Lecture 10:

Drug Discovery and Molecular ML

- AI-powered drug discovery
- Success stories
- Pipeline transformation

Introduction to Biomedical Datascience

Lecture 10:

Drug Discovery and Molecular ML

AI-powered drug discovery

Success stories

Pipeline transformation

Introduction to Biomedical Datascience

Lecture Contents

Part 1: Drug Discovery Pipeline

Part 2: Molecular Machine Learning

Part 3: Practical Applications

Part 1/3

Drug Discovery Pipeline

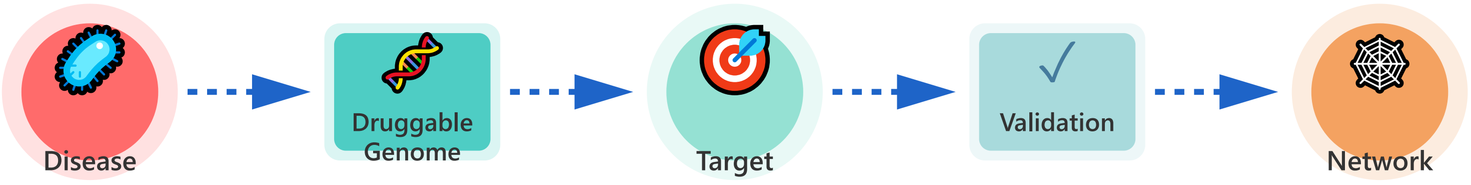
- Traditional vs AI-enhanced
- Time and cost savings
- Success rate improvements

Part 1/3:

Drug Discovery Pipeline

- Traditional vs AI-enhanced approaches
- Time and cost savings
- Success rate improvements

Target Identification



Disease mechanisms

Understanding biological pathways

Druggable genome

Identifying targetable proteins

Target validation

Confirming therapeutic relevance

Genetic evidence

Human genetics support

Network approaches

Systems biology integration

Lead Discovery

High-throughput screening

Automated testing of compounds

Virtual screening

Computational compound filtering

Fragment-based design

Building from molecular fragments

Natural products

Nature-inspired compounds

Diversity libraries

Chemical space exploration

Lead Optimization

SAR analysis

Structure-activity relationships

ADMET optimization

Pharmacokinetic properties

Selectivity improvement

Reducing off-target effects

Patent space

IP landscape navigation

Multi-parameter optimization

Balancing multiple objectives

Preclinical Studies

In vitro assays

Cell-based testing

Animal models

In vivo efficacy testing

Toxicology studies

Safety assessment

PK/PD modeling

Pharmacokinetic/pharmacodynamic

IND preparation

Regulatory submission readiness

Clinical Trials

Phase I-III design

Human testing stages

Biomarker strategies

Patient selection & monitoring

Adaptive trials

Flexible trial designs

Real-world evidence

Post-market data collection

Regulatory submission

FDA/EMA approval process

Computational Approaches

Structure-based design

Protein structure utilization

Ligand-based design

Known active compound patterns

Systems pharmacology

Network-based approaches

ML integration

AI-powered prediction

Quantum computing

Next-generation simulations

Part 2/3

Molecular ML

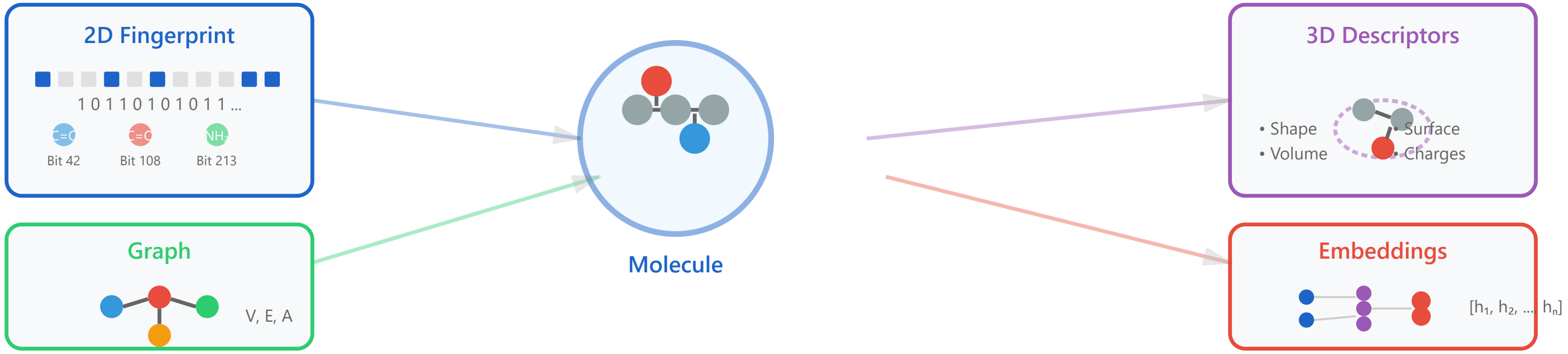
- Representation learning
- Property prediction
- Generative models

Part 2/3:

Molecular Machine Learning

- Representation learning
- Property prediction
- Generative models

Molecular Representations



2D fingerprints

Binary feature vectors

3D descriptors

Geometric and conformational features

Graph representations

Molecular graph structures

Learned embeddings

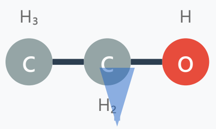
Deep learning representations

Multi-view learning

Combining multiple representations

SMILES Notation

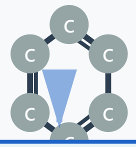
Ethanol



CCO

SMILES String

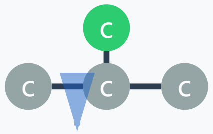
Benzene



c1ccccc1

Ring Closure

Branched



CC(C)C

Branching ()

Syntax rules

String-based molecular encoding

Canonical SMILES

Unique molecular representation

SMARTS patterns

Substructure search patterns

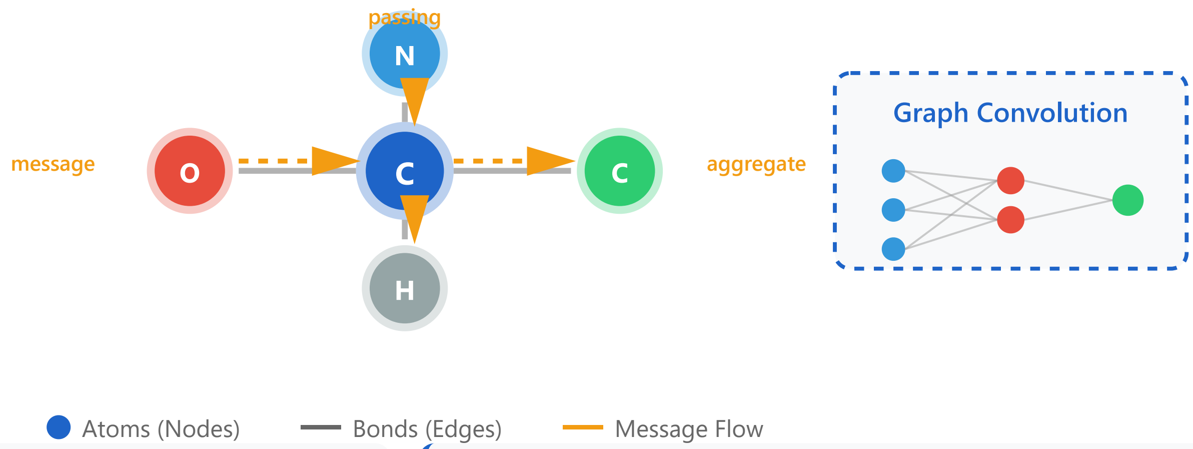
Tokenization

Breaking into meaningful units

Augmentation strategies

Data augmentation techniques

Graph Neural Networks



Molecular graphs

Atoms as nodes, bonds as edges

Message passing

Information flow between atoms

Graph convolutions

Feature aggregation operations

Attention mechanisms

Weighted information aggregation

Pooling strategies

Graph-level representations

Property Prediction

Regression tasks

Continuous property prediction

Classification tasks

Binary and multi-class prediction

Multi-task learning

Joint prediction of properties

Uncertainty quantification

Prediction confidence

Domain adaptation

Transfer across datasets

QSAR Modeling

Descriptor selection

Feature engineering and selection

Model validation

Cross-validation strategies

Applicability domain

Model reliability assessment

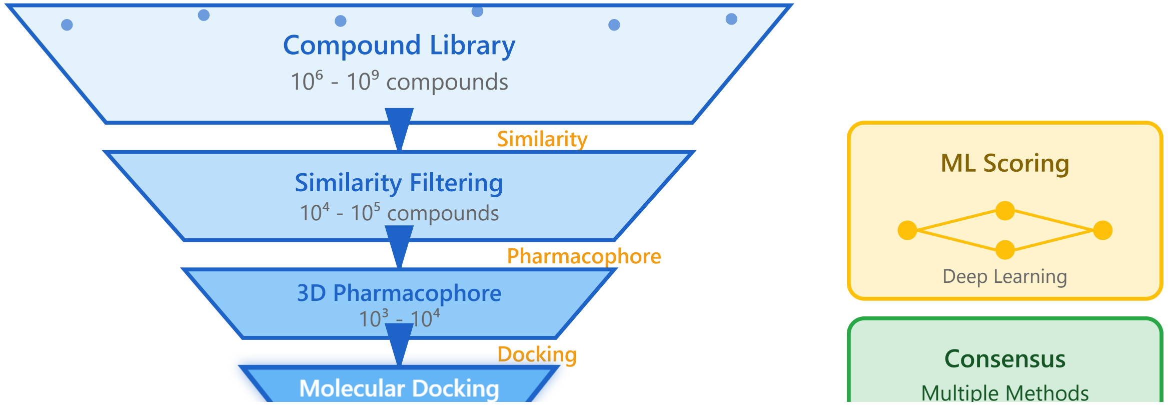
Y-randomization

Statistical significance testing

OECD principles

Regulatory compliance

Virtual Screening



Similarity searching

Finding similar active compounds

Pharmacophore modeling

3D feature-based screening

Docking scores

Protein-ligand binding prediction

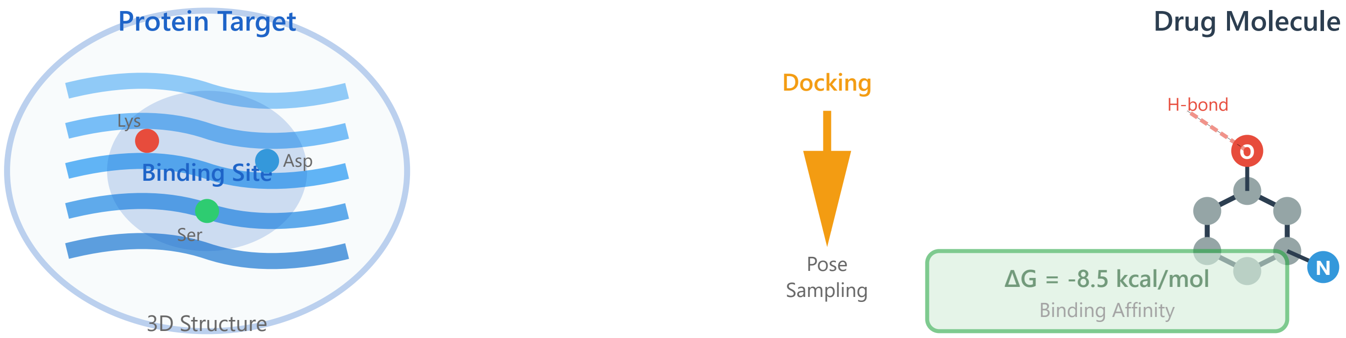
ML scoring functions

Learning-based scoring

Consensus approaches

Combining multiple methods

Docking Simulation



Scoring: vdW + Electrostatic + H-bonds + Solvation + Entropy

Protein preparation

Structure optimization

Binding site detection

Active site identification

Conformational sampling

Exploring binding modes

Scoring functions

Binding affinity estimation

Induced fit

Protein flexibility modeling

Part 3/3

Applications

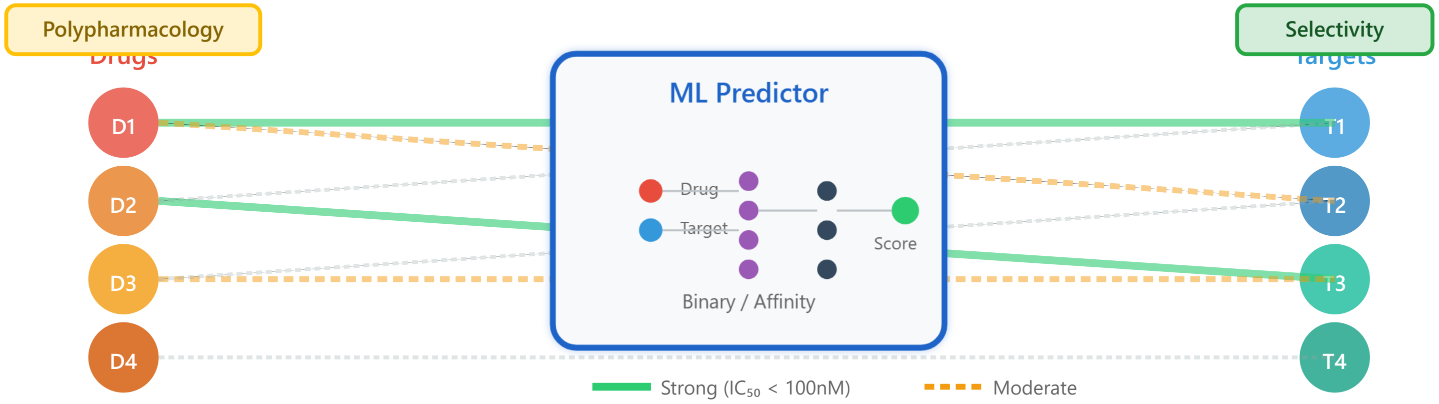
- Practical implementations
- Success metrics
- Future directions

Part 3/3:

Practical Applications

- Practical implementations
- Success metrics
- Future directions

Drug-Target Interaction



Binary classification

Predicting interaction likelihood

Binding affinity

Quantitative affinity prediction

Kinome profiling

Kinase selectivity analysis

Polypharmacology

Multi-target interactions

Off-target prediction

Safety profiling

Side Effect Prediction

ADR databases

Adverse drug reaction resources

Network approaches

Drug-target-disease networks

Chemical similarity

Structure-based prediction

Target-based

Mechanism-based approaches

Clinical translation

Preclinical to clinical

Drug Repurposing

Indication expansion

New therapeutic uses

Signature matching

Disease signature comparison

Network propagation

Disease module identification

Clinical evidence

Real-world validation

IP considerations

Patent and exclusivity

Bioactivity Prediction

Activity cliffs

Small structural changes, large activity differences

Matched pairs

Systematic SAR analysis

Free energy perturbation

Physics-based predictions

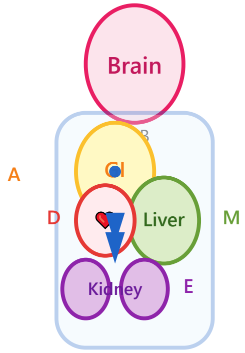
Active learning

Iterative experiment design

Experimental validation

Wet-lab confirmation

ADMET Prediction



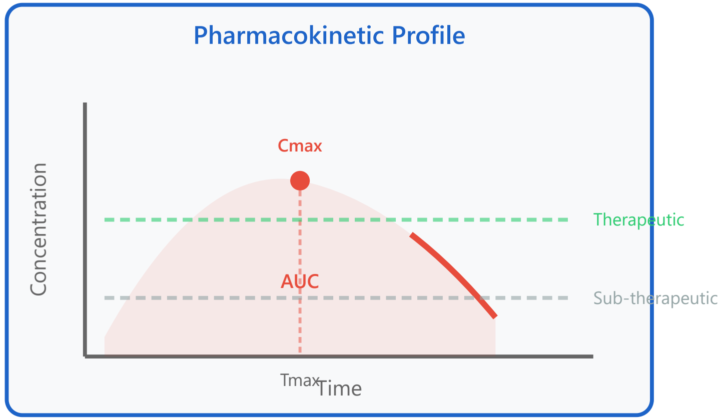
Absorption
Caco-2, LogP

Distribution
Vd, PPB

Metabolism
CYP450

Excretion
Clearance

Toxicity
hERG, LD50


Pharmacokinetic Profile

Absorption models

Oral bioavailability prediction

Distribution (BBB, Vd)

Tissue distribution modeling

Metabolism (CYP)

Drug metabolism prediction

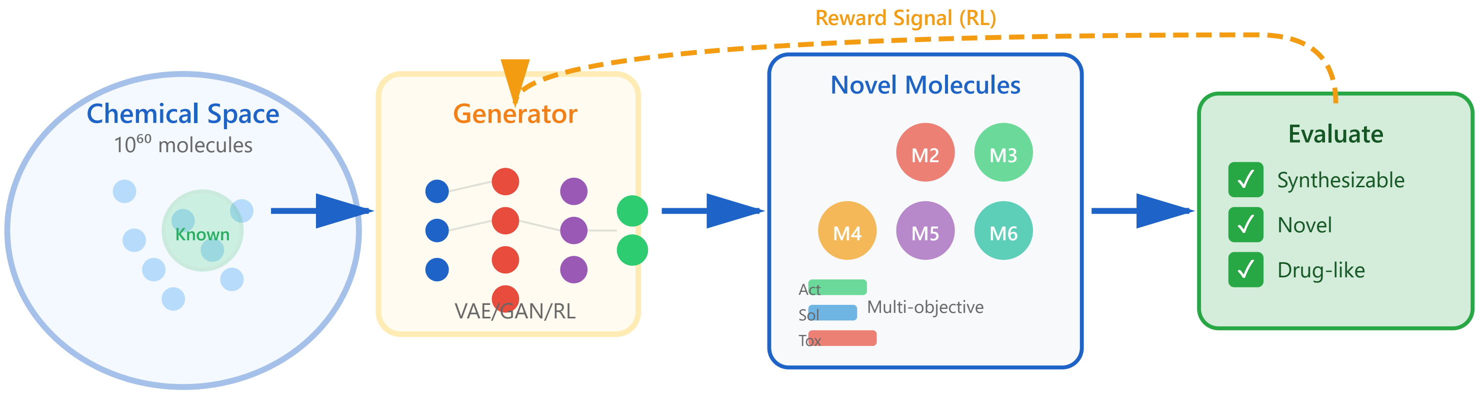
Excretion (clearance)

Elimination pathway modeling

Toxicity endpoints

Safety assessment

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

Generative Models

SMILES generation

String-based generation

Graph generation

Graph-based generation

3D molecule generation

Conformer generation

Conditional generation

Property-guided generation

Multi-objective optimization

Balancing multiple properties

Clinical Trial Optimization

Patient selection

Identifying responders

Dose finding

Optimal dosing strategies

Endpoint prediction

Trial outcome forecasting

Site selection

Geographic optimization

Recruitment optimization

Accelerating enrollment

Pharmacovigilance

Signal detection

Identifying safety signals

Causality assessment

Determining drug-event relationships

Risk-benefit analysis

Therapeutic decision support

Literature mining

Automated safety surveillance

Social media monitoring

Real-time safety signals

Hands-on: RDKit and DeepChem

Molecule manipulation

Reading and writing structures

Descriptor calculation

Computing molecular features

Model training

Building predictive models

Scaffold splitting

Dataset partitioning strategies

Performance evaluation

Metrics and validation

💻 HANDS-ON

RDKit and DeepChem

Molecule manipulation

Loading, parsing, and modifying molecular structures

Descriptor calculation

Computing physicochemical properties and fingerprints

Model training

Building predictive models with DeepChem framework

Scaffold splitting

Creating train/test splits based on molecular scaffolds

Performance evaluation

Assessing model accuracy using appropriate metrics

```
# Example: RDKit & DeepChem workflow
from rdkit import Chem
import deepchem as dc

# Load molecules and compute descriptors
featurizer = dc.feat.CircularFingerprint()
loader = dc.data.CSVLoader(tasks=['activity'], featurizer=featurizer)
```

Hands-on: Molecular Generation

SMILES RNN

Recurrent neural network generation

Graph VAE

Variational autoencoder for graphs

Reinforcement learning

Policy-based optimization

Property optimization

Multi-objective design

Diversity analysis

Chemical space coverage

💻 HANDS-ON

Molecular Generation

SMILES RNN

Recurrent neural networks for sequential generation

Graph VAE

Variational autoencoders for graph-based generation

Reinforcement learning

Policy-based optimization for desired properties

Property optimization

Guiding generation toward specific target profiles

Diversity analysis

Measuring chemical diversity in generated libraries

```
# Example: Generative model workflow
from rdkit import Chem
import torch

# Load pre-trained generative model
model = MolecularRNN.load('pretrained_model.pt')
generated_smiles = model.sample(n_molecules=100)
```

Thank You

- Approved AI-discovered drugs
- Pipeline statistics & success rates
- Investment trends in AI drug discovery
- Future outlook & opportunities

Introduction to Biomedical Datascience