香港中文大學
The Chinese University of Hong Kong

20260423 Journal Club

Lewen Wang

ACS Chem. Neurosci. **2026**, *17*, 666–680.

ACS Chemical Neuroscience

pubs.acs.org/chemneuro

Review

Machine Learning for *De Novo* Molecular Generation: A Comprehensive Review

Yingjun Chen and Weiwei Xue*

Cite This: *ACS Chem. Neurosci.* 2026, 17, 666–680

Read Online

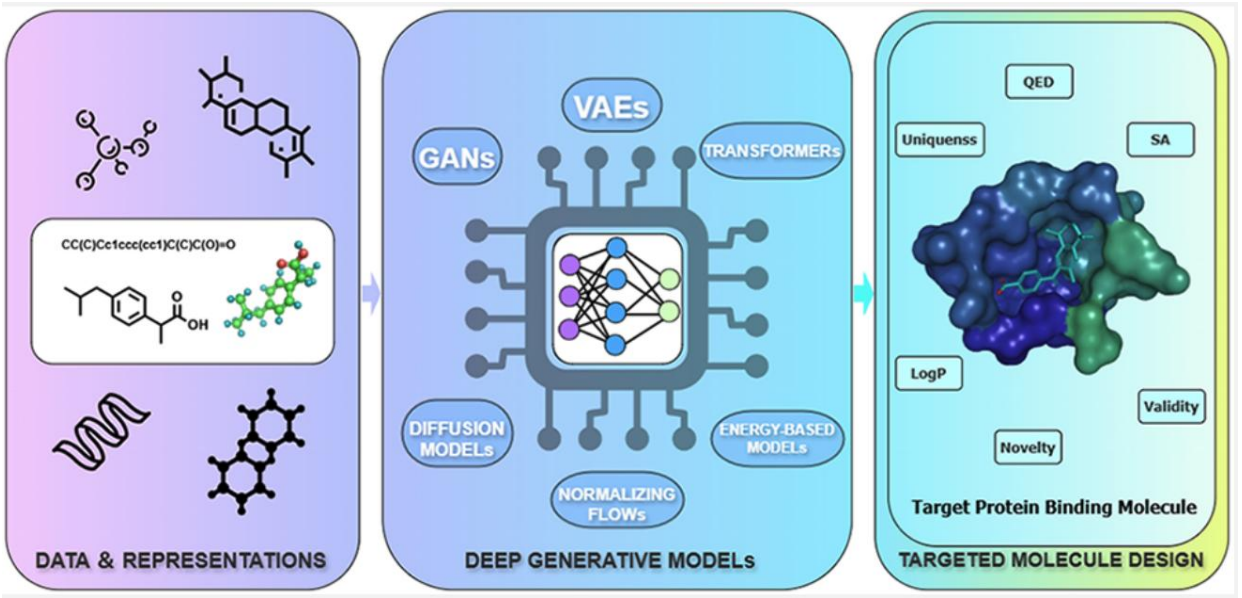
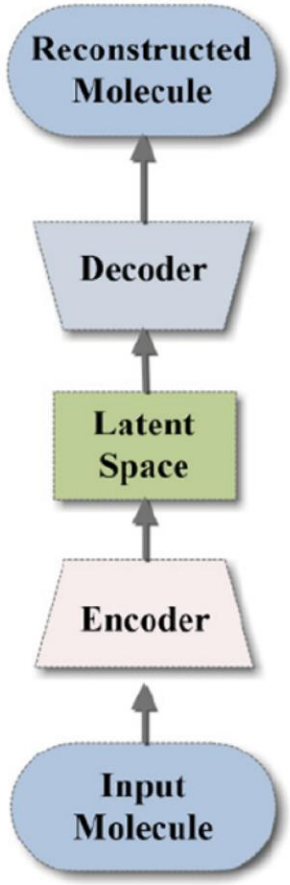


Table 1. Comparison of the Molecular Representations

Representation	Key Advantages	Key Disadvantages	Typical Architectures
1D String-Based*(e.g., SMILES, SELFIES)*	Computationally efficient Compatible with NLP Human-readable	High invalidity rate Nonsmooth latent space Loss of 3D/conformational info	RNNs, Transformers, VAEs, GANs
2D Graph-Based *(e.g., Adjacency Matrix)*	Natural topology Invariant to atom ordering High validity	Neglects 3D spatial info Computationally intensive Complex decoding steps	GNNs, Graph VAEs, Graph GANs
3D Geometry *(e.g., Point Clouds)*	Captures physical realism Essential for SBDD Highest information fidelity	High computational cost Requires E(3)-equivariance Issues with data sparsity	Equivariant GNNs, 3D CNNs, Diffusion Models



(a) Schematic diagram of the VAE model.

VAE-Based Models

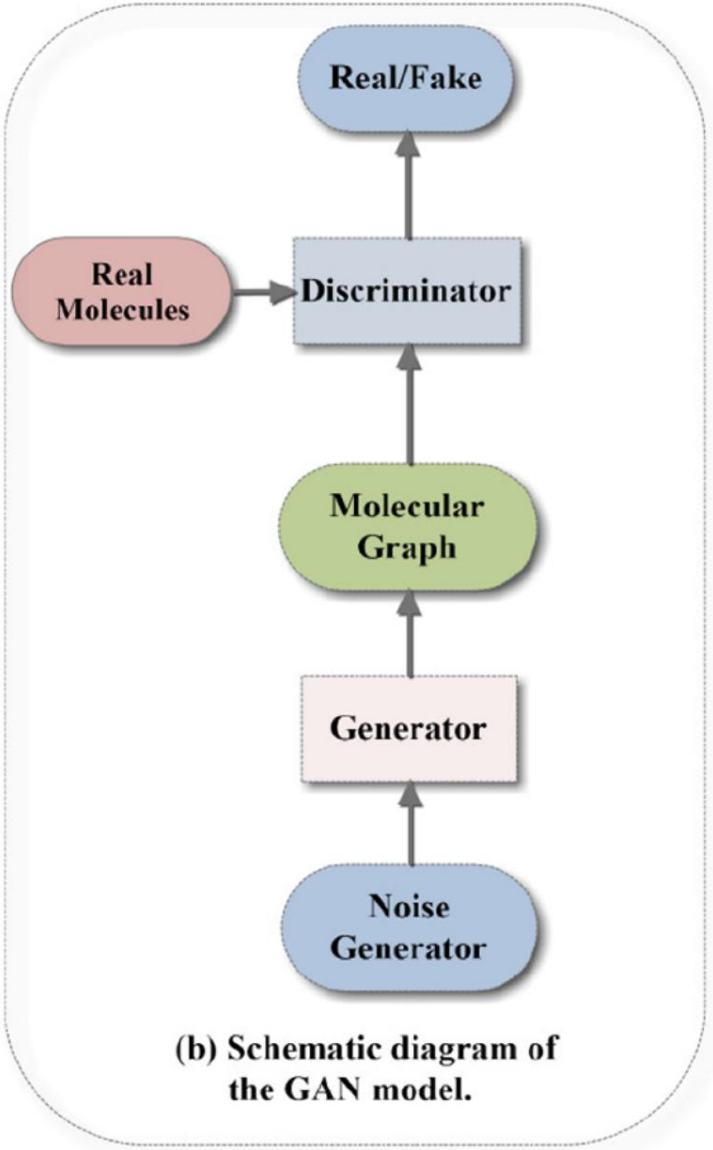
Maps molecules to a continuous latent probability distribution via encoder−decoder frameworks, enabling property optimization and structure interpolation.

Advantage

- The inherent smooth latent space is particularly advantageous for navigating the stringent physicochemical constraints of CNS drug discovery.

Disadvantages

- Validity−quality trade-of.
- Latent optimization landscapes are rarely perfectly smooth.



(b) Schematic diagram of the GAN model.

GAN-Based Models

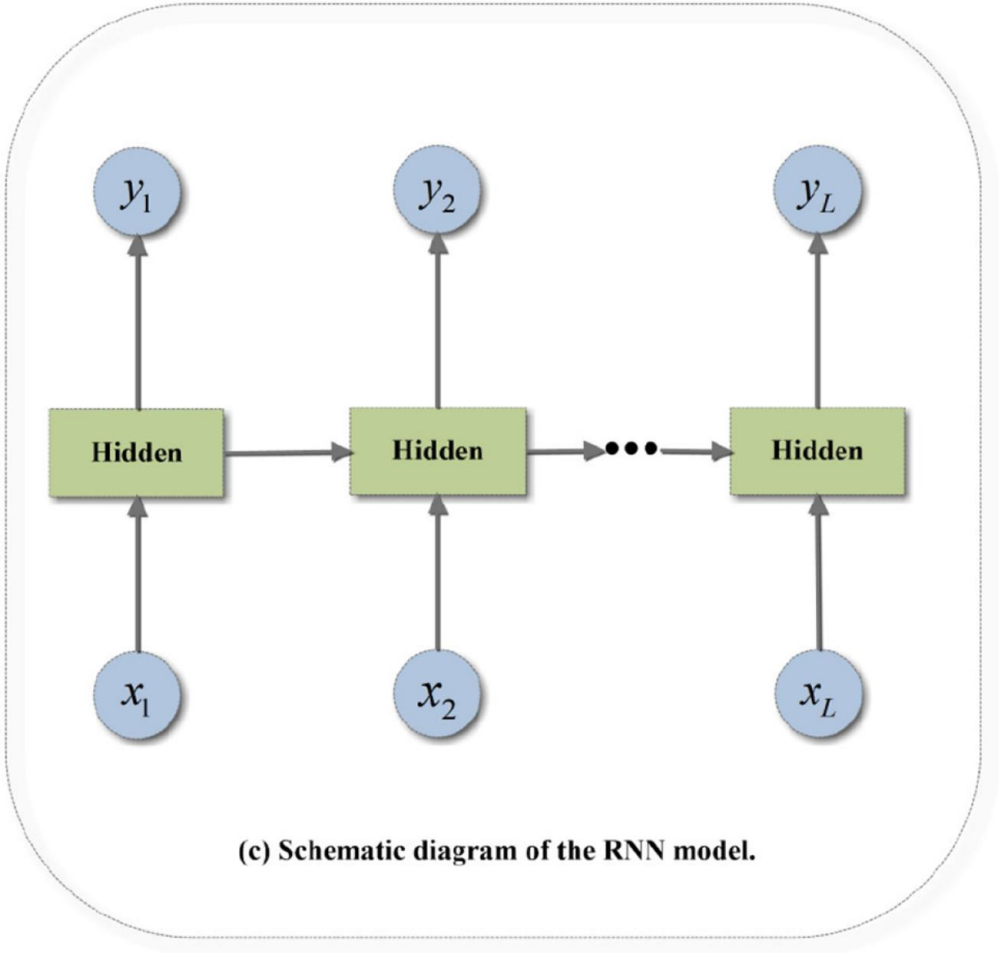
Approximate data distributions via an implicit minimax game between a generator and a discriminator.

Advantage

- Capable of producing high-fidelity structures

Disadvantages

- Mode collapse.
- Reward hacking.



(c) Schematic diagram of the RNN model.

RNN-Based Models

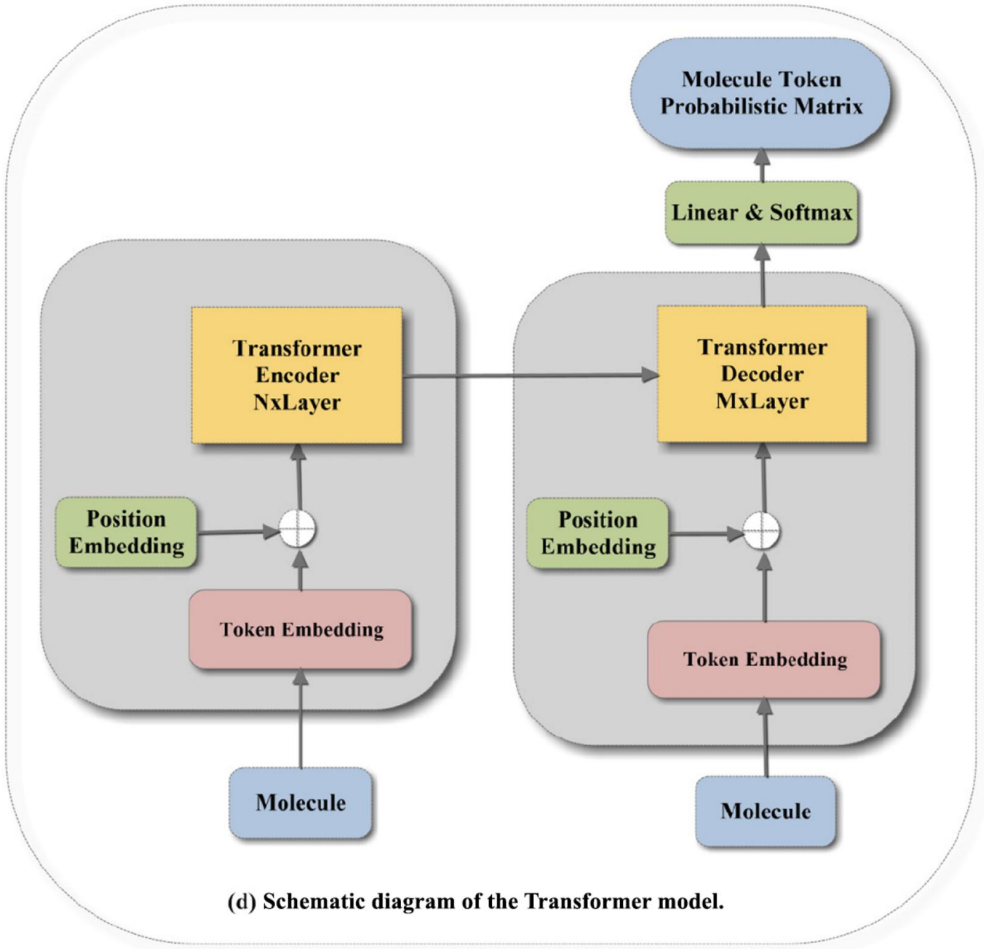
Particularly those using long-term memory (LSTM) units, served as foundational architectures for sequence-based generation. Typically operate on 1D string presentations.

Advantage

- Data efficient.
- Simple mature baseline.

Disadvantages

- Long-term dependency failure.
- Error accumulation.



(d) Schematic diagram of the Transformer model.

Transformer-Based Models

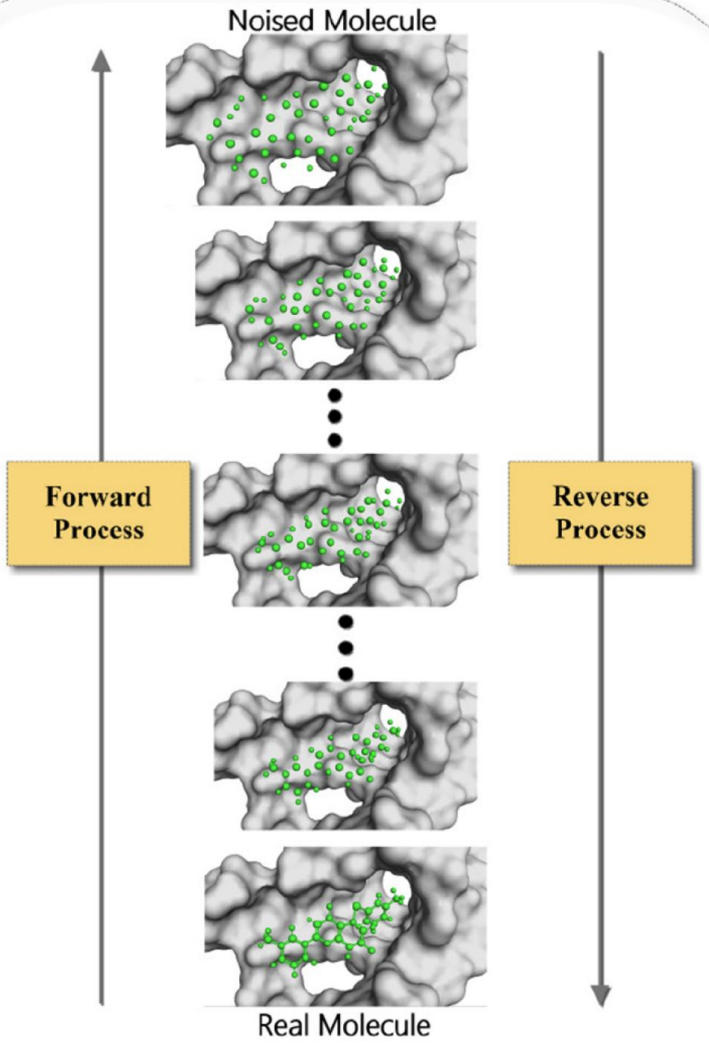
With its self-attention mechanism, the Transformer is particularly powerful at capturing long-range dependencies within molecular structures.

Advantage

- Captures long-range dependencies.

Disadvantages

- Data hungry.
- High compute cost.
- Chemical hallucination.



(e) Schematic diagram of the Diffusion model.

Diffusion-Based Models

They operate in two stages: a fixed forward process that adds noise and a learned reverse process that iteratively denoises samples into valid molecules.

Advantage

- State-of-the- art for geometric deep learning

Disadvantages

- High computational cost due to iterative denoising process.
- Slow sampling speed.
- Memorization.

VAE-Based

Model/Method	Key Contribution
VAE	Foundational generative model with a probabilistic latent space.
Chemical VAE	Latent space interpolation and optimization for molecular properties.
Grammar VAE	Ensures syntactic validity of SMILES via context-free grammar.
JT-VAE	Hierarchical, fragment-based generation ensuring chemical graph validity.
ARAE	Adversarial regularization of latent space to improve sample quality.
PCF-VAE	Mitigates posterior collapse to enhance molecular diversity.
Info-Theoretic VAE	Information-theoretically controlled, gradient-based latent space search.
Adaptive β-VAE	Dynamic tuning of KL-divergence weight (β) for property optimization.
TrustMol	Aligns latent space with molecular dynamics and quantifies uncertainty.
VAE-GMR	Direct inverse QSAR/QSPR by predicting latent vectors from properties.
DarkChem VAE	*In silico* property library generation for molecule identification.
Chemical VAE	Comparative study of VAE generation vs traditional similarity search.
Multimodal VAE	Application to anticancer molecule generation and bioactivity prediction.
Adapted JT-VAE	Inverse design of transition metal complexes with custom encoding.
LC-JT-VAE	Conditioned generation of molecular glues via protein sequence embedding.
Cage-VAE	Generative design of Porous Organic Cages (POCs).
Pretrained JT-VAE	Pretraining and fine-tuning for HOMO prediction and optimization.

GAN-Based

Model/Method	Key Contribution
Feedback GAN	Feedback loop for optimized generation and stereochemistry.
LM-GAN + GA	Hybrid LM-GAN with Genetic Algorithm.
QGAN	Quantum GANs for molecular graph generation.
DNMG (WGAN)	3D grid-based generation with Wasserstein GAN.
ED-GAN	Generation from experimental electron density (ED) maps.
ED-GAN/Pocket-GAN	ED and pocket-based design of M^{pro} inhibitors.
TopMT-GAN	Two-step 3D topology generation and atom assignment.
FS-GDA (WGAN-GP)	Few-shot domain adaptation for low-data scenarios.
BEGAN	Boltzmann reweighting for goal-directed optimization.

RNN-Based

Model/Method	Key Contribution
Fine-tuned RNN	Transfer learning for generating focused molecular libraries.
MolecularRNN (RL)	Objective-based optimization via reinforcement learning.
Bi-RNN	Bidirectional SMILES generation reflecting molecular nature.
Nested LSTM	Nested LSTM structure as an alternative to stacked RNNs.
MAN-RNN	Memory-augmented network for learning long-range dependencies.
Conditional RNN	3D pocket information (Coulomb matrix) to guide generation.
GxRNN	Conditioned on gene expression profiles for functional molecules.
Fragment-based LSTM	Analogue generation via fragment training and fine-tuning.
DeLA-Drug	Analog generation via "sampling with substitutions" strategy.
LSTM for Peptides	*De novo* design of therapeutic peptides.
GRU-based Model	Generation of drug candidates for COVID-19.
LSTM in CycleGAN	Hybrid model integrating RNN generator into CycleGAN framework.

Transformer-Based

Model/Method	Key Contribution
MolGPT	GPT-style decoder for next-token prediction in SMILES generation.
GMTransformer	BERT-style blank-filling objective for learning molecular grammars.
cMolGPT	Conditions generation via modifying keys/values in attention mechanism.
AlphaDrug	Target-specific design using a joint embedding of protein and SMILES.
TransGEM	Generates molecules conditioned on desired gene expression profiles.
Sc2Mol	Hybrid VAE-Transformer framework for scaffold generation and decoration.
FRATTVAE	Hybrid VAE with a Tree-Transformer for fragment-based generation.
DockingGA	Integrates a Transformer with a genetic algorithm and docking scores.
FSM-DDTR	Employs a feedback strategy for multiobjective optimization.

Diffusion-Based Models

Model/Method	Key Contribution
GCDM	Geometry-complete GNN for improved 3D validity, especially for large molecules.
SA-DM	Structure-aware generation using a k-NN module to capture local geometry.
EC-Conf	Single-step conformation generation via an ultrafast consistency model.
DTF-diffusion	Stepwise fusion of ligand and target information for better interaction modeling.
DiffSBDD	Versatile SE(3) model for multiple SBDD tasks (e.g., pocket-conditioning, inpainting).
PMDM	Dual (local/global) dynamics for pocket-conditioned generation and lead optimization.
PIDiff	Physics-informed guidance by integrating physicochemical principles into reverse diffusion.
SILVR	Fragment-based guidance for pretrained models at sampling time without retraining.
DrugDiff	Latent diffusion with flexible classifier guidance for property optimization.
GLDM	Diffusion process within a graph autoencoder's latent space for efficiency.
GCLDM	Diffusion within a "geometry-complete" autoencoder's latent space.

Table 2. Comparative Analysis of the Generative Model Architectures

Model	Mechanism	Key Strengths	Critical Limitations	Primary Use
VAEs	Probabilistic Encoder−Decoder	Smooth continuous latent space Fast inference	Blurry samples Posterior collapse issues	Scaffold hopping, optimization
GANs	Adversarial Minimax Game	High sample fidelity Sharp distribution matching	Mode collapse Unstable training Hard to optimize discrete data	Unconditional generation, 3D shapes
RNNs	Sequential Autoregression	Data efficient Simple mature baseline	Long-term dependency failure Error accumulation	SMILES library generation
Transformers	Self-Attention Mechanism	Captures long-range dependencies Parallel training	Data hungry High compute cost Chemical hallucination	Chemical language models (LLMs)
Diffusion	Iterative Denoising (SDE/ODE)	SOTA for 3D structures Stable training	Slow sampling speed Complex conditioning	Structure-based design (SBDD)

Category	Metric/Platform	Description
Foundational Metrics	Validity	Measures the proportion of generated molecules that are chemically plausible.
	Uniqueness	Quantifies the fraction of nonduplicate molecules among the valid outputs.
	Novelty	Assesses the percentage of generated molecules that are not present in the training set.
Physicochemical Properties	Drug-likeness	Evaluates molecules against empirical rules and scores like Lipinski's Rule of 5 and QED.
	Diversity	Measures the structural dissimilarity within a set of generated molecules (e.g., Tanimoto similarity).
Benchmarking Platforms	MOSES	A standardized evaluation suite focused on assessing how well a model reproduces the property distribution.
	GuacaMol	A comprehensive platform offering both distribution-learning and goal-directed benchmarks.
Synthesizability	SA/SC Score	Heuristic metrics that provide a computational estimate of a molecule's synthetic accessibility.
	CASP Tools	Computer-Aided Synthesis Planning tools for rigorous retrosynthetic analysis.

Core Applications Of Generative Models

1. Distribution Learning

- Capture the probability distribution of large molecular corpora.
- Generate entirely novel molecules that are statistically indistinguishable from the training data.
- A model that successfully approximates this distribution demonstrates a foundational understanding of chemical plausibility.

2. Goal-directed Generation

- Property-conditioned generation.
- Substructure-conditioned generation.
- Target-conditioned generation.
- Molecule optimization.

Thank You