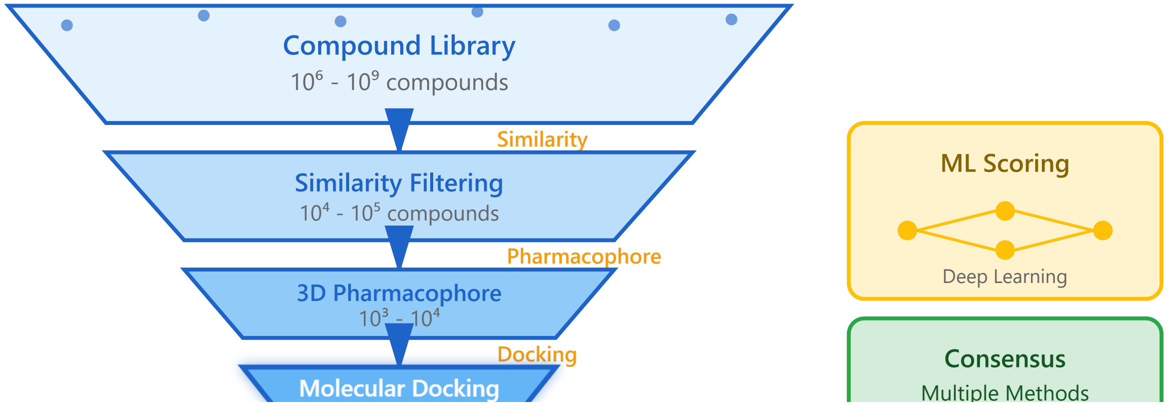
Virtual Screening



Similarity searching
Finding similar active compounds

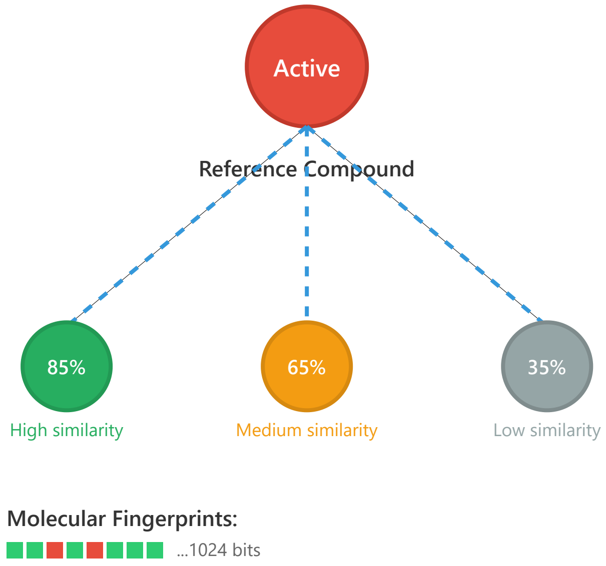
Pharmacophore modeling
3D feature-based screening

Docking scores
Protein-ligand binding prediction

ML scoring functions
Learning-based scoring

Consensus approaches
Combining multiple methods

1. Similarity Searching



- **Principle:** Compounds with similar structures tend to have similar biological activities (Similar Property Principle)
- **Method:** Compare molecular fingerprints using Tanimoto coefficient or other similarity metrics
- **Speed:** Very fast - can screen millions of compounds in minutes
- **Input required:** One or more known active compounds

Typical Workflow

Generate fingerprints → Calculate similarity scores → Rank compounds → Select top candidates (typically Tanimoto > 0.7)

Common Fingerprints

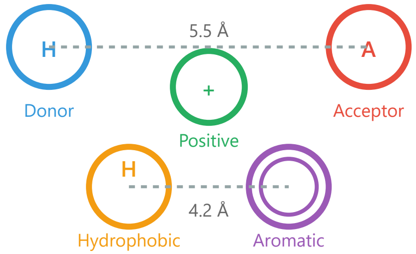
- ECFP (Extended Connectivity Fingerprints)
- MACCS keys (166-bit structural keys)
- Atom pairs and topological torsions

Key Advantages & Limitations

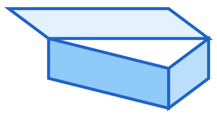
✓ Pros: Extremely fast, simple to implement, good for scaffold hopping

✓ Cons: 2D only (no 3D conformational info), may miss structurally diverse actives

2. Pharmacophore Modeling



3D Spatial Arrangement



Common Features:
- H-bond donor
- H-bond acceptor
- Hydrophobic
- Aromatic

- **Principle:** Identifies essential 3D chemical features required for biological activity
- **Generation:** Can be ligand-based (from active compounds) or structure-based (from protein-ligand complex)
- **Features:** H-bond donors/acceptors, hydrophobic centers, aromatic rings, charged groups
- **Constraints:** Spatial distances and angles between features

Screening Process

Generate conformers → Map features → Check spatial constraints → Score matches → Filter hits

Software Tools

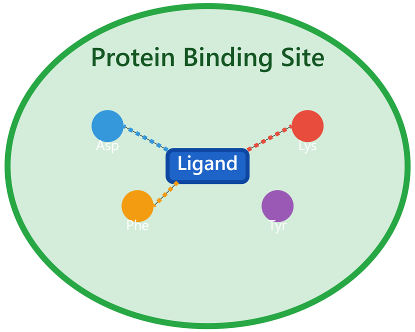
• LigandScout (structure-based)
• Phase (Schrödinger)
• Discovery Studio CATALYST

Key Advantages & Limitations

✓ Pros: Captures 3D information, allows scaffold hopping, interpretable results

✓ Cons: Computationally intensive, requires conformer generation, sensitive to feature selection

3. Molecular Docking



Scoring Function

$\Delta G = \Delta G\_vdW$	van der Waals
$+ \Delta G\_elec$	Electrostatic
$+ \Delta G\_hbond$	H-bonding

- **Principle:** Predicts the binding mode and affinity of small molecules to protein targets
- **Components:** Search algorithm (pose generation) + Scoring function (affinity estimation)
- **Search algorithms:** Genetic algorithms, Monte Carlo, incremental construction
- **Scoring:** Force field-based, empirical, or knowledge-based functions

Docking Protocol

Prepare protein & ligands → Define binding site → Generate poses → Score and rank → Analyze interactions

Popular Programs

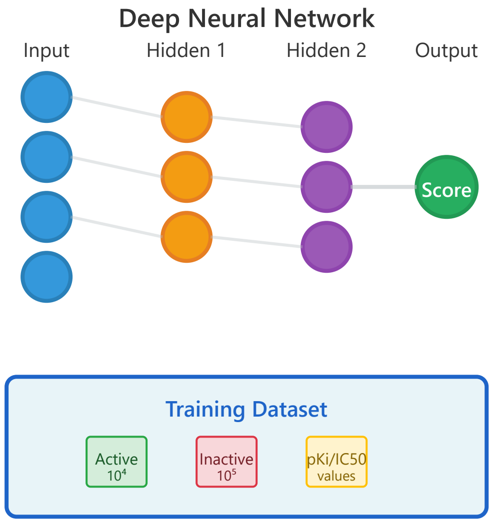
• AutoDock Vina (open source)
• Glide (Schrödinger)
• GOLD, DOCK, FlexX

Key Advantages & Limitations

✓ Pros: Structure-based, provides binding mode, widely validated, considers protein flexibility

✓ Cons: Computationally expensive, accuracy depends on scoring function, protein flexibility challenges

4. Machine Learning Scoring Functions



Input Features:

- Protein-ligand interaction fingerprints
- 3D molecular descriptors

- **Principle:** Learn complex patterns from experimental binding data using machine learning models
- **Architectures:** Random Forest, Gradient Boosting, Deep Neural Networks, Graph Neural Networks
- **Features:** Molecular descriptors, interaction fingerprints, 3D coordinates, graph representations
- **Training:** Requires large datasets with binding affinity measurements

ML Workflow

Collect data → Extract features → Train model → Validate → Apply to screening → Post-process predictions

State-of-the-art Models

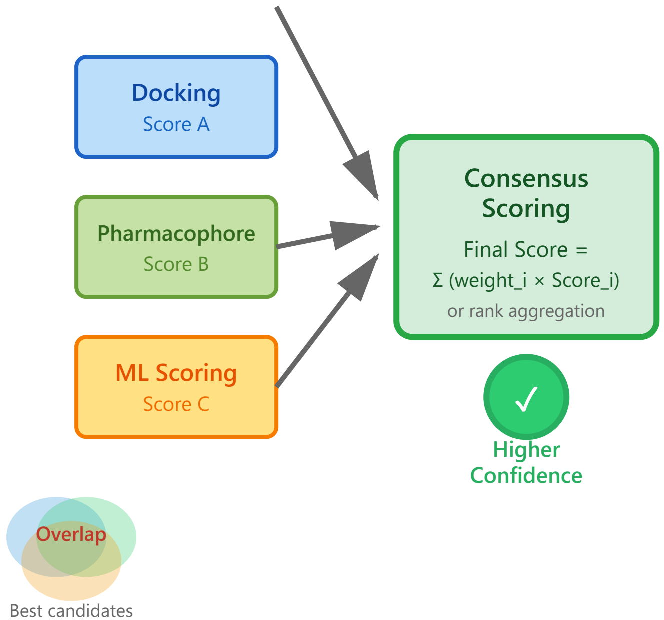
- DeepDTA (binding affinity)
- KDEEP, OnionNet
- Graph neural networks (GAT, GCN)

Key Advantages & Limitations

✓ Pros: Learns from data, often outperforms classical scoring, can capture complex patterns

✓ Cons: Requires large training datasets, potential overfitting, interpretability challenges

5. Consensus Approaches



- **Principle:** Combines predictions from multiple independent methods to improve accuracy and reduce false positives
- **Strategies:** Rank-by-rank, score-by-score, voting schemes, machine learning ensembles
- **Rationale:** Different methods have complementary strengths and weaknesses
- **Result:** Higher enrichment of true positives in top-ranked compounds

Implementation Approaches

1. Intersection: Select compounds ranked high by ALL methods
2. Weighted scoring: Combine scores with optimized weights
3. Rank aggregation: Merge ranking lists

Common Combinations

- Docking + Pharmacophore
- Multiple docking programs
- Classical + ML scoring

Key Advantages & Limitations

✓ Pros: Improved accuracy, reduces method-specific biases, more robust predictions

✓ Cons: Computationally expensive (multiple methods), requires careful weight optimization

Virtual Screening: Summary & Best Practices

Method	Speed	Accuracy	3D Info	Best Use Case
Similarity	★★★★★	★★★☆☆	✗	Large library screening
Pharmacophore	★★★☆☆	★★★★☆	✓	Feature-based filtering
Docking	★★☆☆☆	★★★★☆	✓✓	Structure-based screening

Recommended Workflow

- ✓ Stage 1: Similarity filtering (fast pre-filter)
- ✓ Stage 2: Pharmacophore screening (3D constraints)
- ✓ Stage 3: Molecular docking (binding mode)
- ✓ Stage 4: Consensus scoring + visual inspection

Critical Success Factors

- ✓ Quality of input structures (protein & ligands)
- ✓ Appropriate method selection for target
- ✓ Validation with known actives/inactives
- ✓ Experimental validation of predictions

Performance Metrics

Enrichment Factor (EF): Measures how well actives are enriched in top-ranked compounds
ROC-AUC: Overall discriminatory power between actives and decoys
Success Rate: Typical hit rates range from 1-10% depending on target and method