OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae690

https://doi.org/10.1093/bib/bbae690

Case Study

STNGS: a deep scaffold learning-driven generation and screening framework for discovering potential novel psychoactive substances

Dongping Liu[1,‡], Dinghao Liu[1,‡], Kewei Sheng[1,‡], Zhenyong Cheng[1], Zixuan Liu[1], Yanling Qiao[2,3,4], Shangxuan Cai[5,6], Yulong Li[5,6], Jubo Wang[4], Hongyang Chen[7], Chi Hu[8], Peng Xu[2,3,4,*], Bin Di[3,4,*], Jun Liao[1,3,7,*]

[1]School of Science, China Pharmaceutical University, Nanjing 211198, China
[2]Key Laboratory of Drug Monitoring and Control, Drug Intelligence and Forensic Center, Ministry of Public Security, Beijing 100193, China
[3]Office of China National Narcotics Control Commission, China Pharmaceutical University Joint Laboratory on Key Technologies of Narcotics Control, Beijing 100193, China
[4]School of Pharmacy, China Pharmaceutical University, Nanjing 211198, China
[5]State Key Laboratory of Membrane Biology, Peking University School of Life Sciences, Beijing 100871, China
[6]PKU-IDG/McGovern Institute for Brain Research, Beijing 100871, China
[7]Zhejiang Lab, Hangzhou 311500, China
[8]School of Engineering, China Pharmaceutical University, Nanjing 211198, China

*Corresponding authors. Peng Xu, Key Laboratory of Drug Monitoring and Control, Drug Intelligence and Forensic Center, Ministry of Public Security, Office of China National Narcotics Control Commission, China Pharmaceutical University Joint Laboratory on Key Technologies of Narcotics Control and School of Pharmacy, China Pharmaceutical University, Nanjing 211198, China. Tel.: +025-86185328; E-mail: pengxu750@163.com; Bin Di, Office of China National Narcotics Control Commission, China Pharmaceutical University Joint Laboratory on Key Technologies of Narcotics Control and School of Pharmacy, China Pharmaceutical University, Nanjing 211198, China. Tel.:+86-25-83271202; E-mail: dibin@cpu.edu.cn; Jun Liao, Office of China National Narcotics Control Commission, China Pharmaceutical University Joint Laboratory on Key Technologies of Narcotics Control, Zhejiang Lab and School of Science, China Pharmaceutical University, Nanjing 211198, China. Tel.: +025-86185160; E-mail: liaojun@cpu.edu.cn

‡Dongping Liu, Dinghao Liu and Kewei Sheng are Co-first authors.

Abstract

The supervision of novel psychoactive substances (NPSs) is a global problem, and the regulation of NPSs was heavily relied on identifying structural matches in established NPSs databases. However, violators could circumvent legal oversight by altering the side chain structure of recognized NPSs and the existing methods cannot overcome the inaccuracy and lag of supervision. In this study, we propose a scaffold and transformer-based NPS generation and Screening (STNGS) framework to systematically identify and evaluate potential NPSs. A scaffold-based generative model and a rank function with four parts are contained by our framework. Our generative model shows excellent performance in the design and optimization of general molecules and NPS-like molecules by chemical space analysis and property distribution analysis. The rank function includes synthetic accessibility score and frequency score, as well as confidence score and affinity score evaluated by a neural network, which enables the precise positioning of potential NPSs. Applied STNGS framework with molecular docking and a G protein-coupled receptor (GPCR) activation-based sensor (GRAB), we successfully identify three novel synthetic cannabinoids with activity. STNGS constrains the chemical space to generate NPS-like molecules database with diversity and novelty, which assists in the ex-ante regulation of NPSs.

Keywords: deep scaffold learning; generative framework; ensemble learning; novel psychoactive substance; synthetic cannabinoids

Introduction

Novel psychoactive substances (NPSs), also known as "designer drugs" or "laboratory drugs" [1, 2], are drug analogs that have been chemically altered by illicit actors with the intention of evading law enforcement targeting controlled substances. These substances produce similar or increased levels of euphoria, hallucinations, and narcotic effects when compared to controlled drugs [3]. Regarding the documented NPSs, the majority are stimulants, followed by synthetic cannabinoid receptor agonists. The secrecy of NPSs, combined with the common practice of modifying existing structures to circumvent legal restrictions [4], greatly hinders the identification and examination of these substances by law enforcement officials. Therefore, identifying potential NPSs proactively and improving existing drug databases can prevent issues and improve the effective-ness of anti-drug agencies in detecting and addressing NPS-related offenses.

The conventional experimental synthesis approach for NPSs is slow and inefficient [5]. Therefore, the rapid and cost-effective acquisition of potential NPS molecules for research is a pressing concern. The application of deep learning techniques to generate potential NPSs and analogs has the potential to provide valuable insights to researchers and regulators. Deep learning has achieved

Received: September 25, 2024. **Revised:** November 23, 2024. **Accepted:** December 19, 2024

significant milestones in the domains of computer vision [6] and natural language processing. A deep generative model is a type of deep learning model that randomly generates samples by capturing the probability density inherent in observable data [7]. Skinnider *et al*. [8] proposed the DarkNPS methodology. This approach employed a deep generative model to understand the structural distributions within the HighResNPS database [9]. The molecules generated through model sampling serve as structural priors for elucidating the structures of other molecules. Notably, this method effectively identified the majority of recently discovered NPS molecules during validation. However, this approach had certain limitations. Firstly, the study lacked constraints on the conditions for sampling chemical space in the deep generative model. Moreover, the validation process entails sampling billions of molecules, which can result in a significant number of false positives. Zhang *et al*. [10] similarly proposed an approach involving the application of deep generative models (SeqGAN and MolGPT) to synthesize fentanyl analog molecules. The study had expanded the fentanyl analog repository and broadened the investigative scope within the realm of fentanyl analog regulation. However, it exclusively relied on data related to fentanyl and its analogs, limiting a comprehensive exploration of the chemical space surrounding fentanyl. None of these studies effectively address the challenge of generating true positives for NPS molecules, leading to numerous model-generated molecules misclassified as NPS. This may hinder the accurate ancillary supervision of NPSs.

This study presents a new molecular generative framework called scaffold and transformer-based NPS generate and screening (STNGS) (Fig. 1a), which acknowledges the crucial role of molecular scaffolds in drug design and optimization. STNGS can be employed to capture the interconnections between scaffold components, and the incorporation of scaffold-type encoding facilitates enhanced comprehension of the mapping relationship between scaffolds and intact molecules. We have developed a ranking function to assess and order the generated molecules. The ranking function includes four components: (i) the synthetic accessibility (SA) score [11]; (ii) the frequency score; (iii) the confidence score; (iv) the affinity score. We employ a genetically encoded endocannabinoid (eCB) sensor [12] and surface plasmon resonance method for affnity assay. Compounds 26 and 31 exhibit superior agonist activity compared to the known NPS JWH-018. The results demonstrate that high true-positive potential NPS candidates are effectively generated and screened by our framework. The resulting database of potential NPS alerts will assist in the preemptive regulation of NPSs.

Methods and materials

Training data

This study utilizes data pertaining to two discrete components: the NPS molecular generator and the NPS discriminator. Accordingly, the data employed for each of these components will be elaborated upon individually.

HighResNPS is a global database for screening novel psychoactive substances (NPS), contributed by forensic laboratories worldwide. It compiles data on newly detected NPS from biological samples and seizures. As the most comprehensive and up-to-date database on NPS structures, HighResNPS is selected to train our NPS generator, enabling the model to effectively capture the chemical characteristics and explore the potential chemical space of NPS molecules. In this study, data collected before April 2020 was used as the training set for the NPS molecular generation model, while data collected after April 2020 served as the external validation set. Additionally, the study excluded the consideration of chiral representations of molecules; thus, all chiral specifications are removed from the SMILES. Only molecules containing carbon, hydrogen, oxygen, nitrogen, sulfur, phosphorus, fluorine, chlorine, bromine, and iodine atoms are retained for subsequent analysis. Following the completion of the aforementioned data cleansing process, a definitive NPS dataset is obtained, containing 2154 NPS molecules.

The purpose of the NPS discriminator is to classify generated molecules. Consequently, the training data for the NPS discriminator must encompass both positive and negative instances of NPS. Positive samples are drawn from the previously mentioned NPS dataset comprising 2154 NPS molecules. Conversely, for negative samples, to ensure equilibrium within the dataset, this inquiry handpicked 2154 compounds from the antitumor dataset AID248 within the PubChem database [13]. These compounds are incorporated as adversarial instances for training the NPS discriminator (Supplementary Fig. S1, Supplementary Fig. S2 and Supplementary Analyses 2).

NPS generator

NPS generator is comprised of two distinct components: T-Scaffold and Mol-GPT (Fig. 1b). T-Scaffold is responsible for capturing the interconnections between scaffold components, and its output, which is represented by a maximum likelihood objective, serves as the input for the subsequent step of Mol-GPT. The incorporation of scaffold-type encoding in the Mol-GPT model facilitates enhanced comprehension of the mapping relationship between scaffolds and intact molecules. Mol-GPT has been demonstrated to be highly effective at generating new molecules, utilizing predetermined molecular scaffolds while exploring the various possible structures of molecular branched chains via different sample temperature.

The scaffold feature $s_i^t \in \mathbb{R}^{v\times d}$ for the ith scaffold is first obtained after word embedding coding and position coding:

$$s_i^t = s_i \cdot W_E + p_i^s \tag{1}$$

where $s_i \in \mathbb{R}^{v\times 1}$,$W_E \in \mathbb{R}^{d\times d}$ and $p_i^s \in \mathbb{R}^{v\times d}$. The interrel-ationship features of the scaffold are extracted through the multi-head self-attention (MHA) mechanism, as shown in the following equation:

$$q_i^{\kappa,t} = W_Q^{\kappa,t}\mathrm{Norm}\left(s_i^t\right) \tag{2}$$

$$k_i^{\kappa,t} = W_K^{\kappa,t}\mathrm{Norm}\left(s_i^t\right) \tag{3}$$

$$v_i^{\kappa,t} = W_V^{\kappa,t}\mathrm{Norm}\left(s_i^t\right) \tag{4}$$

$$z_i^t = \mathrm{Dropout}\left(\mathrm{Concat}_{\kappa\in 1,\ldots,H}\left(\mathrm{Softmax}\left(\frac{q_i^{\kappa,t}\cdot\left(k_i^{\kappa,t}\right)^T}{\sqrt{d_k}}\right)\cdot v_i^{\kappa,t}\right)\right) \tag{5}$$

$$s_i^p = \mathrm{TDecoder}\left(s_i \cdot E_T + p_i, z_i^t\right) \tag{6}$$

where $W_Q^{\kappa,t}$, $W_K^{\kappa,t}, W_V^{\kappa,t} \in \mathbb{R}^{d\times d}$are learnable parameters from linear layers; H denotes the number of attention heads; dk is the dimension of each head, which equals d divided by H; Norm denotes layer normalization; Concat denotes the concatenation operation; Dropout denotes the dropout operation. The output of the maximum likelihood estimation of the scaffold is input

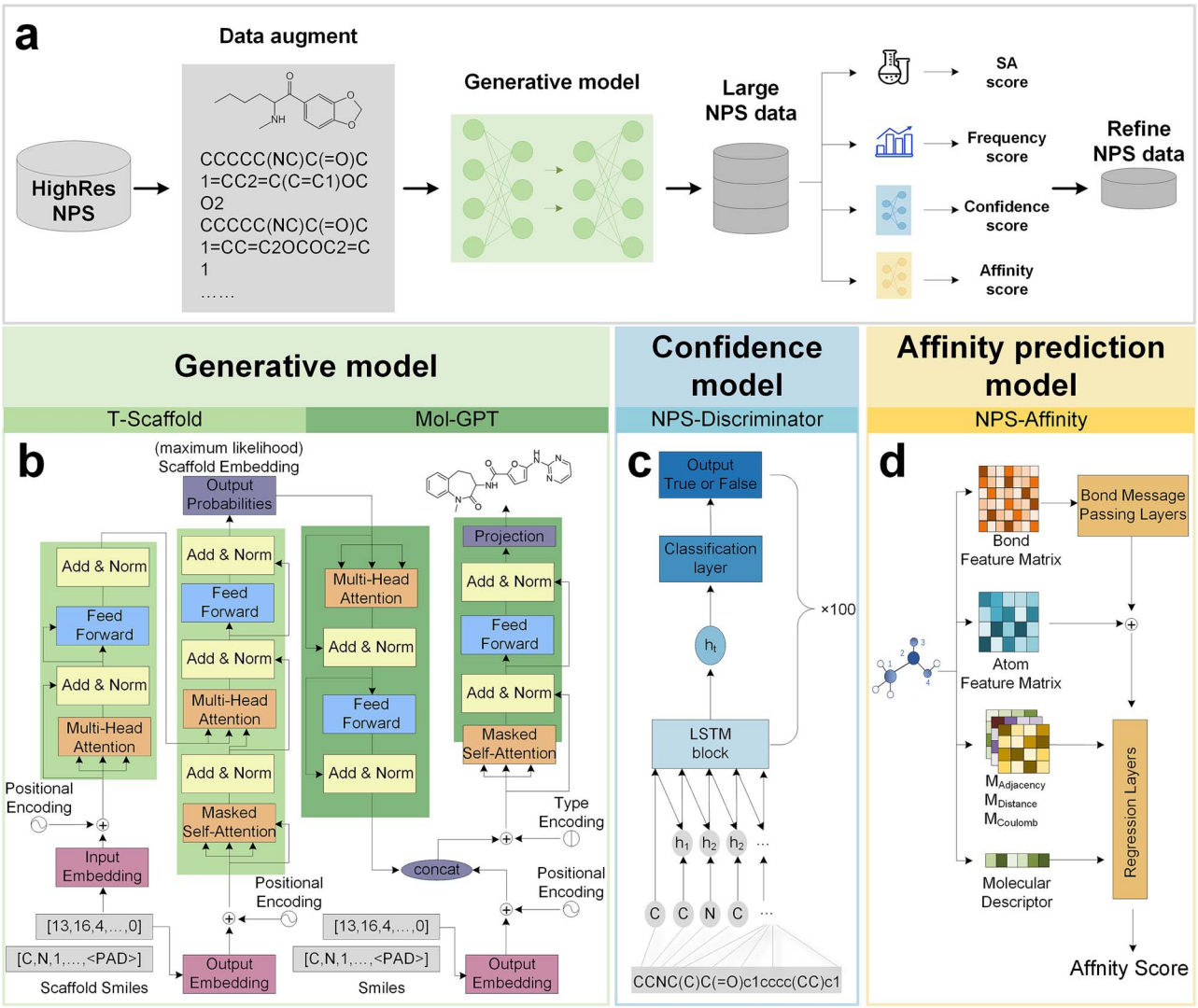

Figure 1. Framework overview. (a) STNGS: Data augment block, generative block and ranking block. (b) Architecture of generative model. The molecular scaffold SMILES is processed by the T-scaffold self-encoder-decoder to extract the latent vector of scaffolds, which is then passed to the Mol-GPT module for molecular generation training. (c) Architecture of NPS-discriminator. This model is used as a sub-model and trained 100 times to obtain 100 models, which are then integrated to form the final NPS discriminator. (d) Architecture of affinity prediction model. The bond feature matrix is concatenated with the atomic feature matrix after passing through the bond message passing layers. It is input to the regression layer with the adjacency feature matrix, the distance feature matrix, the coulomb feature matrix and the molecular description vector to calculate the affinity score. Alt text: Overview of the STNGS framework (a) with three blocks: Data augmentation, generation, and ranking. (b) Generative model: T-scaffold extracts scaffold latent vectors for Mol-GPT molecular generation. (c) NPS-discriminator: 100 two-layer LSTM sub-models integrated for discrimination. (d) Affinity prediction: Bond and atomic features combined for regression based affinity scoring.

into the GPT module of the model, along with the molecular features. This allows the model to explore the chemical space of the molecule, as shown in the following equation:

$$s_i^G = \text{GPTEncoder}\left(s_i^P\right) \tag{7}$$

$$t_c = \begin{cases} 0 & \text{scaffold} \\ 1 & \text{molecule} \end{cases} \tag{8}$$

$$I_i = \text{Concat}\left(s_i^G + t_{c,0}, m_i \cdot W_E + p_i^m + t_{c,1}\right) \tag{9}$$

$$q_i^{\kappa,g} = W_Q^{\kappa,g} \cdot \text{Norm}\left(I_i\right) \tag{10}$$

$$k_i^{\kappa,g} = W_K^{\kappa,g} \cdot \text{Norm}\left(I_i\right) \tag{11}$$

$$v_i^{\kappa,g} = W_V^{\kappa,g} \cdot \text{Norm}\left(I_i\right) \tag{12}$$

$$z_i^g = \text{Dropout}\left(\text{Concat}_{\kappa\in 1,\ldots,H}\left(\text{Softmax}\left(\frac{q_i^{\kappa,g}\cdot\left(k_i^{\kappa,g}\right)^T}{\sqrt{d_k}}\right)\cdot v_i^{\kappa,g}\right)\right) \tag{13}$$

$$d_i = z_i^g + W_{hi}^l \text{Dropout}\left(\text{GELU}\left(W_{mi}^l \text{Norm}\left(z_i^g + I_i\right)\right)\right) \tag{14}$$

$$e_i = \text{Projection}\left(d_i\right) \tag{15}$$

where type coding $t_c \in \mathbb{R}^{v\times d}$ is identified with "0" for scaffolds and "1" for molecules; $W_Q^{\kappa,g}$, $W_K^{\kappa,g}$, $W_V^{\kappa,g} \in \mathbb{R}^{d\times d}$, $W_{mi}^l \in \mathbb{R}^{d\times 4d}$ and $W_{hi}^l \in \mathbb{R}^{4d\times d}$ are learnable parameters from linear layers; Gaussian Error Linear Unit (GELU) represents a type of activation functions; Projection denotes the projection layer. The training processing was carefully designed. We use the cross-entropy loss as the loss

function:

$$L = -\sum Y_i \cdot \text{Logsoftmax}(e_i) \tag{16}$$

The hyperparameters of the model are listed in the Supplementary Table S5.

NPS discriminator

To bolster the reliability of the neural network in NPS discrimination, this study employs an ensemble learning strategy.

Ensemble learning involves combining multiple neural network models to create an ensemble model specifically tailored for NPS discrimination tasks, facilitated by a voting mechanism [14]. The deep ensemble model consists of 100 sub-models, each incorporating a two-layer LSTM (Fig. 1c). After each LSTM layer, a Dropout layer is introduced, followed by an output through a fully connected layer. The model undergoes training for 100 iterations, resulting in the development of 100 distinct models. Subsequently, these 100 models are combined to form the ultimate NPS discriminator using the ensemble methodology.

$$h_t^l = f_l\left(h_{t-1}^l, h_t^{l-1}\right) \tag{17}$$

$$h_t^0 = x_t \tag{18}$$

f_l represents the lth layer of the LSTM network, t is the moment, h is the hidden layer vector, and x_t represents the token at the t moment.

$$\text{output} = \text{MLP}\left(h_t^l\right) \tag{19}$$

MLP is the fully connected layer.

The hyperparameters of the model are listed in the Supplementary Table S5.

Affinity prediction model

Hu *et al.* [15] introduced an atom-bond transformer-based message-passing neural network (Fig. 1d) for predicting molecular properties. We employ the generated molecules for affinity prediction, represented by the affinity score.

$$G = (V, E) \tag{20}$$

Graph G consists of the atom set V and the bond set E.

$$e'_{vw} = \text{BMP}(e_{vw}) \tag{21}$$

$$h_{vw} = \text{Concat}\left(x_v, e'_{vw}\right) \tag{22}$$

$$m_v = \text{ReLU}\left(W_0 \text{Concat}\left(x_v, \Sigma_{w \in \text{Neighbor}(v)} h_{vw}^T\right)\right) \tag{23}$$

BMP is the Bond Message Passing Layers. Bond feature $e_{vw} \in \mathbb{R}^{n \times F_b}$, atom features$x_v \in \mathbb{R}^{m \times F_a}$.

$$h_v = \text{AtomAttention}\left(m_v, M_{\text{Adjacency}}, M_{\text{Distance}}, M_{\text{Coulomb}}\right) + m_v \tag{24}$$

Three inter-atomic matrices:$M_{\text{Adjacency}}, M_{\text{Distance}}, M_{\text{Coulomb}}$.

$$h = \Sigma_{v \in G} h_v \tag{25}$$

$$\hat{y} = \text{FFN}\left(\text{Concat}\left(h, h_f\right)\right) \tag{26}$$

$h_f$represents molecular descriptors.

Model evaluation

The performance of the generative model is assessed in this study on the MOSES dataset, and a comparison is conducted with four baseline models using the subsequent metrics: validity, uniqueness, recovery, similarity and scaffold similarity. The computation of the similarity is grounded in the Tanimoto coefficient of molecular fingerprints. In this specific investigation, extended-connectivity fingerprints featuring a radius of 2 and a length of 1024 bits (ECFP4 [16]) are adopted. Scaffold similarity encompasses the average resemblance between the scaffolds associated with the generated molecules and their corresponding counterparts within the test molecules.

This study evaluates the performance of the ensemble NPS discriminator model by constructing baseline models based on machine learning and employing molecular fingerprints. A comparison is carried out between the constructed baseline models and the submodels within the ensemble. Utilizing a total of 12 model combinations, the study employs three distinct molecular descriptors (ECFP fingerprint, MACCS fingerprint [17], and Topological fingerprint) and four machine learning techniques (Random Forest, Support Vector Machine, k-Nearest Neighbors, and Adaptive Boosting). Evaluation metrics for the models include precision, recall, F1 score, and Area Under the Curve (AUC). The models undergo training using a five-fold cross-validation methodology, and their performance is assessed on both the test set and an external validation set (The results are shown in Supplementary Table S3, Supplementary Table S4 and Supplementary Analyses 2).

This study systematically assessed the molecular properties generated by the model, with a focus on distinct physical and chemical attributes (SA Score, LogP, NP Score [18], BertzTC [19], TPSA [20, 21], quantitative estimate of drug-likeness (QED) [22], percent of carbons and the number of aromatic rings).

NPS rank function

Distinguishing potential NPS molecules from the generated pool remains a challenging task. To address this challenge, our research introduces a ranking function that considers two metrics: the composite deep ensemble network scores derived from the NPS discrimination model and synthetic accessibility score.

In this approach, the confidence in the potential of the molecule as a NPS is determined by the ratio of positive votes from the NPS discriminator. The confidence is calculated as the ratio of sub-models classified positively to the total number of models. The confidence ranges from 0 to 1, with higher values indicating a stronger resemblance between the molecule and those in the training set. Consistent with chemical similarity principles, increased confidence signifies a heightened likelihood of the molecule being a potential NPS candidate. Moreover, to guarantee the practicality of synthesizing the generated molecules, the research utilizes the synthetic accessibility score (SA score) to impose constraints. Only molecules with an SA score of 3 or below are retained.

In the initial step, we compute the Tanimoto coefficient between the generated compounds and the nearest neighbor structures derived from the collection of known NPSs molecules. Following this, we categorize the compounds based on their sampling frequency to identify potential dissimilarities in their

resemblance to established NPS molecules across various groups. Our study incorporates a molecular sampling frequency score into the previously discussed ranking function.

Normalization of confidence scores and frequency scores:

$$\text{MinMaxNormalisation}(x) = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{27}$$

$$\text{Norm Confidence Score} = \text{MinMaxNormalisation}(\text{Confidence Score}) \tag{28}$$

$$\text{Norm Frequence Score} = \text{MinMaxNormalisation}(\text{Frequence}) \tag{29}$$

$$\text{Norm SA score} = \frac{\text{SA}}{10} \tag{30}$$

We design three ranking functions:

$$\text{Rank Funtion } 1 = \text{Norm Confidence Score} - \text{Norm SA score} \tag{31}$$

$$\text{Rank Funtion } 2 = \text{Norm Confidence Score} - \text{Norm SA score} + \text{Norm Frequence Score} \tag{32}$$

$$\text{Rank Funtion } 3 = \text{Norm Confidence Score} - \text{Norm SA score} + \text{Norm Frequence Score} * 2 \tag{33}$$

Molecular biology

To generate stable cell lines for synthetic cannabinoids efficacy evaluation, sequences encoding GRABeCB2.5 is cloned into a vector via Gibson [23] assembly called pPacific, containing a 3′ terminal repeat, IRES, the puromycin gene, and a 5′ terminal repeat. Two mutations (S103P and S509G) are introduced into pCS7-PiggyBAC to generate hyperactive piggyBac transposase (ViewSolid Biotech) [24].

Cell lines

HEK293T cells are purchased from ATCC (CRL-3216) and verify based on their morphology and growth rate. Stable cell line expressing GRABeCB2.5 is generated by co-transfecting HEK293T cells with the pPacific plasmids encoding sensors and the pCS7-PiggyBAC plasmid encoding the transposase. The GRAB sensor cell line is cultured at 37°C in 5% CO_2 in DMEM (Biological Industries) supplemented with 10% (v/v) fetal bovine serum (GIBCO) and 1% penicillin–streptomycin (GIBCO).

Fluorescence imaging of cultured cells

Before imaging, the culture medium is replaced with Tyrode's solutions containing: 150 mM NaCl, 4 mM KCl, 2 mM MgCl2, 10 mM HEPES and 10 mM glucoses (pH adjusted to 7.35–7.45 with NaOH). The cells then are imaged in an Operetta CLS high-content screening system (PerkinElmer) with a 488-nm laser. Green fluorescence is collected using 525/50-nm emission filter.

To measure GRABeCB2.5 responses induced by various chemicals, solutions containing the indicated concentrations of synthetic cannabinoids are administered to the cells via bath application.

Data for 96-well plate imaging are collected and analyzed using Harmony high-content imaging and analysis software (PerkinElmer). In brief, membrane regions are selected as ROI and the green fluorescence channel (that is, the sensor) is measured. The ΔF/F0 values (F0 represents baseline fluorescence) are then calculated using the formula ΔF/F0 = (F − F0)/F0.

Results

Exploring key factors for optimizing generative model performance

Schneider's investigation [25] demonstrates the feasibility of constructing a resilient chemical language model using a limited set of modest datasets by employing data augmentation techniques. The present study also employs data augmentation for the purpose of model training.

The optimal data augmentation factor is determined by conducting a comparative analysis of data augmentation multiples (Fig. 2a). The scaffold ratio represents the proportion of molecular scaffolds in the generated molecules that align with the provided molecular scaffolds, while validity indicates the adherence to the SMILES syntax in the generated molecules. At a data enhancement factor of 200, the validity value is 0.9853, and the scaffold ratio is 0.9340. This investigation is focusing on two prevalent molecular scaffolds: the generic scaffold and the Bemis-Murcko (BM) scaffold [26] (Fig. 2b). Training procedures are conducted independently on established MOSES datasets [27] using the BM molecular scaffolds. The results (Fig. 2c) show that the model on the BM scaffold has a higher molecular uniqueness, which is favorable for our task of discovering potential NPSs. Therefore, BM scaffold is selected for NPSs discovery and design.

During the molecular sampling process using the model, we have observed a tendency for sampled molecules to repeat, leading to a decrease in molecular distinctiveness. To address this issue, we conduct molecular sampling at different temperatures using the model and calculate validity, uniqueness, and novelty metrics at each temperature (Fig. 2d). Uniqueness increases as the sampling temperature rises, while validity decreases but remains within acceptable limits. Four distinct scaffolds, based on different levels of complexity, are selected to randomly sample a single molecule for all scaffolds at each of the five designated sampling temperatures. This procedure results in a total of 20 unique molecules. We visualize the molecular structures of these 20 entities (Fig. 2e). The chemical diversity of the sampled molecules is markedly enhanced by the presence of alkyne, nitrile, sulfonyl, and even sulfonamide groups as the sampling temperature increases, expanding its coverage of chemical space. At a sampling temperature of 1.5, the sampled molecules contain various functional groups, including phenylacetylene and methylxanthamides. The findings indicate that utilizing a higher sampling temperature results in the generation of a more expansive range of molecules.

Evaluate the ability of the model to generate molecules

To assess the performance of the STNGS architecture in generating general molecules, we train the model on the MOSES dataset and compare its results with those of recently published models with similar objectives: (CharVAE [28], FragLinker [29], MoFlow and Sc2Mol [30]). The evaluation includes assessing the ability of the model to generate molecules, considering criteria such as validity, uniqueness, novelty, scaffold similarity (see Supplementary Tables S1 and S2; Supplementary Analyses 1),

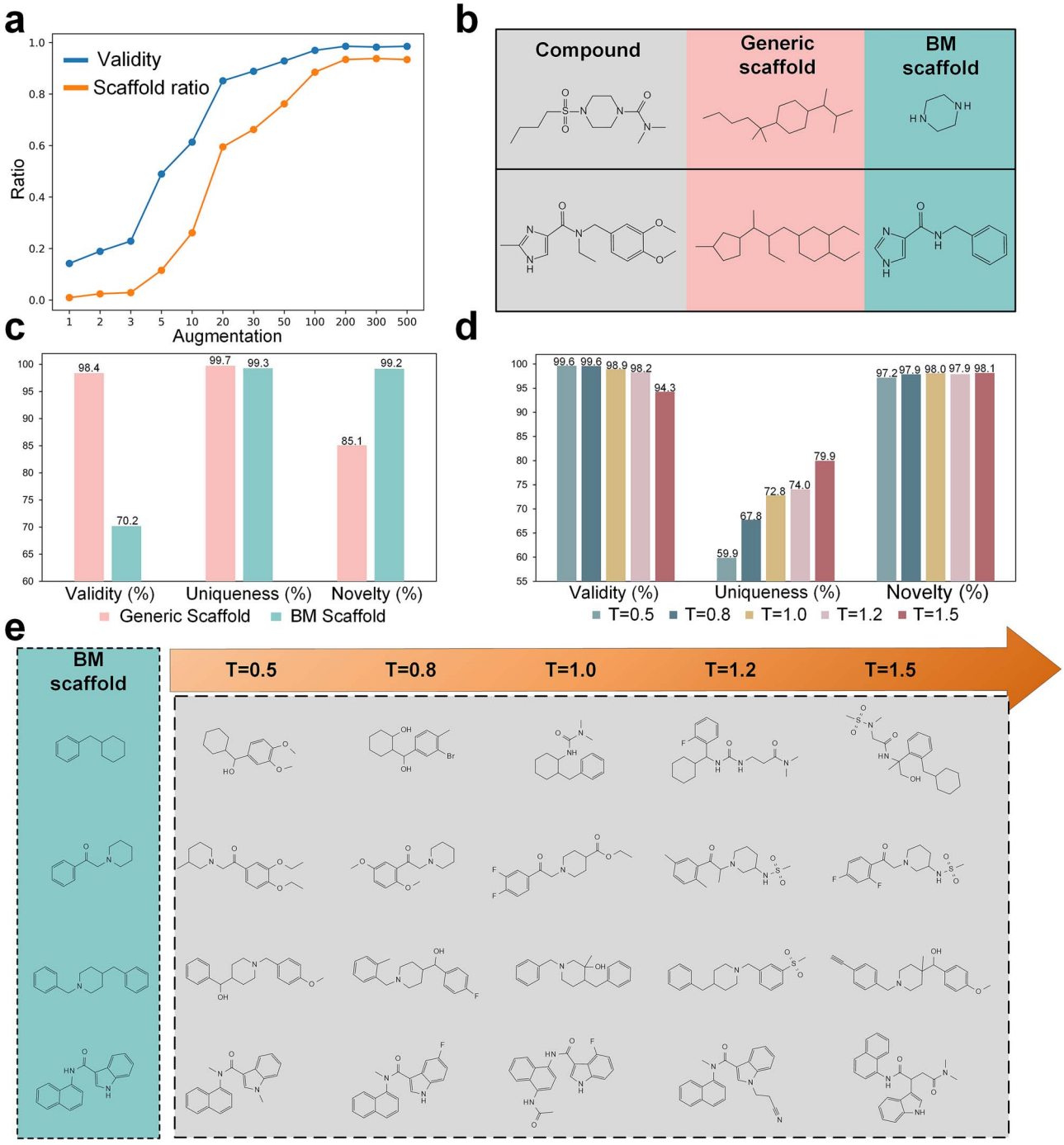

Figure 2. Effect of data augmentation multiplier, scaffold type and sampling temperature on model performance. (a) Association of Data Enhancement Multiples with validity and scaffold ratio. (b) Two compounds and their corresponding two scaffolds. (c) Impact of two scaffolds on validity, uniqueness and novelty. (d) Impact of sampling temperature on validity, uniqueness, and novelty. (e) Molecules generated by four scaffolds at different sampling temperatures. Alt text: Impact of data augmentation, scaffold type, and sampling temperature on model performance. (a) Data enhancement multiples affect validity and scaffold ratio. (b) Two compounds and their scaffolds. (c) BM scaffold improves uniqueness. (d) Sampling temperature influences validity, uniqueness, and novelty. (e) Molecules generated by four scaffolds at different temperatures.

chemical space and property distribution. The model-generated molecules effectively cover the chemical space of the entire MOSES dataset (Fig. 3a). We select nine physicochemical indicators to assess the similarity between the properties of the generated molecules and those of the training set. Specifically, NP score, BertzTC, QED, the fraction of sp^3 carbons, molecular weight, and the number of aromatic rings are primarily used to assess the structural similarity between generated molecules and those in the training set. LogP and TPSA evaluate molecular lipophilicity and cell membrane permeability, while the SA score reflects synthetic feasibility. At the sampling temperature of 1.0 the model-generated molecules closely resemble those in the MOSES dataset across all nine property distributions (Fig. 3b).

To assess the performance of the STNGS architecture in generating NPS molecules, we train the model on the HighResNPS dataset. We demonstrate the superiority of the STNGS framework

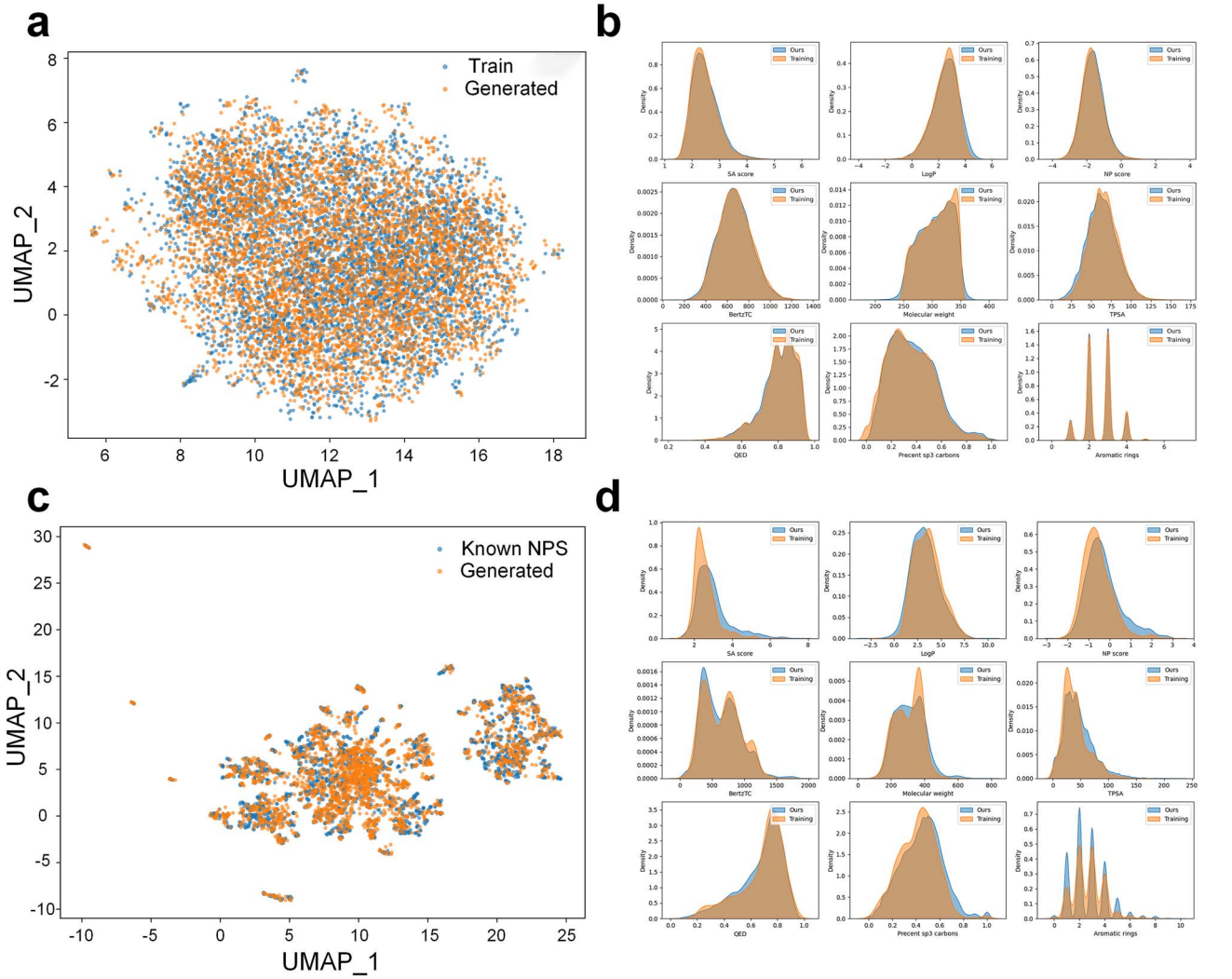

Figure 3. Chemical space and property distribution of generated molecules. (a) UMAP visualization of 5000 randomly selected molecules form MOSES dataset and 5000 randomly selected molecules from the trained generative model. (b) Properties distribution of 30 000 randomly selected molecules form MOSES dataset and 30 000 randomly selected molecules from the trained generative model. (c) UMAP visualization of 2154 molecules form NPS dataset, and all generated NPS-like molecules from the trained generative model. (d) Properties distribution of 2154 molecules form NPS dataset and all generated NPS-like molecules from the trained generative model. Alt text: Chemical space and property distribution of generated molecules. (a) UMAP of 5000 MOSES and 5000 model-generated molecules. (b) Property distributions of 30 000 MOSES and model-generated molecules. (c) UMAP of 2154 NPS dataset molecules and generated NPS-like molecules. (d) Property distributions of NPS dataset molecules and generated NPS-like molecules.

in predicting the potential of emerging NPS molecules through comparative and ablation experiments (with DarkNPS [8], Multi-step Decorator [31] and results in Table 1). We analyze the chemical space and physicochemical property distributions of potential NPS molecules generated by the model. The generated potential NPS molecules exhibite significant overlap with the chemical space of NPS molecules in the HighResNPS dataset (Fig. 3c). This indicates that the model effectively explores the chemical space of potential NPSs. As shown in Fig. 3d, the distribution of physicochemical properties of potential NPS molecules generated by the model closely aligns with those of NPS molecules in the HighRes NPS dataset. This consistency demonstrates the model's ability to capture the physicochemical characteristics of NPS molecules. While the similarity between the properties of the generated molecules in Fig. 3d and the training set is lower compared to those in Fig. 3b, this difference arises from setting the sampling temperature to 1.5 to prioritize the generation of more novel potential NPS molecules. Fig. 3 demonstrates the exceptional performance of the STING architecture in generating both general molecules and potential NPS candidates. By leveraging a sampling temperature strategy, the model effectively captures molecular features and thoroughly explores chemical space.

Constructing ranking functions to screening potential NPSs

To achieve a wide range of molecular structures while ensuring the accuracy of the sampled SMILES, the sampling temperature is set to 1.5 for each scaffold and the results of each scaffold are resampled 1000 times to generate the data set for screening. After removing any illegitimate SMILES and NPSs that appeared in the training set, we are left with a total of 127 739 unique molecules.

This study presents a ranking function designed to identify potential NPSs from a large number of generated NPS-like compounds. The ranking function consists of two key components: an evaluated NPS discriminator model (details in the NPS discriminator section of the Methods) and a synthetic accessibility (SA) score assigned to the molecules. The provided ranking function (See methods NPS rank function section for details) is

Table 1. Performance of each model on NPS molecule generation

Model	Validity	Uniqueness	Novelty	Number of NPS hits
DarkNPS	0.9268	0.7089	0.9497	58
Multi-step Decorator	0.9103	0.7117	**0.9913**	69
STNGS (without T-Scaffold)	0.9102	0.7664	0.9665	71
STNGS (without Type-Encoding)	0.9064	0.7595	0.9805	69
STNGS (without T-Scaffold and Type-Encoding)	0.8843	0.7117	0.9515	57
STNGS	**0.9426**	**0.7995**	0.9814	**75**

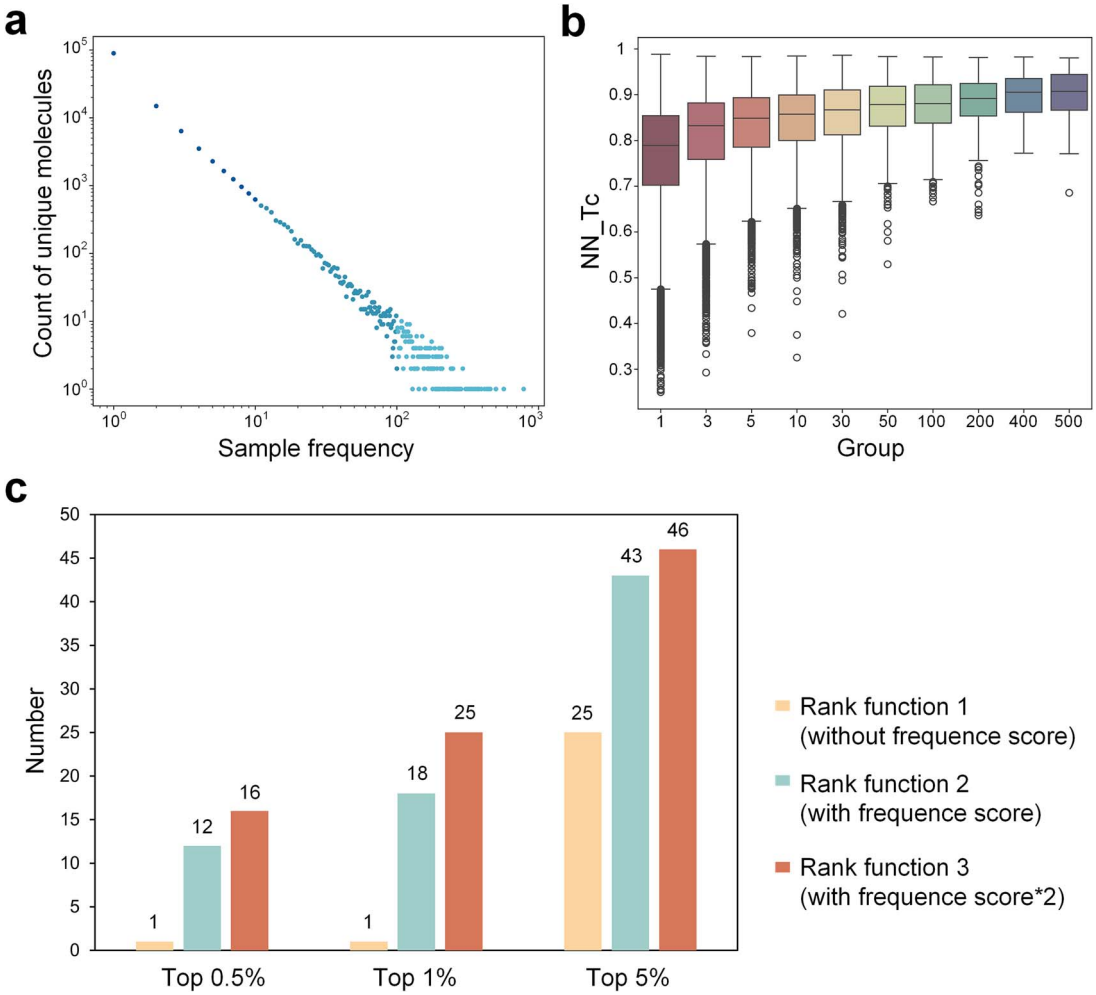

Figure 4. Effect of sampling frequency on the rank function. (a) Relationship between sampling frequency and number of molecules. (b) Nearest-neighbor Tanimoto coefficients to known NPS for all molecules in different sampling frequency groups. (c) Number of emerging NPS molecules hit by different scoring functions. Alt text: Effect of sampling frequency on rank function. (a) Sampling frequency versus number of molecules. (b) Nearest-neighbor Tanimoto coefficients for molecules in different sampling frequency groups. (c) Number of emerging NPS molecules identified by different scoring functions.

used to evaluate and rank a collection of >120 000 generated molecules.

After processing the model-generated molecules, we conduct a thorough investigation into the recurrence of specific molecules in the final outcomes. We then perform a statistical analysis to evaluate the frequency of occurrence of the model-generated molecules (Fig. 4a). It is worth noting that a significant number of molecules are sampled either individually or infrequently, in contrast to a smaller subset of molecules that are subjected to repetitive sampling, with frequencies reaching thousands of instances. The highest sampling frequency observed is 786, while the lowest is only once.

Our hypothesis suggests that molecules with higher sampling frequencies may be due to the model assigning greater weights to certain tokens, resulting in their more frequent generation. In other words, molecules with higher sampling frequencies may indicate a greater likelihood of being potential NPS entities.

The results (Fig. 4b) show that the higher the sampling frequency, the more similar the molecules are to the NPS molecules. Based on this pattern, we propose to refine the screening of potential NPSs by increasing the frequency score.

The 146 new NPS entries added to the HighResNPS database are employed as the test set. The efficacy of different ranking functions is evaluated by comparing the number of molecules

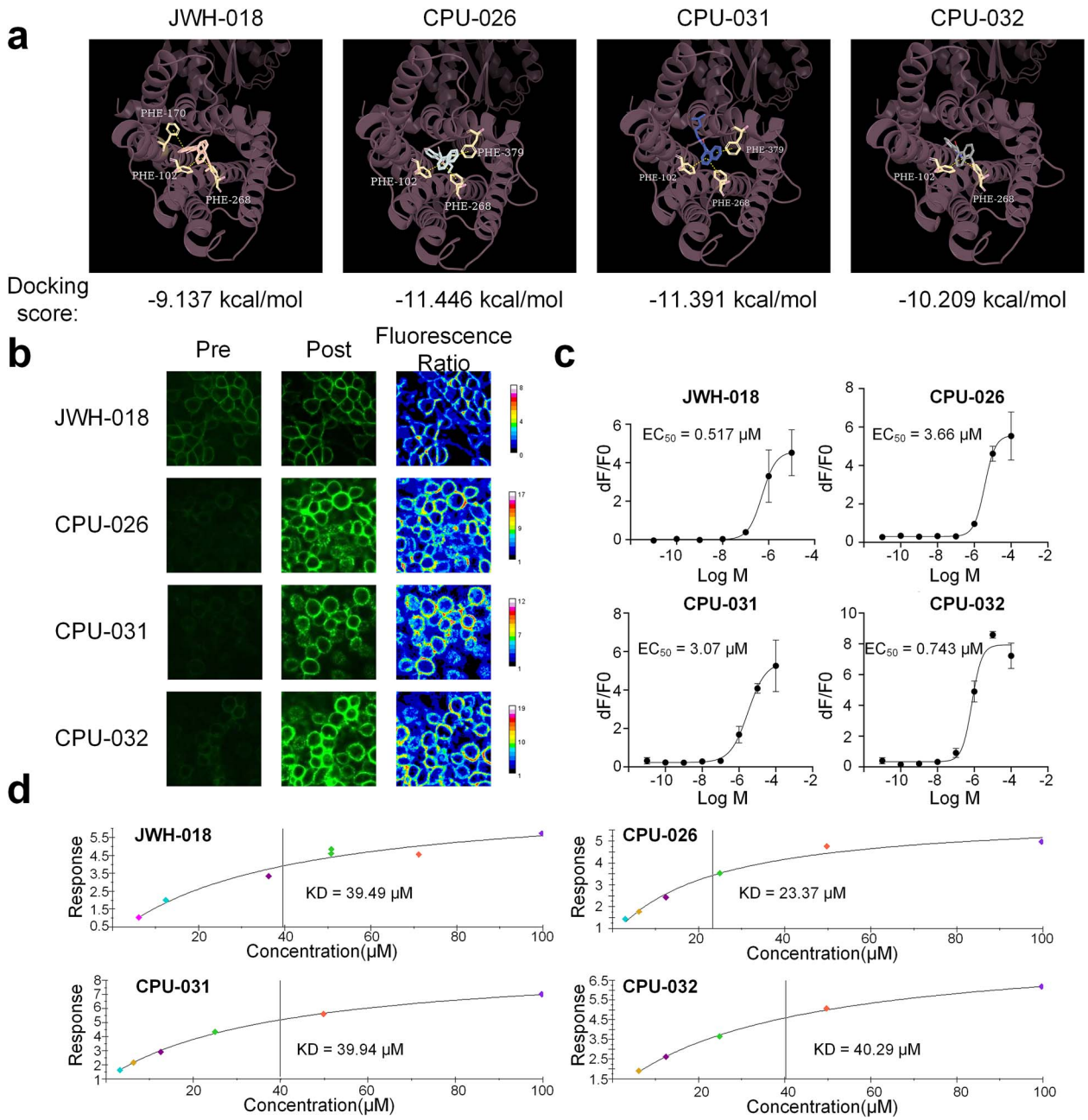

Figure 5. Experimental validation of the generated molecular activity. (a) The binding poses and interaction modes of the target CB1 with three compounds. The dashed line represents the π-π stacking interaction. (b) Expression and fluorescence change in response to three compounds in cells. (c) Effectiveness curves for JWH-018 and three compounds. The effectiveness on cells is shown as the fluorescence expression intensity. (d) Affinity assay and KD values of JWH-018 and three compounds. Alt text: Experimental validation of generated molecular activity. (a) Binding poses and interactions with CB1 receptor. (b) Fluorescence response in cells. (c) Effectiveness curves for JWH-018 and three compounds. (d) Affinity assay results and KD values.

from the top 0.5%, top 1%, and top 5% of the screened and sorted set that hit the test set. The results (Fig. 4c) show that the ranking function 1 based solely on discriminator votes and SA scores hits the fewest molecules in the test set, with only one newly emerged NPS molecule hit in both the top 0.5% and top 1%. In contrast, the scoring function incorporating frequency scores (ranking function 2) hit 12, 18, and 43 newly emerged NPS molecules in the top 0.5%, top 1%, and top 5%, respectively. Moreover, the ranking function 3, which assigns a higher weight to frequency scores, demonstrates an increased number of newly emerged NPS molecules in the top 0.5%, top 1%, and top 5% compared to the other functions. These findings support our hypothesis that including sampling frequency scores is more effective in identifying potential NPSs than without.

Experimental verification of the activity of the generated molecules

To evaluate the screening effectiveness of the STNGS framework for potential NPSs, the activity of the generated molecules is detected. In order to increase the binding activity of the molecules with real receptors, the affinity scores predicted by the affinity prediction model are integrated into the ranking function. For synthetic cannabinoids in NPSs, we select three top-ranked synthetic cannabinoid-like molecules, CPU-026, CPU-031, and CPU-032 (see

Supplementary Figs. S3–S5 and Supplementary Fig. S7), from the generated molecular libraries. A search of the PubChem and Zinc databases using the canonical SMILES of these molecules reveals that all three are unknown compounds. The binding conformations of these three molecules to cannabinoid receptor 1 (CB1) are then analyzed. It is found that all three molecules interact with the CB1 in a similar manner to NPS JWH-018 [32, 33] (Fig. 5a), forming a π-π stacking interaction with active site residues PHE-102 and PHE-268. Additionally, the three compounds form a π-π stacking interaction with active site residues PHE-170 or PHE-379, further enhancing the binding strength. We also use molecular docking to validate other NPSs (results in Supplementary Fig. S6). The STNGS model is considered effective in learning the interaction patterns between binding sites and molecules.

To demonstrate the ability of the ranking function to screen for potentially active NPS, the GRABeCB2.5 sensor are used to determine the binding affinity of the three compounds to the CB1 receptor. All three compounds effectively stimulated the GRABeCB2.5 (Fig. 5b). The EC_{50} value is calculated from the fluorescence intensity (Fig. 4c). Although the EC50 values of compounds CPU-26 (3.66 μM) and CPU-31 (3.07 μM) are higher than that of NPS JWH-018, they remain within an acceptable range. Compound CPU-032 exhibits a comparable affinity to JWH-018. In the Surface Plasmon Resonance (SPR) affinity assay (Fig. 4d), CPU-026 demonstrates the lowest KD value, while the KD values of CPU-31 and CPU-032 are similar to those of JWH-018. This result suggests that the ranking function, which considers affinity scores, can effectively screen potential NPSs and construct a database of potential NPSs with high true positive rates.

Discussion

In this work, we propose a novel molecular generation model based on Transformer and molecular scaffolds. This model is capable of generating complete molecules from a given molecular scaffold, focusing on the identification of potential NPSs. The chemical space of these substances is comprehensively explored using a dataset of over 2000 known NPSs. Through temperature sampling, our generative model produces numerous distinctive potential NPSs.We develop a ranking function for NPS screening based on a neural network classifier, synthetic accessibility score, and sampling frequency score. From the over 120 000 unique molecules generated, we select the top 0.5%, top 1%, and top 5% to form three final virtual libraries of NPSs. These libraries yielded hits for 16, 25, and 46 emerging NPS molecules, respectively.To further refine the results, we incorporated affinity scores into the ranking function and re-evaluated the generated molecules. The top three ranked synthetic cannabinoid-like molecules were selected for experimental validation. Our analysis demonstrates that all three exhibit significant affinity for the CB1 receptor.

In real-world scenarios, violators often evade legal controls by making subtle modifications to the structures of known NPSs, creating new variants. This strategy significantly hinders anti-narcotics efforts in identifying and studying these substances. Our generative model mimics this behavior by preserving the molecular scaffold while systematically exploring the chemical space of NPSs. We developed a self-encoding and decoding scaffold learning module that effectively captures the chemical information of molecular scaffolds and identifies modification-prone sites using an attention mechanism. Unlike the standard Transformer model, which transfers encoder information to the decoder as hidden vectors, our approach integrates molecular scaffold information with inputs from the molecular processor after joint encoding by the scaffold processor. This method mitigates information loss from hidden vector inputs and ensures adequate attention to scaffolds.

The STNGS architecture offers significant advantages over existing NPS generative models and scaffold-based molecular generative models retrained with NPS datasets. Our model identifies the highest number of emerging NPS molecules beyond the training set. It effectively generates and screens high true-positive potential NPS candidates. The resulting database of potential NPS alerts will support proactive regulation efforts. This capability has important implications for NPS research and broader initiatives to combat drug-related crime.

Key Points

- This study first integrates deep scaffold learning and molecular generative language model to develop a novel computational approach STNGS for exploring potential NPS molecules.
- With a well-designed multi-angle ensemble learning scoring function, STNGS can efficiently and accurately screen NPS molecules, substantially accelerating the construction of a database of potential NPS molecules.
- To accurately determine alterations in activity for the generated molecules with the CB1 receptor, we elaborate an affinity assay based on a genetically encoded Endocannabinoid (eCB) sensor known as GRABeCB2.5 and surface plasmon resonance.

Acknowledgments

This work was supported by funding from the program of A study on the diagnosis of addiction to synthetic cannabinoids and methods of assessing the risk of abuse (2022YFC3300905), the research on key technologies for monitoring and identifying drug abuse of Narcotic drugs and psychotropic substances and intervention for addiction (2023Y FC3304200) and ZHEJIANG LAB, the program of Ab initio design and generation of AI models for small molecule ligands based on target structures (2022PE0AC03).

Supplementary Data

Supplementary data is available at Briefings in Bioinformatics online.

Conflict of interest

There are no conflicts to declare.

Funding

None declared.

Data availability

Due to the sensitivity of the data and the potential for misuse, HighResNPS and the databases of generated molecules are not available to the public for unrestricted download. However, the data can be requested from the corresponding authors and will be made available to all qualified researchers in the field upon request. The codes and execution details of STNGS can be found at https://github.com/cpuliaojun/STNGS.

References

1. Peacock A, Bruno R, Gisev N. *et al.* New psychoactive substances: challenges for drug surveillance, control, and public health responses. *The Lancet* 2019;**394**:1668–84. https://doi.org/10.1016/S0140-6736(19)32231-7.
2. Baumann MH, Solis E, Watterson LR. *et al.* Baths salts, spice, and related designer drugs: the science behind the headlines. *J Neurosci* 2014;**34**:15150–8. https://doi.org/10.1523/JNEUROSCI.3223-14.2014.
3. Smith JP, Sutcliffe OB, Banks CE. An overview of recent developments in the analytical detection of new psychoactive substances (NPSs). *Analyst* 2015;**140**:4932–48. https://doi.org/10.1039/C5AN00797F.
4. Yang Y, Liu D, Hua Z. *et al.* Machine learning-assisted rapid screening of four types of new psychoactive substances in drug seizures. *J Chem Inf Model* 2023;**63**:815–25. https://doi.org/10.1021/acs.jcim.2c01342.
5. Nichols D. Legal highs: the dark side of medicinal chemistry. *Nature* 2011;**469**:7–7. https://doi.org/10.1038/469007a.
6. Chai J, Zeng H, Li A. *et al.* Deep learning in computer vision: a critical review of emerging techniques and application scenarios. *Machine Learning with Applications* 2021;**6**:100134. https://doi.org/10.1016/j.mlwa.2021.100134.
7. Salakhutdinov R. Learning deep generative models. *Annu Rev Stat Appl* 2015;**2**:361–85. https://doi.org/10.1146/annurev-statistics-010814-020120.
8. Skinnider MA, Wang F, Pasin D. *et al.* A deep generative model enables automated structure elucidation of novel psychoactive substances. *Nat Mach Intell* 2021;**3**:973–84. https://doi.org/10.1038/s42256-021-00407-x.
9. Mardal M, Andreasen MF, Mollerup CB. *et al.* HighResNPS.com: an online crowd-sourced HR-MS database for suspect and non-targeted screening of new psychoactive substances. *J Anal Toxicol* 2019;**43**:520–7. https://doi.org/10.1093/jat/bkz030.
10. Zhang Y, Jiang Q, Li L. *et al.* Predicting the structure of unexplored novel fentanyl analogues by deep learning model. *Brief Bioinform* 2022;**23**:bbac418. https://doi.org/10.1093/bib/bbac418.
11. Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Chem* 2009;**1**:8. https://doi.org/10.1186/1758-2946-1-8.
12. Dong A, He K, Dudok B. *et al.* A fluorescent sensor for spatiotemporally resolved imaging of endocannabinoid dynamics in vivo. *Nat Biotechnol* 2022;**40**:787–98. https://doi.org/10.1038/s41587-021-01074-4.
13. Wang Y, Xiao J., Suzek TO *et al.* PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res* 2009;**37**:W623–33. https://doi.org/10.1093/nar/gkp456.
14. Lam L, Suen SY. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Trans Syst, Man, Cybern A* 1997;**27**:553–68. https://doi.org/10.1109/3468.618255.
15. Liu C, Sun Y, Davis R. *et al.* ABT-MPNN: an atom-bond transformer-based message-passing neural network for molecular property prediction. *J Chem* 2023;**15**:29. https://doi.org/10.1186/s13321-023-00698-9.
16. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742–54. https://doi.org/10.1021/ci100050t.
17. Durant JL, Leland BA, Henry DR. *et al.* Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 2002;**42**:1273–80. https://doi.org/10.1021/ci010132r.
18. Vanii Jayaseelan K, Moreno P, Truszkowski A. *et al.* Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Bioinformatics* 2012;**13**:106. https://doi.org/10.1186/1471-2105-13-106.
19. Bertz SH. The first general index of molecular complexity. *J Am Chem Soc* 1981;**103**:3599–601. https://doi.org/10.1021/ja00402a071.
20. Pajouhesh H, Lenz GR. Medicinal chemical properties of successful central nervous system drugs. *Neurotherapeutics* 2005;**2**:541–53. https://doi.org/10.1602/neurorx.2.4.541.
21. Hitchcock SA, Pennington LD. Structure–brain exposure relationships. *J Med Chem* 2006;**49**:7559–83. https://doi.org/10.1021/jm060642i.
22. Bickerton GR, Paolini GV, Besnard J. *et al.* Quantifying the chemical beauty of drugs. *Nat Chem* 2012;**4**:90–8. https://doi.org/10.1038/nchem.1243.
23. Gibson DG, Young L, Chuang R-Y. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 2009;**6**:343–5. https://doi.org/10.1038/nmeth.1318.
24. Yusa K, Zhou L, Li MA. *et al.* A hyperactive *piggyBac* transposase for mammalian applications. *Proc Natl Acad Sci USA* 2011;**108**:1531–6. https://doi.org/10.1073/pnas.1008322108.
25. Moret M, Friedrich L, Grisoni F. *et al.* Generative molecular design in low data regimes. *Nat Mach Intell* 2020;**2**:171–80. https://doi.org/10.1038/s42256-020-0160-y.
26. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 1996;**39**:2887–93. https://doi.org/10.1021/jm9602928.
27. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B. *et al.* Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Front Pharmacol* 2020;**11**:565644. https://doi.org/10.3389/fphar.2020.565644.
28. Gómez-Bombarelli R, Wei JN, Duvenaud D. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci* 2018;**4**:268–76. https://doi.org/10.1021/acscentsci.7b00572.
29. Yang Y, Zheng S, Su S. *et al.* SyntaLinker: automatic fragment linking with deep conditional transformer neural networks. *Chem Sci* 2020;**11**:8312–22. https://doi.org/10.1039/D0SC03126G.
30. Liao Z, Xie L, Mamitsuka H. *et al.* Sc2Mol: a scaffold-based two-step molecule generator with variational autoencoder and transformer. *Bioinformatics* 2023;**39**:btac814.
31. Arús-Pous J, Patronov A, Bjerrum EJ. *et al.* SMILES-based deep generative scaffold decorator for de-novo drug design. *J Chem* 2020;**12**:38. https://doi.org/10.1186/s13321-020-00441-8.
32. Hua T, Vemuri K, Pu M. *et al.* Crystal structure of the human cannabinoid receptor CB1. *Cell* 2016;**167**:750–762.e14. https://doi.org/10.1016/j.cell.2016.10.004.
33. Hua T, Vemuri K, Nikas SP. *et al.* Crystal structures of agonist-bound human cannabinoid receptor CB1. *Nature* 2017;**547**:468–71. https://doi.org/10.1038/nature23272.

Briefings in Bioinformatics, 2025, **26(1)**, bbae690
https://doi.org/10.1093/bib/bbae690
Case Study