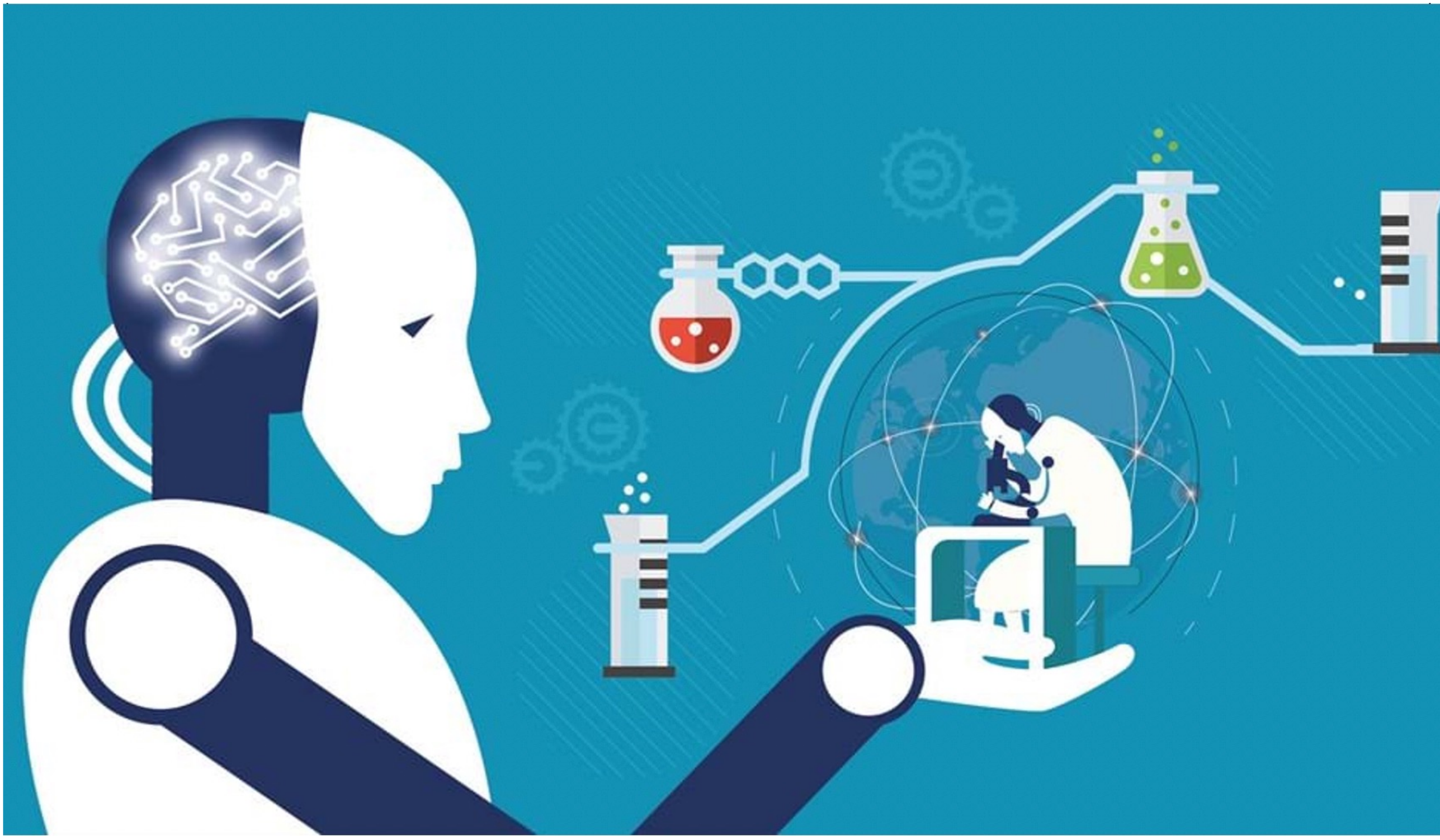
AI for Drug Discovery

2022



Survey: DMPK

Drug Metabolism and Pharmacokinetics

DMPK Review

신약개발 과정에서 약물대사 및 약동학 연구의 역할 변화

강원호* · 황진아** · 채정우*,# · 권광일*,# · 윤휘열*,#
*충남대학교 약학대학, **건양대학교 PRIME 창의융합대학

- 목적: 좋은 약동학적 특성을 보이는 전임상 후보물질 탐색
- 예전의 in-vitro potency/efficacy 위주의 후보물질 탐색에서, 후보물질의 물리화학적 특성을 고려한 in-vitro/vivo ADME (Absorption-Distribution-Metabolism-Excretion)를 **미리 고려**하는 전략으로 변화

→ 낮은 pharmacokinetics(PK) 또는 bioavailability(BA)로 실패율을 낮춤

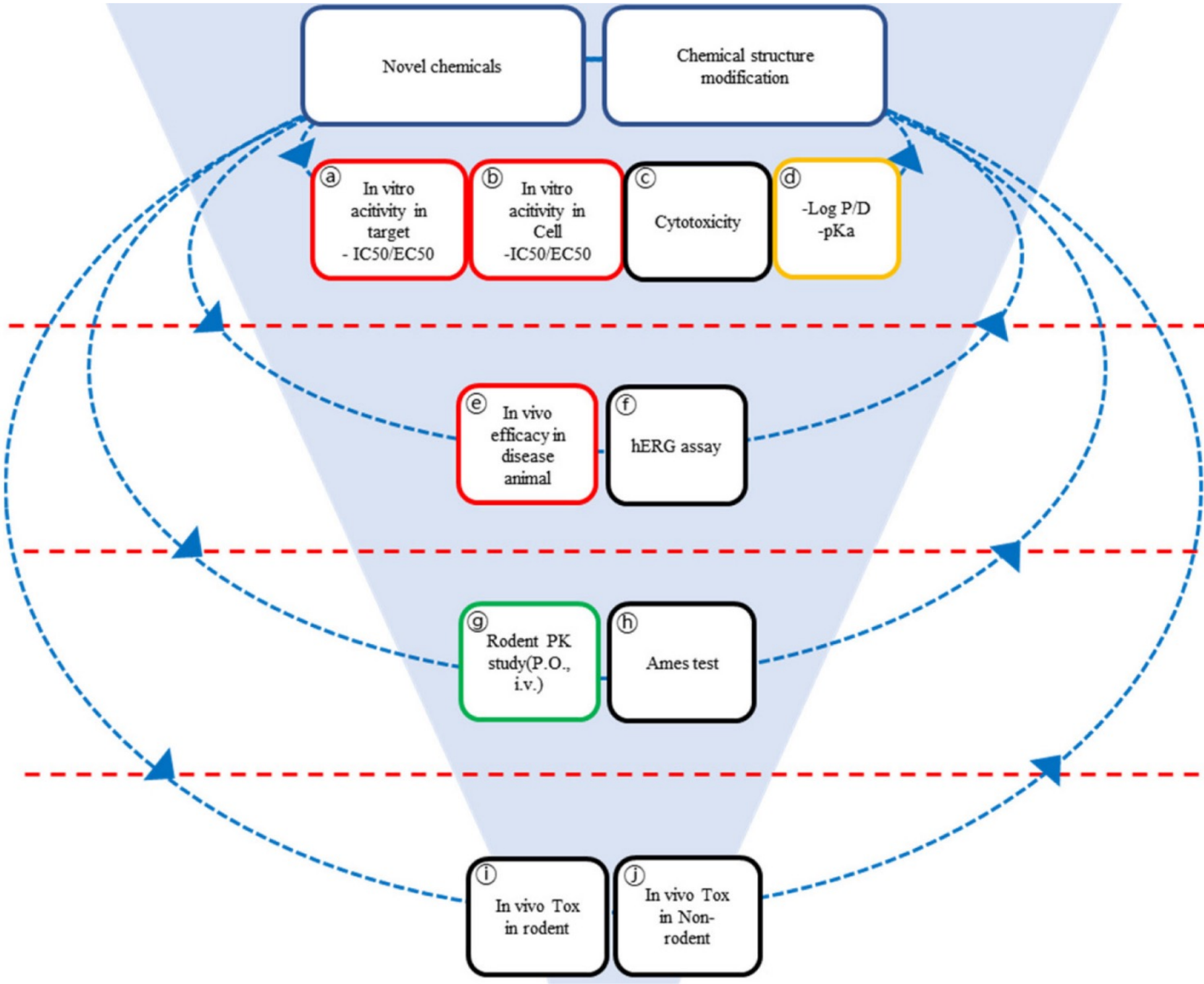
전통적 방식 (~2000)

- 신약개발 비임상 절차
  - 타겟 단백질 및 세포 수준에서 화합물의 activity 확인
  - in-vitro activity가 우수한 물질들을 in-vivo disease model을 통해 치료학적 효과를 파악
  - short term toxicity, in-vivo PK(Pharmacokinetics), BA(Bioavailability) 확인
  - 후보물질을 선별하여 IND (Investigational New Drug application) 준비

→ SAR (Structure-Activity Relationship) 중심의 후보 탐색임

- 대부분의 ADME 시험은 개발 단계에서 검증했음
  - 설치류 및 비설치류를 이용한 용량별 in-vivo PK 시험, 장관 투과도, 약물상호작용(DDI) 위험성 파악시험, 대사체 및 대사효소 규명 시험
  - radiolabeled drug을 이용한 조직분포, QWBA (Quantitative Whole-Body Autoradiography), mass balance 시험 및 in-vivo metabolite tracking 시험 등

Non-clinical work-flow (~2000)



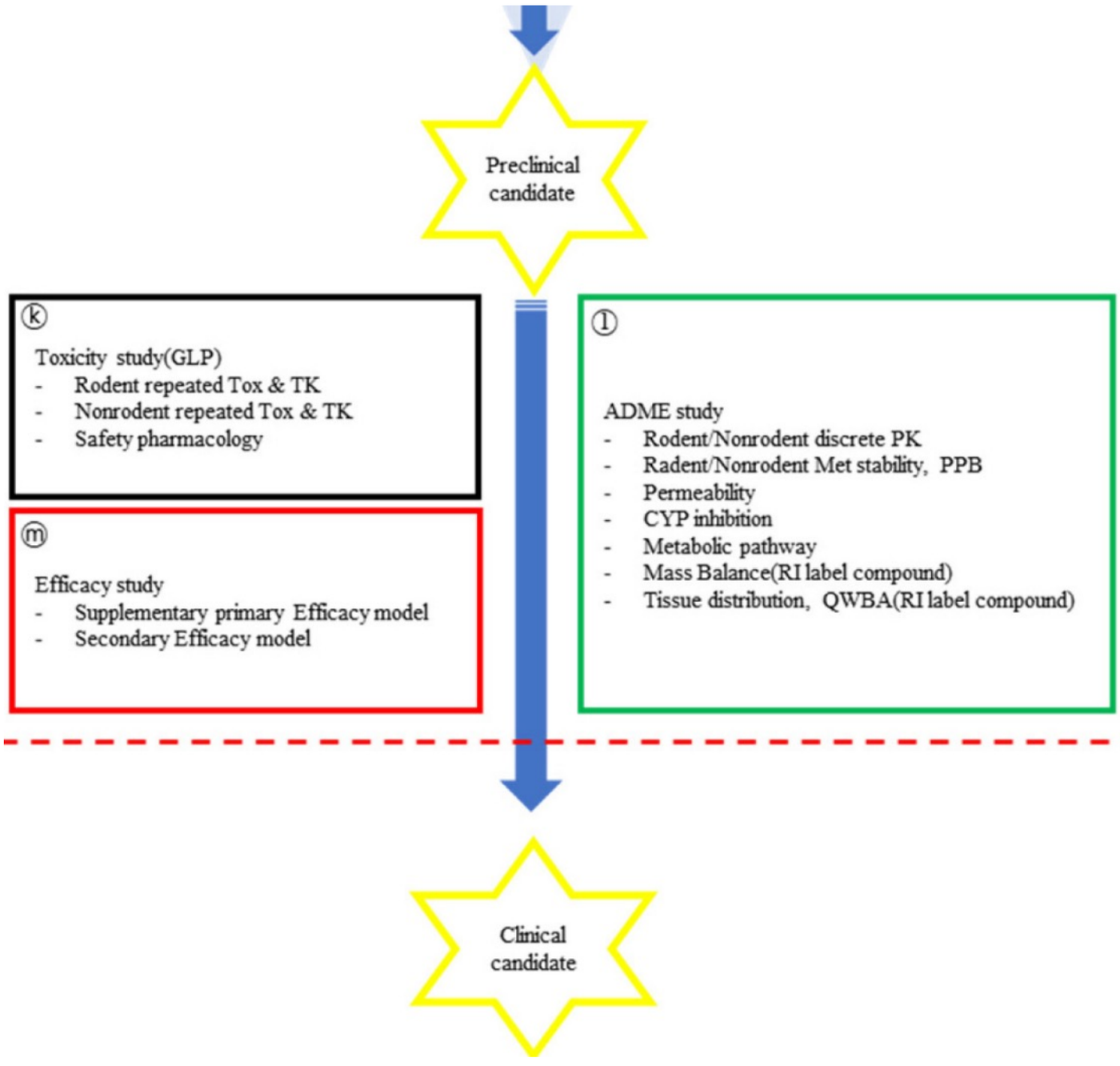
Discovery Tier 1 : SAR
- Identify in vitro Potency(ⓐ, ⓑ)
- Check cytotoxicity(ⓒ)
- Check physicochemical property(ⓓ)

Discovery Tier 2 : in-vivo Efficacy & ADR
- Identify in-vivo efficacy(ⓔ)
- Check hERG inhibition(ⓕ)

Discovery Tier 3 : in-vivo PK & ADR
- Identify in vivo PK, BA (ⓖ)
- Check in-vitro genetic toxicity(ⓗ)

Discovery Tier 4 : Toxicity
- Grasp MTD, NOEL/NOAEL(ⓘ, ⓙ)
- Check safety margin based on dose(ⓘ, ⓙ, ⓔ)

Cont.



Development (Preparation IND filing)

- Reconfirm NOEL/NOAEL(ⓚ)
- Check safety margin based on exposure(ⓚ, ①, ⓔ, ⓜ)
- Check safety pharmacology(ⓚ)
- Supplement Efficacy(ⓜ)
- Rodent/Nonrodent ADME characterization(①)

최근 방식 (2000~)

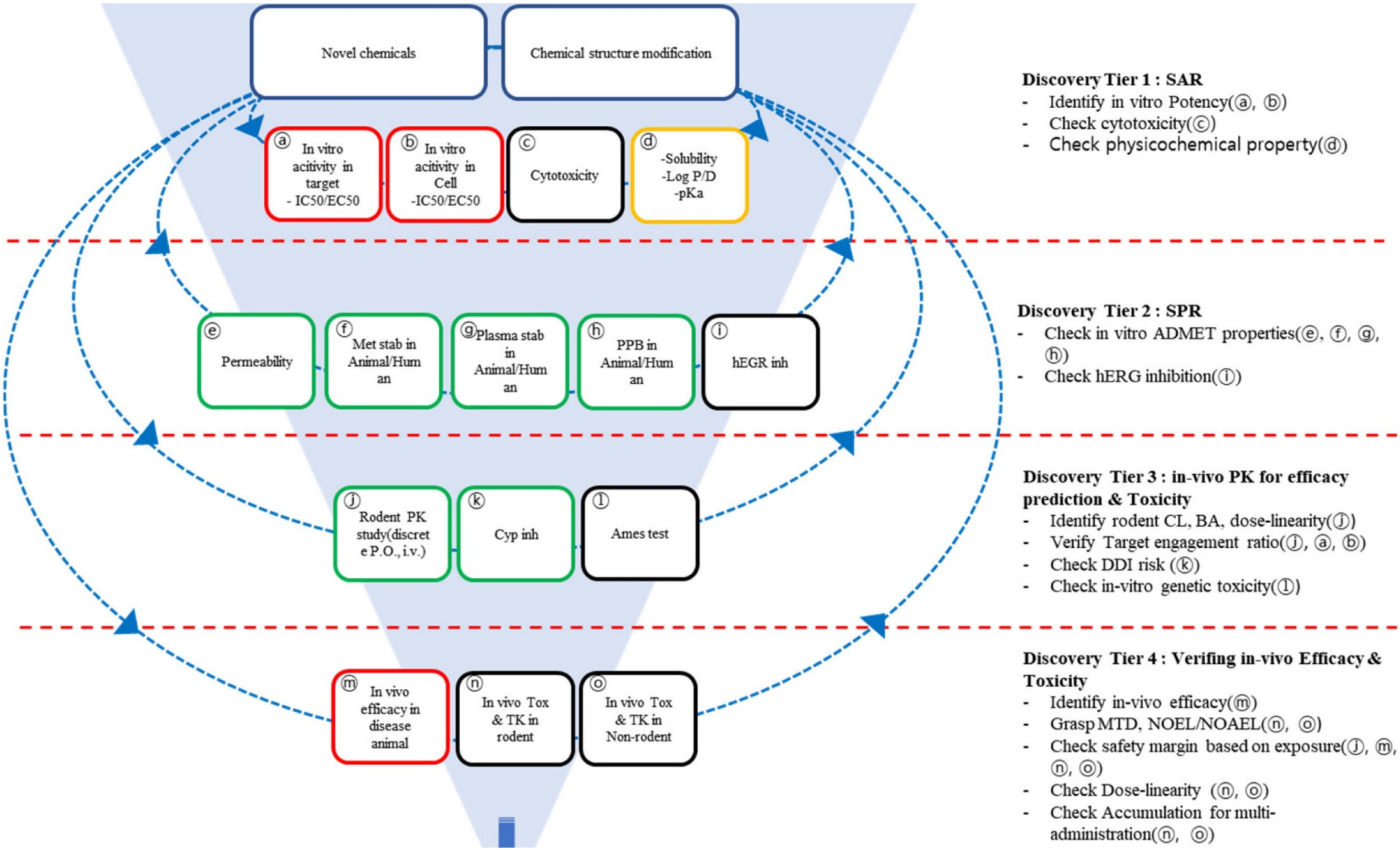
- SAR로 최적화한 화합물의 한계
  - 실제 생체에 적용했을 때, 위장관 용출/용해, 투과도, 단백결합, 대사, 조직 분포 등 타겟 사이트로의 이행과정에서 발생할 수 있는 다양한 변수가 고려되지 못함

→ in-vitro ADMET 시험을 early discovery stage에서 실행

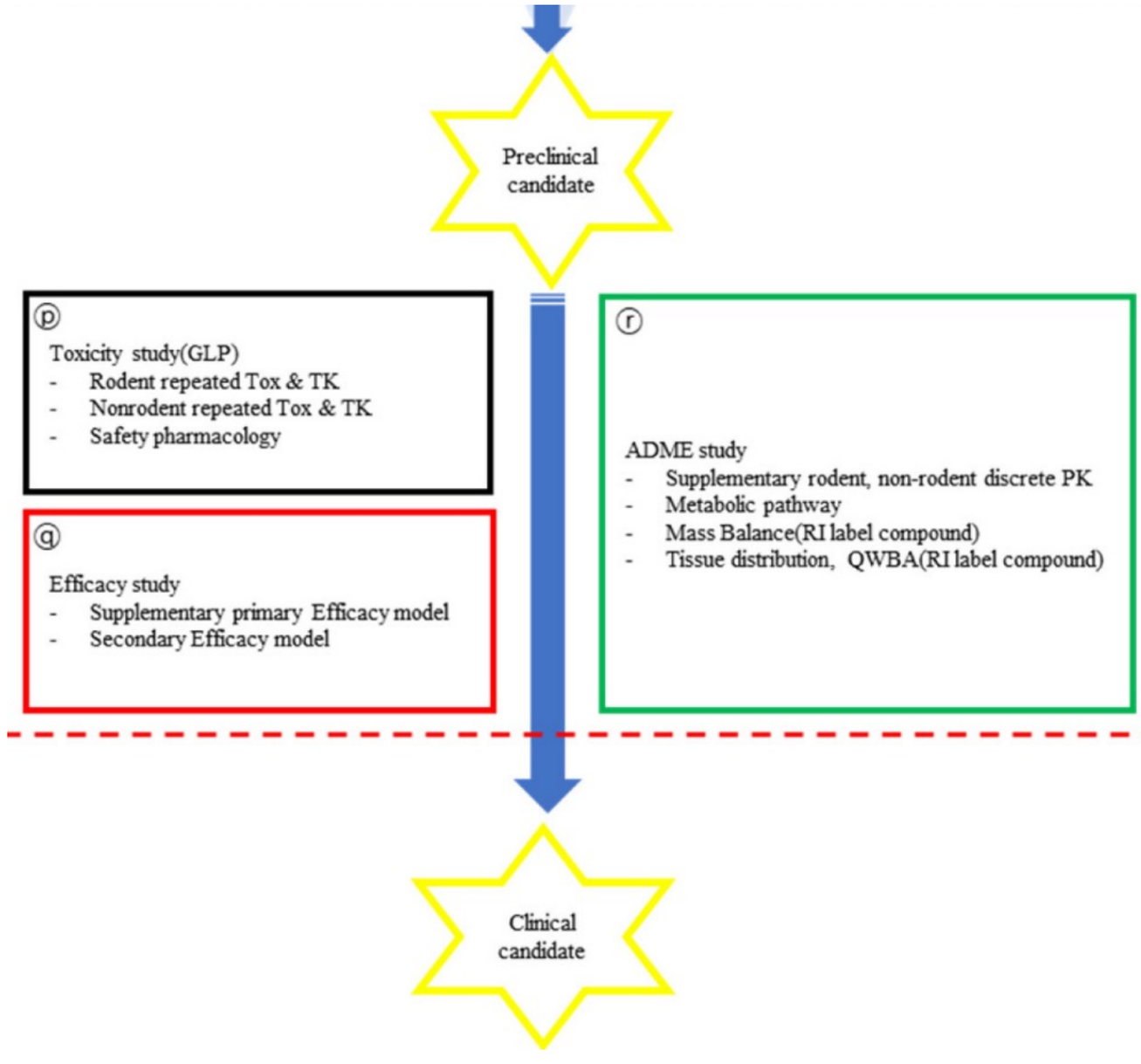
- 투과도, 대사안정성, 혈장내 안정성, 단백 결합률 평가 등
- in-vitro 수준에서 적합한 약동학적 특성을 보이는 화합물을 찾고
- in-vivo PK 시험과 타겟 수준에서의 activity를 연결지어 in-vivo efficacy에 대한 예측을 수행하여 in-vivo efficacy 평가 대상의 우선순위 설정

→ 낮은 생체이용률 및 부적합한 약동학적 특성으로 인한 실패율 감소

최근 방식 (2000~)


Novel chemicals
Chemical structure modification
ⓐ In vitro acitivity in target - IC50/EC50
ⓑ In vitro acitivity in Cell -IC50/EC50
ⓒ Cytotoxicity
ⓓ -Solubility -Log P/D -pKa
ⓔ Permeability
ⓕ Met stab in Animal/Human
ⓖ Plasma stab in Animal/Human
ⓗ PPB in Animal/Human
ⓘ hEGR inh
ⓙ Rodent PK study(discrete P.O., i.v.)
ⓚ Cyp inh
ⓛ Ames test
ⓜ In vivo efficacy in disease animal
ⓝ In vivo Tox & TK in rodent
ⓞ In vivo Tox & TK in Non-rodent
Discovery Tier 1 : SAR
- Identify in vitro Potency(ⓐ, ⓑ)
- Check cytotoxicity(ⓒ)
- Check physicochemical property(ⓓ)
Discovery Tier 2 : SPR
- Check in vitro ADMET properties(ⓔ, ⓕ, ⓖ, ⓗ)
- Check hERG inhibition(ⓘ)
Discovery Tier 3 : in-vivo PK for efficacy prediction & Toxicity
- Identify rodent CL, BA, dose-linearity(ⓙ)
- Verify Target engagement ratio(ⓙ, ⓐ, ⓑ)
- Check DDI risk (ⓚ)
- Check in-vitro genetic toxicity(ⓛ)
Discovery Tier 4 : Verifing in-vivo Efficacy & Toxicity
- Identify in-vivo efficacy(ⓜ)
- Grasp MTD, NOEL/NOAEL(ⓝ, ⓞ)
- Check safety margin based on exposure(ⓙ, ⓜ, ⓝ, ⓞ)
- Check Dose-linearity (ⓝ, ⓞ)
- Check Accumulation for multi-administration(ⓝ, ⓞ)

Cont.



Development (Preparation IND filing)

- Reconfirm NOEL/NOAEL(ⓟ)
- Recheck safety margin based on exposure(ⓟ, ⓙ, ⓡ, ⓜ, ⓠ)
- Check safety pharmacology(ⓟ)
- Supplement Efficacy(ⓠ)
- Supplement ADME characterization(ⓡ)

최근 방식 (2000~)

- Discovery Tier 1. 구조-활성 상관관계(SAR) 연구단계
  - 활성 화합물을 찾아내는 과정
  - 타사 화합물의 특허를 회피할 수 있는 mimic 화합물 찾기
  - CADD (Computer Aided Drug Design)를 이용한 target과 ligand의 structure 연구
  - pharmacophore modeling, ligand docking, QSAR (Quantitative SAR)
  - 타겟 단백질 및 세포 수준에서 활성 평가
  - 화합물의 기본적인 물리화학적 특성 파악
  - 용해도 등 기본적인 생물약제학적 특성 파악으로 경구제 개발 가능성에 대해 빠른 판단
  - in-vitro activity value (IC50 or EC50)

Cont.

- Discovery Tier 2. 구조-특성 관계(SPR)에 대한 연구
  - Tier 1에서 발굴한 in-vitro active compound들의 invitro ADMET property들을 파악하여 drug-likeness (약물유사성, drugability)를 확인
  - 장관 투과도 및 efflux efficiency를 평가해보기 위한 다양한 permeability assay (PAMPA, Caco-2, IAM-PCDD 등) 수행
  - first-pass metabolism을 확인하기 위한 metabolic stability assay(microsome, hepatocyte, S9 fraction 등) 수행
  - 혈장 중 단백결합율을 파악하기 위한 plasma protein binding assay
  - 혈장 내 분해 클리어런스(plasma degradation clearance)를 파악하기 위한 plasma stability 등

Cont.

- 약물 유사성을 평가하기 위한 각 시험 항목별 최적 범위

Table 1. The optimal range for each *in-vitro* ADME assay[13)]

No.	Assay	Goal	Optimal range
1	Permeability	Investigation of gastrointestinal permeability and efflux ratio	$P_{app} > 2\times10^{-6}$ cm/sec Efflux ratio < 2
2	Metabolic stability	Investigation of hepatic metabolic stability	20% of HBF < CL_{hep} < 80% of HBF
3	Plasma stability	Investigation of plasma stability	Stability > 85%
4	Plasma protein binding	Investigation of unbound fraction portion in plasma	No criteria

P_{app}, Apparent permeability; CL_{hep}, Hepatic metabolic clearance; HBF, Hepatic blood flow

Cont.

- Discovery Tier 3. 약효 및 독성 예측을 위한 생체내 약동학 연구
  - 설치류를 이용한 정맥, 경구 in-vivo PK시험과 DDI risk를 판단해 보기 위한 CYP inhibition assay등
  - 정맥 투여시 혈중농도-시간 곡선, 전신 클리어런스(CLtotal)를 간

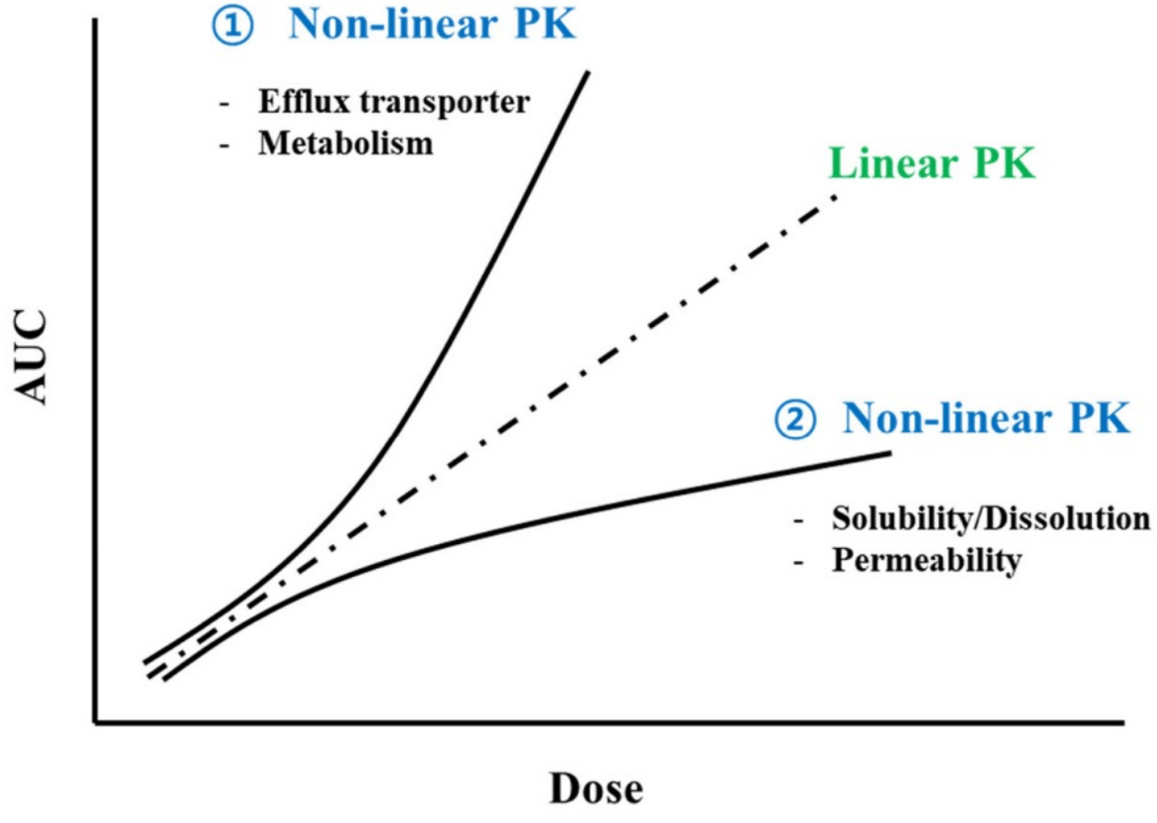


Fig. 3. Linear and non-linear pharmacokinetics in dose escalation.

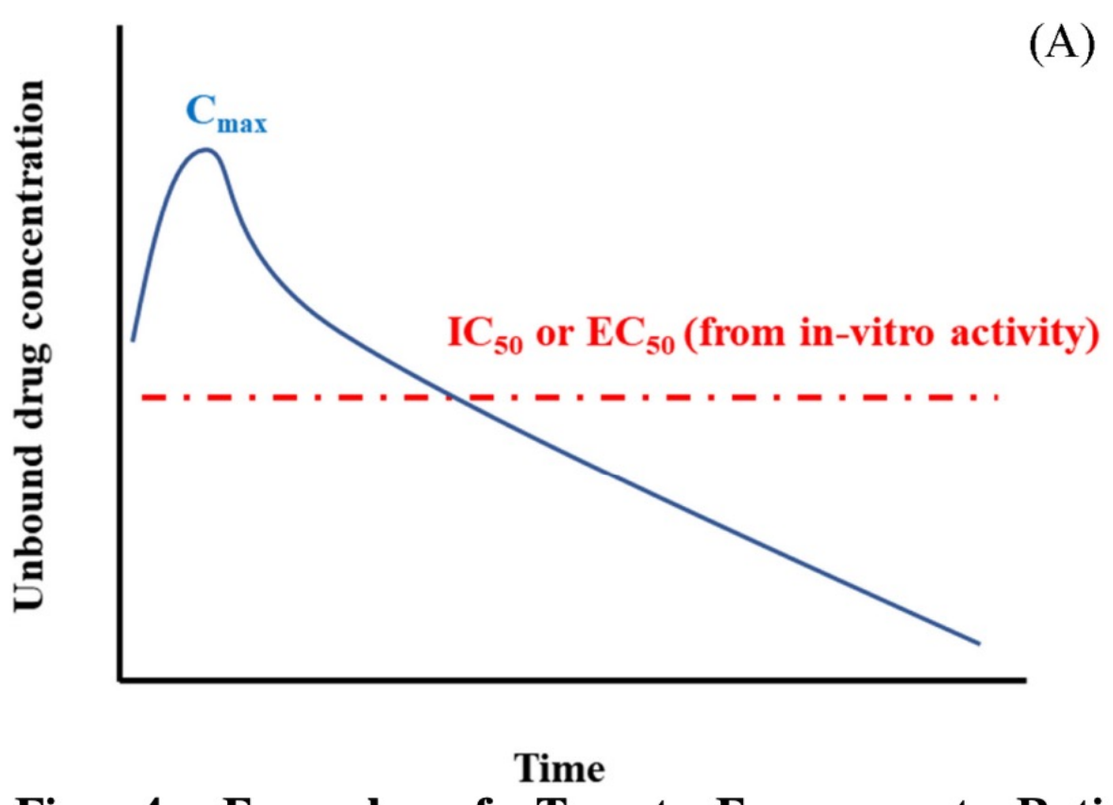


Fig. 4. Example of Target Engagement Ratio
Concentration based TER: C_{max} / IC_{50} or EC_{50}, (B) AUC ba

Cont.

- Discovery Tier 4. 생체내 약효 및 독성 검증 단계
  - 이 과정에서는 in vivo disease animal model을 이용하여 효력에 대한 확인 및 therapeutic dose range에 대한 확증
  - in-vivo toxicity 시험을 통하여 MTD (Maximum Tolerated Dose), NOEL (No Overserved Effect Level) 등을 파악
  - toxicokinetic 시험을 통해 고용량 구간에서의 노출 수준, 반복 투여시 약물 축적 여부 및 dose linearity등을 파악
  - in-vivo efficacy 시험의 MED (Minimum Efficacious Dose) 또는 ED50 (50% efficacious dose)에서의 혈중 약물 노출과 NOEL (or NOAEL)을 비교하여 safety margin 산출

Cont.

- Development (preparation IND filing)
  - IND 자료 제출을 위한 GLP toxicity 시험, secondary efficacy model 평가를 통한 약효 자료의 보강, 미비된 ADME study 등을 수행
  - Non-rodent animal (dog and/or monkey)을 이용한 용량별 in-vivo PK 시험, 대사 경로 확인을 위한 대사체 및 대사효소 규명 시험, radiolabeled drug을 이용한 조직분포, QWBA, mass balance 시험 및 in-vivo metabolite tracking 시험 등
- 비임상 흡수, 분포, 대사, 배설 데이터를 이용한 Human 약동학 프로파일 예측
  - 동물 시험 데이터들을 이용하여 human PK 모델링 및 시뮬레이션을 수행하고 human에서의 유효용량을 예측

Survey: AIDD

Survey Paper

Artificial Intelligence in Drug Discovery: Applications and Techniques

Jianyuan Deng, Zhibo Yang, Iwao Ojima, Dimitris Samaras, Fusheng Wang

- (2021.11)
- https://arxiv.org/abs/2106.05386
- https://github.com/dengjianyuan/Survey_AI_Drug_Discovery
- Contents
  - molecular property prediction
  - molecule generation
  - common data resources
  - molecule representations
  - benchmark platforms

Popular Applications of AIDD

- virtual screening
- de novo drug design (small molecules)
- Retrosynthesis (역합성) and reaction prediction
- de novo protein design
- drug repositioning
- target identification
- exploiting omics data for druggability

→ **Predictive task**

→ **Generative task**

Drug Discovery Overview

- **Basic Hypothesis**:
  - activation or inhibition of a target (e.g., an enzyme, a receptor, an ion channel, etc,) results in therapeutic effects
  - target identification and target validation
- intensive assays to find the hits and leads (i.e., drug candidates)
  - hit discovery, hit-to-lead, lead optimization phases
- preclinical studies
- clinical trials
- medical product

HTS, VS

- high-throughput screening (HTS) increase the discovery efficiency since the 1980s
  - outcome of HTS is the large-scale structure-activity relationship (SAR) datasets
  - chemical databases: PubChem, ZINC
- virtual screening (VS)
  - search the chemical libraries for potentially active molecules to be tested in subsequent **in vitro and in vivo assays**
  - identify active molecules using computational approaches
    - based on knowledge about the target (**structure-based VS**)
    - based on known active ligands (**ligand-based VS**)

Mechanism of Action (MoA)

- agonists and antagonists (작용제, 길항제)
- an agonist
  - activates the target to exert a biologic response
- an antagonist
  - binds to the target to block the response

Measure of activity

- Affinity (친화력)
  - the extent to which a molecule binds to a target at a given concentration
- Potency
  - the necessary amount of a molecule to produce an effect of a given magnitude (inversely proportional to the affinity)
- Efficacy (효능)
  - describes the effect size, such as inhibition of an enzyme to 60%

Table 1: Common Measures of Molecule Activity

Measures	Definition
Kd	Equilibrium dissociation constant
Km	Michaelis constant
Ki	Inhibition constant
IC50	Half maximal inhibitory concentration
EC50	Half maximal effective concentration

Needs more than activity

- To be more than a ligand
- Consider properties:
  - physicochemical (water solubility, acid-base dissociation constant(해리상수), lipophilicity(친유성), permeability(투과성))
  - pharmacokinetic (absorption, distribution, metabolism, excretion)
  - pharmacodynamic (activity, selectivity)
- Compound synthesis
  - Synthetic Accessibility Score (SAS): 1 ~ 10
  - Quantitative Estimation of Drug-likeness (QED): 0 ~ 1

QSAR Model

- Quantitative structure-activity relationship (QSAR) modeling
- A predictive model to map the molecular structure to the property value with either classification or regression
- Can be exploited inversely to reveal the structural features to guide de novo drug design

Drug Design

- Needs to explore the vast chemical space, ~10^60
- Needs to optimize the design-make-test-analysis (DMTA) cycle
- Two questions:
  - Can molecular properties be deduced from molecular structures?
  - Which structural features are relevant for certain molecular properties?
- Drug design can be viewed as an extension to VS
  - involves both molecular property **prediction** and molecule **generation**

→ the major tasks in current AI-driven drug discovery

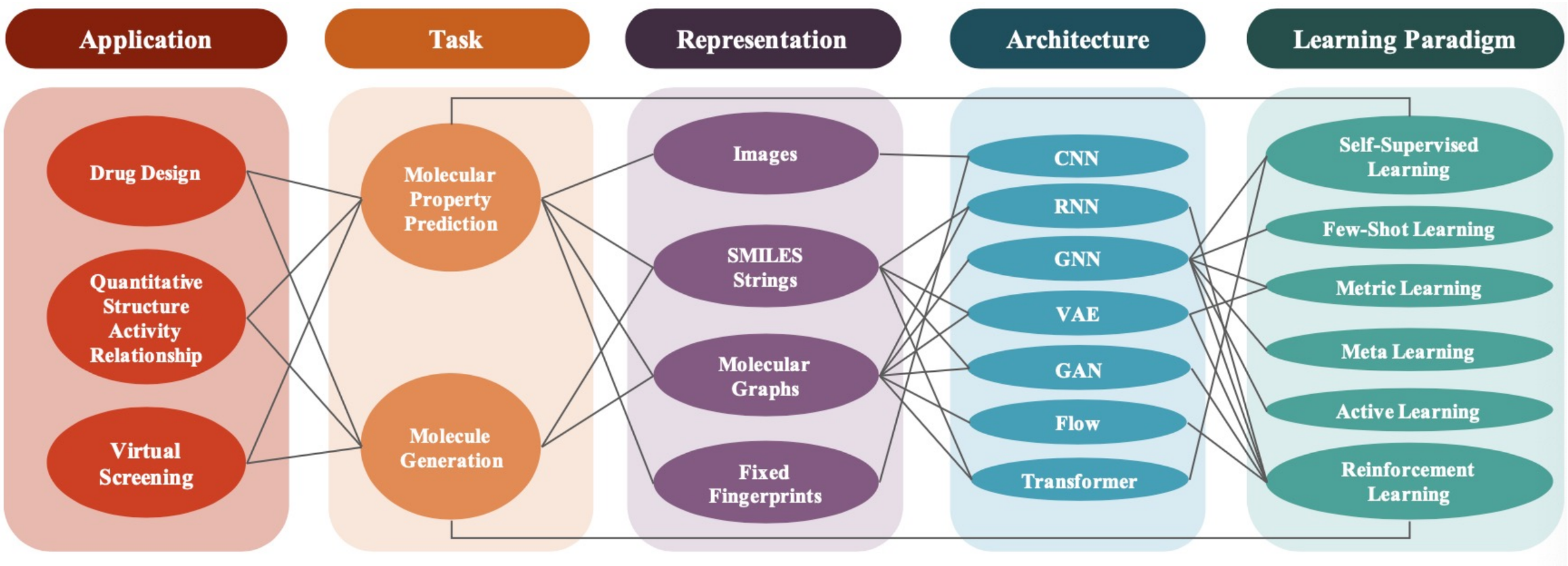
AI-driven DD

- 2000~ : RF models for VS and QSAR
- 2012: Deep Learning models
- 2015: Computer vision AI
- 2018: Natural language processing AI
- 2019: potent inhibitors of DDR1 (discoidin domain receptor 1) were discovered in 21 days by Insilico Medicine

Lead Identification

- Molecular property prediction (VS)
  - predict the property value of a molecule given its structure or learned representation
  - for drug-target interaction (DTI) prediction, toxicity prediction and drug-induced liver injury (DILI) prediction, etc.
- Molecule generation (drug design)
  - generating molecules within constraints imposed by the chemical rules
  - generating chemically valid molecules with desired properties

Summary



- For low-data molecular property prediction
  - self-supervised learning for the pretraining-finetuning practice
- For goal-directed molecule generation
  - reinforcement learning for navigating the chemical space search

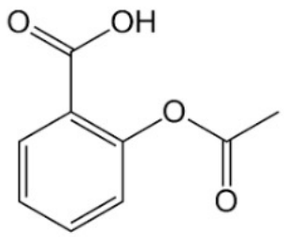
Data Representation

- Needs good machine-readable formats
- PubChem
  - By National Institutes of Health from 2004 (non-curated)
  - (August 2020) contains 111 million unique chemical structures with 271 million activity data points from 1.2 million biological assays experiments
  - curated dataset of 77 million SMILES strings from PubChem
- ChEMBL
  - By European Molecular Biology Laboratory
  - (ChEMBL22) 1.6 million distinct chemical structures with over 14 million activity values
  - manually curated

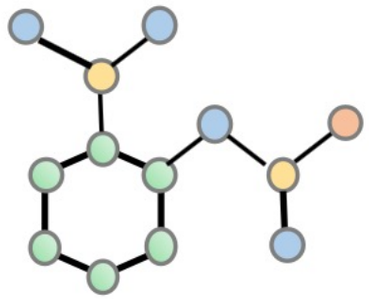
Data Representation

- ZINC
  - by UCSF
  - molecules, annotated ligands and targets as well as the purchasability for 120 million "drug-like" compounds
  - subsets of the ZINC: ZINC-250k, ZINC Clean Leads collections
- Etc.
  - PDBbind, BindingDB, DUD, DUD-E, MUV, STITCH, GLL&GDD, NRLiST BDB, KEGG
- For marketed drugs and their effects
  - adverse drug reactions (ADR) (e.g., DrugBank, SIDER, OFFSIDES and TWO-SIDES) and the datasets for DILI (e.g., DILIrank)

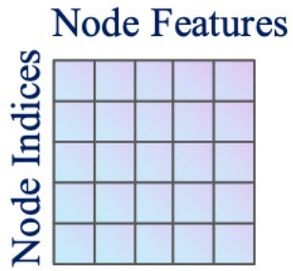
Small Molecule Representation



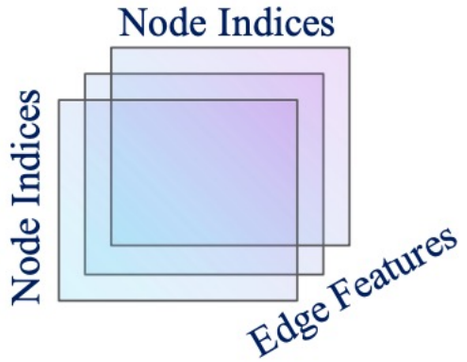
A. Kekulé Diagram



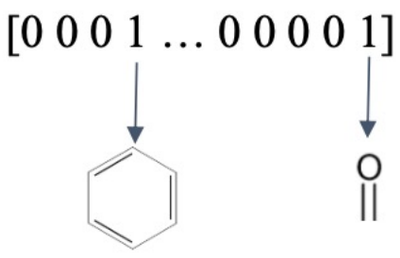
C. Molecular Graph



Node Feature Matrix



Adjacency Tensor



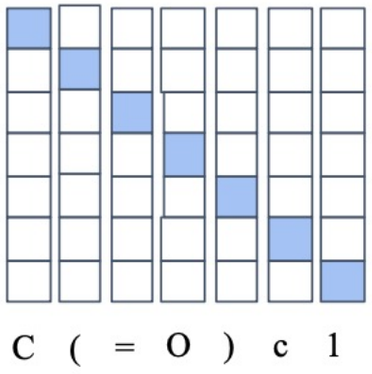
B. Fingerprints

CC(=O)Oc1ccccc1C(=O)O

D. SMILES String

Tokenization

One-Hot Encoding



Molecular Descriptors

- 0D
  - molecular weight (MW), atom number, and atom-type count
- 1D
  - Fingerprints, substituent atoms, chemical bonds, structural fragments, and functional groups
- 2D
  - represent the atom connectivity and molecular topology
  - 1) Keyed fingerprints - molecular access system (MACCS) keys
  - 2) Path- based fingerprints - DayLight fingerprints
  - 3) Circular fingerprints - extended connectivity fingerprints (ECFPs) based on the Morgan algorithm
- 3D
  - steric properties, surface area, volume and binding site properties

Embedding Representation

- For end-to-end (E2E) predictions
  - embedded into a continuous latent space
- molecular graphs
- simplified molecular input entry system (SMILES) strings

Molecular Graphs

- carry more structural information, highly interpretable
- atoms are typically mapped to nodes and bonds to edges
  - node feature matrix
  - edge feature matrix, adjacency tensor
  - common node and edge features

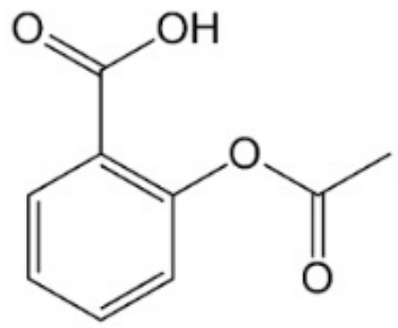
Type	Feature	Notes
Node	Atom type	Element type
Node	Formal charge	Assigned charges
Node	Implicit Hs	Number of bonded hydrogens
Node	Chirality	R or S configuration
Node	Hybridization	Orbital hybridization: sp^x, $sp^x d^y$
Node	Aromaticity	Aromatic atom or not
Edge	Bond type	Single, double, triple or aromatic
Edge	Conjugated	Conjugated or not
Edge	Stereoisomers	cis (Z) or trans (E)

SMILES Strings

- organic subset, B, C, N, O, P, S, F, Cl, Br and I, can be written without brackets
- Others: [Fe2+]
- c is used for the aromatic carbon
- single, double, triple and aromatic bonds: (-), =, # and (:)
- canonicalization methods for unique SMILES
- simple, but lose some structural information

Other Representations

- more sophisticated 3D-atomic coordinates, in structure-based VS or QSAR studies
  - bond lengths, bond angles and torsional angles, can also be incorporated
- images of molecular structures



Benchmark Platforms

- MoleculeNet (2018)
  - 1) Quantum mechanics (QM7, QM7b, QM8, QM9)
  - 2) Physical chemistry (ESOL, FreeSolv, Lipophilicity)
  - 3) Biophysics (PCBA, MUV, HIV, PDDBind, BACE)
  - 4) Physiology (BBBP, Tox21, Tox- Cast, SIDER, ClinTox)
- ChemBench package from MolMapNet
- Chemprop (2019)
- GuacaMol (2019)
- REINVENT, GraphINVENT (2020) for generating model
- MOSES (Molecular sets) (2020)

Common Evaluation Metrics

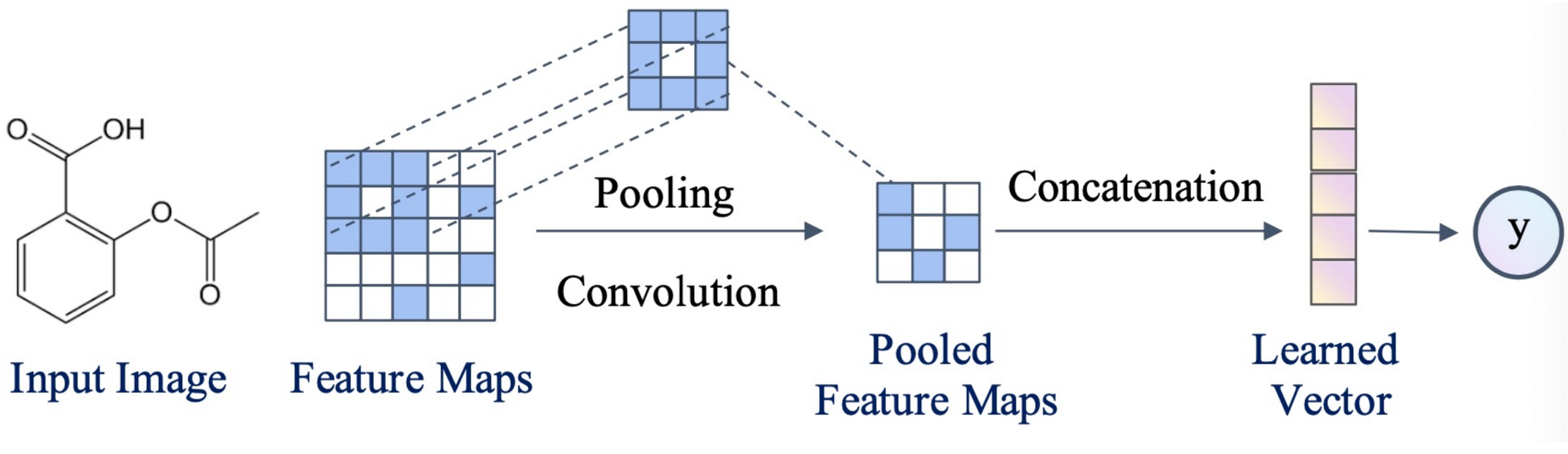
Application	Task	Metric	Purpose
Virtual screening	Molecular property prediction	Recall@k	Retrieval
Virtual screening	Molecular property prediction	Precision@k	Retrieval
Virtual screening	Molecular property prediction	AP@k	Retrieval
QSAR	Molecular property prediction	Accuracy	Classification
QSAR	Molecular property prediction	Recall	Classification
QSAR	Molecular property prediction	Precision	Classification
QSAR	Molecular property prediction	AUROC	Classification
QSAR	Molecular property prediction	AUPRC	Classification
QSAR	Molecular property prediction	MAE	Regression
QSAR	Molecular property prediction	RMSE	Regression
Drug design	Molecule generation	Validity	Distribution learning
Drug design	Molecule generation	Unique@k	Distribution learning
Drug design	Molecule generation	Novelty	Distribution learning
Drug design	Molecule generation	Diversity	Distribution learning
Drug design	Molecule generation	FCD	Distribution learning
Drug design	Molecule generation	KL divergence	Distribution learning
Drug design	Molecule generation	Scaffold similarity	Goal-directed design
Drug design	Molecule generation	Rediscovery	Goal-directed design

Metrics

- **validity** measures how well a model explicitly captures the chemical rules such as valency
- **uniqueness and diversity** examine whether the generative model collapses to producing only a limited set of molecules
- **novelty** indicates whether the model is overfitted to just memorize the training examples
- Fr´echet ChemNet Distance (FCD) is a measure of how close the distributions of the generated set are to the distribution of molecules
- **Kullback- Leibler (KL) divergence** measures the difference between two probability distributions

Machine Learning Models

- Traditional models
  - support vector machines (SVM)
  - K nearest neigh- bors (KNN)
  - random forest (RF)
  - naive bayes (NB), and logistic regression (LR)
- deep neural networks (DNNs)
  - Convolutional neural networks (CNNs) : 2015 ~

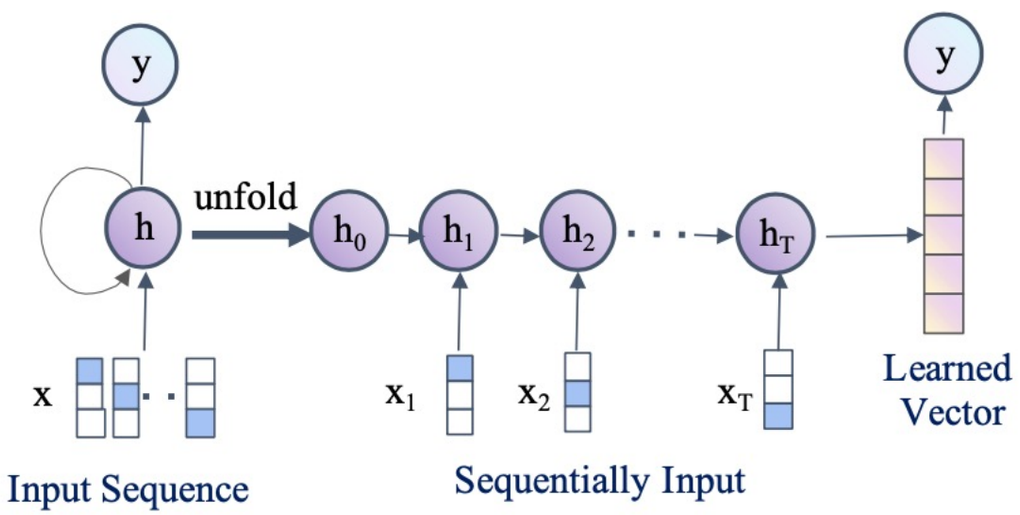


CNN

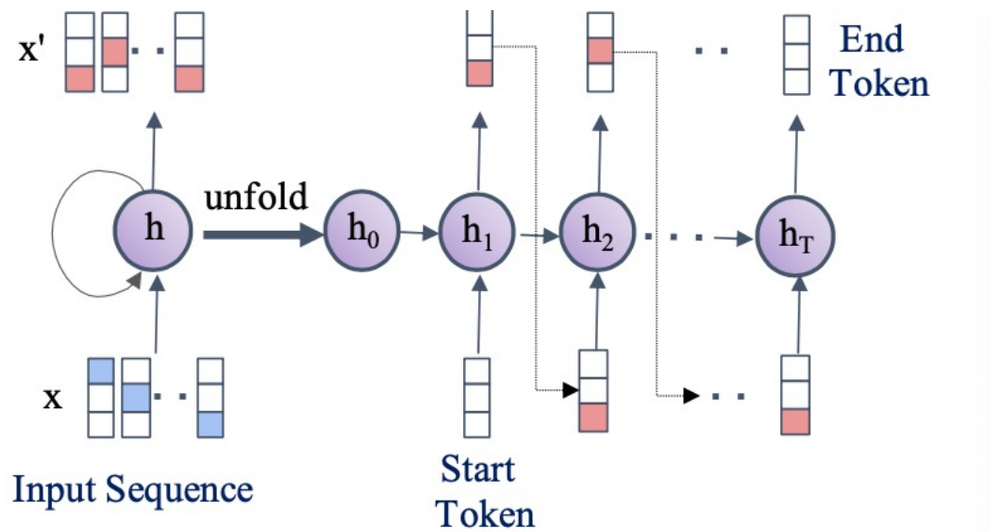
- Chemception
  - trained on images to predict free energy of solvation and inhibition of HIV replication
- Toxic Colors
  - toxicity classification with the images as input
- KekuleScope
  - uses Kekul´e structure images for molecular property prediction
- DEEPScreen
  - a large-scale DTI prediction system
- DECIMER
  - translate bitmap images of a molecule into a SMILES string, as an image captioning task

RNN

- Recurrent neural networks
  - Variations: LSTM, GRU
  - process sequential data. E.g., SMILES strings
- SMILES2Vec
  - learn features from SMILES and predicts chemical properties
- SmilesLSTM
  - perform DTI prediction



A. Recurrent Neural Networks in Prediction Mode



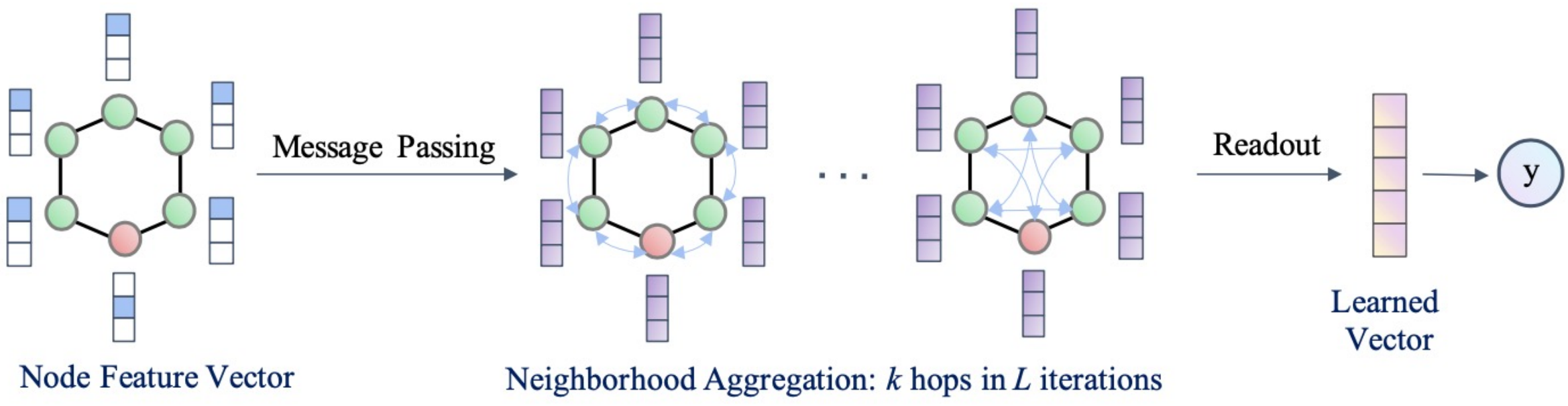
B. Recurrent Neural Networks in Generation Mode

Graph Neural Networks

- data represented in graphs with a set of nodes and edges
- Tasks:
  - node-level (e.g., node classification)
  - edge-level (e.g., link prediction)
  - graph-level (e.g., graph regression)

GNN

- Types of GNN
  - convolutional GNNs (ConvGNNs) and recurrent GNNs
- ConvGNNs
  - 1) Spectral-based: ChebNet, graph convolutional network (GraphConv)
  - 2) Spatial-based: message passing neural networks (MPNN), GraphSAGE, graph attention network (GAT), graph isomorphism network (GIN)
- MPNN:



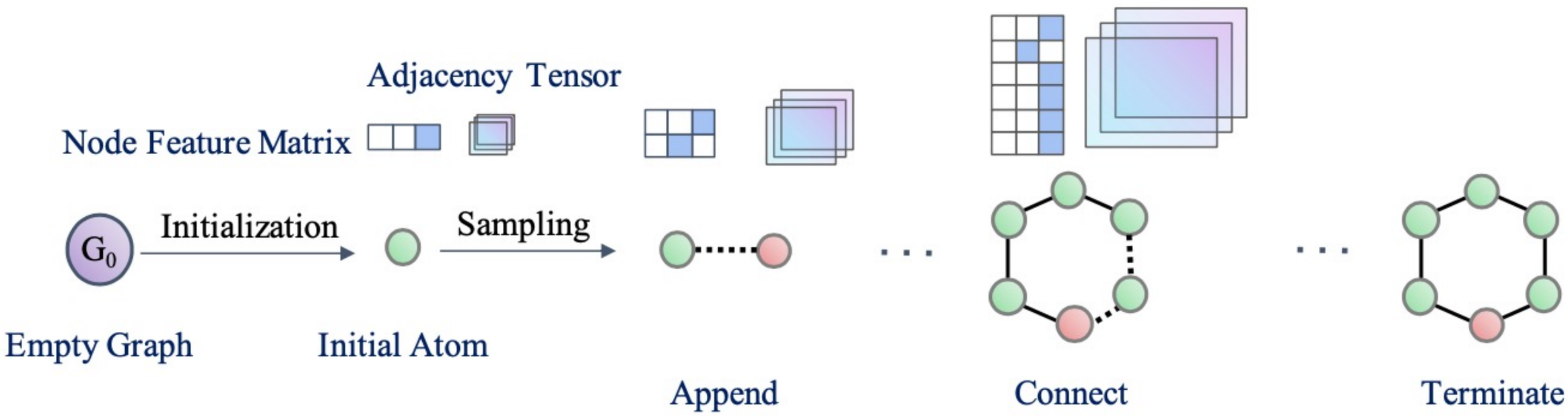
A. Graph Neural Networks in Prediction Mode

ConvGNNs

- directed MPNN (D-MPNN)
  - uses messages associated with directed edges (bonds) instead of nodes (atoms) used in MPNN
  - preventing repeated message passing
  - concatenate the 200 global molecular features calculated by RDKit with the learned features by D-MPNN for downstream predictions
- graph attention GNN
  - developed Attentive FP, which is able capture topologically adjacent atoms' interactions
- attention MPNN (AMPNN)
- edge memory neural network (EMNN)

For Molecule Generation

- MolMP - models the graph generation as a MDP (Markov Decision Process) problem
- for goal-directed drug design:
  - GCPN, MolDQN, DeepGraphMolGen, MNCE-RL.

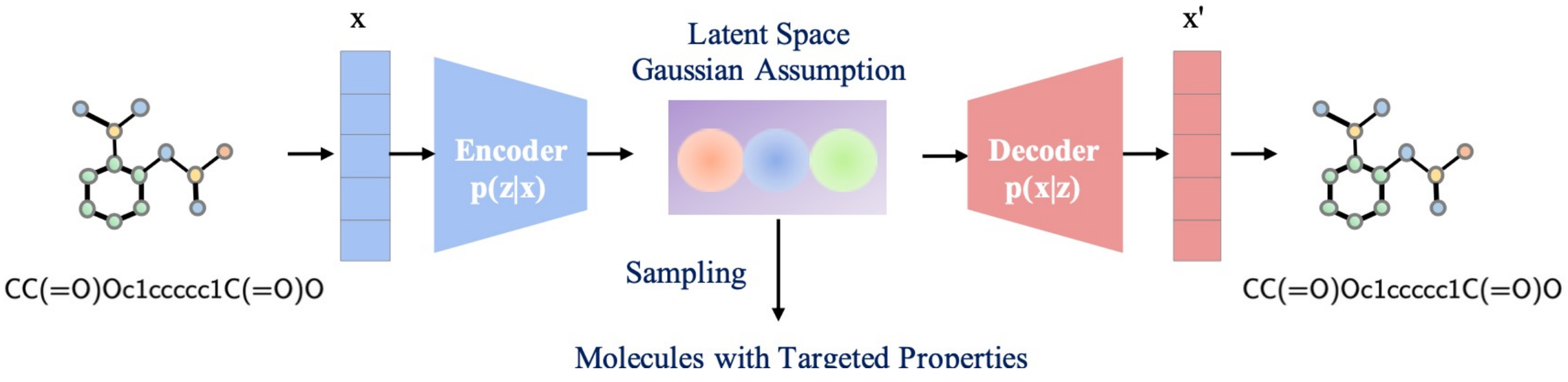


B. Graph Neural Networks in Generation Mode

Variational Autoencoders

- probablistic generative models
- the encoder maps high-dimensional data into latent space
- the latent space is regularized to be organized, ideally, through the KL divergence
- given x, the parameters of VAEs are optimized by minimizing the reconstruction loss and the KL divergence (= maximize the evidence lower bound (ELBO)

$$||x - D(E(x))||^2 + KL\big(N(\mu_x, \sigma_x), N(0,1)\big)$$



VAE Models

- applied on SMILES strings for molecule generation
- a constraint is applied in the autoencoder
  - by jointly training a physical property regression model to organize the VAE's latent space subjected to the desired property value
- GrammarVAE
- syntax-directed VAE
- semi-supervised VAE (SSVAE)
- conditional VAE (CVAE)
- constrained graph VAE (CGVAE)
- NeVAE, GTM VAE, CogMol
- adversarial autoencoder (AAE)
  - which replaces the KL divergence with an adversarial objective

VAE Models

- GraphVAE
  - graph representations for generation purpose
- junction tree VAE (JT-VAE)
  - a molecular graph is first mapped into a junction tree via a tree decomposition algorithm
  - the learned latent space of the junction tree can be used to search for substructures, which then assemble into molecules with specific properties
- Regularized VAE
  - regularize output distribution of the decoder, improving the validity
- molecular hypergraph grammar VAE (MHG-VAE)
  - molecular graph is described as a hypergraph and the grammar VAE is trained by inputting the grammar for sequence production of the hypergraph

VAE Models

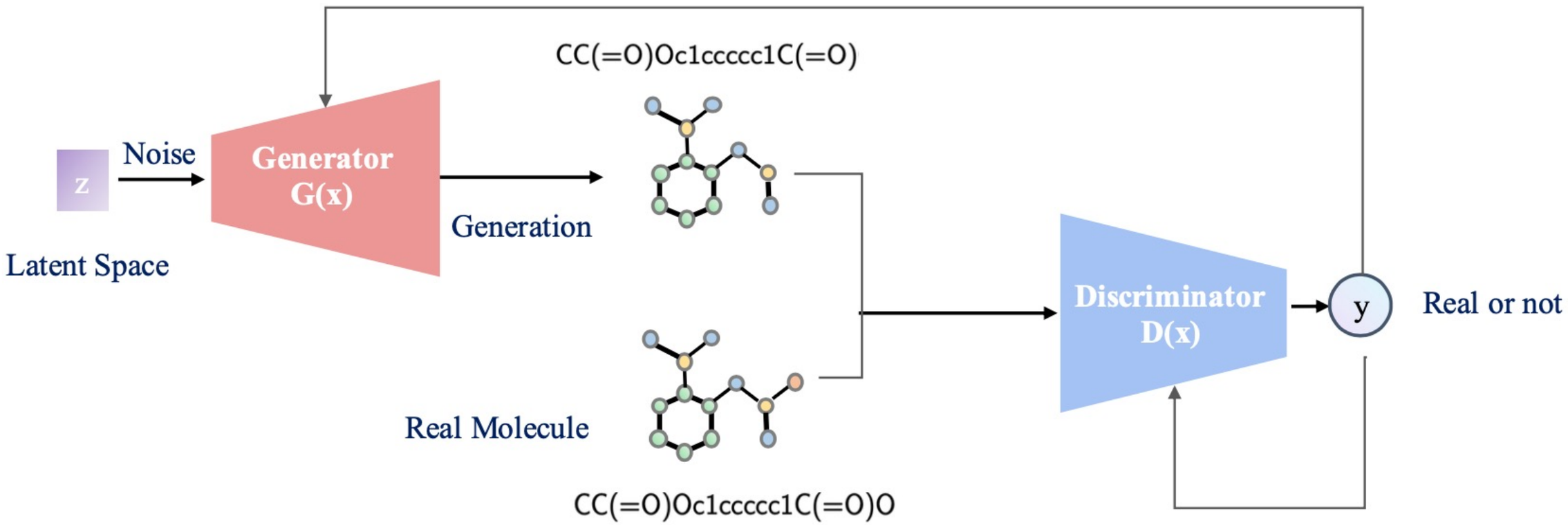
- ScaffoldVAE
  - graph-based VAE to retain a particular scaffold (i.e., substructure)
- CORE
  - combining scaffolding tree generation and adversarial training
- hierarchical graph VAE (HierVAE)
  - employ larger and more flexible graph motifs as building blocks for molecules
  - encoder produces a multi-resolution representation for each molecule in a fine-to-coarse fashion, from atoms to connected motifs

GAN

- trained by the min-max loss, which alternatively optimizes the generator and the discriminator using a min-max objective

$$\min_G \max_D \mathcal{L}(G, D) = \mathbb{E}_{x \sim p_x}[\log(D(x))] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$$

- objective-reinforced GANs (ORGAN), MolGAN, …

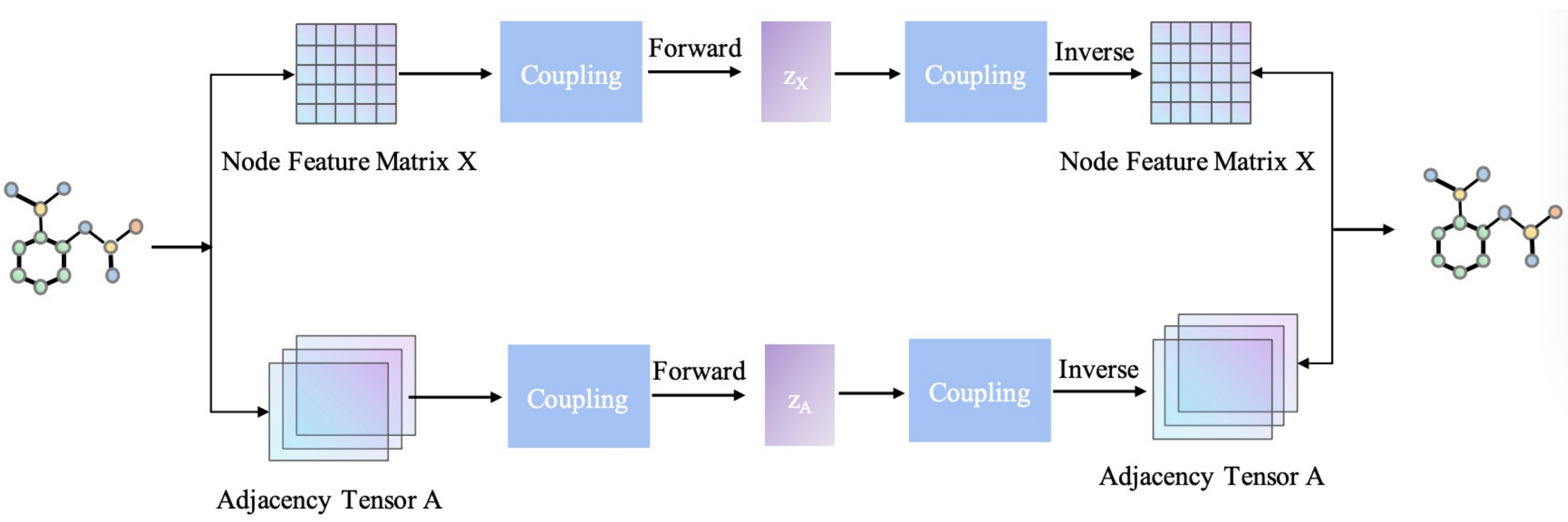


Normalizing Flow Models

- generative model learning an invertible mapping between complex distributions and simple prior distributions
- flow models enable efficient one-shot inference and 100% reconstruction of the training data
- able to exactly reconstruct all the input data without duplicates due to the precise likelihood maximization
- models:
  - Non-linear Independent Component Estimation model (NICE)
    - tractable calculation for reversible transformations; the affine coupling layers
  - Real-valued Non-Volume Preserving model (RealNVP)
  - Glow model

Flow models

- GraphNVP
  - the first flow model for molecular graph generation
  - two steps; generation of an adjacency tensor and generation of node attributes, which yields the exact likelihood maximization on the graph with two reversible flows

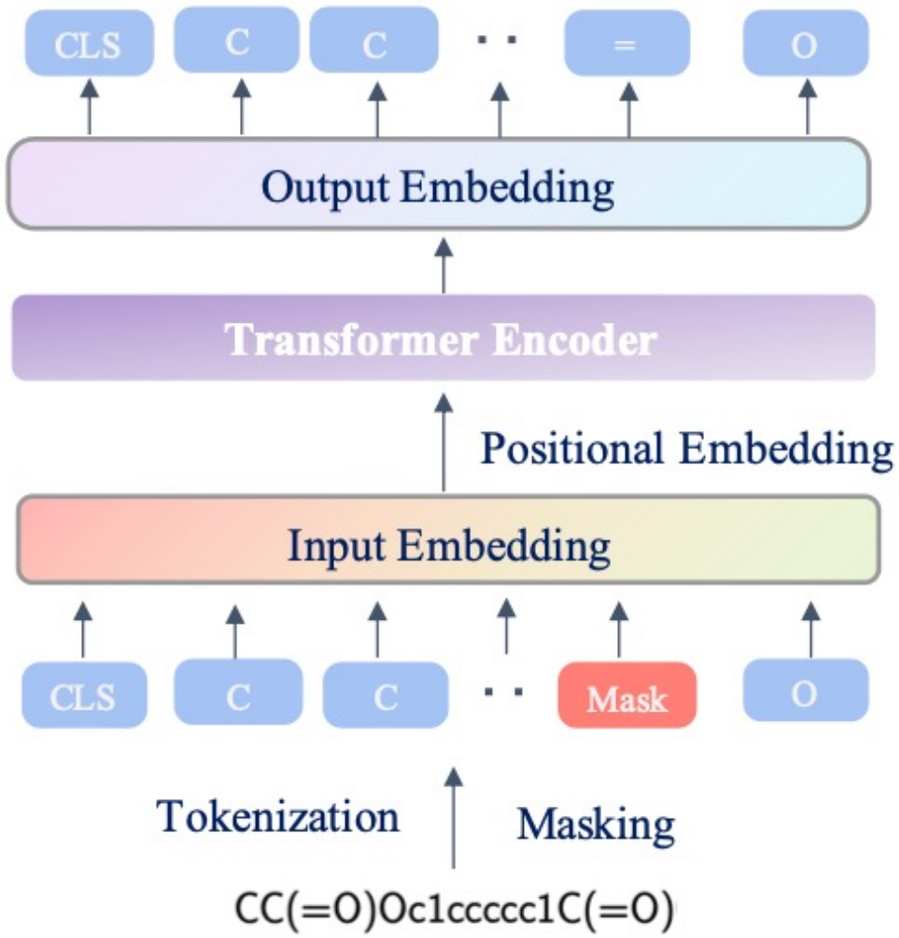


Flow models

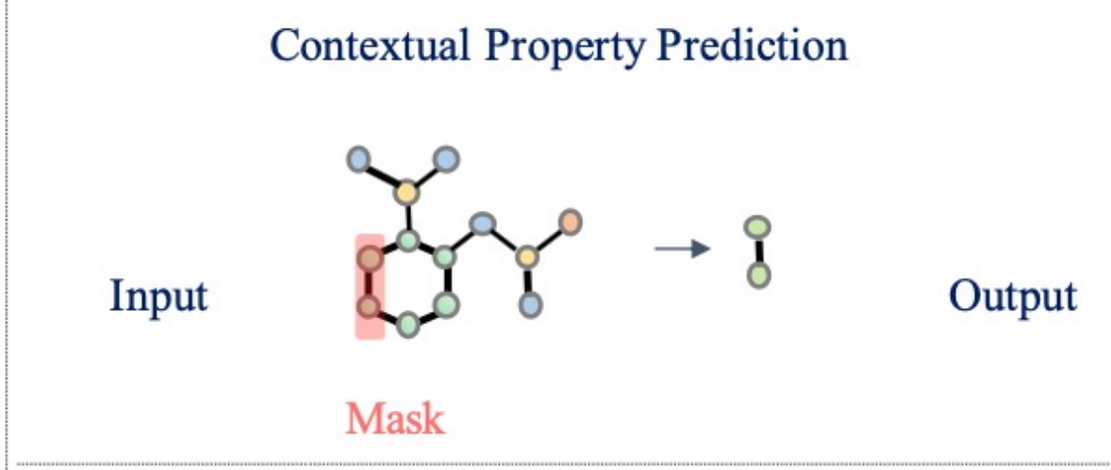
- graph residual flow (GRF)
  - invertible flow model for molecular graph generation based on residual flows
  - needs much less parameters
  - generate molecular graphs in a single-shot manner
- GraphAF (autoregressive flow-based model)
  - adopts an iterative sampling process to leverage chemical domain knowledge - generate molecules of 100% validity
  - can be further finetuned with RL, which achieves better performance on molecular property optimization compared to JT-VAE
- MoFlow : applies a validity correction to the generated graph
- GraphDF : learn a discrete latent representation

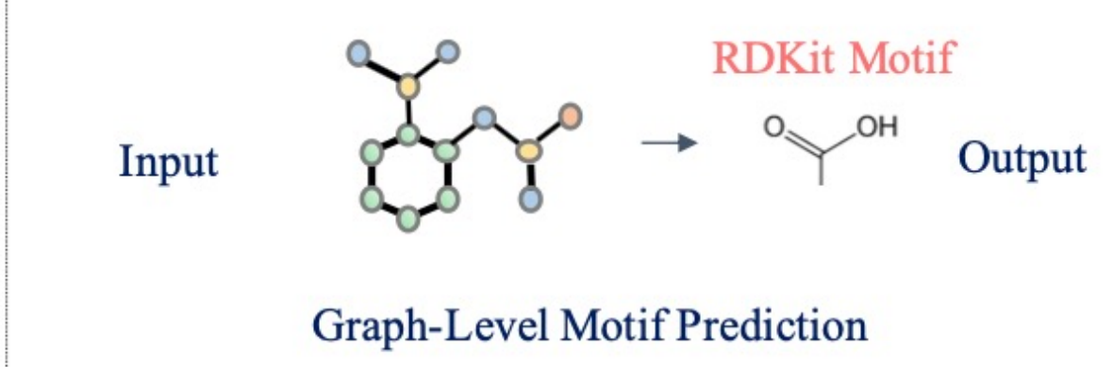
Transformers

- built with the self-attention mechanism
- does not use recurrent connections (unlike RNN)



A. Masked Language Modeling on SMILES Strings





B. Context and Motif Prediction on Molecular Graphs

Models

- SMILES-BERT (2019)
  - For molecular property prediction
- SMILES Transformer
- Chem- BERTa : using RoBERTa
- MolBERT
- GROVER
  - learn graph representations with the message passing transformer
- MoleculeChef
  - generate the reactants for a given product, similar to machine translation
- protein-specific molecule generation
  - input is the amino acid sequence of the target protein and the output are ligands in the SMILES representation

Lack of Labels

- low-data problem
  - exhibit high sparsity, and can be heavily biased and noisy
  - low novelty and diversity for molecule generation

Self-Supervised Learning

- Unsupervised learning
  - focuses on detecting patterns in data without labels, such as clustering
- self-supervised learning
  - aims to recover the data
  - two main types, i.e., generative and contrastive self-supervised learning
- MolCLR (molecular contrastive learning of representations)
  - use molecular graphs for molecular property prediction
  - augmentation : atom masking, bond deletion, and subgraph removal
  - contrastive learning: use positive vs negative data pairs

Meta Learning

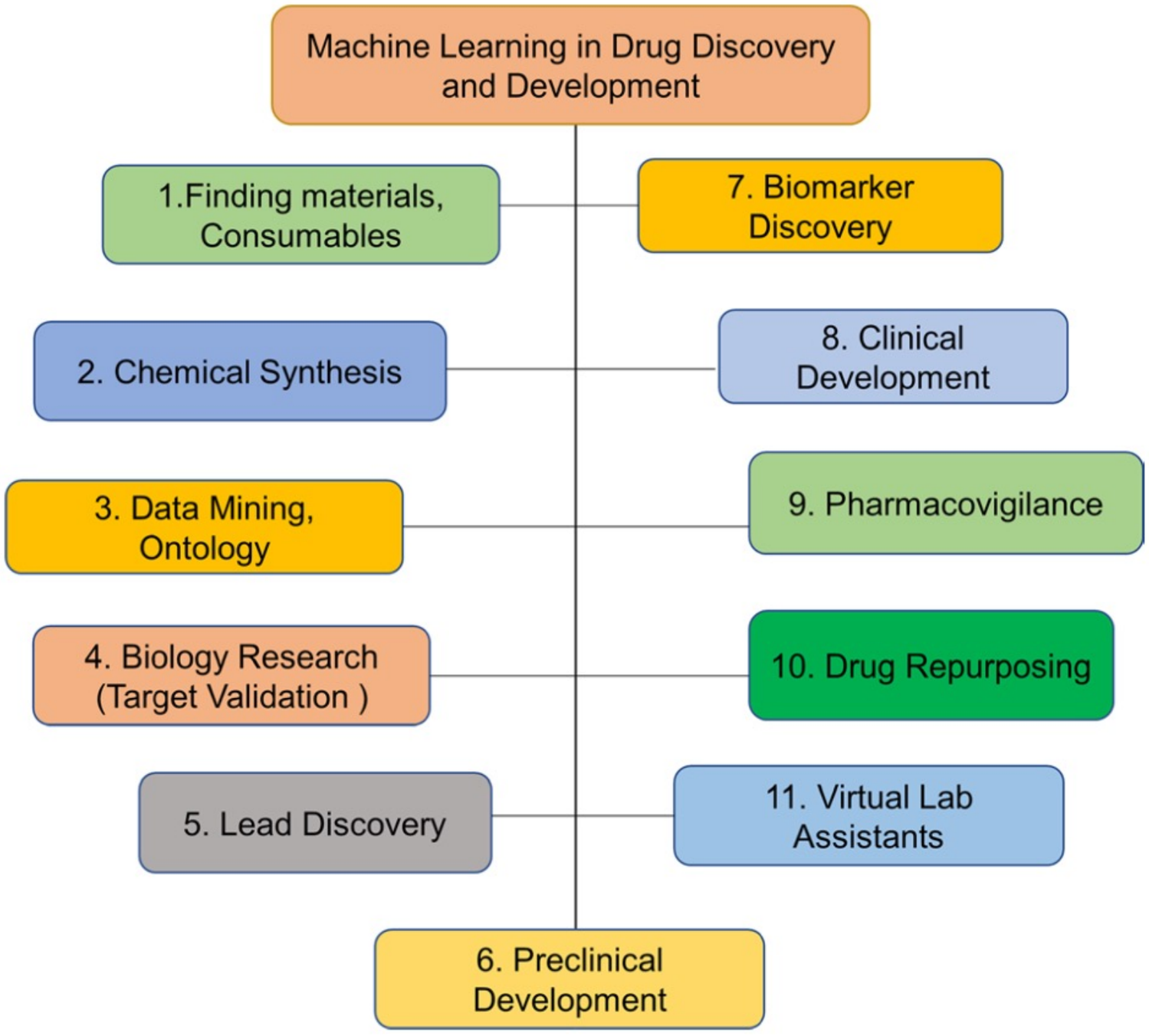
- learn a learner to be adapted to new tasks
- few-shot learning
  - generalize with a few examples
  - embeddings can be compared to the exiting labeled molecules for more accurate prediction
- metric learning
  - deals with data with mixed distribution due to the activity cliffs

Reinforcement Learning

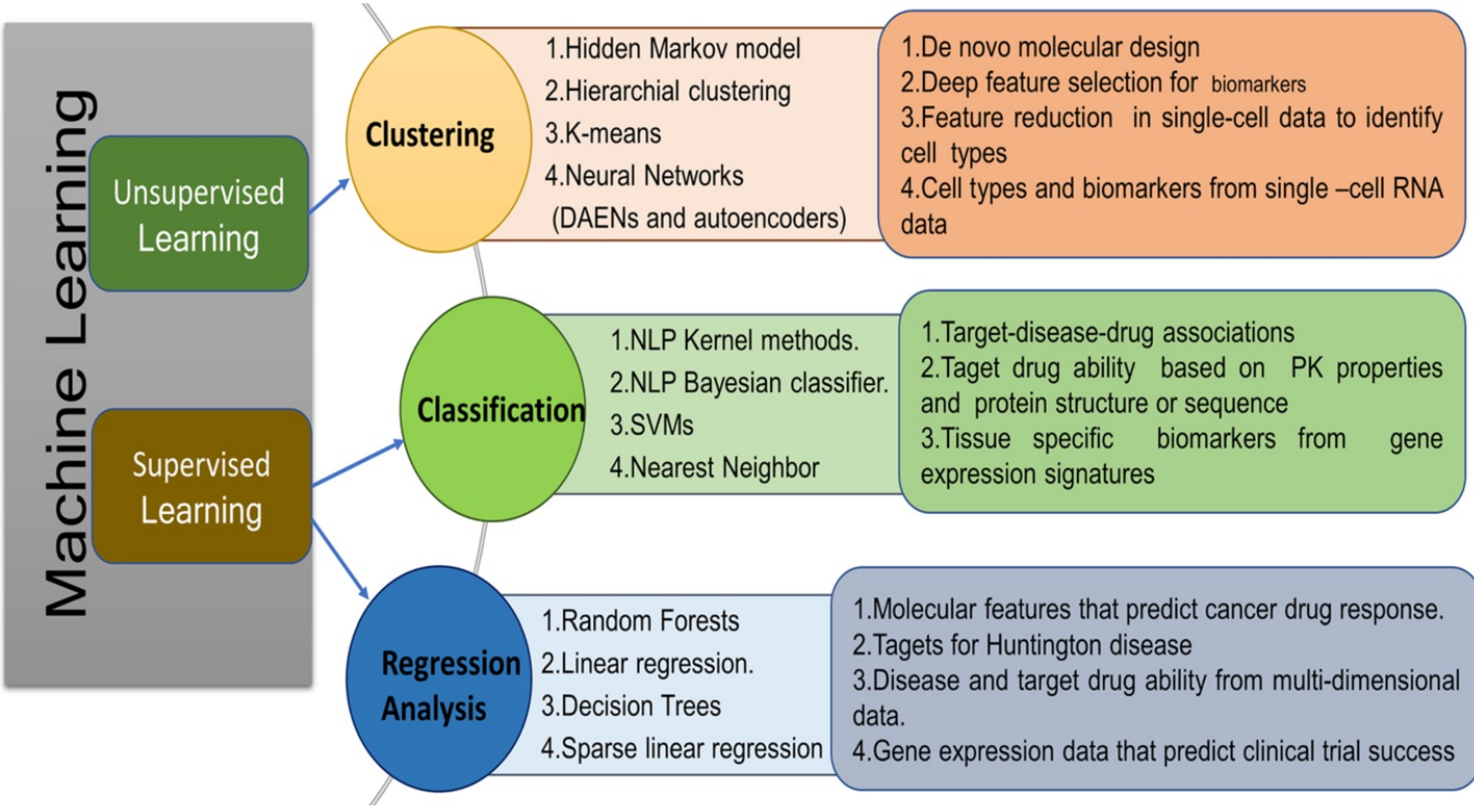
- How to design molecules with the desired properties?
  - the objective functions defined in the latent space is usually non-convex
- RL
  - how an agent should take actions in a certain state so as to maximize a reward or return
  - 1) value-based (e.g., Q-learning)
  - 2) policy-based (e.g., policy gradient)
  - 3) hybrid (e.g., actor-critic)
- However, drug design is a multi-objective optimization problem
  - lack of a complete knowledge of the rewards for all the states
  - a solution for the exploration-exploitation trade-off **is active learning**

Machine Learning in DD

Scope

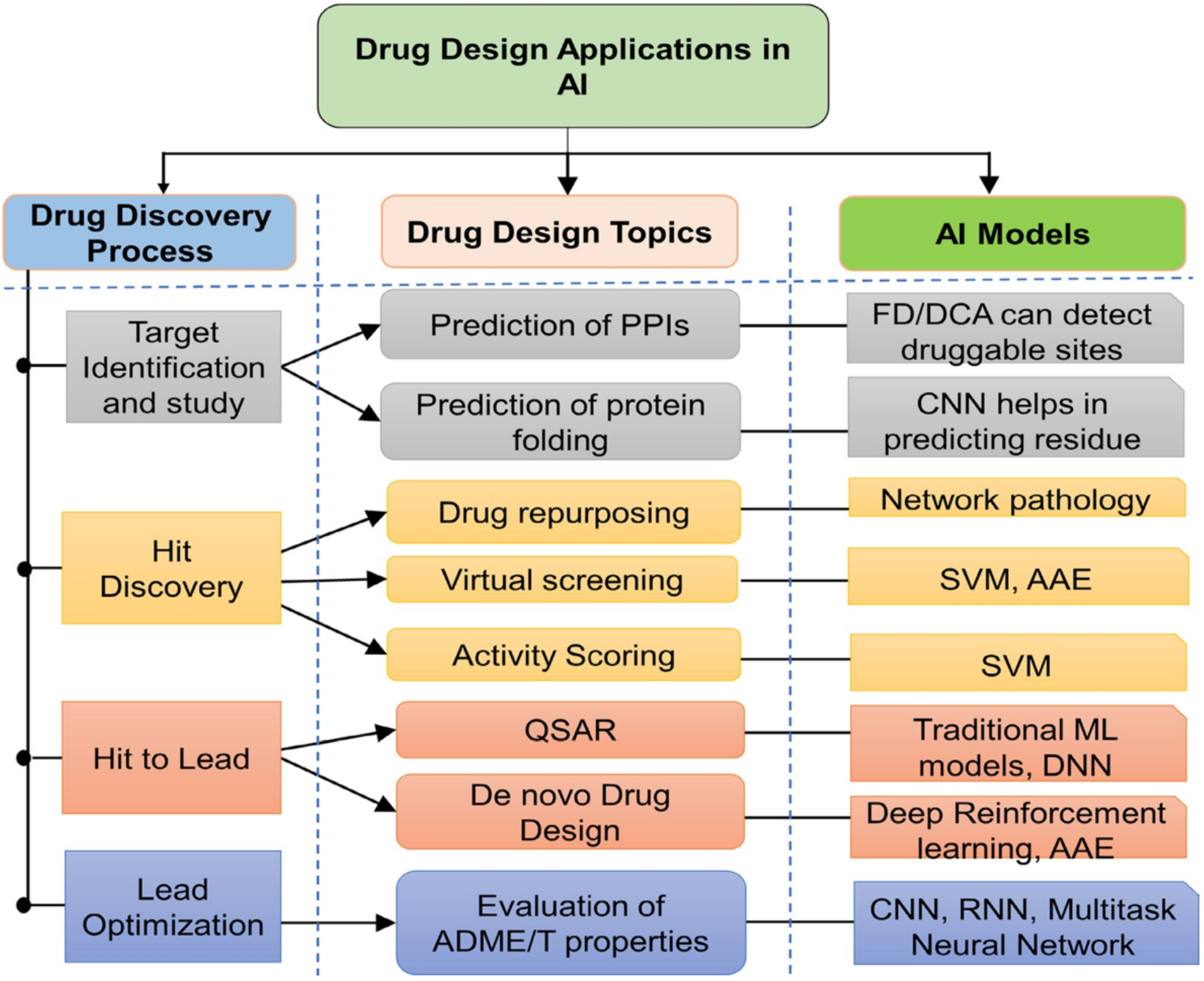

Machine Learning in Drug Discovery and Development
1.Finding materials, Consumables
2. Chemical Synthesis
3. Data Mining, Ontology
4. Biology Research (Target Validation)
5. Lead Discovery
6. Preclinical Development
7. Biomarker Discovery
8. Clinical Development
9. Pharmacovigilance
10. Drug Repurposing
11. Virtual Lab Assistants

Machine learning



Data Types

- textual data
- Images
- assay information
- Biometrics
- high dimensional omics data


Drug Design Applications in AI
Drug Discovery Process
Drug Design Topics
AI Models
Target Identification and study
Hit Discovery
Hit to Lead
Lead Optimization
Prediction of PPIs
Prediction of protein folding
Drug repurposing
Virtual screening
Activity Scoring
QSAR
De novo Drug Design
Evaluation of ADME/T properties
FD/DCA can detect druggable sites
CNN helps in predicting residue
Network pathology
SVM, AAE
SVM
Traditional ML models, DNN
Deep Reinforcement learning, AAE
CNN, RNN, Multitask Neural Network