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 1/3:

Drug Discovery Pipeline

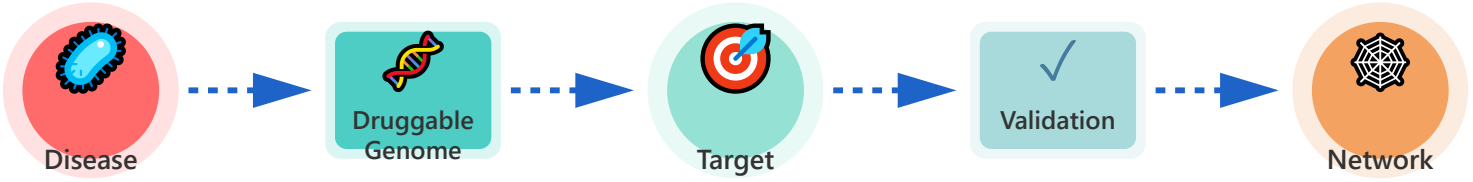
- Traditional vs AI-enhanced approaches
- Time and cost savings
- Success rate improvements

Part 1/3

Drug Discovery Pipeline

- Traditional vs AI-enhanced
- 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

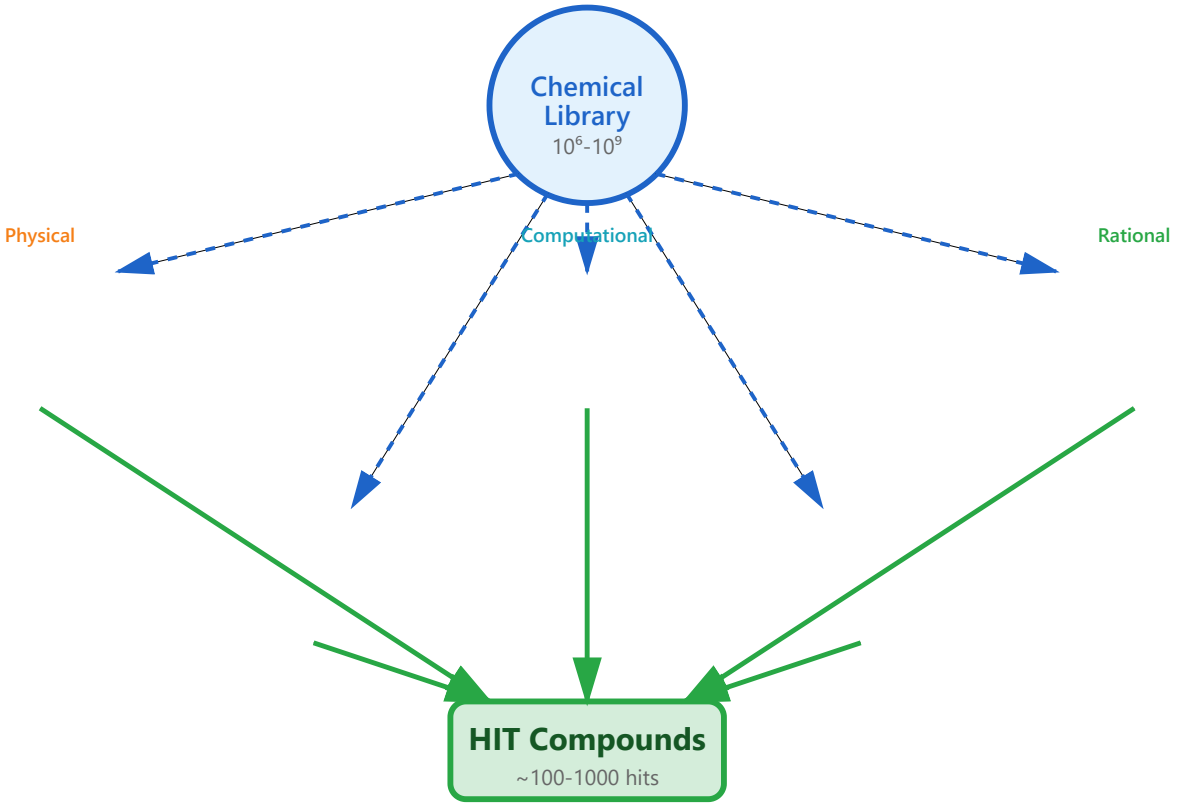
Human genetics support

Network approaches

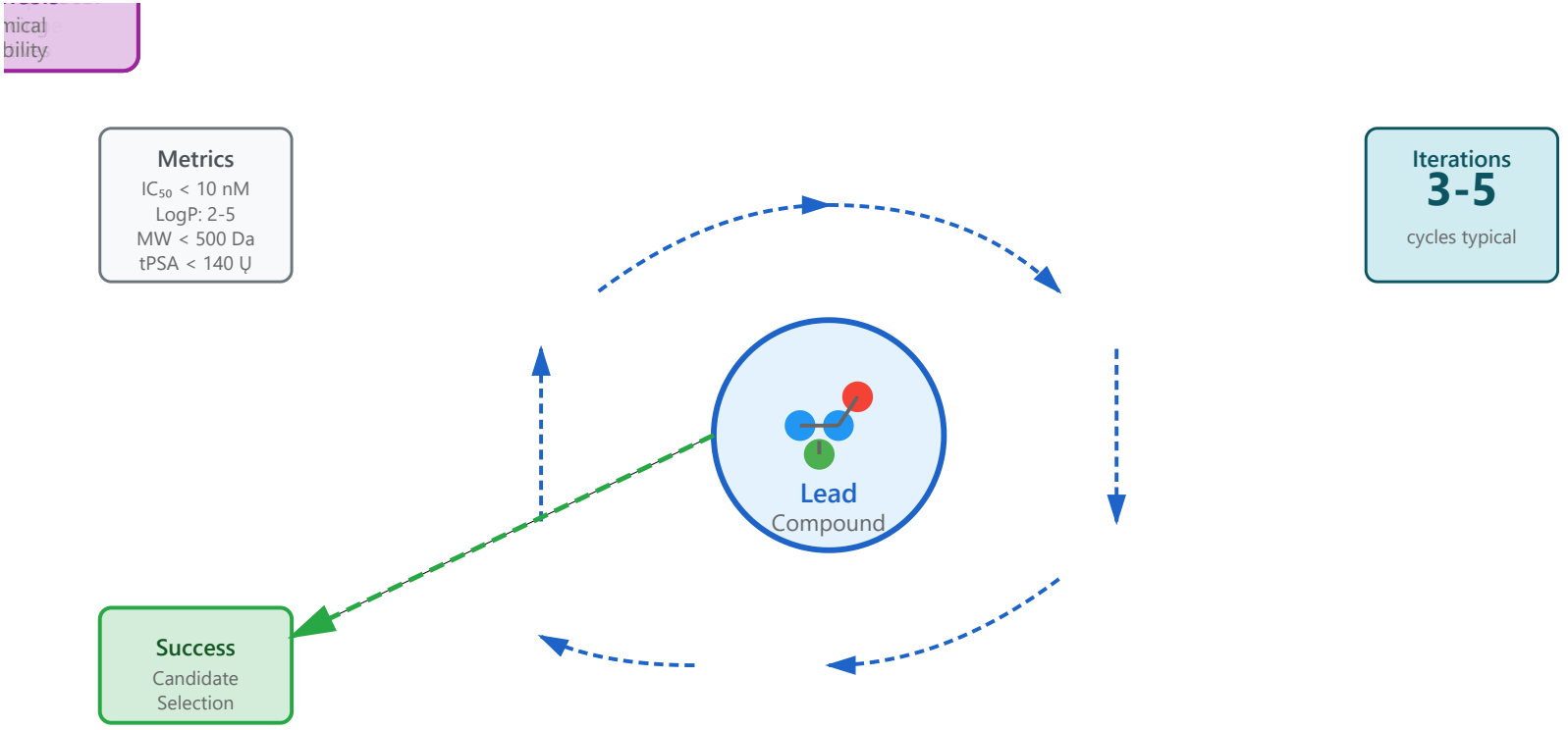
Systems biology integration

Lead Discovery

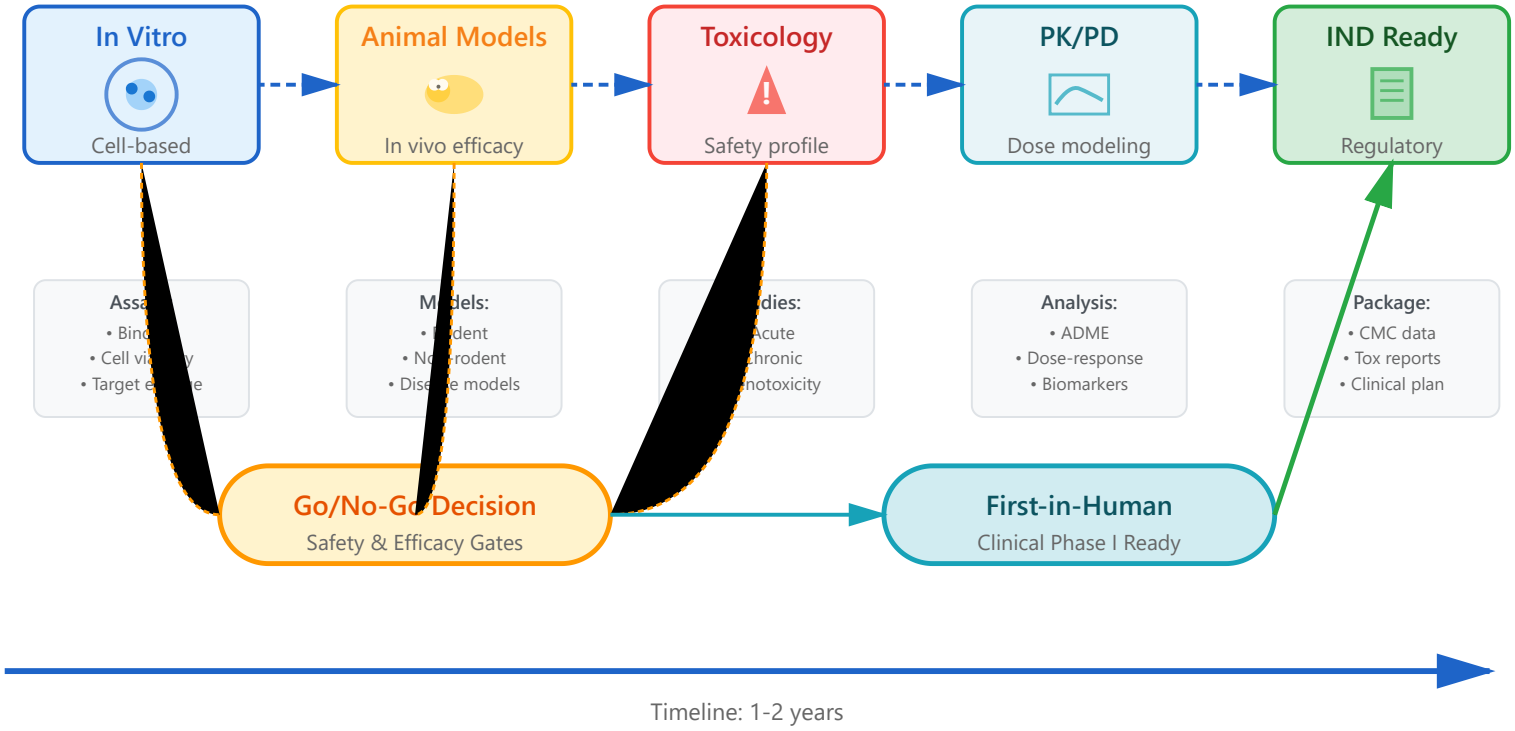
aries
al space



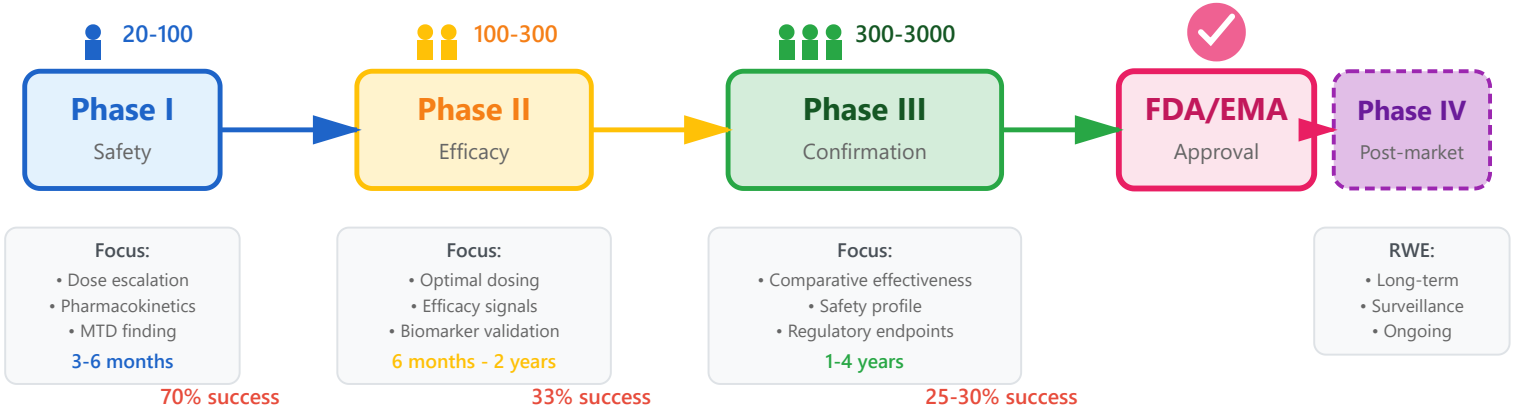
Lead Optimization



Preclinical Studies



Clinical Trials



Biomarker Strategies

- Patient selection
- Response monitoring
- Surrogate endpoints
- Precision medicine

Adaptive Trials

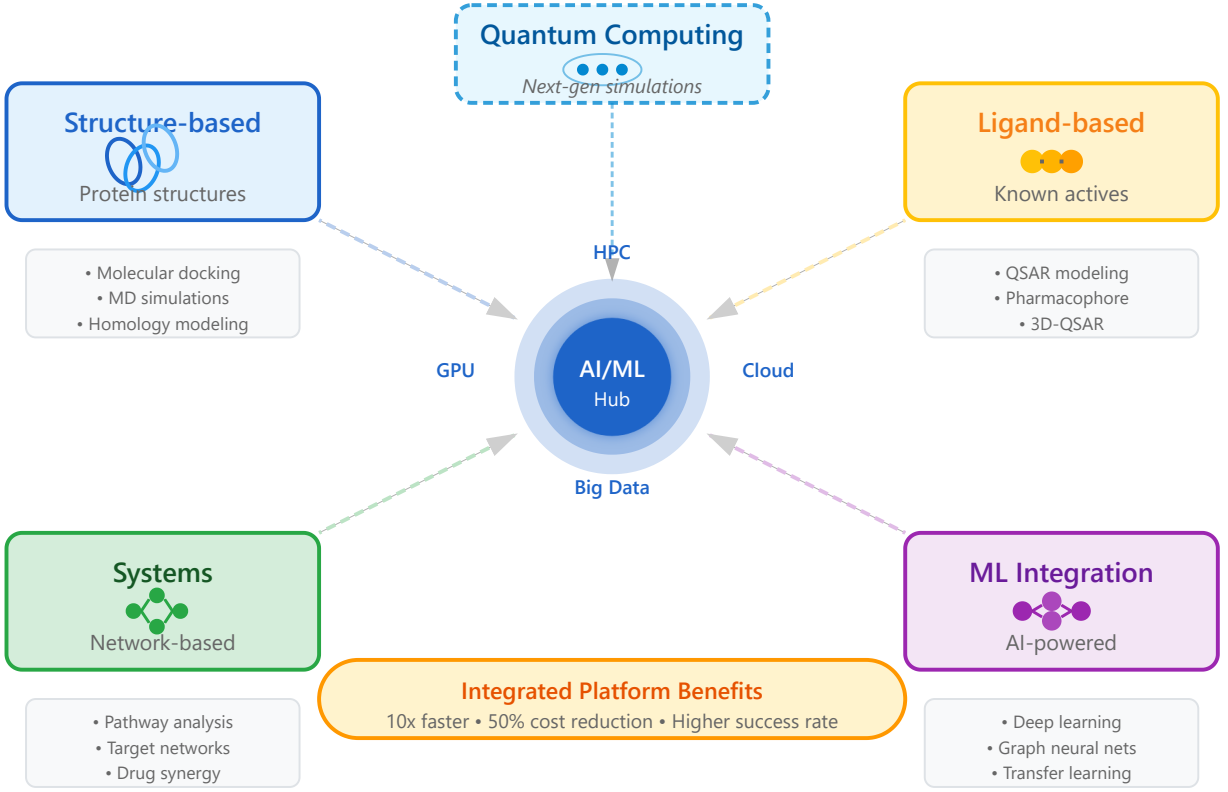
- Dose finding
- Sample size re-estimation
- Seamless phase transition
- Bayesian approaches

Real-World Evidence

- EHR data mining
- Claims analysis
- Registry studies
- Digital biomarkers

Total: 10-15 years, $1-3 billion

Computational Approaches



Part 2/3:

Molecular Machine Learning

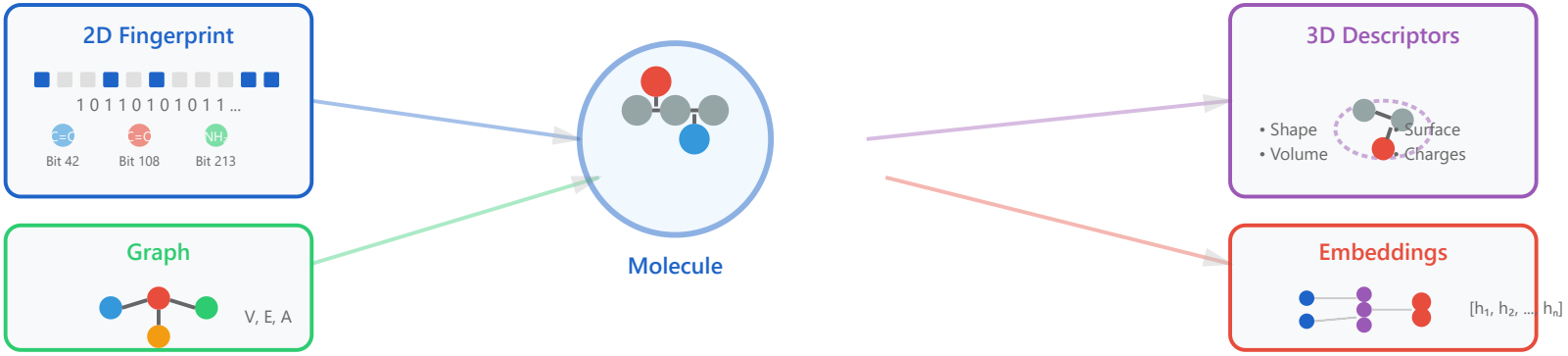
- Representation learning
- Property prediction
- Generative models

Part 2/3

Molecular ML

- Representation learning
- Property prediction
- Generative models

Molecular Representations



2D fingerprints

Binary feature vectors

3D descriptors

Geometric and conformational features

Graph representations

Molecular graph structures

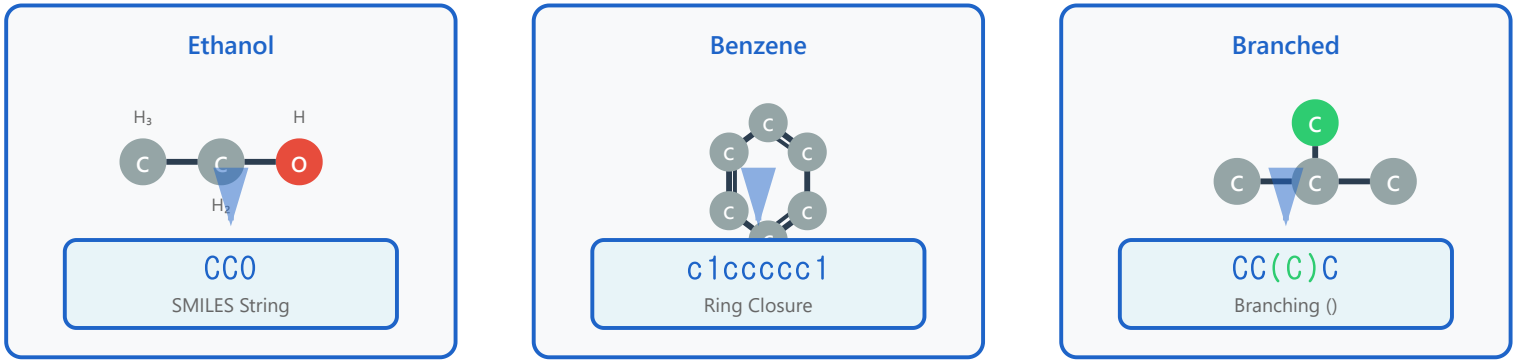
Learned embeddings

Deep learning representations

Multi-view learning

Combining multiple representations

SMILES Notation



Syntax rules

String-based molecular encoding

Canonical SMILES

Unique molecular representation

SMARTS patterns

Substructure search patterns

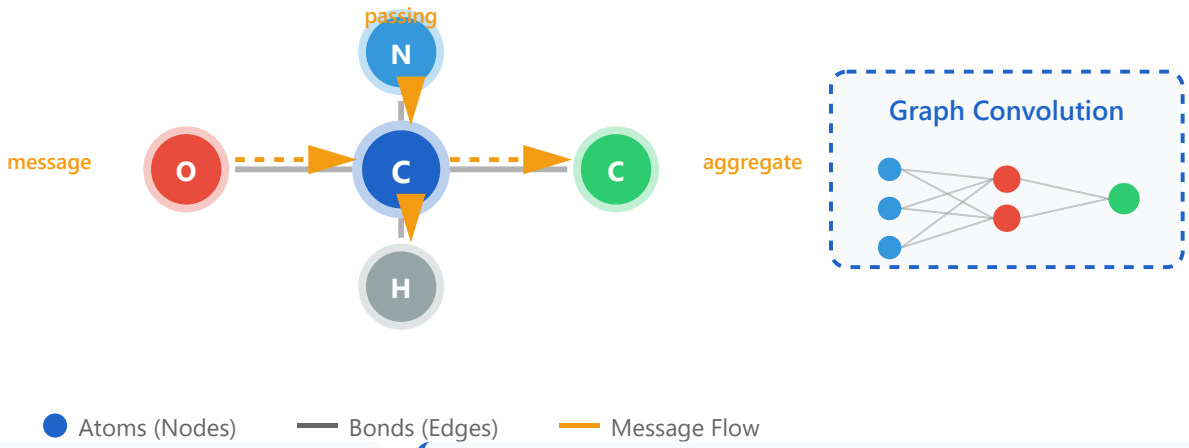
Tokenization

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

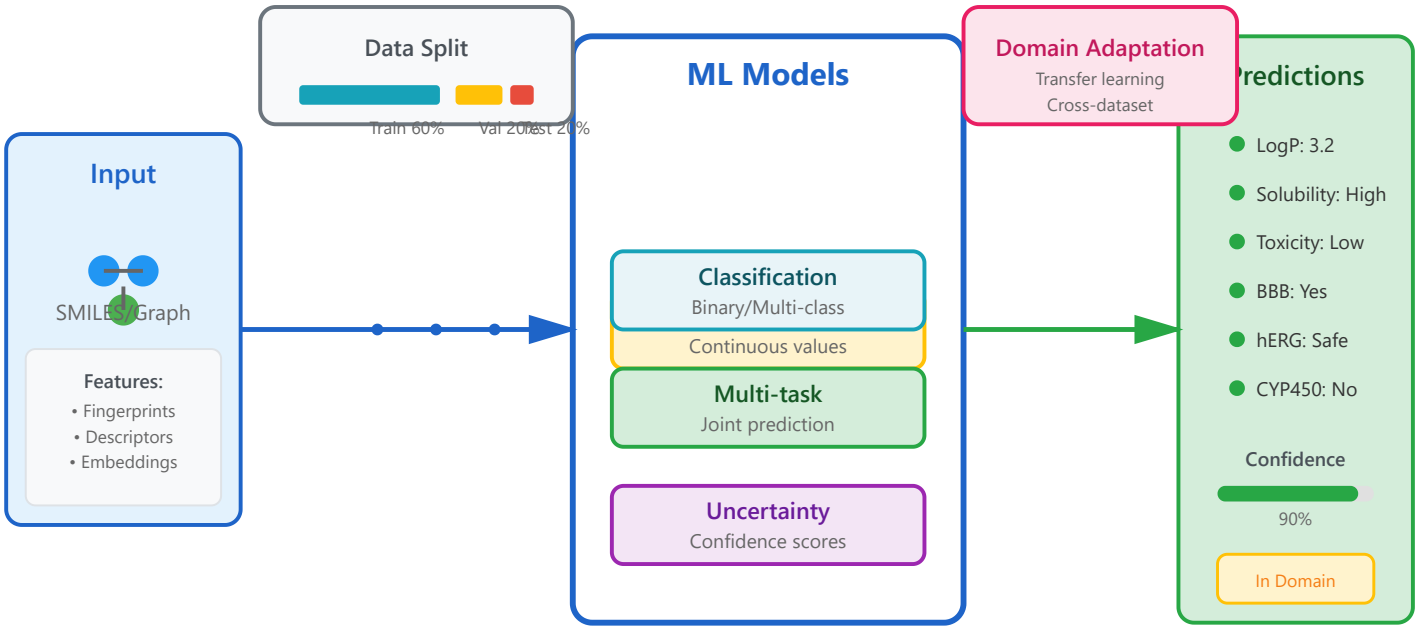
Attention mechanisms

Weighted information aggregation

Pooling strategies

Graph-level representations

Property Prediction



QSAR Modeling

Activity = f(Structure)

Validation Metrics

$R^2 > 0.6$
$Q^2 > 0.5$
RMSE < 0.3
MAE < 0.25

Descriptor
Selection

Feature
Engineering

Model
Validation

QSAR
Cycle

OECD
Principles

Applicability
Domain

Y-random
Test

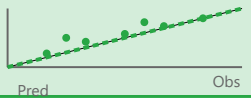
Cross-Validation

5-fold CV

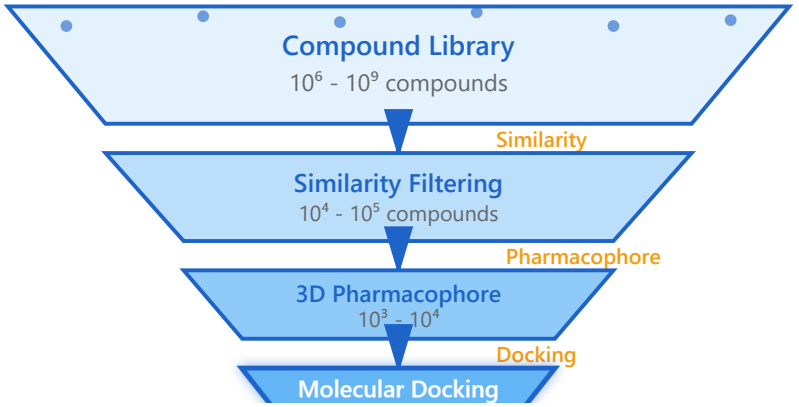
OECD Guidelines

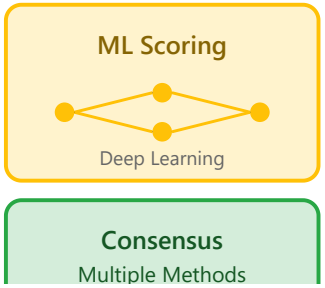
1. Defined endpoint
2. Unambiguous algorithm
3. Applicability domain
4. Statistical validation

Performance Plot



Virtual Screening





Similarity searching

Finding similar active compounds

Pharmacophore modeling

3D feature-based screening

Docking scores

Protein-ligand binding prediction

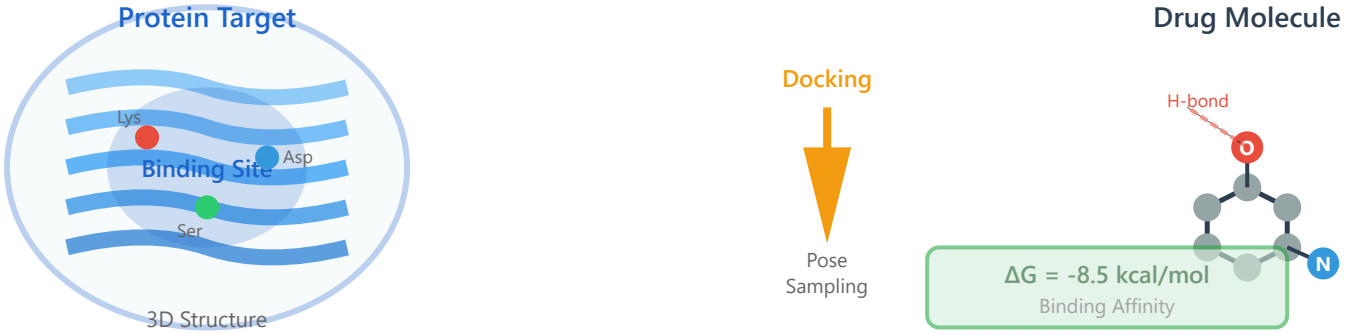
ML scoring functions

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:

Practical Applications

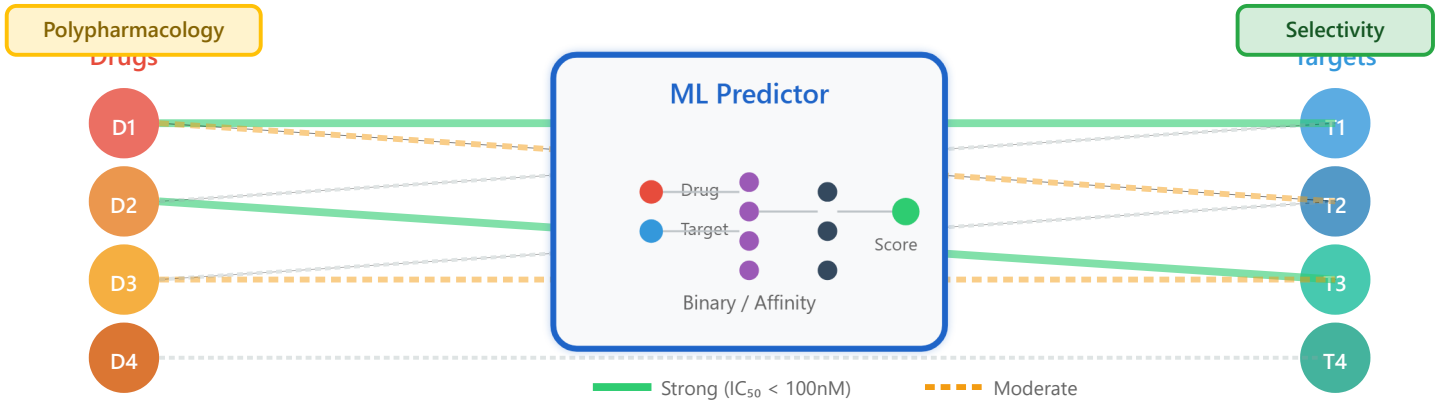
- Practical implementations
- Success metrics
- Future directions

Part 3/3

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

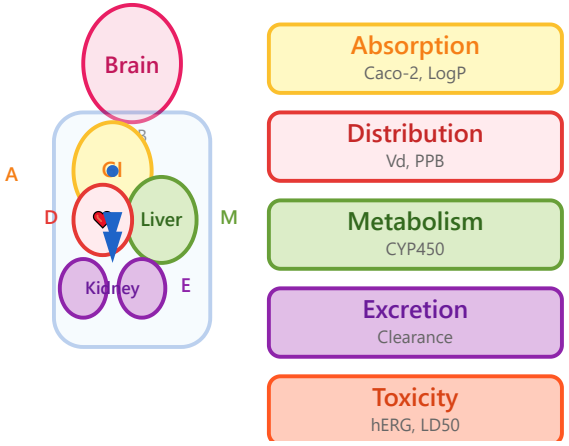
Active learning

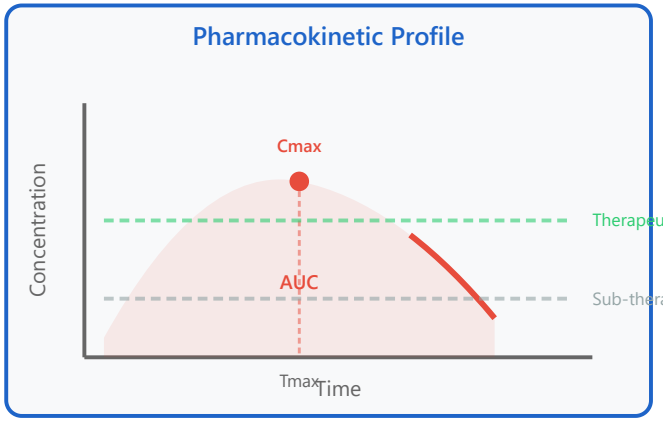
Iterative experiment design

Experimental validation

Wet-lab confirmation

ADMET Prediction





Absorption models

Oral bioavailability prediction

Distribution (BBB, Vd)

Tissue distribution modeling

Metabolism (CYP)

Drug metabolism prediction

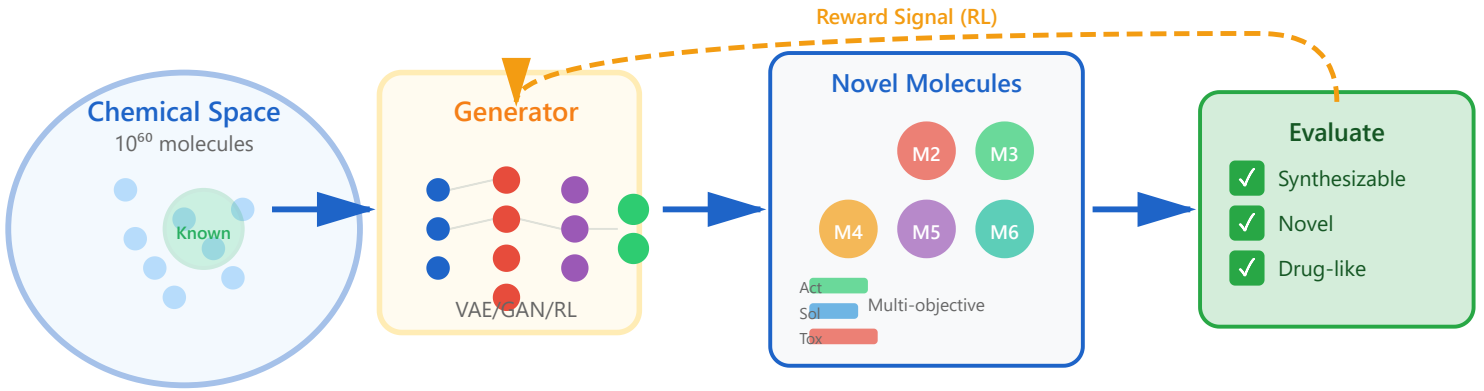
Excretion (clearance)

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

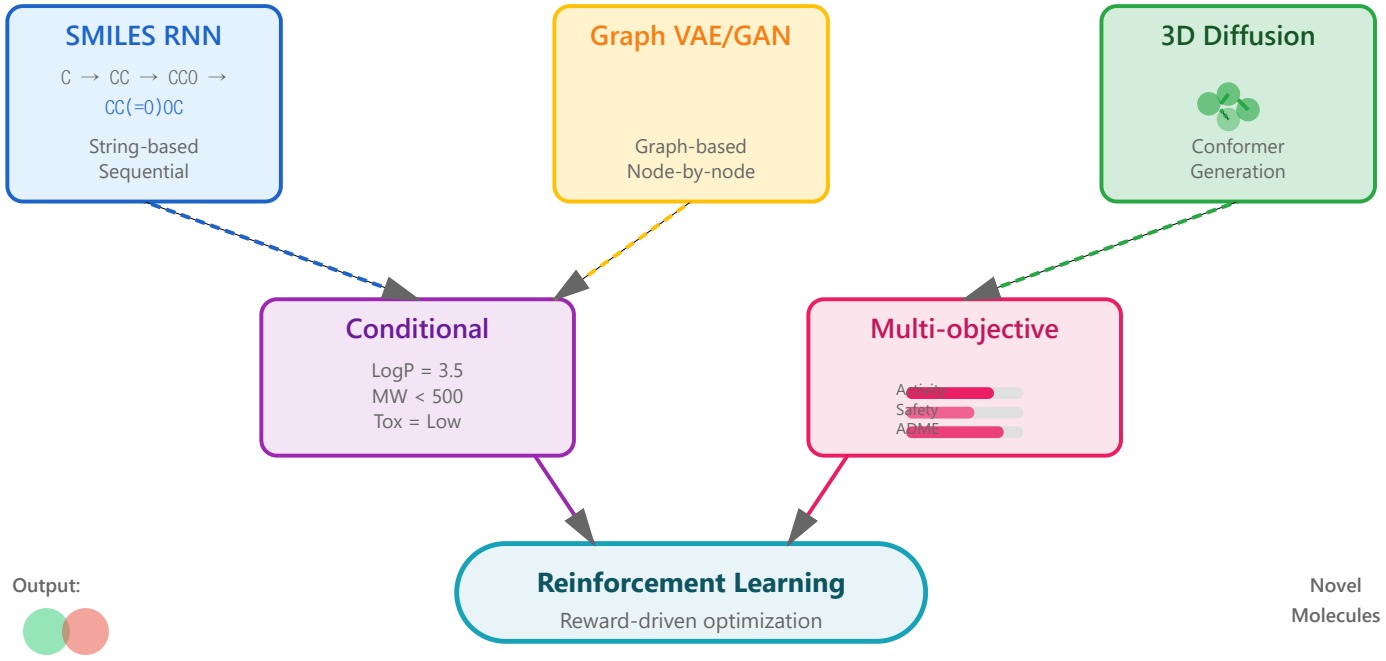
Synthesizability

Chemical feasibility assessment

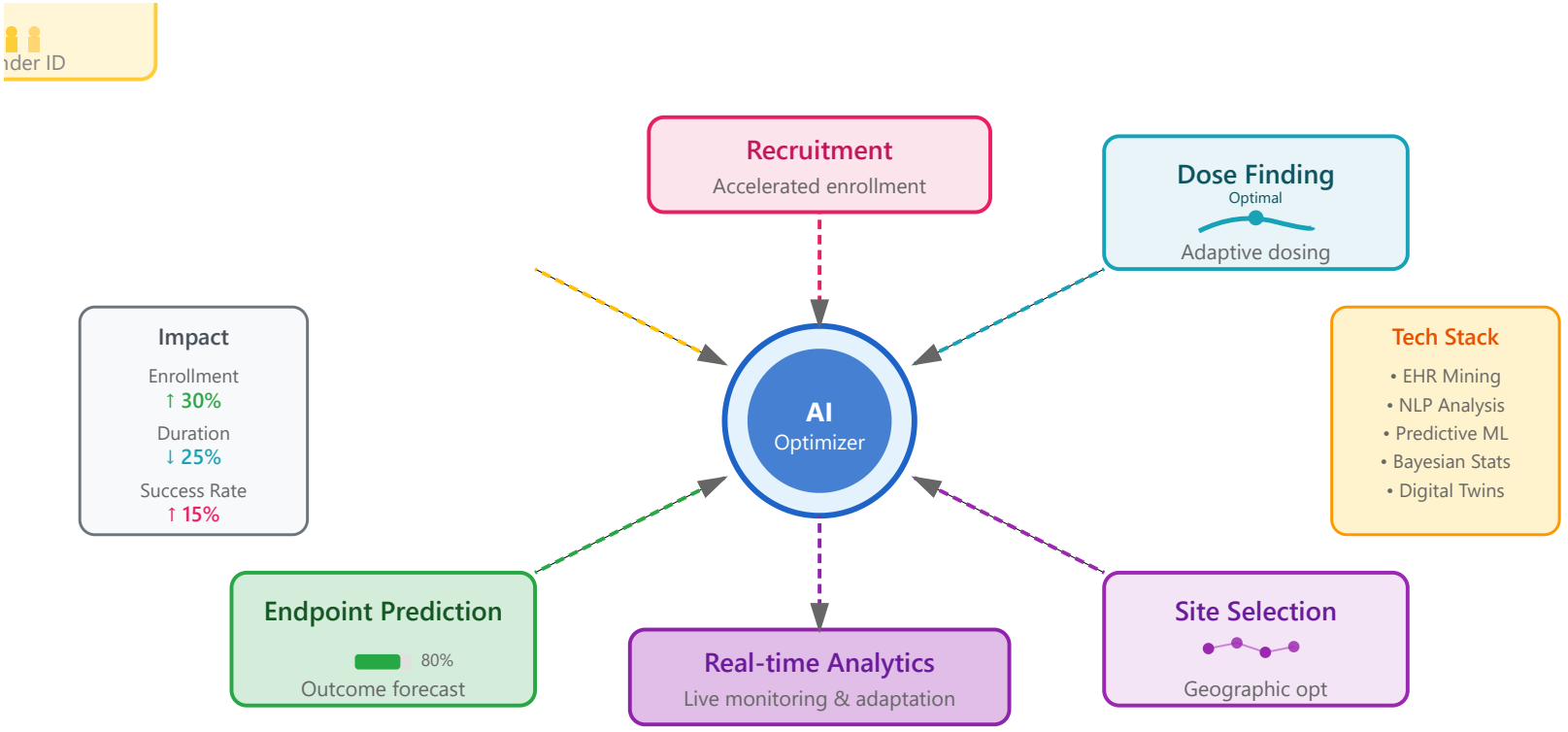
Diversity metrics

Novelty quantification

Generative Models



Clinical Trial Optimization



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

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: 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

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)
```

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

Thank You

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

Introduction to Biomedical Datascience