Multiobjective Molecular Optimization for Opioid Use Disorder Treatment Using Generative Network Complex

Hongsong Feng,
Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States

Rui Wang,
Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States

Chang-Guo Zhan,
Department of Pharmaceutical Sciences, University of Kentucky, Lexington, Kentucky 40506, United States

Guo-Wei Wei
Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States; Department of Electrical and Computer Engineering and Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States

Abstract

Opioid use disorder (OUD) has emerged as a significant global public health issue, necessitating the discovery of new medications. In this study, we propose a deep generative model that combines a stochastic differential equation (SDE)-based diffusion model with a pretrained autoencoder. The molecular generator enables efficient generation of molecules that target multiple opioid receptors, including mu, kappa, and delta. Additionally, we assess the ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties of the generated molecules to identify druglike compounds. We develop a molecular optimization approach to enhance the pharmacokinetic properties of some lead compounds. Advanced binding affinity predictors were built using molecular fingerprints, including autoencoder embeddings, transformer embeddings, and topological Laplacians. Our process yields druglike molecules that can be used in highly focused experimental studies to further evaluate their pharmacological effects. Our

machine learning platform serves as a valuable tool for designing effective molecules to address OUD.
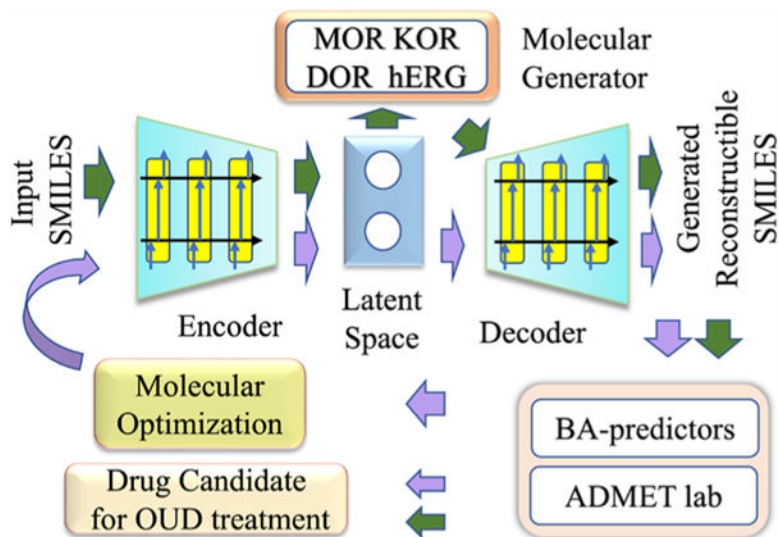
Graphical Abstract

Table of Contents Graphic

1. INTRODUCTION

Opioid use disorder (OUD) is a chronic and intricate condition characterized by the compulsive seeking and use of drugs despite the detrimental effects.[1] It represents a significant public health concern, causing severe consequences for individuals, families, and communities. The opioid epidemic has become a pressing global health crisis, highlighting the urgent need for effective treatments for OUD. Safe and effective medication treatments can alleviate withdrawal symptoms, reduce cravings, and help individuals maintain abstinence from opioids.[2]

The main treatment methods for OUD typically involve a combination of medications and behavioral interventions,[3] aiming to address the physical and psychological aspects of addiction, promote recovery, and prevent relapse. The U.S. Food and Drug Administration (FDA) has approved three medications, including methadone, buprenorphine, and naltrexone for the treatment of OUD.[4] These medications exert their effects by binding to opioid receptors in the brain, namely, mu opioid receptor (MOR), kappa opioid receptor (KOR), and delta opioid receptor (DOR). Methadone is a long-acting opioid agonist that primarily acts on MORs. It helps alleviate withdrawal symptoms and cravings.[5] Buprenorphine, on the other hand, acts as a partial opioid agonist primarily targeting MORs. It eases withdrawal symptoms and cravings while producing less euphoria and carrying a reduced risk of respiratory depression compared to methadone.[6] Naltrexone, classified as an opioid antagonist, blocks the effects of opioids and reduces the rewarding effects. Its mechanism of action primarily involves MORs, but it also exhibits some affinity for KORs.[7]

The pharmaceutical effects of medications for OUD treatment are associated with the functions of opioid receptors in the brain. MOR triggers euphoria and is essential for stimulating the reward system and goal-directed behaviors.[8] In patients with opioid addiction, their goal-directed behaviors shift toward habitual behaviors, leading to compulsive drug use.[9] KOR exhibits anti-reward effects and is capable of inducing dysphoria.[10] Prolonged opioid exposure may lead to stress responses, but KOR blockade can help alleviate these responses, decrease drug cravings, and reduce depressive states. Additionally, DORs can lower anxiety levels and mitigate depressive symptoms.[11]

While current medications effectively address OUD, relapse and remission remain common due to neurobiological changes and opioid receptor tolerance resulting from repeated opioid abuse.[4] Additionally, some patients may not tolerate or respond optimally to the standard medications used for OUD. Alternative medications provide additional options to customize treatment according to individual needs and preferences. The drug discovery process encompasses several stages, including target discovery, lead discovery, lead optimization, preclinical development, and three phases of clinical trials before a new drug can be brought to market.[12] Traditional drug discovery is a time-consuming endeavor that can extend over many years, require significant financial investments amounting to billions of dollars, and entail a substantial failure rate.

Various methods and technologies have emerged to accelerate the drug discovery process. The number of potential druglike molecules is estimated to be between 10^{23} and 10^{60}.[13] High-throughput screening (HTS) allows for the rapid screening of large compound libraries against specific biological targets or disease models, quickly identifying leads for further medicinal chemistry optimization.[14] It allows for effective automated operation, but is associated with high costs of equipment and assay development. Virtual screening involves the use of computational methods to virtually screen large databases of compounds against specific target structures. It employs molecular docking, molecular dynamics simulations, and machine learning algorithms.[15] These methods enable the prediction of compound–target interactions, assessment of physicochemical and pharmacological properties, and identification of compounds with potential therapeutic effects.[16]

De novo drug design (molecular generation) explores the chemical space to generate novel molecules with desirable properties. The advancement of deep learning has opened up new opportunities for innovative drug design and discovery. In recent years, many deep learning-based algorithms have been developed for de novo drug design. These algorithms utilize the power of deep learning to generate novel molecular structures with desired properties. They are trained on large datasets of known molecules, allowing them to learn the intricate patterns and relationships between chemical structures and their corresponding activities.[17] The machine learning algorithms used to construct these generative models can be categorized into four main types: recurrent neural network (RNN), encoder-decoder, reinforcement learning (RL), and generative adversarial network (GAN).[18] For example, an RNN-based generative model has been proposed to generate novel molecules. By employing a fine-tuning strategy with small sets of molecules, the generated compounds can exhibit activity toward specific biological targets.[19,20] Additionally, a variational autoencoder (VAE) has been utilized to encode molecules into a continuous latent space.[17] This enables

operations such as generating new molecules or applying optimization strategies to design compounds with desired properties. Perturbations or interpolations can be performed on molecules' latent space vectors, and gradient-based optimization can be applied to their continuous representations.[17] Furthermore, a generative GAN combined with an autoencoder has been employed for de novo molecular design.[21] This approach allows for the generation of random druglike compounds or compounds biased toward specific targets.

In recent years, diffusion models have gained popularity in various fields, including image synthesis, video generation, and molecule design, due to their ability to produce high-quality and realistic samples. There are a few main subtypes of diffusion models: latent space random noise (LSRN),[20] denoising diffusion probabilistic models (DDPMs),[22] score-based generative models (SGMs),[23] and stochastic differential equations (SDEs).[24] SDEs describe the evolution of a system over time, considering both deterministic and stochastic forces influenced by random noise.[25] Generative diffusion models have been applied in the field of drug discovery. Random noise was introduced to latent space molecular vectors to generate novel druglike molecules.[20] LSRN was compared with two other approaches, latent space controlled output and latent space optimized output for drug generation in a generative network complex (GNC).[20] Another example is DiffLinker,[26] which utilizes diffusion models for the design of molecular linkers. DiffSBDD[27] is used in structure-based drug design to generate high-quality ligands for specific protein targets. Some recent works[28] have also employed diffusion models trained in the latent space of autoencoders.

In this study, we employ deep generative models to design and optimize molecules that have potential applications in the treatment of OUD. Specifically, we utilize diffusion method in the latent space of a pretrained autoencoder (AE) model to generate novel molecules. Our objective is to create molecules that exhibit similar structural and pharmacological properties to known opioids or alternative compounds with therapeutic potential. The development of medications for OUD relies on the binding effects of opioid receptors, particularly MOR, KOR, and DOR. To achieve the molecular design, we combine a stochastic differential equation (SDE)-based diffusion method with the latent space of the pretrained autoencoder model. This enables us to design molecules that are active on MOR, KOR, and DOR. In the diffusion modeling, we incorporate appropriate reference and seed compounds to steer the generation of target-biased molecules. Additionally, we employ accurate binding affinity (BA) predictors to identify potentially effective molecules that interact with these critical targets. Importantly, we consider the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties in selecting druglike compounds for OUD treatment. By integrating ADMET criteria, we identify the generated compounds that posse desirable druglike properties. We conduct extensive experiments to efficiently generate druglike compounds. Furthermore, we employ a molecular optimization approach to discover additional drug candidates with nearly optimal properties. By employing different reference compounds and employing various molecular novelty thresholds, we successfully identify several druglike compounds that exhibit multitarget effectiveness on the critical opioid receptors. Our molecular generation platform serves as a valuable tool for advancing OUD treatment.

2. EXPERIMENTS

2.1. Structure of Multitarget Stochastic Generative Network Complex (MTSGNC).

2.1.1. Sequence-to-Sequence Autoencoder.—An autoencoder is a type of artificial neural network in deep learning that learns to compress and then reconstruct input data, such as images or text. The encoder network maps the input data to latent space, while the decoder network maps the latent space back to the original data. The compressed representation, or latent space between the two networks can capture essential features of the input data. The commonly used network for encoder and decoder can be a gated recurrent unit (GRU) or a long short-term memory (LSTM) network.

In molecular science, autoencoder can be used for molecular representation learning, which involves encoding molecules into a lower-dimensional space while preserving their structural and functional properties. For instance, a sequence-to-sequence (seq2seq) autoencoder achieves this by translating one molecular string representation to another.[29] Simplified molecular-input line-entry system (SMILES) is one of the commonly used molecular representations. The dimension of the latent space is 512. The seq2seq AE was trained with a high reconstruction ratio between inputs and output SMILES, through which latent vectors carry faithful molecular information.[30] Therefore, the latent space can be utilized to represent the chemical space and serve as a molecular fingerprint for machine learning modeling.

The autoencoder also has an application for generative modeling, where the decoder can be used to generate new data points by sampling from the latent space. The autoencoder (AE) structure is illustrated in Figure 1(top left). In the current study, we utilize the pretrained seq2seq autoencoder[29] to design novel molecules for OUD treatment.

2.1.2. Data Preparation.—MOR, KOR, and DOR are three critical pharmaceutical targets for treating OUD. Additionally, hERG is a crucial target that needs to be avoided in drug discovery.[31] We gathered inhibitor datasets for the four targets from the ChEMBL database. The data points include SMILES strings of molecular compounds and corresponding binding labels in the form of IC_{50} or K_i. Augmenting the IC_{50} with K_i can be beneficial to large-scale analysis. As suggested by Kalliokoski,[32] the IC_{50} label can be approximated to K_i value using the relation $K_i = IC_{50}/2$ for data in broad datasets like ChEMBL database. These experimental labels can be converted to binding energy with the formula $BA = 1.3633 \times \log_{10} K_i (kcal/mol)$.[33] The protein-ligand binding affinity is reflected by the dissociation constant $K_d = [L][P]/[LP]$, where $[L]$, $[P]$, and $[LP]$ are the molar concentration of ligand, protein, and protein-ligand complex, respectively. It is noted that Gibbs free energy (kcal/mol) can be derived by $\Delta G = RT \ln K_d$, where R and T are the gas constant and temperature, respectively. At room temperature $T = 298.15$ K, one has $\Delta G = -1.3633 pK_d$.[33] Here, pK_d is $-\log_{10} K_d$ with K_d in the unit of mol. K_i is the inhibition constant. Following the way that PDBbind database mixes K_d and K_i in their refined datasets,[34] we calculate the binding energy with the above BA calculation formula.[33] For compounds with multiple BA values, we calculate their average as the final BA. The details about the four collected datasets can be found in the Supporting Information.

The datasets are utilized in two main aspects. First, the molecular compounds, particularly those exhibiting potency on MOR, KOR, and DOR, are employed as references and seeds for generating new compounds. Second, the four datasets are utilized as training sets to construct machine learning models for predicting binding affinity.

2.1.3.   Stochastic Molecular Generator.—Generative models serve as powerful tools for generating potential new drug molecules. In our previous work, we introduced the GNC for the generation of new druglike molecules in 2019.[20,35] We considered three latent space perturbation models, including (1) a random noise diffusive model called "randomized output", (2) a gradient descending model called "controlled output", and (3) a multiobjective optimization model called "optimized output". The first model perturbs the latent space molecular vectors with the Gaussian white noise and then, selects molecules that have improved properties in a manner similar to the Monte Carlo method. This model ensures the novelty of generated molecules but may not be effective in reaching the desirable druglike properties. The second model improves specific molecular properties of generated molecules by a force-driven term. However, the resulting new molecules may not retain other important properties and lack novelty. The third approach was designed to simultaneously optimize multiple molecular properties of generated molecules via a multi-objective loss function. However, the novelty of generated molecules may not be guaranteed. To improve the performance of our GNC model, we propose to combine our random noise diffusive model with our multiobjective optimization model. The resulting model can be regarded as a drift and diffusion model as described by the Langevin equation.

Langevin equation is a stochastic differential equation (SDE), which is used to describe the diffusion processes such as the random motion of the particles over time in the particle's velocity space, taking into account both deterministic forces and random forces. It can be regarded as stochastic generalization of the Newton's equation of motion. In this work, one of our goals is to employ Langevin equation to optimize the molecular generator in our previous GNC model.

We assume $\mathbf{X}$ is a latent space vector of a molecule with 512 dimensions, and $\mathbf{X}_k$ represents its kth latent space reference vector. Then, the Langevin equation of our drug generator system is:

$$\frac{d\mathbf{X}}{dt} = \alpha \sum_k a_k(\mathbf{X}_k - \mathbf{X}) + \xi(t)$$

(1)

where a_k is a positive weighting parameter corresponding to $\mathbf{X}_k$ satisfying $\sum_k a_k = 1$, $\xi(t)$ is a Gaussian white noise, and α is a hyperparameter. The first term in eq 1 is a force term that gives rise to a gradient descent toward targets $\mathbf{X}_k$. Then, according to eq 5 in Section 2.3.2, the general solution of this system is given by:

$$\mathbf{X}(t) = \mathbf{C}e^{-at} + \int_0^t e^{-\alpha(t-u)}\left(\alpha \sum_k a_k\mathbf{X}_k + \xi(u)\right)du$$

(2)

where the initial state $\mathbf{X}(0) = \mathbf{C}$. The latent vector $\mathbf{C}$ is for an initial seed molecule.

In our implementations, we employ the fourth-order Runge–Kutta method[36] to numerically solve eq 1 and obtain the corresponding molecular representations in the latent space. The Runge–Kutta approximation is achieved using the function solve_ivp from the Python scipy.integrate library. At different time points t, $X(t)$ represents the latent vector of a novel molecule. We limit the time t to the range [0, 1000] and use an appropriate time step to generate different $X(t)$ vectors. The selection of reference molecules and weighting parameters $(\mathbf{X}_k, \alpha_k)$ in eq 2 is crucial. The number of reference molecules and their corresponding weights can significantly influence the molecular novelty and properties inherited from the given reference compounds. Furthermore, the α value is associated with the impact of the initial seed molecule on the generated compounds. A smaller value allows for a more pronounced impact within a certain time frame. This approach enables us to model the diffusion-based generation of molecules using the Langevin eq 1.

It is worth noting that the Langevin equation provides a molecule-wise description of drift and diffusion processes. However, to obtain an understanding of how the overall distribution of molecules evolves over time, a statistical perspective is needed. Therefore, we also discuss the Fokker–Planck equation derived from the Langevin equation (the detailed derivation can be found in Section S7) in the Supporting Information, providing a connection between the dynamics of individual druglike molecules and the statistical behavior of the entire system.

2.1.4.   Multitarget Stochastic Molecular Generation.—Figure 1 presents the compound generation process in our GNC. We aim at designing novel druglike compounds effective on MOR, KOR, and DOR, while having no hERG side effect.[37] The molecular generation process comprises four key steps, which are further elaborated in the subsequent subsections.

1.   Select three compounds that are potent on MOR, KOR, and DOR, respectively, as reference compounds from the collected datasets. Also pick a seed compound that is potent on multiple of the three receptors. Then, encode the SMILES strings of the references and seed compounds into latent vectors through the pretrained encoder.

2.   Input the latent vectors of reference and seed compounds into the stochastic molecular generator, giving rise to a large number of new latent vectors. These new latent vectors, as the representations of potential molecules, are screened with a constraint of binding affinity $\Delta G < -9.54 \, \text{kcal/mol}(K_i = 1 \, \text{nm})$ on MOR, KOR, and DOR, as well as $\Delta G > -8.18 \, \text{kcal/mol}(K_i = 10 \, \text{nm})$ on hERG. Pretrained BA predictors are used in the BA constraint evaluation.

3.   Latent vectors with desired BA properties are decoded into valid SMILES (interpretable by RDKit). These SMILES are subsequently fed back through the pretrained encoder and decoder again, as illustrated in Figure 1, to identify

molecules that can be successfully reconstructed. Those reconstructed molecules are considered to be structurally stable and well interpreted by the autoencoder networks.

4.   Those reconstructed molecules are reevaluated on their binding affinities and ADMET properties to identify druglike compounds. A molecular optimization process can be implemented to design more drug candidates with desired druggable profiles.

To generate novel molecules, different references and seed compounds, or weights for reference latent vectors can be used in Step 2. The hyperparameter α in eq 1 is set as 0.15 in this study. We set the value for α as a relatively small number.

As we consider three reference compounds corresponding to the three receptors, we apply three reference vectors in eq 1. Weight coefficients α_k, k = 1, 2, and 3, are for MOR, KOR, and DOR reference compounds, respectively. Different weight values can be used to emphasize the importance of selected references. The parameter t in eq 2 can be various numbers to generate vectors for potential molecules.

2.2.   Binding Affinity Predictors.

Within the generator of our GNC, we incorporate four BA predictors to assess the BA of the generated potential molecules on the four crucial targets. To construct these predictors, we utilize molecular fingerprints derived from the latent space of an autoencoder network. These fingerprints are referred to as AE-FP, and the resulting four BA predictors are denoted as AE-BPs. The AE-BPs were fulfilled by integrating the AE-FPs with a deep neural network (DNN) algorithm.

Consensus models are used to further evaluate the BAs of those molecules whose SMILES can be reconstructed, as shown in the right bottom of Figure 1. Two more molecular fingerprints, namely, transformer fingerprint (TF-FP) and topology Laplacian fingerprint (TL-FP) are used. They were designed with a pretrained transformer model[38] and our recently proposed topology Laplacian theory.[39] Brief descriptions of the two fingerprints are provided in the following subsections.

BA predictors were constructed by integrating the TF-FP with a deep neural network (DNN) algorithm, and fusing the TL-FP with a gradient boosting decision tree (GBDT) algorithm. We refer to these two models as TF-BP and TL-BP, respectively. The consensus model or prediction is obtained by averaging the predictions from AE-BP, TF-BP, and TL-BP. This strategy can enhance machine learning predictions[37,40] and typically outperforms individual models. Such approaches were employed in our previous studies on OUD, which involved machine learning repurposing of DrugBank compounds for OUD treatment[41] and machine learning analysis of the OUD interactome networks.[42]

Both DNN and GBDT algorithm are popular algorithms in building machine learning models. DNN has advantages of dealing large and complex datasets, constructing hierarchical features and modeling complex nonlinear relationships. GBDT as an ensemble algorithm has the merits of being less sensitive to hyperparameters, less prone to overfitting,

and easy to implement. In building our machine learning BA predictors, AE-FP and TF-FP showed better predictive performance when combined with DNN, while TL fingerprints have better predictive ability when integrated with GBDT. The details about the hyperparameter for AE-BPs, TF-BPs, and TL-BPs can be found in Table S2 in the Supporting Information.

The predictive performance of the consensus models was evaluated using 5-fold cross-validation. The average Pearson correlation coefficients (R) obtained were 0.824, 0.840, 0.845, and 0.756 for the MOR, KOR, DOR, and hERG datasets, respectively. Additionally, the average root-mean-square error (RMSE) values were 1.010, 1.027, 1.006, and 0.801 kcal/mol for the same datasets.

2.2.1. Topological Laplacian (TLs) Molecular Fingerprint.—In this subsection, we give a brief explanation of topological Laplacians (TLs) based on the spectral graph theory proposed by Wang et al.[39] Topology offers significant simplification of biomolecular data by dealing with the connectivity of different components in a space and characterizes independent entities, rings and higher-dimensional faces within the space.[43,44] It can be used for a high level of abstraction to three-dimensional (3D) biomolecular structures. Topological Laplacian can reveal both topological invariants and homotopic shape information through the harmonic and nonharmonic spectra of the Laplacian matrix.[45–48] Intricate shape information can be obtained through evolving manifolds defined under filtration parameters. The topological space is based on geometric components of a dataset, including discrete vertices, edges, triangles, tetrahedrons in the context of 3D molecular structures. TL forms families of persistent q-combinatorial Laplacian operators, providing a powerful multiscale analysis tool. These operators are derived from persistent spectral graph theory, as illustrated below.

The persistent Laplacians are defined under a filtration of an oriented simplicial complex K. A sequence of the subcomplexes $\{K_t\}_{t=0}^{m}$ of K is constructed

$$\emptyset = K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_m = K$$

The notation t in this subsection denotes the filtration process, distinct from the time t mentioned in the previous section. On each simplicial complex K_t, a chain complex is defined as $C_q^t := C_q(K_t)$ and there exists a q-boundary operator $\partial_q^t : C_q(K_t) \to C_{q-1}(K_t)$. For the general case with $0 < q \leq \dim(K_t)$, the q-boundary operator is in the following form:

$$\partial_q^t(\sigma_q) = \sum_{i}^{q} (-1)^i \sigma_{q-1}^i, \quad \text{for } \sigma_q \in K_t$$

where $\sigma_q = [v_0, v_1, \cdots, v_q]$ is an oriented q-complex and $\sigma_{q-1}^i = [v_0, \cdots, \hat{v}_i, \cdots, v_q]$ is an oriented $(q-1)$-simplex by removing vertex v_i. For the case of $q < 0$, the $C_q(K_t) = \{\emptyset\}$ and ∂_q^t is a zero map. The q-adjoint boundary operator is defined as the adjoint operator that corresponds to the q-boundary operator.

$$\partial_q^* : C_{q-1}(K_t) \rightarrow C_q(K_t)$$

We consider $\mathbb{C}_q^{t+p}$, a subset of C_q^{t+p} with its boundary in C_{q-1}^t:

$$\mathbb{C}_q^{t+p} := \left\{ \sigma \in C_q^{t+p} \,\middle|\, \partial_q^{t+p}(\sigma) \in C_{q-1}^t \right\}$$

For this subset, the p-persistent q-boundary operator $\partial_q^{t+p} : \mathbb{C}_q^{t+p} \rightarrow C_{q-1}^t$ and the adjoint boundary operator $(\partial_q^{t+p})^* : C_{q-1}^t \rightarrow \mathbb{C}_q^{t+p}$ are well defined. The p-persistent q-combinatorial Laplacian operator is given as

$$\Delta_q^{t+p} = \partial_{q+1}^{t+p}\left(\partial_{q+1}^{t+p}\right)^* + \left(\partial_q^t\right)^* \partial_q^t$$

together with its matrix representation as

$$\mathscr{L}_q^{t+p} = \mathscr{B}_{q+1}^{t+p}\left(\mathscr{B}_{q+1}^{t+p}\right)^T + \left(\mathscr{B}_q^t\right)^T \mathscr{B}_q^t$$

Matrices $\mathscr{B}_{q+1}^{t+p}$ and $\mathscr{B}_q^t$ are the matrix representations for boundary operators ∂_{q+1}^{t+p} and ∂_q^t, respectively. The row number of $\mathscr{B}_{q+1}^{t+p}$ is equal to the number of oriented q-simplices in K_t, and the column number equals that of oriented $(q+1)$-simplices in $K_{t+q} \cap \mathbb{C}_{q+1}^{t+p}$. In addition, the transposes of $\mathscr{B}_{q+1}^{t+p}$ and $\mathscr{B}_q^t$ are the matrix representation for $(\partial_{q+1}^{t+p})^*$ and $(\partial_q^t)^*$. The topological and spectral information of K_t can be accessed from the Laplacian operator. We denote spectra of $\mathscr{L}_q^{t,p}$ as a set

$$\text{spectra}\left(\mathscr{L}_q^{t+p}\right) = \left\{ (\lambda_1)_q^{t+p}, (\lambda_2)_q^{t+p}, \cdots, (\lambda_N)_q^{t+p} \right\}$$

where N indicates the dimension of $\mathscr{L}_q^{t+p}$. The Betti numbers, the number of zero eigenvalues, of $\mathscr{L}_q^{t+p}$ can reveal q-cycle information. For the p-persistent q-combinatorial Laplacian matrix $\mathscr{L}_q^{t+p}$, the Betti number is defined as

$$\beta_q^{t+p} = \dim\left(\mathscr{L}_q^{t+p}\right) - \text{rank}\left(\mathscr{L}_q^{t+p}\right) = \text{nullity}\left(\mathscr{L}_q^{t+p}\right)$$
$$= \text{number of zero eigenvalues of } \mathscr{L}_q^{t+p}$$

The β_q^{t+p} value indicates the number of q-cycles in simplices K_t that are still alive in simplices K_{t+p}. For the biomolecular data, the order of q ranges from 0 up to 2, as the data is in three-dimensional space. The values of β_q^{t+p} measure the persistence of connected components, tunnels or circles, and cavities or voids. The harmonic persistent spectra track the topological changes while nonharmonic persistent spectra record the geometric changes.

Based on the aforementioned topological Laplacians, we form a set of molecular features by using the eigenvalue statistics of Laplacian matrix $\mathscr{L}_0^{t+p}$. The features are compromised of β_0^{t+p} and the sum, mean, median, maximum, minimum, standard deviation, variance, and sum of the square of the nonharmonic spectra. The representability of TL feature for

molecules depends on the selection of atoms with different combinations of elements, which in turn construct distinct oriented q-simplices in K_t. As a result, element-specific Laplacian matrices are defined under a set of filtration. To enhance the representability of TL features, it is necessary to analyze the given dataset on element types and atomic proportions. Additionally, this study utilizes a filtration radius with a lower bound of 1 angstrom and an upper bound of 10 angstroms. This range is chosen based on the observation that the majority of compounds in each dataset have a three-dimensional size of less than 20Å in each Cartesian direction. More details of element-specific topological Laplacians based on distribution analysis can be found in the Supporting Information.

2.2.2. Bidirectional Transformer Molecular Fingerprint.—In a recent work,[38] a self-supervised learning (SSL) platform was developed to pretrain deep learning models on millions of unlabeled molecules. This platform generated latent space vectors for input SMILES.[38] The pretraining of SSL was accomplished by employing the bidirectional encoder transformer (BET) model. Within the SSL pretraining platform, SMILES strings were encoded by constructing pairs of real SMILES and masked SMILES, with a certain percentage of symbols in the strings hidden. The model was then trained in a supervised manner using these data-mask pairs.[38] The attention mechanism was utilized to capture the significance of each symbol in the SMILES strings. A set of molecular fingerprint can be obtained by averaging 256 embedding vectors associated with a given SMILES string. For the training of the SSL-based BET model, molecular SMILES from ChEMBL databases were employed, and the latent vector transformer fingerprints (TF-FP) generated by the pretrained model were used as molecular fingerprints in this study.

2.3. Langevin Equation.

2.3.1. Random Variables and Expected Value.—A random variable X is a variable whose possible values are outcomes of a random phenomenon. The random variable can be either discrete, taking on a countable number of values, or continuous, taking on any value within a certain range or set. For a discrete random variable, we can write $P(X = x)$ to denote the probability that X takes the value x. The expected value of random variable X is $E(X) = \sum_i x_i P(X = x_i)$. For a continuous random variable, we talk about probability density function (pdf) $p(x)$ such that for any interval $[a, b]$, $P(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x$. Then, the expected value of random variable X is $E(X) = \int_{-\infty}^{+\infty} x p(x)\mathrm{d}x$.

2.3.2. Langevin Equation.—The Langevin equation is a commonly used stochastic differential equation (SDE) in physics that aims to describe the behavior of a system as it evolves over time under the influence of deterministic drift and random (fluctuating) forces. The Langevin equation can describe the motion of a particle in a fluid,

$$m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\lambda \mathbf{v} + \boldsymbol{\eta}(t)$$

(3)

where m is the mass of the particle, $\mathbf{v}$ is the velocity of the particle, λ is its corresponding damping coefficient, and $\boldsymbol{\eta}$ is the noise term which represents the effect of the collisions with

the molecules of the fluid. In many cases, the one-dimensional Langevin equation is written in a general form as:

$$\frac{dx}{dt} = -\gamma x + \xi(t)$$

(4)

where $\xi(t)$ is a Gaussian white noise process with $\langle \xi(t) \rangle = 0$ and $\langle \xi(t)\xi(t') \rangle = \delta(t - t')$. The general solution of the one-dimensional Langevin equation has the form:

$$x(t) = Ce^{-\gamma t} + \int_0^t e^{-\gamma(t - u)}\xi(u)du$$

(5)

where the initial state $x(0) = C$.

2.4. Generating Novel Multitarget Inhibitors for MOR, KOR, and DOR.

The FDA-approved medications for OUD treatment are highly effective on MOR, KOR, and DOR. We utilize our GNC to design more molecules that are simultaneously effective on the three receptors. Meanwhile, we avoid those inhibitors with potential hERG side effects.

Selecting appropriate reference and seed compounds is crucial in generating effective molecules. We prioritize compounds from the collected datasets that show effectiveness on the opioid receptors as our references. This is because the generated compounds, inheriting the pharmacophores of such references, are more likely to exhibit potency on the receptors. Figure 2a displays the binding affinity (BA) distributions of inhibitors in these datasets. It is evident that there is a significant number of effective molecules in the MOR, KOR, and DOR datasets, all having binding affinity values below −9.54 kcal/mol. This threshold is widely accepted for identifying active compounds. The three datasets collectively consist of 2152 common compounds, from which we select our reference or seed compounds

We rely on machine learning models to predict the BA values of the generated compounds. The three datasets demonstrate broad BA distributions, ranging from −14 to −6 kcal/mol, which indicates the presence of highly diverse molecules. Moreover, the BA data exhibit balanced distributions in relation to the BA threshold of −9.54 kcal/mol. This balanced distribution of training data enables unbiased BA prediction.

Each dataset contains a moderate number of molecules with BAs ranging between −12 and −10 kcal/mol. We prioritize selecting compounds within this range as reference or seed compounds for two reasons. First, these compounds increase the likelihood of generating potent molecules. References showing high effectiveness on multiple targets are especially valuable for this purpose. Second, the ample data within this BA range aids in accurately identifying potent inhibitors through machine learning predictions.

Reference compounds are pivotal in drug design as they greatly impact the novelty of generated compounds. Two crucial factors influencing novelty are the number of reference

compounds and the coefficient weights assigned to their AE latent vectors. Novelty is measured by comparing the similarities between the generated compounds and the reference compounds, with lower similarity indicating higher novelty. In the upcoming experiments, we will demonstrate the mechanism of our molecular GNC and specifically investigate the impact of reference numbers on molecular novelty.

2.4.1.   Generation Using Three Distinct Reference Molecules.—In our first experiment, we selected three compounds, namely, ChEMBL2048770, ChEMBL3349979, and ChEMBL494462, from the inhibitor datasets. Each of these compounds demonstrates effectiveness on MOR, KOR, and DOR, with corresponding BA values of −11.51, −11.58, and −11.92 kcal/mol, respectively. Therefore, we utilized them as reference compounds for each respective receptor. The seed compound exhibits BA values of −10.44, −10.5, and −8.96 kcal/mol on MOR, KOR, and DOR, respectively. It acts as a weak inhibitor of DOR. By using ChEMBL494462 as the reference compound for DOR, the generated compounds can incorporate certain moieties present in ChEMBL494462. Consequently, those generated molecules may exhibit effectiveness on DOR. When applying the three references in the molecular generator, the weight coefficients (a_1, a_2, a_3) in eq 1 are set to (0.35, 0.35, 0.3).

Using the reference and seed compounds, our GNC generated over a million novel and valid molecules in just a few hours using supercomputers. Subsequently, these compounds were passed through the encoding-decoding network, and we retained those that could be successfully reconstructed for further BA reevaluation and ADMET analysis. The reconstruction rate of the generated compounds was 90.1%, yielding a vast library of novel molecules.

AE-BPs are initially used to evaluate the BAs of the generated compounds, identifying those multitarget active molecules. Figure 2b indicates the BA distributions of the reconstructed molecules by our AE-BPs, indicating a large number of active compounds for each of the three opioid receptors. In addition, only a very small portion of these molecules can cause hERG side effects. It is promising to find enough multitarget active compounds.

Investigating molecular similarity scores from various perspectives is crucial as they are associated with machine learning predictions. One important similarity score is between the reference compound and the respective source dataset. These reference compounds are selected from the MOR, KOR, and DOR inhibitor datasets, which were utilized as training data for developing BA predictors. The generated molecules exhibit similarities with the reference compounds. Figure 2c presents the similarity score distribution between the reference compounds and their source datasets. A small number of molecules within each training dataset exhibit high similarities with the corresponding reference compounds. The similarity scores are calculated using Tanimoto coefficients, comparing the AE latent vectors of the reference compounds with the molecules in each dataset. On average, the similarity scores between each reference compound and its corresponding dataset remain below 0.4. However, there are still more than ten molecules in each dataset that exhibit similarity scores over 0.85 for the selected reference compound.

Figure 2d showcases the similarity score distribution between the generated molecules and the three inhibitor datasets. Each similarity score is determined by the highest Tanimoto coefficient calculated between the AE latent vectors of a generated molecule and all molecules in the respective dataset. The majority of similarity scores fall within the range of 0.4–0.6, indicating high levels of novelty among the generated molecules. This outcome aligns with expectations, as each new molecule incorporates molecular features from three distinct reference compounds, resulting in unique molecular structures. Notably, the distribution curves of the three datasets exhibit a consistent pattern.

Figure 2e presents the distribution of similarity scores between the generated molecules and the three reference compounds. The majority of similarity scores are below 0.5, further confirming the high novelties of the generated molecules. The similarity distributions of the molecules with MOR and KOR reference compounds demonstrate a consistent pattern, while the average similarity scores with the DOR reference compound are comparatively lower than those for MOR and KOR. This discrepancy is primarily attributed to the weights assigned to eq 1, specifically (0.35, 0.35, 0.3). A higher weight assigned to a reference compound leads to a greater similarity between the generated molecules and the designated reference compound.

The novelties of the generated molecules are observed in Figure 2d,e. High novelties come with the risk of inaccurate BA prediction, as a higher molecular similarity with the training data can have more reliable predictions. To address this issue, we can either use more accurate BA predictors or reduce the molecular novelty.

2.4.2. Generation Using Two Reference Molecules.—We conduct a test using two reference compounds to generate novel molecules, aiming to improve the molecular similarities to the training data and enhance the accuracy of BA predictions. The same two compounds, ChEMBL2048770 and ChEMBL494462, used in the previous experiment were utilized. ChEMBL2048770 exhibits high potency as an inhibitor for both MOR and KOR, with binding affinities of −11.51 and −11.78 kcal/mol, respectively. ChEMBL494462 specifically binds to DOR with a BA value of −11.92 kcal/mol. In this test, ChEMBL2048770 was selected as the reference compound for both MOR and KOR, while ChEMBL494462 served as the reference compound for DOR. We continued to use ChEMBL243195 as the seed compound. The weight coefficient (a_1, a_2, a_3) is set to be (0.35, 0.35, 0.3) in eq 1. We utilized our GNC to generate millions of new valid molecules, and these new molecules have a reconstruction ratio of 86.31% through the autoencoder encoding-decoding network.

Figure 3 presents the results of our molecule generation in the second experiment. Similar to the previous experiment, our GNC successfully generated a significant number of active molecules for MOR, KOR, and DOR targets, while exhibiting weak hERG side effects, as depicted in Figure 3a. As anticipated, the generated molecules showed improved similarity scores with the training data, as observed in Figure 3c. The average similarity score of approximately 0.6 was higher than the score of around 0.45 obtained using three reference compounds in our previous experiment, as shown in Figure 2d. Furthermore, in Figure 3d, we can observe that the similarity scores for the MOR and KOR reference compounds are

higher than those for the DOR reference compound, which aligns with Figure 3b. It is important to note that the distribution curves of the MOR and KOR reference compounds overlap in Figure 3b,d, as ChEMBL2048770 was used as the reference compound for both MOR and KOR targets.

By comparing the two experiments, we can see that utilization of two reference compounds can be an effective approach to design novel molecules yet with high similarity scores to the training data. This, in turn, enables more accurate BA predictions.

2.4.3.   Binding Affinity Reevaluation with Consensus Models.—We use AE-BPs for an initial screening of the generated molecules. By applying BA constraints, we generate a large pool of novel molecules that may function as effective multitarget inhibitors. To refine this pool, we employ our consensus models to reevaluate the BAs of the filtered generated molecules. The BA constraints are applied once again by the consensus reevaluations, resulting in a reduced number of compounds for ADMET analysis. Below, we continue to carry out further investigations on the above experiment using two reference compounds.

2.5.   ADMET Analysis.

ADMET (absorption, distribution, metabolism, excretion, and toxicity) plays a critical role in drug discovery and development as it encompasses various attributes related to a compound's pharmacokinetic studies. A promising drug candidate must demonstrate both efficacy on the therapeutic target and compliance with essential ADMET properties. Accurate ADMET predictions are vital in drug design as they enable the screening of new compounds' properties and help mitigate the risk of late-stage attrition.

For systematic ADMET screening, we examined six indexes: FDAMDD, $T_{1/2}$ and $F_{20\%}$, Log P, Log S, and Caco-2. To assess these ADMET properties, we utilized the machine learning predictions provided by ADMETlab2.0 solvers (https://admetmesh.scbdd.com/).[49,50] The provided documentation offers optimal ranges for various ADMET properties. Furthermore, we evaluated the synthetic accessibility score (SAS) of the compounds, employing RDKit for the evaluation. Table 1 provides the optimal ranges for ADMET properties and SAS. By systematically evaluating the binding effects, ADMET properties, and SAS, we conducted a search for potential compound leads.

2.5.1.   Potential Optimal Drug Candidates.—After performing consensus BA and SAS predictions, we proceeded to evaluate the ADMET properties of the identified multitarget active compounds in the second experiment. Among those, 2155 compounds satisfy the BA constraints and are in the SAS proper range. According to the ADMETlab2 predictions, only a very limited number of compounds can satisfy the various properties. Figure 4a presents the ADMET screening results. The orange frames outline the proper domains for a pair of properties. The color points represent the predicted BA values of the generated compounds on MOR.

The first plot in Figure 4a shows the distributions of FDA maximum recommended daily dose (FDAMDDs), an index of potential for toxicity, and Caco-2, cell permeability of

compounds. The FDAMDD screening reveals that a very small fraction of our generated compounds can satisfy toxicity requirements, despite the fact that nearly half of them can pass the Caco-2 screening. This suggests the necessity of ADMET screening before a new compound is synthesized.

The second plot in Figure 4a displays the screening results based on two parameters: $F_{20\%}$ (bioavailability of 20%) and $T_{1/2}$ (half-life). The plot shows that a significant number of the compounds fall within the optimal domain for both indexes. Especially, almost all of these generated potent inhibitors can pass the $T_{1/2}$ screening.

The third plot in Figure 4a illustrates the screening based on Log P and Log S, which are parameters that relate to the distribution of drugs in the human body. The outline optimal domain covers only a small portion of the plot. While nearly half of the compounds fall within the acceptable range of Log S, very few of these potent inhibitors are within the suggested range for Log P. This suggests that a significant amount of resources are being wasted in early studies.

The screenings for the compound on FDAMDD and Log P indexes impose strict filtering criteria for finding optimal compounds from the 2155 compounds, which limits the number of potential drug candidates. None of the 2155 potent compounds passed the ADMET screening. To increase the pool of possible drug candidates, we relaxed the Log P requirement by adjusting the proper range to 0–5 log mol/L, as Lipinski's rule suggests Log P less than 5 Log mol/L for an orally active drug.[51] Using this new range, we were able to identify three drug candidates, as shown in Figure 4b. The predicted binding affinity values for the four critical targets, as well as their similarity scores to two reference compounds, are provided.

The ADMETlab2 server was used to evaluate a range of other ADMET indexes for the three nearly optimal compounds. Figure 4c shows that the additional physicochemical properties of the two molecules were within the appropriate ranges, except for Log P and Log D. Here, Log D, which is associated with Log P, refers to the logarithmic value of Log P at physiological pH 7.4. Upon structural optimization of the three molecules, it is possible to achieve simultaneous optimization on both Log P and Log D.

2.6. Molecular Optimization.

As noted earlier, the FDAMDD and Log P profiles pose obstacles that prevent the generated molecules from becoming optimal drug candidates in the second experiment. To expand the pool of optimal drug candidates, we consider optimizing the generated compounds that exhibit desired BA values and nearly satisfactory ADMET properties. Log P index is the objective we strive to optimize. We could also carry out molecular optimization using other property indexes, but we limit our efforts on Log P in the current study.

2.6.1. Log P Optimization.—The polarity of a molecule can influence its Log P value. Highly polar molecules may have lower Log P values, as they are more soluble in the aqueous phase and less likely to partition into the lipid phase. The relationship between molecular polarization and Log P can be complex and depends on other factors such as

molecular size, shape, and functional groups. To induce more polarization, we propose replacing a hydrogen atom with a hydroxyl group. These modified molecules are then subjected to scrutiny for molecular validity through our encoding-decoding process. The reconstructed molecules are subsequently evaluated for their BAs and ADMET properties using our consensus BA predictors and ADMETlab2. This optimization process may yield compounds that are closer to being optimal in terms of desired BAs and ADMET profiles. We tested this strategy on the three aforementioned molecules, which were nearly optimal except for their Log P profiles.

Figure 5 displays the molecular optimization on its Log P profile for the first nearly optimal compound in Figure 4b. By replacing one hydrogen atom with a hydroxyl group on the molecule, we obtained 22 new valid molecules. Figure 5a shows the statistics of the 22 new compounds on the magnitude of predicted BA values on MOR, KOR, DOR, and hERG, as well as their Log P values. The original compound with its BAs and Log P value is shown in Figure 5b. The average predicted BA values on MOR, KOR, DOR, and hERG are −10.15, −10.24, −9.49, and −7.65 kcal/mol. The original generated compound has predicted BA values of −10.43, −10.34, −9.62, and −8.0 kcal/mol, as well as Log P of 4.28. These 22 derived compounds exhibited slightly reduced potency on the four critical targets but improved Log P profiles. The optimization process alleviated hERG side effect and reduced Log P values. Figure 5c–h displays the six derived molecules with the highest average BA potency on the MOR, KOR, and DOR. The red circles highlight the positions where a hydrogen atom is replaced. All of the six compounds showed improved Log P profiles. In addition, five of them were all predicted to be effective inhibitors on MOR, KOR, and DOR without hERG side effects. Among the 22 derived compounds, 9 were predicted to be effective on the three targets while their Log P values are less than 4.0. These results demonstrate the effectiveness of the optimization process in offering drug candidates with improved pharmacokinetic profiles. In addition to the potency and Log P, other pharmacokinetics properties are taken into account again to identify nearly optimal compounds. Unfortunately, none of the 22 compounds met the criteria for binding affinity and ADMET properties, thereby ruling out their candidacy as new drugs. This reflects the challenge in molecular optimization, where multiobjective optimization is characterized by the inherent trade-off between improving one property at the expense of another.

We applied the Log P optimization approach to the other two nearly optimal compounds shown in Figure 5b and screened them for potency and ADMET properties. Upon replacing a hydrogen atom with a hydroxyl group, we obtained 20 new valid molecules for the second compound and 21 new valid molecules for the third compound. Among these, 11 inhibitors effectively targeted multiple receptors (MOR, KOR, and DOR) for the second compound, while the third compound yielded 20 such inhibitors. The average predicted BA values for the 20 new compounds on MOR, KOR, DOR, and hERG were −9.98, −10.06, −9.62, and −7.52 kcal/mol, respectively. Similarly, for the 21 new compounds, the average predicted BA values were −10.68, −10.99, −9.92, and −7.38 kcal/mol, respectively. The average Log P values for the 20 compounds were 3.47, and for the 21 compounds, it was 3.85. Among the 20 compounds, 11 were effective on all three opioid receptors, and 18 had Log P values less than 4. Ten compounds exhibited desired BA and Log P values. Among the 21 compounds, 20 were effective on all three opioid receptors, and 14 had Log P values

less than 4. Thirteen compounds showed desired BA and Log P values. Figure 6 illustrates some of our optimization results for the third compound in Figure 5b. The six best derived compounds, in terms of Log P value, are presented. It can be observed that modifications to various functional groups of the molecule contributed to a reduction in Log P values.

Through molecular optimization of the three nearly optimal compounds, we have obtained some new compounds. Among these compounds, a significant number exhibit desired BA and Log P values, thereby providing more promising candidates as nearly optimal compounds. By screening these compounds on their BAs and additional ADMET properties using our BA predictors and ADMETlab2, we have identified two compounds with improved Log P profiles that meet all BA and ADMET requirements. These two compounds, derived from applying Log P optimization to the third nearly optimal compound shown in Figure 4b, are presented in Figure 7. The positions where hydrogen atoms were replaced are highlighted in Figure 7a,b. The screening results for a series of physicochemical properties are depicted in Figure 7c,d. In comparison to the aforementioned three nearly optimal compounds, the two derived compounds exhibit improved Log P and Log D profiles.

The optimization process is beneficial to the generation of additional drug candidates in an efficient way. We screened millions of compounds before we could identify three nearly optimal compounds in Figure 4b. The Log P optimizations of the three compounds provide the two additional compounds with improved Log P profiles in a short time.

2.6.2. Molecular Interactions between Opioid Receptors and Effective Inhibitors.—It is crucial to understand the molecular mechanism of drug–target interactions in identifying desired drug candidates. To predict the docking poses of one drug candidate to opioid receptors, namely, MOR, KOR, and DOR, we utilized the molecular docking software AutoDock Vina.[52] The three receptors are in the same protein family and share high structural similarities. Above we identified five nearly optimal drug candidates including two derived compounds. The compound in Figure 7a showed high potency profiles and improved ADMET properties. Its molecular docking poses on the receptors are depicted in Figure 8.

It is observed that hydrogen bonds play critical roles in molecular interactions. There are at least two hydrogen bonds in each drug–target interaction system, which contribute to the high potency of the molecules on the receptors. The molecule in Figure 7a was derived with the Log P optimization by replacing a hydrogen atom in the methyl group with a hydroxyl group. The hydroxyl group itself plays a critical role in molecular interactions with the three receptors by forming hydrogen bonds. In its interaction with MOR, a hydrogen bond is formed between one oxygen atom on residual Tyr148(A) of MOR and the oxygen in the hydroxyl group. The second hydrogen bond in this interaction occurs between one nitrogen atom of the molecule and an oxygen atom in residual Asp147(A) of MOR. When interacting with KOR, the oxygen atom in the hydroxyl group forms hydrogen bonds with a nitrogen atom and an oxygen atom on residual Cys210(A), respectively. In its interaction with DOR, two hydrogen bonds are formed between the hydroxyl group and two oxygen

atoms on residual Asp128(A) of DOR. Additionally, another hydrogen bond exists between a hydrogen atom on Asp128(A) and a nitrogen atom on the compound.

2.7. Additional Nearly Optimal Compounds.

Above, we have presented a comprehensive overview of the workflow within our complex generative network for the discovery of novel multitarget compounds. These compounds exhibit druglike potential based on machine learning predictions and hold promise for OUD treatment. Utilizing our GNC, we generated additional multitarget compounds using different reference and seed compounds. Through screening their BA values, SAS scores, and ADMET properties, we identified additional compounds with druglike potential for treating OUD. Further details on these compounds can be found in the CSV Supporting Information.

3.   DISCUSSION

3.1.   Designing Analogue Drugs of the Approved Medications.

Currently, the US FDA has approved three medications, namely, methadone, buprenorphine, and naltrexone, for the treatment of OUD. These medications exert their pharmacological effects by targeting MOR, KOR, and DOR. Additionally, naloxone is a crucial medication used for the treatment of opioid overdose. We are interested in generating potential analogues of these four medications. We utilize these medications as reference compounds. Our focus is on designing analogues that exhibit simultaneous activity on MOR, KOR, and DOR. To achieve this, we employ both molecular generation and optimization approaches.

3.1.1.   Designing Buprenorphine Analogues.—Buprenorphine acts as a partial agonist for the MOR receptor and an antagonist for the KOR receptor. It can alleviate opioid withdrawal symptoms, reduce the effects of injected opioids, and provide protection against overdose.[53] Buprenorphine has a ceiling effect on euphoria and carries a lower risk of respiratory depression compared to methadone.[54] Compared to methadone, buprenorphine has lower retention rates. Higher doses of buprenorphine can increase retention and abstinence rates. Improved retention rates can be achieved with buprenorphine analogues, or alternatively, maintaining moderate retention rates through reduced doses.

Buprenorphine exhibits activity on MOR, KOR, and DOR with BA values of −12.55, −12.83, and −11.57 kcal/mol, respectively. To generate analogues of buprenorphine, we use it as the reference compound for both MOR and KOR. For DOR, we utilize ChEMBL494462 as the reference compound, which has a BA value of −11.92 kcal/mol. In applying the molecular generator, a weight coefficient of 0.8 is assigned to the latent vectors of buprenorphine, while the AE latent vector of ChEMBL494462 is given a weight coefficient of 0.2. This weighting scheme increases the likelihood of generating buprenorphine analogues, as the higher weight assigned to buprenorphine promotes a greater resemblance to this reference compound. Our GNC generated millions of novel molecules. From this vast pool, we identified five compounds that were nearly optimal, each possessing Log P values below 5. The synthetic accessibility score (SAS) less than 6 is the suggested proper range. A lower SAS indicates a higher level of ease in synthesizing the compound.

The five compounds have a SAS of less than 5.7, with two scores below 4. Therefore, these five compounds can be synthesized relatively easily. These highly potent compounds on all three receptors are depicted in Figure 9.

Buprenorphine, with BA values being close to −13 kcal/mol on the three receptors, resides at the outer bounds of the BA distributions in the training data. In Figure 9, the identified compounds exhibit predicted BAs around −11 kcal/mol for the receptors. Moreover, they exhibit low predicted hERG side effects and Log P values are below 5. Among the compounds, one has a Log P value below 3, two have Log P values between 3 and 4, and two have Log P values between 4 and 5. In comparison to buprenorphine whose Log P value is 5.56, the five identified compounds possess superior Log P profiles.

The five generated compounds exhibit similarities ranging from 0.57 to 0.74 with buprenorphine. Since novel molecules are generated with potent reference compounds, they can inherit certain moieties from them, which can facilitate effective binding to the three receptors. These new compounds are observed to preserve some functional groups. For instance, buprenorphine contains methyl groups, a benzene ring, trimethylene, and hydroxyl groups. Many of the generated compounds retain these groups, particularly those with higher similarity scores to buprenorphine. Among the five compounds, the one depicted in Figure 9b is the least similar to buprenorphine, with a similarity score of 0.57. However, it introduces two fluorine atoms as new elements while still preserving a benzene ring, several hydroxyl groups, and methyl groups. Notably, it possesses the best Log P profile among the five new compounds. The retention of these functional groups in the five molecules contributes to their binding potency on the receptors.

We apply our Log P optimization strategy to the three compounds depicted in Figure 9f–g, as their Log P values exceed 4. By replacing hydrogen atoms with hydroxyl groups on these three molecules, we generate 14 and 18 new molecules for each respective compound. This optimization approach proves valuable in generating additional nearly optimal compounds based on the molecules in Figure 9f,g. The derived nearly optimal compounds, exhibiting desired binding affinity and ADMET properties, are presented in Figures S2 and S3 in the Supporting Information.

The compound depicted in Figure 9e exhibits the highest similarity to buprenorphine, with a similarity score of 0.74. It demonstrates high potency on all three receptors while exhibiting a low hERG side effect. We also employed Autodock Vina software to predict the molecular interactions with the three receptors. As illustrated in Figure 10, the compound establishes multiple hydrogen bonds with the receptors. In its interaction with the MOR receptor, a hydrogen bond is formed between the oxygen atom on the benzene ring of the molecule and the nitrogen atom on the residue Trp318(A). Regarding its interaction with the KOR receptor, two hydrogen bonds are formed. One is established between an oxygen atom of the molecule and an oxygen atom on the residue Tyr312(A), while the other is formed between an oxygen atom of the molecule and a sulfur atom on the residue Cys210(A). Additionally, the molecule forms two hydrogen bonds with the DOR receptor. One bond is created between an oxygen atom on the molecule and an oxygen atom on the residue

Tyr129(A), while the other bond is formed between an oxygen atom on the molecule and a nitrogen atom on the residue Lys214(A).

3.1.2. Designing Naltrexone Analogues.—Naltrexone has BA values of −12.55, −12.11, and −10.48 kcal/mol on MOR, KOR, and DOR, respectively. It functions as an antagonist for both MOR and KOR. Its KOR antagonist properties have been linked to mood improvements in individuals with OUD.[55] However, it faces challenges regarding low adherence among individuals addicted to opioids.[7] In some studies, the effectiveness of naltrexone treatment was reported to be not encouraging such that there is not much difference between naltrexone and placebo on treatment retention or relapse rates.[56] Despite this, naltrexone still shows benefit in OUD treatment with reduced heroin use and decreased criminal activity.[56] Naltrexone analogues can potentially improve the treatment retention or relapse rates.

We employed our GNC to generate analogues of naltrexone. Naltrexone was used as the reference compound for MOR and KOR, while ChEMBL56585 served as the reference compound for DOR. ChEMBL56585 displayed BA values of −12.26, −13.64, and −12.35 kcal/mol on MOR, KOR, and DOR, respectively. In applying the molecular generator, weight coefficients of 0.8 and 0.2 are assigned to the latent vectors of naltrexone and ChEMBL56585, respectively. From the millions of generated compounds, we identified four analogues that were nearly optimal, as depicted in Figure 11. The four compounds have SAS values close to 4 or 5. Therefore, they can be synthesized easily compared to the above buprenophine analogues.

The similarity of these compounds to naltrexone ranged from 0.61 to 0.76. The least similar compound exhibited molecular novelty while still possessing similar functional groups as naltrexone, such as a benzene ring, hydroxyl group, and other rings. Naltrexone displayed an optimal Log P profile with a value of 2.26, and the resulting five analogues inherited this favorable Log P profile. This emphasizes the importance of selecting reference compounds with desired physicochemical properties, eliminating the need for further optimization to improve the Log P profiles of the derived compounds. The naltrexone analogue shown in Figure 11c demonstrates promising binding effects, with a detailed illustration of its molecular interactions provided in Figure S4 in the Supporting Information.

3.1.3. Designing Methadone Analogues.—Methadone functions as a full agonist on the MOR receptor and can alleviate withdrawal and craving symptoms.[5] Its extended half-life and diminished druglike effects, such as euphoria, result in fewer withdrawal symptoms and reduced potential for reinforcing behavior compared to other opioids.[57] However, methadone carries a risk of respiratory depression in cases of overdose. Methadone is the best option for retaining patients in treatment programs.[4] Methadone analogues with less risk of respiratory depression can be possibly designed.

Methadone demonstrates BA values of −11.84, −8.99, and −8.54 kcal/mol on MOR, KOR, and DOR, respectively. Designing a multitarget methadone analogue presents a challenging task, given that methadone primarily exhibits potency on the MOR receptor. In our molecular generator, methadone serves as the reference compound for MOR and KOR,

while other potent compounds were used as the reference for DOR. To effectively generate analogues of methadone, assigning a high weight coefficient to methadone is necessary. However, this approach results in a limited number of compounds that effectively target all three receptors. Additionally, due to the strict ADMET requirements, none of our generated compounds can be considered nearly optimal.

3.1.4.   Designing Naloxone Analogues.—Naloxone is an opioid antagonist used for reversing respiratory depression in cases of opioid overdose. It exhibits a high affinity that allows it to displace opioid drugs, thereby counteracting their respiratory depression effects. It does not stimulate the opioid receptors and therefore does not cause sedation, analgesia, respiratory depression, and euphoria.[1] Naloxone specifically exhibits BA values of −11.47, −10.96, and −9.8 kcal/mol on the MOR, KOR, and DOR receptors, respectively. Naloxone can reverse opioid-induced respiratory depression in overdose. However, due to its very short half-life of about an hour,[58] naloxone causes patients to revert to the overdosed state once its effects wear off, leading to the need for multiple administrations. Naloxone analogues could have a longer half-life, giving better reversing effect of respiratory depression.

We utilize naloxone as the reference compound for the MOR and KOR receptors, while ChEMBL494462 serves as the reference compound for DOR. Compound ChEMBL494462 demonstrates BA values of −12.26, −13.64, and −12.35 kcal/mol on the MOR, KOR, and DOR receptors, respectively. In applying the molecular generator, weight coefficients of 0.8 and 0.2 are assigned to the latent vectors of naloxone and ChEMBL494462, respectively. From millions of generated compounds, we identified five molecules that exhibit nearly optimal properties, as illustrated in Figure 12. The five compounds, except for compound e, have SAS values close to six. These SAS values indicate a relatively high level of difficulty in synthesizing these compounds.

The high weight of naloxone has a significant impact on the structures of its derivatives. As depicted in Figure 12, there is high degree of similarities between these compounds and naloxone. The smallest similarity value observed with naloxone was 0.72, indicating a strong resemblance in terms of molecular structures and functional groups. All five derivatives contain critical pharmacophores such as methyl groups, hydroxyl groups, benzene rings, and other related ring structures, similar to naloxone. Moreover, these analogues possess optimal Log P profiles, akin to the design of naltrexone analogues. This can be attributed to naloxone's favorable physicochemical properties, which facilitate the derivatives' adherence to ADMET requirements. The naloxone analogue presented in Figure 12c demonstrates promising binding effects, with a detailed depiction of its molecular interactions provided in Figure S4 in the Supporting Information.

3.2.   Key Factors in Designing Optimal Compounds.

We proposed a molecular generator based on stochastic differential equations in the latent space of autoencoder networks. Molecules are represented by the AE latent vectors. Multiple reference compounds are used to guide the design of novel molecules such that the generated compounds will inherit similar structures or function groups from the references.

As demonstrated in the above experiments, our stochastic molecular generator proves to be effective in achieving the goal. Specifically, we aim at designing molecules that are effective on several critical opioid receptors including MOR, KOR, and DOR. The generated compounds are found to share similarities with the selected reference compounds. To generate novel molecules with desired binding affinity and druglike properties, there are several concerns in utilizing our generator.

3.2.1.   Binding Affinity Prediction Reliability Analysis.—We incorporated two layers of BA predictors into our approach. The first layer, AE-BP, is utilized for initial BA screening, while the second layer consists of a consensus BA predictor, which provides more accurate predictions. This two-layered approach helps us identify promising drug candidates.

However, considering the inherent nature of machine learning predictions, we anticipate that the generated potent compounds will exhibit certain similarities to the available the training data. We carefully select reference compounds from the training data with BA values ranging from −10 to −12 kcal/mol. Within the three inhibitor datasets we collected, a considerable number of compounds fall within this specific BA range. If the generated compounds share a certain degree of similarities with the training data, the machine learning models can effectively differentiate these potent compounds from the inactive ones.

3.2.2.   Molecular Novelties.—The molecular novelties are assessed by measuring their similarity to selected reference compounds. The generator can effectively manage the level of novelty by adjusting the number of references or assigning weights to each reference compound. Higher molecular novelties indicate a broader coverage of the chemical space by the generated compounds, as they are different from the reference compounds or the available molecules in the training data. A wide range of chemical space, encompassing high-binding-affinity compounds, is advantageous for drug design. This enables a greater diversity in various druggable properties, such as Log P, Log S, Cano-2, and others. Having more drug candidates with different druggable properties provides a wider range of options for treating patients with varying health conditions. However, it is important to control the novelties of the generated compounds. Higher molecular novelties imply low similarities with the reference compounds or the machine learning training data. To ensure accurate binding affinity predictions, a certain degree of similarity with the available training data is still necessary.

3.2.3.   Importance of Selecting Appropriate References.—When generating novel molecules, we appropriately choose reference compounds from the training data that exhibit the desired potency on specific receptors. Our experiments have demonstrated that appropriate references are beneficial in generating a greater number of candidates with desired BAs. However, it is essential for a promising drug candidate to also meet other crucial druglike properties, particularly the ADMET criteria.

By utilizing reference compounds that demonstrate satisfactory ADMET properties, our generator can generate a substantial pool of compounds that exhibit the desired ADMET characteristics. This is exemplified in the analogue generations for naltrexone and naloxone.

Conversely, when the reference compounds do not meet the essential ADMET criteria, even millions of compounds are generated, they can hardly pass the ADMET screening process. This observation is illustrated in the demonstration example depicted in Figure 4. In this particular example, the majority of the generated compounds do not successfully pass the FDAMDD screening due to the utilization of reference compounds with a low FDAMDD profile.

3.2.4. Agonist/Antagonist for Opioid Receptors.—Approved medications, as mentioned earlier, function as agonists or antagonists on the opioid receptors, playing crucial roles in the treatment of OUD. Methadone and buprenorphine serve as examples of agonist medications that target MOR. Methadone, a long-acting MOR agonist, effectively alleviates withdrawal symptoms and cravings.[5] On the other hand, buprenorphine acts as a partial agonist of MOR, producing milder effects compared to full agonists.[6] Antagonist medications, in contrast, block opioid receptors, thereby preventing the binding of opioids and reducing their reinforcing effects. Buprenorphine, for instance, acts as a KOR antagonist, offering mood improvements for individuals with OUD.[54] Naltrexone and naloxone are antagonists that target all three critical opioid receptors. In particular, naloxone exhibits the highest affinity for MOR and is used to counteract the respiratory and mental depression effects of opioid overdose.[59]

Recent advancements in deep generative research have introduced innovative approaches for the de novo design of improved opioid antagonists,[60] as well as the design for selective KOR antagonists,[61] recognizing the crucial role of antagonists in OUD treatment. Besides, machine learning models have been proposed to predict the agonist or antagonist activities of small molecules on MOR, KOR, and DOR.[62,63] These models offer the potential to prioritize compounds from extensive libraries for subsequent experimental testing.

While it is essential to investigate the agonist/antagonist properties of our generated druglike compound for OUD treatment, including initial machine learning predictions or experimental validations, this study does not encompass that scope. However, we plan to conduct such investigations in the future, either through collaborations with experimentalists or by constructing reliable machine learning predictive models.

4. CONCLUSIONS

We have developed a highly effective deep generative model for generating novel molecules that can be effective on multiple targets, including MOR, KOR, and DOR. The molecular generator is designed by integrating a stochastic differential equation (SDE)-based diffusion approach into the latent space of a pretrained autoencoder model. Through careful selection of appropriate reference compounds and adherence to a series of novelty criteria, a substantial number of novel compounds with desirable binding affinities for MOR, KOR, and DOR, as well as other druglike properties, can be generated.

To predict the binding affinities, we employ advanced machine learning models that integrate autoencoder embeddings, transformer embeddings, and topological Laplacian fingerprints with machine learning algorithms. The incorporation of these diverse molecular

representations enhances the accuracy of the binding affinity predictions. The selection of reference compounds is crucial in two aspects. First, the number of reference compounds influences the novelty of the generated molecules. Second, the use of reference compounds with desired ADMET properties increases the likelihood of generating compounds that satisfy the necessary ADMET requirements.

Extensive experiments have demonstrated the effectiveness of our deep generative models in designing molecules that exhibit structural similarities to known opioid molecules or alternative compounds with therapeutic potential. We utilized our generative network complex to generate a diverse set of druglike molecules, but further experimental studies are needed to evaluate their pharmacological effectiveness for OUD treatment. Our machine learning platform represents a valuable tool in addressing the urgent need for medications in the treatment of OUD. Additionally, our platform has the potential to facilitate the design of molecules that require specific selectivity on multiple targets, making it a promising tool for medication development in various diseases.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

Data Availability Statement

The related datasets studied in this work are available at: https://weilab.math.msu.edu/ DataLibrary/2D/.

ABBREVIATIONS

ADMET	absorption, distribution, metabolism, excretion, and toxicity
AE	autoencoder
AE-BP	autoencoder binding affinity predictor
AE-FP	autoencoder fingerprint
BA	binding affinity
DDPMs	denoising diffusion probabilistic models
DNN	deep neural network
DOR	delta opioid receptor
FDA	U.S. Food and Drug Administration
FDAMDDs	FDA maximum recommended daily dose

GAN	generative adversarial network
GBDT	gradient boosting decision tree
GNC	generative network complex
GRU	gated recurrent unit
HTS	high-throughput screening
KOR	kappa opioid receptor
LSRN	latent space random noise
LSTM	long short-term memory
MOR	mu opioid receptor
MTSGNC	multitarget stochastic generative network complex
R	Pearson correlation coefficients
RL	reinforcement learning
RMSE	root-mean-square error
RNN	recurrent neural network
SAS	synthetic accessibility score
SDE	stochastic differential equation
SGMs	score-based generative models
SMILES	Simplified molecular-input line-entry system
TF-BP	transformer binding-affinity predictor
TF-FP	transformer fingerprint
TL-BP	topology Laplacian binding-affinity predictor
TL-FP	topology Laplacian fingerprint
VAE	variational autoencoder
OUD	opioid use disorder

REFERENCES

1). McLellan AT; Lewis DC; O'brien CP; Kleber HD Drug dependence, a chronic medical illness: implications for treatment, insurance, and outcomes evaluation. Jama 2000, 284, 1689–1695. [PubMed: 11015800]

2). Douaihy AB; Kelly TM; Sullivan C Medications for Substance Use Disorders. In Social Work in Public Health; Taylor & Francis, 2013; Vol. 28, pp 264–278. [PubMed: 23731419]

3). Sofuoglu M; DeVito EE; Carroll KM Pharmacological and behavioral treatment of opioid use disorder. Psychiatr. Res. Clin. Pract 2019, 1, 4–15.

4). Wang S Historical review: opiate addiction and opioid receptors. Cell Transplant 2019, 28, 233–238. [PubMed: 30419763]

5). Fareed A; Vayalapalli S; Steven S; Jennifer C; Karen D; Bailey SP Effect of methadone maintenance treatment on heroin craving, a literature review. J. Addict. Dis 2010, 30, 27–38.

6). Bickel WK; Amass L Buprenorphine treatment of opioid dependence: a review. Exp. Clin. Psychopharmacol 1995, 3, 477.

7). Morgan JR; Schackman BR; Leff JA; Linas BP; Walley AY Injectable naltrexone, oral naltrexone, and buprenorphine utilization and discontinuation among individuals treated for opioid use disorder in a united states commercially insured population. J. Subst. Abuse Treat 2018, 85, 90–96. [PubMed: 28733097]

8). Contet C; Kieffer BL; Befort K Mu opioid receptor: a gateway to drug addiction. Curr. Opin. Neurobiol 2004, 14, 370–378. [PubMed: 15194118]

9). Wassum KM; Cely IC; Balleine BW; Maidment NT μ-opioid receptor activation in the basolateral amygdala mediates the learning of increases but not decreases in the incentive value of a food reward. J. Neurosci 2011, 31, 1591–1599. [PubMed: 21289167]

10). Bruchas MR; Land BB; Chavkin, Ch. The dynorphin/kappa opioid system as a modulator of stress-induced and pro-addictive behaviors. Brain Res 2010, 1314, 44–55. [PubMed: 19716811]

11). Roberts AJ; Gold LH; Polis I; McDonald JS; Filliol D; Kieffer BL; Koob GF Increased ethanol self-administration in δ-opioid receptor knockout mice. Alcoholism 2001, 25, 1249–1256. [PubMed: 11584142]

12). Hughes JP; Rees S; Kalindjian SB; Philpott KL Principles of early drug discovery. Br. J. Pharmacol 2011, 162, 1239–1249. [PubMed: 21091654]

13). Polishchuk PG; Madzhidov TI; Varnek A Estimation of the size of drug-like chemical space based on gdb-17 data. J. Comput.-Aided Mol. Des 2013, 27, 675–679. [PubMed: 23963658]

14). Szyma ski 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. [PubMed: 22312262]

15). Salimi A; Lim JH; Jang JH; Lee JY virtual screening, molecular docking, and molecular dynamics simulations to identify potential vegfr2 kinase inhibitors. Sci. Rep 2022, 12, No. 18825. [PubMed: 36335233]

16). Mensa S; Sahin E; Tacchino F; Barkoutsos PK; Tavernelli I Quantum machine learning framework for virtual screening in drug discovery: a prospective quantum advantage. Mach. Learn 2023, 4, 015023.

17). Gómez-Bombarelli R; Wei JN; Duvenaud D; Miguel Hernández-Lobato J; Sánchez-Lengeling B; Sheberla D; Aguilera-Iparraguirre J; Hirzel TD; Adams RP; Aspuru-Guzik A Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci 2018, 4, 268–276. [PubMed: 29532027]

18). Wang M; Wang Z; Sun H; Wang J; Shen C; Weng G; Chai X; Li H; Cao D; Hou T Deep learning approaches for de novo drug design: An overview. Curr. Opin. Struct. Biol 2022, 72, 135–144. [PubMed: 34823138]

19). Segler MHS; Thierry K; Christian T; Waller MP Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent. Sci 2018, 4, 120–131. [PubMed: 29392184]

20). Grow C; Gao K; Nguyen DD; Wei G-W Generative network complex (gnc) for drug discovery. Commun. Inf. Syst. 2019, 19, 241. [PubMed: 34257523]

21). Prykhodko O; Johansson SV; Kotsias P-C; Arús-Pous J; Bjerrum EJ; Engkvist O; Chen H A de novo molecular generation method using latent vector based generative adversarial network. J. Cheminf 2019, 11, No. 74.

22). Vignac C; Krawczuk I; Siraudin A; Wang B; Cevher V; Frossard P Digress: Discrete Denoising Diffusion for Graph Generation. 2022, arXiv:2209.14734. arXiv.org e-Print archive http://arxiv.org/abs/2209.14734.

23). Niu C; Song Y; Song J; Zhao S; Grover A; Ermon S In *Permutation Invariant Graph Generation via Score-Based Generative* Modeling, International Conference on Artificial Intelligence and Statistics, PMLR, 2020; pp 4474–4484.

24). Huang H; Sun L; Du B; Fu Y; Lv W Graphgdp: Generative Diffusion Processes for Permutation Invariant Graph Generation. 2022, arXiv:2212.01842. arXiv.org e-Print archive http://arxiv.org/abs/2212.01842.

25). Zhang M; Qamar M; Kang T; Jung Y; Zhang C; Bae S-H; Zhang C A Survey on Graph Diffusion Models: Generative AI in Science for Molecule, Protein and Material. 2023, arXiv:2304.01565. arXiv.org e-Print archive http://arxiv.org/abs/2304.01565.

26). Igashov I; Stärk H; Vignac C; Satorras VG; Frossard P; Welling M; Bronstein M; Correia B Equivariant 3d-Conditional Diffusion Models for Molecular Linker Design. 2022, arXiv:2210.05274. arXiv.org e-Print archive http://arxiv.org/abs/2210.05274.

27). Schneuing A; Du Y; Harris C; Jamasb A; Igashov I; Du W; Blundell T; Lió P; Gomes C; Welling M et al. Structure-Based Drug Design with Equivariant Diffusion Models.2022, arXiv:2210.13695. arXiv.org e-Print archive http://arxiv.org/abs/2210.13695.

28). Vahdat A; Kreis K; Kautz J Score-based generative modeling in latent space. Adv. Neural Inf. Process 2021, 34, 11287–11302.

29). 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. [PubMed: 30842833]

30). Winter R; Montanari F; Steffen A; Briem H; Noé F; Clevert D-A Efficient multi-objective molecular optimization in a continuous latent space. Chem. Sci 2019, 10, 8016–8024. [PubMed: 31853357]

31). Timm D; Möller C Early identification of herg liability in drug discovery programs by automated patch clamp. Front. Pharmacol 2014, 5, No. 00203.

32). Kalliokoski T; Kramer C; Vulpetti A; Gedeck P Comparability of mixed ic50 data–a statistical analysis. PLoS One 2013, 8, e61007. [PubMed: 23613770]

33). Zixuan C; Wei G-W Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. Int. J. Numer. Method Biomed. Eng 2018, 34, No. e2914.

34). Wang R; Fang X; Lu Y; Yang C-Y; Wang S The pdbbind database: methodologies and updates. J. Med. Chem 2005, 48, 4111–4119. [PubMed: 15943484]

35). Gao K; Nguyen DD; Tu M; Wei G-W Generative network complex for the automated generation of drug-like molecules. J. Chem. Inf. Model 2020, 60, 5682–5698. [PubMed: 32686938]

36). Dormand JR; Prince PJ A family of embedded runge-kutta formulae. J. Comput. Appl. Math 1980, 6, 19–26.

37). Feng H; Wei G-W Virtual screening of drugbank database for herg blockers using topological laplacian-assisted ai models. Comput. Biol. Med 2023, 153, 106491. [PubMed: 36599209]

38). Chen D; Zheng J; Wei G-W; Pan F Extracting predictive representations from hundreds of millions of molecules. J. Phys. Chem. Lett 2021, 12, 10793–10801. [PubMed: 34723543]

39). Wang R; Nguyen DD; Wei G-W Persistent spectral graph. Int. J. Numer. Method Biomed. Eng 2020, 36, e3376. [PubMed: 32515170]

40). Gao K; Nguyen DD; Sresht V; Mathiowetz AM; Mathiowetz AM; Tu M; Tu M; Wei G-W Are 2d fingerprints still valuable for drug discovery? Phys. Chem. Chem. Phys 2020, 22, 8373–8390. [PubMed: 32266895]

41). Feng H; Jiang J; Wei G-W Machine-learning repurposing of drugbank compounds for opioid use disorder. Comput. Biol. Med 2023, 160, 106921. [PubMed: 37178605]

42). Feng H; Elladki R; Jiang J; Wei G-W Machine-learning analysis of opioid use disorder informed by mor, dor, kor, nor and zorbased interactome networks. Comput. Biol. Med 2023, 157, 106745. [PubMed: 36924727]

43). Mischaikow K; Kaczynski T; Mrozek M Computational Homology; Springer-Verlag: New York, 2004; Vol. 157.

44). Zomorodian A; Carlsson G et al. In Computing Persistent Homology, Proceedings of the Twentieth Annual Symposium on Computational Geometry, 2004; pp 347–356.

45). Chen J; Qiu Y; Wang R; Wei G-W Persistent laplacian projected omicron ba. 4 and ba. 5 to become new dominating variants. Comput. Biol. Med 2022, 151, 106262. [PubMed: 36379191]

46). Jiang P; Chi Y; Li X-S; Meng Z; Liu X; Hua X-S; Xia K Molecular persistent spectral image (mol-psi) representation for machine learning models in drug design. Briefings Bioinf 2022, 23, bbab527.

47). Meng Z; Xia K Persistent spectral-based machine learning (perspect ml) for protein-ligand binding affinity prediction. Sci. Adv 2021, 7, eabc5329. [PubMed: 33962954]

48). Liu X; Feng H; Wu J; Xia K Persistent spectral hypergraph based machine learning (psh-ml) for protein-ligand binding affinity prediction. Briefings Bioinf 2021, 22, bbab127.

49). Xiong G; Wu Z; Yi J; Fu L; Yang Z; Hsieh C; Yin M; Zeng X; Wu C; Lu A; et al. Admetlab 2.0: an integrated online platform for accurate and comprehensive predictions of admet properties. Nucleic Acids Res 2021, 49, 5–14.

50). Lei T; Li Y; Song Y; Li D; Sun H; Hou T Admet evaluation in drug discovery: 15. accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling. J. Cheminf 2016, 8, 1–19.

51). Lipinski CA; Lombardo F; Dominy BW; Feeney PJ Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Delivery Rev 2012, 64, 4–17.

52). Trott O; Olson AJ Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem 2010, 31, 455–461. [PubMed: 19499576]

53). Bell J Pharmacological maintenance treatments of opiate addiction. Br. J. Clin. Pharmacol 2014, 77, 253–263. [PubMed: 23210630]

54). Mattick RP; Breen C; Kimber J; Davoli M Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence. Cochrane Database Syst. Rev 2014, 2, No. D002207.

55). Weerts EM; Kim YK; Wand GS; Dannals RF; Lee JS; Frost JJ; McCaul ME Differences in δ-and μ-opioid receptor blockade measured by positron emission tomography in naltrexone-treated recently abstinent alcohol-dependent subjects. Neuropsycho-pharmacology 2008, 33, 653–665.

56). Minozzi S; Amato L; Vecchi S; Davoli M; Kirchmayer U; Verster A Oral naltrexone maintenance treatment for opioid dependence. Cochrane Database Syst. Rev 1996, 4, No. D001333.

57). Brown R; Kraus C; Fleming M; Reddy S Methadone: applied pharmacology and use as adjunctive treatment in chronic pain. Postgrad. Med. J 2004, 80, 654–659. [PubMed: 15537850]

58). Rzasa Lynn R; Galinkin JL Naloxone dosage for opioid reversal: current evidence and clinical implications. Ther. Adv. Drug Saf 2018, 9, 63–88. [PubMed: 29318006]

59). Algera MH; Kamp J; van der Schrier R; van Velzen M; Niesters M; Aarts L; Dahan A; Olofsen E Opioid-induced respiratory depression in humans: a review of pharmacokinetic-pharmacodynamic modelling of reversal. Br. J. Anaesth 2019, 122, e168–e179. [PubMed: 30915997]

60). Salas-Estrada L; Provasi D; Qiu X; Kaniskan HU; Huang X-P; DiBerto J; Lamim Ribeiro JM; Jin J; Roth BL; Filizola M De novo design of κ-opioid receptor antagonists using a generative deep learning framework. bioRxiv 2023, No. 2023–04.

61). Deng J; Yang Z; Li Y; Samaras D; Wang F Towards Better Opioid Antagonists Using Deep Reinforcement Learning. 2020, arXiv:2004.04768. arXiv.org e-Print archive http://arxiv.org/abs/2004.04768.

62). Sakamuru S; Zhao J; Xia M; Hong H; Simeonov A; Vaisman I; Huang R Predictive models to identify small molecule activators and inhibitors of opioid receptors. J. Chem. Inf. Model 2021, 61, 2675–2685. [PubMed: 34047186]

63). Jia X; Ciallella HL; Russo DP; Zhao L; James MH; Zhu H Construction of a virtual opioid bioprofile: a data-driven qsar modeling study to identify new analgesic opioids. ACS Sustainable Chem. Eng 2021, 9, 3909–3919. [PubMed: 34239782]
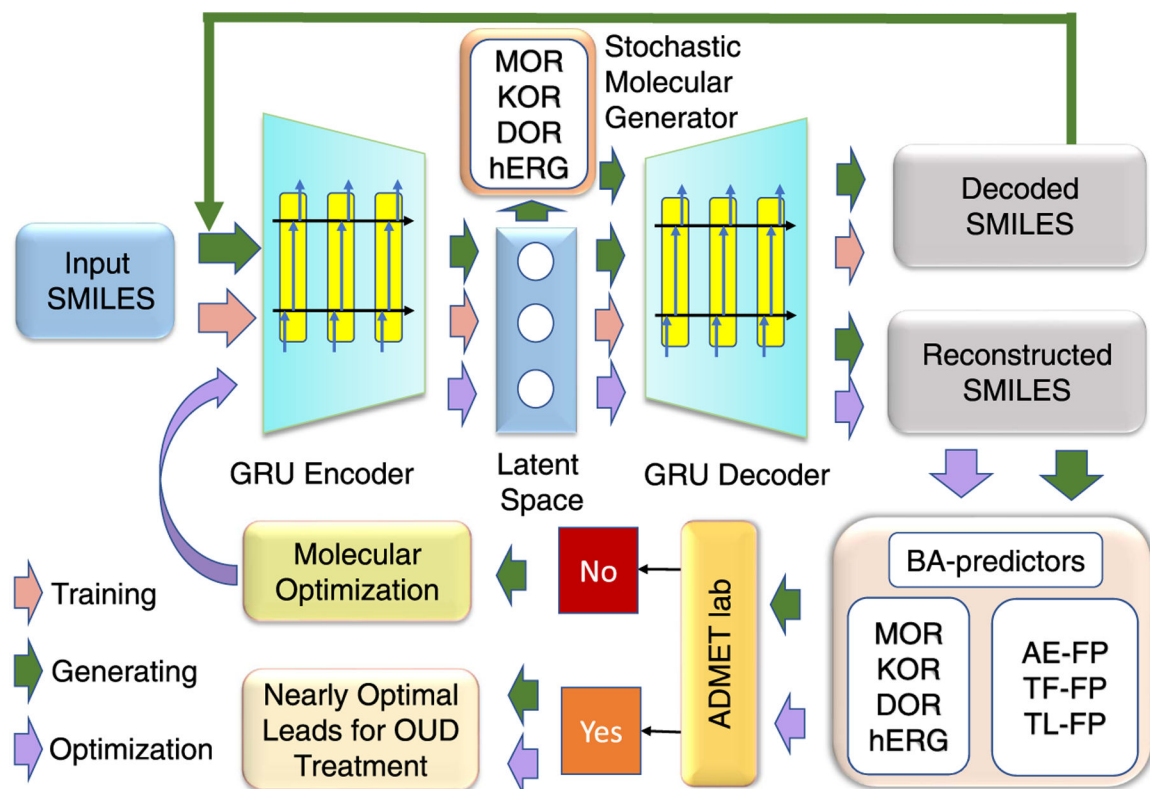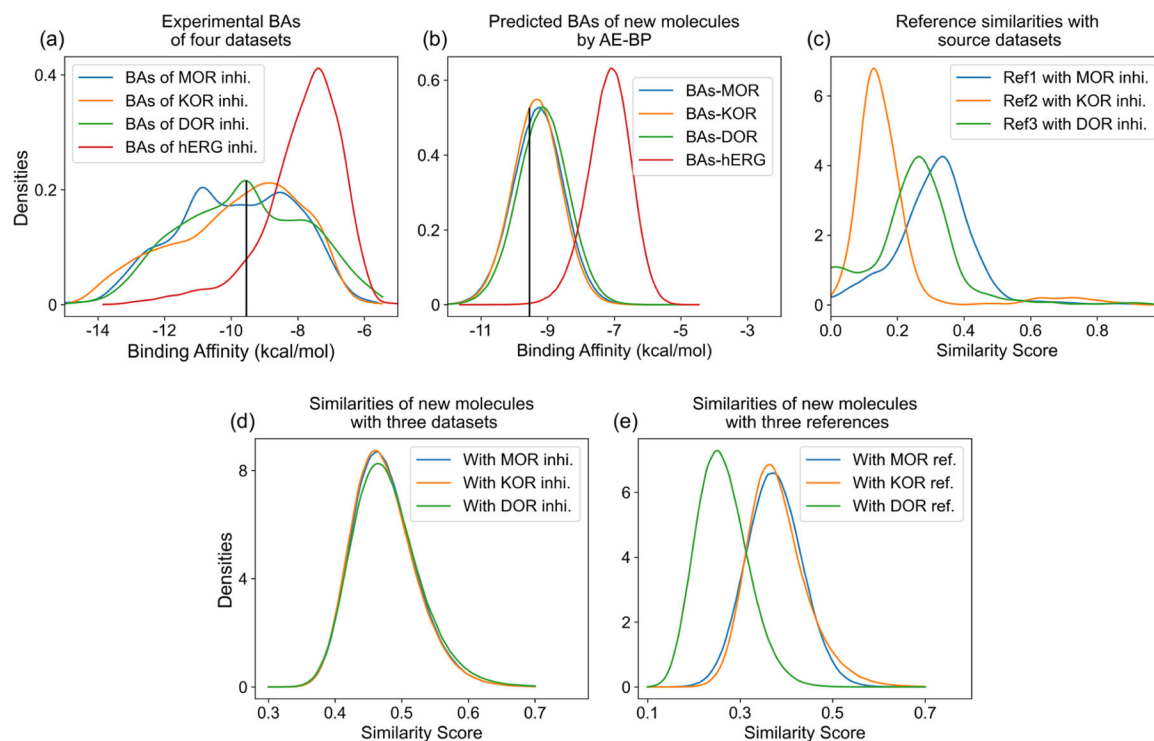
Figure 1.
Schematic illustration of our multitarget stochastic generative network complex, used to design novel compounds for the treatment of OUD. Three different paths, i.e., model training, molecular generation, and lead optimization are colored in pink, green, and purple, respectively. The pink path represents the model training process of the pretrained autoencoder network. The blue path signifies the compound generation process: the SMILES string of a given compound undergoes encoding in the encoding network, and its latent space representation is then fed into the stochastic molecular generator. The generated compounds are subsequently processed through the decoding–encoding network. The resulting molecules that have stable SMILESs are evaluated for their binding affinity (BA) and ADMET properties. Molecules exhibiting desired BA and ADMET properties are regarded as nearly optimal leads. In cases where the properties are not satisfactory, molecular optimization (the purple path) is performed to generate more potential druglike compounds.

Figure 2.
Analysis of binding affinity (BA) distribution of the training data and the generated compounds, as well as the analysis of similarity score distributions. (a) BA distributions of the inhibitors in the four training datasets. The unit for BA is kcal/mol. (b) BA distributions of the generated molecules predicted by AE-BPs. (c) Distributions of similarity scores between reference compounds and corresponding source inhibitor dataset. (d) Distributions of similarity scores between generated compounds and inhibitor datasets for the three opioid receptors. (e) Distributions of similarity scores between generated compounds and three reference compounds.
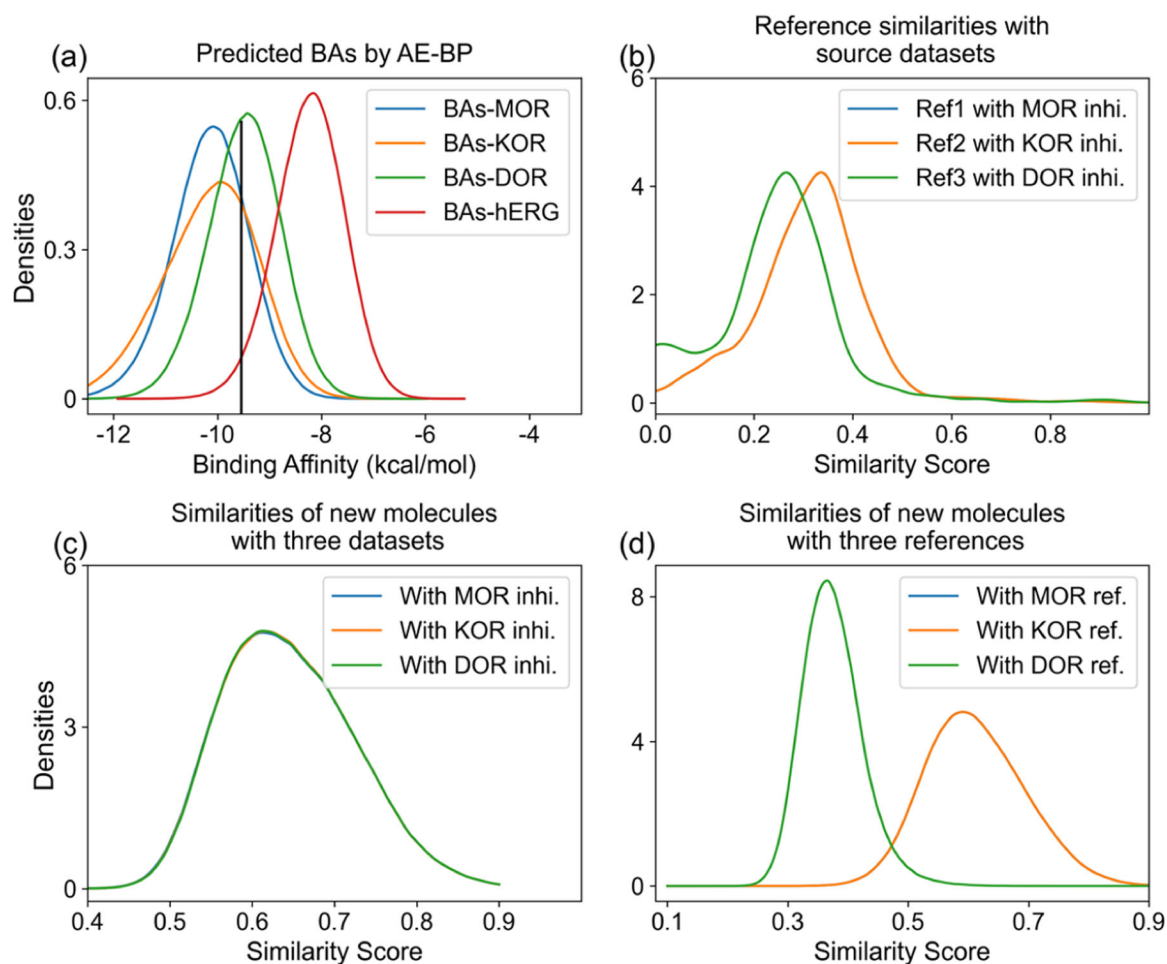
Figure 3.
BA distribution of the generated compounds, as well as the similarity score distributions regarding two reference compounds, training data, and generated compounds. (a) BA distributions of the generated molecules predicted by AE-BPs. The unit for BA is kcal/mol. (b) Distributions of similarity scores between reference compounds and corresponding source inhibitor dataset. (c) Distributions of similarity scores between generated compounds and three inhibitor datasets. (d) Distributions of similarity scores between generated compounds and three reference compounds.
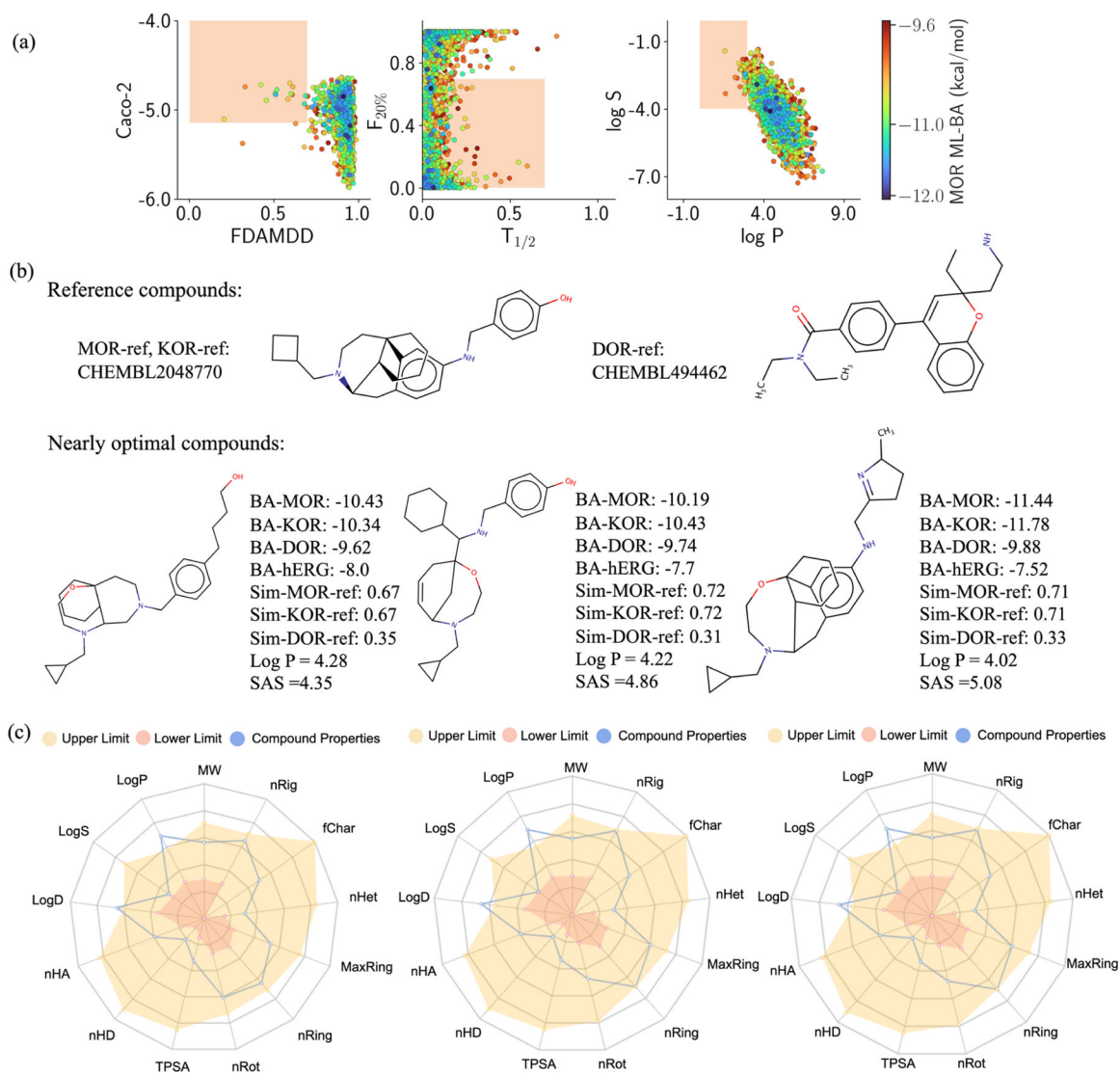
Figure 4.

Identifying nearly optimal compounds. (a) ADMET screening of the multitarget molecules.
(b) Three nearly optimal compounds that satisfy the BA constraint, Log P, SAS, and
ADMET properties. Their BAs and similarity scores with the two references are presented.
(c) Additional ADMET prediction from ADMETlab2. More abbreviations: MW (molecular
weight), log P (log of octanol/water partition coefficient), log S (log of the aqueous
solubility), log D (log P at physiological pH 7.4), nHA (number of hydrogen bond
acceptors), nHD (number of hydrogen bond donors), TPSA (topological polar surface area),
nRot (number of rotatable bonds), nRing (number of rings), MaxRing (number of atoms in
the biggest ring), nHet (number of heteroatoms), fChar (formal charge), and nRig (number
of rigid bonds). The optimal ranges of these indexes are shown in Table S4 in the Supporting
Information.
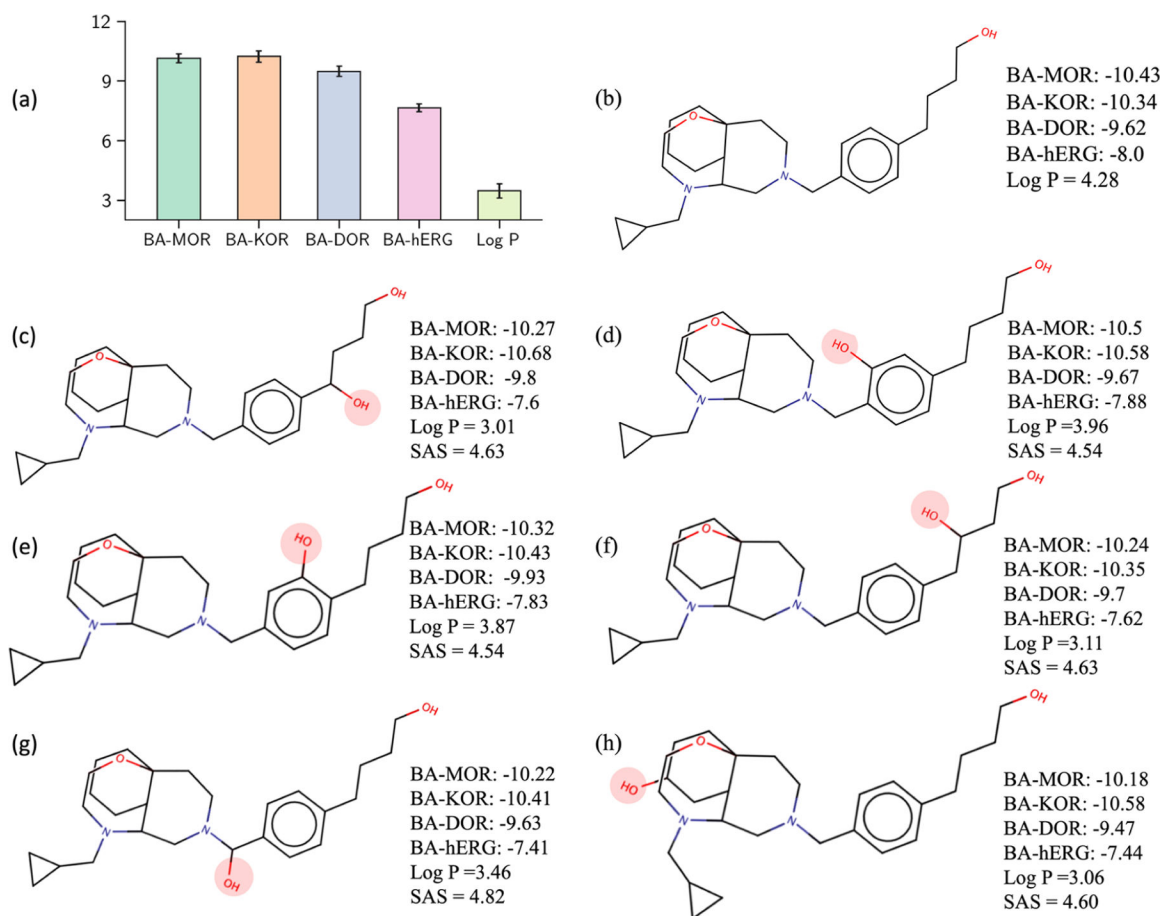
Figure 5.
Results of Log P optimization for one nearly optimal compound. The *y*-axis indicates the magnitude of the BA and Log P value. The BA values are all negative and the Log P values are all positive. (a) Statistics of the predicted BA values on the four critical targets and the predicted log P values of the induced 22 compounds. (b) Original generated compound from our GNC that is considered to be a nearly optimal compound. (c–h) Six of the 22 derived compounds that have high BAs and improved Log P values.
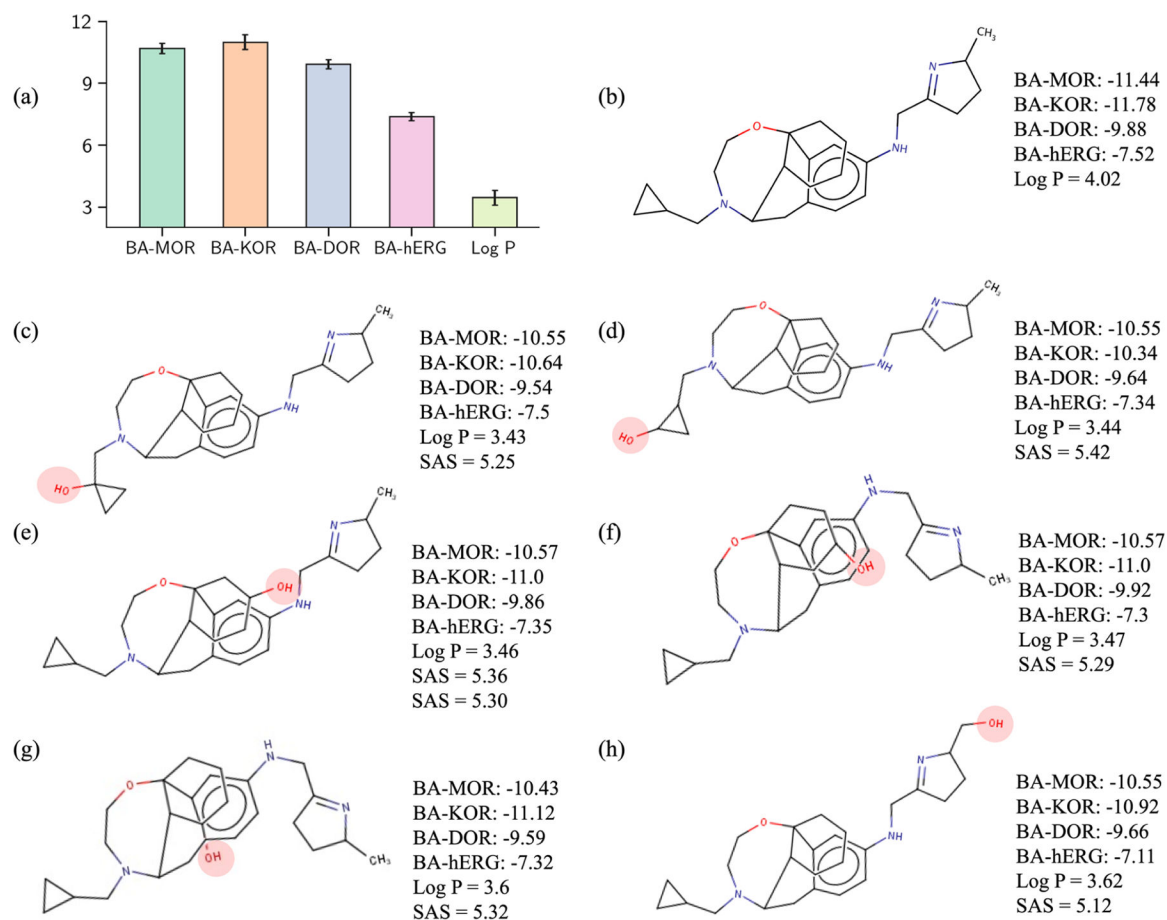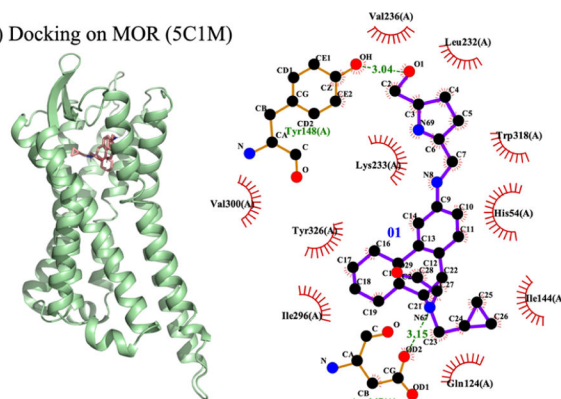
Figure 6.
Results of Log P optimization for one nearly optimal compound. (a) Statistics of the predicted BA values on the four critical targets and the predicted log P values of the derived 22 compounds. The *y*-axis indicates the magnitude of the BA and Log P value. The BA values are all negative, and the Log P values are all positive. (b) Original generated compound from our GNC that is considered to be a nearly optimal compound. (c–h) Six of the 22 derived compounds that have high BAs and improved Log P values.

(a)

BA-MOR: -10.55
BA-KOR: -11.92
BA-DOR: -9.66
BA-hERG: -7.11
Log P = 3.62
SAS = 5.25

(b)

BA-MOR: -10.55
BA-KOR: -10.34
BA-DOR: -9.64
BA-hERG: -7.43
Log P = 3.44
SAS = 5.42

(c) Upper Limit ● Lower Limit ● Compound Properties

(d) Upper Limit ● Lower Limit ● Compound Properties
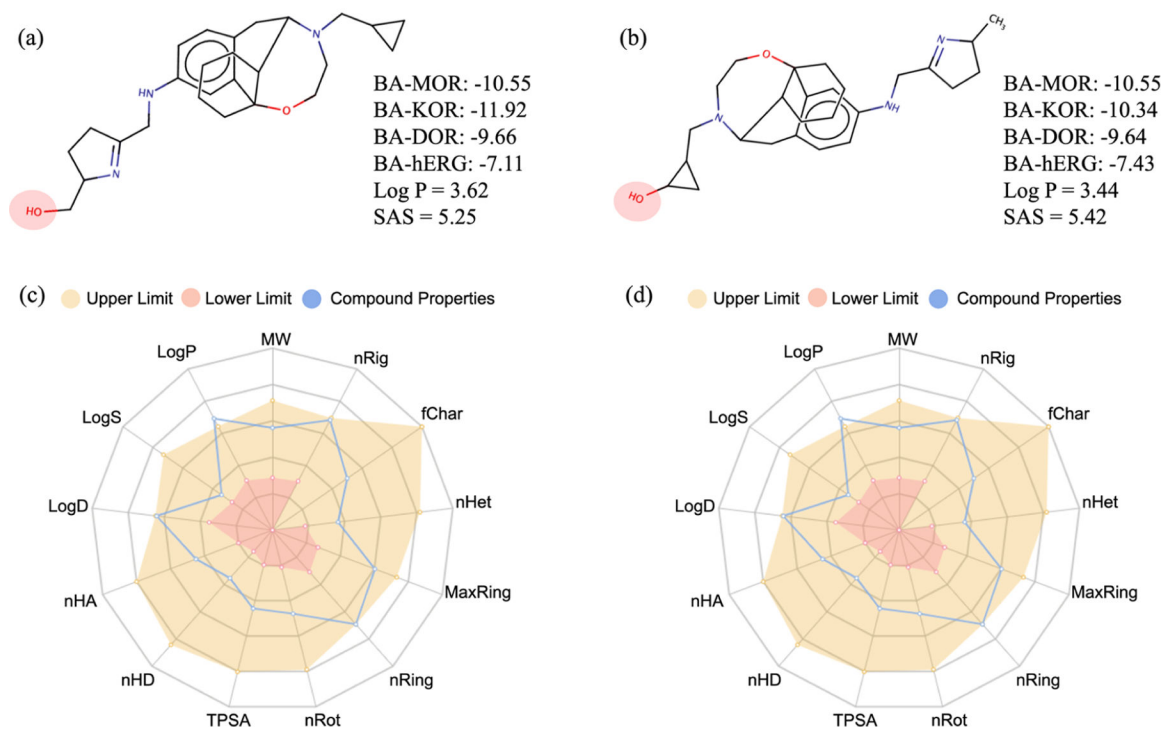
Figure 7.
Two nearly optimal compounds derived from the Log P optimization based on one nearly optimal compound. (a, b) Two-dimensional (2D) structures of the two nearly optimal compounds. The predicted BA values as well as their predicted Log P and SAS are shown. (c, d) ADMET predictions of the two compounds according to ADMETlab server.
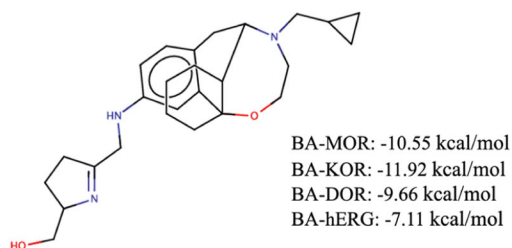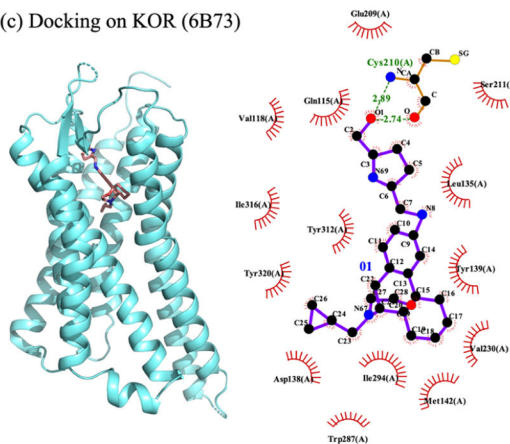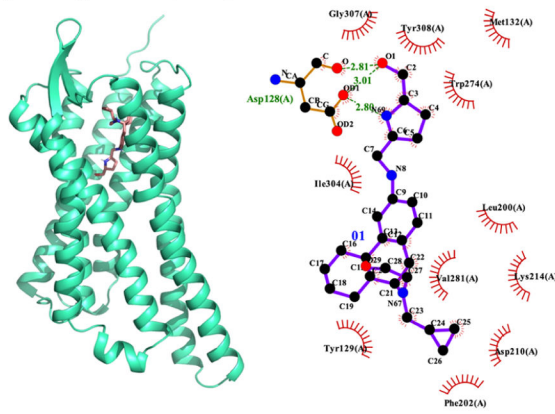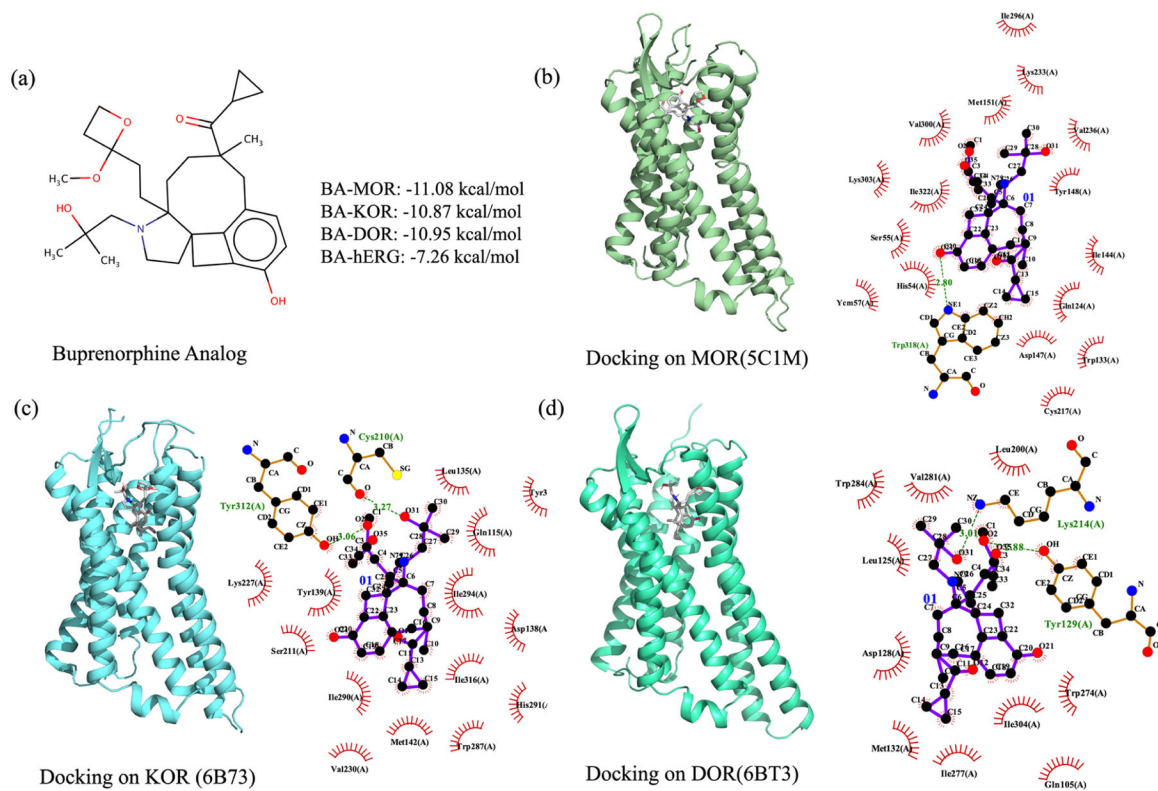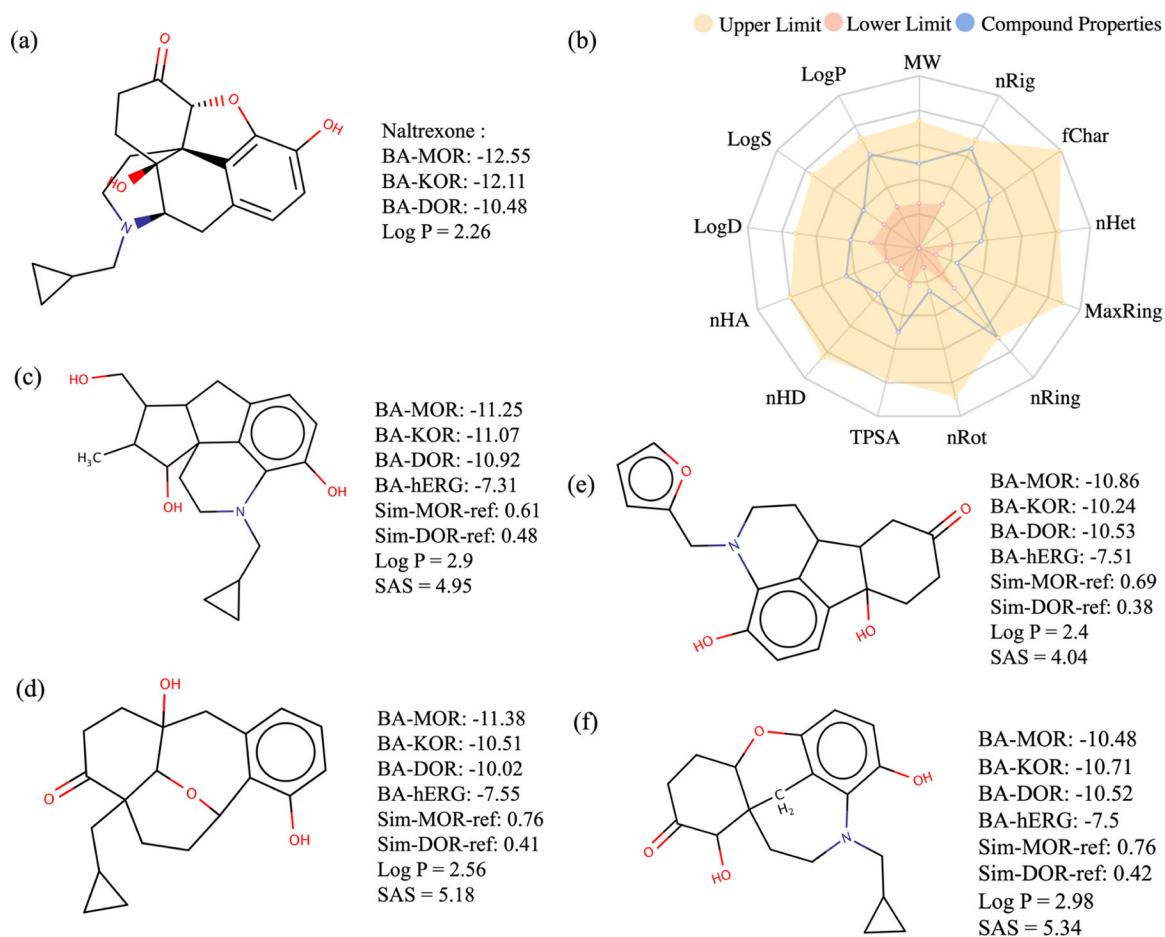
Figure 8.
(a) 2D structure of the optimized compound with the predicted BA values on the four critical targets. (b–d) Compound's docking poses and interactions with the three receptors: MOR (PDB ID: 5C1M), KOR (PDB ID: 6B73), and DOR (PDB ID: 6BT3).
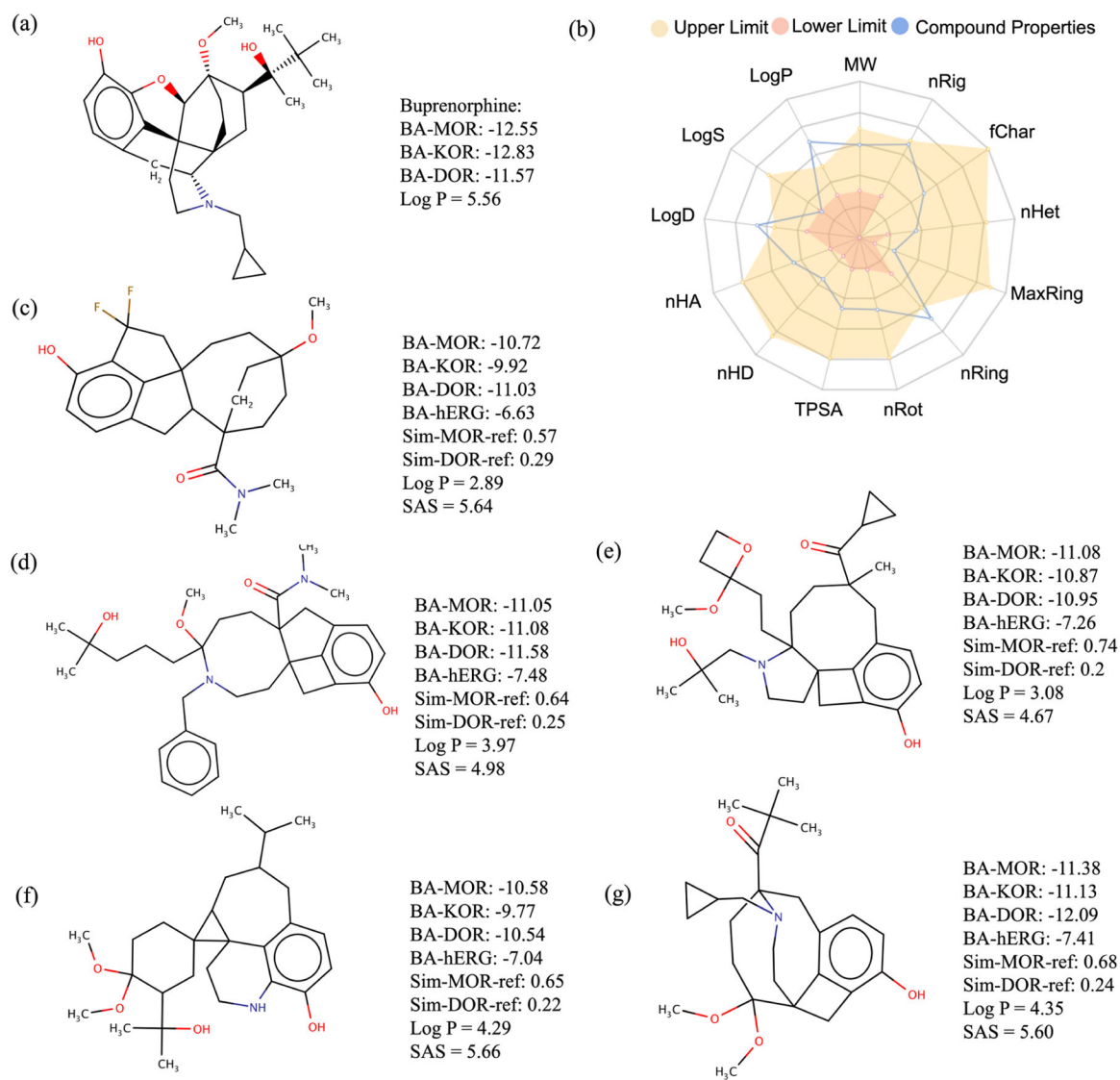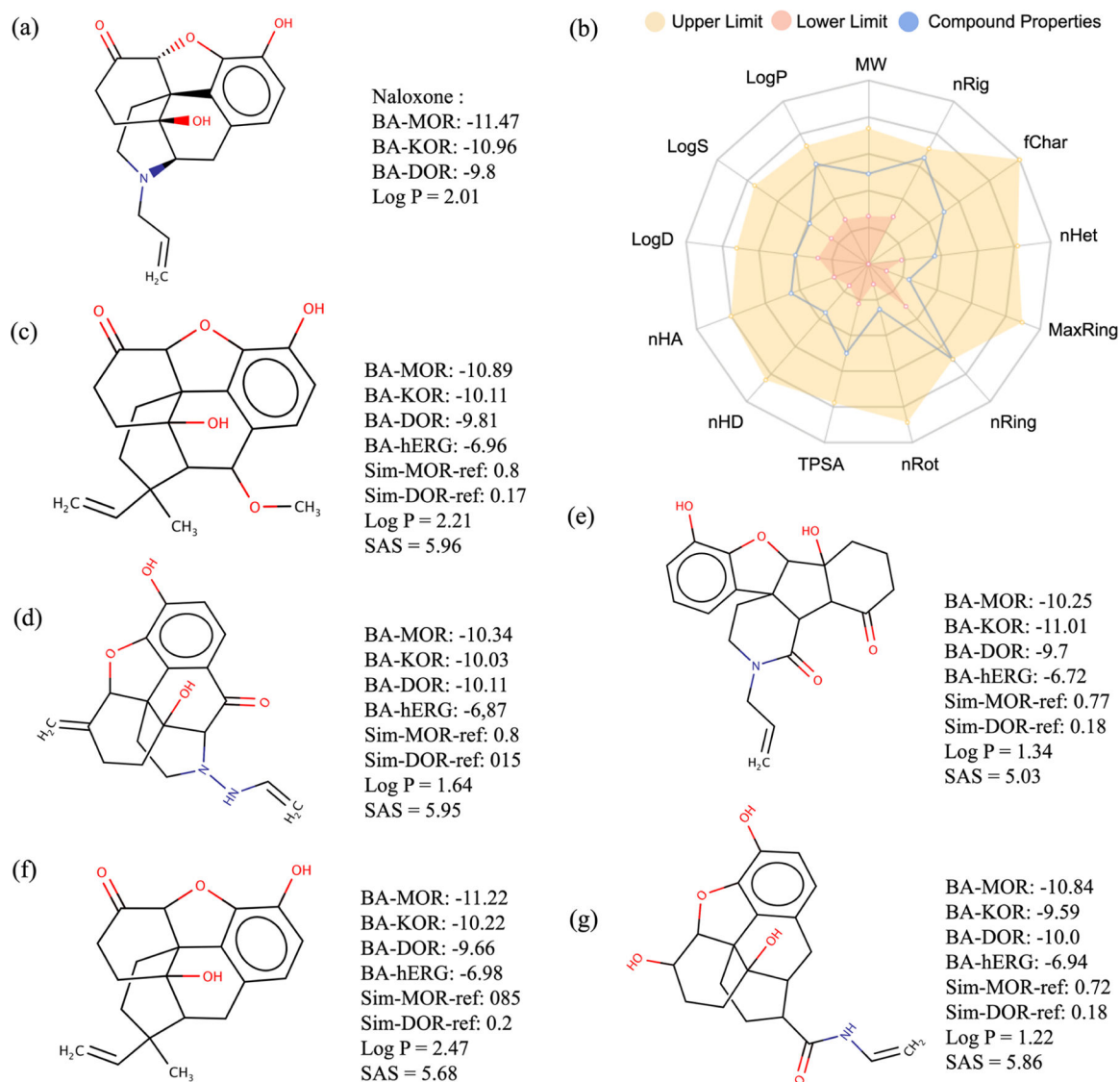
Figure 9.
(a) 2D structure of buprenorphine with the predicted BA values on the four critical targets.
(b) ADMET predictions of buprenorphine according to ADMETlab server. (c–g) Several
nearly optimal buprenorphine analogues that could be effective on MOR, KOR, and DOR.
The predicted BA values on the three receptors and hERG, the similarity scores to reference
compounds, as well as their predicted Log P and SAS values, were provided.

Figure 10.
(a) 2D structure of one nearly optimal buprenorphine analogue in Figure 9e. (b–d) Molecular docking poses and interactions of the buprenorphine analogue with three opioid receptors: MOR (PDB ID: 5C1M), KOR (PDB ID: 6B73), and DOR (PDB ID: 6BT3).

Figure 11.
(a) 2D structure of naltrexone. (b) ADMET predictions of naltrexone according to ADMETlab server. (c–f) Several nearly optimal naltrexone analogues that could be effective on the multiple receptors. The predicted BA values on the critical targets, the similarity scores to reference compounds, as well as their predicted Log P and SAS values were provided.

Figure 12.
(a) 2D structure of naloxone. (b) ADMET predictions of naloxone according to ADMETlab server. (c–g) Several nearly optimal naloxone analogues that could be effective on multiple receptors. The predicted BA values on the critical targets, the similarity scores to reference compounds, as well as their predicted Log P and SAS values were provided.

Table 1.

Optimal Ranges of Six Selected ADMET Properties and Synthesizability (SAS) Used to Screen Nearly Optimal Compounds

property	optimal range
FDAMDD	excellent: 0–0.3; medium: 0.3–0.7; poor: 0.7–1.0
$F_{20\%}$	excellent: 0–0.3; medium: 0.3–0.7; poor: 0.7–1.0
Log P	the proper range: 0–3 log mol/L
Log S	the proper range: −4–0.5 log mol/L
$T_{1/2}$	excellent: 0–0.3; medium: 0.3–0.7; poor: 0.7–1.0
Caco-2	the proper range: >−5.15
SAS	the proper range: <6