

De Novo Generation of Chemical Structures of Inhibitor and Activator Candidates for Therapeutic Target Proteins by a Transformer-Based Variational Autoencoder and Bayesian Optimization

Yuki Matsukiyo, Chikashige Yamanaka, and Yoshihiro Yamanishi*



Cite This: <https://doi.org/10.1021/acs.jcim.3c00824>



Read Online

ACCESS |



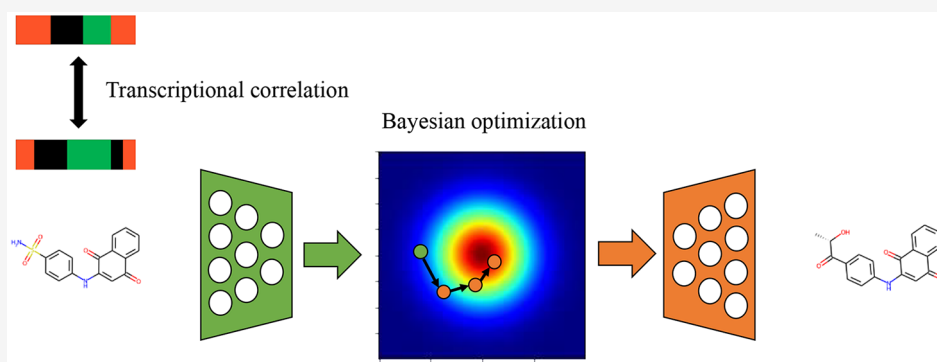
Metrics & More



Article Recommendations



Supporting Information



ABSTRACT: Deep generative models for molecular generation have been gaining much attention as structure generators to accelerate drug discovery. However, most previously developed methods are chemistry-centric approaches, and comprehensive biological responses in the cell have not been taken into account. In this study, we propose a novel computational method, TRIOMPHE-BOA (transcriptome-based inference and generation of molecules with desired phenotypes using the Bayesian optimization algorithm), to generate new chemical structures of inhibitor or activator candidates for therapeutic target proteins by integrating chemically and genetically perturbed transcriptome profiles. In the algorithm, the substructures of multiple molecules that were selected based on the transcriptome analysis are fused in the design of new chemical structures by exploring the latent space of a Transformer-based variational autoencoder using Bayesian optimization. Our results demonstrate the usefulness of the proposed method in terms of having high reproducibility of existing ligands for 10 therapeutic target proteins when compared with previous methods. Moreover, this method can be applied to proteins without detailed 3D structures or known ligands and is expected to become a powerful tool for more efficient hit identification.

INTRODUCTION

The development of new drugs is extremely costly and time consuming. It takes ≥ 10 years from the identification of a therapeutic target to its approval through clinical trials, costing an average of \$2.6 billion.^{1,2} Identifying a molecule with the desired bioactivity in the hit identification process is extremely difficult, as there are theoretically $\geq 10^{60}$ candidate compounds.³ Although large-scale compound screening methods such as high-throughput screening^{4–6} and DNA-encoded libraries^{7–10} have been used to efficiently identify hit compounds, it is impossible to experimentally test all candidate compounds due to their vast number.

In addressing this issue, deep generative models (i.e., structure generators) for *de novo* drug design strategies have been actively developed. There have been many previous studies that use the variational autoencoder (VAE)^{11–17} and generative adversarial network (GAN).^{18–22} However, most of

the previously developed structure generators are chemistry-centric approaches; they focused on generating chemically valid structures, and few considered the comprehensive biological responses caused by interactions between the drug candidate molecules and target proteins. The goal of most previous methods was to optimize properties that could be calculated directly from chemical structures (e.g., the quantitative estimate of drug-likeness²³ and the synthetic accessibility score²⁴) or to improve bioactivity by using

Special Issue: Machine Learning in Bio-cheminformatics

Received: May 30, 2023



ACS Publications

© XXXX American Chemical Society

A

<https://doi.org/10.1021/acs.jcim.3c00824>
J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

quantitative structure–activity relationship (QSAR) models in the molecular generation process. However, QSAR-based generators do not work when information about known ligands is unavailable or insufficient.

With the development of molecular biological analysis methods^{25–28} and biomedical big data,^{29,30} the application of omics data to drug discovery has been actively attempted.^{31,32} In particular, transcriptome profiles, which contain comprehensive biological response information, have been proven to be useful in drug discovery.^{33,34} In this context, structure generators that generate molecules from transcriptome profiles have been developed for hit identification.^{35,36} The introduction of transcriptome profiles into the structure generation algorithm made it possible to generate molecules that are likely to be active against a therapeutic target protein by considering the comprehensive biological responses in the cell, which was not possible with the chemistry-centric approaches. A GAN-based structure generator³⁵ and a VAE-based structure generator³⁶ were proposed to generate new chemical structures of molecules that are likely to be active against a therapeutic target protein based on transcriptome profiles obtained by gene knockdown or overexpression of the therapeutic target proteins. However, these previous studies^{35,36} had some problems in terms of the reproducibility of known ligand structures, and there is much room for improving structure optimization with transcriptional constraints.

In this study, we propose a novel computational method, TRIOMPHE-BOA (transcriptome-based inference and generation of molecules with desired phenotypes using the Bayesian optimization algorithm), to generate new chemical structures of inhibitor or activator candidates for therapeutic target proteins by integrating chemically and genetically perturbed transcriptome profiles. In the algorithm, the substructures of multiple molecules selected based on the transcriptome analysis are fused in the design of new chemical structures by exploring the latent space of Transformer-based VAE using Bayesian optimization. Our results demonstrate the usefulness of the proposed method in terms of its high reproducibility of existing ligands for 10 therapeutic target proteins.

RESULTS

Overview of the Proposed Method. This study aimed to generate new molecules that act as inhibitors or activators against a given therapeutic target protein. Suppose that the transcriptome profiles obtained by gene knockdown or overexpression of a therapeutic target protein (termed target-perturbed transcriptome profiles) are available, and the transcriptome profiles obtained by adding molecules (including drugs) to human cells (termed chemically induced transcriptome profiles) are also available. Target-perturbed transcriptome profiles, which are obtained by gene knockdown and overexpression of a therapeutic target protein, are known to be correlated with the chemically induced profiles of inhibitors and activators of that protein, respectively.³³ Using the correlation between target-perturbed transcriptome profiles and chemically induced transcriptome profiles, we attempted to generate inhibitors or activators against a therapeutic target protein.

Figure 1 shows the schematic diagram of the proposed method, which consists of two phases: (A) first, we selected molecules that were transcriptionally correlated with a therapeutic target perturbation, which we call source

molecules, based on the correlation coefficients between the therapeutic target-perturbed and chemically induced transcriptome profiles; (B) then, we fused the substructures of multiple source molecules and generated new inhibitor and activator candidate molecules using Bayesian optimization.

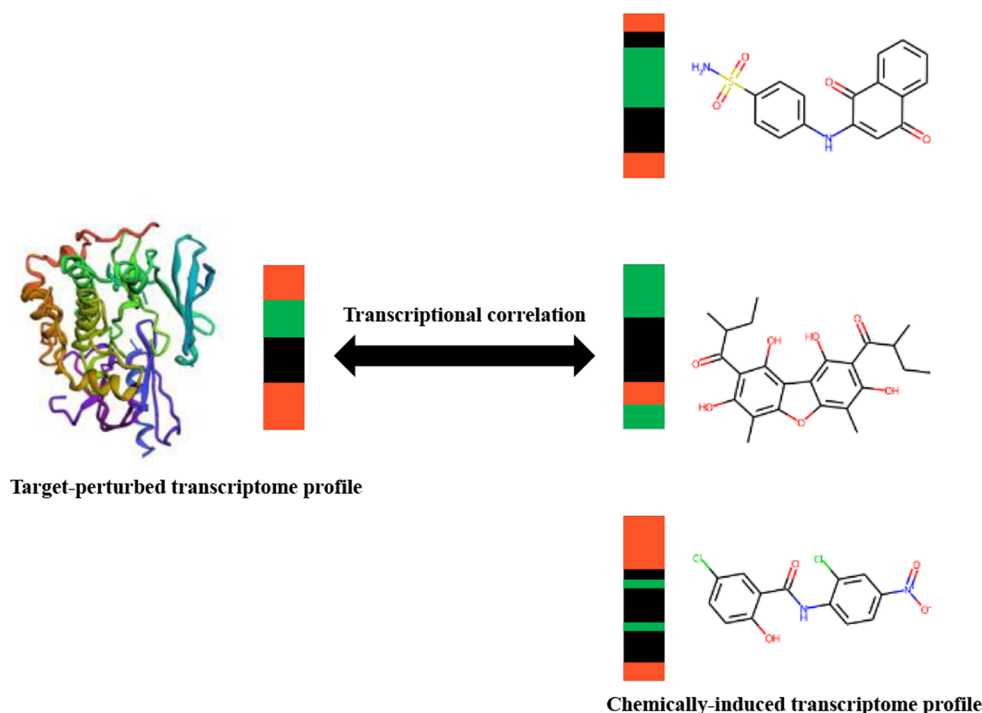
In the first phase, we determined source molecules using therapeutic target-perturbed and chemically induced transcriptome profiles. First, we prepared the target-perturbed transcriptome profiles obtained by gene knockdown or overexpression of a given therapeutic target protein and created target protein-specific transcriptome profiles of each target protein (see the [Target-Perturbed Transcriptome Profiles](#) subsection in the [Experimental Section](#)). Then, we searched for molecules that were transcriptionally correlated with a target perturbation, comparing the target-specific profile with the chemically induced profiles. Lastly, we selected the molecules with high transcriptional correlations as the source molecules.

In the second phase, we generated new molecules by exploring the latent space of a Transformer-based VAE using Bayesian optimization based on the source molecules for each target protein. We adopted TransVAE³⁷ as a molecular generator because VAE enables us to treat chemical structures as real-valued vectors and TransVAE was reported to be able to learn more detailed human-interpretable structural features of SMILES³⁸ than recurrent neural network (RNN)-based VAE with a single attention head.³⁷ First, the source molecule with the highest transcriptional correlation (the first source molecule) was input into the encoder of TransVAE and converted to a latent coordinate. Next, we explored the neighborhood of the latent coordinate of the first source molecule using Bayesian optimization with the objective function of the structural similarities with the second and subsequent source molecules and generated new molecules that incorporated substructures of multiple source molecules. We restricted the search space to the neighborhood of the first source molecule in order to generate molecules similar to the first source molecule, as structurally similar compounds tend to be closely located in the latent space. Thus, we searched for new molecules that were not only structurally similar to the second and subsequent source molecules in the objective function but also similar (i.e., close) to the first source molecule in the latent space in order to incorporate the substructures of the 1st–*k*th source molecules in the design of the new molecules.

Similar to previous studies with the same purpose,^{35,36} we chose 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), the mammalian target of rapamycin (MTOR), and the phosphatidylinositol 3-kinase catalytic subunit (PIK3CA) as examples of inhibitory target proteins and mothers against decapentaplegic homologue 3 (SMAD3) and tumor protein p53 (TP53) as examples of activatory target proteins.

Source Molecule Selection Based on Target-Perturbed Transcriptome Profiles. First, we created the target protein gene knockdown profile by gene knockdown of each inhibitory target protein and the target protein gene overexpression profile by gene overexpression of each activatory target protein (see the [Target-Perturbed Transcriptome Profiles](#) subsection in the [Experimental Section](#)). Subsequently,

A)



B)

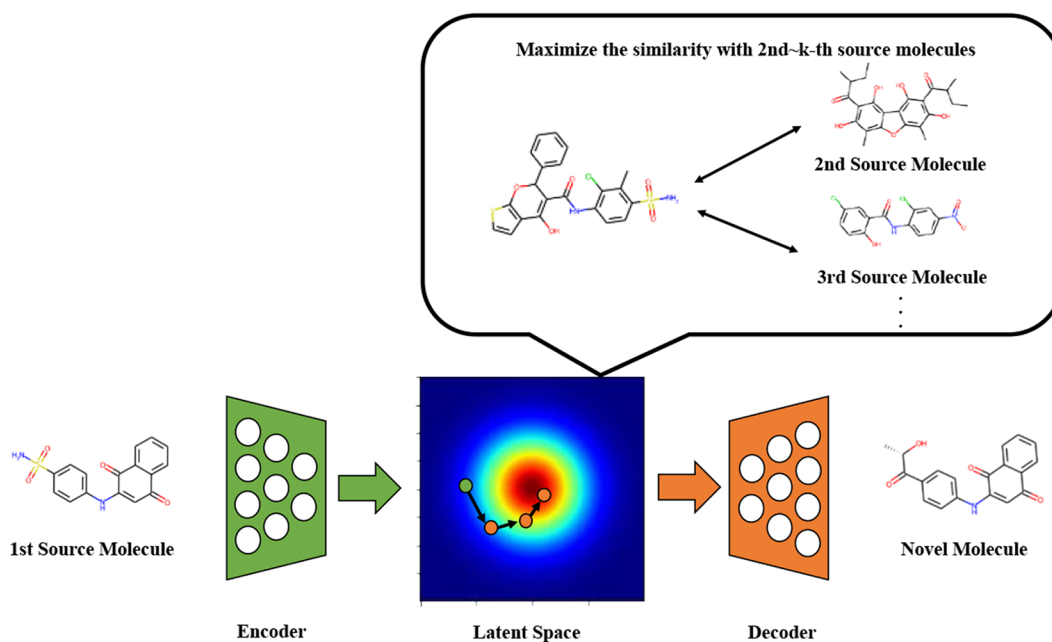


Figure 1. Overview of the proposed method. (A) We searched for molecules that were transcriptionally correlated with a therapeutic target perturbation, comparing the target-perturbed profile with the chemically induced profiles. Specifically, we calculated correlation coefficients between the target-specific profile and the chemically induced profiles and searched for molecules whose chemically induced profiles were correlated with the target-specific profile. Then, molecules with the top k highest correlation coefficients were selected as 1st- k th source molecules. (B) The first source molecule was input into an encoder of TransVAE and converted to a latent coordinate. Subsequently, we explored the neighborhood of the latent coordinate of the first source molecule using Bayesian optimization with the objective function of the similarities with the second and subsequent source molecules and generated new molecules that incorporated substructures of multiple source molecules.

we evaluated the transcriptional correlations between the target protein gene knockdown profile and the chemically induced profiles for each inhibitory target protein and the

transcriptional correlations between the target protein gene overexpression profile and the chemically induced profiles for each activatory target protein. Then, we selected molecules

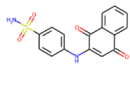
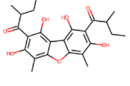
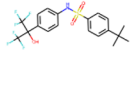
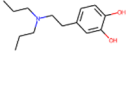
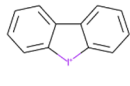
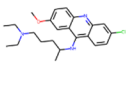
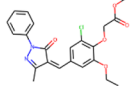
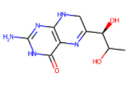
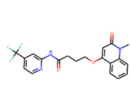
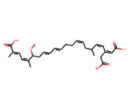
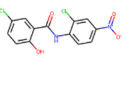
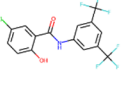
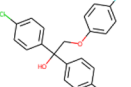
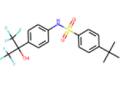
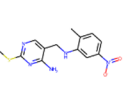
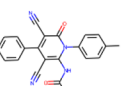
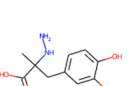
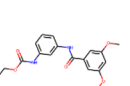
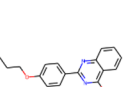
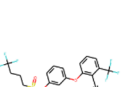
Therapeutic target protein	Source molecules	
AKT1	1st 	2nd 
	cc : 0.568	cc : 0.550
AKT2	1st 	2nd 
	cc : 0.575	cc : 0.519
AURKB	1st 	2nd 
	cc : 0.528	cc : 0.522
CTSK	1st 	2nd 
	cc : 0.457	cc : 0.433
EGFR	1st 	2nd 
	cc : 0.404	cc : 0.399
HDAC1	1st 	2nd 
	cc : 0.341	cc : 0.339
MTOR	1st 	2nd 
	cc : 0.505	cc : 0.486
PIK3CA	1st 	2nd 
	cc : 0.331	cc : 0.320
SMAD3	1st 	2nd 
	cc : 0.679	cc : 0.669
TP53	1st 	2nd 
	cc : 0.334	cc : 0.326

Figure 2. Source molecules selected for each of the 10 target proteins. For the inhibitory target proteins, we calculated the correlation coefficients between the target protein gene knockdown profiles and the chemically induced profiles. For the activatory target proteins, we calculated the correlation coefficients between the target protein gene overexpression profiles and the chemically induced profiles. Then, we selected the molecules with the top k highest correlation coefficients as the source molecules. Here, the top two source molecules used for Bayesian optimization with the parameter $k = 2$ are shown.

with the top k highest correlation coefficients as source molecules (1st– k th source molecules).

Figure 2 shows the selected source molecules, where the first and second source molecules used for Bayesian optimization with the parameter $k = 2$ are shown. Details of the Bayesian optimization and the k parameter are given in the [Bayesian Optimization to Generate New Molecules](#) subsection of the [Experimental Section](#). The maximum values of the correlation coefficients for AKT1, AKT2, AURKB, CTSK, EGFR, HDAC1, MTOR, PIK3CA, SMAD3, and TP53 were 0.568, 0.575, 0.528, 0.457, 0.404, 0.341, 0.505, 0.331, 0.679, and 0.334, respectively.

The source molecules of each target protein had diverse structural features. For AKT1, the first source molecule had a benzenesulfonamide and a 1,4-naphthoquinone, while the second source molecule had a dibenzofuran, carbonyl groups, and hydroxy groups. For AKT2, the first source molecule had a sulfonamide and trifluoromethyl groups, and both the first and second source molecules had a hydroxy group. For AURKB, the first source molecule was a diphenyleneiodonium, while the second source molecule had multiple nitrogen atoms, an acridine, and a chlorine atom. For CTSK, the first source molecule had an ester bond, while the second source molecule had a nitrogen-rich fused ring. For EGFR, the first source

molecule had a 4-trifluoromethylpyridine, an amide bond, and a 1,2-dihydroquinoline, while the second source molecule had a long, linear, unsaturated hydrocarbon and carboxy groups. For HDAC1, the first source molecule had a nitro group, while the second source molecule had trifluoromethyl groups. In addition, both source molecules shared a basic skeleton containing an amide bond. For MTOR, the first source molecule contained multiple halogen elements, while the second source molecule had a sulfonamide and trifluoromethyl groups. For PIK3CA, the first source molecule had a pyrimidine and a nitro group, while the second source molecule had one amide bond and multiple cyano groups.

Similar to the inhibitory target proteins, source molecules with diverse chemical structures were also selected for the activatory target proteins. For SMAD3, the first source molecule had hydroxy groups and a carboxy group, while the second source molecule had two amide bonds. For TP53, the first source molecule had a quinazoline and a *n*-propoxybenzene, while the second source molecule had a diphenyl ether, trifluoromethyl groups, a cyano group, and a sulfonate ester. Interestingly, 4-*tert*-butyl-*N*-[4-(1,1,1,3,3,3-hexafluoro-2-hydroxypropan-2-yl)phenyl]benzenesulfonamide was selected as a source molecule for both AKT2 and MTOR.

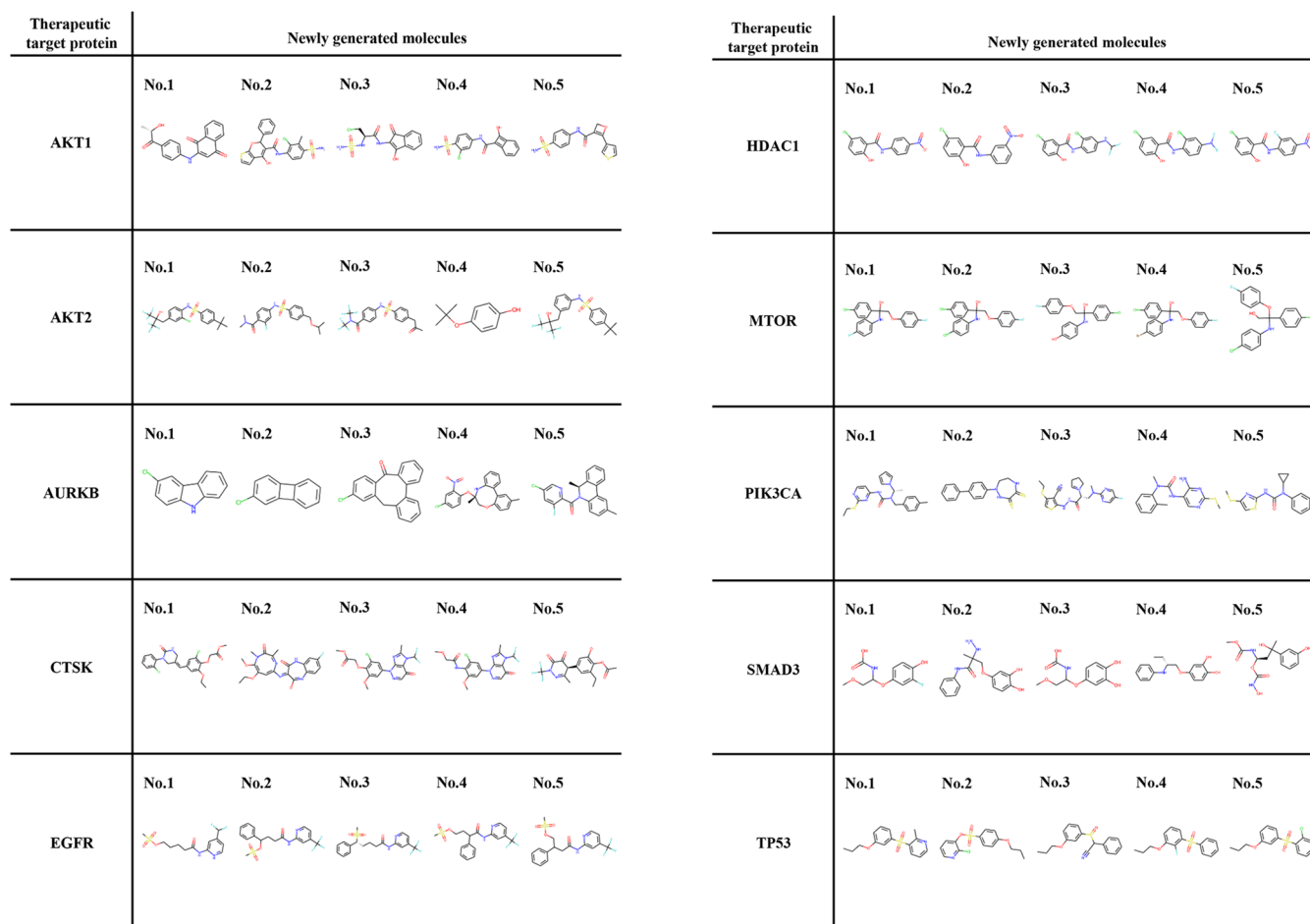


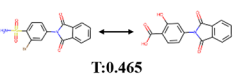
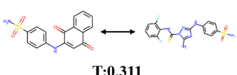
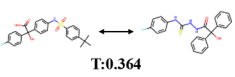
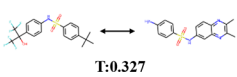
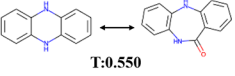
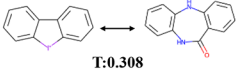
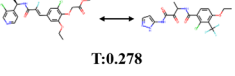
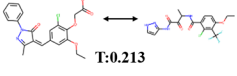
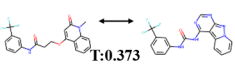
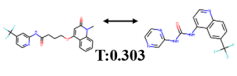
Figure 3. Newly generated molecules for each of the 10 target proteins. No. 1–5 indicate the molecules with the top 1–5 objective function values, respectively.

These source molecules that were derived from the therapeutic target protein transcriptome profiles could undergo further improvements to enhance their function as inhibitors or activators. Therefore, we aimed to generate new molecules that were more likely to work as inhibitors or activators than these source molecules by fusing their substructures using TransVAE and Bayesian optimization.

Bayesian Optimization-Based Molecule Generation of Inhibitors and Activators for Therapeutic Target Proteins. To generate molecules that were more inhibitor-like or activator-like than the source molecules, we first converted the first source molecule to a latent coordinate of TransVAE. Next, we generated new molecules via Bayesian optimization with the parameter $k = 2$.

Figure 3 shows the newly generated molecules with the top five highest scores of the objective function. The newly generated molecules for each target protein had structures derived from two source molecules. For AKT1, the benzenesulfonamide in the first source molecule was commonly found in molecules No. 2, 4, and 5, and molecules No. 1–4 commonly contained a hydroxy group, as did the second source molecule. For AKT2, in all of the molecules other than No. 4, there were structural changes in the side chains while the basic skeleton of the first source molecule including the sulfonamide was retained. Molecules No. 2 and 3 had a nitrogen atom bonded to multiple carbon atoms, as in the second source molecule. Furthermore, the No. 4 molecule

contained a *tert*-butyl group, as in the first source molecule, and a benzene ring with a hydroxy group that the second source molecule had as well. For AURKB, all of the top five molecules had a chlorine atom, as in the second source molecule. In addition, an iodine ion from the first source molecule changed to a nitrogen atom in the No. 1 molecule, which was abundant in the second source molecule, while the structure of a fused ring (the basic skeleton of the first source molecule) was retained. For CTSK, molecules No. 1, 3, and 5 had the same ester bond that the first source molecule contained, and all of the top five molecules had multiple nitrogen atoms, as in the second source molecule. In addition, molecules No. 2–4 had a fused ring containing multiple nitrogen atoms, as in the second source molecule. For EGFR, there were some structural changes, but the basic skeleton of the first source molecule (including a pyridine ring and an amide bond) was retained. Interestingly, all of the molecules other than No. 3 had a sulfonate ester. For HDAC1, despite some structural changes, all of the top five newly generated molecules retained the basic skeleton of two benzene rings connected by an amide bond, which was common to the two source molecules. For MTOR, the basic skeleton of the first source molecule was retained despite some structural changes. For PIK3CA, all of the top five molecules had a sulfur atom (as in the first source molecule) and multiple nitrogen atoms, commonly found in both source molecules. In addition, the No. 3 molecule had a cyano group, as in the second source

Therapeutic target protein	Generated molecule	Known ligand	1st source molecule	Known ligand
AKT1				
AKT2				
AURKB				
CTSK				
EGFR				

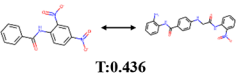
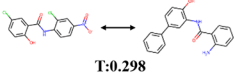
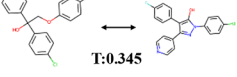
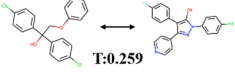
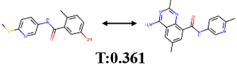
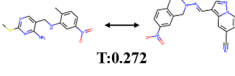
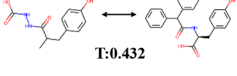
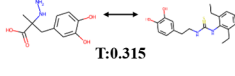
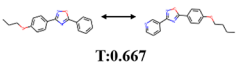
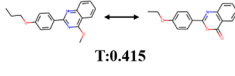
Therapeutic target protein	Generated molecule	Known ligand	1st source molecule	Known ligand
HDAC1				
MTOR				
PIK3CA				
SMAD3				
TP53				

Figure 4. Structural similarity between known ligands and the newly generated molecules and the first source molecules. The first column shows target protein names, the second column shows a comparison of the structures between the newly generated molecules and known ligands, and the third column shows a comparison of the structures between the first source molecules and known ligands. The *T* values indicate the Tanimoto coefficients.

molecule. Furthermore, all of the molecules other than No. 2 contained an amide bond, as in the second source molecule.

Similar to the inhibitory target proteins, molecules containing structures derived from two source molecules were generated for the activatory target proteins. For SMAD3, all of the top five molecules had a hydroxy group, as in the first source molecule. In addition, all of the molecules other than No. 4 had an amide bond, as in the second source molecule. For TP53, all of the top five molecules had a *n*-propoxybenzene, as in the first source molecule. In addition, the No. 2 and 3 molecules had a sulfonate ester and a cyano group, respectively, as in the second source molecule.

These results suggest that, by exploring the neighborhood of the first source molecule in the latent space to increase the similarity with the other source molecule, we are able to generate new molecules that incorporate substructures of multiple source molecules.

Evaluation of Newly Generated Molecules in Terms of Reproducibility of Existing Ligands. To confirm that the proposed method could generate more inhibitor-like or activator-like molecules than the first source molecule, which was not structurally optimized, we calculated the Tanimoto coefficients for structural similarity with known ligands for the newly generated molecules and the first source molecules. If a newly generated molecule shares similar chemical structural

features with a known ligand and has a high structural similarity, we can assume that the molecule is likely to act on the target protein. Note that we followed the same procedure that was used in the previous works with the same objective.^{35,36}

Figure 4 shows comparisons of the structures between the newly generated molecules and known ligands and between the first source molecules and known ligands. For all of the target proteins, the newly generated molecules were more similar to the known ligands than were the first source molecules. In particular, the newly generated molecules for AKT1, AURKB, HDAC1, SMAD3, and TP53 had a similarity to known ligands that was ≥ 0.1 higher than the first source molecules.

In addition, we observed the advantage of fusing substructures of multiple source molecules, especially in the newly generated molecules of AURKB and PIK3CA. For AURKB, the newly generated molecule incorporated nitrogen atoms that were abundant in the second source molecule while retaining the basic skeleton of a fused ring of the first source molecule, resulting in a significant increase in its structural similarity to the known ligand when compared to the first source molecule. Moreover, for PIK3CA, the newly generated molecule retained the two benzene rings and the sulfur atom found in the first source molecule and acquired an amide bond

Table 1. Maximum Structural Similarities between Newly Generated Molecules and Known Ligands^a

therapeutic target protein	proposed method	previous GAN-based method ³⁵	previous VAE-based method ³⁶
AKT1	0.423 ^b (0.389 < <i>x</i> < 0.457) ^c	0.317	0.417
AKT2	0.370 (0.346 < <i>x</i> < 0.394)	0.289	0.351
AURKB	0.500 (0.414 < <i>x</i> < 0.585)	0.364	0.340
CTSK	0.285 (0.259 < <i>x</i> < 0.311)	0.311	0.292
EGFR	0.386 (0.328 < <i>x</i> < 0.444)	0.298	0.306
HDAC1	0.483 (0.400 < <i>x</i> < 0.565)	0.339	0.304
MTOR	0.372 (0.329 < <i>x</i> < 0.416)	0.392	0.686
PIK3CA	0.344 (0.323 < <i>x</i> < 0.364)	0.261	0.324
SMAD3	0.422 (0.393 < <i>x</i> < 0.451)	0.439	0.476
TP53	0.620 (0.540 < <i>x</i> < 0.700)	0.457	0.530

^aBold numbers indicate the highest scores among the three methods. ^bThe mean value of the five runs. ^cThe 95% confidence interval of the five runs.

that the second source molecule contained, resulting in it having a higher structural similarity to the known ligand than the first source molecule. These results reveal that the proposed method of fusing together substructures of multiple source molecules can generate more ligand-like molecules than the first source molecule, which is not structurally optimized.

Performance Comparison between Our Proposed Method and Previous Methods. To evaluate the effectiveness of the proposed method, we compared its performance with two previous methods that generated molecules from transcriptome profiles.^{35,36} We selected the same 10 target proteins as in these two previous studies and adopted the maximum value of structural similarities between the newly generated molecules and known ligands as an evaluation metric, following the previous studies.^{35,36}

Table 1 shows the maximum structural similarities between the new molecules generated by the three methods and the known ligands. We ran the generation five times with different random seeds, and the mean values and 95% confidence intervals of the five runs are shown in Table 1. The detailed results of all five runs can be found in the Supporting Information (Table S1). The structures of the newly generated molecules and known ligands of the two previous studies were borrowed from the original paper,³⁶ and we calculated Tanimoto coefficients between them. Our proposed method outperformed the previous methods for 7 out of the 10 target proteins. This result confirms that the proposed method is more likely to generate ligand-like molecules than these previous methods, which had the same purpose.

Investigation of the Robustness of the Model Performance to the Active Ligand Threshold. To investigate the robustness of the model performance to the active ligand threshold, we moved the threshold to define active ligands from $\text{pXC}_{50} \geq 5.0$ to $\text{pXC}_{50} \geq 6.0$ and compared the results of the two activity thresholds. Table S2 shows the maximum structural similarities between the new molecules generated using our proposed method and the known ligands of the two activity thresholds. For all of the therapeutic target proteins except AKT1, AURKB, and TP53, the differences between the values of $\text{pXC}_{50} \geq 5.0$ and $\text{pXC}_{50} \geq 6.0$ were within 0.05. These results suggest that there is no significant difference in the tendency of most therapeutic target proteins if the activity threshold is moved. Please note that for a fair performance comparison, we used the results of $\text{pXC}_{50} \geq 5.0$ for the Performance Comparison between Our Proposed Method and Previous Methods section, as the two previous studies^{35,36} used $\text{pXC}_{50} \geq 5.0$ as the activity threshold.

DISCUSSION AND CONCLUSIONS

In this study, we developed a novel computational method called TRIOMPHE-BOA to generate new chemical structures of inhibitor or activator candidates for therapeutic target proteins by integrating chemically and genetically perturbed transcriptome profiles. We generated new molecules that incorporated substructures of multiple source molecules by exploring the latent space of TransVAE using Bayesian optimization with transcriptional constraints. The new molecules generated by the proposed method were more similar to known ligands than were the source molecules, which were not structurally optimized. The proposed method is expected to be useful for efficient hit identification.

In the proposed method, molecules with similar transcriptional patterns with a therapeutic target perturbation were selected as the source molecules. The first source molecules and known ligands with the highest similarity to them had many substructures in common, as shown in the third column of Figure 4. A benzenesulfonamide and a sulfonamide appeared in both the first source molecules and the known ligands for AKT1 and AKT2, respectively. For AURKB, both the first source molecule and the known ligand had the basic skeleton of a fused ring containing two benzenes, and for CTSK, a carbonyl group and an ether bond were common to both the first source molecule and the known ligand. In addition, for EGFR, 1,2-dihydroquinoline, an amide bond, and a trifluoromethyl group were common to both the first source molecule and the known ligand, while for HDAC1, both the first source molecule and the known ligand had the basic skeleton of a benzene ring connected by an amide bond. Furthermore, for MTOR, both the first source molecule and the known ligand contained three six-membered rings and fluorine and chlorine atoms. For PIK3CA, both the first source molecule and the known ligand had a nitro group and were nitrogen-rich. A similar trend was observed for the activatory target proteins: for SMAD3, a 1,2-dihydroxybenzene was common to both the first source molecule and the known ligand, and for TP53, both the first source molecule and the known ligand had an oxygen-bonded benzene ring and a fused ring containing a nitrogen atom. These substructures common to both the first source molecules and the known ligands indicate that the selected source molecules were reasonable as initial molecules for ligand design and that our method, which uses target-perturbed transcriptome profiles, is valid for source molecule selection.

The applicability domain (AD) is important in real-world applications. In transcriptome-based molecular generation, AD

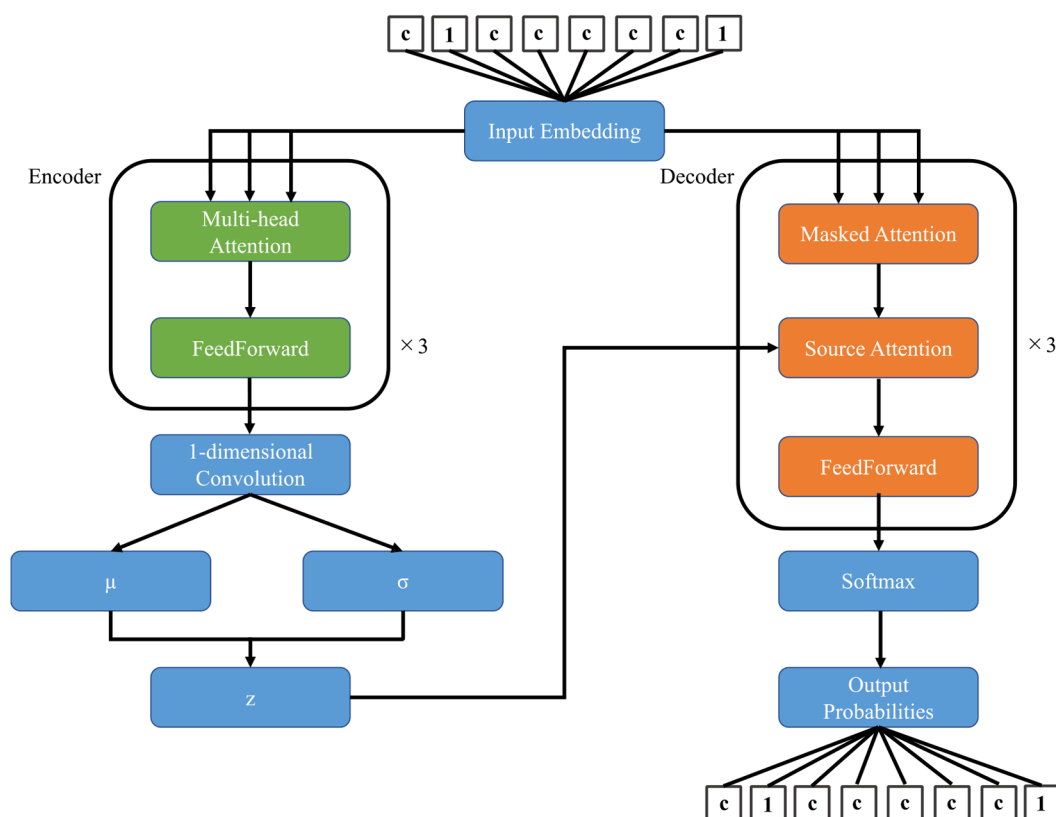


Figure 5. TransVAE architecture. (Normalization layers are omitted in this figure.) Similar to Transformer, TransVAE consists of encoder and decoder layers. The encoder output is passed through the 1-dimensional convolutional layer, converted to the mean and variance vectors of a 128-dimensional normal distribution, and z is calculated. Thereafter, z is decompressed and fed into the decoder's source-attention.

corresponds to the range of correlation coefficients of the gene expression profiles between a given target protein and source molecules. In our proposed method, we always adopted molecules with the top k highest correlation coefficients as the source molecules and did not consider the value of the correlation coefficients themselves. In practice, the proposed method may not work well if source molecules with extremely low correlation coefficients are selected. We believe that it is important to use only source molecules with relatively high correlation coefficients. Therefore, one possible way to consider AD in our proposed method is to establish an appropriate threshold for the value of the correlation coefficients with the source molecules.

One advantage of our proposed method is that it does not require structural information of the therapeutic target proteins; it is able to generate inhibitory or activatory candidate molecules with only a transcriptome profile of the knockdown or overexpression of the target protein gene. In modern drug discovery research, both structure-based drug design (SBDD), which uses the structure of target proteins, and ligand-based drug design (LBDD), which designs better ligands based on known ligand structures, play key roles. However, when the structures of target proteins are not available, SBDD does not work, and LBDD cannot be applied to proteins without known ligands or orphan proteins. Therefore, the fact that the proposed method does not require structural information about target proteins and can be applied to orphan proteins once transcriptome profiles of the target proteins are obtained is a major advantage.

There is room for improvement in our proposed method. One possible improvement is to change the way of measuring similarities between compounds and the molecular generation algorithm. In our proposed method, we measured the similarities between compounds using Tanimoto similarity, and we used TransVAE as a molecular generator. If we were to use latent-space similarity instead of Tanimoto similarity, we would not have to decode the latent coordinates into SMILES strings during Bayesian optimization. Thus, the use of latent-space similarity would reduce the computational cost. Moreover, since our objective function is simple similarity to the source molecules, a simpler molecular generator (e.g., an evolutionary graph or string optimization) could be used instead of TransVAE. Hence, it might be interesting to see how using latent-space similarity and a simpler molecular generator would affect the model performance. Another improvement direction would be to use cell-specific transcriptome profiles. In this study, we used transcriptome profiles from the MCF7 cell line for all 10 target proteins, but the use of cell-specific profiles might result in a better source molecule selection. In future work, we would like to develop a more practical molecule generation model by utilizing cell-specific transcriptome profiles.

EXPERIMENTAL SECTION

Target-Perturbed Transcriptome Profiles. We used the Library of Integrated Network-Based Cellular Signatures (LINCS)³⁹ to collect target-perturbed transcriptome profiles. We constructed 978-dimensional transcriptome profiles obtained by gene overexpression for SMAD3 and TP53 and

obtained by gene knockdown for AKT1, AKT2, AURKB, CTSK, EGFR, HDAC1, MTOR, and PIK3CA. We used transcriptome profiles from the MCF7 cell line, and when there were multiple profiles of the same target protein measured under different experimental conditions, we averaged them to create target protein-specific transcriptome profiles.

Chemically Induced Transcriptome Profiles. We collected 978-dimensional chemically induced transcriptome profiles from LINCS. LINCS stores transcriptome profiles obtained by adding numerous compounds to 77 different cultured human cell lines. In this study, we used 16 441 transcriptome profiles from the MCF7 cell line at a dose of 10 μ M. Then, we extracted 16 303 profiles with valid SMILES strings and excluded the following: (1) compounds with a molecular weight ≥ 500 , (2) compounds containing ≥ 4 ring structures, and (3) compounds containing ≥ 3 chiral centers. Finally, we obtained 7 677 profiles.

Compound Structures for TransVAE Training. We obtained compound chemical structures for training and validation of the TransVAE model from the L1000 and ZINC databases.^{39,40} After excluding duplicates and SMILES strings with string lengths >100 , we obtained 270 099 compounds. Then, we divided them into training data (243 088 compounds) and validation data (27 011 compounds). After excluding “Cl[Co]Cl” and “Cl[Pt+2]Cl” from the validation data, we had a training data set containing 243 088 compounds and a validation data set containing 27 009 compounds.

Known Ligands for Assessing Ligand Likeness of Newly Generated Molecules. We obtained known ligands of each therapeutic target protein from the ExCAPE database.⁴¹ In the ExCAPE database, each compound with a dose–response value equal to or lower than 10 μ M (i.e., $pXC_{50} \geq 5.0$) is annotated as active. In contrast, inactive compounds in concentration–response type assays (confirmatory type in PubChem⁴²) and compounds labeled as inactive in PubChem screening assays are annotated as inactive. To calculate the similarity between the newly generated molecules and the known ligands, we used a filtered data set of known ligands, where compounds included in the training and validation data of TransVAE were excluded.

TransVAE Architecture. TransVAE is a VAE with attention layers that consists of an encoder, latent space, and a decoder. Figure 5 shows the TransVAE architecture. The encoder and decoder are the same as those of Transformer, but the way TransVAE handles the output from the encoder is different than Transformer. While Transformer directly inputs the output from the encoder to the decoder’s source-attention, TransVAE first compresses the output values from the encoder into the latent space probabilistically and then decompresses the compressed information to feed it to the decoder’s source-attention. In this study, we used the hyperparameters with the highest validity as in the original paper (see ref 37): 256 dimensions for the embedding layer, 128 dimensions for the latent space, and 1024 dimensions for the feedforward layers. Details about TransVAE are written in the original paper.³⁷

Bayesian Optimization to Generate New Molecules. We performed Bayesian optimization to generate new molecules based on the latent space of TransVAE. Bayesian optimization is a method for finding explanatory variables that minimize or maximize an objective function.⁴³ We used GPyOpt, a Python library, to implement Bayesian optimization.

The objective function of Bayesian optimization is shown in eq 1:

$$\text{score}(\mathbf{z}_{\text{new}}) = \sum_{i=2}^k T(f(\mathbf{z}_{\text{new}}), \mathbf{u}_i) \quad (1)$$

where \mathbf{z}_{new} is the latent coordinate of a newly generated molecule, f is the function converting a latent coordinate to an ECFP4 fingerprint with 2048 dimensions, \mathbf{u}_i is an ECFP4 fingerprint with 2048 dimensions of the i th source molecule, k is the number of source molecules used for Bayesian optimization, and T represents the Tanimoto coefficient. We restricted the search range to the latent coordinate of the first source molecule ± 1.0 and explored a latent coordinate that maximizes the objective function. We set the number of initial samplings to 5 and the number of maximum iterations to 1000, and we used Expected Improvement as an acquisition function.

Calculation of Chemical Structure Similarity. We used an ECFP4 fingerprint with 2048 dimensions to calculate the structural similarity between compounds. The ECFP4 fingerprint is a binary vector that describes the chemical structure based on substructures at a distance of two atoms in radius. The Tanimoto coefficient was defined using the following equation:

$$T = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

where A and B are the set representation of the ECFP4 fingerprints of two compounds, $|A \cap B|$ is the number of bits that are both 1 in the two fingerprints, and $|A \cup B|$ is the number of bits that are 1 in at least one of the two fingerprints. A Tanimoto coefficient closer to 0 and 1 indicates low and high structural similarity, respectively.

■ ASSOCIATED CONTENT

Data Availability Statement

The source code and data used in this study are available at <https://yamanishi.cs.i.nagoya-u.ac.jp/triompheboa/>.

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.3c00824>.

The maximum structural similarities between the newly generated molecules and known ligands of our proposed method obtained from five runs with different random seeds (Table S1), and the maximum structural similarities between the new molecules generated using our proposed method and known ligands of the two activity thresholds (Table S2) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Yoshihiro Yamanishi – Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan; Graduate School of Informatics, Nagoya University, Nagoya, Aichi 464-8601, Japan; Phone: +81 52 789 5638; Email: yamanishi@i.nagoya-u.ac.jp; Fax: +81 52 789 5638

Authors

Yuki Matsukiyo – Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems

Engineering, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan; orcid.org/0000-0003-0699-9970
Chikashige Yamanaka – Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan; orcid.org/0000-0001-8302-9316

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.jcim.3c00824>

Author Contributions

Y.M. and C.Y. performed the computational analyses and contributed equally to the research. Y.M., C.Y., and Y.Y. wrote the manuscript. All authors approved the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by AMED under Grant JP22nk0101111 and JSPS KAKENHI (Grants 20H05797 and 21H04915).

ABBREVIATIONS

AKT1, RAC- α serine/threonine-protein kinase; AKT2, RAC- β serine/threonine-protein kinase; AURKB, aurora B kinase; CTSK, cysteine synthase A; EGFR, epidermal growth factor receptor; GAN, generative adversarial network; HDAC1, histone deacetylase 1; LBDD, ligand-based drug design; MTOR, mammalian target of rapamycin; PIK3CA, phosphatidylinositol 3-kinase catalytic subunit; QSAR, quantitative structure–activity relationship; SBDD, structure-based drug design; SMAD3, mothers against decapentaplegic homologue 3; TP53, tumor protein p53; VAE, variational autoencoder

REFERENCES

- (1) DiMasi, J. A.; Feldman, L.; Seckler, A.; Wilson, A. Trends in risks associated with new drug development: success rates for investigational drugs. *Clin. Pharmacol. Ther.* **2010**, *87*, 272–277.
- (2) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *J. Health Econ.* **2016**, *47*, 20–33.
- (3) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-based Drug Design: a Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.
- (4) Mayr, L. M.; Bojanic, D. Novel Trends in High-Throughput Screening. *Curr. Opin. Pharmacol.* **2009**, *9*, 580–588.
- (5) Ye, C.; Ho, D. J.; Neri, M.; Yang, C.; Kulkarni, T.; Randhawa, R.; Henault, M.; Mostacci, N.; Farmer, P.; Renner, S.; Ihry, R.; Mansur, L.; Keller, C. G.; McAllister, G.; Hild, M.; Jenkins, J.; Kaykas, A. DRUG-seq for Miniaturized High-Throughput Transcriptome Profiling in Drug Discovery. *Nat. Commun.* **2018**, *9*, 4307.
- (6) Szymański, P.; Markowicz, M.; Mikiciuk-Olasik, E. Adaptation of High-Throughput Screening in Drug Discovery-toxicological Screening Tests. *Int. J. Mol. Sci.* **2012**, *13*, 427–452.
- (7) Clark, M. A.; Acharya, R. A.; Arico-Muendel, C. C.; Belyanskaya, S. L.; Benjamin, D. R.; Carlson, N. R.; Centrella, P. A.; Chiu, C. H.; Creaser, S. P.; Cuzzo, J. W.; Davie, C. P.; Ding, Y.; Franklin, G. J.; Franzen, K. D.; Geffer, M. L.; Hale, S. P.; Hansen, N. J. V.; Israel, D. I.; Jiang, J.; Kavarana, M. J.; Kelley, M. S.; Kollmann, C. S.; Li, F.; Lind, K.; Mataruse, S.; Medeiros, P. F.; Messer, J. A.; Myers, P.; O'Keefe, H.; Oliff, M. C.; Rise, C. E.; Satz, A. L.; Skinner, S. R.; Svendsen, J. L.; Tang, L.; van Vloten, K.; Wagner, R. W.; Yao, G.; Zhao, B.; Morgan, B. A. Design, synthesis and selection of DNA-encoded small-molecule libraries. *Nat. Chem. Biol.* **2009**, *5*, 647–654.
- (8) Kleiner, R. E.; Dumelin, C. E.; Liu, D. R. Small-molecule discovery from DNA-encoded chemical libraries. *Chem. Soc. Rev.* **2011**, *40*, 5707–5717.
- (9) Goodnow, R. A., Jr.; Dumelin, C. E.; Keefe, A. D. DNA-encoded chemistry: enabling the deeper sampling of chemical space. *Nat. Rev. Drug Discovery* **2017**, *16*, 131–147.
- (10) Flood, D. T.; Kingston, C.; Vantourout, J. C.; Dawson, P. E.; Baran, P. S. DNA Encoded Libraries: A Visitor's Guide. *Isr. J. Chem.* **2020**, *60*, 268–280.
- (11) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. In *Conference proceedings: papers accepted to the International Conference on Learning Representations (ICLR)*; 2014.
- (12) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar Variational Autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*; 2017; 1945–1954.
- (13) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.
- (14) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *Proceedings of the 35th International Conference on Machine Learning*; 2018; 2323–2332.
- (15) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chem. Sci.* **2019**, *10*, 1692–1701.
- (16) Boitreau, J.; Mallet, V.; Oliver, C.; Waldispühl, J. OptiMol: Optimization of Binding Affinities in Chemical Space for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 5658–5666.
- (17) Gao, K.; Nguyen, D. D.; Tu, M.; Wei, G.-W. Generative Network Complex for the Automated Generation of Drug-like Molecules. *J. Chem. Inf. Model.* **2020**, *60*, 5682–5698.
- (18) Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*; 2014; 27.
- (19) Cao, N. D.; Kipf, T. MolGAN: An Implicit Generative Model for Small Molecular Graphs. *arXiv* **2018**, DOI: 10.48550/arXiv.1805.11973.
- (20) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharmaceutics* **2017**, *14*, 3098–3104.
- (21) Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. *arXiv* **2017**, DOI: 10.48550/arXiv.1705.10843.
- (22) Li, C.; Yamanaka, C.; Kaitoh, K.; Yamanishi, Y. Transformer-Based Objective-Reinforced Generative Adversarial Network to Generate Desired Molecules. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*; 2022; 3884–3890.
- (23) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90–98.
- (24) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8.
- (25) Qi, H.; Wang, F.; Tao, S.-C. Proteome microarray technology and application: higher, wider, and deeper. *Expert Rev. Proteom.* **2019**, *16*, 815–827.
- (26) Morganti, S.; Tarantino, P.; Ferraro, E.; D'Amico, P.; Duso, B. A.; Curigliano, G. Next Generation Sequencing (NGS): A Revolutionary Technology in Pharmacogenomics and Personalized Medicine in Cancer. *Adv. Exp. Med. Biol.* **2019**, *1168*, 9–30.

- (27) Yang, X.; Kui, L.; Tang, M.; Li, D.; Wei, K.; Chen, W.; Miao, J.; Dong, Y. High-throughput transcriptome profiling in drug and biomarker discovery. *Front. Genet.* **2020**, *11*, 19.
- (28) Pereira, R.; Oliveira, J.; Sousa, M. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. *J. Clin. Med.* **2020**, *9*, 132.
- (29) Dai, L.; Gao, X.; Guo, Y.; Xiao, J.; Zhang, Z. Bioinformatics clouds for big data manipulation. *Biol. Direct* **2012**, *7*, 43.
- (30) Fillinger, S.; de la Garza, L.; Peltzer, A.; Kohlbacher, O.; Nahnsen, S. Challenges of big data integration in the life sciences. *Anal. Bioanal. Chem.* **2019**, *411*, 6791–6800.
- (31) Turanli, B.; Karagoz, K.; Gulfidan, G.; Sinha, R.; Mardinoglu, A.; Arga, K. Y. A network-based cancer drug discovery: from integrated multi-omics approaches to precision medicine. *Curr. Pharm. Des.* **2019**, *24*, 3778–3790.
- (32) Chen, B.; Garmire, L.; Calvisi, D. F.; Chua, M.-S.; Kelley, R. K.; Chen, X. Harnessing big 'omics' data and AI for drug discovery in hepatocellular carcinoma. *Nat. Rev. Gastroenterol. Hepatol.* **2020**, *17*, 238–251.
- (33) Sawada, R.; Iwata, M.; Tabei, Y.; Yamato, H.; Yamanishi, Y. Predicting Inhibitory and Activatory Drug Targets by Chemically and Genetically Perturbed Transcriptome Signatures. *Sci. Rep.* **2018**, *8* (1), 156.
- (34) Namba, S.; Iwata, M.; Yamanishi, Y. From drug repositioning to target repositioning: prediction of therapeutic targets using genetically perturbed transcriptomic signatures. *Bioinformatics* **2022**, *38*, i68–i76.
- (35) 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. *Nat. Commun.* **2020**, *11* (1), 10.
- (36) Kaitoh, K.; Yamanishi, Y. TRIOMPHE: Transcriptome-Based Inference and Generation of Molecules with Desired Phenotypes by Machine Learning. *J. Chem. Inf. Model.* **2021**, *61*, 4303–4320.
- (37) Dollar, O.; Joshi, N.; Beck, D. A. C.; Pfaendtner, J. Attention-based generative models for de novo molecular design. *Chem. Sci.* **2021**, *12*, 8362–8372.
- (38) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- (39) 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.; Sorger, P. K.; Subramanian, A.; Ma'ayan, A. LINCS Canvas Browser: Interactive Web App to Query, Browse and Interrogate LINCS L1000 Gene Expression Signatures. *Nucleic Acids Res.* **2014**, *42*, W449–W460.
- (40) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.
- (41) Sun, J.; Jeliazkova, N.; Chupakhin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; Kochev, N.; Ashby, T. J.; Chen, H. ExCAPE-DB: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. *J. Cheminf.* **2017**, *9*, 1–9.
- (42) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem BioAssay: 2014 update. *Nucleic Acids Res.* **2014**, *42*, D1075–D1082.
- (43) Frazier, P. I. A Tutorial on Bayesian Optimization. *arXiv* **2018**, DOI: 10.48550/arXiv.1807.02811.