De Novo Molecular Generation Augmentation for Drug Discovery Using Deep Learning Approaches: A Comparative Study of Variational Autoencoders

Muzaffar Ahmad Sofi[1], Dhanpratap Singh[1], Tawseef Ahmed Teli [2]*

Abstract

Background: Drugs are defined as chemicals that induce physiological effects when ingested, and their development involves multiple stages, including discovery, design, and development, which are often complex and resource-intensive. To address these challenges, machine learning (ML) and deep learning (DL) techniques have emerged as powerful tools to optimize the drug development pipeline. Methods: This study utilized two distinct variational autoencoders: a convolutional encoder-decoder model and a convolutional-GRU-based encoder-decoder model. Employing a reparameterization technique, we aimed to improve the efficiency of de novo molecular generation. Both models were trained and evaluated on the ZINC dataset, assessing their capability to generate chemically valid and syntactically accurate molecules.Results: The convolution-GRU model demonstrated a synthesis accuracy of 96.79%, matching the performance of the convolutional encoder-decoder model. Additionally, the chemical validity of the generated compounds was notable, with unique chemical validity scores of 90.71% for the convolutional encoder-decoder model and 90.42% for the convolution-GRU model. Conclusion: The results indicate that deep molecular generative models, especially the convolution-GRU approach, significantly advance de novo molecular design. By achieving high levels of accuracy and chemical validity, these models hold promise for enhancing drug discovery processes and expediting the introduction of new therapeutics to the market.

Keywords: Drug discovery, Deep learning, Variational autoencoders, Molecular generation, Chemical validity.

Significance | This study demonstrated how deep learning enhances drug discovery, streamlining molecular design processes and improving accuracy and chemical validity.

*Correspondence. Tawseef Ahmed Teli , Govt Degree College Anantnag, Khanabal, Anantnag, 192101, India.
E-mail: 1muzaffarsofi.g@gmail.com

Editor Md Shamsuddin Sultan Khan And accepted by the Editorial Board October 14, 2024 (received for review August 26, 2024)

1. Introduction

Drug discovery is a profoundly intricate, costly, and multifactor-dependent process. On average, identifying and developing a new drug molecule can cost approximately $2.6 billion (Kiriiri, Njogu, & Mwangi, 2020). The primary purpose of a drug, often a protein (Middaugh & Pearlman, 1999), is to bind to a target protein in the body whose modification can alter the course of a disease. Despite substantial investments in time and financial resources, the success rate in discovering new drugs remains disappointingly low, and the overall pipeline is largely unproductive.

The persistent threat of disease has rendered drug discovery essential, pushing researchers toward optimizing this challenging process. Recent advancements in data-driven and artificial intelligence (AI) approaches, such as machine learning (ML) and deep learning (DL), have begun to yield promising results in what was once considered a barren field in terms of efficiency and output. These AI techniques are now applied at various stages of the drug discovery pipeline, including target identification, drug-target

Author Affiliation.
[1] Lovely Professional University, Jalandhar - Delhi, Grand Trunk Rd, Phagwara, Punjab 144411, India.
[2] Govt Degree College Anantnag, Khanabal, Anantnag, 192101, India.

Please Cite This:
Muzaffar Ahmad Sofi, Dhanpratap Singh, Tawseef Ahmed Teli (2024). "De Novo Molecular Generation Induction for Drug Discovery Using Deep Learning Approaches: A Comparative Study of Variational Autoencoders", Journal of Angiotherapy, 8(10),1-13,9996.

https://doi.org/10.25163/angiotherapy.8109996

interactions, safety biomarker assessment, lead compound optimization, de novo molecular structure design, protein function prediction, and genome association, as well as in data mining tasks like drug efficacy and adverse effect investigation (Dara et al., 2022; Vamathevan et al., 2019; Askr et al., 2023; Nag et al., 2022).

In the last decade, AI-driven approaches have focused primarily on predicting drug side effects (Gao et al., 2017), drug-target interactions (Gupta et al., 2021; Cheng et al., 2017; Yaseen & Kurnaz, 2021), response predictions (Ali & Aittokallio, 2019; Partin et al., 2021; Liu et al., 2019), and drug repurposing. Additionally, these approaches have been instrumental in assessing properties such as solubility, binding, and molecular dynamics modeling. However, designing new compounds with specific properties has historically received less attention due to the vast chemical space and the inherent challenges in exploring it. Recently, latent variable models, including generative networks like Recurrent Neural Networks (RNN), Variational Autoencoders (VAE), and Generative Adversarial Networks (GANs), have enabled the generation of novel molecular structures (Prykhodko et al., 2019; Méndez-Lucio et al., 2020).

While creating syntactically and chemically valid molecules remains challenging (Bilodeau et al., 2022; Chen et al., 2020), deep generative models that integrate spatial and sequential information—through convolutional and recurrent neural network layers in the encoder-decoder architecture—offer potential solutions. Including structural information in models is essential, as a drug's structure influences its pharmacokinetics, pharmacodynamics, and safety profile (Paul et al., 2021). This relationship between molecular structure and activity, termed Quantitative Structure-Activity Relationship (QSAR), is central to predicting a molecule's functionality. Consequently, employing a convolutional-based variational autoencoder is advisable for incorporating structural data, given its capability in capturing complex molecular features (Ekins, 2016; Gómez-Bombarelli et al., 2016).

The following sections detail the literature review and existing limitations in Section 2, molecular representation techniques and variational encoder-based model architectures in Section 3, evaluation criteria in Section 4, experimental findings in Section 5, and finally, conclusions and future research directions in Section 6.

2. **Literature Review**

Pharmaceutical companies have made significant strides in leveraging machine learning (ML) and deep learning (DL) techniques to enhance drug discovery processes (Teli & Masoodi, 2021; Zhu, 2020). Across various drug development tasks, DL algorithms have delivered cutting-edge performance, particularly in the complex challenge of designing de novo drug molecules (Blaschke et al., 2018). Recently, researchers have focused on creating molecules with desired properties such as solubility and minimal toxicity, aiming to design drug molecules that can alter specific pathways and bind effectively to target proteins. Encoder-decoder-based models, including autoencoders and generative adversarial networks (GANs), have proven valuable in this process, as they facilitate the generation of new molecular samples by adjusting their latent representations (Gómez-Bombarelli et al., 2016). These advancements underscore the ability of DL models to capture multidimensional molecular representations, which are vital for novel drug development.

In recent studies, Gupta et al. (2018) employed a Long-Short Term Memory (LSTM) model to generate new drug-like compounds. Their generative recurrent neural network (RNN) included two LSTM layers with 256 hidden units and dropout regularization, followed by a dense layer with softmax activation. The model was trained on SMILES strings from the ChEMBL22 dataset, achieving 58% validity in generating token-wise compounds after 22 training epochs. By refining this approach to allow molecule fragment expansion, they improved accuracy, resulting in molecules with higher validity. However, limitations included relatively low validity rates in some molecules, highlighting the need for more robust methods.

Blaschke et al. (2018) proposed an alternative approach by mapping molecular structures to a continuous latent space using autoencoders. Their results showed that preserving molecular similarity in the latent space facilitated the generation of new molecules with enhanced properties. Specifically, they demonstrated that a variational autoencoder (VAE) with an additional discriminator aligned the encoder output with a user-defined target distribution, achieving 77.4% valid molecules with a Gaussian distribution and 78.3% validity with a uniform distribution on the ChEMBL version 22.34 dataset.

Further development in this area included Kadurin et al. (2017) and Joo et al. (2020), who introduced an architecture utilizing conditional variational autoencoders (CVAE) for designing novel drug candidates with specific attributes. By conditioning the encoder and decoder to target properties, the CVAE model was able to generate anti-cancer compounds, as demonstrated on the NCI-60 dataset. This approach allowed the generation of molecules with high Tanimoto similarity coefficients, indicating strong structural similarity to known compounds, and paved the way for more tailored molecular searches in public datasets.

Despite these advancements, current models, including autoencoders and VAEs, face challenges in generating chemically and syntactically valid molecules consistently. Researchers have also explored adversarial autoencoders (AAEs) to improve novelty and applicability in drug design. For instance, Kadurin et al. (2017) used an AAE with a seven-layer architecture to develop anti-cancer molecules, demonstrating significant potential when evaluated with NCI-60 cell line data.

In addition to the molecular generation models, blockchain technology has shown promise in ensuring data security and enhancing trust in healthcare applications (Teli & Masoodi, 2021; Ahmed Teli, Tawseef, et al., 2022). However, to address the current limitations in molecular generation, the development of DL architectures with higher structural and syntactical validity is essential. This study focuses on enhancing these aspects by leveraging convolutional neural networks (CNNs) and Gated Recurrent Units (GRU) models, which have shown promise in capturing intricate molecular patterns.

3. Methodology

3.1 Molecular Representation

Molecules are typically represented by their chemical structures, which include atoms and the bonds connecting them. However, for computational processing, an effective molecular representation must possess two key characteristics: uniqueness and invertibility. Uniqueness ensures that each molecular structure corresponds to a single, distinct representation, while invertibility guarantees a one-to-one relationship between a molecular structure and its representation (David et al., 2020).

Over the years, numerous molecular representation methods have been proposed. Among these, the Simplified Molecular Input Line Entry System (SMILES) is one of the most prevalent for high-speed machine processing due to its simplicity and efficiency (Weininger, 1988). To construct a SMILES representation of a chemical structure, each atom in the molecule is assigned a unique number, and a graph traversal algorithm generates a sequence of ASCII characters based on this numbering (David et al., 2020). Additionally, SMILES strings can be easily converted into various formats, such as one-hot encoding, word embeddings, and molecular fingerprints, which are directly compatible with downstream computational models.

In this study, we utilize one-hot encoding of SMILES strings at the character level to canonicalize them for unique molecular representation. Specifically, each character in a SMILES string is converted into a one-hot encoded vector whose length equals the number of unique characters in the entire dataset. Each SMILES string is then transformed into a fixed-length tensor with dimensions

$$[1 \times len(unique\ chatacters\ in\ dataset) \times len(longest\ SMILES\ string)],$$

Shorter strings are padded with the letter ‘E’ to maintain uniform tensor dimensions. This approach is consistent with the methodology employed by Lim et al. (2018).

Variational Autoencoder Architecture

The majority of techniques for de novo drug design leverage an encoder-decoder structure, typically variants of the Variational Autoencoder (VAE) architecture. In this work, we employ two variants of the VAE: one solely based on Convolutional Neural Networks (CNNs) and the other combining CNNs with Gated Recurrent Units (GRUs). Unlike traditional VAE models that predominantly use fully connected dense layers or RNN layers for the encoder and decoder, our models incorporate convolutional and GRU layers to better capture structural and sequential information inherent in molecular data.

3.2 Variational Autoencoder (VAE)

The basic workflow of a VAE mirrors that of a classical autoencoder but with a probabilistic twist. The encoder maps the input data into a low-dimensional latent space representation, often referred to as the latent code. The decoder then reconstructs the original input from this latent code. Unlike classical autoencoders, both the encoder and decoder in a VAE are probabilistic. Specifically, the encoder generates a probability distribution for each latent dimension rather than a single deterministic value. The decoder samples from this latent space distribution, using the encoder's output parameters, to generate a plausible reconstruction of the input data.

Mathematically, the objective function of a VAE is defined as:

$$\mathbb{E}[\log P(X|z)] - \mathrm{D}_{\mathrm{KL}}[Q(z|X)||P(z)] \qquad (1)$$

Equation 1 VAE Objective Fun

where $\mathbb{E}$: expectation value; P, Q: probability distributions X: data; z: latent space; D_{KL}:Kullback-Leiber divergence.

The first term in the objective function aims to minimize the reconstruction loss, ensuring that the decoder accurately reconstructs the input data from the latent representation. The second term minimizes the Kullback-Leibler divergence between the encoder's distribution Q(z|X)Q(z|X)Q(z|X) and a prior distribution P(z)P(z)P(z), typically a multivariate normal distribution (Kingma & Welling, 2019).

3.3 Sampling with the Reparameterization Trick

A critical aspect of the VAE is the sampling process. Unlike classical autoencoders, the VAE encoder outputs both the mean (μ\muμ) and standard deviation (σ\sigmaσ) for each latent dimension, enabling the generation of a range of values for each latent variable. The latent vector zzz is then sampled using these parameters:

$$z = \mu + \sigma * \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I)$$

Equation 2: Reparameterization Trick

This reparameterization technique allows for backpropagation through the stochastic sampling process by expressing zzz as a deterministic function of μ\muμ, σ\sigmaσ, and a random variable

https://doi.org/10.25163/angiotherapy.8109996

ϵ\epsilonϵ sampled from a standard normal distribution (Kingma & Welling, 2019).

3.4 Model Architectures

We propose two distinct VAE-based architectures for molecular generation:

1. Convolutional Encoder-Decoder Architecture
2. Convolution-GRU Encoder-Decoder Architecture

3.4.1 Convolutional Encoder-Decoder Architecture

As illustrated in Figure 3, this model employs three convolutional layers within the encoder, each followed by a Rectified Linear Unit (ReLU) activation function. The output from the convolutional layers is flattened and passed through two separate linear layers to compute the latent distribution parameters (μ\muμ and logσ2\log \sigma^2logσ2). The latent dimension size is set to 196, a hyperparameter determined through empirical validation.

The decoder reconstructs the one-hot encoded molecular representations by passing the latent code through a series of transposed convolutional layers, each accompanied by ReLU activations. The final layer outputs the reconstructed molecule using a sigmoid activation function (Hancock & Khoshgoftaar, 2020; Dahouda & Joe, 2021).

3.4.2 Convolution-GRU Encoder-Decoder Architecture

Depicted in Figure 4, the second architecture shares the same convolutional encoder as the first model. However, the decoder integrates Multilayer GRUs instead of transposed convolutional layers. After obtaining the latent code through the reparameterization trick (Equation 2), the GRU layers process this code to reconstruct the molecular representations. A final linear layer with sigmoid activation ensures the output maintains the one-hot encoded format (Hancock & Khoshgoftaar, 2020; Dahouda & Joe, 2021).

Both architectures utilize the same encoder setup, consisting of three convolutional layers with ReLU activations, followed by flattening and linear layers to derive the latent space parameters. The primary distinction lies in the decoder design: the first model leverages transposed convolutions, while the second employs GRUs to handle sequential dependencies in molecular data.

3.5 Training Procedure

The models were trained on the ChEMBL22.34 dataset, which comprises a diverse set of molecular structures represented as SMILES strings. Each SMILES string was one-hot encoded and padded to match the length of the longest string in the dataset. The training process involved optimizing the VAE objective function (Equation 1) using the Adam optimizer with a learning rate of 1×10−31 \times 10^{-3}1×10−3.

During training, the models learned to encode the one-hot encoded SMILES strings into a continuous latent space and subsequently decode them back to their original form. The use of convolutional and GRU layers in the encoder and decoder was instrumental in capturing both the structural and sequential nuances of the molecular data, thereby enhancing the validity and novelty of the generated molecules.

3.6 Evaluation Metrics

To comprehensively evaluate the performance of the proposed molecular generative models, several key metrics were utilized. Validity refers to the percentage of generated SMILES strings that are chemically valid, indicating the model's accuracy in producing molecules that adhere to chemical principles. Uniqueness measures the proportion of distinct molecules among all generated samples, highlighting the model's ability to generate diverse structures. Novelty assesses the extent to which the generated molecules differ from those in the training dataset, demonstrating the model's capability to create new, unseen molecular structures. Additionally, Tanimoto Similarity is used as a quantitative measure of structural similarity between generated molecules and known compounds, which helps determine how closely the new molecules resemble existing ones. Together, these metrics provide a robust framework to evaluate the models' ability to generate meaningful, unique, and chemically diverse molecular structures.

4. Evaluation criteria

To design novel and valuable compounds, several essential properties must be present in the generated molecules. First, the molecules must be both chemically and syntactically valid, as validated by methods such as SMILES representation or SDF format analysis researchers often begin by modifying an existing drug to develop new molecules that improve upon the original compound while retaining structural resemblance. One common metric for measuring structural similarity between generated compounds and existing drug molecules is the Tanimoto similarity. This metric is widely recognized in drug development for its efficiency in comparing molecular fingerprints. For molecules AAA and BBB, Tanimoto similarity can be defined over their fingerprint bit vectors as:

$$\text{Tanimoto similarity}= \frac{A.B}{||A||^2+||B||^2-A.B} \qquad (3)$$

where A·BA represents the dot product of the fingerprint bit vectors of molecules A and B, and $\|A\|^2$ $\|B\|^2$are the squared magnitudes of these vectors.

For evaluating the generated molecules, we use the following metrics, given a set GGG of chemically valid molecules, a training set D, n as the count of syntactically valid generated molecules, as the total number of samples:

4.1 Syntactic Validity Ratio (SVR): Measures the ratio of syntactically valid molecules to total generated samples, calculated as:

$$\text{Syntactic Validity Ratio: } \frac{n}{n_{samp}} \qquad (4)$$

4.2 Chemical Validity Ratio (CVR): Assesses the proportion of generated molecules that are chemically viable, or realistic according to chemical laws:

Chemical Validity Ratio: $\frac{|G|}{n}$

(5)

4.3 ***Uniqueness***: Evaluates the distinctiveness of generated molecules by determining the proportion of unique molecules within the generated set:

Uniqueness: $\frac{|\text{set}(G)|}{n}$

(6)

4.4 Novelty: Measures the proportion of generated molecules that are novel and do not overlap with the training set DDD:

Novelty: $1 - \frac{|G \cap D|}{|G|}$

(7)

4.5 Similarity Ratio: Averages the pairwise Tanimoto similarity between molecules in GGG, providing a measure of how structurally similar generated compounds are:

Similarity

$$\frac{\sum_{i=0}^{|G|} \sum_{j=i+1}^{|G|} Tanimoto\ Similarity\left(G_i, G_j\right)}{\frac{|G|(|G|-1)}{2}}$$

(8)

These metrics are pivotal in drug discovery, as they help guide model development, assess the diversity and quality of generated molecules, and aid in selecting potential drug candidates. Metrics like the Syntactic Validity Ratio ensure that generated molecules follow syntactical norms, while the Chemical Validity Ratio checks their adherence to chemical laws. Uniqueness and Novelty are essential for expanding chemical diversity, and Similarity Ratio allows for comparing generated structures to known compounds, providing insight into the balance between novelty and desired structural characteristics.

5. Results

This experiment uses two deep generative models to synthesize new molecular structures. Specifically, two different types of variational autoencoders (VAEs) were employed, each illustrated in Figures 3 and 4.

5.1 Dataset and Hyperparameters

The ZINC dataset, a publicly accessible resource, was chosen for this experiment due to its extensive collection of over 100 million small compounds. This dataset is valuable in drug discovery as it offers a broad range of chemical structures, properties, and biological activities, providing a foundation for exploring numerous therapeutic possibilities. For this experiment, a subset of the ZINC dataset containing 250,000 SMILES strings was used, with lengths ranging from 9 to 109 characters. Since only a small fraction exceeded 60 characters, only those with lengths of 60 or fewer characters were selected, resulting in a filtered dataset of 235,724 samples. A 70:30 split was applied, yielding 165,006 samples for training and 70,718 for testing.

Training was conducted over 35 epochs, with key hyperparameters including a hidden channel size of 32, a latent space dimension of 196, and a convolutional layer kernel size of 3. Both models were optimized using the Adam optimizer on VAE loss, with the learning rate set at 0.001.

5.2 Molecule Synthesis Process

Following model training, the generative models were evaluated and deployed to synthesize novel molecules. The synthesis process began with a seed molecule, commonly aspirin, as shown in Figure 6. Molecule generation involves calculating the latent space point of the starting molecule and using the following sampling formula:

$s = \sigma * \epsilon + \mu$

(9)

Where: μ is the mean obtained from the latent representation of a given starting point. σ is the standard deviation for sampling. In this work, different values of the standard deviation were manually set to observe the different scores depending on how far you are from the starting point (mean). ϵ belongs to normal distribution. A total of 5000 samples were tested, for different values of the standard deviation. For both types of models, taking the pre-trained models with the best score, 5000 samples were generated starting from the SMILE string representation of the aspirin molecule.

5.3 Convolution-Based Model

Using the convolutional encoder-decoder model, molecules were generated from the aspirin seed molecule with a standard deviation of 0.065. The resulting synthesized molecules are displayed in Figure 7, while Figure 8 shows the synthesis ratios. This setup successfully generated a significant number of chemically valid molecules.

When the standard deviation was adjusted to 0.085, the generated molecules exhibited increased uniqueness, as shown in Figures 9 and 10. However, chemical validity slightly decreased compared to the configuration with a 0.065 standard deviation.

5.4 Chemical Viability

Chemically viable molecules must conform to the laws of chemistry, though not all syntactically valid molecules are chemically feasible. Syntactic validity depends on the representation language (e.g., SMILES, SDF) used, and differences may arise depending on how various SMILES parsers interpret the generated SMILES strings. Ensuring consistency with reliable SMILES parsers is recommended for future work.

5.5 Convolution-GRU Based Model

In this setup, molecules were generated using a Convolution-GRU-based model with the aspirin molecule as the seed and a standard deviation of 0.065. The generated molecules are shown in Figure 11, and Figure 12 presents the results, indicating that this model produced fewer valid and unique molecules.

https://doi.org/10.25163/angiotherapy.8109996

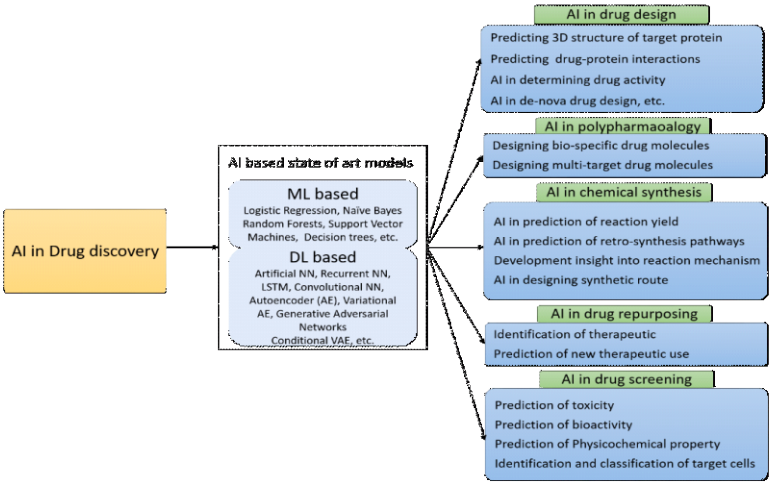
AI in Drug discovery
AI based state of art models
ML based
Logistic Regression, Naïve Bayes Random Forests, Support Vector Machines, Decision trees, etc.
DL based
Artificial NN, Recurrent NN, LSTM, Convolutional NN, Autoencoder (AE), Variational AE, Generative Adversarial Networks Conditional VAE, etc.
AI in drug design
Predicting 3D structure of target protein
Predicting drug-protein interactions
AI in determining drug activity
AI in de-nova drug design, etc.
AI in polypharmaoalogy
Designing bio-specific drug molecules
Designing multi-target drug molecules
AI in chemical synthesis
AI in prediction of reaction yield
AI in prediction of retro-synthesis pathways
Development insight into reaction mechanism
AI in designing synthetic route
AI in drug repurposing
Identification of therapeutic
Prediction of new therapeutic use
AI in drug screening
Prediction of toxicity
Prediction of bioactivity
Prediction of Physicochemical property
Identification and classification of target cells

Figure 1a. A visual representation of the various phases of drug discovery where artificial intelligence (AI) plays a crucial role, highlighting its impact on target identification, lead optimization, and preclinical testing.

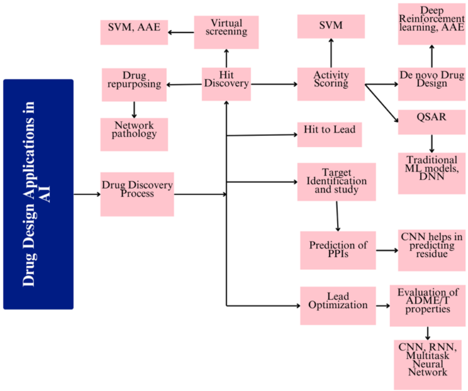
Drug Design Applications in AI
Drug Discovery Process
SVM, AAE
Virtual screening
Drug repurposing
Hit Discovery
Network pathology
SVM
Activity Scoring
Hit to Lead
Target Identification and study
Prediction of PPIs
Lead Optimization
Deep Reinforcement learning, AAE
De novo Drug Design
QSAR
Traditional ML models, DNN
CNN helps in predicting residue
Evaluation of ADME/T properties
CNN, RNN, Multitask Neural Network

Figure. 1b. Detailed illustration of AI integration across different stages of drug discovery, emphasizing the potential for efficiency and innovation in drug development processes.

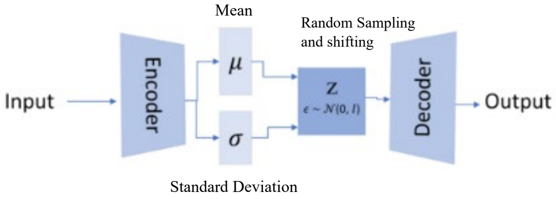

Figure 2. Variational Autoencoder (VAE) Architecture Illustration of the architecture of a variational autoencoder, showcasing the encoder and decoder components, and the latent space representation used for molecular generation.

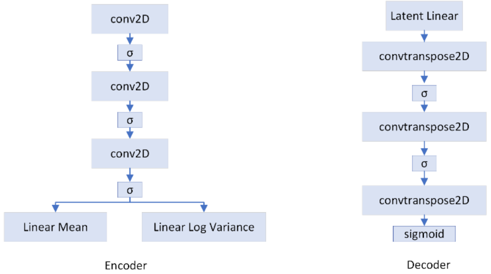

Figure 3. Convolutional Layer-Based Encoder and Decoder Architectures Schematic representation of the encoder and decoder architectures utilizing convolutional layers, demonstrating how these structures contribute to de novo molecular design.

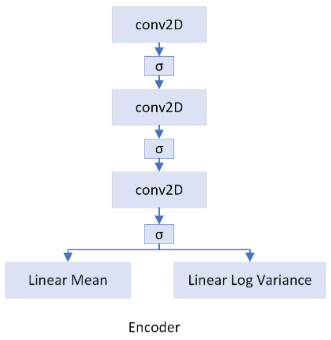

Figure 4. Convolution-Based Encoder and GRU-Based Decoder Model Comparison of the convolutional encoder and GRU-based decoder model architecture, detailing their respective roles in enhancing molecular generation.

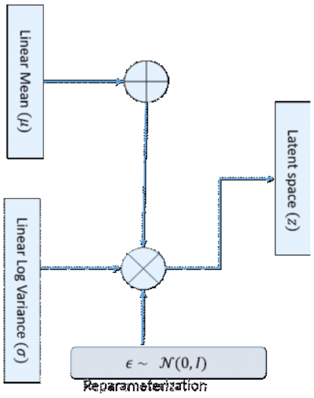

Figure 5. Reparameterization Technique Schematic Illustration of the reparameterization technique used to derive latent space vectors from encoder outputs, facilitating improved sampling for molecular generation.

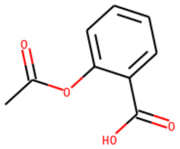

Figure 6. Display of a sample SMILES string alongside its corresponding chemical structure, demonstrating the conversion from text representation to molecular visualization.

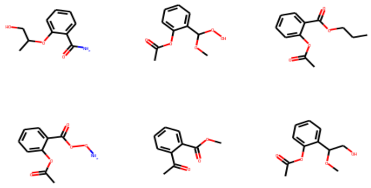

Figure 7. Molecules generated by the convolution-based model with σ=0.06

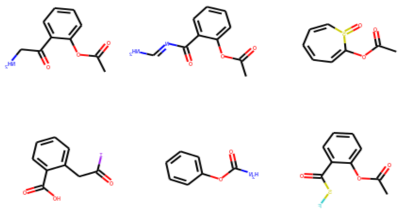

Figure 9. Illustration of additional molecules synthesized by the convolution-based model, highlighting variations at a standard deviation of σ=0.085.

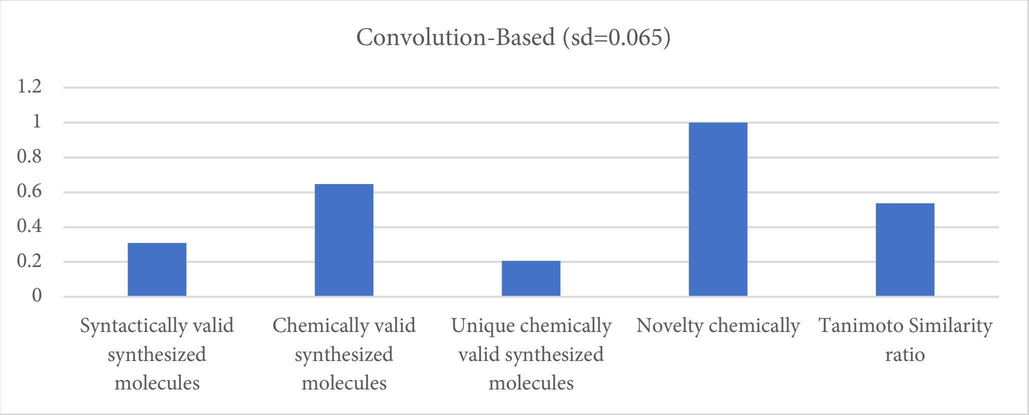

Figure 8. Performance of Convolutional-based encoder-decoder model with σ=0.065

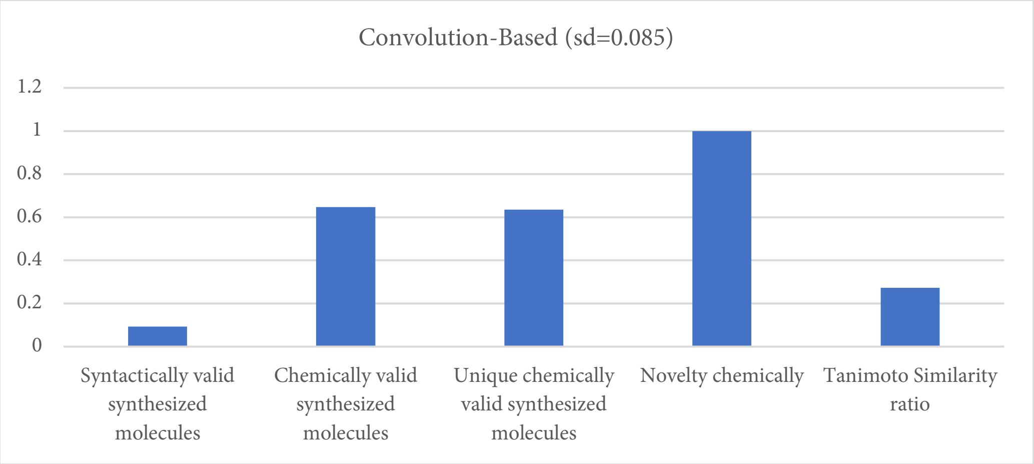

Figure 10. Performance metrics for the convolution-based encoder-decoder model at a standard deviation of σ=0.085, emphasizing improvements in molecular generation.

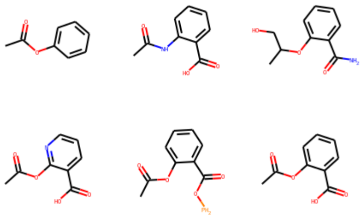

Figure 11. Molecules generated by convolution-GRU-based model with σ=0.065

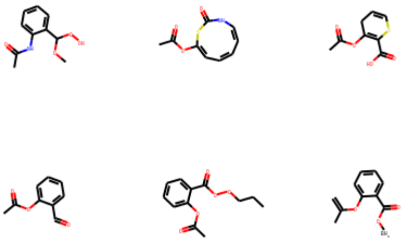

Figure 13. Showcase of molecules synthesized by the convolution-GRU-based model at a standard deviation of σ=0.08, illustrating its generative capabilities.

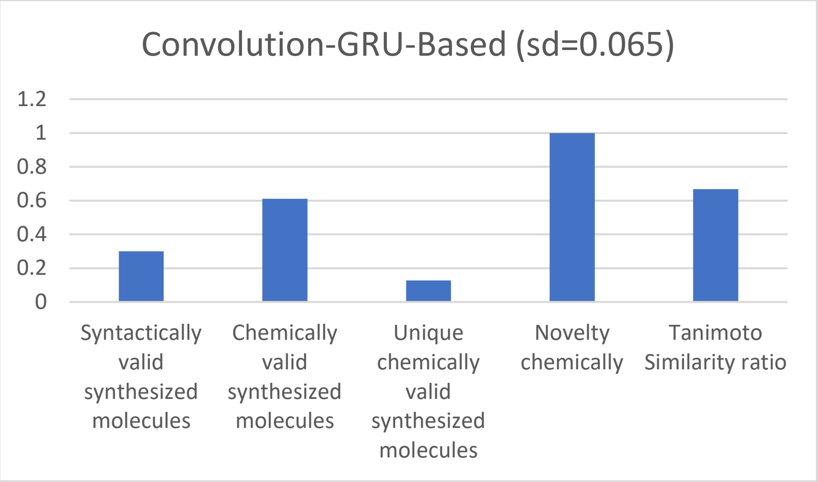

Figure 12. Evaluation of the convolution-GRU-based encoder-decoder model performance at σ=0.065, detailing the accuracy and chemical validity of generated molecules.

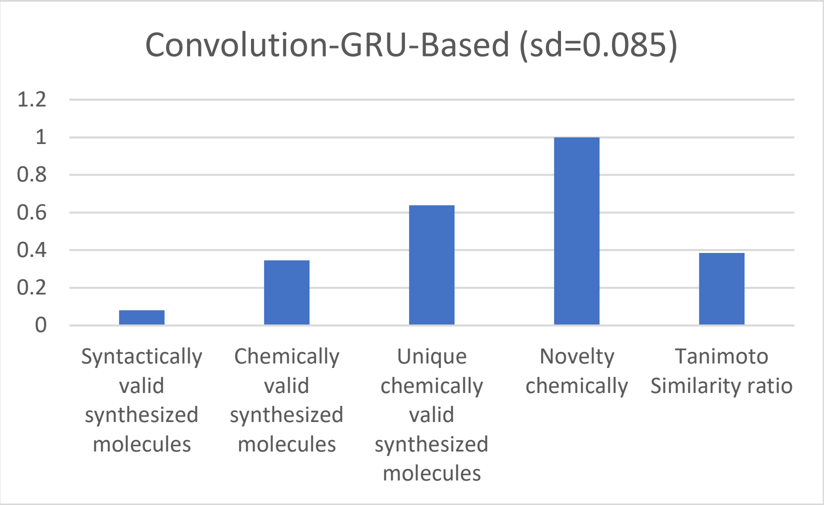

Figure 14. Performance analysis of the convolution-GRU-based encoder-decoder model at a standard deviation of σ=0.085, demonstrating enhancements in molecular synthesis.

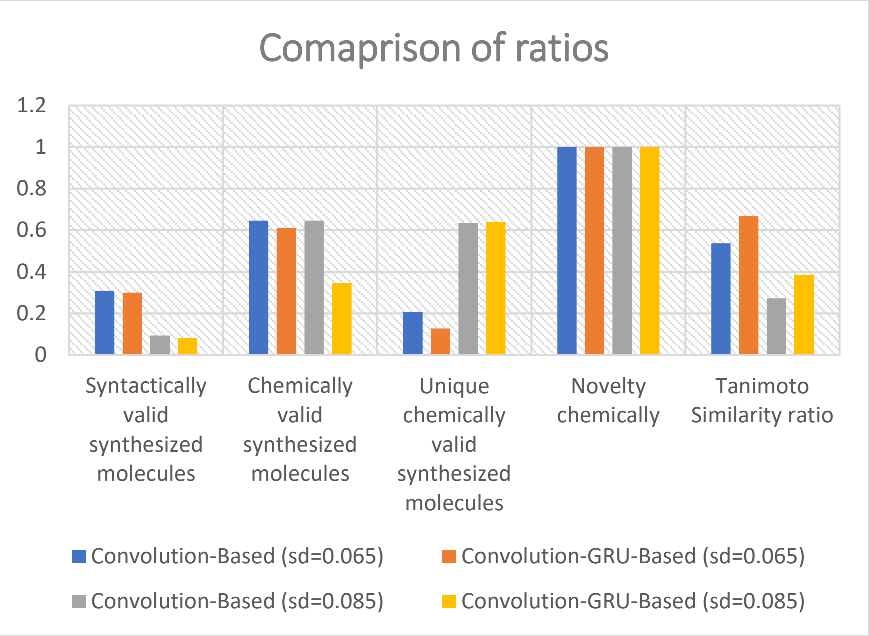

Figure 15. A graphical comparison illustrating the ratios of performance metrics between the convolution-based and convolution-GRU-based models, emphasizing the strengths and weaknesses of each approach.

When the standard deviation was increased to 0.085, the model generated a higher number of valid and unique molecules, but with reduced syntactic validity, as shown in Figures 13 and 14.

5.6 Observations on Model Performance

Testing the performance of the two models under different standard deviations (0.065 and 0.085) revealed that increasing the standard deviation led to more unique molecules. However, other indicators, such as chemical validity and Tanimoto similarity, tended to decrease. Figure 15 presents a comparative graph illustrating these performance trends.

6. Discussion

The recent advancements in molecular generation utilizing convolution-based models have demonstrated a significant ability to produce syntactically and chemically valid molecules derived from the aspirin structure. This research indicates that both convolution-based and convolution-GRU (Gated Recurrent Unit)-based models are highly effective in generating unique, chemically valid molecules. While their performance is largely comparable, the convolution-GRU model exhibits superior results in terms of Tanimoto similarity ratio, which assesses structural similarity between generated compounds and known drugs, making it particularly advantageous in the context of drug discovery.

The achieved test accuracy for the convolution-based encoder and decoder reached an impressive 96.79%, while the convolution + GRU model recorded a slightly lower accuracy of 93.63%. This high level of accuracy signifies the effectiveness of these models in generating relevant molecular structures, although the marginal difference in performance suggests that while the integration of GRU improves structural similarity, it does not drastically alter the overall accuracy of generation.

Despite these promising results, the utilization of Variational Autoencoders (VAEs) for molecule generation poses several inherent challenges that warrant further exploration and refinement. One of the primary issues is the chemical validity and likelihood of generated compounds. VAEs can produce molecules that are chemically improbable or invalid, which undermines their applicability in drug discovery. It is crucial for generated compounds to adhere to established chemical principles and possess a high likelihood of existing within the real chemical space (Jha et al., 2020; Gomez-Bombarelli et al., 2018). This necessitates further research into enhancing the validity of compounds generated by VAEs, ensuring they meet the requisite chemical standards.

In addition to chemical validity, the specificity and diversity of generated molecules pose significant challenges. VAEs may struggle to produce molecules with specific desirable properties, such as selectivity or target binding affinity. This highlights a continuous need to enhance VAEs' capacity to capture and generate a variety of molecular structures that exhibit specific attributes (Chen et al., 2018; Segler et al., 2018). The lack of specificity can impede the efficiency of drug discovery processes, where tailored compounds are often required for successful therapeutic interventions.

Another critical aspect is the management of molecular conformations. Accurately capturing a molecule's bioactivity necessitates representing and generating it in multiple conformations. Enhancing VAE architectures to accommodate various molecular conformations could significantly improve the functionality of generated molecules (Li et al., 2018). A better representation of molecular flexibility is essential for predicting how compounds interact with biological targets, thereby enhancing the drug discovery process.

Moreover, the generation of rare chemical entities presents a challenge for VAEs. Often, compounds that resemble existing drugs constitute a small subset of the overall chemical space, making it difficult for VAEs to generate unique or rare molecules. To overcome this limitation, future research should focus on promoting the synthesis of molecules in less populated regions of chemical space (Zhang et al., 2019). This could involve developing strategies to guide the generative process toward these underrepresented areas.

Multi-objective optimization is another pressing challenge in the context of drug development. The process involves balancing competing goals such as safety, pharmacokinetics, and efficacy. Developing VAE architectures capable of managing these multi-objective optimization tasks is complex and necessitates innovative approaches (Huang et al., 2020). Achieving a balance among these attributes is critical for generating molecules that are not only effective but also safe for clinical use.

The interpretability of latent representations generated by VAEs is also vital for medicinal chemists. Understanding the latent space of produced compounds allows researchers to connect specific molecular features with desired chemical properties. Future research efforts should focus on improving the interpretability of the latent space, thereby enabling better insights into the underlying mechanisms driving molecular generation (Kearnes et al., 2016). This could facilitate the identification of promising drug candidates by linking structural attributes to biological activity.

Additionally, the concept of transfer learning across targets holds significant promise for enhancing drug discovery efforts. By developing VAE models that can leverage information from previously studied drug targets to generate compounds for new targets, the drug development process can be accelerated. This approach requires the formulation of effective transfer learning strategies within the VAE framework (Ramsundar et al., 2017).

Addressing the limitations associated with the availability of extensive datasets in drug discovery is also crucial. Enhancing the data efficiency of VAEs in learning from limited datasets and producing significant compounds is essential, particularly in scenarios where access to large, labeled datasets is restricted (Vinyals et al., 2016; DeWolf et al., 2020). Developing methodologies that optimize the learning process from smaller datasets can expand the applicability of VAEs in diverse drug discovery contexts.

Incorporating safety and toxicity predictions into the VAE architecture represents a further avenue for improving drug discovery applications. By integrating prediction modules that assess the safety and toxicity of generated compounds, VAEs can be made more effective in identifying viable drug candidates (Liu et al., 2018; Xu et al., 2017). This integration ensures that the generated compounds not only demonstrate chemical validity but also possess a favorable safety profile.

To address the challenges inherent in de novo drug generation, several strategies may be implemented in future research. Hybrid models that combine VAEs with other machine learning approaches, such as graph neural networks and reinforcement learning, can capture complex molecular interactions and enhance generation quality (Yang et al., 2017). Additionally, employing hierarchical structures that represent molecules at various levels of abstraction may provide improved control over generated molecular structures. Implementing conditional VAEs that utilize scaffolds and target attributes as conditioning elements could guide the generation process more effectively.

Furthermore, incorporating explicit chemical rules into the VAE architecture can enhance the validity of generated compounds. Utilizing knowledge graphs to direct the generative process and represent chemical information can also contribute to improved outcomes. Additionally, integrating human expertise through expert systems into VAE frameworks may further enrich the generative capabilities of these models.

Establishing criteria for assessing the chemical viability of produced compounds is crucial for advancing drug discovery. This includes developing drug-likeness metrics, predicting target binding affinity, and evaluating potential toxicity and safety of generated compounds. Resolving these challenges requires continuous innovation in model architectures, the integration of domain knowledge, and the establishment of specialized assessment metrics to ensure the chemical validity and suitability of generated molecules for drug discovery applications.

7. Conclusion

In conclusion, the exploration of molecular generation using convolution-based models, particularly those utilizing GRU, demonstrates significant potential in producing chemically valid and syntactically accurate compounds. While both models achieved commendable test accuracies, challenges remain in ensuring the generated molecules adhere to chemical principles, possess desired characteristics, and can be effectively represented in various conformations. Future advancements should focus on refining VAE architectures, enhancing data efficiency, and integrating safety predictions to improve drug discovery outcomes. The proposed hybrid models, conditional VAEs, and explicit chemical rules are promising avenues for addressing these challenges. Ultimately, the successful application of these models hinges on overcoming existing limitations, which will facilitate the identification of effective drug candidates and enhance the overall drug development process. Continuous innovation and integration of domain knowledge will be critical in advancing the field of molecular generation for therapeutic applications.

Author contributions

M.A.S. and D.S. contributed to the conceptualization and design of the study. T.A.T. was responsible for data analysis, manuscript drafting, and provided critical revisions. All authors reviewed and approved the final version of the manuscript.

Acknowledgment

The authors were grateful to their department.

Competing financial interests

The authors have no conflict of interest.

References

Ahmed, T., Teli, T., & Masoodi, F. (2021). Blockchain in healthcare: Challenges and opportunities. In Proceedings of the International Conference on IoT Based Control Networks Intelligent Systems - ICICNIS 2021. SSRN. https://doi.org/10.2139/ssrn.3882744

Ahmed, T., Teli, T., Yousuf, R., & Masoodi, F. (2021). Security concerns and privacy preservation in blockchain-based IoT systems: Opportunities and challenges. ICICNIS 2020. SSRN. https://ssrn.com/abstract=3768235

Ali, M., & Aittokallio, T. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. Biophysical Reviews, 11, 31–39. https://doi.org/10.1007/s12551-018-0446-z

Askr, H., Elgeldawi, E., Aboul Ella, H., et al. (2023). Deep learning in drug discovery: An integrative review and future challenges. Artificial Intelligence Review, 56, 5975–6037. https://doi.org/10.1007/s10462-022-10306-1

Bhadwal, A. S., Kumar, K., & Kumar, N. (2024). NRC-VABS: Normalized reparameterized conditional variational autoencoder with applied beam search in latent space for drug molecule design. Expert Systems with Applications, 240, 122396.

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., & Jensen, K. F. (2022). Generative models for molecular discovery: Recent advances and challenges. Wiley Interdisciplinary Reviews: Computational Molecular Science, 12, e1608. https://doi.org/10.1002/wcms.1608

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., & Chen, H. (2018). Application of generative autoencoder in de novo molecular design. Molecular Informatics, 37(1-2), 1700123. https://doi.org/10.1002/minf.201700123

Brown, N. (2015). In silico medicinal chemistry: Computational methods to support drug design. Royal Society of Chemistry.

Chan, H. S., et al. (2019). Advancing drug discovery via artificial intelligence. Trends in Pharmacological Sciences, 40(8), 592–604.

Chen, G., Shen, Z., Iyer, A., Ghumman, U. F., Tang, S., Bi, J., Chen, W., & Li, Y. (2020). Machine-learning-assisted de novo design of organic molecules and polymers: Opportunities and challenges. Polymers, 12(1), 163. https://doi.org/10.3390/polym12010163

Chen, S., Lin, T., Basu, R., Ritchey, J., Wang, S., Luo, Y., ... & Cheng, X. (2024). Design of target-specific peptide inhibitors using generative deep learning and molecular dynamics simulations. Nature Communications, 15(1), 1611.

Cheng, T., Hao, M., Takeda, T., et al. (2017). Large-scale prediction of drug-target interaction: A data-centric review. AAPS Journal, 19, 1264–1275. https://doi.org/10.1208/s12248-017-0092-6

Ciallella, H. L., & Zhu, H. (2019). Advancing computational toxicology in the big data era by artificial intelligence: Data-driven and mechanism-driven modelling for chemical toxicity. Chemical Research in Toxicology, 32, 536–547.

Dahouda, M. K., & Joe, I. (2021). A deep-learned embedding technique for categorical features encoding. IEEE Access, 9, 114381–114391. https://doi.org/10.1109/ACCESS.2021.3104357

Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2022). Machine learning in drug discovery: A review. Artificial Intelligence Review, 55(3), 1947-1999. https://doi.org/10.1007/s10462-021-10058-4

David, L., Thakkar, A., Mercado, R., et al. (2020). Molecular representations in AI-driven drug discovery: A review and practical guide. Journal of Cheminformatics, 12, 56. https://doi.org/10.1186/s13321-020-00460-5

Diederik, P. K., & Max, W. (2019). An introduction to variational autoencoders.

Diederik, P. K., & Max, W. (2019). An introduction to variational autoencoders.

Ekins, S. (2016). The next era: Deep learning in pharmaceutical research. Pharmaceutical Research, 33(11), 2594-2603. https://doi.org/10.1007/s11095-016-2029-7

Gangwal, A., & Lavecchia, A. (2024). Unlocking the potential of generative AI in drug discovery. Drug Discovery Today, 103992.

Gao, M., Igata, H., Takeuchi, A., Sato, K., & Ikegaya, Y. (2017). Machine learning-based prediction of adverse drug effects: An example of seizure-inducing compounds. Journal of Pharmacological Sciences, 133(2), 70-78. https://doi.org/10.1016/j.jphs.2017.01.003

Gómez-Bombarelli, R., Duvenaud, D., Hernández-Lobato, J., Aguilera-Iparraguirre, J., Hirzel, T., Adams, R., & Aspuru-Guzik, A. (2016). Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science, 4, 268-276. https://doi.org/10.1021/acscentsci.7b00572

Guan, S., & Wang, G. (2024). Drug discovery and development in the era of artificial intelligence: From machine learning to large language models. Artificial Intelligence Chemistry, 2(1), 100070.

Gupta, A., Müller, A. T., Huisman, B. J. H., Fuchs, J. A., Schneider, P., & Schneider, G. (2018). Generative recurrent networks for de novo drug design. Molecular Informatics, 37(1-2), 1700111. https://doi.org/10.1002/minf.201700111

Gupta, R., Srivastava, D., Sahu, M., et al. (2021). Artificial intelligence to deep learning: Machine intelligence approach for drug discovery. Molecular Diversity, 25, 1315–1360. https://doi.org/10.1007/s11030-021-10217-3

Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. Journal of Big Data, 7, 28. https://doi.org/10.1186/s40537-020-00305-w

Hirohara, M., Saito, Y., Koda, Y., Sato, K., Sakakibara, Y., & Yasubumi, K. (2018). Convolutional neural network based on SMILES representation of compounds for detecting chemical motifs. BMC Bioinformatics, 19, 10. https://doi.org/10.1186/s12859-018-2523-5

Joo, S., Kim, M., Yang, J., & Park, J. (2020). Generative model for proposing drug candidates satisfying anticancer properties using a conditional variational autoencoder. ACS Omega. https://doi.org/10.1021/acsomega.0c01149

Kadurin, A., Aliper, A., Kazennov, A., et al. (2017). The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. Oncotarget, 8(7), 10883-10890. https://doi.org/10.18632/oncotarget.14073

Kanakala, G. C., Devata, S., Chatterjee, P., & Priyakumar, U. D. (2024). Generative artificial intelligence for small molecule drug design. Current Opinion in Biotechnology, 89, 103175.

Kiriiri, G. K., Njogu, P. M., & Mwangi, A. N. (2020). Exploring different approaches to improve the success of drug discovery and development projects: A review. Future Journal of Pharmaceutical Sciences, 6(1), 27. https://doi.org/10.1186/s43094-020-00047-9

Lauv Patel, T., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. (2020). Machine learning methods in drug discovery. Molecules.

Lavecchia, A. (2024). Navigating the frontier of drug-like chemical space with cutting-edge generative AI models. Drug Discovery Today, 104133.

Lim, J., Ryu, S., Kim, J. W., et al. (2018). Molecular generative model based on conditional variational autoencoder for de novo molecular design. Journal of Cheminformatics, 10, 31. https://doi.org/10.1186/s13321-018-0286-7

Liu, P., Li, H., Li, S., et al. (2019). Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional networks. BMC Bioinformatics, 20, 408. https://doi.org/10.1186/s12859-019-2910-6

Mariam, Z., Niazi, S. K., & Magoola, M. (2024). Unlocking the future of drug development: Generative AI, digital twins, and beyond. BioMedInformatics, 4(2), 1441–1456.

Méndez-Lucio, O., Baillif, B., Clevert, D. A., et al. (2020). De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nature Communications, 11, 10. https://doi.org/10.1038/s41467-019-13807-w

Middaugh, C. R., & Pearlman, R. (1999). Proteins as drugs: Analysis, formulation, and delivery. In D. L. Oxender & L. E. Post (Eds.), Novel therapeutics from modern biotechnology (pp. 35-60). Handbook of Experimental Pharmacology, vol 137. Springer. https://doi.org/10.1007/978-3-642-59990-3_3

Nag, S., Baidya, A. T. K., Mandal, A., et al. (2022). Deep learning tools for advancing drug discovery and development. 3 Biotech, 12(1), 110. https://doi.org/10.1007/s13205-022-03165-8

Ozawa, M., Nakamura, S., Yasuo, N., & Sekijima, M. (2024). IEV2Mol: Molecular generative model considering protein–ligand interaction energy vectors. Journal of Chemical Information and Modeling.

Partin, A., Brettin, T., Evrard, Y. A., et al. (2021). Learning curves for drug response prediction in cancer cell lines. BMC Bioinformatics, 22, 252. https://doi.org/10.1186/s12859-021-04163-y

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. Drug Discovery Today, 26(1), 80-93. https://doi.org/10.1016/j.drudis.2020.10.010

Pereira, J. C., et al. (2016). Boosting docking-based virtual screening with deep learning. Journal of Chemical Information and Modeling, 56, 2495–2506.

Prykhodko, O., Johansson, S. V., Kotsias, P. C., et al. (2019). A de novo molecular generation method using latent vector based generative adversarial networks. Journal of Cheminformatics, 11, 74. https://doi.org/10.1186/s13321-019-0397-9

Rusdi, N. A., Kasihmuddin, M. S. M., Romli, N. A., Manoharam, G., & Mansor, M. A. (2023). Multi-unit discrete Hopfield neural network for higher order supervised learning through logic mining: Optimal performance design and attribute selection. Journal of King Saud University - Computer and Information Sciences, 35(5), 101554. https://doi.org/10.1016/j.jksuci.2023.101554.

Teli, T. A., Yousuf, R., & Khan, D. A. (2022). Ensuring secure data sharing in IoT domains using blockchain. In M. M. Ghonge, S. Pramanik, R. Mangrulkar, & D.-N. Le (Eds.), Cyber security and digital forensics. https://doi.org/10.1002/9781119795667.ch9.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18(6), 463-477. https://doi.org/10.1038/s41573-019-0024-5

Yang, Y., Adelstein, S. J., & Kassis, A. I. (2009). Target discovery from data mining approaches. Drug Discovery Today, 14, 147–154.

Yaseen, B. T., & Kurnaz, S. (2021). Drug–target interaction prediction using artificial intelligence. Applied Nanoscience. https://doi.org/10.1007/s13204-021-02000-5

Zhang, C., Xie, L., Lu, X., Mao, R., Xu, L., & Xu, X. (2024). Developing an improved cycle architecture for AI-based generation of new structures aimed at drug discovery. Molecules, 29(7), 1499.

Zhu, H. (2020). Big data and artificial intelligence modelling for drug discovery. Annual Review of Pharmacology and Toxicology, 60, 573–589.