ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Enhancing *De Novo* Drug Design across Multiple Therapeutic Targets with CVAE Generative Models

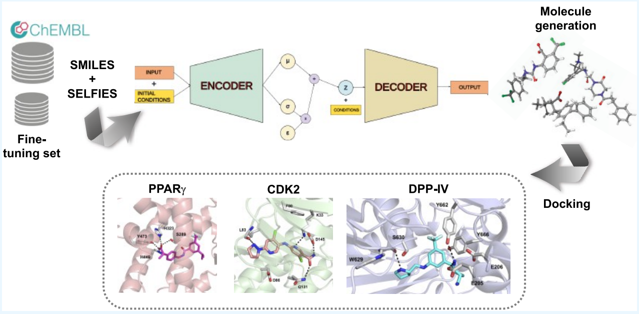
Virgilio Romanelli,[§] Daniela Annunziata,[§] Carmen Cerchia, Donato Cerciello, Francesco Picciali, and Antonio Lavecchia*

Cite This: *ACS Omega* 2024, 9, 43963−43976

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Drug discovery is a costly and time-consuming process, necessitating innovative strategies to enhance efficiency across different stages, from initial hit identification to final market approval. Recent advancement in deep learning (DL), particularly in *de novo* drug design, show promise. Generative models, a subclass of DL algorithms, have significantly accelerated the *de novo* drug design process by exploring vast areas of chemical space. Here, we introduce a Conditional Variational Autoencoder (CVAE) generative model tailored for *de novo* molecular design tasks, utilizing both SMILES and SELFIES as molecular representations. Our computational framework successfully generates molecules with specific property profiles validated though metrics such as uniqueness, validity, novelty, quantitative estimate of drug-likeness (QED), and synthetic accessibility (SA). We evaluated our model's efficacy in generating novel molecules capable of binding to three therapeutic molecular targets: CDK2, PPAR*γ*, and DPP-IV. Comparing with state-of-the-art frameworks demonstrated our model's ability to achieve higher structural diversity while maintaining the molecular properties ranges observed in the training set molecules. This proposed model stands as a valuable resource for advancing *de novo* molecular design capabilities.



INTRODUCTION

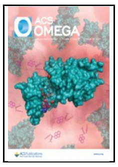
De novo drug design, which involves creating novel molecules with specific molecular properties such as molecular weight (M_W), polarity, and toxicity, poses a significant challenge in drug discovery.[1,2] Over recent years, the integration of artificial intelligence (AI) and deep learning (DL) into computer-aided drug design (CADD) has led to a wealth of groundbreaking discoveries.[3−5]

Generative models, a subset of DL algorithms, have made significant strides across diverse fields such as image generation,[6] speech recognition[7] and translation,[8] surpassing traditional machine learning (ML) approaches. Recent advancements in generative models have enabled the computational design of both chemically novel and synthetically viable compounds, facilitating exploration within the vast chemical space of drug-like molecules.[5,9−14]

The molecular generation process using generative models involves several steps that vary depending on the model's architecture. Typically, it starts with selecting a data set of reference molecular structures, which are then converted into a machine-readable format for training the model to recognize chemical patterns from existing molecules.[5,11,12] This iterative process continues until desired property scores are achieved. Generative models predominantly utilize deep neural networks (DNNs) to generate new compounds based on latent representations of molecular structures learned during model training.

To optimize the design of novel chemical structures, various generative models have been employed, including Recurrent Neural Network (RNN),[15] Generative Adversarial Networks (GAN),[16] Graph Neural Networks (GNNs)[17] and Variational Autoencoders (VAE).[18] Each method has its strengths and limitations, and there is no "champion" model that universally outperforms the others. One pioneering approach in this domain integrated a VAE model, comprising an encoder that translate SMILES strings into continuous vectors in latent space, and a decoder that convert these vectors back into SMILES, along with a molecular properties predictor.[19] However, this model suffered from a high incidence of generating invalid chemical structures. Another widely adopted model for molecular generation is the RNN,[10] which samples from learned distributions of input molecules to generate new SMILES-formatted structures. In the REINVENT model,

Received: August 31, 2024
Revised: September 25, 2024
Accepted: September 30, 2024
Published: October 18, 2024



https://doi.org/10.1021/acsomega.4c08027

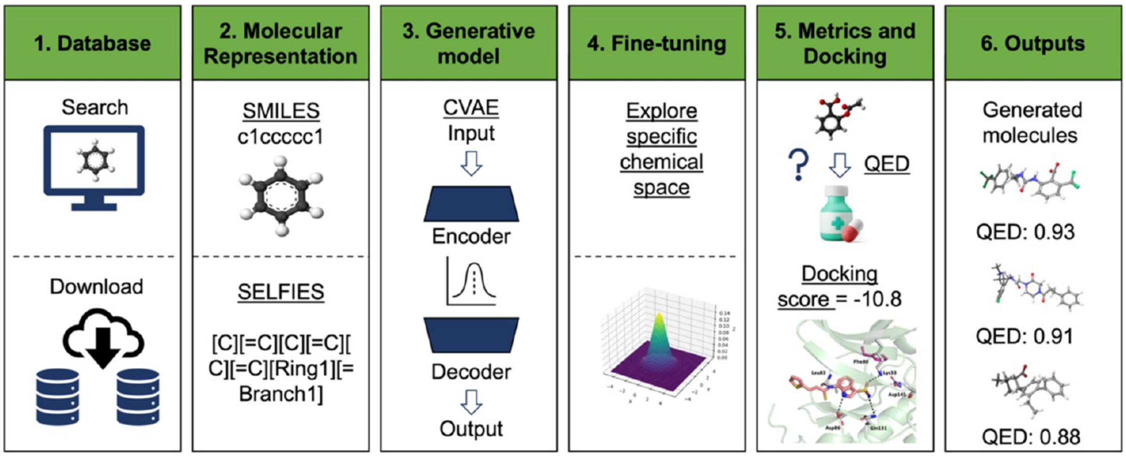


Figure 1. Workflow for molecule generation using the CVAE model.

which is based on RNN, a reinforcement/curriculum-learning (RL/CL) method was employed to fine-tune the pretrained RNN, enhancing its ability to generate structures with desired properties.[18,20,21]

In the MOLGEN model,[22] which is also based on RNNs, a GAN approach was employed. In this framework, the RNN serves as the generator, which is trained to produce molecular structures, while the discriminator evaluates the generated molecules against real samples. The adversarial training process fine-tunes the generator, enhancing its ability to produce valid molecules. This GAN-based approach allows MOLGEN to improve the diversity and quality of the generated molecular structures.

The Conditional VAE (CVAE), an extension of the classical VAE, has proven effective in multivariable control by incorporating molecular properties information into the encoding process. This integration enables manipulation of these properties during decoding within the latent space. This approach facilitates the design of compounds with specific attributes such as M_W, partition coefficient ($\log P$), and topological polar surface area (TPSA), collectively influencing compound's drug-likeness.

Kim et al.[23] introduced a novel approach to molecular design using a CVAE to generate molecules with specific properties. By leveraging CVAE, the authors successfully produced molecules exhibiting properties akin to those of Aspirin and Tamiflu. These molecules varied in structural configurations yet maintained a degree of similarity and comparable characteristics. Joo et al.[24] proposed a CVAE model for generating potential anticancer agents encoded as MACCS fingerprints (FPs). Training utilized a data set derived from NCI-60 drug screening, with normalized GI_{50} (growth inhibitory activity) values employed as conditional vectors. The CVAE effectively captured the distribution of molecular structures associated with anticancer activity, enabling the generation of new FPs with desired traits. Yang et al.[25] introduced a generative model based on a multiobjective CVAE and subsequently implemented Bayesian optimization guided by docking scores to enhance the biological activity of generated molecules.

In this study, we introduce a generative model based on the CVAE architecture, leveraging two prominent molecular representations: Simplified Molecular Input Line Entry System (SMILES) and Self-Referencing Embedded Strings (SELFIES). SMILES captures the sequential arrangement of atoms and bonds in a molecule,[26] while SELFIES represents a novel linear notation for constrained graphs, ensuring the generation of syntactically and semantically valid molecular structures.[27] This notation can be easily converted to and from other molecular representations. To evaluate the real-world potential of our model for *de novo* drug design, we focused on three therapeutic targets: Cyclin-Dependent Kinase 2 (CDK2), Dipeptidyl Peptidase IV (DPP-IV), and Peroxisome Proliferator-Activated Receptor γ (PPARγ). Our goal was to generate compounds capable of modulating these targets while exhibiting drug-like properties. The quality of the generated molecules was evaluated using established metrics to evaluate the model's validity. Additionally, molecular docking was employed to determine the binding affinity of the generated molecules within each target's active site, aiming to replicate the interactions of known active compounds. Our workflow, outlined in Figure 1, demonstrates a robust and adaptable approach to drug discovery, easily generalizable to other molecular targets. This model represents a significant advancement in the field, providing a valuable tool for the development of novel therapeutic agents.

■ EXPERIMENTAL SECTION

Data Sets. For model training, we used the ChEMBL data set (version 22),[28−30] selecting only compounds involved in direct interactions (type "D") with human targets at the highest confidence level (score 9). We included compounds with specified equilibrium constants (K_i values) or IC_{50}, resulting in a set of 327,660 molecules.

A filtering procedure was conducted with the following criteria:

(i) Removal of duplicate entries.
(ii) Standardization using RDKit[31] to remove salts and stereochemical information.
(iii) Exclusion of molecules with SMILES strings outside the 24−82 token range.
(iv) Exclusion of molecules with
- Hydrogen Bond Donors (HBD) ≥10
- Hydrogen Bond Acceptors (HBA) ≥10
- Number of rings ≥8
- Rotatable bonds (RotB) ≥15
- Length of rings ≥9

Molecules were encoded into sequences of characters denoting specific structural attributes, combining tokens for

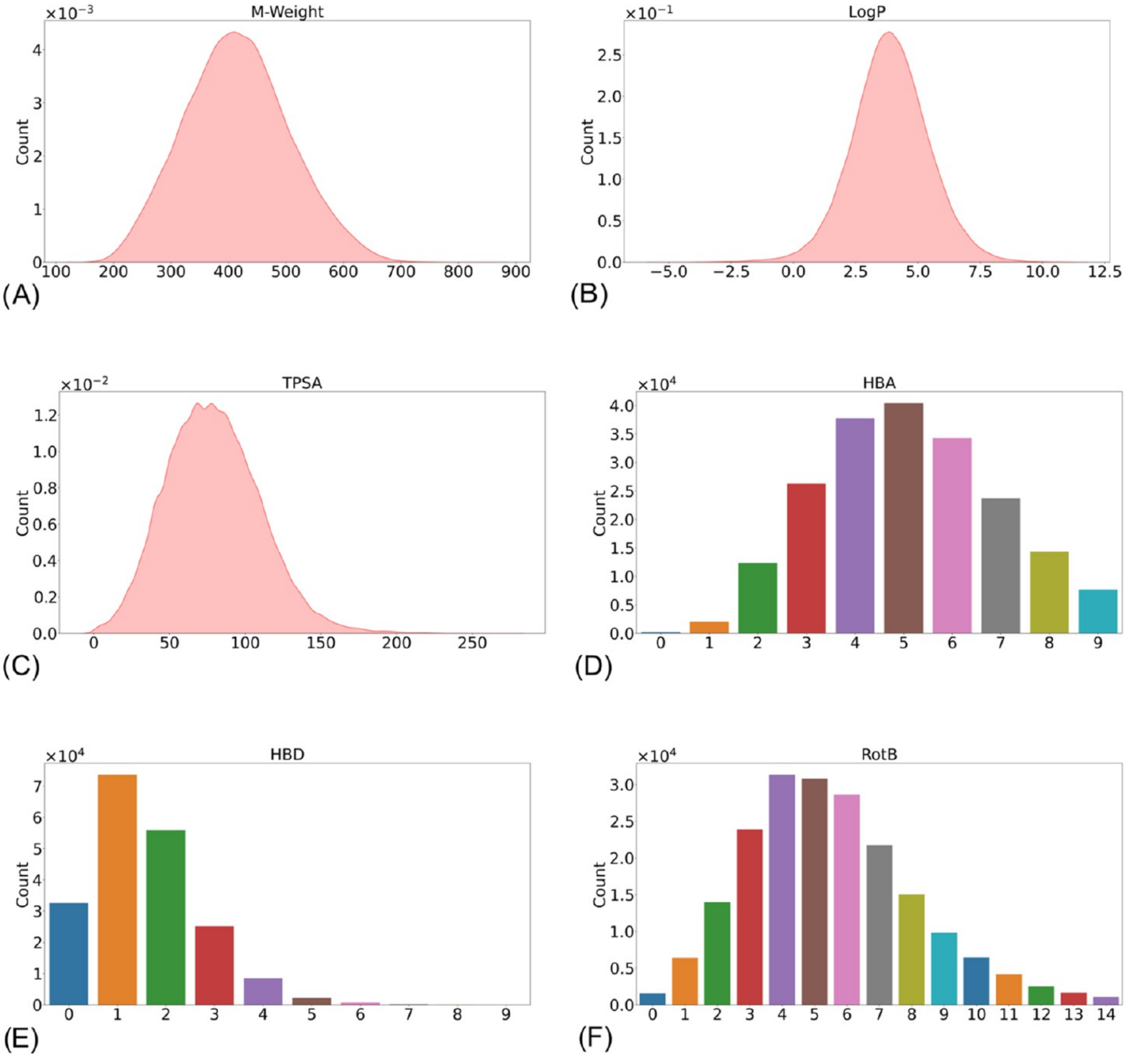


Figure 2. Distribution of calculated physicochemical properties across the final set of ChEMBL compounds. Panels (A−C) display kernel density estimate plots for M_W, log P, and TPSA, respectively. Panels (D−F) depict histograms for HBA, HBD, and RotB, respectively. These visualizations highlight the diversity and distribution range of these properties within the data set.

Table 1. Physicochemical Properties for CDK2, DPP-IV, and PPARγ Datasets after Filtering

data set	length of molecules	M_W	log P	TPSA	HBA	HBD	RotB	length of rings
CDK2	55.85 ± 9.14	393.98 ± 72.05	3.24 ± 1.19	94.02 ± 23.80	5.70 ± 1.79	2.27 ± 0.96	4.82 ± 2.29	3.74 ± 0.75
DDP-IV	52.84 ± 12.07	396.95 ± 80.81	2.25 ± 1.39	86.76 ± 27.34	5.24 ± 2.17	1.53 ± 0.78	4.91 ± 1.99	3.27 ± 1.00
PPARγ	62.39 ± 10.73	457.01 ± 72.39	5.25 ± 1.23	81.98 ± 21.37	5.05 ± 1.73	1.31 ± 0.65	8.85 ± 2.69	3.56 ± 0.97

atoms not found in the organic subset and without formal charge. To simplify the vocabulary and reduce its size, an additional filtering step was implemented:

(i) Removal of molecules with natural numbers ≥6, due to their low data set representation (3902 molecules, 1.86% of the total).

(ii) Encoding all molecules using only capital letters.

These refinements led to a comprehensive "dictionary" including {C, =, (,), N, O, 1, 2, 3, 4, F, S, 5, Cl, [O-1], [NH1], Br, #, [N-1], [N+1], [NH1+1], I, P, [S-1], [NH2+1], [S+1], B, [NH1−1], [Si], [C-1], [NH3+1], [Se], [B-1], [O+1], [PH1], [P+1], [2H], [SH1+1], [CH1−1], [Se+1], [OH1+1], [S+2], [Te+1], [Te], [SH1]}. To ensure uniformity, the length of each molecule was standardized to 120 characters by padding the sequences. The letters "X" and "E" were included in the vocabulary, denoting the start and end of molecules, respectively. The letter "E" is repeated as needed to pad each sequence to the desired length of 120 characters.

The final set consisted of approximately 198,962 molecules. Their physicochemical properties, including M_W, log P, TPSA, HBA, HBD, RotB, were calculated using RDKit (see Figure 2).

For model tuning, we collected three data sets of compounds active against CDK2, DPP-IV and PPARγ from ChEMBL (version 33), including compounds with activity values (IC_{50} for CDK2 and DPP-IV, EC_{50} for PPARγ) of at least 10 μM. The resulting sets included 1352 CDK2 inhibitors, 3911 DPP-IV inhibitors, and 2588 PPARγ agonists. The aforementioned filters were applied to these data sets, with properties detailed in Table 1.

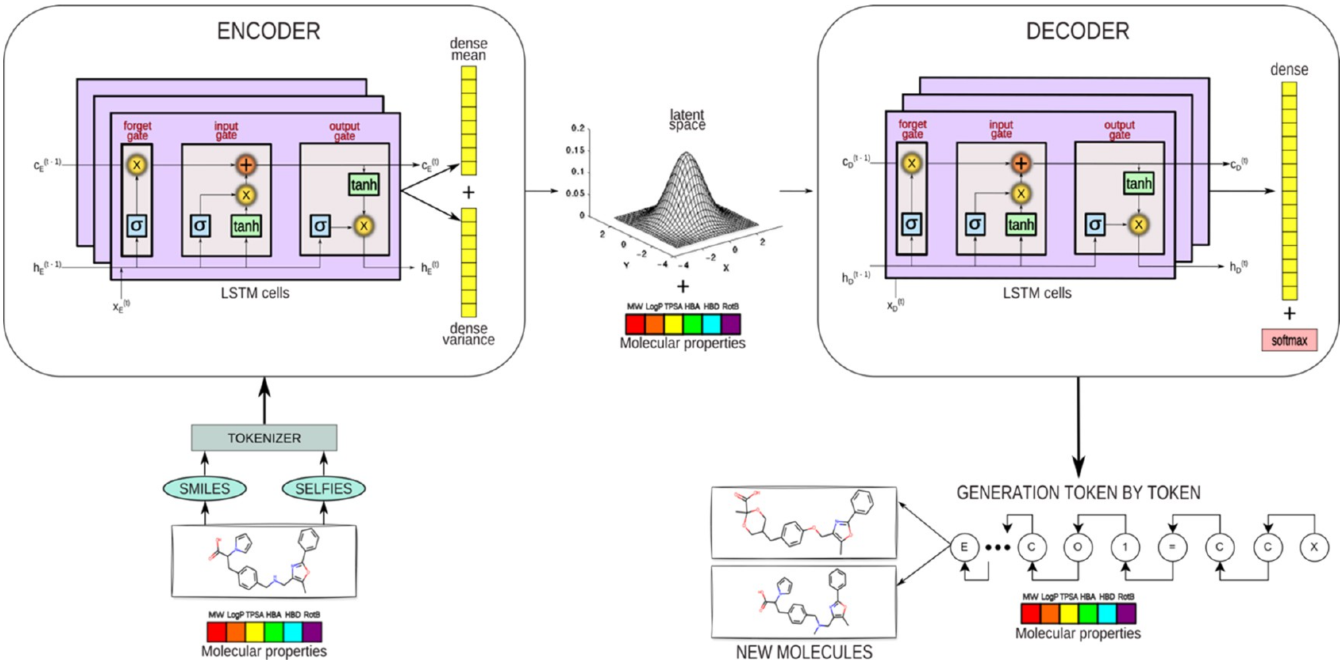


Figure 3. CVAE architecture for molecular generation. The chemical structure of a molecule is initially represented using SMILES or SELFIES, which are converted into a machine-readable sequence of tokens encoding molecular properties such as M_W, log P, TPSA, HBA, HBD, and RotB. The encoder, consisting of three LSTM layers, transforms these descriptors into a latent space representation visualized as a probability distribution function. The decoder, comprising additional LSTM layers, processes the latent representation to generate new chemical structures. A SoftMax layer at the end of the decoder produces a probability distribution.

Similar procedures were followed for SELFIES representations, utilizing the same data set as for SMILES generation and encoding each molecule using the SELFIES library.

CVAE Model Architecture and Training Process. The VAE forms the foundational framework of our model. It consists of an encoder-decoder neural network structure that learns to generate new data samples resembling those in then training data set by identifying a compact and continuous representation known as the latent space. The VAE operates by exploring a space x through a prior distribution over the latent space $\mathbb{P}_\theta(z)$ and a conditional likelihood of generating a data sample from the latent space $\mathbb{P}_\theta(x|z)$. The encoder maps the input to the posterior density $\mathbb{Q}_\phi(z|x)$ over the latent variable z using a multivariate Gaussian distribution $\mathbb{Q}_\phi(z|x) \sim N(\mu_\phi, \sigma_\phi^2)$. The goal of the VAE is to learn the marginal log-likelihood of the observed data in the generative process. Given that directly computing the log-likelihood is impractical, we rely on the Evidence Lower Bound (ELBO). This bound ensures that the posterior distribution $\mathbb{Q}_\phi(z|x)$ approximates the probability $\mathbb{P}_\theta(z|x)$:

$$D_{KL}[\mathbb{Q}_\phi(z|x)||\mathbb{P}_\theta(z|x)] = E_{\mathbb{Q}_\phi(z|x)}[\log \mathbb{Q}_\phi(z|x) - \log \mathbb{P}_\theta(z|x)] \tag{1}$$

where D_{KL} non-negative Kullback–Leibler divergence loss and E represents the expectation value. By applying Bayes' theorem and rearranging terms, eq 1 can be reformulated as

$$E_{\mathbb{Q}_\phi(z|x)}[\log \mathbb{Q}_\phi(z|x)] - D_{KL}(\mathbb{Q}_\phi(z|x)||\mathbb{P}_\theta(z)) = \log \mathbb{P}_\theta(x) - D_{KL}(\mathbb{Q}_\phi(z|x)||\mathbb{P}_\theta(z|x)) \tag{2}$$

The left-hand side of eq 2 is what we refer to as the ELBO. The VAE seeks to minimize the reconstruction term so that the encoder generates meaningful latent vectors for the decoder to reconstruct. In essence, it aims to optimize θ and ϕ to minimize the reconstruction error between the input and output and to make $\mathbb{Q}_\phi(z|x)$ as close as possible to $\mathbb{P}_\theta(z|x)$. Since maximizing the ELBO is equivalent to maximizing the log-likelihood of the observed data and minimizing the divergence between the approximate and true posterior, remembering that the Kullback–Leibler divergence is a non-negative function, the eq 2 can be rewritten as

$$\log(\mathbb{P}_\theta(x|z)) \geq E_{\mathbb{Q}_\phi(z|x)}[\log(\mathbb{P}_\theta(x|z))] - D_{KL}[\mathbb{Q}_\phi(z|x)||\mathbb{P}_\theta(z)] = L(\theta, \phi; x, z) \tag{3}$$

Instead, the decoder in a VAE is responsible for translating the latent variables back into the original data space. This means taking the sampled latent variables and producing an output that must be as close as possible to the original input. To achieve this, it focuses on maximizing the likelihood $\mathbb{P}_\theta(x|z)$ of the observed data x, given the latent variables z. The greater this likelihood, the more proficient the decoder becomes at reconstructing the original data from the latent space. Usually, the most used optimization loss in this part is the sum squared error:

$$\text{SSE}\ (x, x') = \sum_{i=1}^{D} (x_i^2 - x_i'^2) \tag{4}$$

where x represents the initial data, x' the final output and D the size of them. Combining eqs 3 and 4, we derive the loss function for VAE as

$$L(\theta) = \min \text{SSE}\,(x, x') + \max \text{ELBO} \tag{5}$$

Beginning with the VAE architecture, CVAE extends this framework by incorporating explicit condition vectors into the latent space representation, enabling the generation of molecules with specific desired properties. The CVAE adapts the latent vectors based on the condition vector without altering the network architecture or loss function. Consequently, the CVAE can control the dimensions of the latent space corresponding to the target properties defined in the condition vector. The ELBO objective function for CVAE is defined as

$$L(\theta, \phi; x, z, c) = E_{\mathbb{Q}_\phi(z|x)}[\log(\mathbb{P}_\theta(x|z, c))] - D_{\text{KL}}[\mathbb{Q}_\phi(z|x, c)||\mathbb{P}_\theta(z|c)] \tag{6}$$

where the primary distinction from the previous ELBO loss is the inclusion of the condition vector c, representing the desired molecular properties to be learned during molecule generation. Therefore, the decoder of the trained CVAE model can generate molecules with specified properties using both the condition vector and the latent space vectors.

Given our string-based representation, we enhance the CVAE architecture with Long Short-Term Memory (LSTM) cells, a type of Recurrent Neural Network (RNN), tailored to address the vanishing gradient problem inherent in sequential data. LSTMs integrate cell state and three gating mechanisms: the input gate, forget gate, and output gate. These gates regulate the flow of information within the cell, enabling effective handling of long-term dependencies. The input gate controls the amount of new information that is integrated into the cell state, while the forget gate determines which part of the previous cell state should be discarded. On the other hand, the output gate decides which portion of the cell state should be output and passed to the next time step. By incorporating these mechanisms, LSTM cells can effectively maintain and update the cell state over long sequences, enabling the network to capture long-term dependencies more efficiently.

Figure 3 illustrates the architecture of our model. LSTM cells are employed in both the encoder and decoder of the CVAE framework. After molecule embedding, the data traverses through three LSTM layers, each dimensioned according to specified hyperparameters such as unit size and batch size (refer to Table 2). The encoder includes a dense mean layer and a dense logarithmic variance layer to transform data into a latent space representation. Conversely, the decoder process utilizes another set of three LSTM layers, mirroring the encoder's dimensions to reverse this representation and generate new chemical compounds. At final output, a SoftMax layer is applied, and cross-entropy is used as the cost function to measure reconstruction error. Through joint training of the encoder and decoder to minimize the CVAE's cost function, our model enhances its capacity to accurately predict chemical compounds.

Table 2. Hyperparameter Values and Epoch Times for CVAE Model Training and Tuning

	training	tuning
epochs	1000	2000
patience	50	100
batch size	2048	120
unit size	512	512
latent size	200	200
sequence length	120	120
learning rate	1×10^{-4}	1×10^{-5}
number of conditions	6	6
time for epoch (s)	~45	~0.1

In our CVAE model, both the initial training and fine-tuning of the model utilize two main input components: the molecular representation vector x and the condition vector c. The vector x represents the molecular structure in a string-based format, such as SMILES or SELFIES. Each molecule is encoded as a sequence of characters, where each character represents a specific atomic symbol, bond type, or structural feature. The sequence is then tokenized to create a machine-readable input that captures the molecular structure's essential components. This input is further embedded to form a numerical representation that can be processed by the model's encoder. The vector c encodes the desired molecular properties that we aim to impose during the generation process. In our model, this vector includes six molecular properties: M_W, log P, TPSA, HBA, HBD, and RotB. The model is trained to jointly learn the latent z, which captures the underlying distribution of molecular structures, and the condition vector c, which enforces the desired properties. Together, x and c are input into the encoder, where they are processed to generate a latent representation z. Notably, no activity labels are included in these inputs. This design allows the model to autonomously learn additional chemical features while guiding the learning process through the provided conditions, without the need for explicit labels.

To generate a molecule with specified properties from the condition vector, the decoder constructs the molecular structure iteratively, token by token. Each character is sampled based on its probability distribution, which is influenced by the preceding character, until the molecule reaches the fixed length of 120 characters. This process allows a single pair of latent and condition vectors to generate multiple unique molecules. If the termination character "E" is absent within this sequence, the output is deemed invalid. The decoder outputs a probability distribution for the subsequent characters in the sequence, including "E", which is then translated into the molecular structure sequence. It is crucial to highlight that any incorrect character can result in an invalid molecule.

The difference between the training and fine-tuning phases lies in their specific objectives and data handling. During the initial training phase, the model operates in an unsupervised manner. Here, the model learns to capture the underlying distribution of molecular representations and their associated properties without being provided with any explicit activity labels.

In the subsequent fine-tuning phase, the model undergoes further adaptation to better fit a specific fine-tuning data set. While this phase also does not involve direct supervision with activity labels, it allows the model to adjust its learned parameters to align more closely with the characteristics of the fine-tuning data. Although the same six molecular properties are used as conditions, their specific values may differ according to the characteristics of the fine-tuning data set (CDK2, DPP-IV, PPARγ). This variation enables the model to better adapt to the new data while maintaining consistency in the conditions imposed during both training stages.

All training procedures were executed on NVIDIA RTX A6000, NVIDIA GeForce RTX 3090 and NVIDIA GeForce

RTX 4090 GPUs. Table 2 details the configured hyperparameters and epoch durations for both training and tuning phases.

Docking Studies. *Protein and Ligand Preparation.* The crystal structures of CDK2 in complex with Dinaciclib (PDB 4KD1),[32] PPARγ in complex with GL479 (PDB 4CI5),[33] and DPP-IV in complex with a heterocyclic inhibitor (PDB 4A5S)[34] were sourced from the Protein Data Bank (PDB). These structures were selected for docking studies based on their high resolution, completeness, and overall quality.

To prepare these structures for further analysis, we employed the Protein Preparation Wizard within Maestro (Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2021; Impact, Schrödinger, LLC, New York, NY; Prime, Schrödinger, LLC, New York, NY, 2021). This involved adding hydrogen atoms, determining bond orders, charges, and atom types, and extensively sampling rotamers, tautomers, and protonation states of titratable amino acids under neutral pH conditions to optimize the hydrogen bonding network. Subsequently, the protein structures underwent constrained minimization using the Impref module and the OPLS4 force field, maintaining a 0.3 RMSD limit from the original coordinates. The generated compounds were filtered on favorable quantitative estimate of drug-likeness (QED) and synthetic accessibility (SA), resulting in a refined subset (Table S1). These compounds were then prepared using LigPrep (LigPrep, Schrödinger, LLC, New York, NY, 2021) to generate suitable three-dimensional (3D) conformations and tautomerization states at pH 7, followed by energy minimization using the OPLS4 force field.

Docking Simulations. The generated compounds were docked using the Glide algorithm (Glide, Schrödinger, LLC, New York, NY, 2021) in Standard Precision (SP) mode.[35,36] Docking grids were generated with an inner box surrounding the ligand binding cavity site of each target protein. Default docking parameters were applied where not specified. The Glide Score function was used to rank and score the predicted binding poses. The top scoring compounds for each target were carefully evaluated to assess their similarity to the binding mode of cocrystallized ligands and the consistency of protein–ligand interactions with experimental data. Before docking the generated compounds, the entire procedure was validated by redocking the cognate ligands for each target. This step confirmed the ability of the docking protocol to reproduce experimental binding modes. The redocking results displayed favorable RMSD values of 0.498, 1.141, and 1.366 Å for DPP-IV, CDK2, and PPARγ, respectively.

Enrichment Factor Calculation. Additionally, we assessed the ability to prioritize active compounds in the docking screens described above by calculating the Enrichment Factor (EF) at various top score percentages: 1, 2, 5, 10, and 20%. For this purpose, decoy molecules (50 decoys for each active molecule) were generated using the DUD-E server[37] for each set of active compounds (CDK2 and DPP-IV inhibitors and PPARγ agonists). This analysis offers insights into the effectiveness of enriching active molecules over inactive ones within targeted screening regions (see Table S2). Furthermore, the performance of the binary classification model was evaluated using the Receiver Operating Characteristic (ROC) curve (Figure S2), which plots the true positive rate against the false positive rate. The ROC values obtained for the three sets were 0.76 for DPP-IV, 0.79 for PPARγ, and 0.83 for CDK2.

RESULTS

Data Sets and Molecular Representation. Here, we detail the implementation of a CVAE-based generative model and its performance in generating compounds. Rather than focusing solely on generating molecular structure, our aim was to tackle a more complex challenge: generating compounds potentially active against specific targets. Generative models are recognized for their ability to provide viable starting points for drug discovery programs.

During the model training phase, we first collected 327,660 molecules from ChEMBL (version 22). These were subsequently filtered based on various criteria (see Experimental Section), resulting in a data set of 198,962 molecules. We explored two different molecular representations, namely SMILES and SELFIES, to assess potential advantages in terms of performance metrics during the molecular generation phase (see Figure 1, phase 2) and overall molecule quality. During the model fine-tuning phase, we curated three data sets comprising 1352 CDK2 inhibitors, 3911 DPP-IV inhibitors, and 2588 PPARγ agonists, selecting only those compounds with activity thresholds of at least 10 μM. To align with the CVAE architecture, all compound data sets were converted into sequences of symbols using a predefined vocabulary.

Conditional Molecular Design. Before delving into the experimental findings, we outline the metrics employed to evaluate our model's effectiveness. Specifically, we employed the metrics implemented in MOSES to evaluate the generated molecules.[38] Molecule validity assessment involves adherence to organic chemistry principles, ensuring accurate representation as legitimate chemical structures. Invalid molecules may exhibit syntax errors or implausible chemical arrangements. Conversely, uniqueness quantifies the variety of generated samples by calculating the ratio of distinct samples within the generated set. Novelty measures how generated samples differ from those in the training data set. Internal diversity appraises the assortment of the generated molecules based on their chemical properties or structural characteristics. Table 3 presents validation metric outcomes for each target, revealing distinct patterns across DPP-IV, CDK2, PPARγ, and representation formats (SMILES and SELFIES). SELFIES-generated molecules exhibit higher validity values compared to SMILES-generated ones across all targets, implying stronger adherence to syntax and chemical rules. Both formats consistently exhibit high uniqueness (>99%), suggesting a diverse and nonrepetitive collection of molecules. Novelty values are consistently at 100%, indicating uniqueness from the training data. Internal diversity varies among targets and formats, but generally remain high, reflecting a range of molecular structures in each set. Comparison of SELFIES with SMILES highlights significant improvements in validity, reconstruction accuracy, and molecule diversity. During

Table 3. Generated Compounds Metrics for Each Target: CDK2, DPP-IV, PPARγ

files	validity	uniqueness	novelty	internal diversity
SMILES_CDK2	63.91	99.97	100.0	86.89
SELFIES_CDK2	71.11	99.82	100.0	88.75
SMILES_DPP-IV	73.12	99.92	100.0	85.32
SELFIES_DPP-IV	79.51	99.89	100.0	87.84
SMILES_PPARγ	73.88	99.97	100.0	84.55
SELFIES_PPARγ	84.62	99.85	100.0	87.66

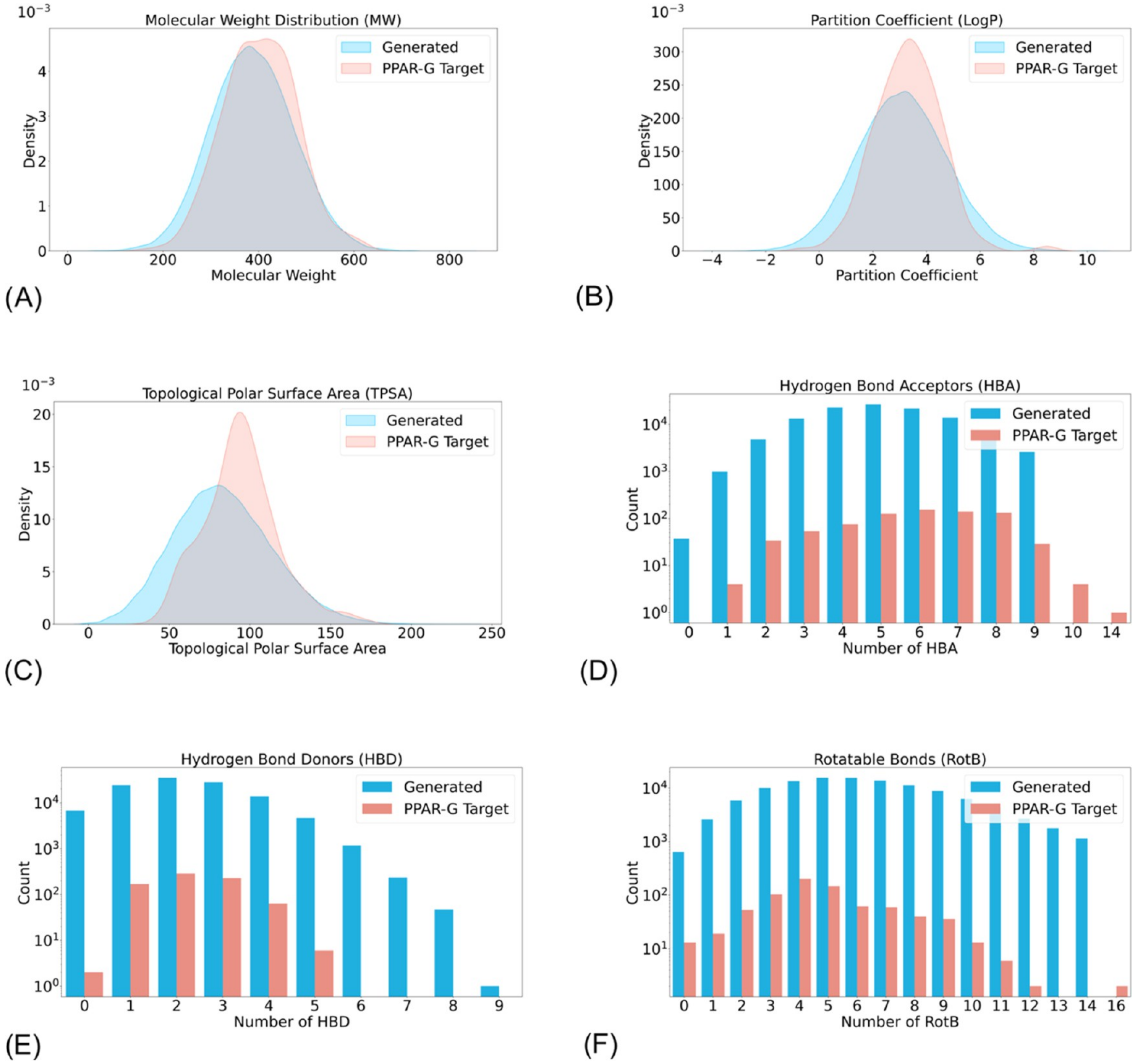


Figure 4. Distribution of physicochemical properties between generated molecules using SELFIES grammar and fine-tuning set ligands (CDK2). Panels (A–C) depict kernel density estimate plots for M_W, log P, and TPSA, respectively. Panels (D–F) show histograms on a logarithmic scale for HBA, HBD, and RotB, respectively.

generation, we applied a "hypervalent" constraint to SELFIES, allowing slight modifications in molecules, where one or more main group elements can carry more than eight electrons in their valence shell. This approach aimed to align molecules closely with the optimization data set. However, as shown in Table 3 SELFIES generation does not achieve 100% validity due to potential deviations from standard chemical rules under this constraint. Alternatively, imposing default or octet rule constraints could alter molecules significantly, potentially leading to a greater number of complex structures that are more challenging to synthesize. In summary, our approach balances generating molecules closer to the fine-tuning data set structurally, while ensuring validity within acceptable limits. This study utilized both molecular representations—SMILES and SELFIES—separately, for comprehensive analysis and validation.

Figure S1 illustrates representative samples of molecules generated by the CVAE model for the three targets in both SMILES and SELFIES format. It also includes QED[39] and SA[40] values. QED measures similarity to known drugs (values between 0 and 1), while SA estimates ease of synthesis (values between 1 and 10). These parameters were used to filter compounds with desirable drug-like properties (QED values between 0.5 and 1) and feasible synthesis potential (SA values between 1 and 5), optimizing them for subsequent docking procedures.

Figure 4 depicts the distributions of molecular properties (M_W, log P, TPSA, HBA, HBD, and RotB) for SELFIES-generated molecules compared to CDK2 inhibitors from the

https://doi.org/10.1021/acsomega.4c08027

fine-tuning data set. This analysis highlights the model's ability to generate molecules with properties closely resembling those in the fine-tuning set, underscoring the efficacy of the CVAE model in capturing and reproducing essential molecular patterns. For a comprehensive analysis across all targets with both grammar formats, refer to Figures S3−S7 in the Supporting Information.

CVAE Model Comparisons with Other Generative Models. We conducted a comparative analysis between our model and the contemporary generative AI framework, REINVENT4.[21] Developed as an open-source platform, REINVENT4 is tailored for creating small molecules in drug discovery. It uses a combination of RNNs and transformer models, structured into four subalgorithms. For a fair comparison, we trained the *De Novo* Design algorithm using the ChEMBL22 data set to teach the model molecular grammar. The tuning phase was performed using the Molecular Optimization subalgorithm. Since the primary competitor to generative VAE model is the generative GAN model, an additional comparative analysis was conducted using the MOLGEN[22] model. Given that both models exclusively employ the SMILES grammar, the comparison was based on this string representation. For each run, a set of 30,000 molecules was generated.

Table 4 shows that REINVENT4 and MOLGEN achieve nearly 100% validity, but there is a notable decrease in the uniqueness (for both REINVENT4 and MOLGEN) and novelty (for REINVENT4) of the generated compounds. This indicates that these models struggled to produce diverse molecules, resulting in a predominantly similar output. In contrast, our CVAE model, while having slightly lower validity, excels in generating more unique and novel compounds. Analyzing Table 5, we observe that the standard deviations of the generated compounds' chemical properties and lengths for each target with REINVENT4 are consistently higher than those generated with CVAE. This variance is due to the inherent nature and methodologies of the two models.

Table 4. Comparative Metrics of Generated Compounds for Each Target Using REINVENT4, MOLGEN, and Our CVAE Model

data set	validity (%)	uniqueness (%)	novelty (%)	internal diversity (%)
CDK2 REINVENT4	99.42	14.35	87.36	86.60
DDP-IV REINVENT4	99.73	14.22	80.00	84.60
PPARγ REINVENT4	99.79	17.36	71.73	82.47
CDK2MOLGEN	93.56	58.42	100	91.24
DDP-IV MOLGEN	97.84	56.68	100	90.31
PPARγ MOLGEN	95.70	65.33	100	90.06
CDK2 Our	65.64	100	100	85.83
DDP-IV Our	73.05	100	100	85.52
PPARγ Our	73.88	100	100	84.77

The lower variability in CVAE-generated compounds is attributed to the model's dependency on initial conditions, such as M_W, log *P*, TPSA, HBA, HBD, and RotB. These conditions are meticulously chosen to align with the statistical median of the training data set, ensuring that CVAE-generated molecules exhibit chemical properties similar to most molecules in the data set, thus minimizing deviation. Conversely, REINVENT4 operates with fewer constraints, allowing for broader exploration of chemical space. Examining the results shown in Tables 1 and 5, it can be observed that the molecules generated by the MOLGEN model are quite diverse with respect to the initial data sets in terms of physicochemical properties. This highlights the model's difficulty in capturing the characteristics of the data set, with a stronger emphasis on generating valid molecules.

As outlined in Section "Conditional Molecular Design", we defined selection criteria based on QED values ranging from 0.5 to 1 and synthetic accessibility (SA) values ranging from 1 to 5. Consequently, in graphs depicting these molecular characteristics (Figure 5), only molecules in the upper-left quadrant are considered potential candidates. The results (in terms of number of generated molecules) within these thresholds are as follows: for REINVENT4, CDK2 = 3059; DDP-IV = 3446; PPARγ = 2060, for MOLGEN CDK2 = 3076; DDP-IV = 6125; PPARγ = 4643; while for our CVAE model, they are CDK2 = 14,741; DDP-IV = 20,230; PPARγ = 14,705. These results indicate that the CVAE model generates a significantly higher number of new molecules conforming to the selected QED and SA values, suggesting a higher likelihood of obtaining desirable molecules using the proposed model compared to REINVENT4. It is worth noting that MOLGEN generated molecules populate much less the upper-left quadrant of the plot, with a higher percentage of molecules falling under the 0.6 value of QED and distributed around values of SA ranging from 2 to 6. On the other hand, there is also a large fraction of molecules generated with CVAE falling

Table 5. Comparative Metrics of Generated Compounds for Each Target Using REINVENT4, MOLGEN, and Our CVAE Model

data set	length of molecules	M_W	log *P*	TPSA	HBA	HBD	RotB	length of rings
CDK2 REINVENT4	44.96 ± 10.55	366.70 ± 84.13	3.25 ± 1.39	82.67 ± 30.34	4.88 ± 1.90	2.05 ± 1.14	4.62 ± 2.41	3.48 ± 0.97
DDP-IV REINVENT4	46.26 ± 11.52	380.94 ± 88.31	2.54 ± 1.58	79.69 ± 30.41	4.54 ± 1.92	1.48 ± 0.97	5.06 ± 2.49	3.07 ± 1.04
PPARγ REINVENT4	51.16 ± 11.78	434.16 ± 94.57	5.02 ± 1.54	77.05 ± 26.53	4.65 ± 1.77	1.31 ± 0.83	8.40 ± 3.36	3.31 ± 1.08
CDK2MOLGEN	14.64 ± 1.77	121.97 ± 10.53	0.14 ± 0.83	46.33 ± 19.13	2.67 ± 0.96	1.07 ± 0.86	1.32 ± 1.30	3.57 ± 2.03
DDP-IV MOLGEN	14.82 ± 1.69	123.52 ± 7.99	0.63 ± 0.97	32.00 ± 18.87	1.92 ± 0.90	0.72 ± 0.73	1.11 ± 1.21	4.02 ± 2.00
PPARγ MOLGEN	15.47 ± 1.73	123.56 ± 8.10	0.31 ± 0.93	35.10 ± 17.69	2.16 ± 0.89	0.65 ± 0.70	0.80 ± 0.94	4.20 ± 1.28
CDK2 Our	46.98 ± 3.86	398.53 ± 16.54	4.37 ± 0.89	61.90 ± 11.89	3.90 ± 0.83	1.98 ± 0.77	6.77 ± 1.74	3.52 ± 0.80
DDP-IV Our	41.22 ± 3.19	352.93 ± 15.71	3.25 ± 0.88	60.11 ± 12.91	3.35 ± 0.70	2.02 ± 0.71	6.25 ± 1.75	2.79 ± 0.73
PPARγ Our	44.91 ± 3.54	389.24 ± 14.85	5.29 ± 0.82	48.12 ± 12.53	3.04 ± 0.72	1.14 ± 0.69	7.54 ± 1.81	3.16 ± 0.81

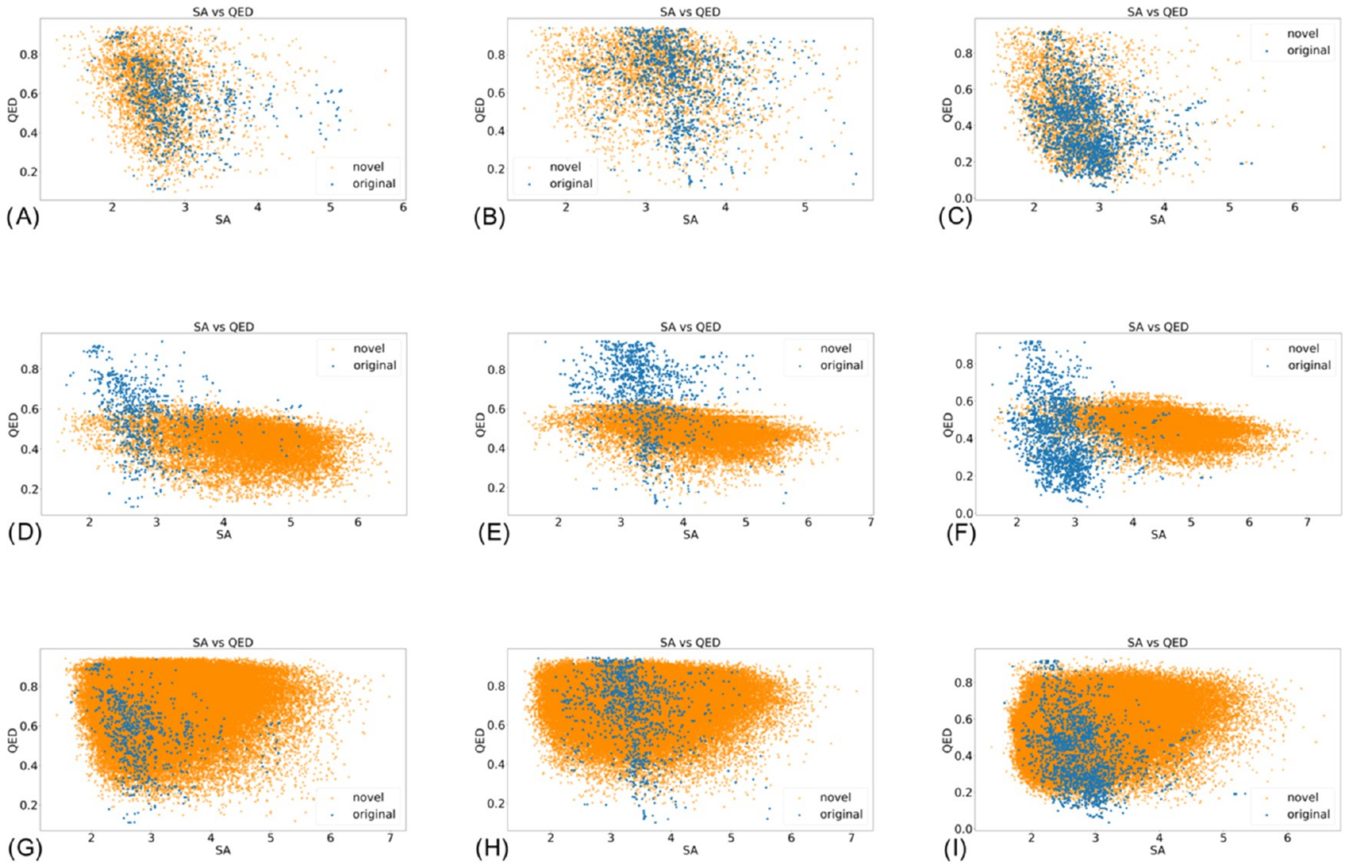


Figure 5. Scatter plots visualizing QED versus SA. Panels (A−C) depict generation results for targets CDK2 (A), DDP-IV (B), and PPARγ (C) using the REINVENT4 model. Panels (D−F) depict generation results for the same targets using MOLGEN, while panels (G−I) using our model. Orange points indicate newly generated molecules (novel), while blue points represent molecules from the original data set (original).

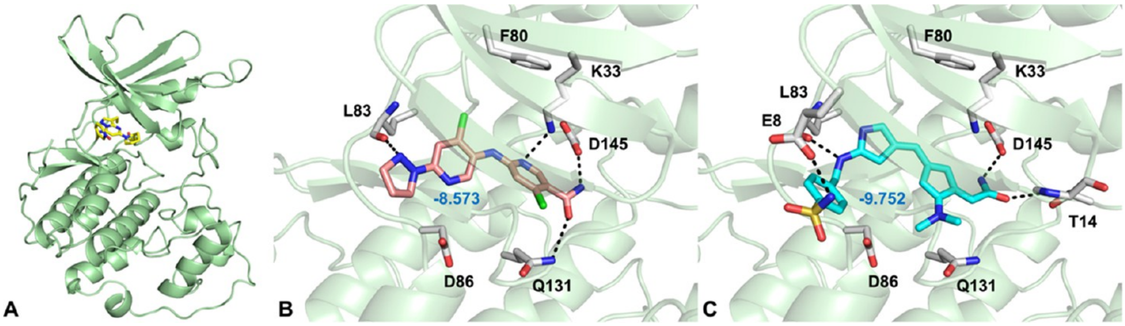


Figure 6. (A) Overview of CDK2 structure depicted as a pale green ribbon model (PDB: 4KD1). The cocrystallized ligand dinaciclib is illustrated as yellow sticks. (B) Predicted binding mode of the generated molecule 1 (represented in SMILES notation, shown as salmon sticks) within CDK2. (C) Predicted binding mode of the generated molecule 2 (represented in SELFIES notation, shown as cyan sticks) within CDK2. Only amino acid residues discussed in the main text are visualized as white sticks and labeled. Hydrogen bonds referenced in the text are depicted as dashed black lines. Docking scores for each molecule are reported in blue.

within a suboptimal area, with a higher SA score. This might be a consequence of the model's ability to navigate different regions of the chemical space with respect to REINVENT4, and to achieve greater novelty. Although such molecules populate an area of the space which is not necessarily the most desirable (including molecules with lower synthetic accessibility), they might still represent a source of "novel chemical matter", which can be a useful starting point for drug discovery.

Case Studies. In this section, we report and analyze the results of docking screens against the three selected targets (CDK2, DPP-IV, and PPARγ) for molecules generated using either SMILES or SELFIES representations. Molecular docking enables us to explore the molecular interactions within the binding pocket of each target. Numerous studies have utilized molecular docking as an additional filter to assess newly generated compounds and guide the generative process, thereby serving as a benchmark in conjunction with the previously described metrics.[41] For each target, we present the binding mode of two representative compounds selected from the top scoring ones and displaying relevant interactions with residues that are known to be crucial for activity.

CDK2. CDK2, a member of the CDK family, is a ubiquitously expressed serine/threonine kinase that regulates cell cycle progression and transcription. Dysregulation of

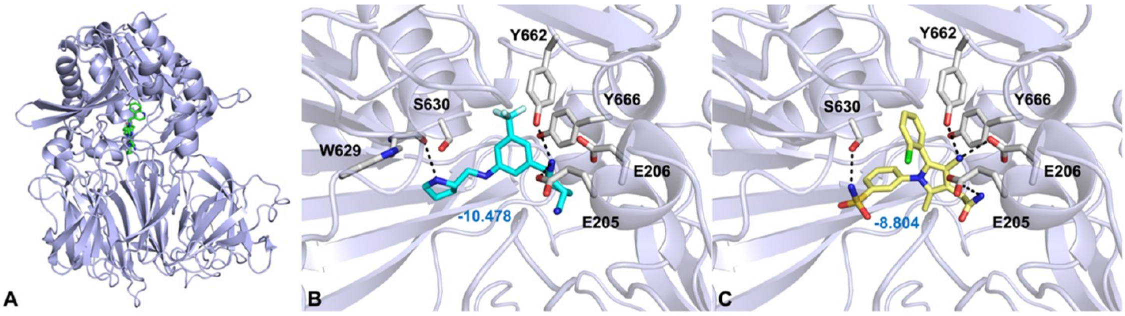


Figure 7. (A) Overview of the DPP-IV structure in its monomeric form depicted as a light-blue ribbon model (PDB: 4A5S). The cocrystallized ligand 4i is illustrated as green sticks. (B) Predicted binding mode of the generated molecule 1 (represented in SMILES notation, shown as cyan sticks) within DPP-IV. (C) Predicted binding mode of the generated molecule 2 (represented in SELFIES notation, shown as yellow sticks) within DPP-IV. Only amino acid residues discussed in the main text are visualized as white sticks and labeled. Hydrogen bonds referenced in the text are depicted as dashed black lines. Docking scores for each molecule are reported in blue.

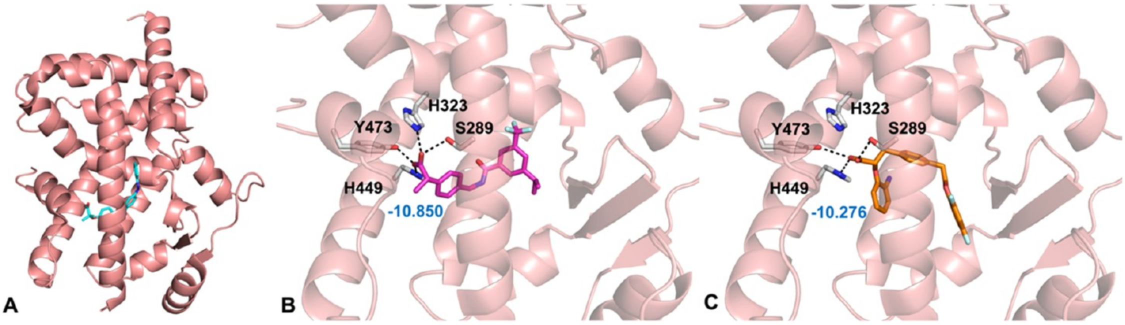


Figure 8. (A) Overview of the PPARγ LBD structure depicted as a salmon ribbon model (PDB: 4CI5). The cocrystallized ligand GL479 is illustrated as cyan sticks. (B) Predicted binding mode of the generated molecule 1 (represented in SMILES notation, shown as magenta sticks) within PPARγ. (C) Predicted binding mode of the generated molecule 2 (represented in SELFIES notation, shown as orange sticks) within PPARγ. Only amino acid residues discussed in the main text are visualized as white sticks and labeled. Hydrogen bonds referenced in the text are depicted as dashed black lines. Docking scores for each molecule are reported in blue.

CDKs has been linked to various medical conditions, underscoring their significance in cellular functions and disease development. The monomeric structure of CDK2 comprises an N-terminal lobe rich in β-sheets (N lobe), a larger C-terminal lobe rich in α-helices (C lobe), and a deep cleft at the interface of the two lobes where ATP binding and catalysis occur (Figure 6A).[42,43] The ATP molecule interacts with three key residues—Lys33, Glu51, and Asp145—that form a conserved catalytic triad found in all eukaryotic kinases, playing a crucial role in positioning the ATP phosphate group for catalysis. The binding pocket consists of five main regions: adenine pocket, ribose pocket, hydrophobic region, phosphate region, and solvent region.[44] Figure 6B illustrates the binding mode of a generated molecule (1) using the SMILES notation within the CDK2 binding pocket.

The molecule achieved a docking score of −8.573 kcal/mol and binds within the adenine pocket, particularly in the hinge region, forming a hydrogen bond between the nitrogen of the pyrazolidine moiety and Leu83. Asp145, located in the phosphate pocket, and Gln131 in the ribose pocket, form hydrogen bonds with the amide moiety, while Lys33 interacts with the pyridine ring. Figure 6C shows the binding mode of the generated molecule 2 using SELFIES notation. This molecule achieved a docking score of −9.752 kcal/mol and is more complex than the one previously analyzed, further highlighting the inherent differences in the generative process. It positions itself within the adenine pocket, forming numerous hydrogen bonds: one between the nitrogen atom of the compound and the CO backbone of Leu83, and two involving the amide group with the backbone of Thr14 and the side chain of Asp145. The latter is part of the DFG motif (composed of Asp145, Phe146, and Glu147), which constitutes a gateway determining the functionally important DFG-in and DFG-out conformations that influence inhibitor binding.[45]

DPP-IV. DPP-IV plays a crucial role in regulating the biological functions of various peptide hormones, chemokines, and neuropeptides, notably in maintaining glucose homeostasis.[46] The structure of DPP-IV (Figure 7A) comprises two subunits forming a dimer, each with an α/β-hydrolase domain and an eight-bladed β-helix domain, connected by a large cavity accessible via two openings. The α/β-hydrolase domain features a central β-sheet structure surrounded by α-helices, which are arranged in contact with the β-helix.[47] The eight-bladed β-helix domain contains eight blades, each composed of four antiparallel β-structures. The catalytic triad (Ser630, His740, and Asp708) is positioned at the interface of the propeller and hydrolase domains, playing a crucial role in enzymatic activity.[48] DPP-IV has five sites: S1, S2, S1′, S2′, and S2 extensive. The S1 and S2 pockets are essential for inhibitory activity, while modulation of the S1′, S2′, and S2 extensive sites can enhance inhibitory potency.[49] Figure 7B shows a molecule generated using SMILES notation (docking score of −10.478 kcal/mol), fitting well into the binding pocket and extending to the S2′ pocket, where the pyrrolidine nitrogen interacts with the CO backbone of Trp629. Critical residues Glu205 and Glu206, part of the S2 subsite, frequently interact with the ligand's amine. The compound generated using

SELFIES notation (Figure 7C) is also well positioned within the binding pocket between the S1 and S2 subunits, achieving a docking score of −8.804 kcal/mol. Ser630 forms a hydrogen bond with the sulfonamide oxygen, while Tyr662 interacts with the amine. Hydrogen bonds with Glu205 and Glu206 are conserved, and $\pi-\pi$ interactions, such as between chlorophenol and Tyr666, are evident.

PPARγ. PPARs, known as lipid and glucose sensors, regulate insulin sensitivity and energy metabolism. These nuclear receptors have three distinct subtypes: PPARα, PPARβ/δ, and PPARγ. Each subtype exhibits unique expression patterns and functions depending on the specific organ and cell type.[50] PPARγ is abundant in adipocytes and macrophages, highlighting its crucial role in various metabolic processes. Beyond adipocyte differentiation and lipid storage, PPARγ modulates insulin sensitivity and maintains glucose homeostasis, underscoring its importance in overall metabolic health.[51] PPARγ has two main functional domains: the central DNA-binding domain, responsible for recognizing specific DNA sequences, and the ligand-binding domain (LBD), essential for receptor activation. The LBD (Figure 8A) consists of 12 α-helices (H1−H12) highly conserved between human and mouse orthologs.[52] Thiazolidinediones, selective PPARγ agonists, function as full agonists by forming hydrogen bonds with PPARγ residues His323, His449, and Tyr473.[53] These interactions stabilize the AF2 surface and the H12 region, enabling the LBD to bind coactivators and promote full gene expression.

Figure 8B shows the predicted binding mode of a molecule generated using SMILES notation, showing a docking score of −10.850 kcal/mol. The carboxylic acid headgroup is oriented toward helix 12, while the hydrophobic tail faces helix 5. Strong hydrogen bonds form between the ligand and Ser289, His449, His323, and Tyr473, highlighting key interactions for receptor activation. The molecule generated using SELFIES notation (docking score = −10.276 kcal/mol) also has a carboxylic acid headgroup oriented toward helix 12. Hydrogen bonds with His449, Ser289, and Tyr473 are preserved, emphasizing these residues' role in stabilizing the ligand within the binding pocket. The ligand's hydrophobic tail is positioned between helix 3 and the β-sheet.

For each of the above-described therapeutic targets, we compared the docking scores obtained for the generated molecules by using our method (both SMILES and SELFIES notation) and REINVENT4, with the docking scores of known actives from ChEMBL. In particular we selected the top ranked 500 molecules resulting from each docking screen; the results are reported in Figures S8−S10 (Supporting Information). In docking screens, molecules with lower docking scores should more likely be active and therefore are put at the top of the hitlist.

As a general trend, the docking score values for known active compounds span a significantly broader range. Interestingly, the docking score median values for the molecules generated with our method using SMILES notation are consistently lower than the other methods for the three targets, possibly denoting a higher possibility of desired biological activity. Lastly, the generated molecules using (i) our method with SELFIES notation and (ii) REINVENT4 showed similar performances for CDK2 and PPARγ. Overall, the three analyzed sets showed remarkably lower docking scores than the known actives, suggesting that they might include good binders.

■ DISCUSSION

The drug discovery process involves optimizing multiple properties to design compounds with the desired characteristics; however, altering one property through structural modifications can inadvertently impact others.[25] To address this challenge, we employed a multiobjective generative model based on the CVAE architecture. CVAE excels at capturing complex data distributions while seamlessly integrating conditional information.[23−25] It extends the VAE framework by incorporating explicit condition vectors into the latent space representation, thereby allowing the generation of molecules with specific properties. Additionally, the model was enhanced with the LSTM cells for the encoder and decoder (Figure 3). Overall, the model is easy to implement and can be trained in a reasonable amount of time and computational power, with the training procedure needing to be performed only once per data set, and thus ensuring optimized operational workflows.

We placed particular emphasis on vocabulary optimization and molecular representation using SMILES and SELFIES. SELFIES exhibited higher validity, while both achieved high uniqueness (>99%) and 100% novelty (Table 3), indicating broad and nonrepetitive molecule generation. A comparison with the REINVENT4 framework, using the same training set to generate 30,000 molecules, revealed that while REINVENT4 achieved nearly 100% validity, it exhibited lower uniqueness and novelty compared to our model (Table 4). Our CVAE model demonstrated higher structural diversity despite slightly lower validity, with the generated compounds closely matching the median properties of the training data set. In contrast, REINVENT4 showed higher variability in chemical properties. These differences can be attributed to the intrinsic nature and methodologies of the models. The comparison with the MOLGEN framework used the same initial training set, and 30,000 molecules were generated. While the model also achieved nearly 100% validity, it produced a significant number of duplicates, resulting in approximately 55% uniqueness. Additionally, the generated molecules exhibited structural forms that were substantially different from those in the training data set. The diversity of generated compounds may be influenced by the model's dependence on initial conditions (M_W, log P, TPSA, HBA, HBD, and RotB), which were chosen to closely align with the statistical median of the training data set. This ensures that CVAE-generated molecules exhibit chemical properties similar to most molecules in the data set, minimizing deviation from the norm. Conversely, REINVENT4 operates with less stringent constraints, allowing for a broader exploration of chemical space.

Finally, molecular docking analysis of the generated molecules demonstrated that their predicted binding modes closely resemble those of cocrystallized ligands. This similarity suggests that our generated molecules can be accommodated within the active sites of the selected target proteins, underscoring their potential as biologically relevant compounds. This methodology can be extended to other therapeutically relevant targets.

■ CONCLUSIONS

In this work, we developed a CVAE model for the efficient and accurate generation of drug-like molecules. We employed both SMILES and SELFIES representations to evaluate potential advantages in terms of performance metrics and overall molecule quality. The CVAE model successfully designed novel compounds, and evaluation metrics such as validity, uniqueness, novelty, and internal diversity were used to assess the effectiveness of both SELFIES and SMILES representations. Our results indicate that the CVAE model is robust and versatile, capable of generating a diverse array of chemically valid and novel molecules with high uniqueness.

Future advancements in model architecture could explore integrating more sophisticated neural network designs or ensemble approaches to enhance coverage of chemical space and molecular interactions. Additionally, leveraging transfer learning or active learning techniques could further improve the model's performance, particularly in scenarios with limited data, such as rare disease research or early stage drug discovery. Currently, the CVAE is considered a "data-hungry" model, requiring substantial training data to achieve optimal results. This characteristic underscores the importance of robust data collection and preprocessing methodologies to ensure the model's effectiveness. Incorporating additional conditioning properties such as pharmacokinetic parameters, toxicity profiles, or molecular stability indicators could further refine the generative process, resulting in molecules with improved drug-like properties. Furthermore, integrating real-time feedback loops from biological assays could enhance the iterative design process, making the model even more responsive and precise in generating candidate molecules. These enhancements would collectively contribute to a more efficient and targeted drug discovery process, expanding the potential applications of the CVAE model in pharmaceutical research.

■ ASSOCIATED CONTENT

Data Availability Statement

The ChEMBL database (https://www.ebi.ac.uk/chembl/) is a public domain data resource. Schrödinger Suite (https://www.schrodinger.com), a licensed software for biomolecular simulation and analysis, was used for docking studies. PyMOL (https://pymol.org/), a molecular visualization tool distributed under a license was used for displaying and analyzing 3D structures and for figures preparation. The code and data sets for the case studies are available at the following link: https://github.com/MODAL-UNINA/Enhancing-De-Novo-Drug-Design-Across-Multiple-Therapeutic-Targets-with-CVAE-Generative-Models.git

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c08027.

> Representative samples of molecules generated using the CVAE model for the three targets (Figure S1); generated compounds and filtered data set for docking screens (Table S1); enrichment factor calculations for the three targets (Table S2); ROC curves for DPP-IV, PPARγ, CDK2 (Figure S2); properties distributions of calculated physicochemical properties between generated molecules with the SMILES grammar and fine-tuning set ligands (DDP-IV) (Figure S3); properties distributions of calculated physicochemical properties between generated molecules with the SMILES grammar and fine-tuning set ligands (PPARγ) (Figure S4); properties distributions of calculated physicochemical properties between generated molecules with the SMILES grammar and fine-tuning set ligands (CDK2) (Figure S5); properties distributions of calculated physicochemical properties between generated molecules with the SELFIES grammar and fine-tuning set ligands (DDP-IV) (Figure S6); properties distributions of calculated physicochemical properties between generated molecules with the SELFIES grammar and fine-tuning set ligands (PPARγ) (Figure S7); box plot comparing the docking score values for the top ranked 500 molecules generated by our method (using SMILES or SELFIES) and REINVENT4, with respect to CDK2 known actives (Figure S8); box plot comparing the docking score values for the top ranked 500 molecules generated by our method (using SMILES or SELFIES) and REINVENT4, with respect to DPP-IV known actives (Figure S9), and box plot comparing the docking score values for the top ranked 500 molecules generated by our method (using SMILES or SELFIES) and REINVENT4, with respect to PPARγ known actives (Figure S10) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Antonio Lavecchia − *Department of Pharmacy, "Drug Discovery Laboratory", University of Naples Federico II, Naples 80131, Italy;* orcid.org/0000-0002-2181-8026; Email: antonio.lavecchia@unina.it

Authors

Virgilio Romanelli − *Department of Pharmacy, "Drug Discovery Laboratory", University of Naples Federico II, Naples 80131, Italy*

Daniela Annunziata − *Department of Mathematics and Applications "R. Caccioppoli", University of Naples Federico II, Naples 80126, Italy*

Carmen Cerchia − *Department of Pharmacy, "Drug Discovery Laboratory", University of Naples Federico II, Naples 80131, Italy;* orcid.org/0000-0002-6631-5000

Donato Cerciello − *Department of Mathematics and Applications "R. Caccioppoli", University of Naples Federico II, Naples 80126, Italy*

Francesco Piccialli − *Department of Mathematics and Applications "R. Caccioppoli", University of Naples Federico II, Naples 80126, Italy*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c08027

Author Contributions

[§]V.R. and D.A. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.L. acknowledges funding from the Italian Ministry of Education, University and Research (MIUR), Progetti di Rilevante Interesse Nazionale (PRIN), grant no. 2022P5LPHS. V.R. acknowledges the PhD scholarship funded by DM351/2022 PNRR, CUP: E66E22000360006. C.C. acknowledges PON R&I 2014−2020 Asse IV "Istruzione e Ricerca per il recupero—REACT-EU" Azione IV.4 "Contratti di Ricerca su tematiche dell'Innovazione" and support from the University of Naples "Federico II" (Grant FRA 2022, Line C, INITIATIVE).

■ REFERENCES

(1) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.

(2) Schneider, G.; Clark, D. E. Automated de Novo Drug Design: Are We Nearly There Yet? *Angew. Chem., Int. Ed.* **2019**, *58*, 10792−10803.

(3) Lavecchia, A. Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discovery Today* **2015**, *20*, 318−331.

(4) Lavecchia, A. Deep Learning in Drug Discovery: Opportunities, Challenges and Future Prospects. *Drug Discovery Today* **2019**, *24*, 2017−2032.

(5) Cerchia, C.; Lavecchia, A. New Avenues in Artificial-Intelligence-Assisted Drug Discovery. *Drug Discovery Today* **2023**, *28*, No. 103516.

(6) Radford, A.; Metz, L.; Chintala, S. In *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*, 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings; ICLR, 2016.

(7) van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio, arXiv:1411.3232. arXiv.org e-Print archive, 2016. https://arxiv.org/abs/1609.03499.

(8) Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Jozefowicz, R.; Bengio, S. In *Generating Sentences from a Continuous Space*, 20th SIGNLL Conference on Computational Natural Language Learning, Proceedings; CoNLL, 2016; pp 10−21.

(9) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675−679.

(10) Tong, X.; Liu, X.; Tan, X.; Li, X.; Jiang, J.; Xiong, Z.; Xu, T.; Jiang, H.; Qiao, N.; Zheng, M. Generative Models for de Novo Drug Design. *J. Med. Chem.* **2021**, *64*, 14011−14027.

(11) Romanelli, V.; Cerchia, C.; Lavecchia, A. Deep Generative Models in the Quest for Anticancer Drugs: Ways Forward. *Front. Drug Discovery* **2024**, *4*, No. 1362956.

(12) Romanelli, V.; Cerchia, C.; Lavecchia, A. Unlocking the Potential of Generative Artificial Intelligence in Drug Discovery. In *Applications of Generative AI*; Lyu, Z., Ed.; Springer International Publishing: Cham, 2024; pp 37−63.

(13) Gangwal, A.; Lavecchia, A. Unleashing the Power of Generative AI in Drug Discovery. *Drug Discovery Today* **2024**, *29*, No. 103992.

(14) Lavecchia, A. Advancing Drug Discovery with Deep Attention Neural Networks. *Drug Discovery Today* **2024**, *29*, No. 104067.

(15) Mikolov, T.; Karafiát, M.; Burget, L.; Jan, C.; Khudanpur, S. In *Recurrent Neural Network Based Language Model*, Proceedings of the 11th Annual Conference of the International Speech Communication Association; Interspeech, 2010; pp 1045−1048.

(16) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inf.* **2018**, *37*, No. 1700111.

(17) Jiang, D.; Wu, Z.; Hsieh, C. Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *J. Cheminf.* **2021**, *13*, No. 12.

(18) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminf.* **2017**, *9*, No. 48.

(19) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(20) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for de Novo Drug Design. *J. Chem. Inf. Model.* **2020**, *60*, 5918−5922.

(21) Loeffler, H. H.; He, J.; Tibo, A.; Janet, J. P.; Voronov, A.; Mervin, L. H.; Engkvist, O. Reinvent 4: Modern AI−Driven Generative Molecule Design. *J. Cheminf.* **2024**, *16*, No. 20.

(22) de Masson d'Autume, C.; Mohamed, S.; Rosca, M.; Rae, J. In *Training Language Gans from Scratch*, Advances in Neural Information Processing Systems; NeurIPS, 2019.

(23) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular Generative Model Based on Conditional Variational Autoencoder for de Novo Molecular Design. *J. Cheminf.* **2018**, *10*, No. 31.

(24) Joo, S.; Kim, M. S.; Yang, J.; Park, J. Generative Model for Proposing Drug Candidates Satisfying Anticancer Properties Using a Conditional Variational Autoencoder. *ACS Omega* **2020**, *5*, 18642−18650.

(25) Yang, Y.; Hsieh, C. Y.; Kang, Y.; Hou, T.; Liu, H.; Yao, X. Deep Generation Model Guided by the Docking Score for Active Molecular Design. *J. Chem. Inf. Model.* **2023**, *63*, 2983−2991.

(26) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(27) Krenn, M.; Häse, F.; Nigam, A. K.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, No. 045024.

(28) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic Acids Res.* **2015**, *43*, W612−W620.

(29) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.

(30) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.

(31) RDKit, 2021. https://www.rdkit.org/. (accessed May 28, 2021).

(32) Martin, M. P.; Olesen, S. H.; Georg, G. I.; Schönbrunn, E. Cyclin-Dependent Kinase Inhibitor Dinaciclib Interacts with the Acetyl-Lysine Recognition Site of Bromodomains. *ACS Chem. Biol.* **2013**, *8*, 2360−2365.

(33) dos Santos, J. C.; Bernardes, A.; Giampietro, L.; Ammazzalorso, A.; De Filippis, B.; Amoroso, R.; Polikarpov, I. Different Binding and Recognition Modes of GL479, a Dual Agonist of Peroxisome Proliferator-Activated Receptor α/γ. *J. Struct. Biol.* **2015**, *191*, 332−340.

(34) Sutton, J. M.; Clark, D. E.; Dunsdon, S. J.; Fenton, G.; Fillmore, A.; Harris, N. V.; Higgs, C.; Hurley, C. A.; Krintel, S. L.; MacKenzie, R. E.; Duttaroy, A.; Gangl, E.; Maniara, W.; Sedrani, R.; Namoto, K.; Ostermann, N.; Gerhartz, B.; Sirockin, F.; Trappe, J.; Hassiepen, U.; Baeschlin, D. K. Novel Heterocyclic DPP-4 Inhibitors for the Treatment of Type 2 Diabetes. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 1464−1468.

(35) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(36) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.

(37) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.

(38) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*, No. 565644, DOI: 10.3389/fphar.2020.565644.

(39) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90−98.

(40) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1*, No. 8.

(41) Ciepliński, T.; Danel, T.; Podlewska, S.; Jastrzębski, S. Generative Models Should at Least Be Able to Design Molecules That Dock Well: A New Benchmark. *J. Chem. Inf. Model.* **2023**, *63*, 3238−3247.

(42) G ray, N.; Detivaud, L.; Doerig, C.; Meijer, L. ATP-Site Directed Inhibitors of Cyclin-Dependent Kinases. *Curr. Med. Chem.* **1999**, *6*, 859−875.

(43) Pavletich, N. P. Mechanisms of Cyclin-Dependent Kinase Regulation: Structures of Cdks, Their Cyclin Activators, and Cip and INK4 Inhibitors. *J. Mol. Biol.* **1999**, *287*, 821−828.

(44) Cheng, W.; Yang, Z.; Wang, S.; Li, Y.; Wei, H.; Tian, X.; Kan, Q. Recent Development of CDK Inhibitors: An Overview of CDK/Inhibitor Co-Crystal Structures. *Eur. J. Med. Chem.* **2019**, *164*, 615−639.

(45) Łukasik, P.; Baranowska-bosiacka, I.; Kulczycka, K.; Gutowska, I. Inhibitors of Cyclin-dependent Kinases: Types and Their Mechanism of Action. *Int. J. Mol. Sci.* **2021**, *22*, No. 2806.

(46) Mentlein, R. Dipeptidyl-Peptidase IV (CD26)-Role in the Inactivation of Regulatory Peptides. *Regul. Pept.* **1999**, *85*, 9−24.

(47) Rasmussen, H. B.; Branner, S.; Wiberg, F. C.; Wagtmann, N. Crystal Structure of Human Dipeptidyl Peptidase IV/CD26 in Complex with a Substrate Analog. *Nat. Struct. Biol.* **2003**, *10*, 19−25.

(48) Ogata, S.; Misumi, Y.; Tsuji, E.; Takami, N.; Oda, K.; Ikehara, Y. Identification of the Active Site Residues in Dipeptidyl Peptidase IV by Affinity Labeling and Site-Directed Mutagenesis. *Biochemistry* **1992**, *31*, 2582−2587.

(49) Juillerat-Jeanneret, L. Dipeptidyl Peptidase IV and Its Inhibitors: Therapeutics for Type 2 Diabetes and What Else? *J. Med. Chem.* **2014**, *57*, 2197−2212.

(50) Michalik, L.; Auwerx, J.; Berger, J. P.; Chatterjee, V. K.; Glass, C. K.; Gonzalez, F. J.; Grimaldi, P. A.; Kadowaki, T.; Lazar, M. A.; O′Rahilly, S.; et al. International Union of Pharmacology. LXI. Peroxisome Proliferator-Activated Receptors. *Pharmacol. Rev.* **2006**, *58*, 726−741.

(51) Willson, T. M.; Lambert, M. H.; Kliewer, S. A. Peroxisome Proliferator−Activated Receptor *γ* and Metabolic Disease. *Annu. Rev. Biochem.* **2001**, *70*, 341−367.

(52) Holzer, G.; Markov, G. V.; Laudet, V. Evolution of Nuclear Receptors and Ligand Signaling: Toward a Soft Key−Lock Model?. In *Current Topics in Developmental Biology*; Elsevier, 2017; Vol. *125*, pp 1−38.

(53) Mazumder, M.; Ponnan, P.; Das, U.; Gourinath, S.; Khan, H. A.; Yang, J.; Sakharkar, M. K. Investigations on Binding Pattern of Kinase Inhibitors with PPAR *γ*: Molecular Docking, Molecular Dynamic Simulations, and Free Energy Calculation Studies. *PPAR Res.* **2017**, *2017*, No. 6397836, DOI: 10.1155/2017/6397836.