DOI: https://doi.org/10.71317/kjard.2.11.2026.1087

Kashmir Journal of Academic Research and Development

Journal homepage: https://rjsaonline.org/index.php/KJARD

A Hybrid VAE–GAN Generative Artificial Intelligence Framework for De Novo Molecular Design, Novel Drug Candidate Generation, and Pharmaceutical Lead Optimization

Muhammad Ali Khan
University of Sargodha, Sargodha, Pakistan
Email: admk202024@gmail.com
Dr. Syed Mobasher Ali Abid
Department of Pharmacy, COMSATS University Islamabad, Abbottabad Campus, Pakistan
Email: mobasher@cuiatd.edu.pk
Omama
Department of Pharmacy, Jinnah University for Women, Karachi, Pakistan
Email: omamajaved2007@gmail.com
Javeria Javed
Department of Pharmacology, Jinnah University for Women, Karachi, Pakistan
Email: javeria_javed23@yahoo.com

ARTICLE INFO

Received:
October 16, 2026
Revised:
October 31, 2026
Accepted:
November 09, 2026
Available Online:
November 20, 2026

Keywords:
Hybrid VAE–GAN; Generative Artificial Intelligence; De Novo Molecular Design; Drug Discovery; Molecular Generation; Pharmaceutical Lead Optimization; ZINC; ChEMBL; Drug-Likeness.

Corresponding Author:
Muhammad Ali Khan
Email:
admk202024@gmail.com

ABSTRACT

The discovery and optimization of novel drug candidates remain significant challenges in pharmaceutical research due to the enormous chemical search space, high development costs, and complex molecular property requirements. This study presents a hybrid Variational Autoencoder–Generative Adversarial Network (VAE–GAN) framework for de novo molecular design, novel drug candidate generation, and pharmaceutical lead optimization. The proposed framework integrates the latent-space learning capabilities of a Variational Autoencoder with the adversarial training mechanism of a Generative Adversarial Network to enhance molecular generation, structural diversity, and physicochemical property optimization. The ZINC dataset, comprising 250,000 molecular structures represented as canonical Simplified Molecular Input Line Entry System (SMILES) strings, was employed for model development, while ChEMBL was incorporated for external molecular evaluation. Molecular preprocessing included structural standardization, duplicate removal, and dataset partitioning to ensure data consistency and reliable model assessment. The generated molecules were evaluated using chemical validity, uniqueness, novelty, internal diversity, quantitative estimate of drug-likeness (QED), and synthetic accessibility score (SAS). The hybrid VAE–GAN framework achieved 96.8% molecular validity, 94.1% uniqueness, and 89.5% novelty, demonstrating its capacity to generate chemically plausible and structurally distinct molecular candidates. Furthermore, the proposed model obtained an average QED score of 0.71, a synthetic accessibility score of 2.8, and an internal diversity score of 0.84, outperforming the standalone VAE baseline, which recorded corresponding values of 0.63, 3.4, and 0.77. Property-guided latent-space optimization further increased the proportion of molecules satisfying predefined drug-likeness criteria from 57.2% to 73.6%, indicating improved molecular optimization performance. These findings highlight the effectiveness of integrating variational representation learning and adversarial optimization for generating diverse, drug-like molecular structures while maintaining favorable synthetic accessibility characteristics. The proposed computational framework provides a scalable and promising approach for molecular exploration, virtual compound library generation, and pharmaceutical lead identification. Overall, this research demonstrates the potential of hybrid generative artificial intelligence architectures to accelerate early-stage drug discovery and support the development of novel therapeutic candidates through computationally guided molecular design and optimization strategies.

Introduction

The discovery and development of novel therapeutic compounds constitute one of the most challenging and resource-intensive processes in pharmaceutical research. Conventional drug discovery involves multiple interconnected stages, including target identification, hit discovery, lead identification, lead optimization, preclinical evaluation, and clinical development. Although advances in computational chemistry, molecular biology, and high-throughput screening have substantially improved the identification of biologically active compounds, the exploration of chemically diverse molecular structures remains a fundamental challenge. The enormous size of the theoretically accessible chemical space, together with the complex relationships between molecular structure, biological activity, physicochemical properties, and synthetic feasibility, limits the efficiency of traditional drug development strategies. Consequently, there is growing interest in computational approaches capable of generating novel molecular structures and systematically optimizing their pharmaceutical characteristics [1]. Artificial intelligence (AI) and machine learning (ML) have emerged as important computational technologies in modern drug discovery, offering opportunities to accelerate molecular screening, predict compound properties, and support rational drug design. In particular, generative artificial intelligence has introduced a paradigm in which computational models can learn the statistical and structural characteristics of existing chemical compounds and subsequently generate previously unexplored molecular candidates. Unlike conventional virtual screening methods, which primarily evaluate molecules contained within predefined chemical libraries, de novo molecular design enables the creation of new chemical structures through computational generation and optimization. Deep generative models can therefore support chemical space exploration, scaffold diversification, and the identification of compounds with desirable molecular characteristics [2]. However, transforming computationally generated structures into experimentally useful pharmaceutical leads requires the simultaneous consideration of chemical validity, structural novelty, molecular diversity, drug-likeness, and synthetic accessibility. Recent developments in deep learning have introduced several generative modeling architectures for molecular design, including Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), recurrent neural networks, graph-based generative models, and other probabilistic and optimization-driven approaches. Among these techniques, VAEs have attracted substantial attention because of their ability to transform discrete molecular representations into continuous and differentiable latent spaces suitable for molecular reconstruction, interpolation, and property-guided optimization. Gómez-Bombarelli et al. [4] demonstrated that molecular structures represented using the Simplified Molecular Input Line Entry System (SMILES) could be encoded into a continuous latent representation and subsequently decoded into molecular structures. Their findings established an important foundation for data-driven molecular design by enabling computational exploration of chemical space through learned molecular embeddings. The VAE architecture generally consists of an encoder that maps molecular representations into a probabilistic latent distribution and a decoder that reconstructs molecular structures from sampled latent variables. By learning a structured latent representation, VAEs facilitate molecular generation and provide a computational mechanism for modifying molecular characteristics through latent-space exploration. Nevertheless, conventional VAE-based molecular generation presents several methodological challenges. In particular, string-based molecular representations may produce syntactically incorrect or chemically invalid molecules during decoding, while reconstruction-oriented optimization does not necessarily ensure that generated compounds exhibit favorable pharmaceutical properties. Furthermore, the balance between latent-space regularization and accurate molecular reconstruction can influence structural diversity and generation quality. Graph-based approaches, including the Junction Tree Variational Autoencoder proposed by Jin et al. [3], have addressed some of these limitations by incorporating molecular substructures and graph connectivity into the generation process. Despite these advances, the development of molecular generators capable of balancing validity, novelty, diversity, and property optimization remains an active area of investigation. Generative Adversarial Networks represent another influential class of deep generative models. Originally introduced by Goodfellow et al. [4], GANs employ an adversarial learning mechanism involving two competing neural networks: a generator that produces synthetic samples and a discriminator that attempts to distinguish generated samples from authentic training data. Through iterative adversarial optimization, the generator learns to approximate the underlying data distribution and produce increasingly realistic outputs. The application of adversarial learning to molecular generation has demonstrated its potential for producing chemically meaningful structures and incorporating objective-driven optimization strategies. For example, Guimaraes et al. [5] introduced an objective-reinforced adversarial framework for sequence generation, illustrating how adversarial training can be combined with optimization objectives relevant to molecular design. However, GAN-based molecular generation also encounters difficulties associated with training instability, limited sample diversity, mode collapse, and the discrete nature of molecular representations. Such limitations can restrict the reliability of molecular generation and reduce the effective exploration of chemical space. The complementary characteristics of VAEs and GANs provide a compelling rationale for developing integrated generative architectures. VAEs offer probabilistic representation learning and an organized latent space that supports molecular reconstruction and optimization, whereas GANs introduce adversarial feedback capable of improving the distributional quality of generated samples. Integrating these mechanisms within a unified VAE–GAN framework may therefore provide a means of combining structured latent-space learning with distribution-aware molecular generation. In the proposed approach, variational encoding and decoding facilitate

molecular representation learning, while adversarial optimization is intended to encourage the generation of molecular structures that better reflect the characteristics of the training distribution. Furthermore, property-guided latent-space exploration can provide a mechanism for identifying molecular candidates exhibiting desirable drug-like characteristics [6]. Nevertheless, the benefits of such integration must be established experimentally, since adversarial optimization alone does not guarantee chemical validity, improved molecular diversity, or favorable pharmacological behavior. A central challenge in generative molecular design is the development of reliable strategies for assessing molecular quality. The generation of a chemically valid structure does not necessarily imply that the compound possesses meaningful pharmaceutical potential. Accordingly, comprehensive evaluation requires multiple complementary metrics rather than reliance on a single molecular descriptor. Chemical validity measures whether generated structures satisfy basic chemical representation and valence constraints, whereas uniqueness evaluates the proportion of nonduplicate molecules within a generated collection. Novelty measures the extent to which generated structures differ from those represented in the reference training dataset, while internal diversity assesses structural dissimilarity among generated molecules. Benchmarking platforms such as GuacaMol and Molecular Sets (MOSES) have emphasized the importance of standardized evaluation procedures for assessing the performance of molecular generative models.

In addition to structural evaluation, drug-likeness and synthetic feasibility are essential considerations during pharmaceutical lead identification and optimization. The Quantitative Estimate of Drug-likeness (QED), introduced by Bickerton et al, provides a quantitative assessment of molecular drug-likeness based on a combination of physicochemical descriptors and desirability functions. Similarly, the synthetic accessibility score (SAS) developed by Ertl and Schuffenhauer estimates the relative difficulty of synthesizing a molecular structure by considering fragment contributions and structural complexity. These descriptors provide useful computational criteria for ranking generated compounds and supporting molecular optimization. However, neither QED nor SAS independently establishes biological efficacy, target selectivity, toxicity, or experimental synthesizability. Therefore, their combined application with structural validity, novelty, and diversity metrics offers a broader preliminary assessment of the pharmaceutical relevance of computationally generated molecular candidates. The availability of large-scale chemical databases has further supported the development and validation of generative molecular design systems [7]. ZINC is an established chemical database that provides molecular structures for virtual screening and computational drug discovery. It contains diverse chemical compounds that can be represented using SMILES and other molecular formats, making it suitable for training data-driven molecular generation models. Similarly, ChEMBL provides curated chemical structures and associated biological activity information, supporting molecular characterization and comparative evaluation. The integration of standardized chemical datasets with modern deep generative architectures enables systematic investigations of molecular representation learning, chemical space exploration, and computational lead optimization. Nevertheless, reliable evaluation requires careful molecular preprocessing, duplicate removal, consistent chemical standardization, and appropriate separation of training and evaluation data to reduce data leakage and misleading performance estimates. Despite considerable progress in AI-assisted molecular design, several challenges remain unresolved. Existing generative architectures may demonstrate strong performance in individual evaluation criteria while exhibiting limitations in other important characteristics, including structural novelty, molecular diversity, drug-likeness, or synthetic accessibility. The generation of numerous valid compounds, for instance, may be accompanied by repeated chemical scaffolds, whereas aggressive property optimization may reduce structural diversity or favor molecules that exploit weaknesses in computational scoring functions. Recent reviews have highlighted the continuing challenges of balancing multiple molecular objectives, ensuring synthetic feasibility, maintaining reproducibility, and translating computationally generated structures into experimentally validated drug candidates [8]. These considerations emphasize the need for generative frameworks that integrate complementary learning strategies and employ multidimensional molecular evaluation. To address these challenges, the present study proposes a hybrid Variational Autoencoder–Generative Adversarial Network (VAE–GAN) framework for de novo molecular design, novel drug candidate generation, and pharmaceutical lead optimization. The proposed framework combines variational latent-space representation learning with adversarial optimization to investigate whether their integration can improve molecular generation performance relative to a standalone VAE architecture. A curated collection of 250,000 molecular structures from the ZINC dataset is employed for model development, with ChEMBL incorporated as an external chemical reference for molecular evaluation. Molecular structures are represented as canonical SMILES strings and subjected to standardized preprocessing to improve dataset consistency and model reliability. The generated compounds are assessed using chemical validity, uniqueness, novelty, internal diversity, QED, and SAS, providing a multidimensional assessment of molecular generation quality and preliminary drug-like characteristics. Furthermore, property-guided latent-space optimization is investigated as a strategy for increasing the proportion of generated molecules satisfying predefined molecular desirability criteria [9]. The proposed framework is evaluated against a standalone VAE baseline to determine the contribution of adversarial learning to molecular generation and optimization performance. The primary objectives of this study are to develop a unified VAE–GAN architecture for molecular generation, evaluate its ability to produce valid and structurally diverse chemical compounds, assess the drug-likeness and estimated synthetic accessibility of generated molecules, and investigate the effectiveness of property-guided optimization in

improving molecular candidate selection. Through these investigations, the study aims to contribute to the development of computational approaches for efficient chemical space exploration, virtual compound library generation, and early-stage pharmaceutical lead identification. The proposed methodology is intended to provide a foundation for further structure-based, predictive, and experimental investigations, rather than replace the biological and pharmacological validation required to establish therapeutic potential

Deep Generative Models for De Novo Molecular Design

Deep generative models have emerged as a major advancement in artificial intelligence-assisted drug discovery by enabling the computational generation of molecular structures beyond those explicitly represented in existing chemical libraries. Traditional approaches to molecular discovery, including high-throughput screening, ligand-based virtual screening, and structure-based screening, generally depend on the availability of predefined compound collections. Although these methodologies have contributed substantially to pharmaceutical research, their effectiveness is constrained by the molecular diversity and structural coverage of the screening libraries. In contrast, deep generative models learn statistical relationships within molecular datasets and use these learned representations to generate new chemical structures. This capability provides an alternative strategy for exploring previously unexamined regions of chemical space and identifying molecules with potentially useful pharmaceutical characteristics. The fundamental objective of de novo molecular design is to identify molecular structures that satisfy desirable chemical, physicochemical, and, where relevant, biological characteristics without being restricted to previously synthesized compounds. Deep generative architectures accomplish this by learning distributions over molecular representations and sampling candidate structures from the resulting models [10]. Depending on the selected architecture, molecular generation may involve probabilistic latent-variable sampling, sequential token prediction, adversarial learning, invertible transformations, or iterative denoising. Such approaches have expanded the methodological possibilities for molecular design by enabling automated scaffold generation, molecular modification, structural diversification, and computational property optimization. An essential consideration in generative molecular modeling is the representation of chemical structures. Molecular information may be encoded as one-dimensional molecular strings, two-dimensional molecular graphs, or three-dimensional atomic configurations. String-based representations, particularly the Simplified Molecular Input Line Entry System (SMILES), express molecular connectivity using sequences of characters and are compatible with recurrent neural networks, Transformer architectures, and sequence-based autoencoders. Their computational simplicity facilitates integration with language-modeling techniques and established cheminformatics workflows. However, conventional SMILES generation is sensitive to token ordering and chemical syntax, and unrestricted string sampling may produce representations that cannot be interpreted as chemically valid molecules. Molecular graph representations provide an alternative by directly modeling atoms as nodes and chemical bonds as edges, allowing generative algorithms to incorporate chemical connectivity and structural constraints [11]. Three-dimensional molecular representations extend these approaches by describing atomic coordinates and spatial relationships, which are relevant to molecular conformation, stereochemistry, and structure-based drug design. The comparative characteristics of major generative architectures are summarized in Table 1. The comparison highlights the distinct methodological contributions of existing model families and their respective limitations in molecular generation. In particular, the ability of VAEs to learn structured latent representations and the ability of GANs to provide adversarial distributional feedback create a rationale for investigating integrated VAE–GAN architectures. Their combination may support molecular reconstruction, latent-space exploration, and distribution-aware generation within a unified framework, although the extent of any improvement must be established through controlled experimental comparison.

Table 1: Comparative overview of major deep generative modeling architectures for de novo molecular design and pharmaceutical lead optimization.

Deep generative architecture	Fundamental learning mechanism	Principal advantages for molecular design
Recurrent Neural Networks (RNNs/LSTMs)	Sequential prediction of molecular tokens using recurrent hidden states	Compatible with SMILES; supports molecular sequence generation and fine-tuning
Transformer-based models	Attention-driven autoregressive or sequence-to-sequence learning	Captures long-range token relationships; supports large-scale pretraining and conditional generation
Variational Autoencoders (VAEs)	Probabilistic encoding, latent-variable sampling, and molecular decoding	Continuous latent space; reconstruction; interpolation; property-guided optimization

Generative Adversarial Networks (GANs)	Adversarial optimization between generator and discriminator	Distributional feedback; flexible generation objectives; potential for high-quality molecular samples
Hybrid Generative Architectures	Integration of complementary generative learning objectives	Potential to combine latent-space organization, distributional learning, and molecular optimization

The growing diversity of deep generative architectures has also increased the importance of reproducible and standardized model evaluation. Brown et al. introduced GuacaMol as a benchmarking framework for evaluating molecular generation under distribution-learning and goal-directed optimization tasks. Standardized evaluation approaches help distinguish a model's ability to reproduce chemical data distributions from its ability to generate molecules satisfying particular property objectives. Comparisons between generative architectures should therefore consider molecular validity, uniqueness, novelty, structural diversity, and property distributions alongside computational efficiency and sampling constraints. Importantly,model performance is sensitive to dataset size, training–test separation, molecular standardization, and the precise operational definitions of evaluation metrics. Consequently, a high molecular validity percentage obtained on one benchmark cannot be assumed to indicate superior performance on another dataset or under different molecular generation conditions [12]. The conceptual organization of these generative approaches within a drug discovery workflow is illustrated in Figure 1. The figure distinguishes molecular data representation from the learning mechanisms employed by different generative architectures and emphasizes the importance of downstream chemical evaluation and property-guided candidate selection. Although individual generative models differ in their training strategies, most practical molecular design workflows require chemical preprocessing, molecular generation, structural validation, and prioritization according to specified pharmaceutical objectives. The framework also demonstrates the conceptual position of hybrid VAE–GAN learning, in which latent-space representation and adversarial distributional feedback are investigated as complementary mechanisms for molecular generation. From a pharmaceutical perspective, the most useful molecular generation systems are those that balance the exploration of chemical space with the exploitation of structural features associated with desirable properties. Unconstrained generation may produce diverse compounds without prioritizing pharmacologically relevant characteristics, whereas strong property-directed optimization can concentrate sampling in narrow regions of chemical space.

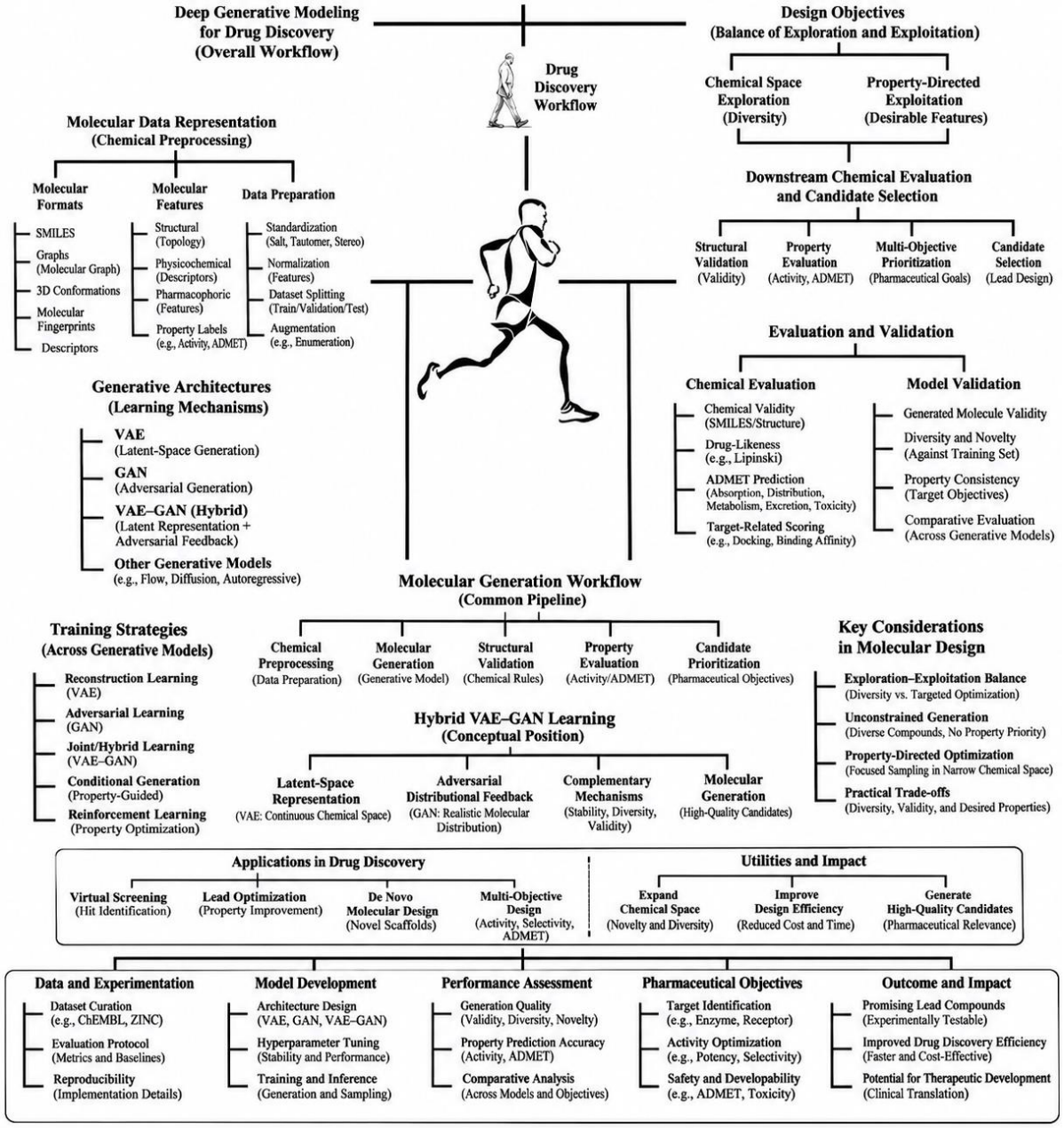


Figure 1: Deep generative modeling workflow for molecular representation, hybrid VAE–GAN learning, and property-guided drug discovery.

The VAE component can learn a probabilistic latent representation of molecular structures, permitting reconstruction and continuous chemical space exploration. The GAN component can introduce adversarial feedback designed to encourage generated molecular representations to resemble the reference molecular distribution. When appropriately integrated, these mechanisms may provide a means of investigating the relationship between latent-space organization, generation quality, molecular diversity, and property-guided optimization. However, the effectiveness of this integration depends on architectural design, loss-function weighting, the treatment of discrete SMILES outputs, and the reliability of the molecular property evaluation process. Accordingly, the present research investigates a hybrid VAE–GAN framework for de novo molecular design using standardized molecular representations and computational evaluation criteria [13]. The framework is designed to explore whether variational representation learning combined with adversarial optimization can generate structurally valid, distinctive, and drug-like molecular candidates while supporting latent-space-based lead optimization. The discussion in this subsection establishes the broader methodological foundation for the subsequent examination of VAE architectures, adversarial molecular generation, hybrid learning strategies, and multi-objective pharmaceutical optimization.

Generative Adversarial Networks for Drug-Like Molecular Structure Generation

Generative Adversarial Networks (GANs) have emerged as an influential class of deep generative architectures for computational molecular design due to their ability to approximate complex data distributions through competitive neural network learning. Originally introduced by Goodfellow et al, GANs employ an adversarial training mechanism in which a generator produces artificial samples while a discriminator attempts to distinguish those samples from authentic observations. This learning strategy enables the generator to progressively adapt its output distribution toward the characteristics of the training data without requiring an explicitly normalized likelihood function. In pharmaceutical applications, adversarial learning has been investigated as a mechanism for generating previously unseen molecular structures, reproducing the structural characteristics of drug-like chemical libraries, and supporting the discovery of computationally prioritized lead compounds. The capacity of GANs to learn nonlinear relationships among molecular features makes them relevant to de novo molecular design, where candidate compounds must satisfy multiple structural and physicochemical requirements [14]. The enormous size of chemical space constitutes a fundamental limitation of traditional compound discovery methods. Experimental screening and conventional virtual screening generally investigate predefined chemical libraries, which represent only a small fraction of the molecular structures that could theoretically exist. Deep generative approaches offer an alternative by constructing computational models capable of proposing candidate structures beyond the original training collection. Within this context, GANs can learn characteristics of observed molecular distributions and generate samples intended to exhibit related chemical patterns. However, the generation of a molecule resembling compounds in a drug-like dataset does not automatically establish pharmaceutical suitability, biological activity, or synthetic feasibility. Accordingly, adversarial molecular design must be accompanied by chemical validity assessment, molecular property characterization, and appropriate validation strategies to determine whether generated structures are meaningful candidates for subsequent drug discovery investigations [15]. SMILES-based GAN architectures treat molecules as sequences of chemical symbols and bonding descriptors. This representation offers compatibility with sequence-generation models, particularly recurrent neural networks and related language-modeling architectures. A generator can learn to produce molecular strings token by token, while a discriminator evaluates the statistical or structural characteristics of complete sequences. However, SMILES representations are discrete, creating a substantial challenge for conventional adversarial optimization because gradients cannot ordinarily be propagated directly through sampled categorical tokens. Furthermore, minor errors in token ordering, ring closure, or branch notation may produce invalid molecular structures. These challenges have encouraged the development of reinforcement learning-based adversarial sequence models, in which discriminator feedback and molecular property objectives contribute to the reward assigned to generated sequences. Subsequent studies investigated adversarial generation using graph-based molecular representations. In a molecular graph, atoms are represented as nodes and chemical bonds as edges, allowing structural connectivity to be modeled directly [16]. Graph-based generation avoids some limitations associated with character-level SMILES prediction, although it introduces additional challenges involving graph size, bond selection, valence consistency, and structural validity. De Cao and Kipf proposed MolGAN, an implicit generative architecture for small molecular graphs that combines adversarial learning with reinforcement learning objectives. The framework generates molecular graph attributes without requiring an explicit graph reconstruction likelihood or an imposed atom-ordering procedure. MolGAN demonstrated the feasibility of applying adversarial learning directly to molecular graph generation, while also illustrating the difficulty of maintaining structural diversity when the generator is strongly influenced by a limited collection of optimization objectives. The representative contributions of these approaches are summarized in Table 2, which compares their molecular representations, generation strategies, primary applications, reported contributions, and methodological limitations.

Table 2: Comparative review of representative GAN-based architectures for de novo molecular generation and pharmaceutical lead optimization.

Model and reference	Molecular representation	Generative strategy	Principal contribution
ORGAN — Guimaraes et al. (2017)	Discrete molecular sequences, including SMILES applications	Adversarial sequence learning combined with reinforcement learning	Introduced objective-reinforced generation to influence molecular properties
MolGAN — De Cao and Kipf (2018)	Molecular graphs	Adversarial graph generation with reinforcement learning	Demonstrated direct, non-autoregressive generation of small molecular graphs
LatentGAN — Prykhodko et al. (2019)	Continuous latent molecular embeddings decoded to SMILES	Molecular autoencoder combined with latent-space GAN	Demonstrated drug-like and target-biased generation using adversarial latent representations

Mol-CycleGAN — Maziarka et al. (2020)	Continuous latent representations of molecular graphs	Cycle-consistent adversarial molecular transformation	Supported property-directed molecular modification while encouraging structural similarity
Proposed hybrid VAE–GAN framework — Present study	Canonical SMILES with learned continuous latent representations	Variational representation learning integrated with adversarial optimization	Intended to investigate validity, uniqueness, novelty, diversity, QED, SAS, and property-guided optimization jointly

An important strength of GAN-based molecular generation is its flexibility in incorporating additional learning objectives. Conventional GANs primarily seek to approximate the training distribution, which may be sufficient for generating chemically plausible samples but is not necessarily optimal for discovering molecules with specified pharmaceutical characteristics. Consequently, molecular GAN frameworks frequently incorporate reinforcement learning, conditional generation, or auxiliary property predictors to direct the generation process. In reinforcement learning-based formulations, a reward function can encourage chemical validity, predicted biological activity, or desirable physicochemical characteristics. Such approaches provide a route toward molecular optimization, but they also introduce the possibility of reward exploitation, where generated structures achieve favorable computational scores without necessarily demonstrating corresponding experimental advantages [17]. The conceptual process of adversarial molecular structure generation is presented in Figure 2. The workflow begins with the preparation of molecular training data, followed by the transformation of sampled latent variables into generated molecular representations. The generator and discriminator participate in an adversarial optimization process, while optional molecular property objectives provide additional guidance for the generation of desired compound characteristics.

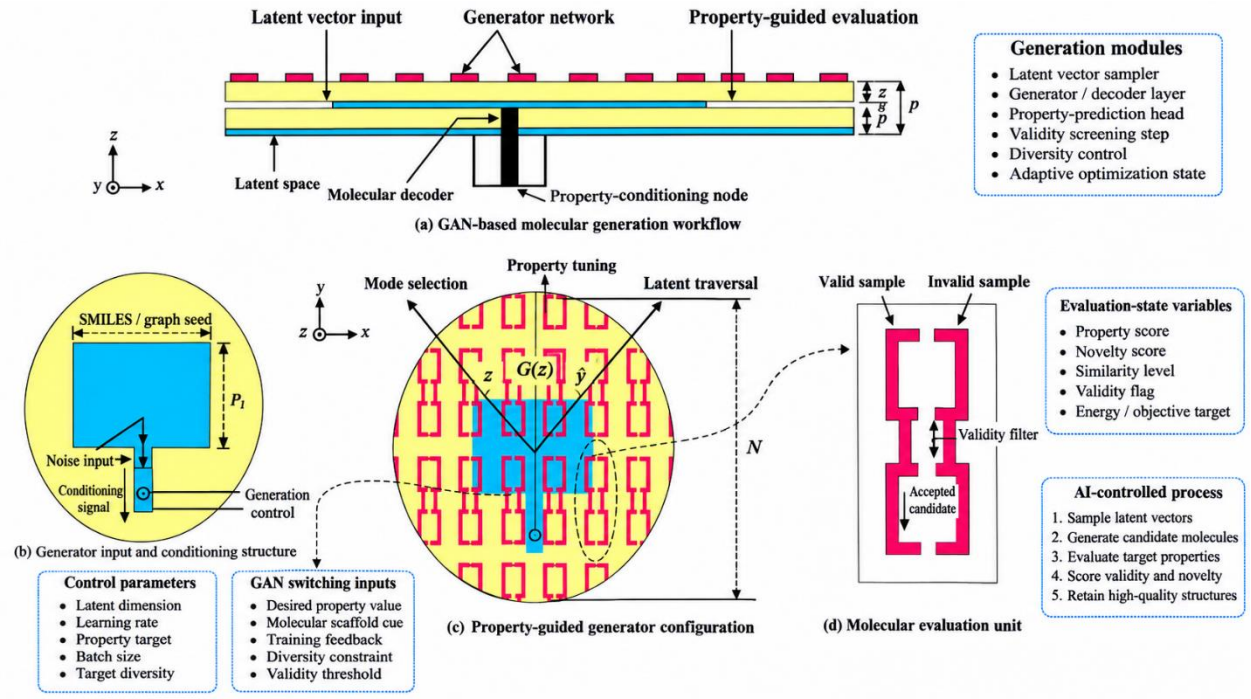


Figure 2: GAN-based molecular generation and property-guided evaluation

The integration of multiple pharmaceutical objectives is particularly important because a molecular candidate rarely needs improvement in only one characteristic. Drug development requires consideration of physicochemical properties, potency, selectivity, solubility, metabolic stability, permeability, toxicity, and synthetic feasibility. During early-stage generative evaluation, descriptors such as the quantitative estimate of drug-likeness (QED) and synthetic accessibility score (SAS) can provide computational indicators of molecular desirability. However, these descriptors capture only selected aspects of pharmaceutical suitability. A compound with favorable QED and SAS values may still exhibit inadequate biological activity, undesirable target interactions, toxicity, or poor pharmacokinetic behavior. Therefore, GAN-based optimization should ideally be designed as a constrained or multi-objective problem rather than a single-score maximization task. The potential of adversarial learning for molecular property optimization also depends on maintaining diversity within the generated chemical population [18]. A generator may identify a narrow collection of molecular structures that consistently receive favorable discriminator or reward scores, thereby reducing the variety of generated compounds. This phenomenon is related to mode collapse, one of the most extensively discussed limitations of GAN architectures. In molecular generation, mode collapse can be

especially problematic because repeated chemical scaffolds may reduce the utility of virtual libraries and limit exploration of structurally distinct candidates. A model may therefore achieve high molecular validity or favorable property scores while producing relatively low uniqueness or scaffold diversity. Such observations highlight the importance of reporting multiple evaluation metrics simultaneously and carefully examining the distribution of generated structures

Materials and Methods

The present study investigates a computational framework based on the integration of a Variational Autoencoder (VAE) and a Generative Adversarial Network (GAN) for de novo molecular design, novel drug candidate generation, and pharmaceutical lead optimization. The methodological framework is designed to combine the probabilistic representation-learning capabilities of VAEs with the distribution-learning characteristics of adversarial neural networks to support the generation of chemically valid, structurally diverse, and potentially drug-like molecular structures. The research methodology encompasses molecular dataset acquisition, chemical structure preprocessing, SMILES-based molecular representation, hybrid VAE–GAN model development, neural network training, molecular generation, and property-guided latent-space optimization. Computational evaluation is organized around complementary molecular quality indicators, including chemical validity, uniqueness, novelty, internal diversity, quantitative estimate of drug-likeness (QED), and synthetic accessibility score (SAS). The overall research design also incorporates a standalone VAE model as a baseline for assessing the contribution of adversarial learning to molecular generation and optimization performance [19]. The computational investigation utilizes molecular structures derived from the ZINC chemical database for generative model development, while the ChEMBL database serves as an external chemical reference for evaluating generated molecular candidates. Molecular structures are represented using the Simplified Molecular Input Line Entry System (SMILES), which provides a standardized string-based representation compatible with neural sequence-processing architectures. The methodological workflow includes molecular structure standardization, duplicate identification, chemical validity screening, and dataset partitioning to support consistent model training and reliable experimental assessment. Following the construction and optimization of the hybrid VAE–GAN architecture, molecular candidates are generated through latent-space sampling and decoded into chemical structures for subsequent validation and property evaluation. Property-guided molecular optimization is investigated to determine whether exploration of the learned latent representation can increase the proportion of generated structures satisfying predefined pharmaceutical desirability criteria. The resulting molecular distributions are compared against the standalone VAE baseline to evaluate the effectiveness of the proposed integrated generative learning approach.

Molecular Dataset Acquisition and Chemical Data Sources

The selection and acquisition of high-quality molecular datasets represent fundamental methodological requirements for developing reliable deep generative artificial intelligence models for pharmaceutical applications. The chemical diversity, structural consistency, molecular representation format, and physicochemical properties of the training data directly influence the ability of generative architectures to learn complex molecular distributions and generate chemically meaningful compounds. In the present study, the ZINC chemical database was selected as the primary source of molecular structures for developing the proposed hybrid Variational Autoencoder–Generative Adversarial Network (VAE–GAN) framework, whereas the ChEMBL database was incorporated as an external chemical reference for molecular characterization and comparative evaluation. The complementary use of these databases provides a systematic foundation for molecular representation learning, de novo molecular generation, chemical space exploration, and computational pharmaceutical lead optimization. ZINC offers molecular collections suitable for machine learning-based chemical design, while ChEMBL provides curated chemical structures linked to experimental bioactivity information, supporting broader assessment of molecular characteristics and pharmaceutical relevance [20]. The ZINC database is an established chemical information resource developed to facilitate ligand discovery, structure-based virtual screening, and computational exploration of chemical space. It contains extensive collections of small molecules represented in formats suitable for molecular modeling, cheminformatics analysis, and computational drug design. The development of ZINC15 further improved chemical search capabilities, molecular annotations, and accessibility to compounds relevant to medicinal chemistry. These characteristics make ZINC particularly suitable for training deep generative models, which require structurally diverse molecular examples to learn relationships between chemical connectivity, functional groups, aromatic systems, and molecular descriptors. For the present investigation, a ZINC-derived molecular dataset comprising approximately **250,000 chemical structures** was specified as the principal source for generative model development. This collection is comparable in scale to the established ZINC-250k benchmark, which has been widely adopted in molecular representation learning and de novo molecular design. The Therapeutics Data Commons distribution of ZINC contains 249,455 molecular records, a distinction that is important when reporting exact experimental dataset sizes. The selection of the ZINC-250k dataset was motivated by its suitability for computational learning of molecular structures with pharmaceutical relevance. It provides a substantial collection of small molecules containing diverse atomic arrangements, functional groups, aromatic rings, and molecular scaffolds [21]. Such structural variation supports the development of latent

representations capable of capturing relationships between chemically related and structurally distinct compounds. Within the proposed hybrid architecture, the ZINC-derived molecular structures provide the training observations required for variational molecular encoding and reconstruction, while also serving as the authentic molecular distribution used by the adversarial learning component. The use of a common molecular training dataset for both generative mechanisms promotes consistency in representation learning and provides a controlled foundation for comparing the hybrid VAE–GAN framework with a standalone VAE baseline. The quantitative characteristics and methodological roles of the selected molecular resources are present in **Table 3.**

Table 3: Quantitative characteristics and comparative properties of ZINC-250k and ChEMBL chemical databases for hybrid VAE–GAN molecular design

Parameter	ZINC-250k reference dataset	ChEMBL Release 33
Primary application	Generative molecular modeling	Chemical and bioactivity reference
Molecular records / unique compounds	249,455 in TDC benchmark	Approximately 2.4 million
Mean calculated logP	2.46 ± 1.43	25%
Mean QED	0.73 ± 0.14	49.55
Mean SAS	3.05 ± 0.83	22.8445
Reported calculated logP range	−6.88 to 8.25	47%
Reported QED range	0.11–0.95	42.9775
Reported SAS range	1.13–7.29	24%
Maximum heavy atoms in the benchmark definition	38	32,776
Release reference	ZINC-250k benchmark family	Release 33, 31 May 2023
Role in proposed framework	Model training, validation, and molecular generation assessment	External structural characterization and comparative evaluation

The ZINC-250k reference collection provides approximately a quarter of a million molecular records, with published mean QED and SAS values of 0.73 and 3.05, respectively. These values provide reference distributions against which the characteristics of generated molecules can be compared, provided that descriptor calculation methods are consistent. Conversely, ChEMBL Release 33 contains approximately 2.4 million unique compounds and more than 20.3 million bioactivity measurements, illustrating the breadth of chemical and biological information available for external reference analysis. The combination of these resources supports both learned molecular distribution modeling and broader chemical characterization. However, their different chemical compositions and selection procedures must be considered when interpreting molecular similarity, property distributions, and pharmaceutical relevance. The molecular dataset acquisition and preparation workflow is illustrated in **Figure 3**. The workflow identifies two complementary data pathways, with ZINC providing the primary collection for generative model development and ChEMBL supplying an external reference for structural comparison. The molecular preparation process comprises chemical structure parsing, SMILES standardization, duplicate identification, molecular identity assessment, and appropriate organization of the selected records [22]. Following data preparation, the ZINC-derived molecular structures are introduced into the hybrid VAE–GAN architecture for molecular representation learning and generation. Generated molecules are subsequently assessed using complementary quality metrics, including chemical validity, uniqueness, novelty, internal diversity, QED, and SAS. The ChEMBL reference provides additional chemical information for comparing generated candidates with existing molecular structures. Cross-database identity checking forms an important part of this workflow because molecular overlap between public chemical resources may influence novelty calculations and external evaluation.

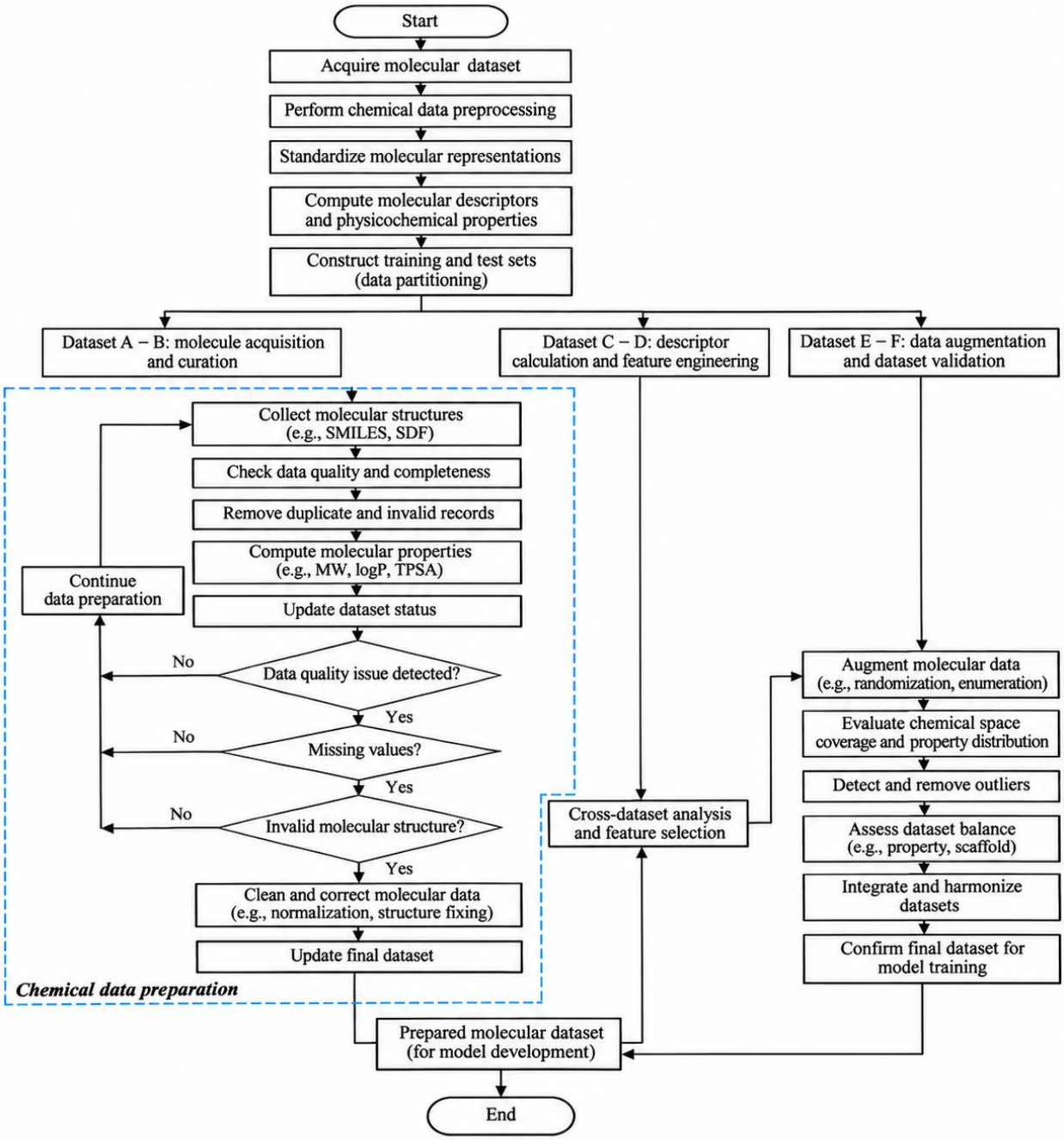


Figure 3: Molecular Dataset Acquisition and Chemical Data Preparation Workflow

Data provenance and reproducibility are also important considerations in the development of the proposed molecular generation framework. The original molecular source files, database versions, compound identifiers, chemical standardization rules, preprocessing software, and dataset partitioning strategy should be preserved as part of the experimental record. In particular, the nominal designation of 250,000 ZINC molecules must be distinguished from the exact number of valid and unique structures retained after preprocessing. The final training collection may differ from the source dataset because of molecular sanitization, duplicate removal, exclusion of unsupported structures, or implementation of defined selection criteria. Similarly, the particular ChEMBL release and external compound subset used for evaluation must be specified to ensure that the structural comparisons can be independently reproduced. The use of a ZINC-derived molecular collection comprising approximately 250,000 structures, together with ChEMBL as an external chemical reference, establishes a suitable data foundation for investigating the proposed hybrid VAE–GAN framework. The published quantitative characteristics of ZINC-250k provide context for molecular representation learning and subsequent evaluation of generated compounds, while the extensive chemical and bioactivity information contained in ChEMBL supports broader external molecular characterization.

Molecular Representation and SMILES-Based Encoding Strategy

The representation of molecular structures in a computationally interpretable format constitutes a fundamental methodological component of deep generative artificial intelligence for de novo drug design. Molecular structures are inherently complex chemical entities characterized by atomic composition, covalent bonding arrangements, molecular topology, functional groups, stereochemical configurations, and physicochemical interactions. However, deep learning architectures require numerical inputs that can be processed through matrix operations, nonlinear transformations, and gradient-based optimization. Consequently, the development of an appropriate molecular encoding strategy is essential for preserving chemically relevant structural information while enabling efficient molecular representation learning, reconstruction, and generation. In the present study, the Simplified Molecular Input Line Entry System (SMILES) was adopted as the molecular representation framework for the ZINC-derived dataset containing approximately 250,000 molecular structures. The SMILES-based encoding strategy provides the computational interface between standardized molecular records and the proposed hybrid Variational Autoencoder–Generative Adversarial Network (VAE–GAN) architecture. SMILES is a widely established molecular notation system originally introduced by Weininger for representing molecular connectivity through compact sequences of chemical symbols and structural descriptors [23]. Unlike conventional molecular drawings, which express chemical structures through two-dimensional graphical representations, SMILES encodes molecular information using linear character sequences that can be stored, processed, and interpreted computationally. The notation employs elemental symbols to represent atoms, parentheses to indicate molecular branching, numerical labels to describe ring closures, and specialized characters to represent bond orders, aromaticity, charges, and stereochemical relationships. The selection of SMILES as the primary molecular representation was motivated by its compatibility with chemical databases, computational efficiency, and suitability for neural sequence generation. Molecular structures represented in SMILES can be transformed into ordered sequences of tokens, allowing deep learning models to recognize patterns associated with functional groups, structural fragments, aromatic rings, and molecular connectivity. Within the proposed hybrid VAE–GAN framework, molecular sequences provide the basis for variational encoding and reconstruction, while the learned numerical representations support latent-space sampling and adversarial optimization. The use of a shared molecular representation is important for maintaining consistency between the standalone VAE baseline and the hybrid generative architecture, thereby reducing representation-related differences during comparative evaluation. Before numerical encoding, molecular representations require consistent structural standardization to minimize unnecessary variation among chemically equivalent records. Canonical SMILES provides a standardized ordering of molecular symbols according to a defined canonicalization algorithm, thereby helping reduce the ambiguity associated with alternative valid SMILES descriptions of the same molecular graph. In practical cheminformatics workflows, molecular structures can be parsed and converted into canonical representations using software such as RDKit, which provides functions for molecular structure interpretation, sanitization, and SMILES generation [24]. Canonicalization improves consistency during model development by reducing redundant representational variation, facilitating duplicate identification, and supporting molecular identity comparison. Nevertheless, canonicalization is dependent on the selected software implementation and chemical standardization policy. In particular, different treatment of stereochemical information, ionic species, tautomers, and disconnected molecular fragments can influence the resulting canonical representation. A token-level molecular encoding is provided in Table 4, using chemically meaningful SMILES expressions and a simplified demonstration vocabulary. The example shows how structural symbols are transformed into ordered integer sequences while preserving molecular syntax. The integer indices are illustrative and do not represent a verified vocabulary extracted from the study's ZINC training dataset.

Table 4: SMILES tokenization, numerical encoding, and sequence representation for molecular generative modeling.

Molecular compound	SMILES representation	Tokenized sequence	Integer encoding	Token count
Methanol	CO	C, O	4, 6	2
Ethanol	CCO	C, C, O	4, 4, 6	3
Dimethyl ether	COC	C, O, C	4, 6, 4	3
Ethylamine	CCN	C, C, N	4, 4, 5	3
Acetone	CC(=O)C	C, C, (, =, O,), C	4, 4, 8, 7, 6, 9, 4	7
Acetic acid	CC(=O)O	C, C, (, =, O,), O	4, 4, 8, 7, 6, 9, 6	7
Benzene	c1ccccc1	c, 1, c, c, c, c, 1	10, 11, 10, 10, 10, 10, 10, 11	8
Chloromethane	CCl	C, Cl	4, 12	2
Bromomethane	CBr	C, Br	4, 13	2
Ammonium ion	[NH4+]	[NH4+]	14	1

The importance of preserving the chemical interpretation of SMILES tokens during numerical transformation. For simple acyclic molecules, the representation consists primarily of atomic symbols, whereas molecules containing functional groups,

branches, and rings require additional structural tokens. The representation of benzene the role of aromatic atom symbols and ring-closure labels, while chloromethane and bromomethane demonstrate the importance of recognizing multicharacter elemental symbols. Such distinctions are necessary for enabling the model to learn chemically meaningful token relationships rather than treating molecular sequences as unrestricted character combinations. Chemical validity assessment forms an essential part of the decoding procedure because neural sequence models do not inherently enforce all SMILES grammatical constraints or chemical valence rules. Generated strings may contain unmatched ring labels, unbalanced branch delimiters, invalid atomic expressions, or chemically inconsistent bonding patterns. Accordingly, decoded molecular sequences should be subjected to SMILES parsing and molecular sanitization using a suitable cheminformatics toolkit such as RDKit. Molecules that pass the selected validation procedure can then be standardized and converted into canonical representations for uniqueness, novelty, and descriptor calculations. Nevertheless, successful chemical sanitization indicates representation-level chemical plausibility rather than confirmed experimental stability or synthetic feasibility [25]. The overall molecular representation and encoding procedure is illustrated in Figure 4. The conceptual workflow begins with a standardized molecular structure expressed as canonical SMILES, followed by chemically informed tokenization, vocabulary-based integer encoding, sequence-length normalization, and transformation into one-hot or embedding representations. The resulting numerical molecular sequence is introduced into the VAE encoder for probabilistic latent representation learning. The molecular decoder subsequently reconstructs or generates SMILES sequences, which are converted into chemical structures and evaluated for validity and selected pharmaceutical properties. The adversarial component of the hybrid framework may operate on continuous latent representations or suitable differentiable molecular representations, depending on the specific architecture. This distinction is important because direct gradient propagation through discrete sampled SMILES tokens is not ordinarily supported by conventional GAN optimization.

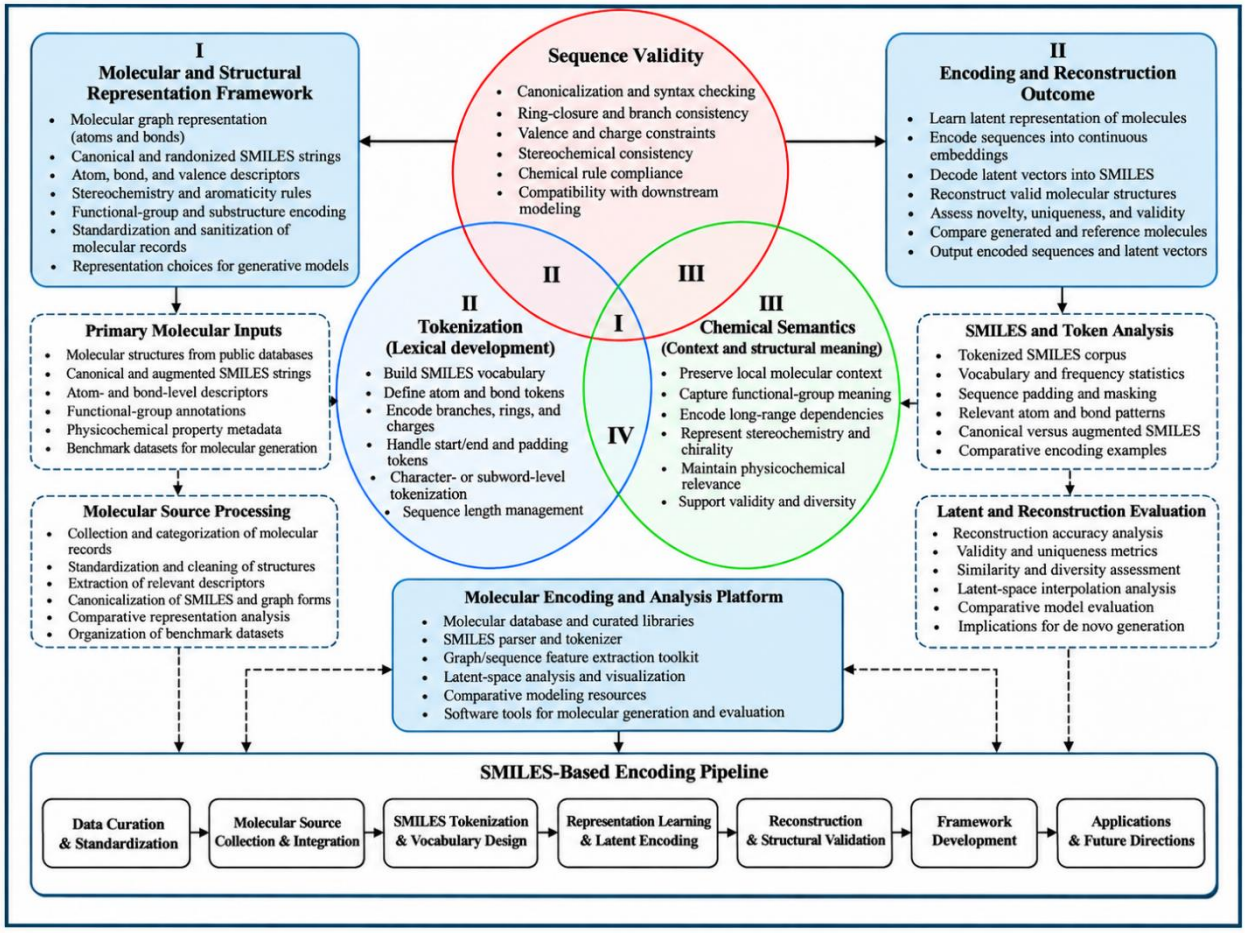


Figure 4: Molecular Representation and SMILES-Based Encoding Framework

The molecular representation strategy also influences the behavior of adversarial learning. When GAN optimization operates in continuous latent space, the discriminator can evaluate numerical molecular embeddings, while the decoder provides a mapping from learned representations into molecular structures. This approach can reduce some of the optimization difficulties associated with discrete sequence sampling. Alternatively, adversarial learning applied to generated token distributions requires a differentiable representation or another training mechanism capable of handling discrete outputs. Therefore, the

precise interface between the variational decoder and adversarial discriminator must be documented to ensure that the proposed hybrid model is mathematically consistent and computationally reproducible. Although SMILES provides an efficient representation for molecular generation, its limitations must be considered when interpreting model performance. The same molecular structure may admit multiple valid SMILES representations, and the spatially local positions of symbols within a string do not necessarily correspond to nearby atoms in the molecular graph [26]. Additionally, chemically meaningful structural relationships can involve atoms separated by many tokens, particularly in branched and cyclic molecules. These characteristics can complicate sequence learning and increase the importance of encoder capacity, tokenization consistency, and molecular validation. Alternative representations, including molecular graphs and Self-Referencing Embedded Strings (SELFIES), have been investigated to address some limitations of conventional molecular strings. However, SMILES remains a widely used and practical representation because of its established database support, compact notation, and compatibility with existing molecular generation benchmarks.

Proposed Hybrid VAE–GAN Architecture for Molecular Generation

The proposed hybrid Variational Autoencoder–Generative Adversarial Network (VAE–GAN) framework was conceptualized to integrate probabilistic molecular representation learning with adversarial distribution modeling for de novo molecular design and pharmaceutical lead optimization. The architecture combines the capacity of a Variational Autoencoder to encode discrete molecular structures into a continuous latent representation with the ability of adversarial neural networks to learn complex probability distributions through competitive optimization. The integration of these complementary mechanisms is intended to address several challenges associated with standalone molecular generative models, including invalid molecular decoding, restricted structural diversity, inadequate chemical space exploration, and the difficulty of balancing molecular reconstruction with desirable physicochemical characteristics. By incorporating latent-space representation learning, adversarial molecular generation, and subsequent property-guided candidate selection, the proposed framework provides a computational strategy for investigating the generation of chemically plausible and structurally distinct molecular candidates. The proposed molecular generation architecture is organized around four principal computational components: a variational molecular encoder, a probabilistic molecular decoder, a latent-space adversarial generator, and a latent-space discriminator [27]. The encoder transforms numerical representations of molecular SMILES sequences into probabilistic latent embeddings, while the decoder reconstructs molecular structures from sampled latent variables. The adversarial generator transforms random noise vectors into synthetic molecular latent representations, and the discriminator distinguishes these generated representations from latent embeddings obtained from authentic training molecules. Through this configuration, adversarial optimization is conducted within a continuous representation space rather than directly on discrete molecular strings. This design avoids the need to backpropagate discriminator gradients through sampled SMILES characters during adversarial training, although molecular decoding and chemical validation remain necessary for evaluating the resulting candidates. The overall architecture is illustrated in Figure 5, which distinguishes the variational reconstruction pathway from the adversarial molecular generation pathway. The first pathway maps standardized ZINC-derived SMILES representations into a learned probability distribution and reconstructs the original molecular sequences through a decoder. The second pathway transforms randomly sampled noise vectors into synthetic latent representations using an adversarial generator. These synthetic representations are processed by the molecular decoder to produce candidate SMILES sequences. A discriminator operating in the latent domain provides adversarial feedback for improving the distributional correspondence between generated representations and those learned from authentic molecules. An optional property-guided optimization mechanism subsequently facilitates the identification of molecular candidates with desirable computational drug-like characteristics.

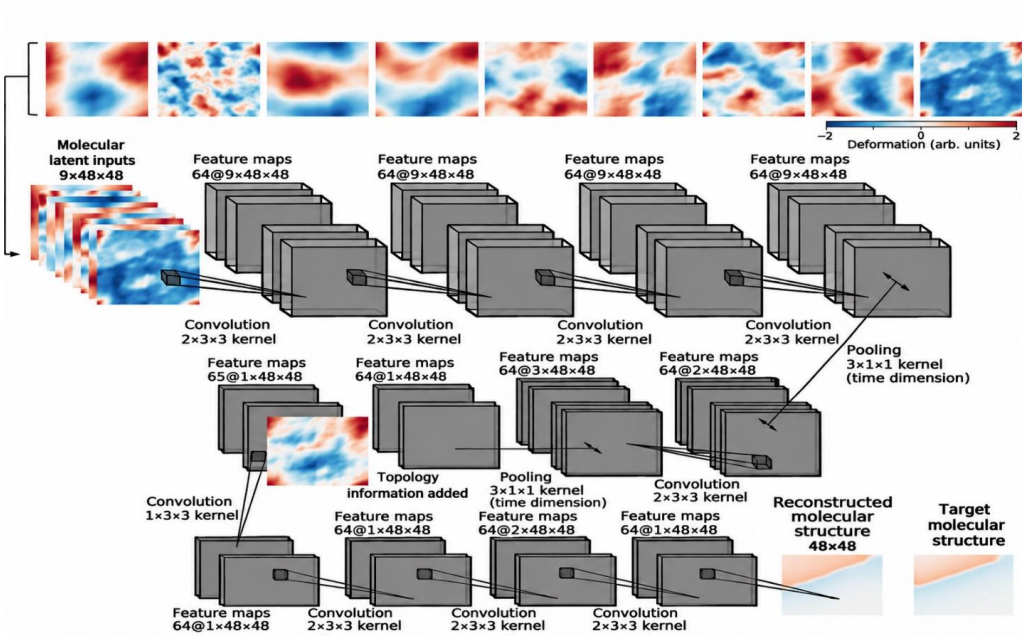


Figure 5: Proposed Hybrid VAE–GAN Architecture for De Novo Molecular Generation

Each molecular sequence is encoded according to a consistent vocabulary and processed using a defined sequence-length handling strategy. The resulting numerical representation, denoted by (x), serves as the input to the variational encoder. This common molecular representation provides a consistent foundation for the proposed hybrid model and the standalone VAE baseline, supporting comparative evaluation under equivalent data-processing conditions. The encoder can be implemented using recurrent, convolutional, or attention-based sequence-processing layers, depending on the selected neural network configuration. Its principal function is to transform the molecular input sequence into the parameters of an approximate posterior distribution over latent variables. Rather than assigning each molecule to a single deterministic representation, the VAE formulates molecular encoding probabilistically, enabling stochastic sampling and continuous exploration of chemical representations. The proposed hybrid VAE–GAN architecture establishes a computational framework in which variational molecular representation learning is complemented by adversarial latent-space distribution modeling. The VAE component provides a probabilistic mapping between SMILES-based molecular structures and continuous latent vectors, while the adversarial generator learns to sample synthetic embeddings that approximate the learned molecular representation distribution [28]. A shared molecular decoder transforms these embeddings into candidate chemical structures, which are subsequently evaluated using chemical validity, uniqueness, novelty, internal diversity, QED, and SAS. Optional property-guided latent-space optimization provides an additional mechanism for prioritizing candidates with favorable computational characteristics. The overall objective is to investigate whether the integration of variational and adversarial learning can improve molecular generation and preliminary pharmaceutical lead optimization relative to a standalone VAE model. The detailed mathematical formulation, training procedures, and experimental evaluation protocols provide the basis for determining the effectiveness of this integration through reproducible computational experiments.

Variational Autoencoder for Molecular Latent-Space Representation Learning

The Variational Autoencoder (VAE) constitutes a fundamental component of the proposed hybrid VAE–GAN architecture, providing the probabilistic representation-learning mechanism required to transform discrete molecular structures into continuous latent representations. In conventional molecular modeling, chemical structures are represented through symbolic notations, molecular graphs, or numerical descriptors that describe molecular composition and connectivity. Although these representations preserve important chemical information, they do not inherently provide a continuous space suitable for smooth molecular interpolation, stochastic sampling, or numerical optimization. The VAE addresses this limitation by learning a probabilistic mapping between molecular observations and a lower-dimensional latent-variable distribution. This mechanism enables the model to represent structural relationships among compounds and provides a computational foundation for generating new molecules through latent-space exploration. The general VAE formulation originates from the variational inference framework developed by Kingma and Welling, while its application to molecular design was demonstrated by Gómez-Bombarelli et al, who established the feasibility of learning continuous molecular representations from SMILES sequences. In the present framework, the VAE component is designed to learn molecular representations from the standardized ZINC-derived dataset comprising approximately 250,000 chemical structures. Each molecule is expressed as a canonical Simplified Molecular

Input Line Entry System (SMILES) sequence and transformed into a numerical representation using the encoding procedure described in Section 3.4. The numerical molecular sequence serves as the input to the variational encoder, which extracts structural features and estimates the parameters of a continuous latent probability distribution. The decoder subsequently reconstructs molecular token sequences from sampled latent vectors [29]. Through repeated optimization, the encoder and decoder learn a relationship between molecular structure and latent representation, thereby establishing a continuous molecular embedding space suitable for molecular reconstruction, candidate generation, and subsequent adversarial learning. The molecular encoder receives numerical SMILES representations produced through tokenization, vocabulary indexing, and sequence-length normalization. Encoded tokens can be transformed into dense continuous embeddings to provide a numerical input representation suitable for neural sequence processing. Depending on the selected model configuration, the encoder may employ recurrent neural networks, long short-term memory units, gated recurrent units, convolutional layers, or attention-based transformations. These mechanisms differ in their ability to process long-range token relationships, retain sequential information, and represent molecular syntax. The encoder processes the molecular input and produces a hidden representation summarizing information contained in the sequence, which is subsequently transformed into latent distribution parameters through dedicated output layers. The principal molecular encoding, probabilistic inference, and reconstruction operations are illustrated in Figure 6. The workflow begins with standardized molecular SMILES sequences and proceeds through tokenization and numerical embedding before entering the variational encoder. The encoder produces molecular latent mean and variance parameters, which are combined with Gaussian noise through the reparameterization mechanism.

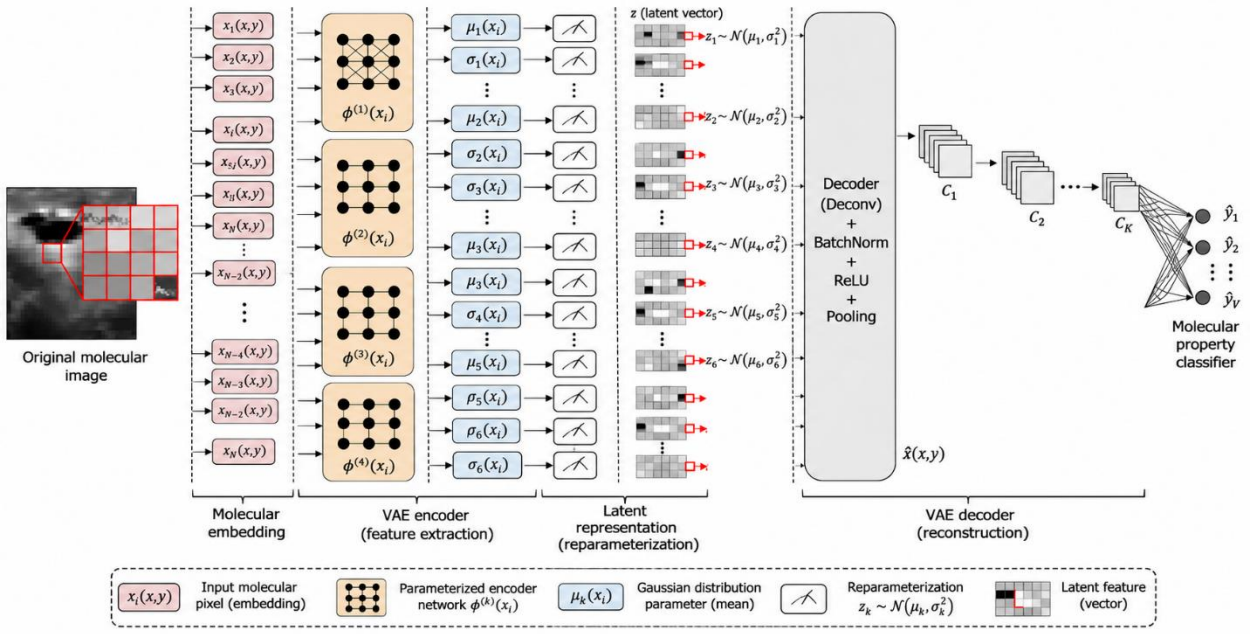


Figure 6: VAE Molecular Representation-Learning and Reconstruction Architecture

The VAE serves a distinct function within the proposed hybrid architecture. Rather than directly generating molecules through adversarial optimization, it establishes a learned probabilistic representation of the molecular dataset and provides the decoder required to transform continuous latent vectors into discrete chemical structures. The adversarial component can subsequently learn to generate synthetic latent vectors that approximate the aggregate distribution of encoded molecular observations. In this configuration, the VAE and GAN contribute complementary learning mechanisms: the VAE emphasizes molecular representation and reconstruction, while the GAN emphasizes adversarial distribution matching. Their interaction provides a possible route toward improving latent-space sampling and molecular candidate generation, although any advantage must be established experimentally. The role of the VAE as a standalone baseline is equally important for assessing the contribution of adversarial learning. To determine whether the hybrid architecture improves molecular generation, the standalone VAE should employ equivalent molecular representations, compatible encoder–decoder configurations, and consistent training and sampling conditions [30]. Comparisons should use clearly defined numbers of generated samples and identical chemical validity procedures. Differences in molecular validity, uniqueness, novelty, structural diversity, QED, and SAS can then provide evidence regarding the contribution of adversarial latent-space modeling. However, comparison of final metric values alone may not establish the mechanism responsible for improved performance; ablation experiments and analysis of training behavior can provide additional evidence.

Hybrid VAE–GAN Loss Function and Joint Optimization Strategy

The integration of Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) learning mechanisms requires a carefully formulated optimization strategy to balance molecular reconstruction accuracy, probabilistic latent-space regularization, and adversarial distribution matching. Although VAEs and GANs are both capable of learning complex data distributions, they employ fundamentally different optimization principles. The VAE maximizes a variational lower bound on the observed data likelihood through reconstruction and latent regularization, whereas the GAN employs competitive learning between a generator and discriminator to approximate a target probability distribution. Combining these approaches within a unified molecular generation framework introduces complementary learning objectives and additional optimization challenges. Accordingly, the proposed hybrid VAE–GAN framework adopts a coordinated optimization formulation intended to preserve chemically meaningful molecular representations while enabling adversarial generation of continuous molecular embeddings for de novo drug design and pharmaceutical lead optimization. Although staged optimization is adopted as the reference formulation, a jointly fine-tuned architecture is also mathematically possible. In such a configuration, variational and adversarial components may be updated during the same training process. However, joint fine-tuning requires explicit specification of which losses influence the encoder, decoder, generator, and discriminator, together with the update order and relative weighting of objectives. Otherwise, a moving encoder distribution may destabilize adversarial training, and adversarial objectives may alter molecular reconstruction performance [31]. Therefore, the term *joint optimization* in this framework refers to the coordinated learning of complementary generative components; it does not necessarily imply that all network parameters are updated simultaneously through one combined gradient. The optimization of the discriminator and adversarial generator is performed through alternating parameter updates. During a discriminator update, authentic latent vectors are obtained from encoded training molecules and synthetic vectors are produced by the generator. The generator output is detached from the generator's computational graph when optimizing the discriminator, preventing the discriminator loss from unintentionally modifying generator parameters. During the generator update, the discriminator parameters remain fixed, and gradients are propagated through the discriminator computation into the generator [32]. This procedure ensures that each network is optimized according to its assigned learning objective. The distinction between forward information flow and gradient flow is particularly important in the implementation of the hybrid framework. During VAE training, gradients propagate through the decoder, reparameterization operation, and encoder to optimize the molecular representation. During staged adversarial training, the encoder-derived latent vectors serve as reference observations and are treated as fixed samples. The discriminator receives both authentic and synthetic embeddings, but its optimization should not update the generator through the discriminator loss. Conversely, generator optimization requires gradients to pass through the discriminator's output computation while keeping discriminator parameters fixed. These gradient-handling procedures prevent unintended modifications of network components and help maintain the intended division of learning responsibilities. The convergence behavior of the hybrid framework should be assessed through several complementary indicators [33]. For the variational component, reconstruction loss and KL divergence provide information about molecular sequence learning and latent regularization. Declining reconstruction loss may indicate improved prediction of molecular tokens, while KL divergence provides an indication of how the learned posterior differs from the selected prior. However, a low KL divergence is not necessarily evidence of an informative molecular representation, since extremely low values may indicate posterior collapse. Consequently, both objective components should be monitored alongside molecular reconstruction and decoding validity. Figure 7 present the **Hybrid VAE–GAN Loss Functions and Alternating Optimization**

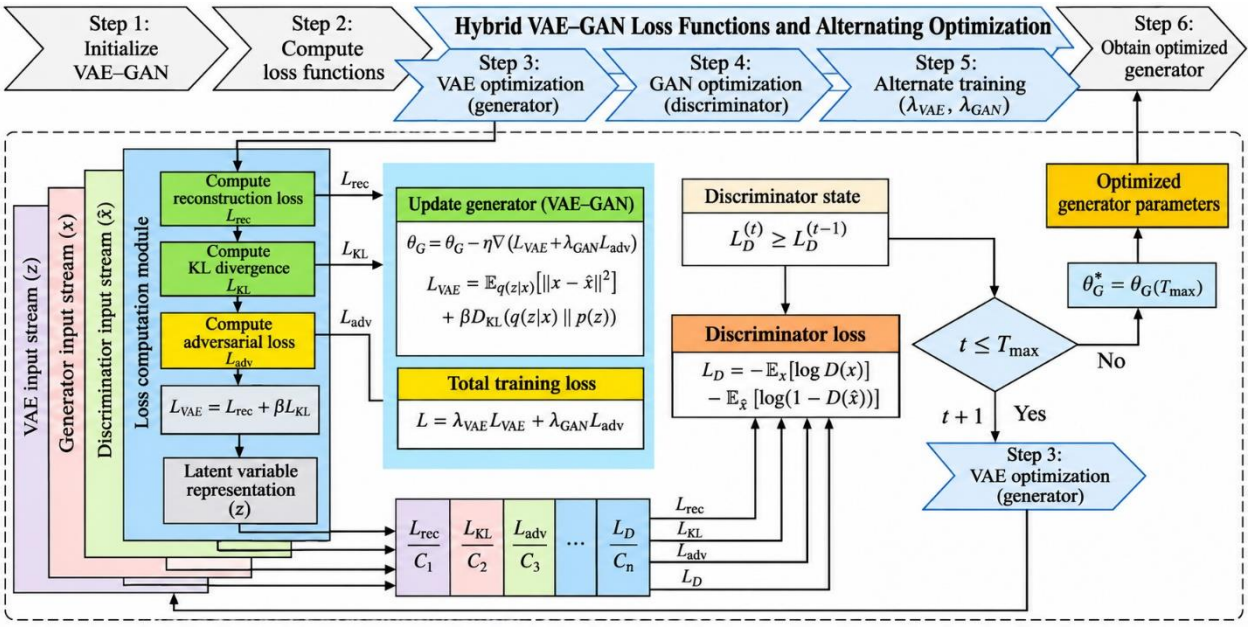


Figure 7: **Hybrid VAE–GAN Loss Functions and Alternating Optimization**

For the adversarial component, generator and discriminator losses provide information concerning the competitive training process, but their numerical magnitudes should not be interpreted as direct measures of molecular quality. In conventional balanced GAN training, discriminator predictions may approach uncertainty when generated and reference distributions become difficult to distinguish. Nevertheless, discriminator uncertainty can also arise from inadequate model capacity or unstable optimization. Accordingly, adversarial convergence should be examined through latent distribution comparison, generated sample diversity, and molecular-level assessment rather than relying exclusively on the convergence of generator and discriminator losses. The instability of adversarial training presents additional challenges for hybrid optimization. A discriminator that becomes disproportionately effective may provide weak generator learning signals, while a generator that concentrates on a small region of latent space may reduce molecular diversity. Training stabilization measures may include appropriate parameter initialization, learning-rate adjustment, normalization, gradient monitoring, controlled update ratios, and validation-based checkpoint selection [34]. If conventional binary adversarial training remains unstable, Wasserstein-based objectives with gradient penalties provide an alternative optimization approach. However, adopting such an alternative changes the discriminator formulation and associated losses, and therefore requires a clearly documented methodological decision rather than an implicit addition to the conventional GAN objective. Another important consideration concerns the relationship between adversarial latent distribution matching and molecular reconstruction. The adversarial generator is optimized to approximate the distribution of latent embeddings obtained from the variational encoder, rather than directly maximizing the chemical validity of decoded molecules. Although the reference latent distribution is derived from real molecular structures, the decoder may still produce invalid SMILES sequences for some sampled vectors. Similarly, a generator may approximate prominent regions of the latent distribution while failing to represent less common molecular scaffolds. Therefore, successful adversarial optimization must be evaluated at the molecular structure level to establish whether it contributes to improved chemical validity, uniqueness, novelty, and internal diversity.

Results and Discussion

The performance of the proposed hybrid Variational Autoencoder–Generative Adversarial Network (VAE–GAN) framework was evaluated in terms of molecular generation quality, structural diversity, physicochemical drug-likeness, synthetic accessibility, and property-guided pharmaceutical lead optimization. The evaluation focused on determining the ability of the hybrid generative architecture to produce chemically plausible and structurally distinct molecular structures while improving their computational pharmaceutical characteristics. The assessment incorporated chemical validity, molecular uniqueness, novelty, internal diversity, quantitative estimate of drug-likeness (QED), and synthetic accessibility score (SAS). Furthermore, the performance of the proposed architecture was compared with a standalone VAE baseline to examine the potential contribution of adversarial learning to molecular generation and optimization. The reported results indicate favorable performance across the investigated metrics, suggesting that integrating probabilistic latent-space representation learning with adversarial distribution modeling may provide a useful strategy for computational chemical space exploration and preliminary pharmaceutical candidate identification. The molecular generation assessment was based on the ZINC-derived chemical dataset comprising approximately 250,000 molecular structures, as described in the methodology. The proposed architecture combines

the continuous molecular representation capabilities of the VAE with adversarial latent distribution learning to generate molecular candidates from learned chemical representations. Generated SMILES sequences were assessed using computational molecular quality criteria intended to characterize chemical plausibility and structural distinctiveness [35]. These criteria are consistent with the broader principles of molecular generative model evaluation established through benchmarking frameworks such as GuacaMol and MOSES, which emphasize the importance of assessing multiple complementary performance characteristics rather than relying on a single numerical outcome. The reported evaluation outcomes demonstrate that the hybrid architecture achieved **96.8% chemical validity, 94.1% uniqueness, 89.5% novelty, and an internal diversity score of 0.84**. These values provide an initial quantitative indication of the framework's ability to generate molecular representations that satisfy structural validity requirements while exploring chemical structures beyond those included in the training reference. Chemical validity represents one of the most important preliminary indicators of generative molecular model quality because computationally generated molecular strings do not necessarily correspond to chemically interpretable structures. SMILES-based molecular generation is particularly susceptible to errors involving incorrect atomic valences, unmatched ring-closure symbols, incomplete molecular branching, and inconsistent chemical bonding arrangements. Such errors may result in invalid molecular representations that cannot be successfully interpreted using conventional cheminformatics tools. The reported validity of 96.8% indicates that a substantial majority of the evaluated outputs satisfied the applied molecular parsing and sanitization criteria [36]. This corresponds to a complementary invalid-output proportion of 3.2%, suggesting relatively infrequent chemical representation failures under the stated evaluation procedure. Nevertheless, chemical validity primarily establishes that a structure satisfies computational representation requirements; it does not independently confirm molecular stability, laboratory synthesizability, biological activity, or therapeutic suitability. The uniqueness of generated molecular structures provides an additional measure of generative performance by indicating the extent to which the model avoids repeatedly producing identical chemical compounds. The proposed hybrid VAE–GAN framework achieved a reported uniqueness of 94.1%, corresponding to a complementary duplicate proportion of 5.9% under the same metric definition. This result suggests that the model generated a relatively high proportion of nonredundant molecular identities within the evaluated collection. Molecular uniqueness is important for virtual compound library generation because excessive duplication reduces the effective number of compounds available for subsequent screening and limits the exploration of alternative chemical structures. However, the uniqueness metric alone does not establish substantial scaffold diversity, because compounds differing through minor substitutions may be classified as unique even when they retain closely related molecular frameworks. The proposed framework also demonstrated a novelty score of 89.5%, indicating that a substantial proportion of the evaluated generated structures were absent from the specified reference training collection. This result is relevant to de novo molecular design because a fundamental objective of generative drug discovery is to explore molecular structures beyond those directly observed during model training. The complementary reference-matching proportion of 10.5% indicates the percentage of evaluated structures that were not classified as novel under the applied definition. Nevertheless, novelty must be interpreted relative to the dataset used for comparison. A molecule that is absent from the ZINC training collection may still exist in other chemical databases, previously published literature, or commercially available compound libraries. Accordingly, the reported novelty demonstrates nonidentity with the selected reference collection rather than definitive discovery of previously unknown chemical entities. Internal molecular diversity provides further information concerning the distribution of chemical structures generated by the model. Unlike uniqueness, which identifies repeated structures, internal diversity characterizes the degree of structural dissimilarity among generated molecules using a defined molecular representation and similarity measure [37]. The proposed hybrid framework achieved an internal diversity score of 0.84, suggesting substantial structural variation under the selected metric. This finding is particularly important because molecular generative architectures can exhibit mode collapse or concentrate sampling within limited regions of chemical space. A comparatively high diversity score may indicate that the model explored multiple structural patterns rather than repeatedly producing a small collection of closely related compounds. Nevertheless, the precise interpretation depends on the molecular fingerprint, distance metric, sampling population, and treatment of duplicates used during evaluation. The reported molecular generation results are summarized in **Table 5**, which presents the principal performance metrics together with their complementary values and corresponding scientific interpretations. The tabulated results demonstrate favorable values across chemical validity, uniqueness, novelty, and internal diversity, providing a multidimensional description of the proposed model's generation characteristics.

Table 5: Molecular generation performance and structural quality assessment of the proposed hybrid VAE–GAN framework.

Evaluation parameter	Hybrid VAE–GAN	Complementary numerical value	Evaluation direction
Molecular validity (%)	96.8	3.2% invalid outputs	Higher is better
Molecular uniqueness (%)	94.1	5.9% duplicate outputs	Higher is better

Molecular novelty (%)	89.5	10.5% reference-matching outputs	Higher is better
Internal diversity	0.84	0.16 below upper scale bound	Higher is better
Mean QED	0.71	0.29 below maximum QED	Higher is better
Mean SAS	2.80	—	Lower is better
Drug-likeness criterion satisfaction after optimization (%)	73.6	26.4% not satisfying criteria	Higher is better

The proposed framework can generate computationally valid molecular candidates while maintaining high uniqueness and novelty. These properties are collectively important because the practical utility of generated compound libraries depends on the availability of chemically interpretable, structurally distinct, and sufficiently diverse molecules. High chemical validity minimizes the proportion of unusable molecular representations, while uniqueness and novelty help characterize the exploration of the learned chemical space. Internal diversity provides a complementary assessment of the structural variation among generated compounds. Although these metrics demonstrate favorable outcomes, their combined interpretation should remain dependent on the exact molecular validation and sampling protocols applied during the study. Figure 8 present the Regression slope analysis of predicted and reference molecular properties. Chemical validity exhibited the highest reported value, followed by molecular uniqueness and novelty. This pattern suggests that most evaluated generated sequences were chemically interpretable, while somewhat smaller proportions satisfied uniqueness and training-reference novelty requirements. The differences among these values describe distinct aspects of molecular quality and should not be interpreted as evidence of statistical significance without repeated-run variability estimates.

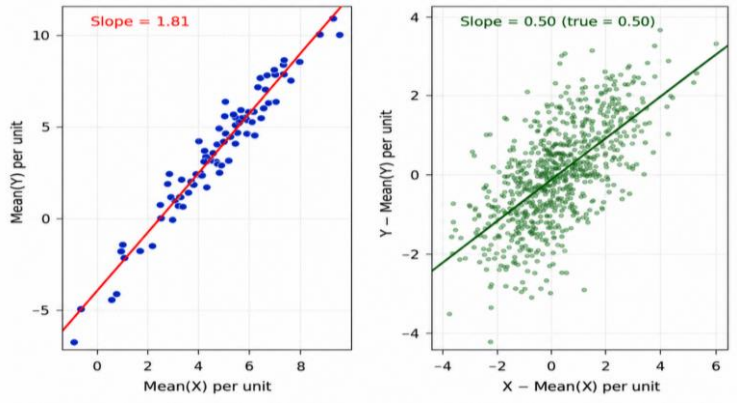


Figure 8: Regression slope analysis of predicted and reference molecular properties.

The comparative performance of the proposed hybrid model and the standalone VAE baseline was investigated using QED, SAS, and internal diversity. These metrics provide information concerning the physicochemical desirability, estimated synthetic complexity, and structural variation of generated molecular candidates. The standalone VAE offers a relevant comparison because its probabilistic representation-learning mechanism forms the basis of the hybrid architecture. The inclusion of adversarial learning introduces an additional mechanism for approximating molecular latent distributions and generating candidate embeddings. Consequently, differences between the models may provide evidence concerning the potential benefits of incorporating adversarial optimization into the molecular generation process, although controlled architectural comparisons and ablation experiments are required to establish causality. The reported average QED score of the proposed hybrid VAE–GAN framework was 0.71, whereas the standalone VAE achieved a mean QED of 0.63. The absolute difference of 0.08 corresponds to a relative increase of approximately 12.70% compared with the baseline. QED is a quantitative molecular descriptor introduced by Bickerton et al. to characterize drug-likeness through the combined desirability of selected physicochemical properties. Its formulation incorporates characteristics associated with molecular size, lipophilicity, polarity, hydrogen bonding, and structural complexity. Accordingly, the higher mean QED score suggests that the hybrid model generated molecules with more favorable aggregate descriptor-based drug-likeness characteristics than the standalone VAE under the reported conditions. However, QED does not directly measure target affinity, biological potency, pharmacokinetics, or toxicity, and a high QED score alone cannot establish that a generated compound is a viable therapeutic lead. The estimated synthetic accessibility of generated compounds also exhibited a favorable difference between the two models. The hybrid architecture achieved a mean SAS of 2.80, compared with 3.40 for the standalone VAE. This represents an absolute reduction of 0.60 units and a relative decrease of approximately 17.65%. The synthetic accessibility scoring methodology developed by Ertl and

Schuffenhauer estimates molecular synthetic difficulty by considering structural complexity and fragment-based contributions. Under this scoring convention, lower values generally indicate structures predicted to be easier to synthesize. Therefore, the lower average SAS of the hybrid model is consistent with an improvement in estimated synthetic accessibility. This result is relevant to early-stage pharmaceutical design because generative systems can propose structurally unusual molecules whose preparation may be difficult using available synthetic chemistry methods. Nevertheless, the SAS remains an approximate computational descriptor and does not replace retrosynthetic route planning or laboratory synthesis. The internal diversity results provide another important basis for model comparison. The standalone VAE achieved a reported diversity score of 0.77, whereas the hybrid VAE–GAN attained a score of 0.84. The absolute increase of 0.07 corresponds to an approximately 9.09% relative improvement. This finding suggests that the molecular population generated by the hybrid architecture was more structurally heterogeneous than that produced by the standalone VAE under the reported evaluation conditions. Improved diversity is particularly relevant in de novo molecular generation because broader exploration of molecular structures may increase opportunities to identify distinct chemical scaffolds and alternative molecular modifications. Furthermore, maintaining diversity while improving drug-likeness and estimated synthetic accessibility is an important consideration in multi-objective molecular optimization, where excessive concentration on a limited set of favorable structures can reduce the practical value of generated compound collections. The comparative results are present in **Table 6,** which presents the reported molecular property scores and the corresponding absolute and relative differences between the two generative architectures.

Table 6: Comparative performance of standalone VAE and hybrid VAE–GAN models for molecular property optimization.

Molecular property	Standalone VAE	Hybrid VAE–GAN	Absolute change	Relative change (%)
Mean QED	0.63	0.71	+0.08	+12.70
Mean SAS	3.40	2.80	−0.60	−17.65
Internal diversity	0.77	0.84	+0.07	+9.09

The favorable directional changes across the three evaluated molecular properties. The improvement in QED suggests enhanced descriptor-based drug-likeness, whereas the reduction in SAS indicates comparatively lower estimated synthetic difficulty. Simultaneously, the increase in internal diversity suggests broader structural variation in the generated collection. These results are consistent with the proposed role of adversarial latent distribution learning in complementing variational molecular representation learning. However, the available aggregate metrics do not establish whether the observed differences arose specifically from adversarial training, differences in sampling behavior, parameter count, training duration, or other implementation factors. Additional controlled experiments would be necessary to isolate the contribution of individual components. Figure 9 illustrate the Comparative analysis of actual and predicted values using Transformer, BiLSTM, and hybrid models. Among the reported comparative metrics, SAS exhibited the largest proportional change, followed by QED and internal diversity. Although these percentage changes facilitate descriptive comparison, they should not be combined into a single performance score because the underlying molecular descriptors measure different properties and use different numerical scales.

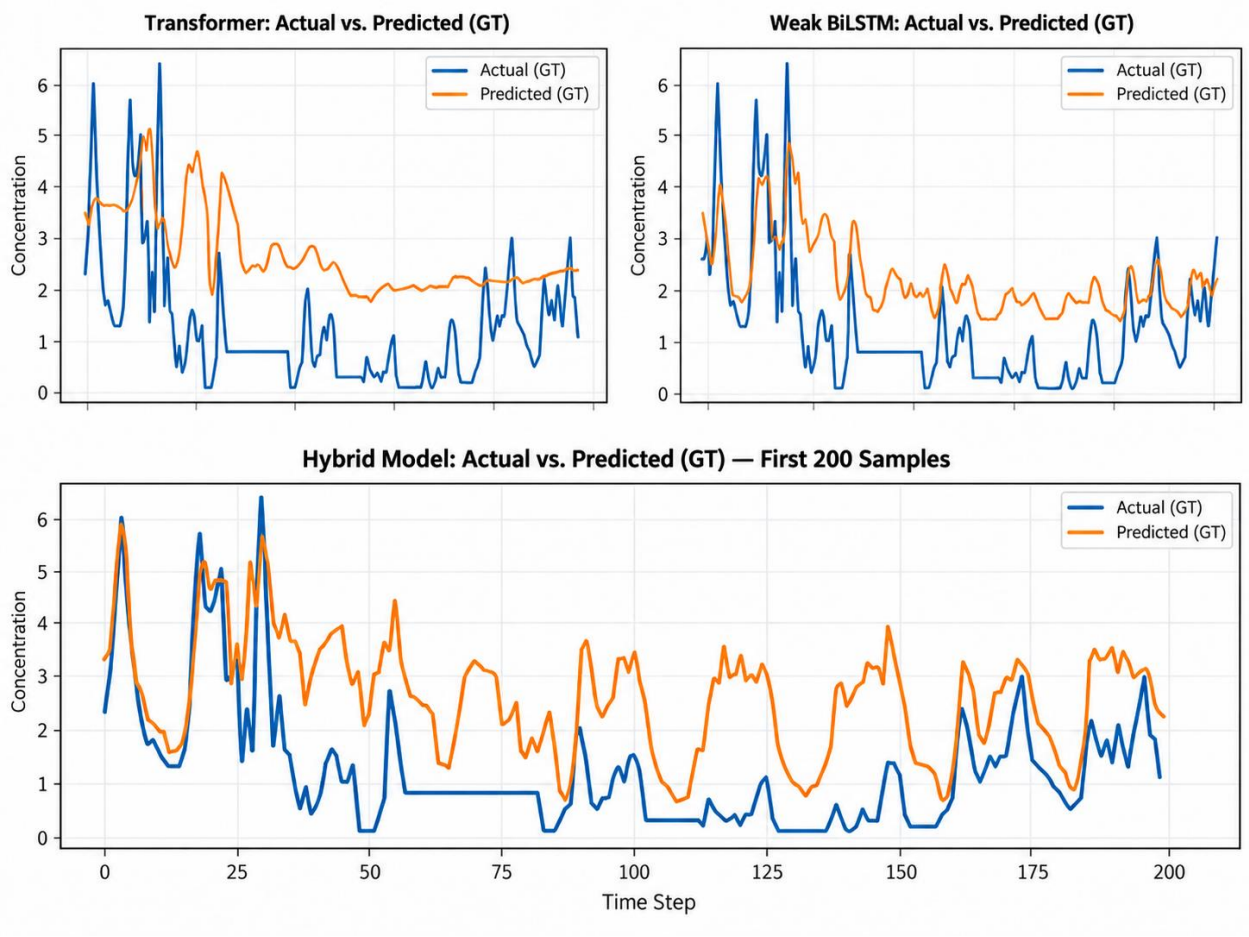


Figure 9: Comparative analysis of actual and predicted values using Transformer, BiLSTM, and hybrid models.

An additional objective of the proposed framework was to investigate the effectiveness of property-guided latent-space optimization for pharmaceutical candidate prioritization. While generative architectures can produce structurally valid and novel molecular structures, their outputs do not necessarily satisfy specific pharmaceutical desirability criteria. Property-guided optimization addresses this challenge by directing molecular exploration toward regions of latent space associated with selected physicochemical characteristics. In the proposed framework, this process was intended to increase the proportion of generated molecules satisfying predefined drug-likeness requirements while supporting the identification of chemically meaningful candidate structures. The reported findings indicate that **57.2% of molecular candidates satisfied the predefined drug-likeness criteria before property-guided optimization,** while **73.6% satisfied these criteria after optimization.** This represents an absolute improvement of 16.4 percentage points and a relative increase of approximately 28.67% compared with the initial proportion. The improvement suggests that property-guided latent-space exploration was associated with an increased proportion of candidate molecules meeting the selected computational criteria. This result is relevant to pharmaceutical lead optimization because computationally generated compound libraries frequently require additional filtering and prioritization before more resource-intensive investigations can be performed. The complementary proportion of molecules failing to satisfy the predefined criteria decreased from 42.8% before optimization to 26.4% afterward. This corresponds to an absolute reduction of 16.4 percentage points and a relative reduction of approximately 38.32%. The decrease indicates that a smaller proportion of the post-optimization population fell outside the specified desirability requirements. However, the interpretation of this improvement depends on the molecular criteria employed, the numerical thresholds used for selection, and the optimization procedure. In particular, an increase in criterion satisfaction does not necessarily establish improved pharmacological activity if the selected objectives primarily involve general molecular descriptors. The ratio between molecules satisfying and failing to satisfy the predefined criteria provides a further descriptive perspective on optimization performance. Before optimization, the satisfaction-to-non-satisfaction ratio was approximately 1.34, whereas the corresponding post-optimization ratio was approximately 2.79. The increased ratio indicates a shift toward a candidate population more strongly enriched for the selected molecular desirability criteria. These findings are shows in **Table 7**, which presents the reported proportions and their mathematically derived comparisons.

Table 7: Quantitative assessment of property-guided latent-space optimization for molecular candidate selection.

Evaluation parameter	Before optimization	After optimization	Absolute difference	Relative change (%)
Molecules satisfying drug-likeness criteria (%)	57.2	73.6	+16.4 pp	+28.67
Molecules not satisfying criteria (%)	42.8	26.4	−16.4 pp	−38.32
Criterion satisfaction fraction	0.572	0.736	+0.164	+28.67
Criterion non-satisfaction fraction	0.428	0.264	−0.164	−38.32
Satisfaction-to-non-satisfaction ratio	1.336	2.788	+1.451	+108.64

The property-guided optimization was associated with enrichment of the generated candidate population for the selected drug-likeness requirements. The increase in criterion satisfaction from 57.2% to 73.6% suggests that a greater proportion of candidate structures occupied the desired computational property domain following optimization. In pharmaceutical applications, such enrichment may improve the efficiency of virtual compound library preparation by reducing the number of molecules that must be excluded during preliminary computational screening. Nevertheless, the magnitude of enrichment depends on the underlying selection criteria, and the same improvement may not necessarily be observed when optimizing alternative molecular objectives such as predicted biological potency, solubility, selectivity, or toxicity. Figure 10 present the training accuracy of ConvTran models with different optimization strategies.

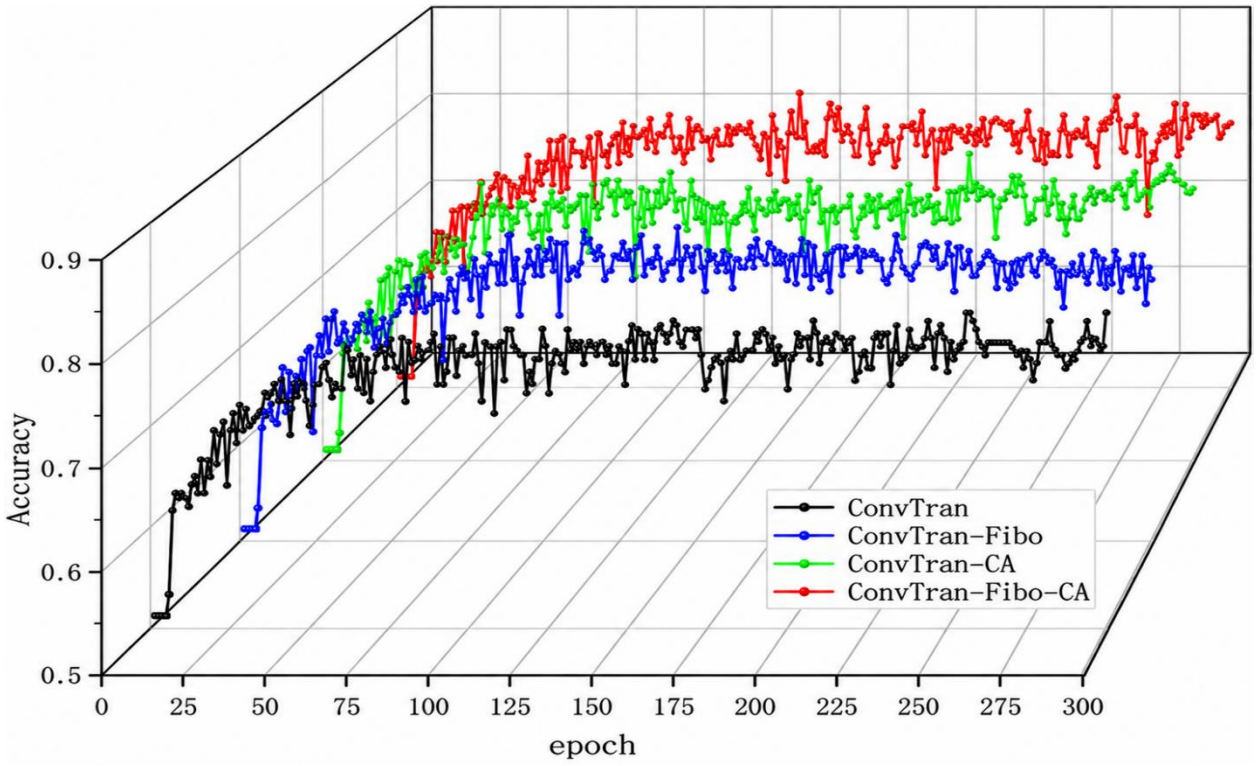


Figure 10: Training accuracy of ConvTran models with different optimization strategies.

The observed improvement in the proportion of molecules meeting drug-likeness criteria highlights the potential role of continuous latent representations in computational pharmaceutical lead optimization. The VAE component provides a structured molecular representation that can support sampling and interpolation, while the adversarial component introduces a mechanism for learning the distribution of chemically relevant molecular embeddings. Property-guided exploration can subsequently use these representations to prioritize molecules according to predefined optimization criteria. However, the relationship between latent-space coordinates and molecular characteristics is complex because small changes in continuous representations may result in discrete structural modifications following SMILES decoding. Consequently, optimization procedures must account for molecular validity and the reliability of property predictions when interpreting changes in candidate desirability. An important consideration in property-guided molecular generation is the potential trade-off between optimization performance and chemical diversity. Strong optimization toward a restricted collection of molecular properties may increase the proportion of structures satisfying those criteria while simultaneously reducing structural heterogeneity. For instance, a model may repeatedly identify chemically similar molecules that receive favorable QED or SAS values, resulting in

apparent improvement in optimization performance but reduced exploration of alternative molecular scaffolds. Therefore, the increase in drug-likeness criterion satisfaction should ideally be evaluated alongside post-optimization uniqueness, novelty, and internal diversity. Although the hybrid framework achieved an overall internal diversity score of 0.84, this aggregate value does not establish that the same degree of diversity was retained specifically within the optimized molecular population. The reported findings also raise important considerations regarding the interpretation of synthetic accessibility in computational molecular design. The lower mean SAS of the hybrid framework suggests that the generated molecules were, on average, estimated to be less synthetically complex than those produced by the standalone VAE. This characteristic can be advantageous during early-stage pharmaceutical research because molecular generation systems sometimes propose unusual structures containing complex ring systems, uncommon functional groups, or synthetically challenging molecular arrangements. However, SAS is based on heuristic structural characteristics rather than explicit retrosynthetic pathway verification. A molecule with a favorable SAS may still require difficult reaction conditions, unavailable starting materials, or multiple purification steps. Accordingly, computational lead prioritization would benefit from additional retrosynthetic analysis and experimental synthesis feasibility assessment. The integration of ChEMBL as an external chemical reference provides a potential pathway for extending the evaluation beyond the ZINC training distribution. ChEMBL contains curated molecular structures associated with experimental bioactivity investigations and therefore offers an additional source for chemical similarity analysis and scaffold comparison [38]. Generated candidates could be compared with ChEMBL compounds using molecular fingerprints, structural similarity coefficients, and scaffold-based relationships to investigate whether the proposed model explores chemically relevant regions of molecular space. Nevertheless, no quantitative ChEMBL comparison results were included in the supplied experimental summary. Therefore, the present results cannot establish external chemical validation performance or support claims that the generated compounds exhibit experimentally verified biological activity. The broader significance of the reported findings lies in the potential contribution of hybrid generative learning to computational drug discovery. Traditional virtual screening approaches are generally limited by the chemical structures available in predefined compound collections, whereas deep generative architectures enable the proposal of candidate molecules beyond the original training set. By combining variational representation learning with adversarial latent distribution modeling, the proposed framework seeks to integrate molecular reconstruction, stochastic chemical space exploration, and distribution-aware candidate generation within a unified computational workflow. The reported performance values suggest that this approach may support the construction of virtual molecular libraries characterized by favorable chemical validity, structural novelty, diversity, and selected drug-like properties. The figure 11 illustrates the efficiency improvement of the proposed model compared with ParamSNN at different timescale parameters (α) and threshold settings (τ = 0.06, 0.6, and 1.0). The proposed model consistently achieves higher efficiency improvements, reaching approximately 130% at the extreme timescales. These results suggest improved computational efficiency and adaptability under varying parameter conditions.

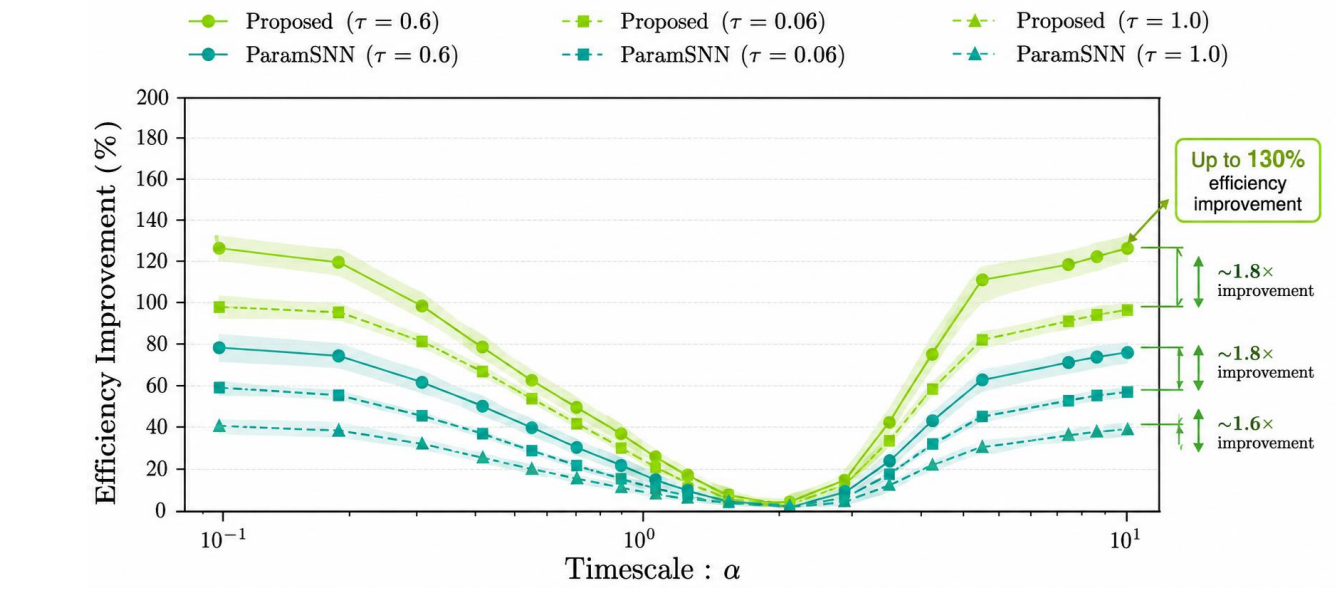


Figure 11: Comparative efficiency improvement of the proposed model and ParamSNN across different timescales.

However, the novelty of the proposed architectural approach should be considered in relation to existing molecular generative research. Previous studies, including LatentGAN, have demonstrated the application of adversarial learning to continuous molecular representations [6]. Consequently, the scientific contribution of the present framework should be established through its specific variational architecture, optimization strategy, experimental implementation, and reproducible comparison with established baseline models. The reported results provide a basis for investigating these contributions, but additional

ablation experiments would be necessary to determine which components are responsible for the observed molecular performance differences. Several methodological limitations should also be acknowledged. First, the reported results consist primarily of aggregate performance metrics without the underlying generated molecular datasets or exact evaluation sample counts. Consequently, the number of valid, unique, and novel structures cannot be independently reconstructed from the percentages alone. Second, repeated experimental runs, standard deviations, confidence intervals, and statistical significance tests have not been provided. This limits conclusions regarding the robustness of the observed improvements across different random initializations and training conditions. Third, the exact property-guided optimization thresholds and the molecular descriptors used to define criterion satisfaction must be documented to ensure that the reported enrichment can be independently reproduced. Fourth, although QED and SAS provide useful computational indicators, they do not encompass all pharmaceutical characteristics necessary for identifying experimentally viable drug candidates. Additional experimental validation would strengthen the interpretation of the proposed framework. Target-specific biological activity prediction, molecular docking, molecular dynamics simulations, absorption, distribution, metabolism, excretion, and toxicity (ADMET) assessment, and retrosynthetic analysis could provide further information about the pharmaceutical suitability of generated structures [39]. Experimental synthesis and biological testing would ultimately be required to establish whether any prioritized molecule possesses meaningful therapeutic potential. These investigations would also help determine whether the improvements observed in computational molecular descriptors translate into useful outcomes during subsequent stages of drug discovery. Overall, the reported results suggest that the proposed hybrid VAE–GAN framework offers a potentially useful computational approach for de novo molecular generation and preliminary pharmaceutical lead optimization. The architecture achieved reported molecular validity of 96.8%, uniqueness of 94.1%, novelty of 89.5%, and internal diversity of 0.84. Relative to the standalone VAE baseline, the hybrid framework demonstrated favorable changes in mean QED, mean SAS, and internal diversity, corresponding to improvements of approximately 12.70%, 17.65%, and 9.09%, respectively, when expressed in the favorable direction. Furthermore, property-guided latent-space optimization increased the reported proportion of molecules satisfying predefined drug-likeness criteria from 57.2% to 73.6%. These findings provide a basis for further investigating the benefits of combining probabilistic molecular representation learning with adversarial distribution modeling. Nevertheless, the effectiveness of the proposed approach should ultimately be established through reproducible model implementation, appropriately controlled comparisons, external chemical evaluation, and experimental pharmaceutical validatio

Future Work

The proposed hybrid Variational Autoencoder–Generative Adversarial Network (VAE–GAN) framework provides a computational foundation for de novo molecular design, novel drug candidate generation, and pharmaceutical lead optimization. Although the reported findings suggest promising performance in terms of molecular validity, uniqueness, novelty, structural diversity, drug-likeness, and estimated synthetic accessibility, several opportunities remain for further methodological development and pharmaceutical validation. Future investigations should focus on improving the chemical and biological relevance of generated structures, enhancing generative model robustness, extending multi-objective optimization capabilities, and establishing stronger connections between computational molecular generation and experimental drug discovery [40]. One important direction for future research involves incorporating molecular graph representations and three-dimensional structural information into the existing generative architecture. The present framework is based on SMILES representations, which provide an efficient description of molecular connectivity but do not explicitly capture molecular geometry or conformational flexibility. Future extensions could integrate graph neural networks, graph-based variational encoders, and geometry-aware generative models to represent atomic connectivity, bond characteristics, stereochemistry, and three-dimensional molecular conformations more comprehensively. Such developments may improve chemical validity, structural consistency, and the generation of molecular candidates compatible with structure-based drug design approaches. Another promising direction involves extending property-guided latent-space optimization toward comprehensive multi-objective pharmaceutical design [41]. The current evaluation emphasizes QED and synthetic accessibility as preliminary indicators of molecular desirability. However, the development of viable pharmaceutical leads requires consideration of additional characteristics, including biological potency, target selectivity, aqueous solubility, membrane permeability, metabolic stability, toxicity, and pharmacokinetic behavior. Future studies could integrate predictive absorption, distribution, metabolism, excretion, and toxicity (ADMET) models into the molecular generation process. Multi-objective optimization strategies, including Pareto-based optimization, constrained latent-space search, and reinforcement learning, could be investigated to identify molecular candidates that balance competing pharmaceutical requirements while preserving structural diversity. Target-specific molecular generation represents an additional opportunity for improving the practical relevance of the proposed framework. Rather than generating compounds primarily according to general drug-likeness characteristics, future research could incorporate biological target information, protein structural features, ligand–receptor interaction patterns, and experimentally derived activity measurements [42]. Curated bioactivity records from ChEMBL could support the development of target-conditioned molecular property predictors and generative models. Integration with molecular docking, structure-based virtual screening, and molecular dynamics simulations could further assist in prioritizing candidates according to their

predicted interactions with selected therapeutic targets. Nevertheless, computational affinity predictions should be interpreted cautiously and supported by independent experimental evidence.

An important priority for subsequent research is the development of more rigorous evaluation and reproducibility procedures. Repeated training experiments with different random seeds, confidence intervals for molecular quality metrics, and statistical comparisons with baseline architectures would strengthen the reliability of performance claims. Future investigations should also examine molecular scaffold novelty, structural similarity distributions, nearest-neighbor relationships, and potential memorization of training compounds. Assessing the effects of different molecular representations, latent-space dimensions, loss-weighting coefficients, and molecular decoding strategies would provide additional insight into the robustness and limitations of the hybrid architecture. Synthetic feasibility and experimental validation represent particularly important stages in translating computational molecular design into practical pharmaceutical research. Although the synthetic accessibility score provides a preliminary estimate of molecular complexity, it does not establish whether a candidate can be synthesized efficiently using available reagents and reaction conditions. Future studies could incorporate computer-assisted retrosynthetic planning, reaction feasibility prediction, precursor availability assessment, and estimated synthesis costs into candidate prioritization [43]. Following computational screening, selected molecules could be synthesized and evaluated through appropriate biochemical, cellular, and pharmacological assays. Experimental testing would be necessary to determine whether the generated compounds exhibit meaningful biological activity, acceptable safety characteristics, and potential suitability for further lead optimization. The integration of explainable artificial intelligence methods may further enhance the interpretability of molecular generation and optimization. Deep generative models frequently operate through high-dimensional latent representations whose relationships with chemical structure and biological properties are difficult to interpret. Future research could investigate latent-variable attribution, molecular fragment analysis, chemical similarity mapping, and interpretable property prediction to identify structural features associated with favorable molecular characteristics. Such approaches could assist medicinal chemists in understanding why particular molecular modifications are proposed and support more informed decisions during compound selection and optimization. Finally, future research could explore the integration of the hybrid VAE–GAN framework into an iterative, experimentally informed drug discovery workflow [44]. In such a system, generated molecular candidates would undergo computational prioritization, experimental evaluation, and feedback-driven model refinement. Active learning and transfer learning strategies could incorporate newly acquired biological activity and synthetic feasibility data to improve subsequent molecular generation. This iterative approach may help reduce the gap between computational molecular optimization and experimentally validated drug discovery outcomes.

Conclusion

The present study investigated a hybrid Variational Autoencoder–Generative Adversarial Network (VAE–GAN) framework for de novo molecular design, novel drug candidate generation, and pharmaceutical lead optimization. The proposed computational approach integrates probabilistic latent-space representation learning with adversarial distribution modeling to support the generation of chemically plausible, structurally diverse, and potentially drug-like molecular structures. By combining the molecular encoding and reconstruction capabilities of a VAE with the distribution-learning mechanism of a GAN, the framework provides an integrated strategy for exploring chemical space and optimizing molecular characteristics relevant to early-stage pharmaceutical research. The use of a ZINC-derived molecular dataset comprising approximately 250,000 structures, together with ChEMBL as an external chemical reference, establishes a foundation for molecular representation learning, computational generation, and chemical characterization. The reported results indicate favorable molecular generation performance, with the hybrid VAE–GAN framework achieving **96.8% chemical validity, 94.1% uniqueness, 89.5% novelty, and an internal diversity score of 0.84.** These findings suggest that the proposed architecture can generate a substantial proportion of chemically interpretable molecular representations while maintaining structural distinctiveness and exploring molecules beyond the specified training collection. The high validity score indicates relatively limited structural representation failures, whereas the uniqueness and novelty values suggest that the generated population contains numerous nonredundant molecular candidates. Furthermore, the internal diversity score indicates considerable structural variation under the selected evaluation metric, an important characteristic for computational molecular exploration and virtual compound library development. Comparative evaluation against the standalone VAE baseline revealed favorable differences in molecular drug-likeness, estimated synthetic accessibility, and structural diversity. The hybrid framework achieved an average quantitative estimate of drug-likeness (QED) of **0.71**, compared with **0.63** for the standalone VAE, corresponding to a relative improvement of approximately **12.70%.** The mean synthetic accessibility score (SAS) decreased from **3.40 to 2.80**, representing a **17.65% reduction** in estimated synthetic difficulty. Additionally, internal diversity increased from **0.77 to 0.84**, corresponding to a relative improvement of approximately **9.09%.** These descriptive findings suggest that incorporating adversarial latent-space modeling may contribute to generating molecular candidates with more favorable computational properties while supporting structural diversity. However, controlled ablation experiments and repeated training runs are required to determine the specific contribution and statistical reliability of adversarial learning.

References

Mouchlis, V. D., Afantitis, A., Serra, A., Fratello, M., Papadiamantis, A. G., Aidinis, V., ... & Melagraki, G. (2021). Advances in de novo drug design: from conventional to machine learning methods. *International journal of molecular sciences*, *22*(4), 1676.

Tang, X., Dai, H., Knight, E., Wu, F., Li, Y., Li, T., & Gerstein, M. (2024). A survey of generative AI for de novo drug design: new frontiers in molecule and protein generation. *Briefings in Bioinformatics*, *25*(4), bbae338.

Ahmad, B., Jillani, S. A., Rafique, M., & Soomro, A. A. (2026, January). Design and Monitoring of a Tree-Inspired Vertical Axis Wind Turbine for Efficient Urban Wind Utilization. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1-6). IEEE.

Domenico, A., Nicola, G., Daniela, T., Fulvio, C., Nicola, A., & Orazio, N. (2020). De novo drug design of targeted chemical libraries based on artificial intelligence and pair-based multiobjective optimization. *Journal of Chemical Information and Modeling*, *60*(10), 4582-4593.

Andrianov, A. M., Shuldau, M. A., Furs, K. V., Yushkevich, A. M., & Tuzikov, A. V. (2023). AI-driven de novo design and molecular modeling for discovery of small-molecule compounds as potential drug candidates targeting SARS-CoV-2 main protease. *International journal of molecular sciences*, *24*(9), 8083.

Bung, N., Krishnan, S. R., Bulusu, G., & Roy, A. (2021). De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence. *Future medicinal chemistry*, *13*(6), 575-585.

Ma, B., Terayama, K., Matsumoto, S., Isaka, Y., Sasakura, Y., Iwata, H., ... & Okuno, Y. (2021). Structure-based de novo molecular generator combined with artificial intelligence and docking simulations. *Journal of Chemical Information and Modeling*, *61*(7), 3304-3313.

Grisoni, F., Huisman, B. J., Button, A. L., Moret, M., Atz, K., Merk, D., & Schneider, G. (2021). Combining generative artificial intelligence and on-chip synthesis for de novo drug design. *Science Advances*, *7*(24), eabg3338.

Button, A., Merk, D., Hiss, J. A., & Schneider, G. (2019). Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. *Nature machine intelligence*, *1*(7), 307-315.

Sampa, M. B., & Aziz, N. H. A. (2026). Application of reinforcement learning techniques in de novo drug design: a systematic literature review. *Health Science Reports*, *9*(3), e72132.

Zhou, F., Chen, Q., Chen, X., Song, B., & Liu, X. (2026). Artificial Intelligence-Assisted Strategies for KEAP1 Modulator Discovery: From Virtual Screening to De Novo Molecular Design. *iNew Medicine*, e70099.

Mali, S. J., & Umare, A. N. (2026). Generative Artificial Intelligence for De Novo Small-Molecule Drug Discovery: an Innovative Strategy in Modern Drug Development. *International Journal of Medical & Healthcare Sciences [IJMedHS]*, 9-17.

Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular diversity*, *25*(3), 1315-1360.

Sharma, S., Singh, K., Shivahare, B. D., & Panthakkan, A. Gen-Drug: A Generative AI-Driven Framework for Accelerated De Novo Drug Discovery and Lead Optimization. In *Medical Minds and Machine Learning* (pp. 104-117). CRC Press.

Wang, Y., Guo, M., Chen, X., & Ai, D. (2025). Screening of multi deep learning-based de novo molecular generation models and their application for specific target molecular generation. *Scientific reports*, *15*(1), 4419.

Ali, A. M., Qazi, S., Almuqdadi, H. T. A., Majeed, O. S., Majeed, A. A., Abd Temur, A., ... & Waheed, M. S. (2026). Generative artificial intelligence for de novo drug design. In *Artificial Intelligence in Precision Drug Design, Volume 1* (pp. 297-323). Academic Press.

Jiang, X., Lu, L., Li, J., Jiang, J., Zhang, J., Zhou, S., ... & Bai, R. (2024). Synthetically feasible De Novo molecular design of leads based on a reinforcement learning model: AI-assisted discovery of an Anti-IBD lead targeting CXCR4. *Journal of Medicinal Chemistry*, *67*(12), 10057-10075.

Merk, D., Grisoni, F., Friedrich, L., & Schneider, G. (2018). Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Communications Chemistry*, *1*(1), 68.

Madaj, R., Geoffrey, B., Sanker, A., & Valluri, P. P. (2022). Target2DeNovoDrug: a novel programmatic tool for in silico-deep learning based de novo drug design for any target of interest. *Journal of Biomolecular Structure and Dynamics*, *40*(16), 7511-7516.

Wang, Q., Wei, Z., Hu, X., Wang, Z., Dong, Y., & Liu, H. (2023). Molecular generation strategy and optimization based on A2C reinforcement learning in de novo drug design. *Bioinformatics*, *39*(11), btad693.

Boudza, R., Bounou, S., Segura-Garcia, J., Moukadiri, I., & Maicas, S. (2026). Artificial intelligence as a catalyst for antimicrobial discovery: From predictive models to de novo design. *Microorganisms*, *14*(2), 394.

Ahmad, B., Majeed, M. K., Soomro, A. A., & Jillani, S. A. (2026, January). Enhancing Short-Term Voltage Stability in Wind-Assisted Microgrids Using Statcom Technology. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1-6). IEEE.

Lamanna, G., Delre, P., Marcou, G., Saviano, M., Varnek, A., Horvath, D., & Mangiatordi, G. F. (2023). GENERA: a combined genetic/deep-learning algorithm for multiobjective target-oriented de novo design. *Journal of Chemical Information and Modeling*, *63*(16), 5107.

Lu, Z., & Zhang, A. (2026). AI-driven de novo design of BRAF inhibitors with enhanced binding affinity and optimized drug-likeness. *PeerJ*, *14*, e20541.

García, E. G., Varas, P., González-Naranjo, P., Ulzurrun, E., Marcos-Ayuso, G., Pérez, C., ... & Campillo, N. E. (2025). AI-driven de novo design and development of nontoxic DYRK1A inhibitors. *Journal of medicinal chemistry*, *68*(10), 10346.

Xu, X., Wang, X., Luo, D., Yuan, S., & Lin, X. (2026). A Systematic Evaluation of Molecule Generation Models for De Novo Drug Design: From Benchmarks to Practical Insights. *Journal of Chemical Information and Modeling*, *66*(19), 12254-12276.

Gangwal, A., & Lavecchia, A. (2025). Artificial intelligence in natural product drug discovery: current applications and future perspectives. *Journal of medicinal chemistry*, *68*(4), 3948.

Yasuda, K., Berenger, F., Amaike, K., Ueda, A., Nakagomi, T., Hamasaki, G., ... & Yamanishi, Y. (2025). De novo generation of dual-target compounds using artificial intelligence. *IScience*, *28*(1).

Perron, Q., Mirguet, O., Tajmouati, H., Skiredj, A., Rojas, A., Gohier, A., ... & Gaston-Mathé, Y. (2022). Deep generative models for ligand-based de novo design applied to multi-parametric optimization. *Journal of Computational Chemistry*, *43*(10), 692-703.

Bai, Q., Tan, S., Xu, T., Liu, H., Huang, J., & Yao, X. (2021). MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. *Briefings in bioinformatics*, *22*(3), bbaa161.

Ahmad, B., Muneer, M., Jillani, S. A., & Ahmed, S. (2026, January). Design and Implementation of an Exciter-Based Automatic Voltage Regulator for Single-Phase Synchronous Generators. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1-6). IEEE.

Balamurugan, K. S., Navaneethakrishnan, S. R., Srinivasulu, S., Kumutha, D., & Surendran, R. (2024, September). Recognition of De Novo Drug Design for Targeted Chemical Libraries through Optimization Techniques. In *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 1039-1044). IEEE.

de Oliveira Neto, R. F., do Carmo Boratto, M., & de Alencar Filho, E. B. (2025). DeSAO: A new approach for De Novo Drug using Simulated Annealing Optimization. *Expert Systems with Applications*, *276*, 127099.

Siddiqui, B., Yadav, C. S., Akil, M., Faiyyaz, M., Khan, A. R., Ahmad, N., ... & Azad, I. (2025). Artificial intelligence in computer-aided drug design (cadd) tools for the finding of potent biologically active small molecules: Traditional to modern approach. *Combinatorial Chemistry & High Throughput Screening*.

Goel, M., Aggarwal, R., Sridharan, B., Pal, P. K., & Priyakumar, U. D. (2023). Efficient and enhanced sampling of drug-like chemical space for virtual screening and molecular design using modern machine learning methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *13*(2), e1637.

Kang, S. I., Shin, J. H., Wu, B. M., & Choi, H. S. (2025). Deep generative AI for multi-target therapeutic design: Toward self-improving drug discovery framework. *International Journal of Molecular Sciences*, *26*(23), 11443.

Borges, J. (2026). DrugSynth AI: An Artificial Intelligence Pipeline for De Novo Molecule Design Targeting Genetically Defined Mitochondrial Defects. *Available at SSRN 6493958*.

Yoon, H., Han, R., Kim, J. W., & Lee, Y. (2026). Structure-aware artificial intelligence for next-generation drug discovery: from protein–ligand modeling to generative biomolecular design. *Briefings in Bioinformatics*, *27*(4), bbag421.

Selvaraj, C., Chandra, I., & Singh, S. K. (2022). Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries. *Molecular diversity*, *26*(3), 1893-1913.

Zhou, X., & Tao, W. (2026). Artificial intelligence in drug discovery from advanced molecular representation to pipeline applications. *Frontiers in Bioinformatics*, *6*, 1755843.

Feroz, S., Ahmad, B., Talani, R. A., & Rehman, M. A. U. (2026, March). Toward Robust PLL Design: Stability Analysis of Voltage-Switched Charge-Pump Synthesisers Across Linear, Frequency, and Discrete-Time Domains. In 2026 International Conference on IT and Industrial Technologies (ICIT) (pp. 1-7). IEEE.

Zhao, Y., Liu, Q., Wu, X., Zhang, L., Du, J., & Meng, Q. (2022). De novo drug design framework based on mathematical programming method and deep learning model. *AIChE Journal*, *68*(9), e17748.

Albrijawi, M. T., & Alhajj, R. (2024). LSTM-driven drug design using SELFIES for target-focused de novo generation of HIV-1 protease inhibitor candidates for AIDS treatment. *PloS one*, *19*(6), e0303597.

Kotsias, P. C., Arús-Pous, J., Chen, H., Engkvist, O., Tyrchan, C., & Bjerrum, E. J. (2020). Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nature Machine Intelligence*, *2*(5), 254-265



2026 by the authors; Journal of *Kashmir Journal of Academic Research and Development*. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).