Gx2Mol: De Novo Generation of Hit-like Molecules from Gene Expression Profiles

Chen Li[1,2] ✉[0000−0002−8784−8148] and Yoshihiro Yamanishi[2][0000−0003−2279−8773]

[1] D3 Center, Osaka University, Ibaraki, Osaka, 567-0047, Japan
li.chen.d3c@osaka-u.ac.jp
[2] Graduate School of Informatics, Nagoya University, Nagoya, 464-8601, Japan
yamanishi@i.nagoya-u.ac.jp

Abstract. *De novo* generation of hit-like molecules is a challenging task in the drug discovery process. Most methods in previous studies learn the semantics and syntax of molecular structures by analyzing molecular graphs or simplified molecular input line entry system (SMILES) strings; however, they do not take into account the drug responses of the biological systems consisting of genes and proteins. In this study we propose a deep generative model, Gx2Mol, which utilizes **g**ene e**x**pression profiles **to** generate **mol**ecular structures with desirable phenotypes for arbitrary target proteins. In the algorithm, a variational autoencoder is employed as a feature extractor to learn the latent feature distribution of the gene expression profiles. Then, a long short-term memory is leveraged as the chemical generator to produce syntactically valid SMILES strings that satisfy the feature conditions of the gene expression profile extracted by the feature extractor. Experimental results demonstrate that Gx2Mol produces new molecules with potential bioactivities and drug-like properties. The source code is available at: https://github.com/naruto7283/Gx2Mol.

Keywords: Gene expressions · Molecular generation · Deep learning.

1 Introduction

Exploring the chemical space to discover molecules with therapeutic effects (e.g., anticancer drug production) is a time-consuming, costly, and high-risk task in the drug discovery field. Despite extensive premarket drug testing, the failure rate is still $> 90\%$ [13]. In general, drug development takes over 12 years and costs greater than \$1.3 billion [3]. After identification of therapeutic target proteins for a disease of interest, researchers search for potential drug candidate molecules that can interact with the therapeutic target proteins. This process is referred to as hit identification [29]. The high-throughput screening of large-scale chemical compound libraries with various biological assays is often performed for the hit identification, but the experimental approach is expensive.

As an alternative to hit identification, computational methods such as virtual screening [9] and *de novo* molecular generation [20] can be used to accelerate the production of drug candidates. Virtual screening attempts to explore chemical

databases containing massive volumes of molecules at minimal cost and obtain hit-like molecules through docking simulation [27]. *De novo* molecular generation attempts to generate new molecules with desired chemical properties or similar to known ligands [24]. Recently, artificial intelligence and deep learning-based generative models such as variational autoencoders (VAEs) [22] and generative adversarial networks (GANs) [18] have emerged for the *de novo* molecular generation. However, most methods in the previous studies focused on learning the syntax and semantics of molecular structures by analyzing molecular graphs or simplified molecular input line entry system (SMILES) strings.

The biological system is perturbed by drug treatment, thus, the use of biological data in addition to chemical data is desired for drug discovery. Omics data including transcriptome offer a comprehensive molecular landscape that can describe the cellular responses of human cells to drug treatment and the pathological histories of disease patients. Thus, omics data representing drug activities are important resources for current drug development. For example, the use of gene expression data in the preliminary stage of drug discovery is a promising approach [30], because it does not depend on prior knowledge of ligand structures or three-dimensional (3D) structural information of therapeutic target proteins [4]. However, omics-based drug discovery approach has severe limitations. The number of molecules with omics information is quite limited; thus, the method is applicable only to molecules for which omics data are measured. Deep learning methods using GANs [21] and VAEs [12] have been applied to generate molecules from gene expression data, but many generated compounds are chemically invalid or unrealistic, suggesting accuracy needs improvement.

In this study, we present a deep generative model, Gx2Mol, to analyze omics data and design new hit-like molecules. Specifically, a VAE is first used to extract low-dimensional features from gene expression profiles. These features then condition an LSTM-based generator [16] to produce valid SMILES strings aligned with the input profile. Gene expression features are used as conditions during LSTM training to guide the generation of molecules aligned with the target profile. The main contributions are as follows:

- **A novel idea**: unlike the previous methods on the generation of molecular chemical structures (e.g., SMILES strings and graphs), this study attempts to generate hit-like molecules from scratch using gene expression profiles.
- **A concise model**: combining simple generative models (i.e., VAE and LSTM) achieves the goal of molecular generation considering biological information.
- **Superior performance**: the experimental results demonstrate that the proposed method yields new molecules with potential bioactivities and drug-likeness properties, which can be utilized for further structure optimization.

2 Related Works

Traditional drug discovery relies on chemical intuition, medicinal chemistry, and structure-based design [1]. Chemists design molecules, build libraries, and use structural data for drug development. However, such methods are limited by high

costs and time demands. Predicting bioactivity remains difficult, as conventional approaches struggle to capture complex structure-activity relationships [13].

2.1 Graph-based Molecular Generation

Molecular graphs contain rich structural information and are often used for molecular generation [10]. Typically, a molecular graph is usually represented by an ensemble of atom vectors and bond matrices. VAEs attempt to approximate the distribution of molecular graphs to learn latent variables [6]. Generally, VAE-based models construct molecular graphs with a tree structure and employ an encoder to extract the molecular graph features and represent them as low-dimensional latent vectors. Then, the VAE decoder is employed as a molecular generator to reconstruct atoms in the tree into molecules via the latent vector representation. The design of graph-based generators is challenging; thus, GAN-based molecular generation models are rare. MolGAN [5] generates new graphs with the maximum likelihood of atoms and chemical bonds by sampling atomic features and chemical bond feature matrices. In addition, an actor-critic [19] reward network is used to calculate the property scores of the generated graphs. However, MolGAN suffers from a severe mode collapse, thereby causing its uniqueness to be less than 5%.

Flow-based molecular generative models, exemplified by MoFlow [34], initially produce bonds (edges) using a Glow-based model. Subsequently, atoms (nodes) are generated based on the established bonds through a novel graph conditional flow. Finally, these components are assembled into a chemically valid molecular graph, with posthoc validity correction. Diffusion models, such as DiGress [31], are based on a discrete diffusion process. Graphs are iteratively modified with noise through the addition or removal of edges and changes in categories.

2.2 SMILES-based Molecular Generation

De novo drug design using SMILES strings attempts to generate new molecules with desired properties. For example, GrammarVAE [14] is a SMILES-based model that is used to generate molecular structures, where a VAE is used with a grammar-based decoder that generates syntactically valid SMILES strings. This model is trained on a dataset of existing molecules and generates new molecules with high structural diversity. In addition, TransORGAN [17] is a transformer-based GAN model designed to generate diverse molecules that are similar to the source molecules. The transformer and a one-dimensional convolutional neural network are employed as the generator and discriminator, respectively, and the Monte Carlo tree search-based policy gradient reinforcement learning algorithm [28] is used to explore new molecules with desired chemical properties.

2.3 Omics Data-driven Molecular Generation

To date, most methods in previous studies generated hit-like molecules based on a learning set of ligand structures and bioactivities, where the structures are

represented by graphs or SMILES strings. Diverging from conventional approaches, omics data-driven hit-like molecular generation endeavors to leverage omics data, specifically gene expression profiles. The overarching goal is to generate hit molecules that exhibit promising biological activities against specific targets, such as proteins associated with particular diseases. To our knowledge, there are limited studies that have explored drug design directly from omics data [12,21].

Generally, omics-based methods can generate hit-like molecules without prior knowledge of ligand structures and the 3D structure of the target proteins. A conditional Wasserstein GAN combined with a gradient penalty was proposed to generate hit-like molecules from noise using gene expression profile data [21], which is referred to as ConGAN in this study. However, the validity of the generated candidate molecules is not guaranteed, thereby limiting the hit identification ability. In addition, the prediction process of transcriptional correlation between ligands and targets is unclear. TRIOMPHE [12] is a VAE-based molecular generation model using transcriptional correlation between the gene expression profile with the perturbation of a therapeutic target protein and the gene expression profile with the treatment of small molecules. The most similar molecule is selected as the source molecule, the source molecule is projected to the latent space using a VAE encoder, and a decoder is used to sample and decode the latent vectors into new molecules. However, in their work, gene expression profiles were solely employed in correlation calculations for selecting SMILES strings before inputting them into the VAE model. During the molecular generation phase, gene expression profiles were not utilized to guide the generation of hit-like molecules. Consequently, the molecules generated using TRIOMPHE exhibited low Tanimoto coefficients compared to the corresponding known ligands. DRAGONET [33] generates drug candidates from patient gene expression profiles via a transformer-based VAE, integrating disease-related molecular substructures. It demonstrated effectiveness for diseases such as gastric cancer, atopic dermatitis, and Alzheimer's by producing molecules similar to approved drugs.

Unlike previous approaches, Gx2Mol generates hit-like compounds that exhibit potential biological activity against specific target proteins or therapeutic efficacy for particular diseases, leveraging gene expression profiles. Gx2Mol first extracts biological features from gene expression data using a VAE. Subsequently, these extracted features are utilized as conditional inputs to an LSTM, guiding the generation of hit-like molecules.

3 Gx2Mol

3.1 Extraction of Biological Features

The architecture of the Gx2Mol model is illustrated in Figure 1. In phase (A), we initiate the process by training a VAE model, extracting essential biological features from gene expression profiles. The encoder network transforms the features of gene expression profiles into a low-dimensional latent space, which is subsequently reconstructed by the decoder. Post the training phase, only the encoder is utilized for subsequent downstream tasks.

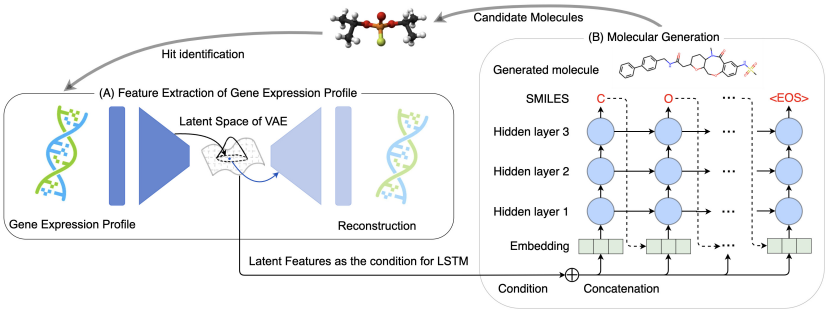

Fig. 1. Architecture of the Gx2Mol model. (A) A VAE is trained to extract the biological features of gene expression profiles. Here, a VAE encoder attempts to extract the latent feature vector of a gene expression profile, and a VAE decoder attempts to reconstruct the gene expression profile from the latent vector. (B) After the VAE training, the latent vector is utilized as a condition to an LSTM to generate SMILES strings. An extracted latent vector and a vector representation of a start token are concatenated to generate the first atom of a SMILES string. Then, the generated atom and the condition generate the next atom iteratively. This iterative process ends when the defined end token (i.e., <EOS>) is generated. Finally, all atoms are assembled into a SMILES string, which serves as a candidate molecule for hit identification in disease treatment.

Formally, let $\boldsymbol{G} = [g_1, g_2, \cdots, g_T]$ indicate the gene expression profile, where g_i represents the i-th gene with the maximum gene number of T. The VAE serves as a feature extractor in Gx2Mol, tasked with learning a latent feature distribution denoted as $p(z|\boldsymbol{G})$. The objective is to align this distribution as closely as possible to the reference distribution $p(z)$, characterized as an isotropic normal distribution. This alignment occurs through the approximation of observed gene expression profiles, while reinforcing the stochastic independence among latent variables. The utilization of the VAE in this manner facilitates the extraction of meaningful latent features from the input data, as demonstrated in Figure 1 (A). This visualization offers a concrete representation of how the VAE captures key features within the gene expression profiles. High-dimensional gene expression profile reconstruction can be modeled by the integration of the low-dimensional feature space $p(z)$ and conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{G}|z)$ parameterized by $\boldsymbol{\theta}$:

$$p_{\boldsymbol{\theta}}(\boldsymbol{G}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{G}|z)p(z)dz. \tag{1}$$

To address the intractable issue of the posterior distribution $p_{\boldsymbol{\theta}}(z|\boldsymbol{G})$, the feature extractor replaces $p_{\boldsymbol{\theta}}(z|\boldsymbol{G})$ by an approximate variational distribution $q_{\boldsymbol{\theta}'}(z|\boldsymbol{G})$. Typically, $q_{\boldsymbol{\theta}'}(z|\boldsymbol{G})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{G}|z)$ are used as the encoder and decoder of a VAE, respectively. According to the evidence lower bound [26], the loss function of the feature extractor can be formulated as

$$\mathcal{L}_F(\boldsymbol{\theta}, \boldsymbol{\theta}') = -\mathbb{E}_{z\sim q_{\boldsymbol{\theta}'}(z|\boldsymbol{G})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{G}|z)] + \beta \cdot D_{KL}(q_{\boldsymbol{\theta}'}(z|\boldsymbol{G})||p(z)), \tag{2}$$

where $\mathbb{E}[\cdot]$ and β indicate an expectation operation and the weight of the Kullback-Leibler divergence D_{KL} [11], respectively. The VAE encoder generates both a

mean (μ) and a variance (σ^2) for each point in the latent space, typically following a Gaussian distribution. For a given gene expression profile $\boldsymbol{G}$, the approximate posterior distribution can be calculated as follows:

$$q_{\boldsymbol{\theta}'}(z|\boldsymbol{G}) = \mathbf{N}\left(\mu(\boldsymbol{G}), \sigma^2(\boldsymbol{G})\right), \tag{3}$$

where $\mu(\boldsymbol{G})$ and $\sigma^2(\boldsymbol{G})$ are the mean and variance functions parameterized by the encoder. The VAE then samples a point z from this distribution. Finally, the extracted latent vector of the gene expression profiles is as follows:

$$\boldsymbol{F}_{Gx} = \text{Encoder}(\boldsymbol{G}). \tag{4}$$

3.2 Generation of Hit-like Molecules

Here, an LSTM model is used as the chemical generator to produce syntactically valid SMILES strings that satisfy the feature conditions of the gene expression profiles extracted by the feature extractor. During phase (B), we incorporate the corresponding SMILES strings as inputs for LSTM training. The extracted features from gene expression profiles are fused with each SMILES token, serving as input for the model to iteratively generate the subsequent token.

Formally, let $\boldsymbol{X}_{1:n} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$ denote a SMILES string of length n, where $\boldsymbol{x}_i$ is the i-th embedding vector of the SMILES string with the size of M. Then, $\boldsymbol{x}_i$ is concatenated with $\boldsymbol{F}_{Gx}$ as the input to the generator. The generator iteratively generates a character of the SMILES string at the current time step based on the previous time step. Let $\boldsymbol{Y}_{1:n} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n]$ indicate the predicted SMILES string for $\boldsymbol{X}_{1:n}$. According to the negative log likelihood, the loss function of the generator can be calculated as follows:

$$\mathcal{L}_G(\boldsymbol{X}_{1:n}, \boldsymbol{Y}_{1:n}) = -\sum_{i=1}^{n} \log p(\boldsymbol{y}_i|\boldsymbol{X}_{1:i-1}; \boldsymbol{\phi}), \tag{5}$$

where $\boldsymbol{\phi}$ is the parameters of the chemical generator.

During the generation phase, the input to the VAE encoder exclusively comprises gene expression profiles for feature extraction. The resulting extracted features are subsequently employed to steer the process of generating hit-like molecules. Algorithm 1 summarizes the procedure of the Gx2Mol model. Here, sets of the gene expression profiles and SMILES strings are first used to train the feature extractor and chemical generator. In the training phase, the features of gene expression profiles are learned from a VAE-based feature extractor. The extracted features are used as conditions of the LSTM-based molecular generator. In the testing phase, the gene expression profile is employed to generate new hit-like molecules.

4 Experiments

Datasets. In this study, we used chemically induced gene expression profiles as training data to train the Gx2Mol model. In addition, we analyzed target protein-perturbed expression profiles with eight knockdown genes and two overexpressed

Algorithm 1 Procedure for the Gx2Mol model

Data: Gene expression profiles $\boldsymbol{G}$ and SMILES strings $\boldsymbol{X}_{1:n}$
Initialization: the feature extractor F_θ, the molecule generator G_ϕ
// Train the feature extractor.
for $i = 1 \rightarrow f\_epochs$ **do**
    Update F_θ using $\boldsymbol{G}$ according to the loss function of Eq. (2).
end for
// Train the molecule generator.
for $i = 1 \rightarrow g\_epochs$ **do**
    Update G_ϕ using $\boldsymbol{X}_{1:n}$ according to the loss function of Eq. (5).
end for
// Generate hit-like molecules from scratch.
Extract the features $\boldsymbol{F}_{Gx}$ using $\boldsymbol{G}$ according to Eq. (4).
Generate the corresponding SMILES representation from $\boldsymbol{F}_{Gx}$.
// Test the generation task.
Calculate the Tanimoto coefficient using known ligands.
Select the molecule with the maximum Tanimoto coefficient score as the candidate molecule.

genes to generate hit-like molecules, and disease reversal gene expression profiles as a case study to generate therapeutic molecules.

- **Chemically-induced gene expression profiles** were collected from the Library of Integrated Network-based Cellular Signatures (LINCS) database [7]. LINCS database stores the gene expression profiles with a dimension of 978 for 77 human cultured cell lines exposed to various molecules. We analyzed the gene expression profiles of the MCF7 cell line treated with 13,755 molecules whose SMILES string lengths were less than 80 at a concentration of 10 µM.
- **Target protein-perturbed gene expression profiles** were collected from the LINCS database. We analyzed the RAC-alpha serine / threonine-protein kinase (AKT1), RAC-beta 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 homologue 3 (SMAD3), and tumor protein p53 (TP53), which have been verified to be useful therapeutic target proteins against cancers. The gene expression profiles for the first eight proteins were obtained from gene knockdown profiles of the MCF7 cell line, while those for the latter two proteins were obtained from gene overexpression profiles. When multiple profiles were measured under different experimental conditions for a single protein, we averaged the multiple profiles of the same target protein to create target protein-specific profiles.
- **Disease-specific gene expression profiles** were obtained from the crowd extracted expression of differential signatures (CREEDS) database [32], which contains the expression profiles of 14,804 genes for 79 diseases. The disease-specific gene expression profiles were acquired by averaging the gene expression

profiles from multiple patients with the same disease. Here, we extracted the most relevant 884 genes for gastric cancer, atopic dermatitis, and Alzheimer's disease from the disease-specific gene expression for model validation, and we created the disease reversal profiles by multiplying the disease-specific gene expression by -1. Note that the disease reversal profiles of a disease are considered to be associated with a therapeutic effect on that disease.

Hyperparameters. For the feature extractor, the encoder of the VAE included three feedforward layers with dimensions of 512, 256, and 128. The latent vector dimension was set to 64. Note that the dimensions of the decoder were the opposite dimensions of the encoder, i.e., 128, 256, and 512. The dropout probability and learning rate were set to 0.2 and 1e−4, respectively. The training of gene expression profiles was conducted with a batch size set at 64. For the generator, the embedding size was set to 128. The LSTM model contained three hidden layers with dimensions of 256. The dropout probability and learning rate were set to 0.1 and 5e−4, respectively. The maximum length of the generated SMILES strings was fixed to 100. The batch size for training LSTM was set to 64. In addition, the feature extractor and generator used the Adam optimizer, and the number of training epochs for the feature extractor and generator was set to 2000 and 300, respectively. All experiments were conducted on GPUs using CUDA.
Dataset splitting and model selection. The dataset was partitioned into distinct sets for training (80%), validation (10%), and testing (10%) to ensure a robust evaluation of our Gx2Mol model. This division allows for effective model training on the training set, tuning of hyperparameters based on the validation set, and unbiased assessment of model performance on the test set. The selection of the optimal model was determined by monitoring the convergence of the total loss function of Gx2Mol during training. Convergence of the loss function indicates stability and optimal performance. This approach ensures the selection of a well-performing model based on its ability to minimize the defined loss and generalize effectively to unseen data.

4.1 Evaluation Measures

In this study, two essential chemical properties (quantitative estimate of drug-likeness (QED) [2] and synthesizability (SA) [8]) and the Tanimoto coefficient [25] were employed to assess hit-like molecules generated by the Gx2Mol.

– **QED** can be calculated by assigning different weights to eight molecular descriptors (i.e., 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). where d_i and W_i represent the desirability function and weight of the i-th descriptor, respectively. Typically, the weights of the eight molecular descriptors were obtained through chemical experiments. In practice, the QED score was calculated by a function in the RDKit tool. The larger the QED score, the more drug-like the molecule.

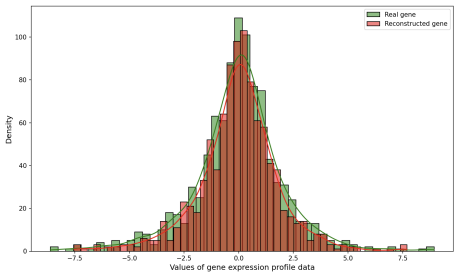

Fig. 2. Distribution of fold change values in the gene expression profile of the molecule "C17H25ClN2O3" exposed in the MCF7 cell. The original gene expression profile of "C17H25ClN2O3" (green) and the reconstructed gene expression profiles (red) have similar distributions.

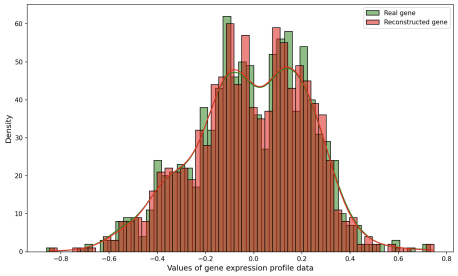

Fig. 3. Distribution of fold change values in the average gene expression profile of all molecules exposed in the MCF7 cell. The original gene expression profiles of the training set (green) and the reconstructed gene expression profiles (red) have similar distributions.

- **Synthesizability (SA)** is assessed through the SA score, denoted as $\text{SA} = r_s - \sum_{i=1}^{5} p_i$. Here, r_s signifies the "synthetic knowledge," representing the ratio of contributions from all fragments to the total number of fragments in the molecule. In this study, r_s is computed from experimental results [8]. Each p_i ($i \in \{1, \cdots, 5\}$) corresponds to the ring complexity, stereo complexity, macrocycle penalty, size penalty, and bridge penalty, computed using the RDKit tool [15]. A higher SA score indicates greater ease of synthesizing the molecule.
- **Tanimoto coefficient**, which is calculated from the ECFP4 fingerprint [23] with a dimension of 2048. In practice, the ECFP4 and Tanimoto coefficients were calculated using the "GetMorganfingerprintAsBitVect" and "BulkTanimotoSimilarity" functions of the RDKit tool.

4.2 Gx2Mol Training

We evaluated the effectiveness of the VAE model in extracting the biological features from gene expression profiles and the capability of the LSTM model to generate new molecules experimentally.

Figure 2 shows a comparison of the distribution of fold change values in the gene expression profile of a molecule between the training set and the reconstructed set. Figure 3 shows a comparison of the distribution of fold change values in the average gene expression profile of all reconstructed molecules between the original set and the reconstructed set. Note that Figure 2 shows the distribution of a gene expression profile of the molecule "C17H25ClN2O3" exposed in the MCF7 cell, whose SMILES representation is denoted as "CCC1=CC(=C(C(=C1O)C(=O)NC[C@@ H]2CCCN2CC)OC)Cl." The distribution of the original gene expression profiles was similar to that of the reconstructed gene expression profiles acquired using the Gx2Mol. In other words, the VAE utilized in the Gx2Mol captures the biological features of the gene expression profiles and successfully reconstructs them into the original distribution.

Table 1. Assessment of QED and SA scores for the top-k generated molecules.

Chemical Property	Data Source	Top-1	Top-10	Top-100	Top-1000
Drug-likeness (QED)	Compounds in training dataset	0.94	0.92	**0.85**	0.64
	Compounds generated by Gx2Mol	**0.95**	**0.93**	0.84	**0.65**
Synthesizability (SA)	Compounds in training dataset	1.00	0.94	0.85	0.47
	Compounds generated by Gx2Mol	**1.00**	**0.99**	**0.88**	**0.48**

Figure A.1 in the appendix [1] shows the training loss and the ratio of the generated valid molecules of the LSTM in the Gx2Mol. The loss decreased smoothly over the 300 training epochs and finally converges under 0.1. In contrast, the validity of the molecules generated by the conditional LSTM model gradually increased as training proceeds, with the final validity ratio converging at approximately 90%. Overall, the results indicate that the conditional LSTM utilized in the Gx2Mol can generate valid molecules effectively.

To further explore the ability of the Gx2Mol to generate molecules, we also compared the distribution of the QED scores of the molecules generated by Gx2Mol with molecules in the training data. Figure A.2 in the appendix shows that the generated molecules and the original molecules have similar QED distributions. The average QED scores of molecules in the training dataset and molecules generated by Gx2Mol were 0.60 and 0.61, respectively. The violin plots of the QED scores indicate that the Gx2Mol did not change the potential chemical property characteristics of the training data during the generation process, which demonstrates the LSTM's ability to generate molecules effectively.

Figures A.3 and A.4 in the appendix show the top-12 molecular structures with their QED scores for molecules in the training dataset and molecules generated by the Gx2Mol, respectively. It seems that all of the molecules are chemically valid and exhibit high QED scores.

Furthermore, we evaluated the QED scores for the top-k generated molecules using the Gx2Mol model. The results are presented in Table 1. The molecules generated by Gx2Mol exhibited QED scores that were higher yet comparable to those of the training data. These findings demonstrate that the Gx2Mol model generated molecules while preserving the QED properties.

Similarly, we present the top-12 molecular structures along with their synthesizability (SA) scores for molecules in the training dataset and those generated by Gx2Mol in Figures A.5 and A.6, respectively. The generated molecular structures indicate that our proposed Gx2Mol can produce valid molecules that

[1] Additional appendices are available at: `https://yamanishi.cs.i.nagoya-u.ac.jp/gx2mol/`

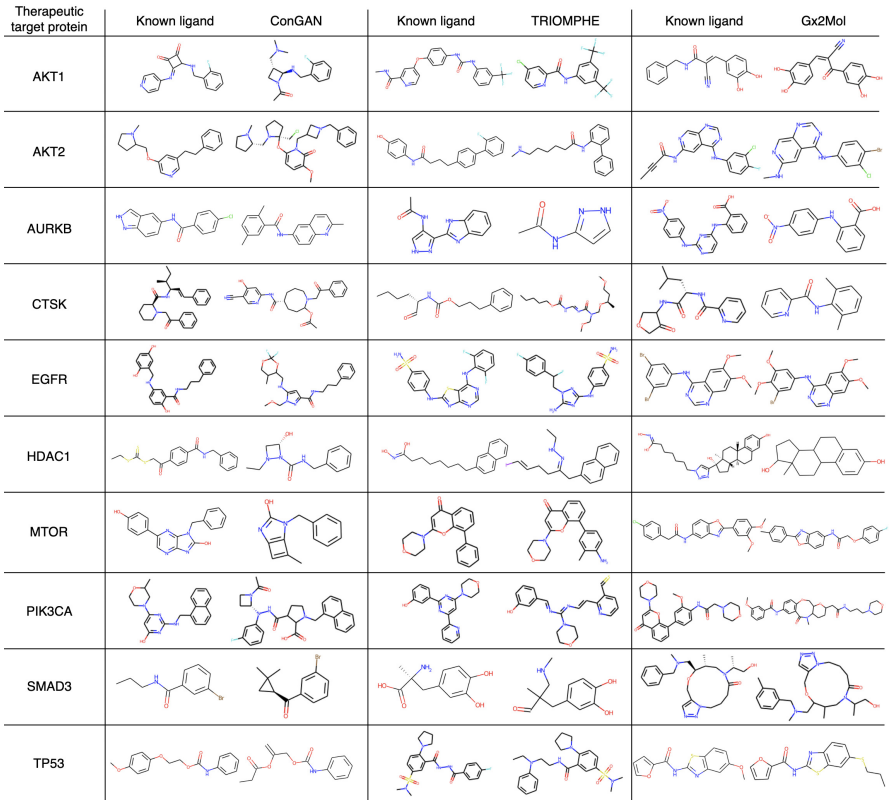

Fig. 4. Comparison of newly generated molecules from the baseline models and Gx2Mol with known ligands for each therapeutic target protein.

are easy to synthesize. Moreover, the SA scores for the top-k generated molecules in Table 1 show that Gx2Mol effectively generated molecules with high SA scores.

4.3 Gx2Mol Generation

Generally, the gene expression profiles of knockdown and overexpression of target proteins correlate with the gene expression profiles of inhibitors and activators, respectively [12]. To generate molecules as candidates for inhibitory and activatory ligands of target proteins, the gene expression profiles of the eight knockdown and two overexpressed target proteins were considered in this study. The former includes AKT1, AKT2, AURKB, CTSK, EGFR, HDAC1, MTOR, and PIK3CA. The latter includes SMAD3 and TP53.

We conducted experiments on the newly generated molecules by comparing their molecular structures with those of the known ligands (inhibitors and activators). If the newly generated molecules are meaningful, the newly generated molecules should be structurally similar to known ligands of each target protein to some extent. To ensure a fair comparison with the TRIOMPHE baseline, the

Table 2. Comparison of structural similarity scores of new molecules with known ligands for each target protein between baselines and Gx2Mol.

Therapeutic target protein	ConGAN	TRIOMPHE	Gx2Mol
AKT1	0.32	0.42	**0.53**
AKT2	0.29	0.35	**0.53**
AURKB	0.36	0.34	**0.67**
CTSK	0.31	0.29	**0.34**
EGFR	0.30	0.31	**0.72**
HDAC1	0.34	0.30	**0.42**
MTOR	0.39	**0.69**	0.46
PIK3CA	0.26	**0.32**	0.30
SMAD3	0.44	0.48	**0.85**
TP53	0.46	0.53	**0.55**

⋆ The values in bold in gray cells are the maximum values.

default sampling number for each gene expression profile of the target protein was set to 1000, consistent with the setting used in TRIOMPHE. Subsequently, we only retained the valid molecules from the 1000 generated samples to calculate structural similarity using Tanimoto coefficients. The results are presented in Table 2. ConGAN [21] and TRIOMPHE [12] are the two state-of-the-art (STOA) baselines that are related to the Gx2Mol. For the former eight knockdown target proteins, six of the calculated Tanimoto coefficients for the molecules generated by the Gx2Mol with inhibitory ligands (i.e., AKT1, AKT2, AURKB, CTSK, EGFR, and HDAC1) outperformed the baselines. For MTOR and PIK3CCA, the Tanimoto coefficients performed second only to TRIOMPHE. In addition, for both 2SMAD3 and TP53, i.e., the target proteins with gene overexpression perturbations, the Tanimoto coefficients of the generated molecules by the Gx2Mol were higher than those obtained by the baseline methods.

Furthermore, we analyzed the diversity metrics of the newly generated molecules. The diversity was computed based on molecular fingerprints generated using the Morgan algorithm (radius = 2, 2048 bits) as implemented in RDKit. The results are summarized in Table B.1. Notably, the maximum diversity values for all ten target proteins reach 1.0, while the average diversity values are consistently high (above 0.82) with low standard deviations, indicating a broad structural variety across the generated molecules.

Figure 4 shows the molecules generated by the baseline and Gx2Mol models and known ligands for each therapeutic target protein, where the newly generated molecular structures with the highest Tanimoto coefficients to the corresponding known ligands are shown. For the 10 target proteins, all generated molecules were structurally similar to the known ligands, compared with the baseline models. In summary, the Gx2Mol exhibited superior performance in terms of generating hit-like molecules from gene expression profiles via deep learning, and the proposed model outperformed the current SOTA baselines in most metrics.

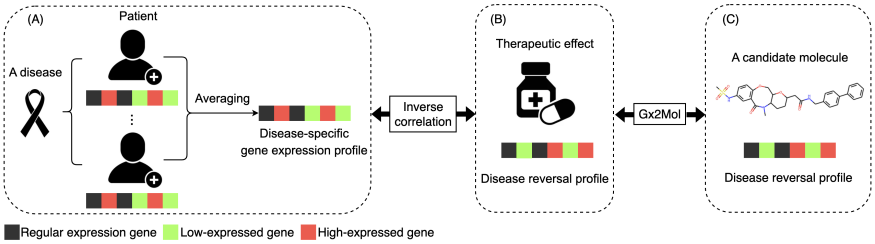

Fig. 5. Data processing of gene expression profiles for therapeutic molecular generation.

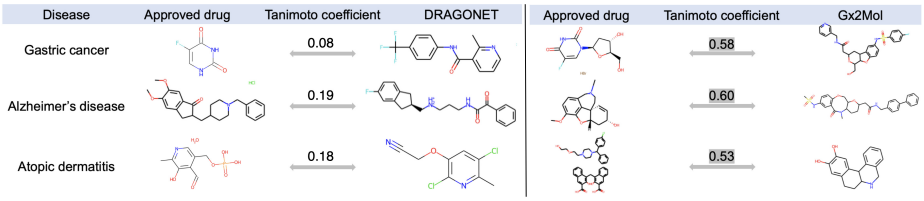

Fig. 6. Comparison of newly generated therapeutic molecules with approved drugs for each disease.

4.4 Case Studies

Generally, gene expression profiles are altered in a patient with a disease state. A molecule that counteracts the disease state is considered to have therapeutic effects on the disease. As a case study, we generated molecules with therapeutic effects on a disease by considering disease-specific gene expression profiles.

Figure 5 illustrates the data processing of a gene expression profile for the generation of molecules with therapeutic effects on a disease. First, as shown in Figure 5 (A), a disease-specific gene expression profile is constructed by averaging the gene expression profiles of patients with a certain disease. Then, a gene expression profile that is inversely correlated with the disease-specific gene expression profile is constructed and defined as the disease reversal profile, as shown in Figure 5 (B). Finally, the disease reversal profile is used as an input to the Gx2Mol to generate molecules with therapeutic effects (Figure 5 (C)). The disease-specific gene expression profiles were obtained from the CREEDS database for patients with three diseases, i.e., gastric cancer, atopic dermatitis, and Alzheimer's disease.

We examined the validity of the newly generated molecules by comparing the newly generated molecular structures with those of the approved drugs. If the newly generated molecules are meaningful, the newly generated molecules should be structurally similar to the approved drugs of each disease to some extent. We calculated the structural similarity using Tanimoto coefficients. Figure 6 illustrates the Tanimoto coefficients between approved drugs and newly generated molecules, comparing the results obtained from the SOTA DRAGONET [33] and our proposed Gx2Mol model, for each of the three diseases. Our proposed

Gx2Mol model surpassed the SOTA DRAGONET in the therapeutic molecule generation for three diseases. Gx2Mol exhibited improved Tanimoto coefficients to approved drugs, reaching 0.58, 0.60, and 0.53 for gastric cancer, Alzheimer's disease, and atopic dermatitis. Additionally, fluorouracil (D04197) can be used in the treatment of liver metastases from gastrointestinal adenocarcinomas and also in the palliative treatment of liver and gastrointestinal cancers. When using the disease reversal profile of gastric cancer patients, the Tanimoto coefficient between the molecule generated by the Gx2Mol and fluorouracil was the largest. The Tanimoto coefficient of the Gx2Mol model-generated molecule with floxuridine was maximum using the disease reversal profile of gastric cancer patients. These results suggest that the generated molecules effectively capture the structural features of approved anti-gastric cancer drugs. In addition, the molecules generated for the other two diseases demonstrate structural features that are similar to those of the approved drugs. As a result, the molecules generated using the Gx2Mol have potential drug-like properties.

5 Conclusion

This study introduced the Gx2Mol model, designing to generate potential chemical structures of hit-like molecules from gene expression profiles using deep learning techniques. In the training phase, the Gx2Mol model first employed a VAE for feature extraction from high-dimensional gene expression profiles, and then the low-dimensional extracted features guided the generation of syntactically valid SMILES strings. In the generation phase, the VAE encoder served as the sole feature extractor, seamlessly combined with the generator to facilitate the generation of hit-like molecules. The results demonstrated the effectiveness of Gx2Mol in generating hit-like molecules from gene expression profiles. Additionally, a case study illustrates the model's ability to generate potential chemical structures for therapeutic drugs related to gastric cancer, stress dermatitis, and Alzheimer's disease using patients' disease reversal profiles.

This study has a primary limitation. Since LSTMs are frequently employed in auto-regressive generation tasks, wherein the token at the next time step is generated based on the token at the current time step, there exists a potential constraint on the diversity of generated molecules when using LSTMs as generators. In future research, we aim to explore strategies to enhance the diversity of molecular generation within the Gx2Mol. Furthermore, the envisaged application of the Gx2Mol model involves integration into practical AI systems to assist chemists in generating diverse drug candidate hit-like molecules tailored for various diseases. This integration is anticipated to leverage the strengths of the Gx2Mol model and contribute to the advancement of drug discovery processes.

Acknowledgments This research was supported by the International Research Fellow of the Japan Society for the Promotion of Science (Postdoctoral Fellowships for Research in Japan [Standard]), AMED under Grant Number JP23nk0101111 and JSPS KAKENHI [grant numbers 20H05797, 21H04915].

References

1. Akaji, K., Konno, H., Mitsui, H., Teruya, K., Shimamoto, Y., Hattori, Y., Ozaki, T., Kusunoki, M., Sanjoh, A.: Structure-based design, synthesis, and evaluation of peptide-mimetic SARS 3CL protease inhibitors. Journal of medicinal chemistry **54**(23), 7962–7973 (2011)
2. Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., Hopkins, A.L.: Quantifying the chemical beauty of drugs. Nature chemistry **4**(2), 90–98 (2012)
3. Bongini, P., Bianchini, M., Scarselli, F.: Molecular generative graph neural networks for drug discovery. Neurocomputing **450**, 242–252 (2021)
4. Bung, N., Krishnan, S.R., Roy, A.: An in silico explainable multiparameter optimization approach for de novo drug design against proteins from the central nervous system. Journal of Chemical Information and Modeling **62**(11), 2685–2695 (2022)
5. De Cao, N., Kipf, T.: MolGAN: An implicit generative model for small molecular graphs. arxiv 2018. arXiv preprint arXiv:1805.11973 (2019)
6. Du, Y., Guo, X., Shehu, A., Zhao, L.: Interpretable molecular graph generation via monotonic constraints. Proceedings of the 2022 SIAM International Conference on Data Mining (SDM) pp. 73–81 (2022)
7. 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.: LINCS canvas browser: interactive web app to query, browse and interrogate lincs l1000 gene expression signatures. Nucleic acids research **42**(W1), W449–W460 (2014)
8. Ertl, P., Schuffenhauer, A.: Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of cheminformatics **1**(1), 1–11 (2009)
9. Gimeno, A., Ojeda-Montes, M.J., Tomás-Hernández, S., Cereto-Massagué, A., Beltrán-Debón, R., Mulero, M., Pujadas, G., Garcia-Vallvé, S.: The light and dark sides of virtual screening: what is there to know? International journal of molecular sciences **20**(6), 1375 (2019)
10. Jin, W., Barzilay, R., Jaakkola, T.: Junction tree variational autoencoder for molecular graph generation. International conference on machine learning pp. 2323–2332 (2018)
11. Joyce, J.M.: Kullback-leibler divergence. International encyclopedia of statistical science pp. 720–722 (2011)
12. Kaitoh, K., Yamanishi, Y.: TRIOMPHE: Transcriptome-based inference and generation of molecules with desired phenotypes by machine learning. Journal of Chemical Information and Modeling **61**(9), 4303–4320 (2021)
13. Kale, B., Clyde, A., Sun, M., Ramanathan, A., Stevens, R., Papka, M.E.: ChemoGraph: Interactive visual exploration of the chemical space. Computer Graphics Forum **42**, 13–24 (2023)
14. Kusner, M.J., Paige, B., Hernández-Lobato, J.M.: Grammar variational autoencoder. International conference on machine learning pp. 1945–1954 (2017)
15. Landrum, G.: Rdkit documentation. Release **1**(1-79), 4 (2013)
16. Li, C., He, M., Qaosar, M., Ahmed, S., Morimoto, Y.: Capturing temporal dynamics of users' preferences from purchase history big data for recommendation system. 2018 IEEE International Conference on Big Data (Big Data) pp. 5372–5374 (2018)
17. Li, C., Yamanaka, C., Kaitoh, K., Yamanishi, Y.: Transformer-based objective-reinforced generative adversarial network to generate desired molecules. IJCAI pp. 3884–3890 (2022)

18. Li, C., Yamanishi, Y.: SpotGAN: A reverse-transformer GAN generates scaffold-constrained molecules with property optimization. Joint European Conference on Machine Learning and Knowledge Discovery in Databases pp. 323–338 (2023)
19. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015)
20. Lin, X., Li, X., Lin, X.: A review on applications of computational methods in drug screening and design. Molecules **25**(6), 1375 (2020)
21. Méndez-Lucio, O., Baillif, B., Clevert, D.A., Rouquié, D., Wichard, J.: De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nature communications **11**(1), 10 (2020)
22. Oliveira, A.F., Da Silva, J.L., Quiles, M.G.: Molecular property prediction and molecular design using a supervised grammar variational autoencoder. Journal of Chemical Information and Modeling **62**(4), 817–828 (2022)
23. Ortiz, A., Gorriz, J.M., Ramírez, J., Salas-Gonzalez, D., Initiative, A.D.N., et al.: Improving MRI segmentation with probabilistic GHSOM and multiobjective optimization. Neurocomputing **114**, 118–131 (2013)
24. Payne, C., Awalt, J.K., May, L.T., Tyndall, J.D., Jörg, M., Vernall, A.J.: Bifunctional tools to study adenosine receptors. Topics in Medicinal Chemistry pp. 1–43 (2022)
25. Rácz, A., Bajusz, D., Héberger, K.: Life beyond the tanimoto coefficient: similarity measures for interaction fingerprints. Journal of cheminformatics **10**(1), 1–12 (2018)
26. Ramapuram, J., Gregorova, M., Kalousis, A.: Lifelong generative modeling. Neurocomputing **404**, 381–400 (2020)
27. Shen, J., Jiang, J., Kuang, G., Tan, C., Liu, G., Huang, J., Tang, Y.: Discovery and structure–activity analysis of selective estrogen receptor modulators via similarity-based virtual screening. European journal of medicinal chemistry **54**, 188–196 (2012)
28. Silver, D., Tesauro, G.: Monte-carlo simulation balancing. Proceedings of the 26th Annual International Conference on Machine Learning pp. 945–952 (2009)
29. Stecula, A., Hussain, M.S., Viola, R.E.: Discovery of novel inhibitors of a critical brain enzyme using a homology model and a deep convolutional neural network. Journal of Medicinal Chemistry **63**(16), 8867–8875 (2020)
30. Thomas, C.E., Will, Y.: The impact of assay technology as applied to safety assessment in reducing compound attrition in drug discovery. Expert Opinion on Drug Discovery **7**(2), 109–122 (2012)
31. Vignac, C., Krawczuk, I., Siraudin, A., Wang, B., Cevher, V., Frossard, P.: DiGress: Discrete denoising diffusion for graph generation. Proceedings of the 11th International Conference on Learning Representations (2023)
32. 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.: Extraction and analysis of signatures from the gene expression omnibus by the crowd. Nature communications **7**(1), 12846 (2016)
33. Yamanaka, C., Uki, S., Kaitoh, K., Iwata, M., Yamanishi, Y.: De novo drug design based on patient gene expression profiles via deep learning. Molecular Informatics **42**(8-9), 2300064 (2023)
34. Zang, C., Wang, F.: MoFlow: an invertible flow model for generating molecular graphs. Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining pp. 617–626 (2020)