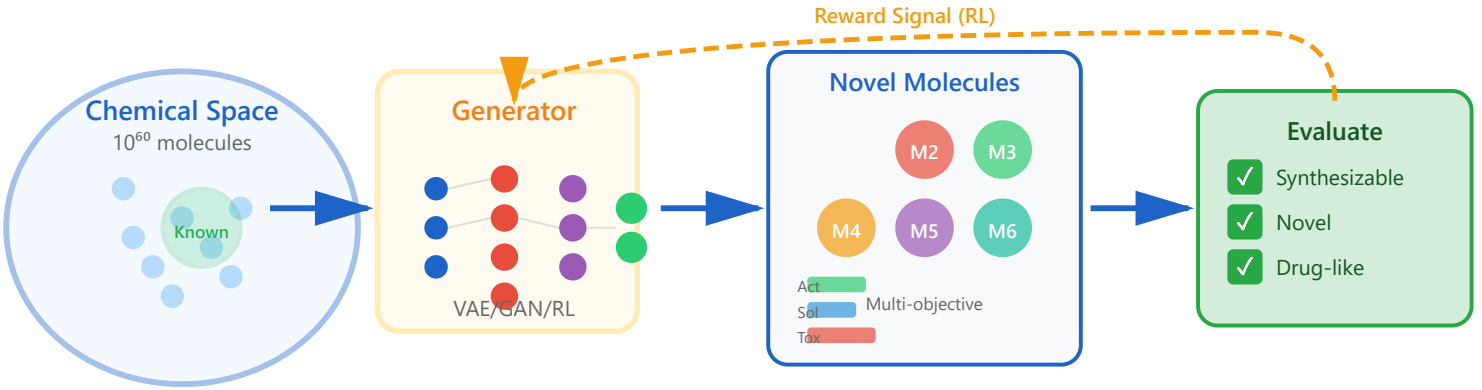
De Novo Design



Chemical space exploration

Novel compound generation

Reinforcement learning

Goal-directed optimization

VAE/GAN approaches

Generative architectures

Synthesizability

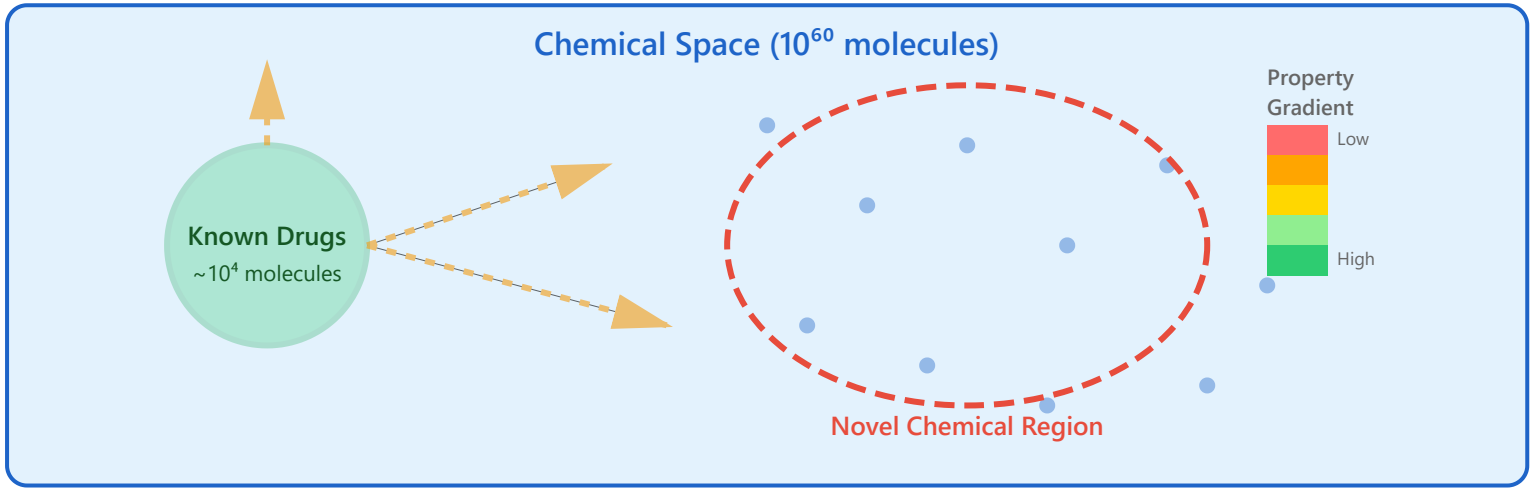
Chemical feasibility assessment

Diversity metrics

Novelty quantification

1. Chemical Space Exploration

Chemical space exploration involves navigating the vast universe of possible molecular structures to discover novel compounds with desired properties. The drug-like chemical space is estimated to contain 10^{60} possible molecules, far exceeding the number of atoms in the observable universe.



> **Key Concept:** Generative models learn the distribution of known molecules and can generate novel structures in unexplored regions of chemical space while maintaining drug-like properties.

Exploration Strategies

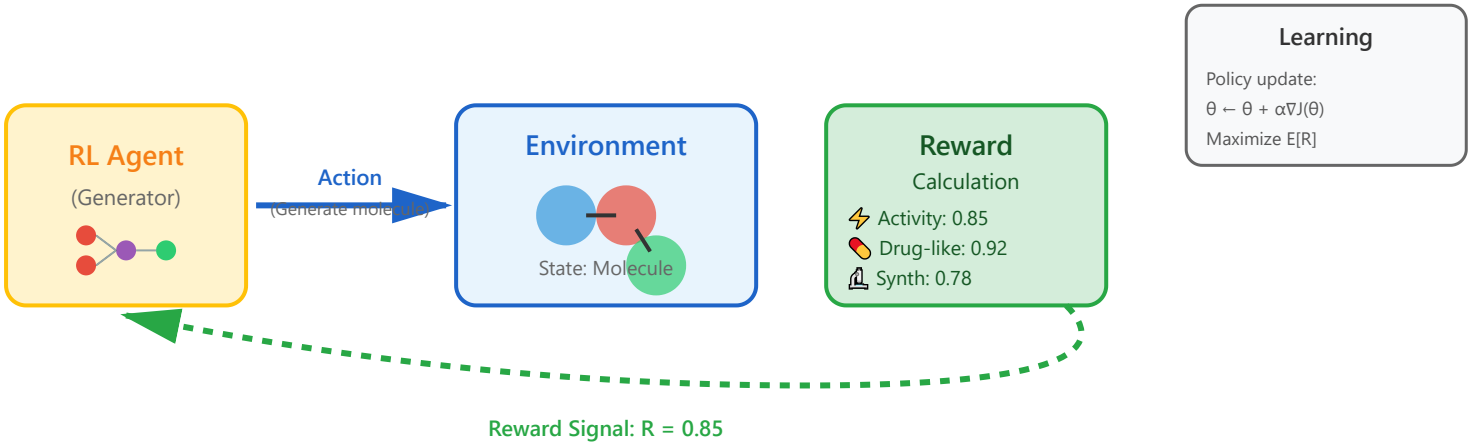
- **Latent space interpolation:** Navigate smoothly between known molecules to discover intermediate structures with novel properties
- **Scaffold hopping:** Replace molecular scaffolds while maintaining biological activity, leading to novel intellectual property
- **Property-guided sampling:** Direct generation toward regions with desired physicochemical or biological properties
- **Multi-objective optimization:** Simultaneously optimize multiple properties such as potency, selectivity, and ADMET characteristics

Molecular Representations

- **SMILES strings:** Sequential text representation enabling language model approaches
- **Molecular graphs:** Nodes represent atoms, edges represent bonds, capturing structural information
- **3D conformations:** Spatial arrangements critical for protein-ligand interactions
- **Fingerprints:** Binary vectors encoding structural features for similarity calculations

2. Reinforcement Learning

Reinforcement Learning (RL) enables goal-directed molecular design by training agents to generate molecules that maximize a reward function. The agent learns to navigate chemical space through trial and error, receiving rewards for generating molecules with desired properties.



Reward Function Design: The reward function combines multiple objectives including biological activity, synthetic accessibility, drug-likeness (Lipinski's rules), and novelty. Proper reward shaping is critical for successful optimization.

RL Algorithms for Molecular Design

- **Policy Gradient Methods:** REINFORCE algorithm optimizes the generator policy directly based on reward signals
- **Actor-Critic:** Combines policy optimization with value function estimation for more stable training
- **Proximal Policy Optimization (PPO):** Prevents large policy updates that could destabilize training
- **Monte Carlo Tree Search (MCTS):** Explores molecular construction paths systematically

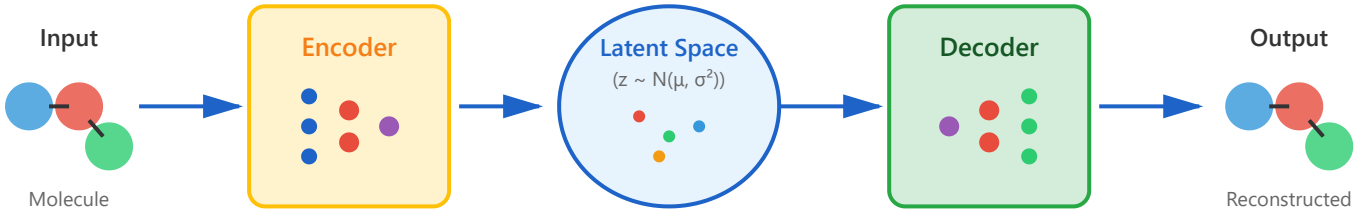
Reward Components

- **Target activity:** Predicted binding affinity or activity against target protein (0.0-1.0 scale)
- **Synthetic accessibility score (SA):** Estimates how difficult the molecule is to synthesize (1-10 scale)
- **QED (Quantitative Estimate of Drug-likeness):** Composite score of drug-like properties
- **Novelty bonus:** Rewards molecules that are structurally distinct from known compounds

3. VAE/GAN Approaches

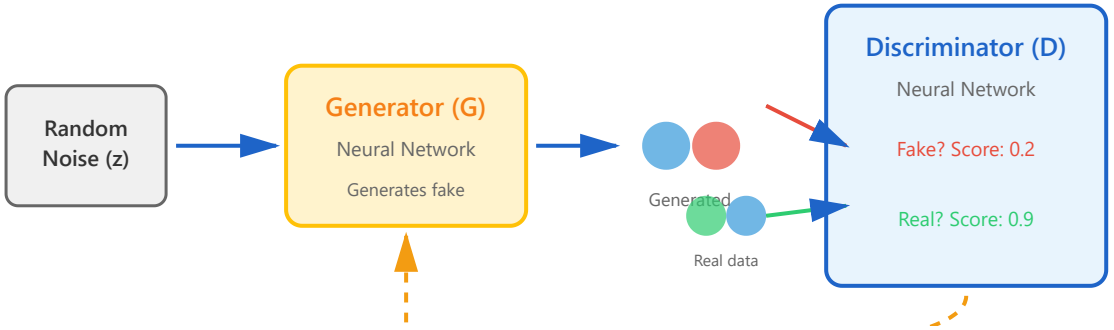
Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) are powerful generative architectures that learn to model the distribution of molecular structures and generate novel compounds with similar properties.

Variational Autoencoder (VAE)



VAE Loss Function: Combines reconstruction loss (how well the output matches input) with KL divergence (regularizes the latent space to follow a normal distribution): L = Reconstruction_Loss + β × KL_Divergence

Generative Adversarial Network (GAN)

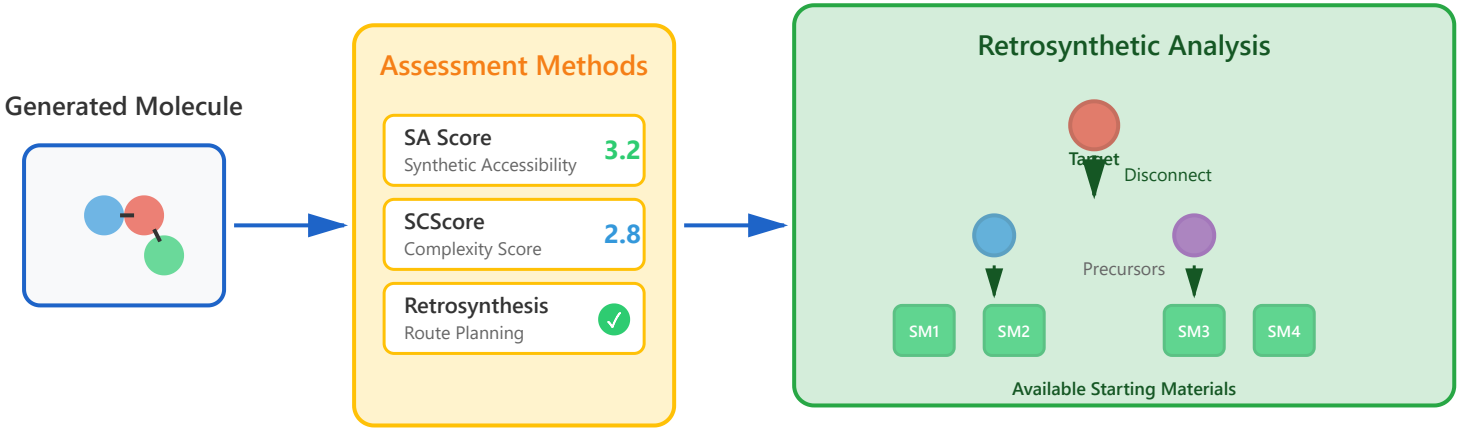


Key Advantages

- **VAE advantages:** Continuous latent space enables smooth interpolation, explicit probabilistic formulation, stable training dynamics
- **GAN advantages:** Can generate sharper, more realistic molecular structures, no need to explicitly define likelihood function
- **Conditional generation:** Both architectures can be conditioned on desired properties for targeted molecular design

4. Synthesizability

A crucial aspect of de novo drug design is ensuring that generated molecules can actually be synthesized in the laboratory. Synthesizability assessment evaluates whether a proposed molecule can be practically made using available chemical reactions and starting materials.



SA Score Range: Scores range from 1 (very easy to synthesize) to 10 (very difficult). Molecules with SA scores below 4 are generally considered synthetically accessible. The score considers molecular complexity, stereochemistry, and availability of building blocks.

Synthesizability Metrics

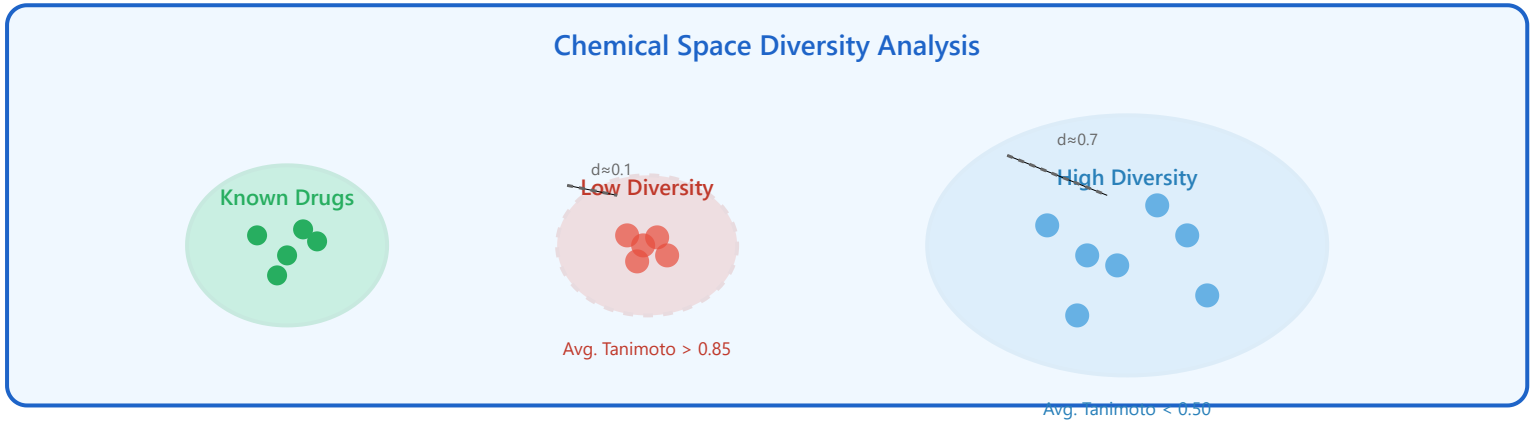
- **SA Score (Synthetic Accessibility):** Rule-based score considering fragment contributions and molecular complexity penalties
- **SCScore (Synthetic Complexity):** Machine learning model trained on reaction data to predict synthesis difficulty
- **RA Score (Retrosynthetic Accessibility):** Evaluates whether a valid retrosynthetic pathway can be found
- **RAscore:** Combines retrosynthetic analysis with availability of starting materials from chemical catalogs

Retrosynthetic Planning Tools

- **Computer-Aided Synthesis Planning (CASP):** Automated tools that propose synthesis routes by working backward from target to starting materials
- **Reaction template matching:** Identifies known chemical transformations applicable to the target molecule
- **Forward synthesis validation:** Verifies proposed routes by simulating forward reactions
- **Commercial availability check:** Ensures starting materials are purchasable from chemical suppliers

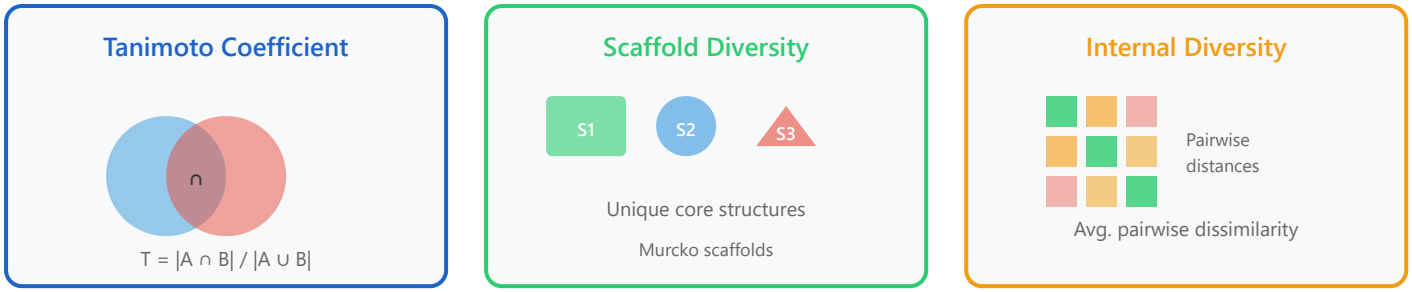
5. Diversity Metrics

Diversity metrics quantify the structural novelty and chemical variety of generated molecules. High diversity ensures exploration of different regions of chemical space while avoiding redundant similar structures. These metrics guide the generation process toward novel chemical matter.



Tanimoto Similarity: Measures structural similarity between two molecules based on fingerprint overlap. Values range from 0 (completely different) to 1 (identical). A Tanimoto coefficient below 0.5 typically indicates structurally distinct molecules.

Key Diversity Metrics



Diversity Calculations

- **Tanimoto coefficient:** Structural similarity based on molecular fingerprints (ECFP, MACCS keys). Lower values indicate greater diversity
- **Scaffold diversity:** Counts unique Murcko scaffolds (core ring systems) in generated set. Higher scaffold count indicates greater chemical diversity
- **Internal diversity:** Average pairwise dissimilarity within generated set. ID = (1/N(N-1)) Σ(1 - Tanimoto_ij)
- **Novelty score:** Minimum Tanimoto similarity to nearest neighbor in reference database. High novelty indicates genuinely new structures

Applications in De Novo Design

- **Diversity penalty in loss function:** Encourages generator to produce structurally distinct molecules rather than minor variations
- **Batch diversity optimization:** Generate sets of molecules that collectively explore different chemical regions
- **Novelty-guided search:** Prioritize regions of chemical space distant from known compounds
- **Scaffold hopping strategies:** Systematically replace core structures while maintaining activity to discover new intellectual property