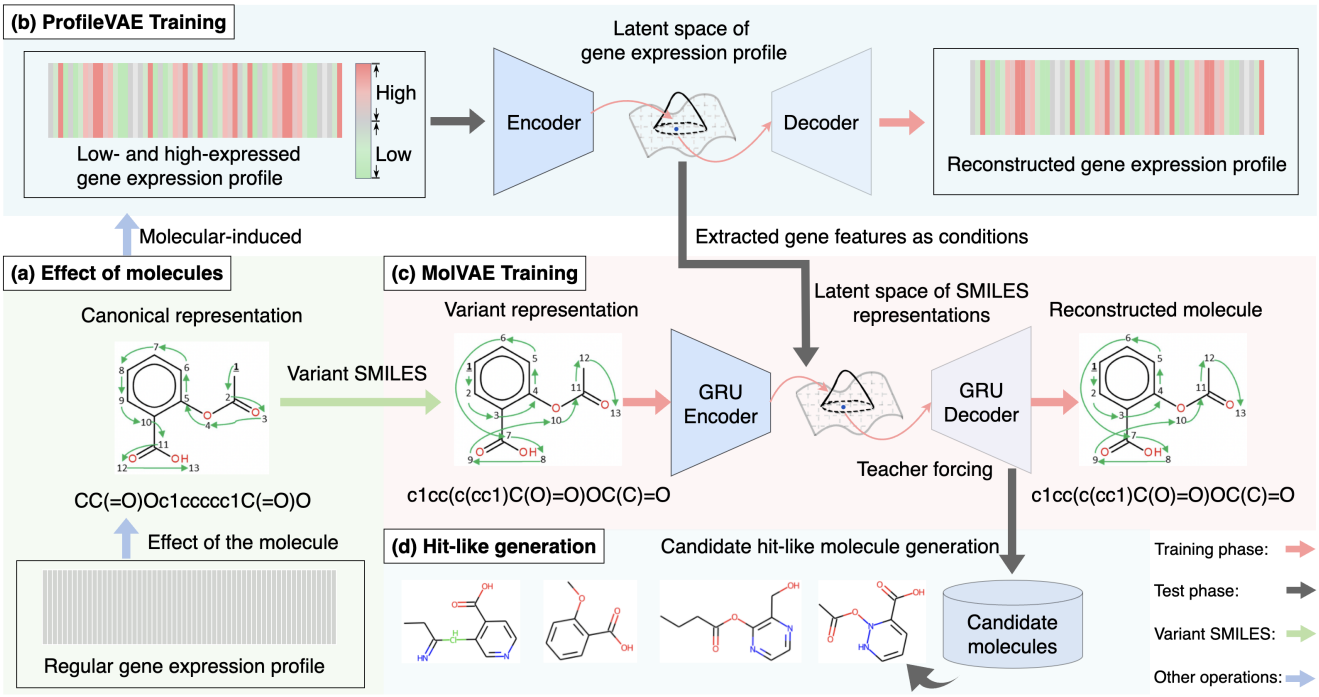

Figure 1: Overview of the proposed GxVAEs. (a) Exposure to a molecule in a cell induces a specific gene expression profile. The regular gene expression profile is perturbed by the molecule "CC(=O)oc1ccccc1C(=O)O", which results in a profile of low or high gene expression values. (b) The gene expression profile that is directly affected by the molecule is input into the ProfileVAE encoder and is mapped to the latent space of gene expression profiles. Then, the feature vector is used to reconstruct the gene expression profile by the ProfileVAE decoder. (c) The extracted features of the gene expression profile are combined with the variant SMILES representation of the molecule (i.e., "c1cc(c(cc1)C(O)=O)OC(C)=O") as the input to the bidirectional gated recurrent units (GRU) encoder of MolVAE. The GRU encoder of MolVAE extracts condition-constrained molecular features and maps them to the latent space of the SMILES representation of molecules. Then, the bidirectional GRU decoder reconstructs the feature vector into the original molecular structure. In the training phase, teacher forcing is employed to stabilize learning and accelerate the convergence of MolVAE. In the inference phase (the process shown by black arrows), ProfileVAE's decoder and MolVAE's encoder are discarded, and only ProfileVAE's encoder and MolVAE's decoder are used to generate hit-like molecules. An unknown gene expression profile is input to ProfileVAE's encoder. The extracted features are input into MolVAE's decoder. (d) Finally, the candidate molecules with the desired gene expression profile are newly generated.

apeutic molecules, GxVAEs generate molecules that are biologically meaningful in the context of specific diseases using gene expression profiles.

GxVAEs

Figure 1 shows an overview of the GxVAEs. GxVAEs mainly consists of two joint VAEs, namely ProfileVAE and MolVAE. ProfileVAE acts as a feature extractor to extract low-dimensional feature vectors from high-dimensional gene expression profiles. MolVAE consists of bidirectional GRUs that are conditioned on the features of gene expression profiles using non-canonical SMILES (i.e., variant SMILES (Li and Yamanishi 2023)) strings to generate hit-like molecules. Note that teacher forcing (Yan et al. 2023) is used to stabilize the training and accelerate the convergence of the GxVAEs during the training phase of the MolVAE.

ProfileVAE. Suppose a gene expression profile containing K genes can be denoted as $\boldsymbol{X} = [x_1, x_2, \cdots, x_K]$, where x_k is the value of the k-th gene expression. The goal of ProfileVAE is to learn the marginal likelihood of gene expression profiles during the following generative process:

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

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

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

where β denotes the weight of the KL divergence D_{KL} (Joyce 2011). In the inference phase, the latent vector $\boldsymbol{C}$ of a given gene expression profile is approximated by the reparameterization trick that uses a unit normal distribution as

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

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ denote the mean value and standard deviation in the Gaussian distribution, respectively.

MolVAE. Let $\boldsymbol{S} = [s_1, s_2, \cdots, s_T]$ be a SMILES string, where s_t is the t-th atom. Assume that $\boldsymbol{S}$ is reconstructed from the latent variable $\boldsymbol{Z}$ and condition using a random process $\boldsymbol{S} \sim p_{\boldsymbol{\varphi}}(\boldsymbol{S}|\boldsymbol{Z})$ that is parameterized by $\boldsymbol{\varphi}$. Then,

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

Note that the feature vectors $\boldsymbol{C}$ of the extracted gene expression profiles are used as the conditions for MolVAE. MolVAE aims to maximize the lower bound of the true log-marginal likelihood as follows:

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

Variant SMILES can traverse a molecular graph using starting atoms that have different positions in the graph. A molecular structure can have various non-canonical SMILES representations. Variant SMILES has been validated as sufficient for training models and is therefore used in the training phase to guarantee the diversity of the generated molecules. An example of variant SMILES is provided in Appendix A[1]. Furthermore, teacher forcing is employed in the training phase of MolVAE to stabilize the training and accelerate convergence (see more details in Appendix B). Algorithm 1 in Appendix C shows the procedures of GxVAEs. Initially, we train ProfileVAE and retain only its encoder to extract features of gene expression profiles. Subsequently, these features are concatenated with their SMILES strings and utilized as input for training MolVAE. After training, only the decoder of MolVAE is retained. Finally, we combine the encoder from ProfileVAE with the retained decoder from MolVAE for GxVAEs, facilitating the generation of hit molecules from gene expression profiles.

Experiments

Experimental Setup

Datasets. Three types of gene expression profiles were used to validate the effectiveness of the proposed GxVAEs.

- **Chemically induced profiles** were collected from the LINCS L1000 database (Duan et al. 2014). Molecules were represented by the gene expression profiles of 77 human cell lines in response to chemical perturbations such as compound additions. Here, gene expression profiles of MCF7 cell line treated with $13,755$ molecules were used.
- **Target perturbation profiles** were obtained from the LINCS database. Target proteins were represented by the gene expression profiles of 77 human cell lines in response to genetic perturbations, such as the gene knockdown and gene overexpression of target proteins. For a fair comparison during the performance evaluation, we used the same target proteins as were used in previous works having the same goal. For example, we analyzed 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 are therapeutic proteins for cancers.
- **Disease-specific profiles** were collected from the crowd extracted expression of differential signatures (CREEDS) database (Wang et al. 2016). Diseases were represented by gene expression profiles that consisted of $14,804$ genes. For example, we obtained disease-specific profiles for gastric cancer, atopic dermatitis, and Alzheimer's disease. We averaged the gene expression profiles for a single disease from multiple patients.

[1]Appendices are available at https://yamanishi.cs.i.nagoya-u.ac.jp/gxvaes/

	AvgLen	MaxLen	MinLen	MW	QED
Validation	57	79	10	429	0.69
GxVAEs	57	83	10	430	0.61

Table 1: Statistics for the dataset and molecules generated by GxVAEs. AvgLen, MaxLen, and MinLen indicate the average, maximum, and minimum lengths of the SMILES strings. MW indicates the average molecular weights.

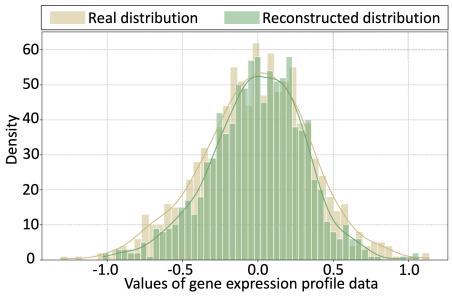

Figure 2: Reconstructed distribution of gene expression profiles of molecule "$C_{16}H_{18}N_2O_2$" by ProfileVAE.

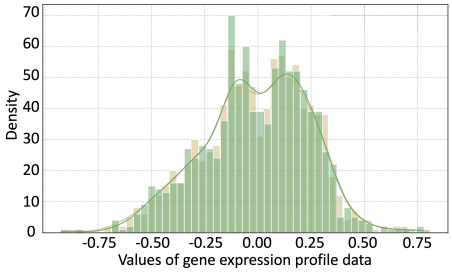

Figure 3: Average distribution of all gene expression profiles exposed in the MCF7 cell by ProfileVAE.

Implementation details. In ProfileVAE, the encoder and decoder contained three feedforward layers of size of 512, 256, and 128. The learning rate and dropout probability were set to $1e{-}4$ and 0.2, respectively. MolVAE had an embedding size of 128 and three hidden layers of size 256. The

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

Table 2: Comparative evaluation of TRIOMPHE baseline and GxVAEs. The values in gray cells indicate maximum values.

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

Table 3: Comparison of the Tanimoto coefficients for the baselines and the proposed GxVAEs.

learning rate, dropout probability, and temperature β were set to $5e{-}4$, 0.1, and 1.0. The maximum length of generated SMILES strings was 100. The dimensionality and batch size of the latent vectors for the two VAEs were 64. ProfileVAE and MolVAE were trained for 2000 and 200 epochs, respectively. All the experiments were run on GPUs with CUDA[2].

Evaluation Measures

The following measures were used to evaluate the quality of the generated hit-like molecules. **Validity** refers to the ratio of chemically valid molecules generated, which can be verified in practice using the RDKit tool (Landrum 2013). Low validity indicates inadequate learning using the model. **Uniqueness** refers to the fraction of non-repeated molecules among the valid molecules generated. Low uniqueness indicates a mode collapse problem. **Novelty** is defined as the ratio of unique molecules that do not appear in the training set. Low novelty indicates overfitting. A **quantitative estimate of drug-likeness (QED)** quantifies the likelihood that a molecule is a drug (Appendix D). The **Tanimoto coefficient** was used to calculate the similarity of the generated molecules to the target proteins, which was calculated using the ECFP4 fingerprint (Rogers and Hahn 2010). Molecules that are structurally similar to a ligand have a mode of action similar to that of the ligand. The details for the calculation of the Tanimoto coefficient are given in Appendix E.

Validity, uniqueness, and novelty were the statistics for the generated molecules. The QED and Tanimoto coefficients were used to examine changes in the chemical properties of the generated hit-like molecules and their quality.

Validation and Evaluation Results of ProfileVAE

To demonstrate that ProfileVAE has the ability to extract the biological features of gene expression profiles, we compared the distributions of the input gene expression profiles and the reconstructed profiles. Figure 2 shows the distribution of the gene expression profiles induced by the molecule "$C_{16}H_{18}N_2O_2$" (in yellow), as well as the distribution of its reconstructed profiles (in green). Figure 3 shows the average distribution of all gene expression profiles in the validation set (yellow) and the average distribution of the corresponding reconstructed profiles (green). The distributions of the reconstructed profiles approximate those of the corresponding validation data, which indicates that ProfileVAE extracted the features of the gene expression profiles well, and proved the effectiveness of the proposed GxVAEs.

Validation and Evaluation Results of MolVAE

Figure F.1 in Appendix presents the change curves for the reconstruction loss and ratio of valid molecules generated by ProfileVAE. We found that with an increase in the number of training epochs, the loss gradually decreased and the validity of the generated molecules gradually increased.

Figure F.2 in Appendix F shows the violin plots for the QED scores from the validation set (red) and the molecules generated by MolVAE (green). These two violin plots have similar distributions, indicating that the proposed GxVAEs can produce valid molecules while ensuring the consistency of their intrinsic molecular chemical properties. To further explore the structure of the generated molecules, Figs. F.3 and F.4 show the 12 molecules with the highest QED in the validation set and MolVAE. The molecules generated by MolVAE have structures and QED scores that are similar

[2]Code is available at: https://github.com/naruto7283/GxVAEs

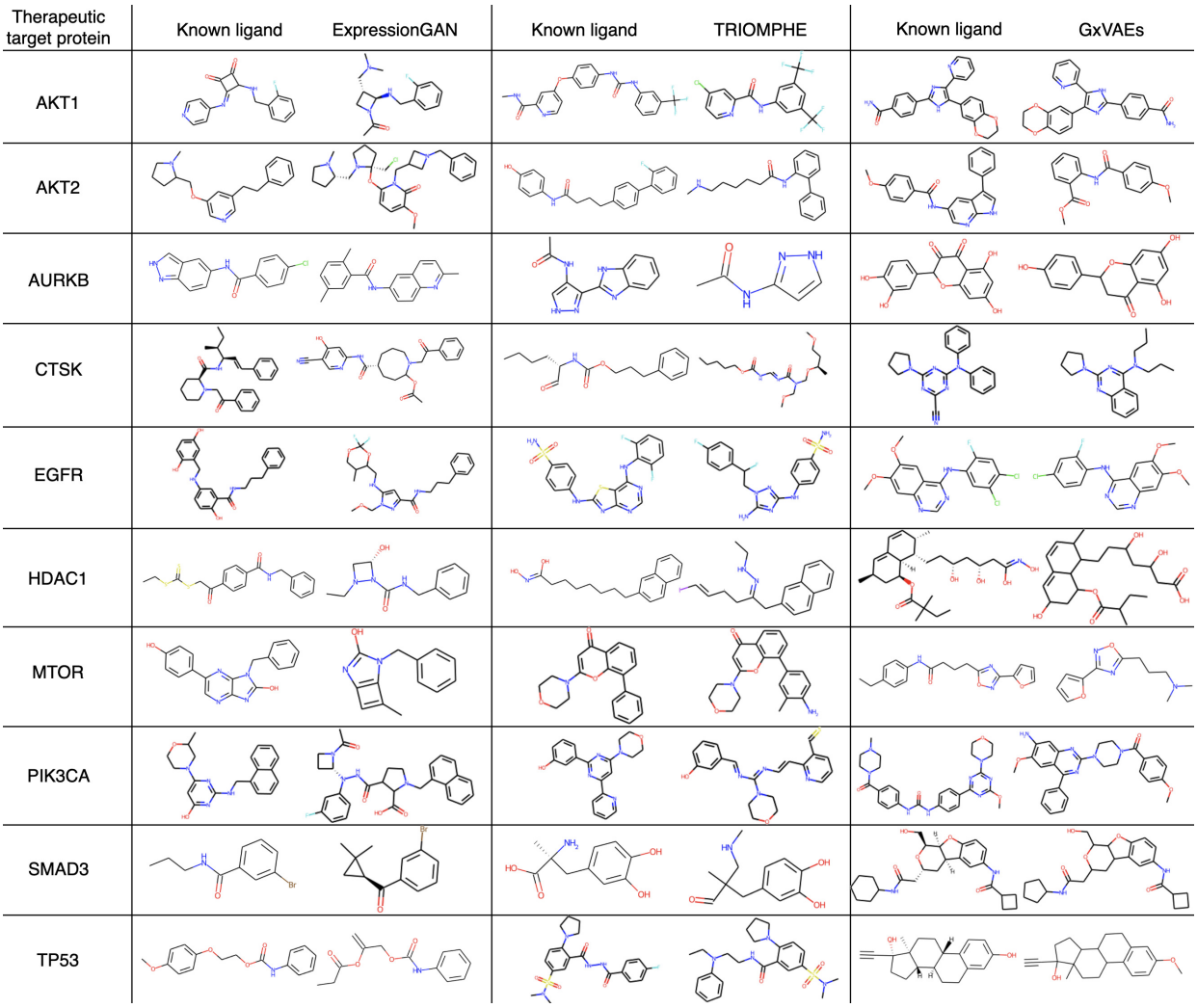

Figure 4: Comparison of hit-like molecules generated by ExpressionGAN and TRIOMPHE baselines with those by GxVAEs, focusing specifically on molecules with the highest Tanimoto coefficients for corresponding known ligands.

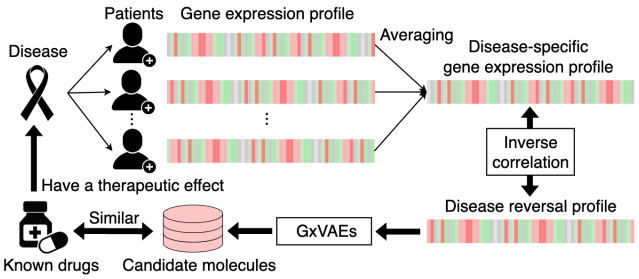

Figure 5: GxVAEs for therapeutic molecular generation.

to those in the validation set. Furthermore, Table 1 lists the statistics for the validation set and molecules generated by the proposed GxVAEs. The average, maximum, minimum lengths, and the average molecular weight of the generated SMILES strings were basically consistent with the SMILES strings in the original validation set, indicating that GxVAEs learned the data distributions of the SMILES strings well.

Evaluation Results of GxVAEs

In biochemistry, when a ligand binds to a target protein, the activity of the ligand is altered to initiate a cellular reaction. This study generated ligand candidate molecules for ten target proteins by considering the gene expression profiles of the target proteins. To generate inhibitors, we used knockdown gene expression profiles for the eight proteins AKT1, AKT2, AURKB, CTSK, EGFR, HDAC1, MTOR, and PIK3CA. To generate activators, we used overexpression gene expression profiles for the two proteins SMAD3 and TP53. Table 2 shows the ability of the proposed GxVAEs to generate candidate molecules. Note that ExpressionGAN has limited ability to generate valid molecules, thus we only compared GxVAEs to the TRIOMPHE baseline. The experimental results demonstrate that the validity of the proposed GxVAEs is three times higher than that of TRIOMPHE for generating ligand-like molecules using the

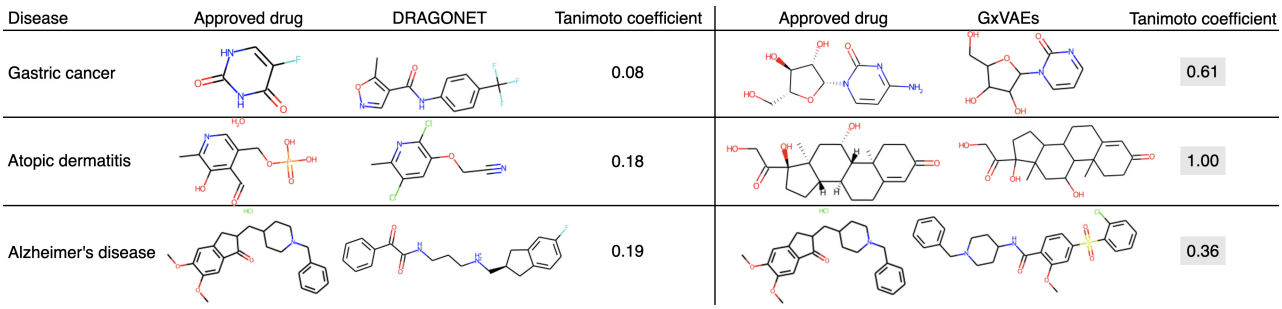

Figure 6: Therapeutic molecule generation by DRAGONET baseline and GxVAEs for three diseases.

gene expression profiles of the ten target proteins, reaching at least 78.0% (HDAC1). Moreover, the uniqueness of the GxVAEs exceeds that of TRIOMPHE except for PIK3CA ($93.5\% < 97.2\%$). Additionally, the novelty of the GxVAEs (97.7%) is close to that of TRIOMPHE. Overall, the proposed GxVAEs provided a sufficient number of candidate ligands for the ten target proteins.

Table 3 compares the Tanimoto coefficients of GxVAEs with those of the two SOTA models. The higher the Tanimoto coefficient, the easier a generated hit-like molecule binds to a target protein. The experimental results illustrate that the Tanimoto coefficients of the candidate ligands generated by GxVAEs for the ten target proteins were much higher than those of the two SOTA models. For example, the Tanimoto coefficient of GxVAEs for AKT1 was 2.7 and 2.0 times higher than those of the two baselines.

Furthermore, to intuitively compare the known ligands with the generated ligands, we show their molecular structures in Fig. 4. The chemical structures of the hit-like molecules generated by GxVAEs are all similar to the structures of the known ligands. Overall, the proposed GxVAEs show excellent performance in generating hit-like molecules from the gene expression profiles, and the biological activity of the generated molecules far exceeded the SOTA baselines.

Case Studies on Therapeutic Molecule Generation

In a disease state, there exists a complicated combination of multiple gene and protein abnormalities, and the gene expression system does not show normal behavior. As a case study, we aimed to generate therapeutic molecules from the gene expression profiles of real patients. Figure 5 shows the process undertaken by GxVAEs for therapeutic molecular generation. By averaging the gene expression profiles of multiple patients with the same disease, we obtained a disease-specific gene expression profile, that reflects the transcriptome landscape of a particular disease. The molecules that counteract the disease-specific gene expression profile are considered to have a therapeutic effect on the disease. Thus, we multiplied the disease-specific gene expression profile by -1 to obtain the disease reversal profile. The molecules that induce a gene expression pattern that is similar to the disease reversal profile are considered to have a therapeutic effect on the disease. Finally, the disease reversal profiles were input into GxVAEs to generate candidate therapeutic molecules. The generated molecules that are structurally similar to drugs approved for treating the disease are expected to have therapeutic effects on the disease. The Tanimoto coefficients, which reflect structural similarity of molecules, were used to evaluate drug-likeness relative to drugs approved for disease treatment.

We attempted to generate new therapeutic drugs for three diseases (i.e., gastric cancer, atopic dermatitis, and Alzheimer's disease). We compared the results generated by GxVAEs with those generated by DRAGONET (Yamanaka et al. 2023). For a fair comparison, we used the same datasets on patient gene expression profiles and molecule chemical structures in DRAGONET. Figure 6 shows the therapeutic molecules generated by DRAGONET and GxVAEs, as well as the Tanimoto coefficients calculated relative to known approved drugs. For example, hydrocortisone (DB00741) is a glucocorticoid that is commonly used to treat atopic dermatitis, immune disorders, and allergies. The Tanimoto coefficients between hydrocortisone and the molecules generated by GxVAEs using the disease reversal profile of patients having atopic dermatitis reached 1.0. This result indicates that GxVAEs effectively captured the structural features of the drugs approved for the treatment of atopic dermatitis. Furthermore, the molecules generated by GxVAEs for treating gastric cancer and Alzheimer's disease showed structural features that were similar to those of known approved drugs. Thus, the proposed GxVAEs can generate molecules that have higher therapeutic properties than DRAGONET.

Conclusion

We proposed GxVAEs, which consisted of two joint VAEs (i.e., ProfileVAE and MolVAE), to generate hit-like molecules from gene expression profiles. ProfileVAE extracted the features of the gene expression profiles, which were then used as conditions to guide MolVAE in producing hit-like molecules. The experimental results showed that GxVAEs outperformed the current SOTA baselines and efficiently generated hit-like molecules from gene expression profiles. Furthermore, we showed the capability of GxVAEs to create molecular structures with the potential therapeutic effects for various diseases from patients' disease profiles.

One limitation of GxVAEs is that the diversity may be influenced by the size of the latent space. If an implementation of MolVAE adopts a fixed latent vector, it could potentially constrain the diversity of newly generated molecules. Addressing this limitation will be a focus of our future research.

Acknowledgments

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].

References

Asyali, M. H.; Colak, D.; Demirkaya, O.; and Inan, M. S. 2006. Gene expression profile classification: a review. *Current Bioinformatics*, 1(1): 55–73.

Born, J.; Manica, M.; Oskooei, A.; Cadow, J.; Markert, G.; and Martínez, M. R. 2021. PaccMannRL: de novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *Iscience*, 24(4): 102269.

Chen, B.; Garmire, L.; Calvisi, D. F.; Chua, M.-S.; Kelley, R. K.; and Chen, X. 2020. Harnessing big 'omics' data and AI for drug discovery in hepatocellular carcinoma. *Nature Reviews Gastroenterology & Hepatology*, 17(4): 238–251.

Dai, H.; Dai, B.; and Song, L. 2016. Discriminative embeddings of latent variable models for structured data. In *International Conference on Machine Learning*, 2702–2711. PMLR.

De Cao, N.; and Kipf, T. 2018. MolGAN: an implicit generative model for small molecular graphs. *ArXiv preprint ArXiv:1805.11973*.

Dobson, C. M.; et al. 2004. Chemical space and biology. *Nature*, 432(7019): 824–828.

Dollar, O.; Joshi, N.; Beck, D. A.; and Pfaendtner, J. 2021. Attention-based generative models for de novo molecular design. *Chemical Science*, 12(24): 8362–8372.

Duan, Q.; Flynn, C.; Niepel, M.; Hafner, M.; Muhlich, J. L.; Fernandez, N. F.; Rouillard, A. D.; Tan, C. M.; Chen, E. Y.; Golub, T. R.; et al. 2014. LINCS canvas browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Research*, 42(W1): W449–W460.

Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; and Aspuru-Guzik, A. 2017. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. *arXiv preprint arXiv:1705.10843*.

Hertzberg, R. P.; and Pope, A. J. 2000. High-throughput screening: new technology for the 21st century. *Current Opinion in Chemical Biology*, 4(4): 445–451.

Jin, W.; Barzilay, R.; and Jaakkola, T. 2018. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, 2323–2332. PMLR.

Joyce, J. M. 2011. Kullback-leibler divergence. In *International Encyclopedia of Statistical Science*, 720–722. Springer.

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

Kang, M.; Ko, E.; and Mersha, T. B. 2022. A roadmap for multi-omics data integration using deep learning. *Briefings in Bioinformatics*, 23(1): bbab454.

Kusner, M. J.; Paige, B.; and Hernández-Lobato, J. M. 2017. Grammar variational autoencoder. In *International Conference on Machine Learning*, 1945–1954. PMLR.

Landrum, G. 2013. Rdkit documentation. *Release*, 1(1-79): 4.

Li, C.; Yamanaka, C.; Kaitoh, K.; and Yamanishi, Y. 2022. Transformer-based objective-reinforced generative adversarial network to generate desired molecules. In *IJCAI*, 3884–3890.

Li, C.; and Yamanishi, Y. 2023. SpotGAN: a reverse-transformer GAN generates scaffold-constrained molecules with property optimization. In *ECML-PKDD*, 3884–3890.

Manolopoulos, D. E.; and Fowler, P. W. 1992. Molecular graphs, point groups, and fullerenes. *Journal of Chemical Physics*, 96(10): 7603–7614.

Méndez-Lucio, O.; Baillif, B.; Clevert, D.-A.; Rouquié, D.; and Wichard, J. 2020. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature Communications*, 11(1): 10.

Morganti, S.; Tarantino, P.; Ferraro, E.; D'Amico, P.; Duso, B. A.; and Curigliano, G. 2019. Next generation sequencing (NGS): a revolutionary technology in pharmacogenomics and personalized medicine in cancer. *Translational Research and Onco-Omics Applications in the Era of Cancer Personal Genomics*, 9–30.

Mun, J.; Choi, G.; and Lim, B. 2020. A guide for bioinformaticians:'omics-based drug discovery for precision oncology. *Drug Discovery Today*, 25(11): 1897–1904.

Oliveira, A. F.; Da Silva, J. L.; and Quiles, M. G. 2022. Molecular property prediction and molecular design using a supervised grammar variational autoencoder. *Journal of Chemical Information and Modeling*, 62(4): 817–828.

Pereira, R.; Oliveira, J.; and Sousa, M. 2020. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. *Journal of Clinical Medicine*, 9(1): 132.

Rahman, A. Z.; Liu, C.; Sturm, H.; Hogan, A. M.; Davis, R.; Hu, P.; and Cardona, S. T. 2022. A machine learning model trained on a high-throughput antibacterial screen increases the hit rate of drug discovery. *PLOS Computational Biology*, 18(10): 1–22.

Rogers, D.; and Hahn, M. 2010. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5): 742–754.

Scannell, J. W.; Bosley, J.; Hickman, J. A.; Dawson, G. R.; Truebel, H.; Ferreira, G. S.; Richards, D.; and Treherne, J. M. 2022. Predictive validity in drug discovery: what it is, why it matters and how to improve it. *Nature Reviews Drug Discovery*, 21(12): 915–931.

Shaker, B.; Ahmad, S.; Lee, J.; Jung, C.; and Na, D. 2021. In silico methods and tools for drug discovery. *Computers in Biology and Medicine*, 137: 104851.

Turanli, B.; Karagoz, K.; Gulfidan, G.; Sinha, R.; Mardinoglu, A.; and Arga, K. Y. 2018. A network-based cancer drug discovery: from integrated multi-omics approaches to precision medicine. *Current Pharmaceutical Design*, 24(32): 3778–3790.

Wang, Z.; Monteiro, C. D.; Jagodnik, K. M.; Fernandez, N. F.; Gundersen, G. W.; Rouillard, A. D.; Jenkins, S. L.; Feldmann, A. S.; Hu, K. S.; McDermott, M. G.; et al. 2016. Extraction and analysis of signatures from the gene expression omnibus by the crowd. *Nature Communications*, 7(1): 12846.

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

Yamanaka, C.; Uki, S.; Kaitoh, K.; Iwata, M.; and Yamanishi, Y. 2023. De novo drug design based on patient gene expression profiles via deep learning. *Molecular Informatics*.

Yan, C.; Yang, J.; Ma, H.; Wang, S.; and Huang, J. 2023. Molecule sequence generation with rebalanced variational autoencoder loss. *Journal of Computational Biology*, 30(1): 82–94.

Zhao, J.; Feng, Q.; and Wei, W.-Q. 2022. Integration of omics and phenotypic data for precision medicine. In *Systems Medicine*, 19–35. Springer.