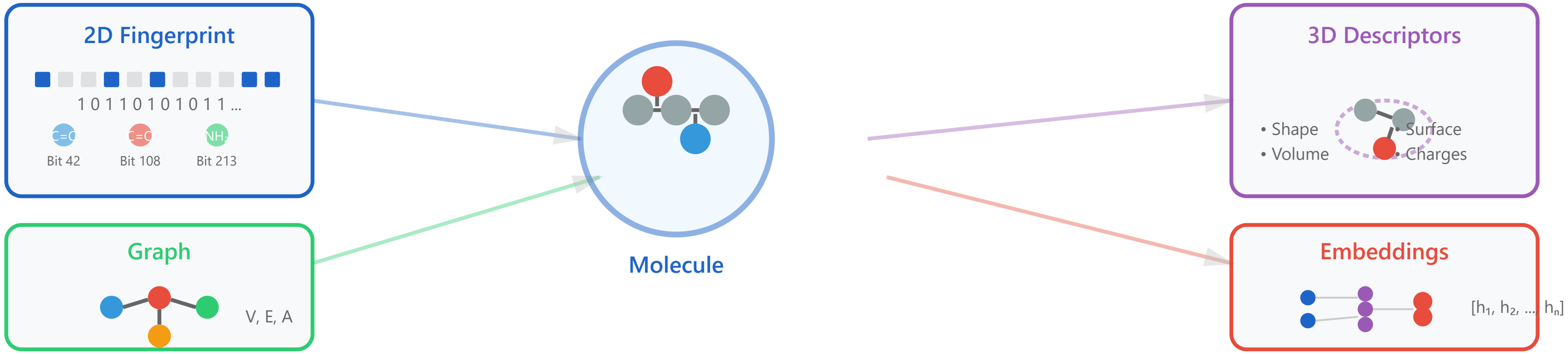
Molecular Representations



2D fingerprints

Binary feature vectors

3D descriptors

Geometric and conformational features

Graph representations

Molecular graph structures

Learned embeddings

Deep learning representations

Multi-view learning

Combining multiple representations

1 2D Molecular Fingerprints

Molecular fingerprints are fixed-length binary vectors that encode the presence or absence of specific structural features in a molecule. Each bit in the fingerprint corresponds to a particular substructure or molecular pattern.

Key Concepts

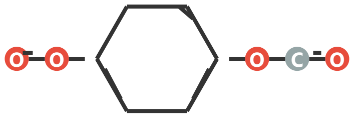
- **Hashing:** Chemical substructures are mapped to specific bit positions using hash functions
- **Fixed Length:** Typical fingerprint sizes range from 512 to 2048 bits
- **Collision:** Multiple substructures may hash to the same bit position
- **Similarity:** Tanimoto coefficient measures molecular similarity

Common Types

- **ECFP (Extended Connectivity FP):** Circular fingerprints based on atom neighborhoods
- **MACCS Keys:** 166 predefined structural keys
- **Daylight:** Path-based fingerprints
- **RDKit:** Open-source implementations

> **Use Cases:** Virtual screening, similarity search, compound clustering, QSAR modeling

ECFP4 EXAMPLE - ASPIRIN



Aspirin (Acetylsalicylic acid)

Generated Fingerprint (2048 bits):

■□■■□■□□■■ ... □■□

Captured Features:
- Benzene ring → Bits 42, 127, 891
- Carboxyl group → Bits 234, 1023

Properties:
- 148 bits set to 1
- 1900 bits set to 0

Similarity Calculation:

Tanimoto = |A ∩ B| / |A ∪ B|

Measures overlap between fingerprints

2 3D Molecular Descriptors

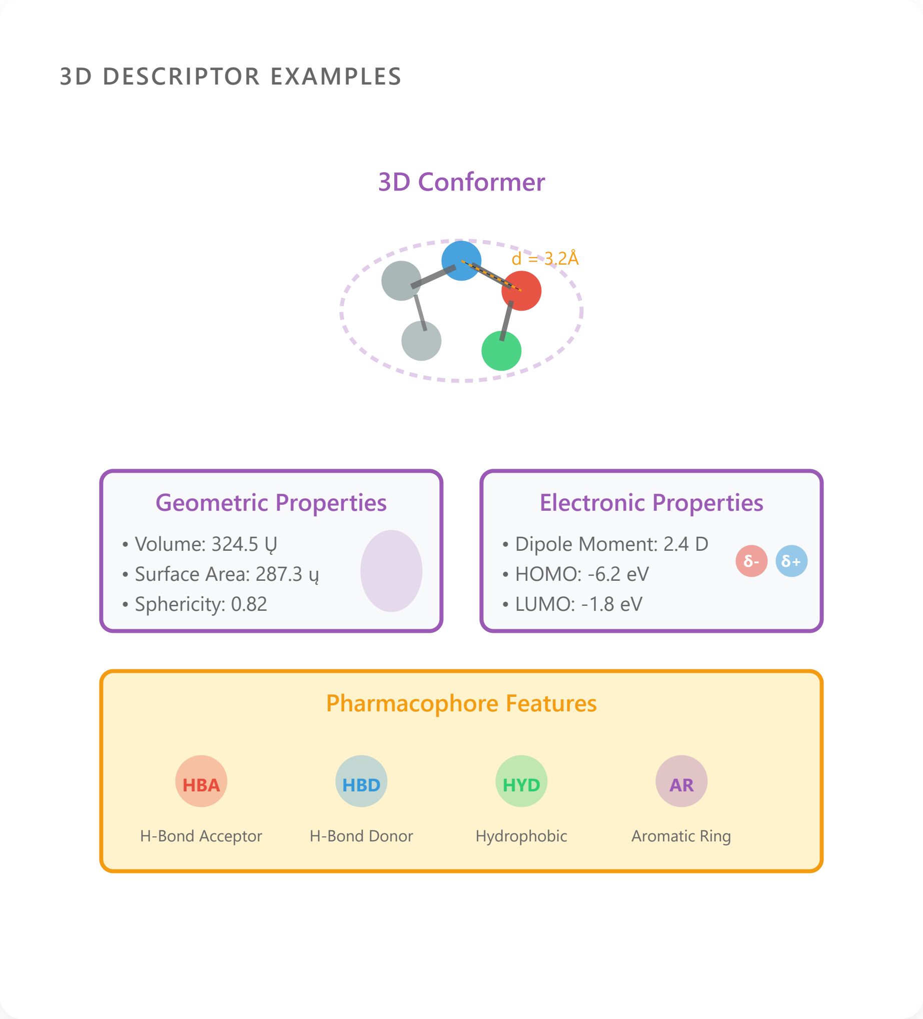
Three-dimensional descriptors capture the spatial arrangement and geometric properties of molecules. These representations account for molecular conformations, which are critical for understanding biological activity and molecular interactions.

Key Features

- **Geometric Properties:** Molecular shape, volume, and surface area
- **Electronic Features:** Partial charges, electrostatic potentials
- **Pharmacophoric:** Spatial arrangement of functional groups
- **Conformation-Dependent:** Properties vary with 3D structure

Types of 3D Descriptors

- **Shape-based:** Molecular volume, surface area, principal moments
- **Field-based:** Molecular electrostatic potential (MEP)
- **Pharmacophore:** Distance geometry, spatial patterns
- **Surface properties:** Hydrophobic/hydrophilic regions



3D DESCRIPTOR EXAMPLES

3D Conformer

Geometric Properties

- Volume: 324.5 Ų
- Surface Area: 287.3 ų
- Sphericity: 0.82

Electronic Properties

- Dipole Moment: 2.4 D
- HOMO: -6.2 eV
- LUMO: -1.8 eV

Pharmacophore Features

H-Bond Acceptor
H-Bond Donor
Hydrophobic
Aromatic Ring

Use Cases: Protein-ligand docking, structure-based drug design, conformer analysis, binding affinity prediction

3 Graph Representations

Graph representations treat molecules as mathematical graphs where atoms are nodes and bonds are edges. This natural representation preserves the connectivity and topological structure of molecules, making it ideal for graph neural networks.

GRAPH REPRESENTATION - ETHANOL (CH_3CH_2OH)

Graph Components

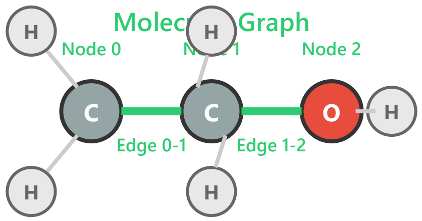
- **Nodes (V):** Atoms with features (element type, charge, hybridization)
- **Edges (E):** Bonds with attributes (bond type, stereochemistry)
- **Adjacency Matrix (A):** Connectivity information
- **Node Features (X):** Atomic properties matrix

Advantages

- **Permutation Invariant:** Same molecule regardless of atom ordering

- **Size Flexible:** Handles molecules of varying sizes
- **Interpretable:** Clear mapping to chemical structure
- **GNN Compatible:** Direct input for graph neural networks

> **Use Cases:** Graph neural networks (GNNs), message passing, molecular property prediction, reaction prediction, retrosynthesis



Adjacency Matrix (A)

	C_0	C_1	O_2	H...
C_0	0	1	0	1
C_1	🟩	⬜	🟩	🟩
O_2	⬜	🟩	⬜	🟩
H...				

Node Features (X)

Node 0 (C):

- Atomic number: 6
- Degree: 4
- Hybridization: sp^3

Node 1 (C):

- Atomic number: 6
- Degree: 4

Graph Neural Network Processing

$h^{(t+1)}_i = \text{UPDATE}(h^{(t)}_i, \text{AGGREGATE}(\{h^{(t)}_j : j \in N(i)\}))$

Nodes aggregate information from neighbors

4 Learned Embeddings

Learned embeddings are continuous vector representations automatically learned by neural networks through training on molecular data. Unlike hand-crafted features, these embeddings capture complex patterns and relationships in a data-driven manner.

NEURAL NETWORK EMBEDDING PIPELINE

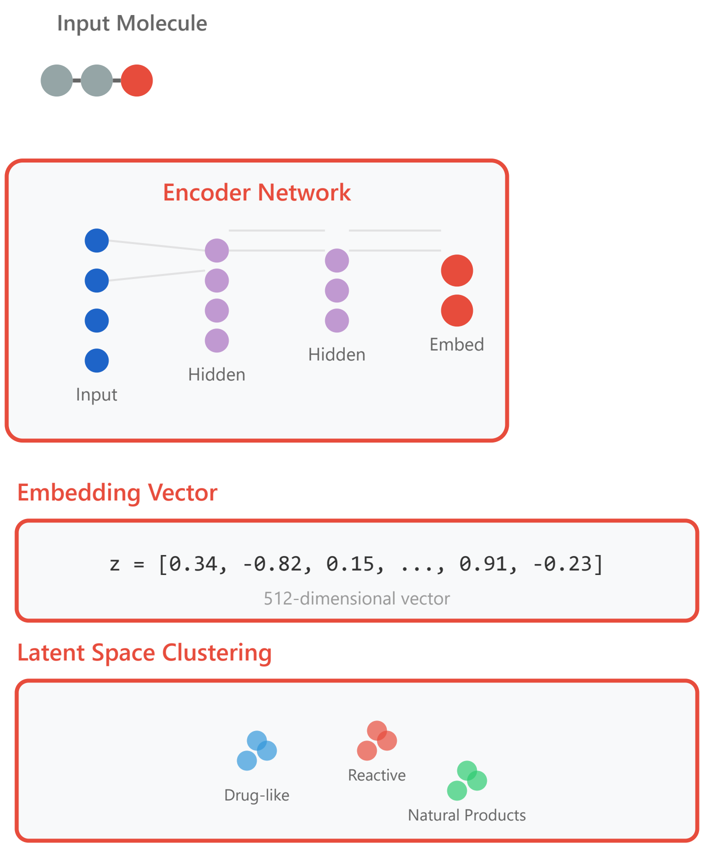
Key Concepts

- **Dense Vectors:** Continuous-valued representations (e.g., 128-512 dimensions)
- **Learned Features:** Automatically discovered patterns from data
- **Task-Specific:** Optimized for particular prediction objectives
- **Semantic Meaning:** Similar molecules have similar embeddings

Common Architectures

- **VAE:** Variational autoencoders for generative modeling
- **GNN:** Graph neural network node/graph embeddings
- **Transformers:** Self-attention based molecular encoders
- **Pre-trained Models:** ChemBERTa, MolBERT, UniMol

> **Use Cases:** Transfer learning, molecular generation, similarity search in latent space, multi-task learning, zero-shot prediction



Input Molecule

Encoder Network

Input

Hidden

Hidden

Embed

Embedding Vector

```
z = [0.34, -0.82, 0.15, ..., 0.91, -0.23]
```

512-dimensional vector

Latent Space Clustering

5 Multi-view Learning

Multi-view learning combines multiple molecular representations to leverage complementary information from different perspectives. This approach recognizes that no single representation captures all relevant molecular properties.

Integration Strategies

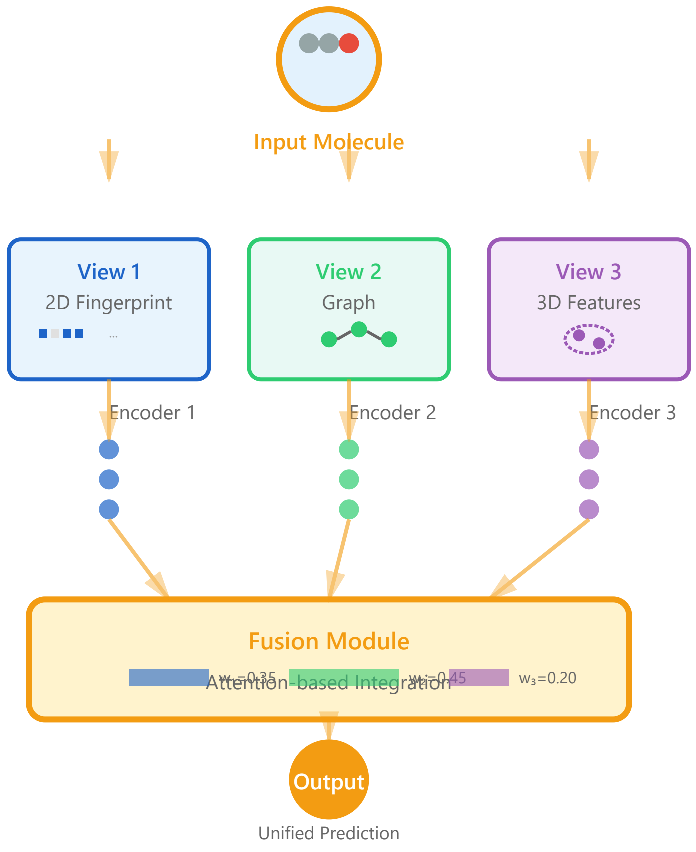
- **Early Fusion:** Concatenate features before model input
- **Late Fusion:** Combine predictions from separate models
- **Intermediate Fusion:** Merge representations at hidden layers
- **Attention-based:** Learn optimal weighting of views

Common View Combinations

- **2D + 3D:** Topology and conformational information
- **Graph + Fingerprint:** Structure and patterns
- **Sequence + Structure:** SMILES and spatial features
- **Multiple Conformers:** Ensemble of 3D structures

> **Use Cases:** Robust property prediction, improved generalization, handling missing modalities, cross-modal retrieval, comprehensive molecular understanding

MULTI-VIEW INTEGRATION ARCHITECTURE



> **Key Advantage:** Each view captures different molecular aspects—topology, geometry, and chemical patterns—creating a comprehensive representation that outperforms single-view approaches for complex prediction tasks.