Novel Molecular Design via a Scaffold-Aware Transformer with Multi-Scale Attention Mechanisms

Junyoung Park [a,b], Sunyong Yoo [a,b*]

[a] Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju, Republic of Korea

[b] R&D Center, MATILO AI Inc., Gwangju, Republic of Korea

*** Corresponding author. Chonnam National University, 77 yongbong-ro, Buk-gu, Gwangju, 61186, Korea, College of Engineering, Building 7, Republic of Korea**

E-mail addresses : **sss206391@gmail.com (J. Park), syyoo@jnu.ac.kr (S. Yoo)**

Highlights

- This study proposes a scaffold-conditioned generative framework where structural control is optimized through continuous bioactivity feedback.
- Scaffold-aware transformer generates molecules that explicitly reflect user-specified scaffolds.
- Multi-scale attention and loss selection improved generation capability.
- Attention-based substructure analysis identifies motifs including binding patterns.

Abstract

Recent advancements in artificial intelligence have demonstrated great potential in accelerating drug discovery by exploring vast chemical spaces and predicting molecular properties. However, conventional molecular generation models have limitations in reflecting desired molecular structures, as they often fail to incorporate specific structural constraints or target properties directly into the generation process. To overcome these limitations, we propose a novel framework that integrates a transformer-based generative model and a graph attention network-based predictive model. The generative model produces molecules with desired structural characteristics by explicitly incorporating scaffold information, while the predictive model estimates the biological activity of the generated molecules. A cyclic learning structure enables the generative and predictive models to interact iteratively, facilitating continuous evaluation and feedback during training. In addition, a multi-stage tournament selection with experience memory guides the subsequent training process. Our approach accelerates the identification of scaffold-consistent, high-affinity candidates by exploring novel chemical variations around a user-specified scaffold. Experimental results show that the proposed scaffold-aware transformer achieves competitive validity, uniqueness, and novelty, and effectively generates novel compounds with high predicted binding affinity for biological targets. An attention-based analysis extracts atom-level importance scores and highlights the substructures that contribute to the predicted binding affinity, providing interpretable insights into structure-activity relationships. This study provides a practical and interpretable tool for scaffold-conditioned molecular generation.

Keywords

Generative Model, *De novo* Molecular Design, Drug Discovery, Attention, Transformer

Introduction

Drug development is a complex and resource-intensive process that faces multiple interconnected challenges. The conventional drug development pipeline takes 10–15 years from concept to approval and costs billions of dollars [1]. In particular, designing molecular structures in the early stages is a challenging task because the chemical search space for new molecules is vast [2]. The number of potential drug-like molecules is estimated to be around 10^{60}, yet only about 10^{8} molecules have been synthesized to date [3]. Despite extensive screening efforts in which thousands of candidate compounds are synthesized, only a small fraction possesses sufficient biological activity and safety to advance into clinical trials [4]. The challenges persist even at later stages, with approximately 90% of drug candidates that enter clinical testing ultimately failing due to insufficient efficacy, toxicity, or lack of drug-like properties [5]. These issues underscore the urgent need for diverse drugs and improved discovery strategies [6-9].

Artificial intelligence is emerging as a means to overcome the limitations of molecular design by accelerating the exploration of the vast chemical space, reducing the time and cost required to derive candidate substances [1, 2, 10-13]. *De novo* molecular design is a computational paradigm for generating novel chemical structures with desired properties. Recently, research efforts to apply generative models have been actively pursued in this field [14-17]. These approaches leverage generative models to learn the distribution of molecules with target-specific activity and to sample novel chemical structures [17]. They can discover promising candidate substances much faster and more efficiently than traditional synthetic chemistry methodologies. Many of these approaches utilize the Simplified Molecular Input Line Entry System (SMILES), a text-based representation that encodes molecular structures as strings, enabling the application of natural language processing techniques to chemistry [18]. Generative models trained on SMILES can learn chemical syntax and generate syntactically valid molecules, with some frameworks incorporating reinforcement learning or evolutionary algorithms to optimize generated compounds [19, 20]. However, most traditional molecular generation models either evaluate the properties of generated molecules separately or cannot directly incorporate target properties during the generation process. To address this limitation, a previous study proposed a framework that optimizes generated molecules by utilizing tournament selection and experience memory [15]. By integrating generative and predictive models, this approach enables property prediction during generation and provides a method to learn optimized distributions of bioactive molecules. However, this approach still has the limitation that it cannot explicitly control the scaffold structure. In medicinal chemistry, the scaffold is the core framework that defines molecular topology and guides key substituent vectors. Early selection and explicit control of the scaffold are central to steering potency, selectivity, and developability, because scaffold changes are labor intensive and often erode activity [21, 22]. Without such scaffold control, the generated molecules may lack the structural characteristics necessary for lead quality.

To address these limitations, we propose a scaffold-aware generative framework that integrates structural control with continuous bioactivity optimization. Our approach integrates a transformer-based generative model and a graph attention network (GAT)-based predictive model in a cyclic learning architecture [15, 23, 24]. Scaffold information is integrated into generation through multi-scale attention, and the model reflects scaffold. This mechanism enables simultaneous control of local atom-bond neighborhoods and global scaffold topology. The GAT-based predictive model estimates the biological activity of generated molecules. Through iterative interaction between the generator and predictor, the framework enables continuous evaluation and refinement throughout the training process. In addition, a tournament-based selection mechanism with experience memory directs subsequent learning iterations. The framework allows for the exploration of new variations while maintaining the core structure of the molecule.

Materials and methods

Datasets

This study utilizes two distinct types of datasets: a molecular design benchmark dataset for training the generative model and a biological activity-labeled dataset for training the predictive model. The statistical characteristics of the datasets are summarized in Table 1.

Table 1. Statistical overview of the datasets utilized in this study

Datasets	Model	Train	Validation	Test	Total
GuacaMol	Generator	1,260,532	78,762	236,374	1,575,668
KOR	Predictor	2,674	573	574	3,821
PIK3CA	Predictor	1,023	219	220	1,462

We used the molecular design benchmark dataset, GuacaMol, for training the generative model [25]. GuacaMol comprises 1,591,378 molecules extracted from the ChEMBL database [26]. Data preprocessing for training the generator was performed as follows. First, SMILES strings were standardized and deduplicated to enhance the training efficiency of the generative model. Second, Bemis-Murcko scaffolds were extracted from each SMILES string to identify the core structural framework of the molecules [27]. RDKit was used in both steps [28]. To ensure structural consistency and relevance, we removed molecules from the dataset for which scaffolds could not be extracted—typically those with simple linear structures or lacking a basic framework. Third, we defined a vocabulary to facilitate tokenization of SMILES in the model [29]. Fourth, we defined and applied a regular expression pattern to tokenize SMILES strings into meaningful units [30]. Since SMILES includes a wide range of chemical symbols and bond representations, precise tokenization is crucial for the model to learn molecular structural information. Fifth, special tokens—[SOS] (start of sequence) and [EOS] (end of sequence)—were appended to each tokenized SMILES string to clearly denote the beginning and end of each molecular sequence. Finally, padding tokens were added to ensure that all SMILES strings within the dataset had uniform lengths. This uniformity was necessary for efficient batch processing during model training, allowing the generator to handle sequences of consistent dimensions.

We used biological activity datasets for KOR (κ-opioid receptor) and PIK3CA (Phosphatidylinositol 3-Kinase Catalytic Subunit Alpha). Both datasets were sourced from preprocessed sets provided by previous study [15]. The original preprocessing pipeline comprised SMILES canonicalization and removal of tokens that were absent from the vocabulary. For example, tokens such as '[Br-]', '[I-]', and '[Cl-]' were removed. Additionally, when a compound had multiple bioactivity measurements, the median value was chosen as its representative label. For KOR, bioactivity was supplied as pIC50. For PIK3CA, bioactivity values were provided as pKi and pKd—metrics that represent the negative logarithms of the inhibition (Ki) and dissociation (Kd) constants, respectively [31-33]. To enable unified activity prediction, the pKi and pKd values were merged into a single measure, pKx. Activity thresholds were established to distinguish between active and inactive compounds. For KOR, molecules with a pIC50 value of 7.0 or higher were classified as active, and for PIK3CA, molecules with a pKx value of 8.0 or higher were classified as active [15].

Framework architecture

An overview of the framework's architecture is presented in Fig. 1. This study focuses on the interaction between the generative and predictive models. The generative model produces new SMILES based on input scaffolds. The predictive model evaluates the generated molecules by estimating their binding affinities.

Molecules with high predicted binding affinity are selected and fed back into the training process. Thus, the generation and prediction processes alternate iteratively, progressively optimizing toward high-affinity molecular candidates.

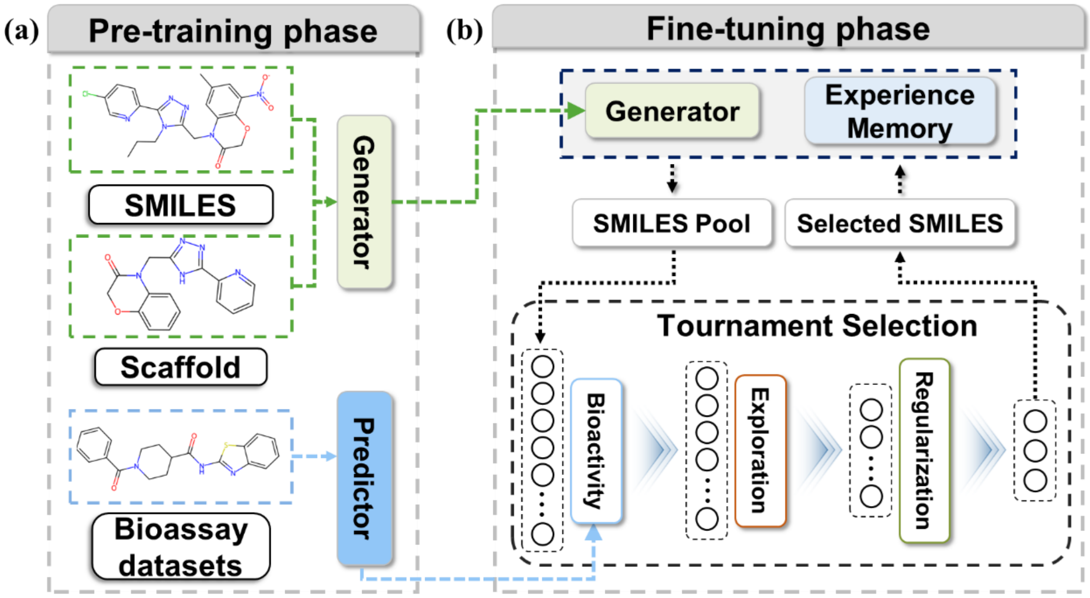

Fig. 1. Overall training framework with pre-training and fine-tuning. (a) Pre-training phase. The generator is trained on SMILES and the corresponding scaffolds, while the predictor is trained on bioassay datasets. (b) Fine-tuning phase. The generator samples candidate SMILES and the experience memory extracts previously stored SMILES; together they form a SMILES pool. A multi-stage tournament selection is then applied to the pool to obtain selected SMILES, which are used to retrain the generator and to update the experience memory.

Generative model

The generative model is based on the transformer architecture, which utilizes a self-attention mechanism to process sequential data (Fig. 2). The transformer overcomes the limitations inherent in conventional sequence models, such as RNN and LSTM, by enabling parallel processing [23, 34, 35]. This effectively addresses long-term dependency issues. The core of the transformer is its attention mechanism, which learns how various positions in the input sequence relate to one another, allowing the model to focus on important information. Each input token is converted into three vectors, including query, key, and value. The attention score is derived by first calculating the dot product of the query and key vectors, then normalizing the result with the softmax function. The final representation for each token is obtained by multiplying the normalized attention scores with the value vectors. This process can be represented by the following formula:

$$A(Q,K,V) = softmax(QK^T)V \tag{1}$$

In Equation 1, $A(Q,K,V)$ is the attention value, Q is the query vector, K is the key vector, V is the value vector. This mechanism operates in parallel across multiple attention heads.

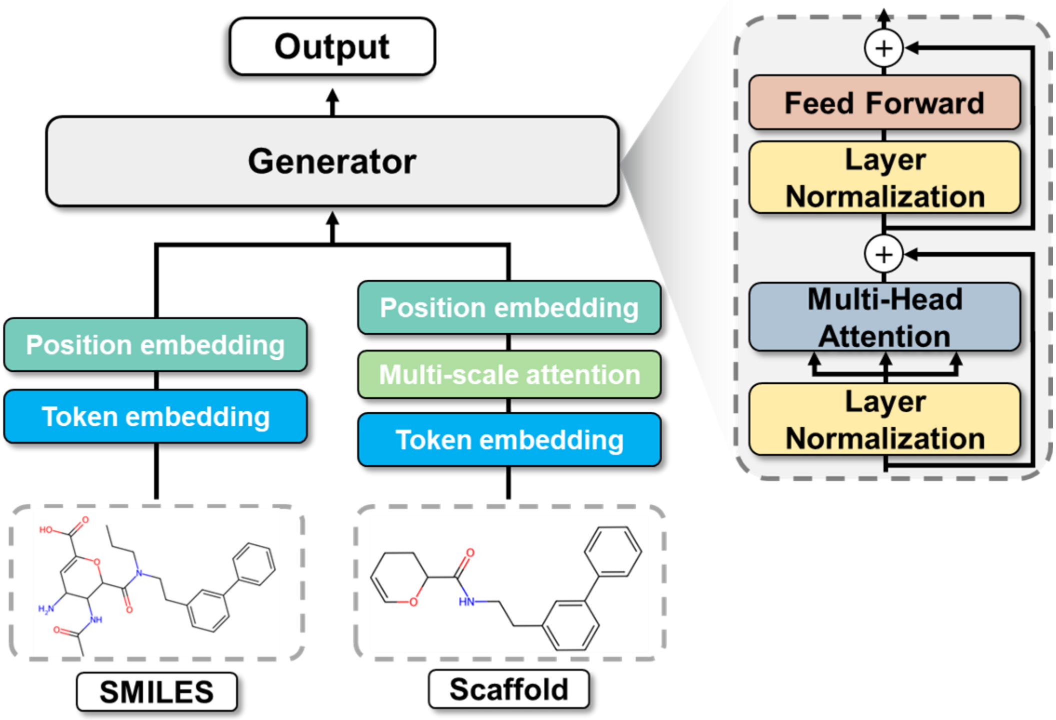

Fig. 2. The architecture of generative model. Input SMILES and scaffolds are converted into numerical representations through token embedding. For scaffolds, multi-scale attention extracts structural features at different scales. Both inputs are combined with positional embeddings to preserve sequential information before being fed into the generator. The generator comprises multiple decoder blocks, each containing layer normalization, multi-head attention, and feed-forward network with residual connections. Output is the sampled from the generator.

The input SMILES and scaffolds are converted into token embeddings. For scaffold inputs, we introduced a multi-scale attention mechanism to effectively generate molecules with desired structural characteristics [36]. This mechanism extracts structural information from the scaffold at various scales and incorporates it into the model. For example, when the scale is 1, the original scaffold embedding is used; when the scale is 2, grid sampling is applied to reduce its length by half; and when the scale is 4, the length is reduced one-quarter. In this study, scales of 1, 2, and 4 were used to consider both local and global patterns of the scaffold. The attention outputs at each scale are linearly interpolated back to the original scaffold embedding length. Subsequently, their mean is computed to provide an integrated scaffold attention representation. This process can be formulated as follows:

$$X_s = Downsample(X_{scaffold}, s) \tag{2}$$

$$Z = \frac{1}{3} \sum_{s \in \{1,2,4\}} Interpolate\ (MHA_s(X_s), L) \tag{3}$$

In Equations 2 and 3, $X_{scaffold}$ represents the scaffold token embeddings, s denotes the scale factor, X_s is the downsampled scaffold embedding at scale s, MHA_s is the multi-head attention operation at scale s, and L is the original scaffold sequence length. Positional encoding is then employed to provide the model with information about the order of the sequence**.** The token embeddings and positional embeddings are combined and used as inputs to the generator.

The generator consists of multiple transformer decoder blocks. In each decoder block, the multi-head attention enables the learning of relationships between each position in the input sequence and other positions simultaneously across various representation spaces. Following the multi-head attention layer, a feed-forward neural network is applied, consisting of two linear transformations and a gaussian error linear unit function placed in between [37]. Layer normalization is employed before each sub-layer to normalize the input features. Additionally, residual connections are incorporated around each sub-layer, which mitigate the vanishing gradient problem and improve information flow through the network. The focal loss is applied to mitigate class imbalance and improve training efficiency [38]. During the molecular generation, certain tokens or patterns may appear disproportionately frequently, and focal loss mitigates this imbalance by down-weighting these frequent cases. The focal loss function is defined as follows:

$$FL(p_{\mathrm{t}}) = -\alpha \times (1 - p_t)^{\gamma} \times log(p_t) \tag{4}$$

In Equation 4, p_t is the predicted probability for the target class. α is the weighting factor that addresses class imbalance by assigning more weight to the minority class. γ is the focusing parameter; it functions to reduce the weighting on well-classified samples and increase the weighting on misclassified samples. The SMILES sequence generation process is described in Section 1 and Fig. S1 of Supplementary materials.

Predictive model

The predictive model consists of three GAT convolutional layers and fully connected layers (Fig. 3). SMILES strings are converted into graph form, where atoms serve as nodes and bonds as edges. Each node in the graph has a vector representing the features of that atom, such as atom type, charge state, and other physicochemical properties. These graph representations are then fed into the predictor. Each GAT convolutional layer learns interactions between nodes from multiple perspectives through a multi-head attention mechanism, effectively capturing complex features of the molecular structure. The core of GAT involves calculating attention coefficients that determine the importance of neighboring nodes when updating a node's feature vector [24]. This is accomplished by first applying a linear transformation to each node feature vector and then computing attention scores between neighboring nodes. The attention scores are then normalized and used to weight the feature vectors of neighboring nodes, which are subsequently aggregated to update the node's feature vector. The ReLU activation function is applied after each GAT layer. Additionally, dropout is used to prevent overfitting. After the three GAT layers, global max pooling is performed to extract a graph-level feature vector, which is then passed through fully connected layers to yield the final prediction.

During training, labeled SMILES data are used to predict the binding affinity between molecules and targets. Techniques such as weight decay and learning rate scheduling are applied during training. The mean squared error (MSE) is used as the loss function to formulate the task as a regression problem. The MSE quantitatively evaluates the model’s prediction performance by squaring and averaging the differences between predicted and true activity values. It is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

In Equation 5, n is the total number of samples, y_i is the true activity value of the i-th sample, and $\hat{y}_i$ is the predicted activity value for the i-th sample. By minimizing the MSE, the predictor updates its parameters to improve activity predictions, gradually reducing the loss.

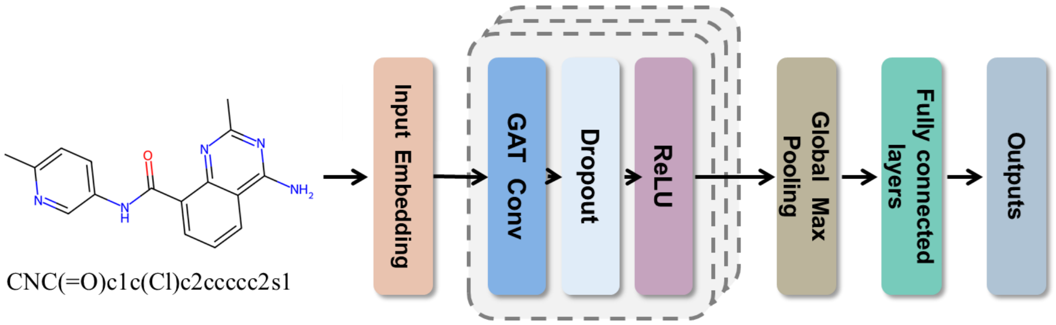

Fig. 3. Workflow of molecular prediction. The prediction process begins by converting input SMILES strings into molecular graphs. These graphs are passed through multiple GAT convolutional layers, which extract graph-level features. Features are processed with global max pooling and then fed to fully connected layers for binding affinity prediction.

Fine-tuning process and tournament selection

The fine-tuning process utilizes a framework that integrates generative and predictive models, with experience memory and tournament selection as its core components (Fig. 4). Experience memory serves as a repository for molecules. Initially, the experience memory is populated with unique, chemically valid molecules generated by the generative model before the fine-tuning loop begins. In the fine-tuning loop, the final winning molecules from tournament selection are stored in the experience memory. The competitor pool is formed by merging SMILES sampled from the generator with an equal number sampled from the experience memory. Tournament selection is used as a strategy to select superior molecules from the pool of sampled candidates [15]. At each stage, molecules compete based on specific criteria, and the winners survive to the next stage. The tournament selection process consists of three stages: the first stage evaluates the predicted binding affinity of the molecules; the second stage evaluates the negative log-likelihood from the generator; and the third stage evaluates the positive log-likelihood from the prior model. In each stage, two molecules are randomly selected from the pool, and the molecule with the higher score is selected as the winner. The winner advances to the next stage and the loser is excluded, so that only half of the molecules survive at each stage. This iterative process enables the generator to focus on molecules with high binding affinities. The framework allows the generator to explore new molecular structures while maintaining structural features that contribute to high biological activity.

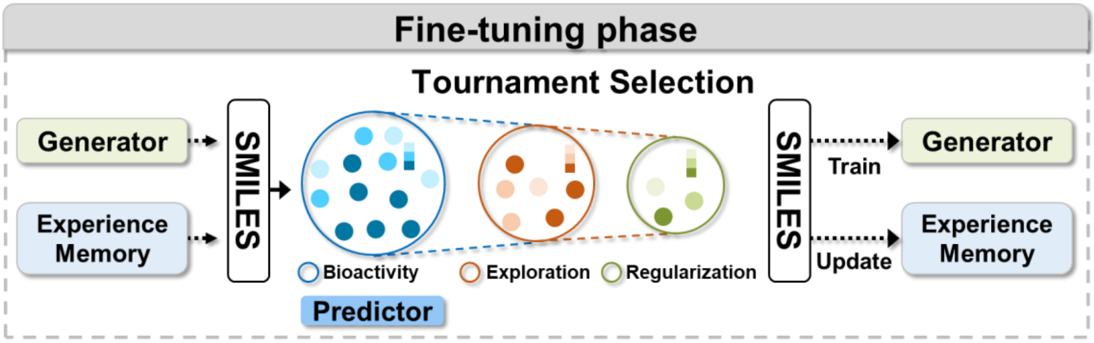

Fig. 4. The flowchart of fine-tuning. SMILES generated by the generator are combined with SMILES sampled from the experience memory to form the competitor pool. Then, through three stages of tournament selection, the final SMILES are selected. The selected SMILES are used to retrain the generator and update the experience memory.

Evaluation metrics

Performance evaluation in *de novo* molecular design uses metrics that differ from those employed in traditional machine-learning tasks such as regression and classification. In this study, we used eight metrics, grouped into two categories: (1) generative quality metrics and (2) distribution similarity metrics. Generative quality metrics include validity, uniqueness, novelty, internal diversity, and predicted bioactivity (PredAct), while distribution similarity metrics include pairwise similarity (PwSim), fréchet chemnet distance (FCD), and optimal transport distance (OTD). The generative quality metrics are defined as follows:

$$Validity = \frac{V}{20000} \tag{6}$$

$$Uniqueness = \frac{U}{V} \tag{7}$$

$$Novelty = \frac{T}{U} \tag{8}$$

$$Internal\ Diversity = \frac{1}{|V_{1000}|^2} \sum_{m_1 \in V_{1000}, m_2 \in V_{1000}} 1 - sim(m_1, m_2) \tag{9}$$

Validity assesses whether the generated molecules are chemically valid. In this process, RDKit was used [28]. We generated 20,000 molecules and calculated the proportion (V) of SMILES representing chemically valid structures. Uniqueness measures the diversity of the generated SMILES and is expressed as the ratio (U) of unique molecules among the valid set (V). A low uniqueness score suggests that the model repeatedly generates the same molecules, indicating a limited capacity for learning the distribution. Novelty is defined as the proportion (T) of valid, unique molecules that do not exist in the training dataset. A low novelty score indicates that the model is overfitting. Internal diversity assesses structural diversity within the generated SMILES. Among the valid molecules, 1,000 molecules were randomly selected, and the similarity between these molecules was calculated using the Tanimoto similarity. In Equation 9, $sim(m_1, m_2)$ represents the Tanimoto similarity between molecules m_1 and m_2, calculated using their Morgan fingerprint with a radius of 2 and 2048 bits. V_{1000} represents the subset of 1,000 randomly selected molecules from V. PredAct is defined as the average predicted bioactivity of molecules generated by the model. The distribution similarity metrics are defined as follows:

$$PwSim = \frac{1}{|V_{1000}||T|}\sum_{m_1 \in V_{1000},\, m_2 \in T} sim(m_1, m_2) \tag{10}$$

$$FCD = ||\ \mu_v\ -\ \mu_T\ ||^2 + Tr(C_V\ +\ C_T\ -\ 2(C_V C_T)^{1/2}) \tag{11}$$

$$OTD\ =\ argmin_{T \in R}\sum_{x_i \in A\ ,y_j \in B} T_{ij} dist(x_i, y_j) \tag{12}$$

$$dist(x_i, y_j) = 10^{1-sim(x_i, y_j)} - 1 \tag{13}$$

PwSim measures the average pairwise similarity between the generated SMILES and active molecules in the test dataset (Equation 10). In this equation, T denotes the set of target active molecules in the test set. FCD measures the difference between two probability distributions, specifically between the generated molecule set and the target molecule set, using the Fréchet distance (Equation 11). This metric quantifies how dissimilar the two sets are by comparing the means and covariance of their distributions, assuming both follow gaussian distributions. Here, μ_V and μ_T represent the means of the feature vectors for the generated molecule set V and the target molecule set T, respectively, while C_V and C_T represent their covariance matrices. Tr represents the trace of a matrix, which is the sum of its diagonal elements. OTD calculates the optimal transport cost between the two probability distributions of the generated molecule set and the target molecule set, thereby measuring the distance between them. This method is defined in terms of the similarity between probability distributions. Higher similarity results in lower OTD values (Equation 12). In this formulation, T_{ij} represents the amount of mass transported from molecule x_i to molecule y_j. R is the set of all possible transport plans between the generated molecule set A and the target active set B. $dist\ (x_i, y_j)$ represents the distance used for OTD calculation. The performance evaluation metrics used in this study are similar to those used in previous studies [2, 15]. Except for OTD and FCD, higher metric values indicate better performance.

Results

Performance of pre-trained generator and predictor

We evaluated the performance of the pre-trained generative and predictive models before fine-tuning. Specifically, we enhanced the basic SMILES generation capability of the generative model by experimenting with different loss functions and applying multi-scale attention mechanism. We also adjusted the temperature parameter to control sampling stochasticity and identified the optimal balance between validity and uniqueness during SMILES generation. We conducted an ablation study that evaluated how each module affected the generator's ability. The results provided insights into how these modifications affected the model's performance in terms of validity, uniqueness, and novelty.

We randomly selected 50 scaffolds from the test set as input conditions to comprehensively evaluate the model’s generalization capability. We generated 10,000 SMILES for each scaffold and evaluated the resulting molecules using the chosen metrics. The performance for SMILES generated from each scaffold is presented in Table 2, showing the top 10 scaffolds. Across all scaffolds, the pre-trained generator achieved validity greater than 0.9, uniqueness greater than 0.9, and novelty of 1.0. Additionally, it achieved an average validity of 0.968, uniqueness of 0.966, and novelty of 1.0. The generator's performance varied significantly depending on the input scaffold. This variation occurred because each scaffold has distinct structural and chemical properties that affect the complexity and feasibility of generating valid molecules. Scaffolds with more flexible structures tended to accommodate a wider variety of substituents and chemical modifications, leading to higher validity and uniqueness in the generated SMILES. Conversely, rigid or highly constrained scaffolds limited the diversity of feasible molecules, affecting the generation performance. For instance, the scaffold 'O=C(Cc1ccccc1)NCc1ccccc1' contains multiple rotatable single bonds that connect two phenyl rings through a benzyl–amide linkage, conferring higher conformational flexibility. This scaffold achieved a higher validity of 0.98 and uniqueness of 0.983. In contrast, the scaffold 'O=C1CCC(c2ccc(NCc3ccccc3)cc2)=NN1' features ring closure and unsaturation within the ring system that reduce the number of rotatable bonds and impose conformational restriction, thereby yielding a more rigid framework. Consequently, it resulted in a lower validity of 0.958 and uniqueness of 0.954. This result supports the assertion that scaffold flexibility positively affects the generator's performance in terms of validity and uniqueness.

Table 2. Performance evaluation of SMILES generated for each scaffold

Scaffold	Validity	Uniqueness	Novelty
O=C(Cc1ccccc1)NCc1ccccc1	0.980	0.983	1.0
c1ccc(-c2ccnnc2)cc1	0.985	0.944	1.0
c1ccc(C2=NOCC2)cc1	0.968	0.982	1.0
c1ccc(Cc2cc3c(CNC4CCCCC4)cccc3o2)cc1	0.964	0.957	1.0
c1ccc(OCCn2ccnc2)c(CNCc2nccs2)c1	0.961	0.975	1.0
C(=NCC(c1ccccc1)N1CCCCC1)c1ccccc1	0.967	0.961	1.0
O=C(Nc1ccccc1)NC1CCC(OCc2ccccc2)CC1	0.981	0.955	1.0
O=C(COC(=O)c1ccccc1)Nc1ccccc1N1CCOCC1	0.962	0.962	1.0
O=C(Nc1ccccc1)c1cccc(N2C=CNN2)c1	0.954	0.991	1.0
O=C1CCC(c2ccc(NCc3ccccc3)cc2)=NN1	0.958	0.954	1.0
Average	0.968	0.966	1.0

We compared three generator variants, including a generator trained with cross-entropy as the loss function, a generator with multi-scale attention using scales 1 to 5, and a generator without multi-scale attention, to isolate the effects of the loss function and multi-scale attention. The cross-entropy variant achieved slightly higher validity but markedly lower uniqueness, whereas extending the multi-scale attention beyond the optimal range or removing it reduced both metrics. All models maintained a novelty of 1.0. Details are provided in Section 2 and Fig. S2 of Supplementary materials. In addition, we conducted experiments to identify the optimal temperature that achieves the best trade-off between validity and uniqueness. The temperature parameter modulates the probability distribution over candidate tokens during generation and thus controls randomness in sampling. As temperature increases, randomness rises and uniqueness improves at the expense of validity, whereas lower temperatures have the opposite effect. In our experiments, a temperature of 0.9 provided the most balanced result. Probability profiles and full results are provided in Section 3, Fig. S3, and Table S1 of Supplementary materials.

The pre-trained predictor for PIK3CA achieved an MSE of 0.444 and an R^2 of 0.744, while for KOR, the MSE was 0.416 and the R^2 was 0.788. Following this assessment of the predictive models for PIK3CA and KOR, we present visual analyses to illustrate their performance (Fig. 5). Fig. 5a and Fig. 5b show scatter plots of predicted versus actual values, demonstrating the correlation and predictive accuracy. In both datasets, the predicted and actual biological activity values show a high overall correlation. The data points are densely clustered around the red solid line, which suggests that the models predict the actual values well. Fig. 5c and Fig. 5d show residual histograms, illustrating the distribution of prediction errors and the consistency of the predictors. In both datasets, the residuals are symmetrically distributed around a mean of zero, indicating that the models' predictions are generally unbiased. For KOR, the distribution is narrower, with over 95% of the residuals distributed between -1 and 1. For PIK3CA, over 95% of the residuals are distributed between -1.5 and 1.5.

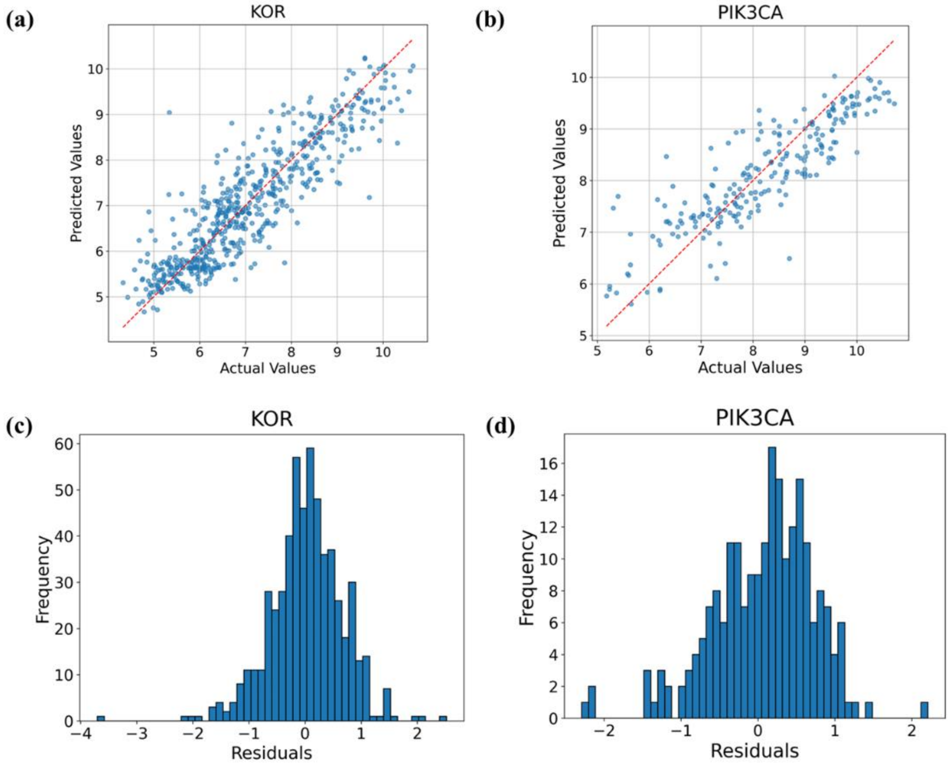

Fig. 5. Visualization of the predictor performance on the KOR and PIK3CA datasets. (a, b) Predicted and actual biological activity values for the test datasets of KOR and PIK3CA. The x-axis represents the actual values, and the y-axis represents the predicted values. The red line indicates the ideal case where the predicted values perfectly match the actual values. (c, d) Distribution of residuals (differences between actual and predicted biological activity values) for the KOR and PIK3CA datasets. The x-axis represents the residual values, and the y-axis represents the frequency of occurrence for each residual value.

Performance of fine-tuned generative model

We compared the results of our fine-tuned model with those of the previously proposed LOGICS framework [15]. We performed pre-training on identically preprocessed datasets for a fair comparison. Specifically, the generator and predictor used in the LOGICS framework were pretrained on identical datasets before fine-tuning, allowing for a direct performance comparison**.** The performance of both models on bioactivity datasets is summarized in Table 3.

Table 3. Performance evaluation of fine-tuning based on bioactivity

Performance Metrics	**KOR**		**PIK3CA**	
	Scaffold-aware Transformer	**LOGICS (KOR)**	**Scaffold-aware Transformer**	**LOGICS**
Validity	0.9802	0.9614	0.9803	0.9645
Uniqueness	0.9865	0.9997	0.9755	0.9994
Novelty	0.9998	0.9810	0.9993	0.9749
Internal Diversity	0.8477	0.8779	0.8645	0.8778
PredAct	6.7877	6.3272	7.6213	7.285
PwSim	0.1319	0.1249	0.1027	0.1085
FCD	25.7713	21.7733	36.0337	38.3491
OTD	5.4998	5.1118	6.186	5.829

The scaffold-aware transformer showed competitive results compared to LOGICS across multiple metrics [15]. Specifically, it achieved validity scores of 0.9802 (KOR) and 0.9803 (PIK3CA), surpassing LOGICS's 0.9614 and 0.9645. Novelty scores were also higher, with our model obtaining 0.9998 (KOR) and 0.9993 (PIK3CA) compared to LOGICS's 0.9810 and 0.9749. Additionally, our model generated molecules with higher predicted biological activity, achieving values of 6.7877 (KOR) and 7.6213 (PIK3CA) compared to LOGICS's 6.3272 and 7.2850. These results suggest that the scaffold-aware transformer outperforms LOGICS in terms of validity, novelty, and predicted biological activity, while LOGICS demonstrates higher uniqueness and internal diversity. Overall, the scaffold-aware transformer shows balanced performance and effectively generates valid and novel molecules with high predicted activity. This capability is crucial for discovering potential drug candidates.

We assessed the drug-likeness and synthetic accessibility of the 15,000 molecules generated by the fine-tuned model using QED and SAS [39, 40]. For KOR, around 20% of all molecules had a QED score above 0.6 and about 99% had an SAS below 5. For PIK3CA, around 46% had a QED above 0.6 and about 99% had an SAS below 5. Detailed summaries and distributions are provided in Section 4 and Fig. S4 of Supplementary materials.

Attention based substructure analysis

We identified substructures associated with the model’s prediction by using attention coefficients from a GAT. This approach offers substructure-level interpretability for the target and identifies molecular motifs that the model considers most important for predicting binding affinity. We cross-validated the highlighted substructures with prior studies on mechanisms of action, including reported binding modes and pharmacophores, to support the model’s interpretation. We performed this analysis on the PIK3CA-targeted drugs Copanlisib and Alpelisib, and the KOR-targeted drugs Nalmefene and Buprenorphine. The attention-highlighted substructures for all drugs are presented in Fig. 6. All drugs used in the analysis have received FDA approval [41-43].

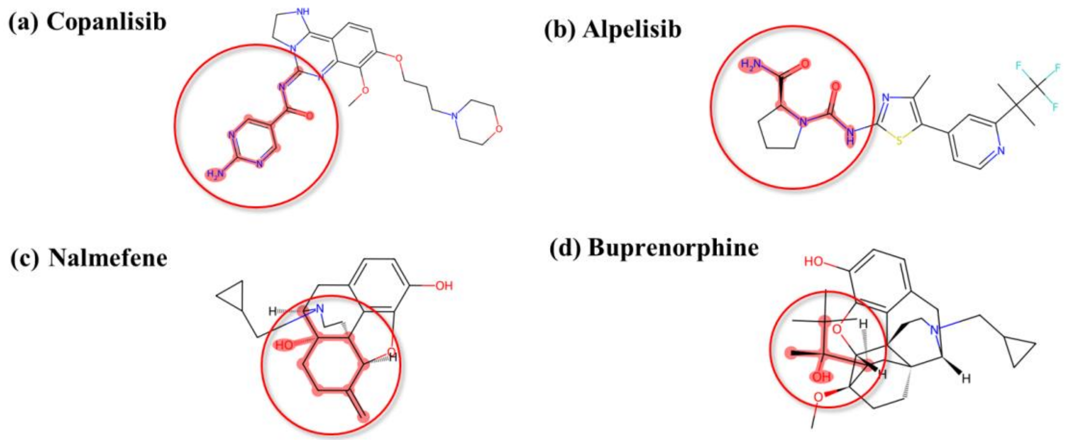

Fig. 6. Attention based substructure analysis for PI3Kα and KOR reference drugs. (a) Copanlisib (b) Alpelisib (c) Nalmefene (d) Buprenorphine. Red highlighted regions indicate substructures with high attention scores from the graph attention network.

Copanlisib attention analysis highlighted the aminopyrimidine group linked to the C5 amide of the dihydroimidazoquinazoline core. In the crystallographic binding model, this group occupies the affinity pocket and forms a three-point hydrogen bond network in which the exocyclic amine donates to Asp836 and Asp841 and a ring nitrogen accepts a hydrogen bond from Lys833 [44]. The fused bicyclic core also engages the hinge valine through a ring nitrogen and anchors the ligand within the ATP pocket. As reported in previous studies, the morpholinopropyl side chain extends toward solvent and primarily improves solubility with limited direct contribution to binding. Findings from lead optimization indicate that the C5 aminopyrimidine is the preferred substitution for potency, the C7 methoxy preserves pocket fit, and the C8 substituent was optimized to a morpholinylpropoxy group to tune properties and pharmacokinetics [45]. Taken together, these observations indicate that the aminopyrimidine-amide ensemble is a key substructure that influences PIK3CA inhibition.

For Alpelisib, the attention analysis highlighted the proline-derived carboxamide linked through a urea and the adjacent 2-aminothiazole motif. Prior structure-activity relationship (SAR) studies show that Alpelisib donates and accepts hydrogen bonds to Gln859 and to the backbone carbonyl of Ser854, engages the hinge Val851, and makes a water-mediated contact from Asp810 and Asp933 to the pyridine nitrogen, with the charged Lys802 positioned near the trifluoromethyl group [41]. Complementary SAR and docking studies further indicate that contacts with Gln859, Ser854, and Val851 are central to selectivity and binding, and that the 2-aminothiazole scaffold with an L-proline-derived carboxamide promotes selectivity for the PI3Kα subtype. These cross-validated findings support the highlighted motif as the decisive substructure for α-selective binding [41, 46, 47].

The attention analysis for Nalmefene highlights the phenolic ring and the adjacent hydroxyl-substituted ring. According to prior docking studies on KOR, Nalmefene forms three hydrogen bonds in the KOR site. The hydroxyl group of Tyr139 forms a hydrogen bond with a ligand oxygen, and the ligand hydroxyl forms hydrogen bonds with the nitrogen of Gln115 and the oxygen of Asp138 [43]. In our visualization, the highlighted substructure captures the hydroxyl-rich region that can interact with Gln115 and Asp138, while the Tyr139 contact is not prioritized by the model's attention. The model therefore regards this highlighted moiety as the key motif that most strongly explains Nalmefene's binding affinity to KOR.

Buprenorphine attention analysis highlighted the tertiary-alcohol motif and focused on the oxygen-containing segment. Prior docking work on KOR reports a single hydrogen bond in which the drug's hydroxyl hydrogen interacts with the oxygen of residue Ile304 [43]. The highlighted substructure captures the alcohol functionality capable of mediating this contact.

Discussion

This study contributes to the field of molecular generation by integrating scaffold-conditioned generation with an attention-based predictor [15, 23, 24]. Our approach demonstrated high validity and novelty in generating molecules and provided interpretable structure-activity explanations. These findings suggest that our model can generate new molecules with desired properties and has the potential to advance *de novo* molecular design.

Our ablation study shows that architectural choices and sampling control are key determinants of generation quality. Using focal loss with multi-scale scaffold attention improved uniqueness relative to cross-entropy, and a sampling temperature of 0.9 provided the most balanced validity–uniqueness trade-off (Supplementary materials, Sections 2–3). Attention-based substructure analysis provided target-level interpretability. For PI3Kα, highlighted motifs aligned with literature-reported contacts at Val851, Ser854, and Gln859. For KOR, the model emphasized hydroxyl-rich regions consistent with contacts to Gln115 and Asp138 for Nalmefene, and the Ile304 hydrogen bond for Buprenorphine, while deprioritizing the Tyr139 interaction. These comparisons with prior reports indicate that the features highlighted by the model are consistent with reported chemical interactions, rather than incidental. The generated molecules also showed practical chemistry profiles. Most had SAS values within ranges consistent with feasible synthesis, and many exhibited moderate to high QED (Supplementary materials, Section 4).

There are several limitations to this study. First, the generalization capability of the model is limited. This study was conducted using only two datasets, KOR and PIK3CA. Such a restricted range of datasets may limit the evaluation of the model's generalizability. Applying the model to targets related to various diseases, such as cancer and metabolic disorders, could provide a more comprehensive assessment of its performance. Second, there is a risk of overfitting to scaffold structures. Scaffold-based generation offers the advantage of generating new molecules while maintaining desired scaffold structures; however, there is a risk of the model overfitting to specific scaffold configurations. This overfitting can diminish the diversity of generated molecules and reduce the overall effectiveness of the model. Since the selection and definition of scaffolds directly affect the model's performance and generation outcomes, strategies to increase scaffold diversity and prevent overfitting are necessary. For example, using multiple scaffolds simultaneously or adopting training methods that consider the structural diversity of scaffolds could mitigate this issue.

This study also suggests several ways for extension. First, multi-objective fine-tuning that optimizes predicted bioactivity together with ADMET-related proxies such as permeability, clearance, and safety risk could bring the generated molecules closer to downstream developability needs. Second, incorporating structure-based signals such as receptor-specific constraints and physics-guided priors into the training loop may further strengthen the link between attention-derived motifs and true binding determinants. Overall, our integration of scaffold-conditioned generation with a cyclic learning mechanism represents a novel contribution to the field, potentially advancing the development of more effective drug discovery methods.

Conclusion

In this study, we introduced a scaffold-aware generative framework that integrates a transformer-based generator and a GAT-based predictor [23, 24]. In this study, we introduced a scaffold-aware generative framework that integrates a transformer-based generator and a GAT-based predictor [23, 24]. By incorporating multi-scale attention mechanisms, our approach enables explicit scaffold control while exploring chemical diversity around user-specified core structures. A cyclic learning mechanism with tournament selection and experience memory facilitates continuous optimization toward high-affinity, scaffold-consistent candidates [15]. Experimental results on KOR and PIK3CA targets demonstrated that the proposed method achieves high validity and novelty while generating molecules with higher predicted biological activity compared to baseline approaches. Attention-based analysis of FDA-approved drugs revealed that the model highlights substructures consistent with known binding interactions, providing interpretable insights into structure-activity relationships. Assessment of drug-likeness and synthetic accessibility revealed that the generated molecules exhibit practical chemistry profiles, with the majority showing favorable synthetic feasibility. Ablation studies confirmed that the combination of focal loss and multi-scale attention mechanisms significantly improves generation quality, and demonstrated that appropriate temperature control achieves an optimal balance between validity and uniqueness. This study presents a balanced and effective approach for generating novel bioactive molecules, highlighting its potential applicability in drug discovery and material design. Future research is expected to contribute to the generation of more complex molecular structures and the expansion of models to consider a broader range of biological properties.

CRediT authorship contribution statement

Junyoung Park : Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Sunyong Yoo : Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Ministry of Food and Drug Safety (MFDS) grants RS-2024-00332003 (2024) and RS-2025-02215961 (2025), the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare (MOHW), Republic of Korea, grant number RS-2025-19252970, and the Ministry of Science and ICT (MSIT) support program grant number RS-2025-16063391.

Appendix A. Supplementary data

The following is the supplementary data to this article.

References

1. Berdigaliyev N, Aljofan M: An overview of drug discovery and development. Future medicinal chemistry 2020, 12(10):939-947.
2. Bagal V, Aggarwal R, Vinod P, Priyakumar UD: MolGPT: molecular generation using a transformer-decoder model. Journal of Chemical Information and Modeling 2021, 62(9):2064-2076.
3. Polishchuk PG, Madzhidov TI, Varnek A: Estimation of the size of drug-like chemical space based on GDB-17 data. Journal of computer-aided molecular design 2013, 27:675-679.
4. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL: How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature reviews Drug discovery 2010, 9(3):203-214.
5. Food, Administration D: Safety clinical trial shows possible increased risk of cancer with weight-loss medicine Belviq, Belviq XR (lorcaserin). In.; 2020.
6. Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P: Drug target identification using side-effect similarity. Science 2008, 321(5886):263-266.
7. Aleo MD, Luo Y, Swiss R, Bonin PD, Potter DM, Will Y: Human drug-induced liver injury severity is highly associated with dual inhibition of liver mitochondrial function and bile salt export pump. Hepatology 2014, 60(3):1015-1022.
8. Sun D, Gao W, Hu H, Zhou S: Why 90% of clinical drug development fails and how to improve it? Acta Pharmaceutica Sinica B 2022, 12(7):3049-3062.
9. Mohs RC, Greig NH: Drug discovery and development: Role of basic biological research. Alzheimer's & Dementia: Translational Research & Clinical Interventions 2017, 3(4):651-657.
10. Mak K-K, Pichika MR: Artificial intelligence in drug development: present status and

future prospects. Drug discovery today 2019, 24(3):773-780.

11. Elton DC, Boukouvalas Z, Fuge MD, Chung PW: Deep learning for molecular design—a review of the state of the art. Molecular Systems Design & Engineering 2019, 4(4):828-849.

12. Blanco-Gonzalez A, Cabezon A, Seco-Gonzalez A, Conde-Torres D, Antelo-Riveiro P, Pineiro A, Garcia-Fandino R: The role of AI in drug discovery: challenges, opportunities, and strategies. Pharmaceuticals 2023, 16(6):891.

13. Hinkson IV, Madej B, Stahlberg EA: Accelerating therapeutics for opportunities in medicine: a paradigm shift in drug discovery. Frontiers in pharmacology 2020, 11:770.

14. Qureshi R, Irfan M, Gondal TM, Khan S, Wu J, Hadi MU, Heymach J, Le X, Yan H, Alam T: AI in drug discovery and its clinical relevance. Heliyon 2023, 9(7).

15. Bae B, Bae H, Nam H: LOGICS: Learning optimal generative distribution for designing de novo chemical structures. Journal of Cheminformatics 2023, 15(1):77.

16. Gupta A, Müller AT, Huisman BJ, Fuchs JA, Schneider P, Schneider G: Generative recurrent networks for de novo drug design. Molecular informatics 2018, 37(1-2):1700111.

17. Meyers J, Fabian B, Brown N: De novo molecular design and generative models. Drug discovery today 2021, 26(11):2707-2715.

18. Weininger D: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences 1988, 28(1):31-36.

19. Olivecrona M, Blaschke T, Engkvist O, Chen H: Molecular de-novo design through deep reinforcement learning. Journal of cheminformatics 2017, 9(1):48.

20. Devi RV, Sathya SS, Coumar MS: Evolutionary algorithms for de novo drug design–A survey. Applied Soft Computing 2015, 27:543-552.

21. Zhao H: Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. Drug discovery today 2007, 12(3-4):149-155.

22. Langevin M, Minoux H, Levesque M, Bianciotto M: Scaffold-constrained molecular generation. Journal of chemical information and modeling 2020, 60(12):5637-5646.

23. Vaswani A: Attention is all you need. Advances in Neural Information Processing Systems 2017.

24. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y: Graph attention networks. arXiv preprint arXiv:171010903 2017.

25. Brown N, Fiscato M, Segler MH, Vaucher AC: GuacaMol: benchmarking models for de novo molecular design. Journal of chemical information and modeling 2019, 59(3):1096-1108.

26. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M: ChEMBL: towards direct deposition of bioassay data. Nucleic acids research 2019, 47(D1):D930-D940.

27. Bemis GW, Murcko MA: The properties of known drugs. 1. Molecular frameworks. Journal of medicinal chemistry 1996, 39(15):2887-2893.

28. Landrum G: Rdkit: Open-source cheminformatics software. 2016.

29. Mazuz E, Shtar G, Shapira B, Rokach L: Molecule generation using transformers and policy gradient reinforcement learning. Scientific Reports 2023, 13(1):8799.

30. Wang Y, Zhao H, Sciabola S, Wang W: cMolGPT: A conditional generative pre-trained transformer for target-specific de novo molecular generation. Molecules 2023, 28(11):4430.

31. Gu Y, Li J, Kang H, Zhang B, Zheng S: Employing molecular conformations for ligand-based virtual screening with equivariant graph neural network and deep multiple instance learning. Molecules 2023, 28(16):5982.

32. Bai X, Yin Y: Exploration and augmentation of pharmacological space via adversarial auto-encoder model for facilitating kinase-centric drug development. Journal of Cheminformatics 2021, 13:1-15.

33. Strömbergsson H, Kryshtafovych A, Prusis P, Fidelis K, Wikberg JE, Komorowski J, Hvidsten TR: Generalized modeling of enzyme–ligand interactions using proteochemometrics and local protein substructures. Proteins: Structure, Function, and Bioinformatics 2006, 65(3):568-579.

34. Rumelhart DE, Hinton GE, Williams RJ: Learning representations by back-propagating errors. nature 1986, 323(6088):533-536.

35. Hochreiter S: Long Short-term Memory. Neural Computation MIT-Press 1997.

36. Guo Q, Qiu X, Liu P, Xue X, Zhang Z: Multi-scale self-attention for text classification. In: Proceedings of the AAAI conference on artificial intelligence: 2020. 7847-7854.

37. Hendrycks D, Gimpel K: Gaussian error linear units (gelus). arXiv preprint arXiv:160608415 2016.

38. Lin T: Focal Loss for Dense Object Detection. arXiv preprint arXiv:170802002 2017.

39. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL: Quantifying the chemical beauty of drugs. Nature chemistry 2012, 4(2):90-98.

40. Ertl P, Schuffenhauer A: Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of cheminformatics 2009, 1:1-11.

41. Vanhaesebroeck B, Perry MW, Brown JR, André F, Okkenhaug K: PI3K inhibitors are finally coming of age. Nature reviews Drug discovery 2021, 20(10):741-769.

42. Nallani SC, Li Z, Florian J, Xu Y, Sabarinath S, Brescia-Oddo T, Roca RA, Uppoor RS, Mehta MU: FDA Approval Summary: Nalmefene Nasal Spray for the Emergency Treatment of Known or Suspected Opioid Overdose. Clinical Pharmacology &

Therapeutics 2025, 117(3):620-626.

43. Feng H, Jiang J, Wei G-W: Machine-learning repurposing of DrugBank compounds for opioid use disorder. Computers in biology and medicine 2023, 160:106921.

44. Scott WJ, Hentemann MF, Rowley RB, Bull CO, Jenkins S, Bullion AM, Johnson J, Redman A, Robbins AH, Esler W: Discovery and SAR of novel 2, 3-dihydroimidazo [1, 2-c] quinazoline PI3K inhibitors: identification of copanlisib (BAY 80-6946). ChemMedChem 2016, 11(14):1517-1530.

45. Krause G, Hassenrück F, Hallek M: Copanlisib for treatment of B-cell malignancies: the development of a PI3K inhibitor with considerable differences to idelalisib. Drug design, development and therapy 2018:2577-2590.

46. Jin R-Y, Tang T, Zhou S, Long X, Guo H, Zhou J, Yan H, Li Z, Zuo Z-Y, Xie H-L: Design, synthesis, antitumor activity and theoretical calculation of novel PI3Ka inhibitors. Bioorganic Chemistry 2020, 98:103737.

47. Yin Z, Hu W, Zhang W, Konno H, Moriwaki H, Izawa K, Han J, Soloshonok VA: Tailor-made amino acid-derived pharmaceuticals approved by the FDA in 2019. Amino Acids 2020, 52(9):1227-1261.

Supplementary Material

Novel Molecular Design via a Scaffold-Aware Transformer with Multi-Scale Attention Mechanisms

Junyoung Park [a,b], Sunyong Yoo [a,b*]

[a] Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju, Republic of Korea

[b] R&D Center, MATILO AI Inc., Gwangju, Republic of Korea

*** Corresponding author. Chonnam National University, 77 yongbong-ro, Buk-gu, Gwangju, 61186, Korea, College of Engineering, Building 7, Republic of Korea**

E-mail addresses : **sss206391@gmail.com (J. Park), syyoo@jnu.ac.kr (S. Yoo)**

Section 1) SMILES sequence generation process

The SMILES sequence generation process is depicted in Fig. S1. Before sequence generation, the generator requires predefined initial tokens and scaffold conditions. The generation process starts by providing the start token [SOS] and the desired scaffold condition as inputs to the generator. The generator predicts the next-token probability distribution by applying the softmax function to its output logits. According to this probability distribution, the next token is sampled and added to the current sequence. The extended sequence is then used again as input to the generator to predict the subsequent token. This process repeats until the [EOS] token is generated or the length of the generated sequence reaches the predefined maximum length of 100 tokens. Once generation is complete, the generator returns a sequence of token indices. Using the predefined vocabulary, these indices are converted into their corresponding tokens, which are then concatenated to obtain the final SMILES string.

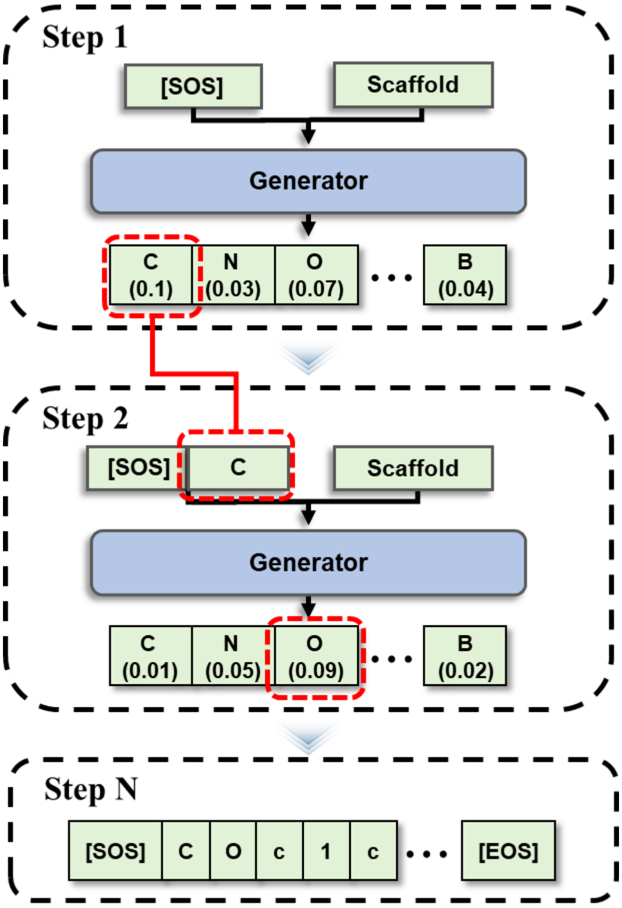

Fig. S1. The process of new molecular generation. This figure illustrates the step-by-step process of generating new molecules using the generator. In the first step, the next token (C) is predicted based on the [SOS] token and the input scaffold. The predicted token is concatenated with the [SOS] token. In the second step, the next token (O) is predicted based on the concatenated sequence and the scaffold. This process continues until an [EOS] token is generated or up to 100 iterations.

Section 2) Ablation study to examine the impact of different components on the performance of the generative model

We conducted an ablation study to examine the impact of different components on the performance of the generative model, comparing three variants: (i) generator trained using cross-entropy as the loss function, (ii) generator trained with multiscale attention using scales 1 through 5, and (iii) generator trained without applying multiscale attention. We compared these models to evaluate how the loss function and the application of multiscale attention affect the quality and diversity of the generated molecules. The results highlight the significance of multiscale attention mechanisms and appropriate loss functions in enhancing the model's ability to generate molecules with desired properties. The performance differences between the proposed model and these ablated models are presented in Fig. S2. The generator using cross-entropy achieved a validity 0.014 higher than the proposed model but recorded a uniqueness 0.17 lower. The generator with multiscale attention using scales 1 through 5 achieved validity and uniqueness 0.07 lower than the proposed model. Lastly, the generator without applying multiscale attention recorded a validity 0.019 lower and a uniqueness 0.159 lower than the proposed model. All models achieved a novelty of 1.0. While the model using cross-entropy achieved higher validity than the proposed model, its uniqueness was markedly lower. The two remaining variants recorded lower values in both validity and uniqueness.

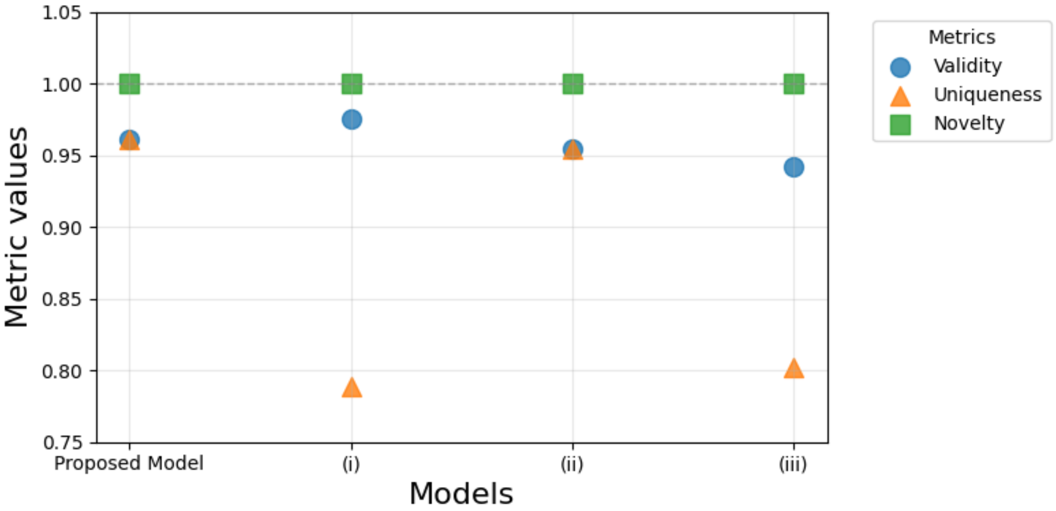

Fig. S2. Results of the ablation study on the proposed model. The figure compares the performance of three variants: (i) generator trained using cross-entropy as the loss function, (ii) generator trained with multiscale attention using scales 1 through 5, and (iii) generator trained without applying multiscale attention. The x-axis represents each specific model, and the y-axis represents the metric values.

Section 3) Comparison of token-wise probability distributions at different temperatures during SMILES generation

Validity and uniqueness have a trade-off relationship. When validity decreases, the denominator in the calculation of uniqueness becomes smaller, leading to an increase in uniqueness. Temperature plays a critical role in this balance because it controls randomness during SMILES generation. In our model, the temperature parameter controls the randomness in selecting the next token. A higher temperature flattens this distribution, allowing the model to explore a wider range of possible tokens. This exploration increases diversity and uniqueness in the generated molecules but can lead to syntactically incorrect or chemically invalid SMILES, thereby reducing validity. Conversely, a lower temperature sharpens the probability distribution, making the model more likely to select the most probable tokens. This increases the likelihood of generating valid molecules but may result in repetitive or similar structures, decreasing uniqueness. As depicted in Fig. S3, the probability distributions at different temperatures demonstrate this effect. When the temperature is 0.7, the differences in probabilities among tokens are large, resulting in a sharper distribution. In contrast, at a temperature of 1.3, the differences in token probabilities decrease, leading to a flatter distribution. Thus, finding the optimal temperature parameter is crucial. We compared the performance of the pre-trained generator at different temperature settings. When the temperature was 0.7, it achieved a high validity of 0.982 but recorded a uniqueness of 0.875. At temperature 1.0, validity dropped to 0.942, whereas uniqueness rose to 0.971. When the temperature was 0.9, both validity and uniqueness were 0.961, showing the most balanced performance.

The performance comparison of the generative model at different temperature settings is summarized in Table S1.

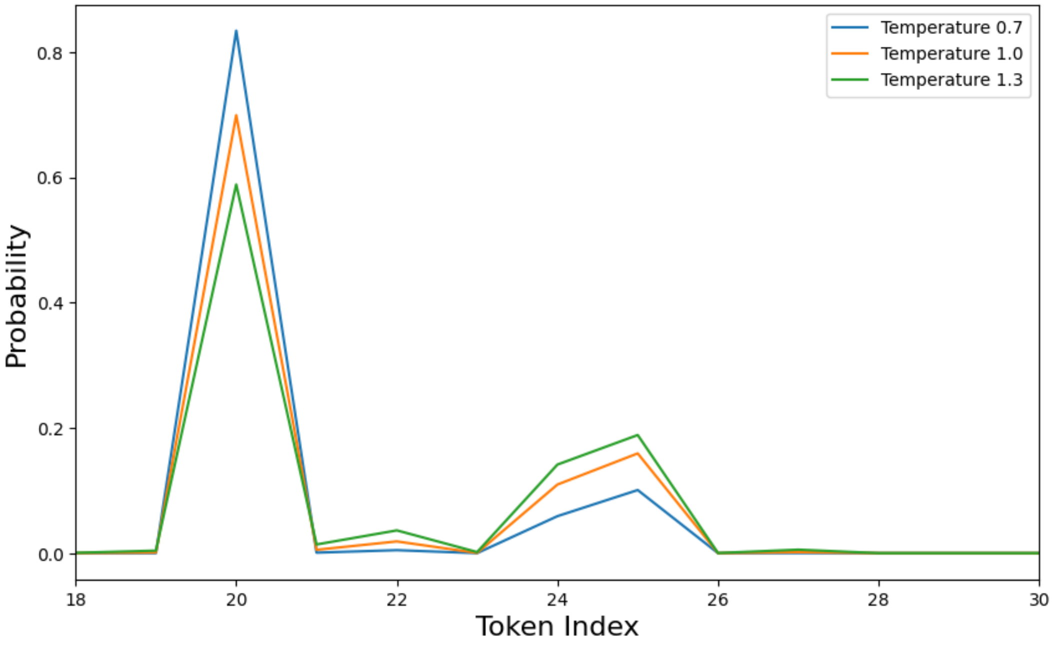

Fig. S3. Probability distributions over tokens at different temperature settings during SMILES generation. The x-axis represents the token index, and the y-axis represents the probability assigned to each token. The blue line represents the probability distribution at temperature 0.7, the red line represents the distribution at temperature 1.0, and the green line represents the distribution at temperature 1.3.

Table S1. The performance comparison of the generative model at different temperatures

Temperature	**Validity**	**Uniqueness**	**Novelty**
0.7	0.982	0.875	1.0
0.8	0.968	0.912	1.0
0.9	0.961	0.961	1.0
1.0	0.942	0.971	1.0

Section 4) Chemical properties of the generated molecules

We calculated the quantitative estimate of drug-likeness (QED) and the synthetic accessibility score (SAS) to evaluate the chemical properties of the generated molecules. QED evaluates the likelihood that a molecule is a potential drug candidate on a scale from 0 to 1 SAS evaluates the synthetic feasibility of a molecule on a scale from 1 to 10. Higher QED values indicate greater drug-likeness, suggesting that these molecules are more promising as drug candidates. Lower SAS values suggest that the molecules can be synthesized with relative ease. We subsequently generated 15,000 molecules using the fine-tuned model and calculated QED and SAS for each one. The distributions of QED and SAS for the generated molecules are presented in Fig. S4. In the case of KOR, the QED values ranged from a minimum of 0.016 to a maximum of 0.947. Around 20% of all molecules had a QED score greater than 0.6, and around 7% had QED greater than 0.8. The SAS values ranged from a minimum of 1.83 to a maximum of 5.91. Around 99% of all molecules had SAS less than 5, and around 32% had SAS less than 3. In the case of PIK3CA, the QED values ranged from a minimum of 0.028 to a maximum of 0.947. Around 46% of all molecules had QED greater than 0.6, and around 11% had QED greater than 0.8. The SAS values ranged from a minimum of 1.66 to a maximum of 5.04. Around 99% of all molecules had SAS less than 5, and 60% had SAS less than 3.

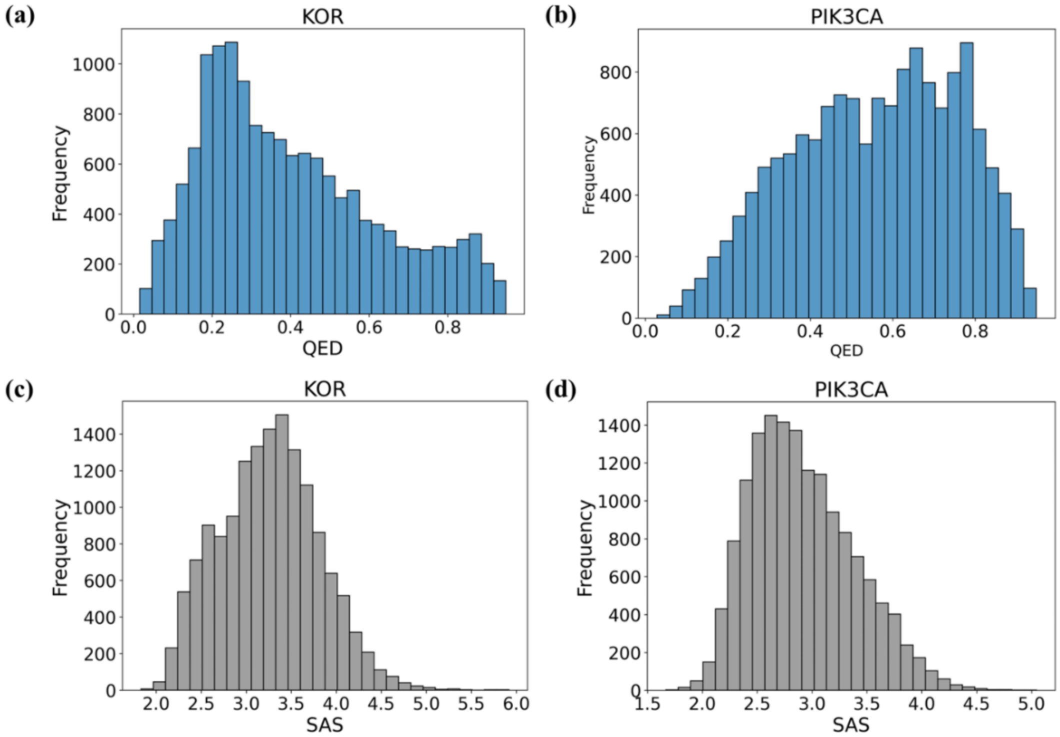

Fig. S4. Distributions of QED and SAS of the generated molecules. (a, b) QED distributions for KOR and PIK3CA datasets. (c, d) SAS distributions for KOR and PIK3CA datasets. The x-axis represents the QED or SAS, and the y-axis represents the frequency of molecules.