ICASSP 2024
IEEE International Conference on Acoustics, Speech and Signal Processing
14~19 April 2024 | Seoul, Korea

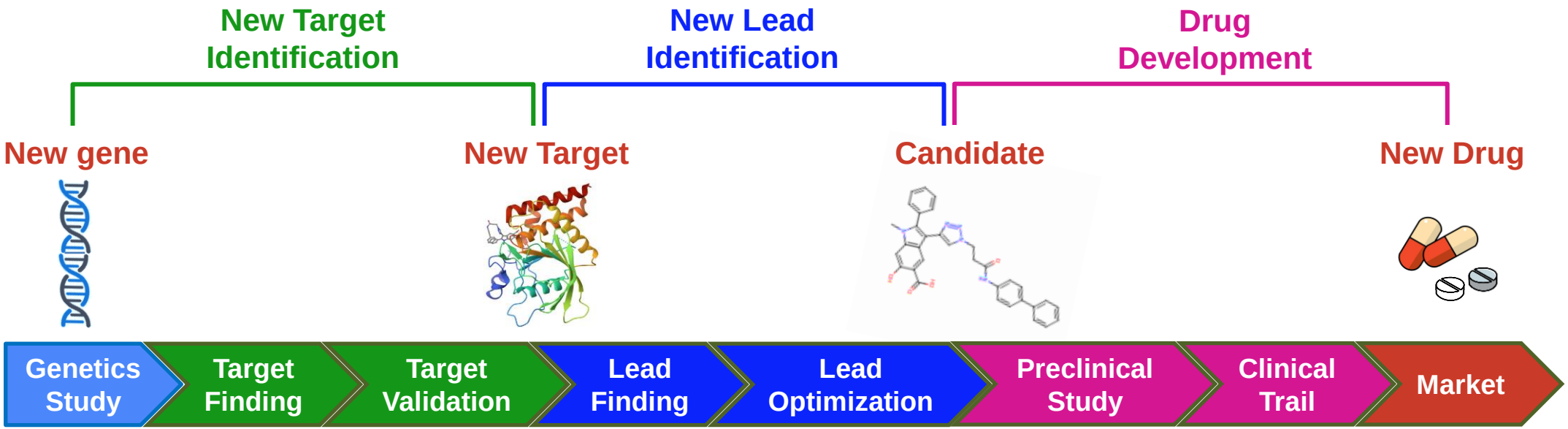
TD-GPT: TARGET PROTEIN-SPECIFIC DRUG MOLECULE GENERATION GPT

Zheng-da HE[1,2] , Linjie Chen[2] ,
Jiaying Xu[2], Hao Lv[2], Rui-ning Zhou[2], Jianhua Hu[2], Yadong Chen[2] , Yang Gao[1]
1 State Key Laboratory for Novel Software Technology, Nanjing University
2 Laboratory of Molecular Design and Drug Discovery, China Pharmaceutical University

OUTLINE

1 INTRODUCTION

- **The Drug Development Process**



$ 2.6 billion

10 - 15 years

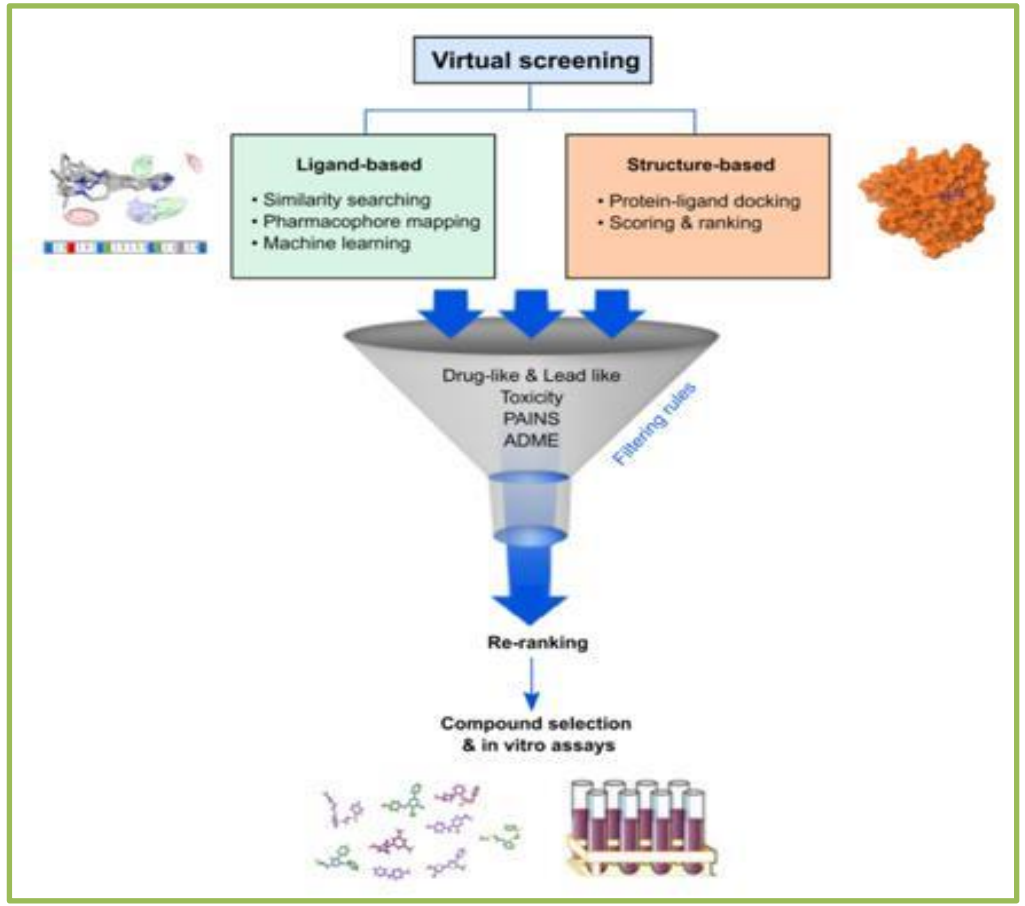
- **High costs, High risks, Long cycles, and Low success rates**
- **Lack of reliable key technologies for discovering lead structures**

Journal of Health Economics, 2016, 47, 20-33.

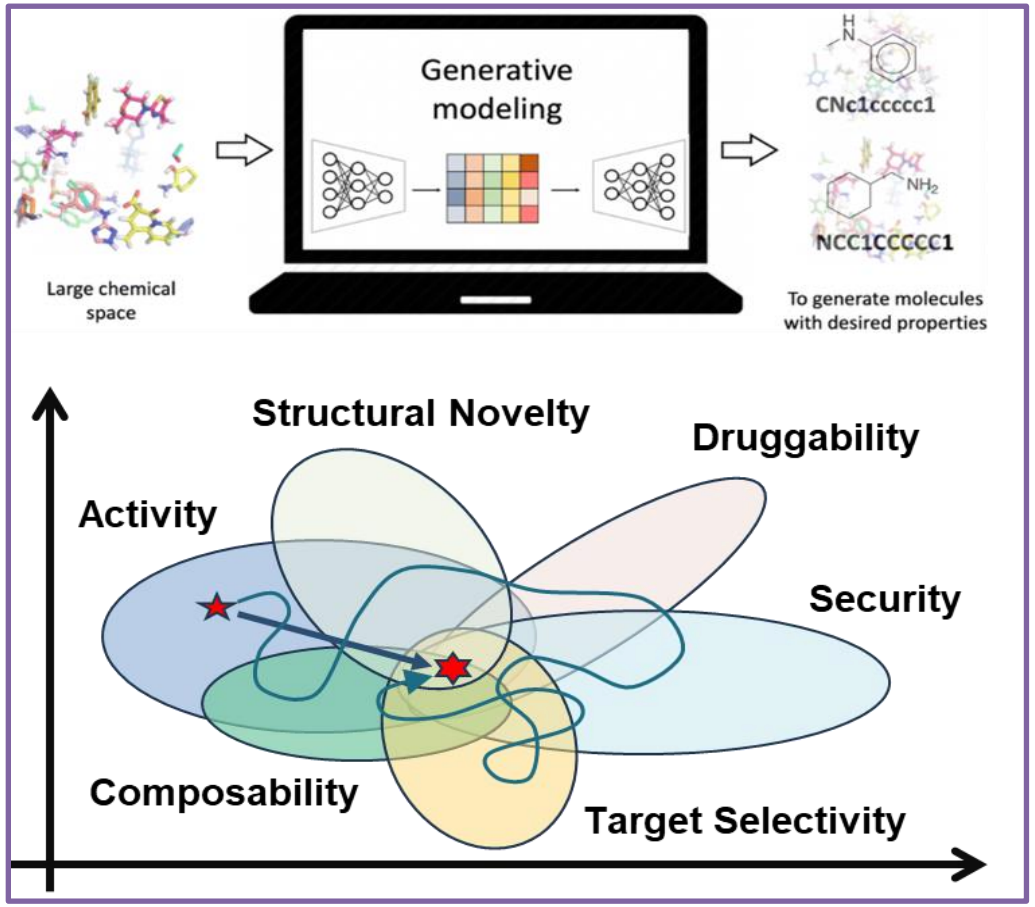
Future Medicinal Chemistry, 2020, 12, 939-947

1 INTRODUCTION

- **Computational Intelligence in the Discovery of Lead**



Virtual Screening



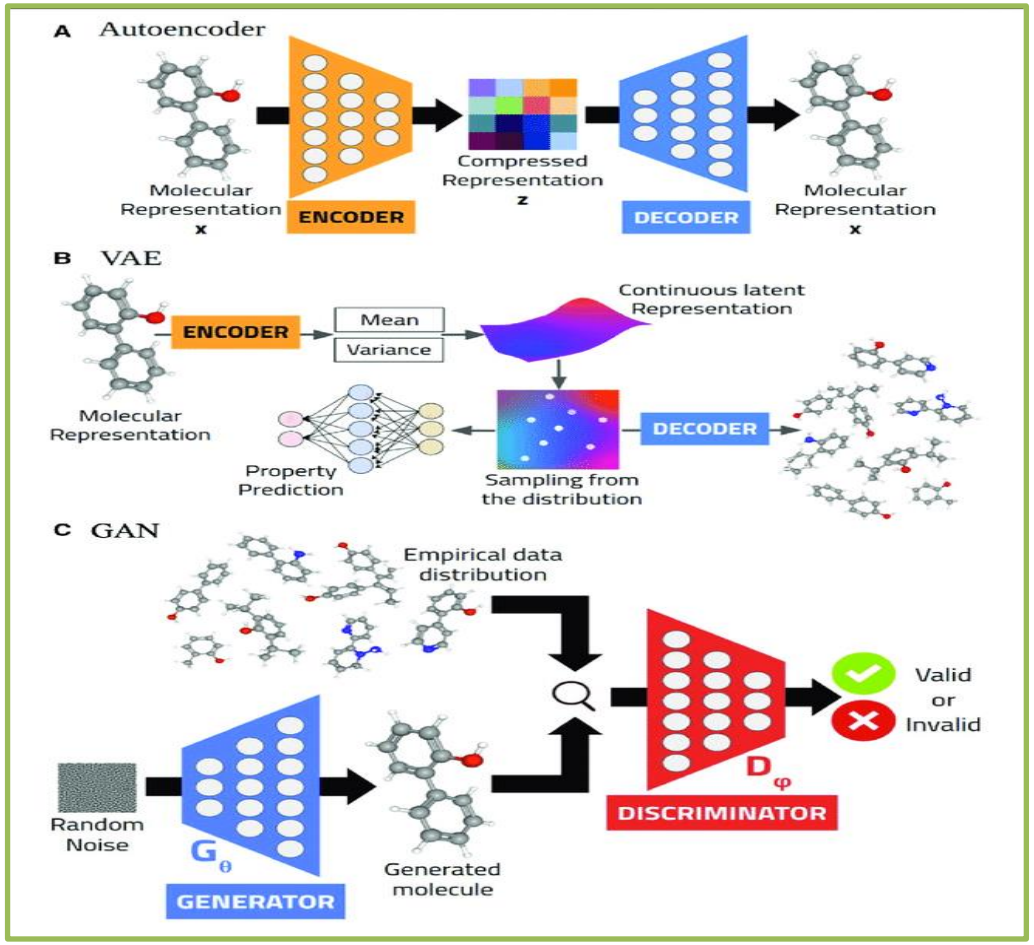
Drug Molecule Generation

J Mol Model, *2021*, 27: 71

https://www.profacgen.com/computer-aided-drug-design.htm

1 INTRODUCTION

- **Current Deep Learning Techniques for Molecular Generation**



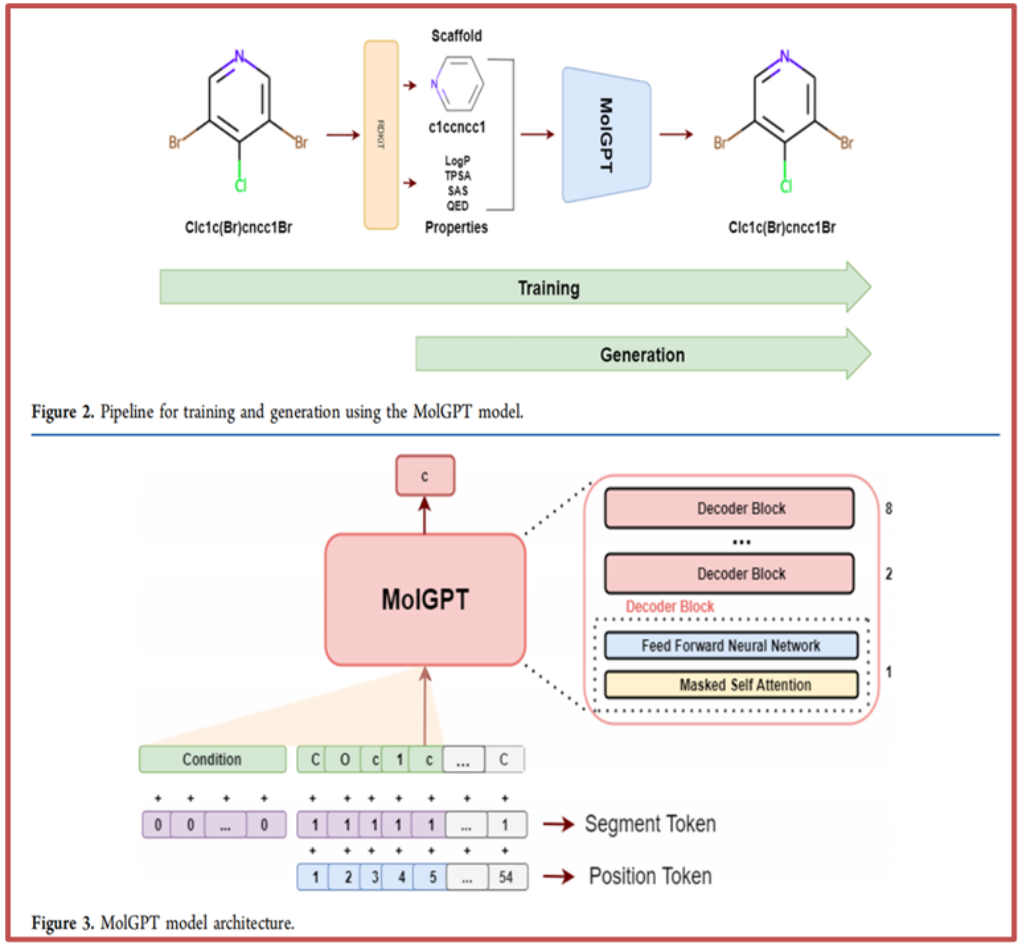


Figure 2. Pipeline for training and generation using the MolGPT model.

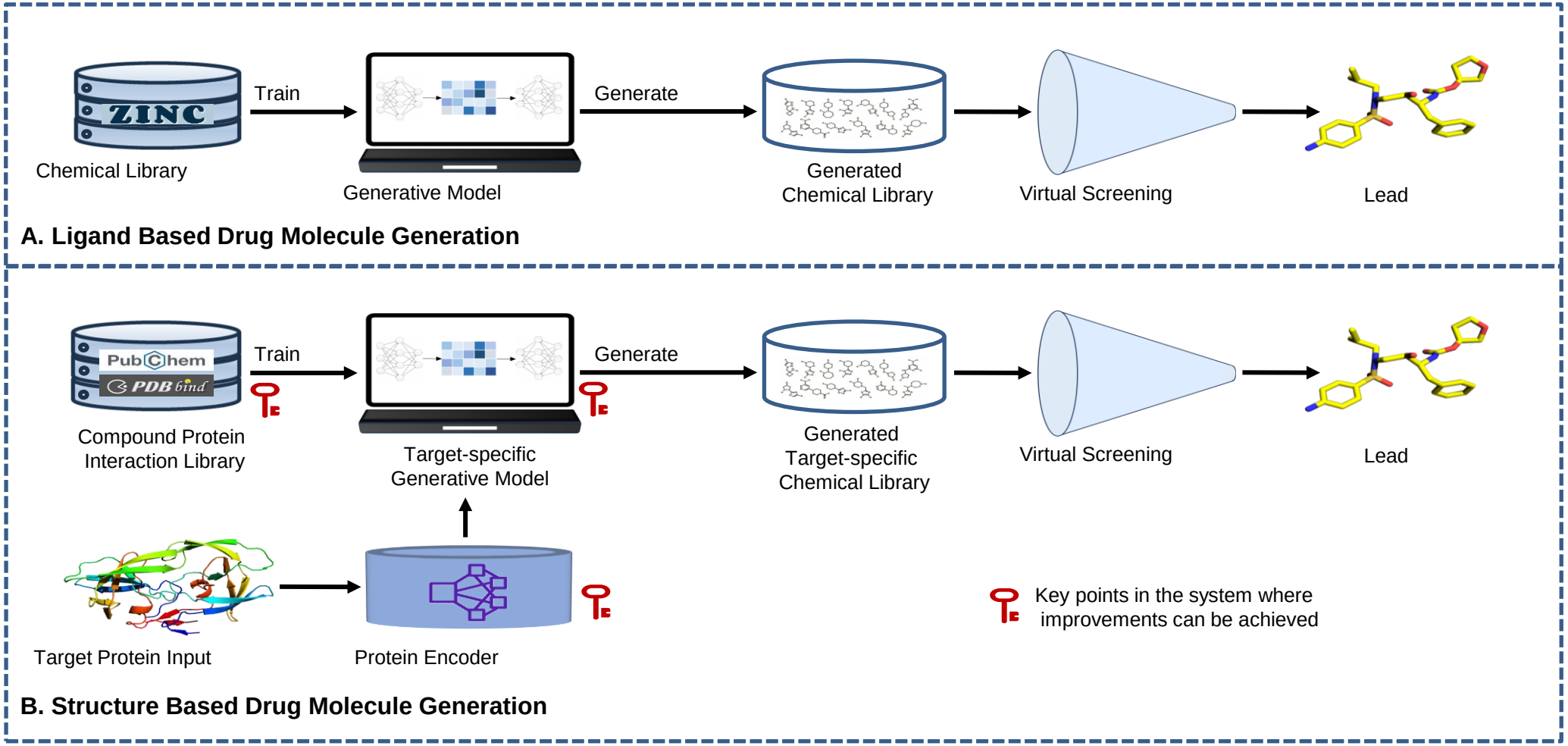
Figure 3. MolGPT model architecture.

Frontiers in Materials, 2022.

J. Chem. Inf. Model. 2022, 62, 2064–2076

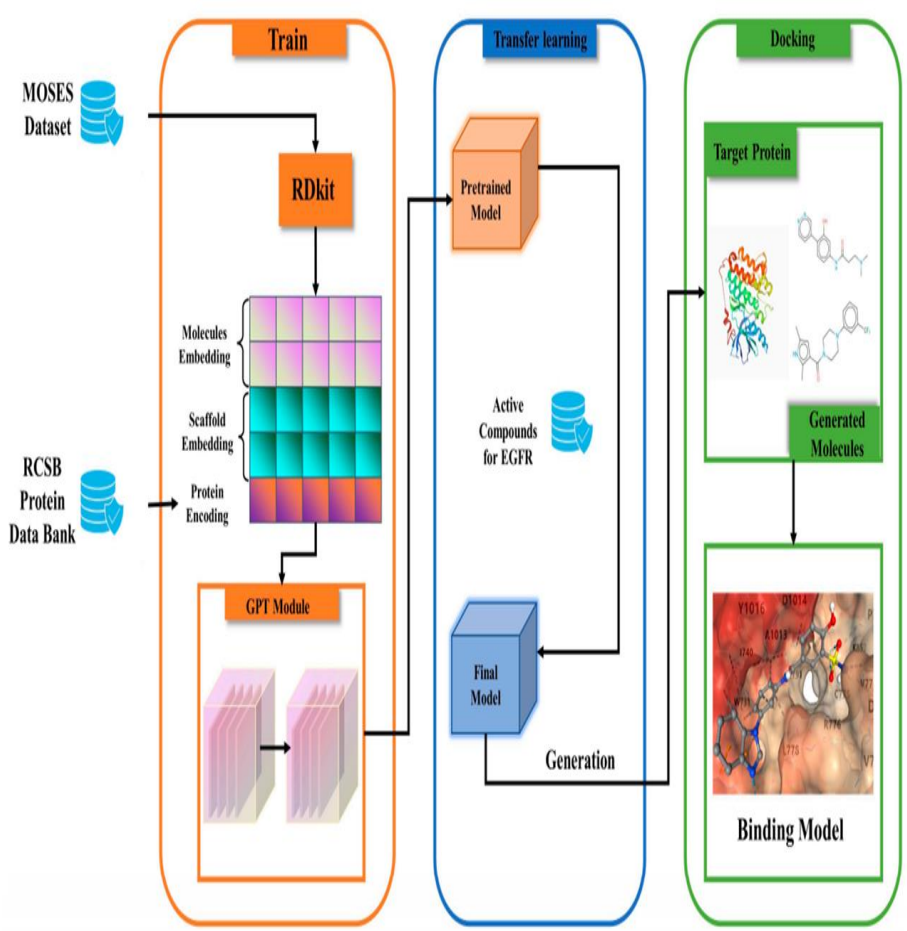
1 INTRODUCTION

- **Ligand Based and Structure Based Drug Molecule Generation**

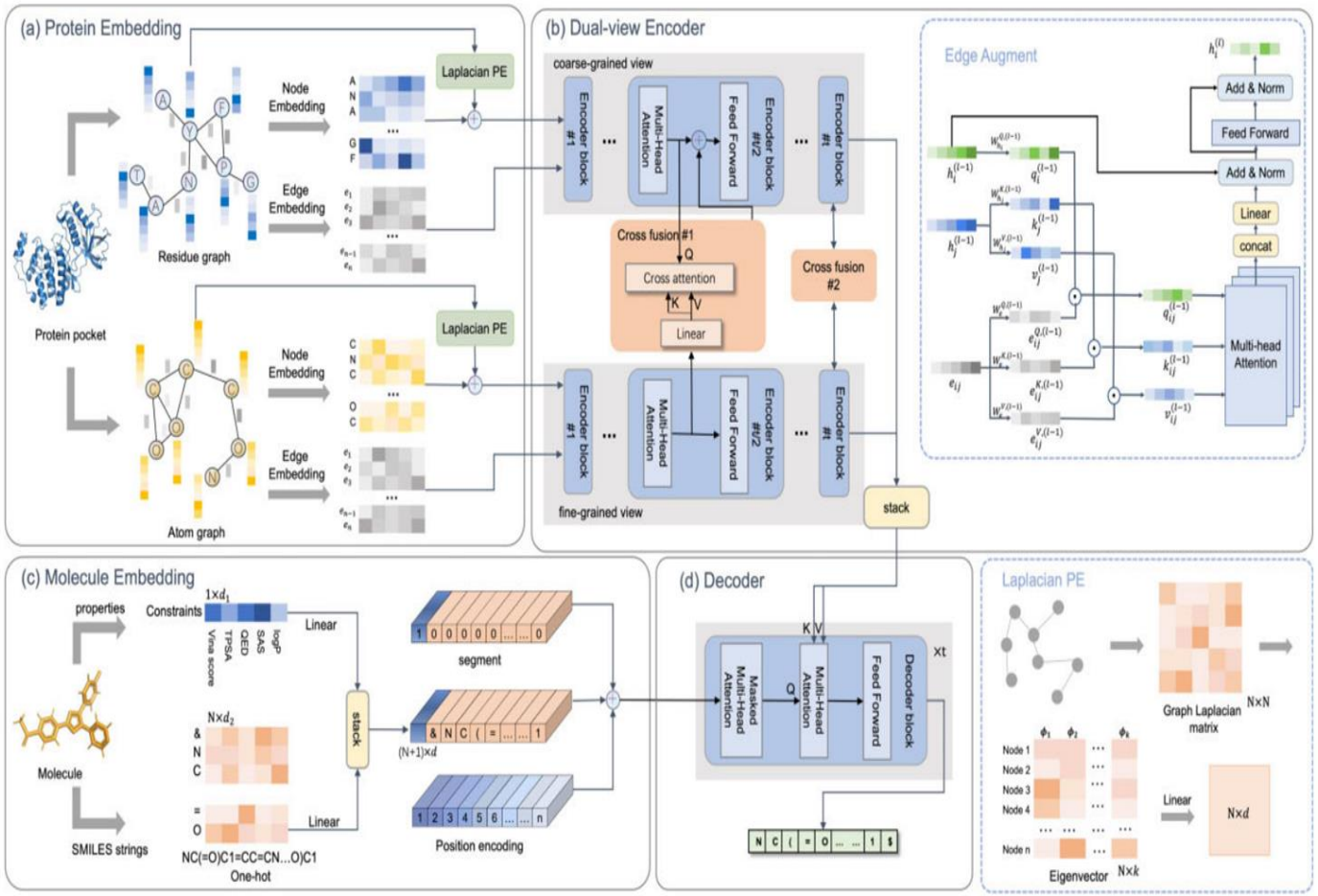


1 INTRODUCTION

- **Related Works – PETrans, CProMG**



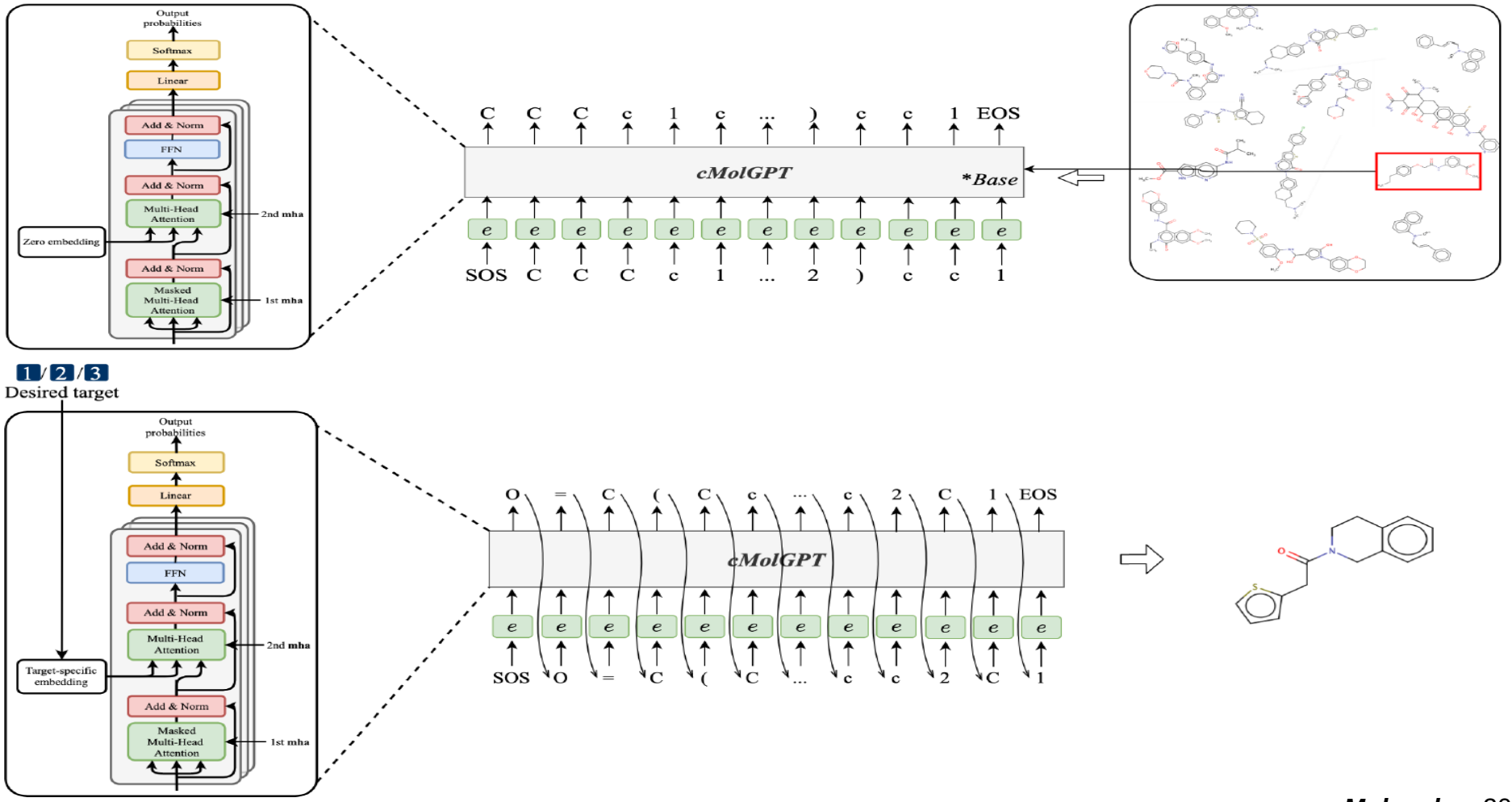
Int. J. Mol. Sci. 2023, *24*, 1146



Bioinformatics, 2023, 39, i326-i336

1 INTRODUCTION

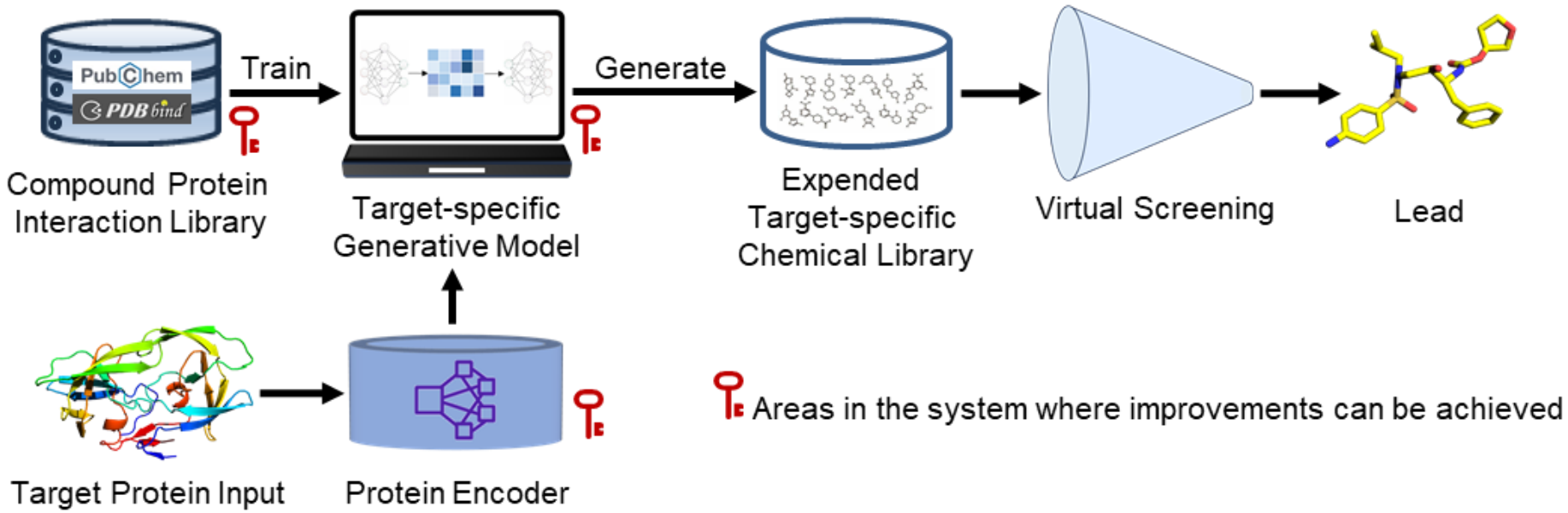
- **Related Works - cMolGPT**



Molecules, 2023, 28, 4430

1 INTRODUCTION

- **Challenges**

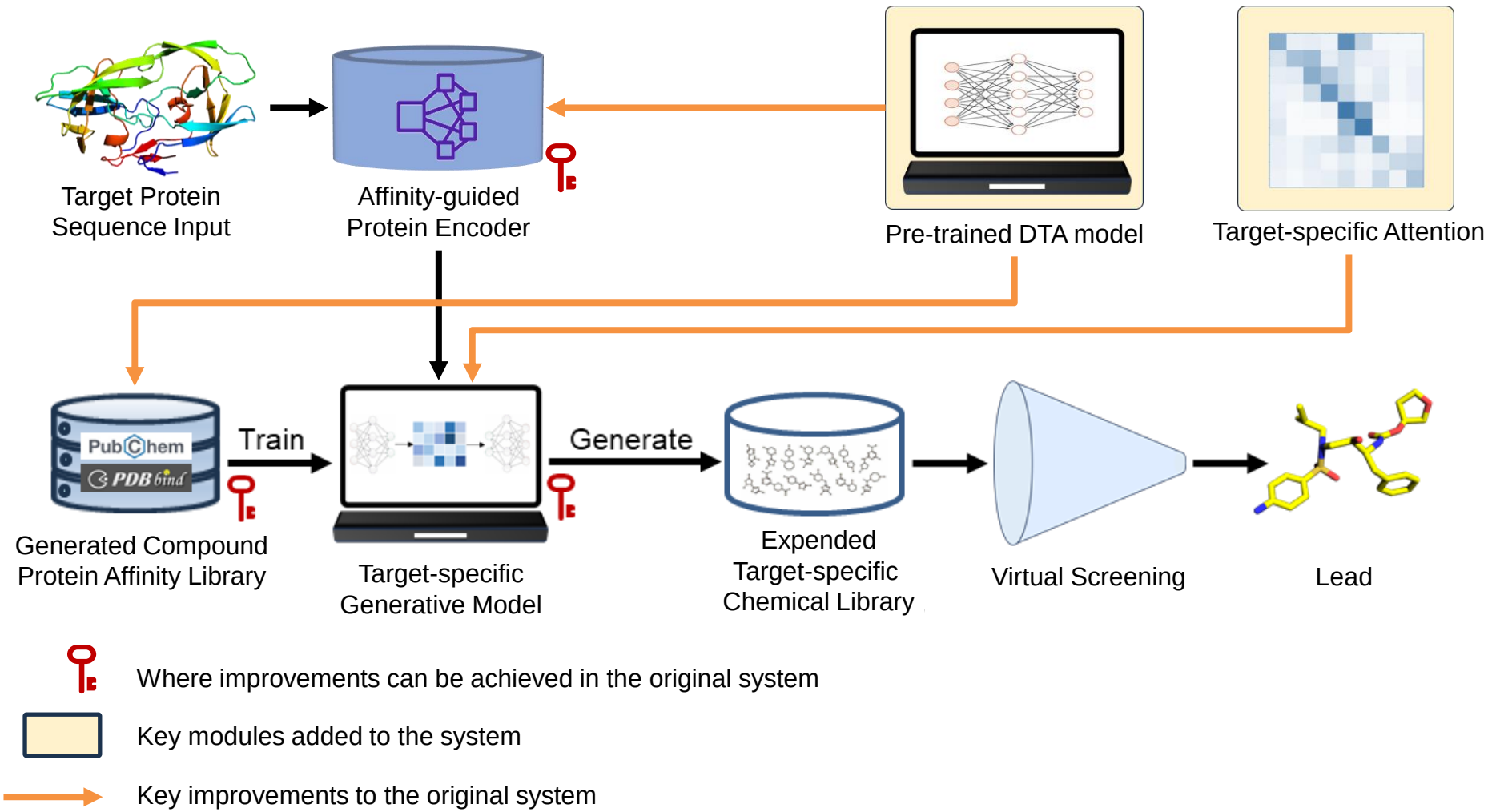


Three key points in the structure-based molecular generation model workflow correspond to three key challenges:

- **The protein-ligand binding database**
- **The type of target protein information**
- **The generative model architecture design**

1 INTRODUCTION

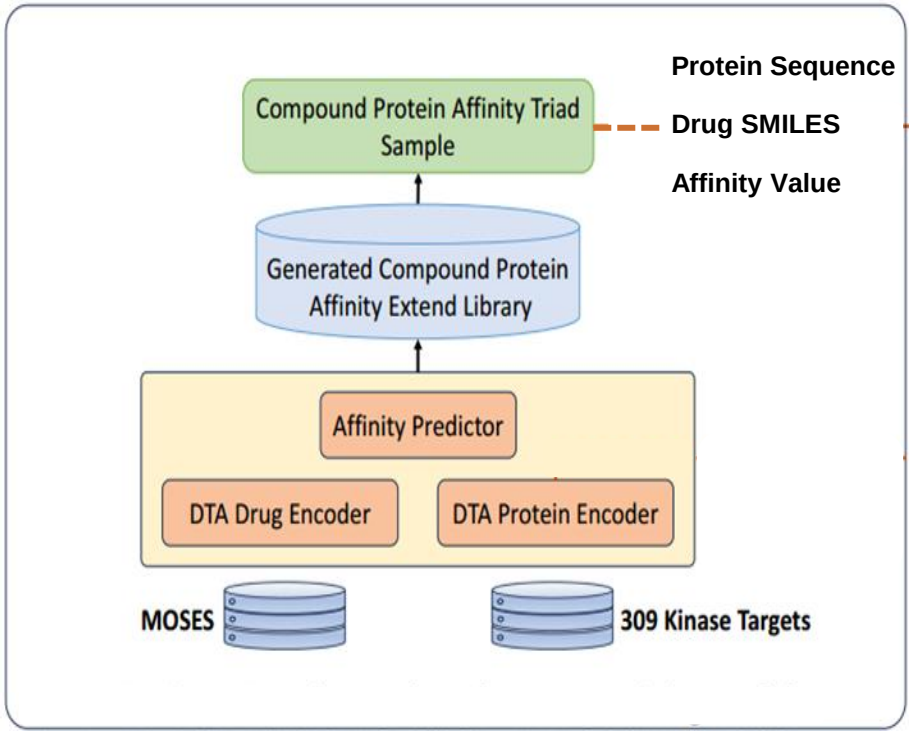
- **Solutions**

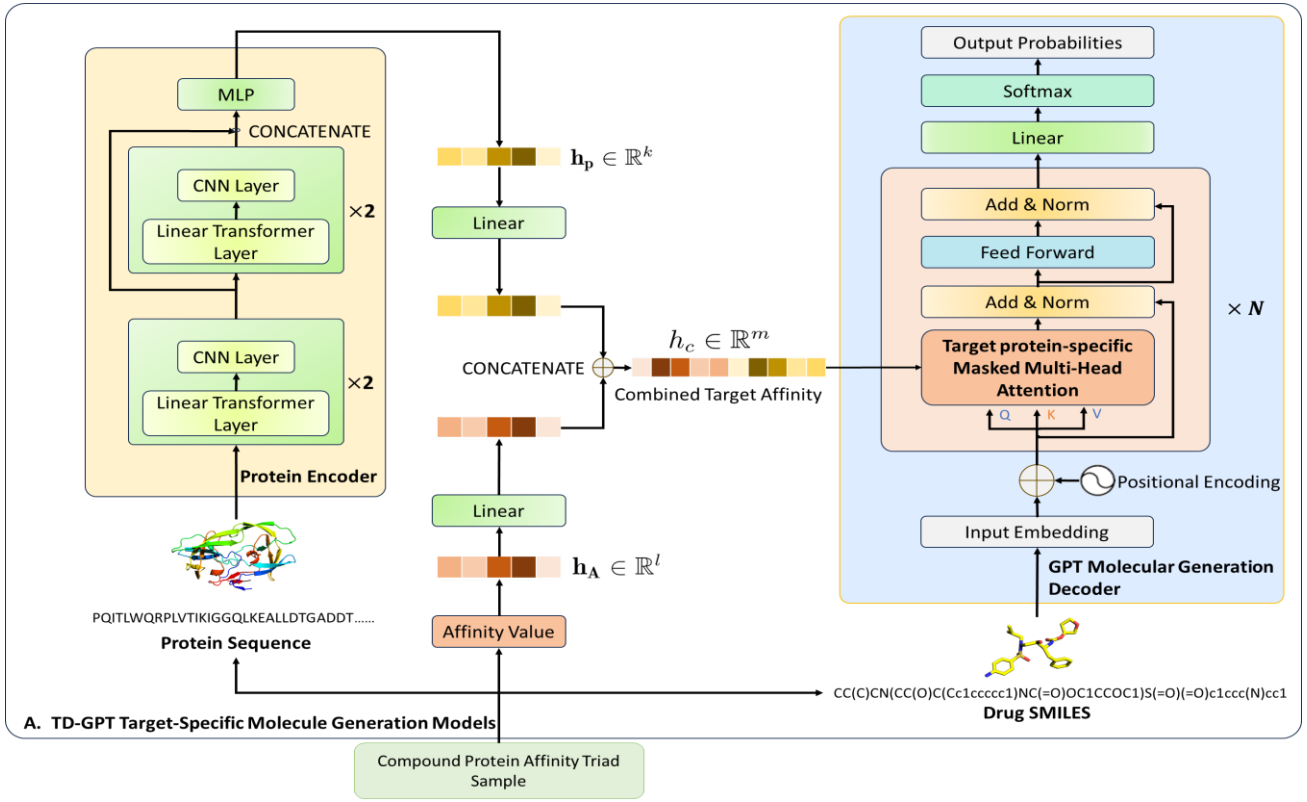


2 METHOD

- ## 2.1 Two Models and a Four-stage Workflow

Two Models





1. LT-DTA drug-target affinity prediction model

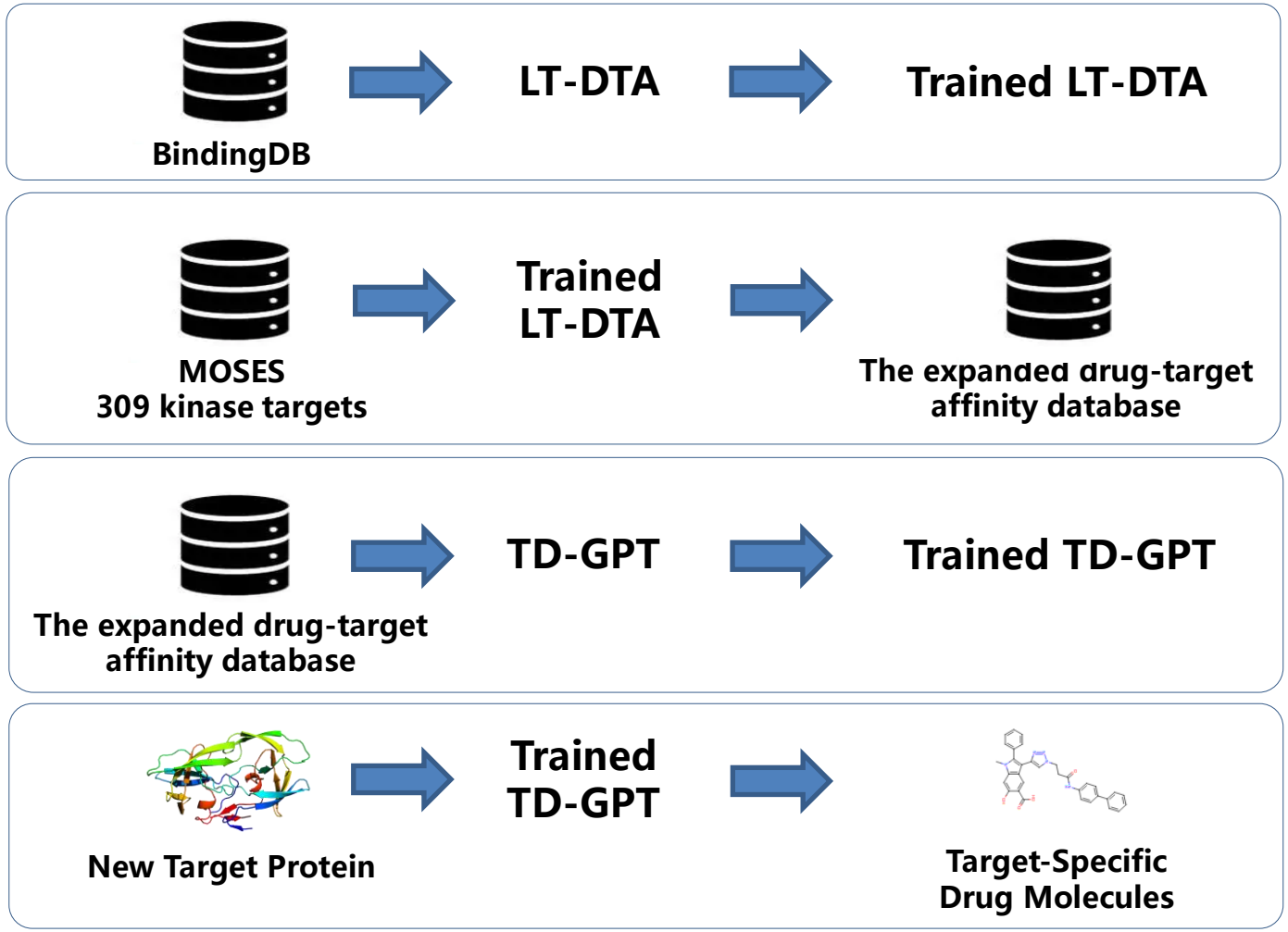
2. TD-GPT Targeted Molecular Generation Model

2 METHOD

• 2.1 Two Models and a Four-stage Workflow

Four-stage workflow

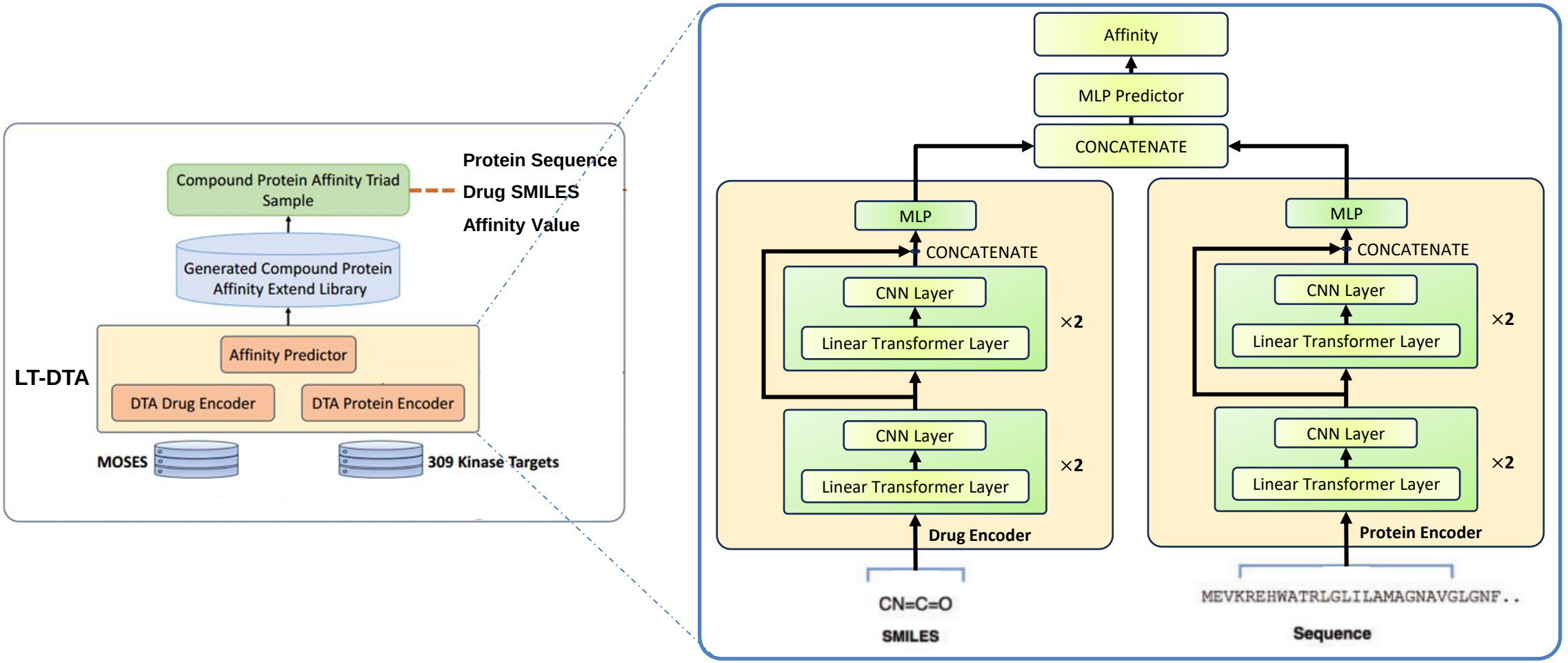
- **Training LT-DTA**
  on BindingBD for affinity prediction;
- **Using LT-DTA**
  to expand the drug-target affinity database for TD-GPT training;
- **Training TD-GPT**
  on the expanded database;
- **Using TD-GPT**
  to generate target-specific molecules.



2 METHOD

- ## 2.2 The LT-DTA Drug-Target Affinity Prediction Model

The Linear Transformer and its Computational Flow Diagram



2 METHOD

• 2.2 The LT-DTA Drug-Target Affinity Prediction Model

The Linear Transformer and its Computational Flow Diagram

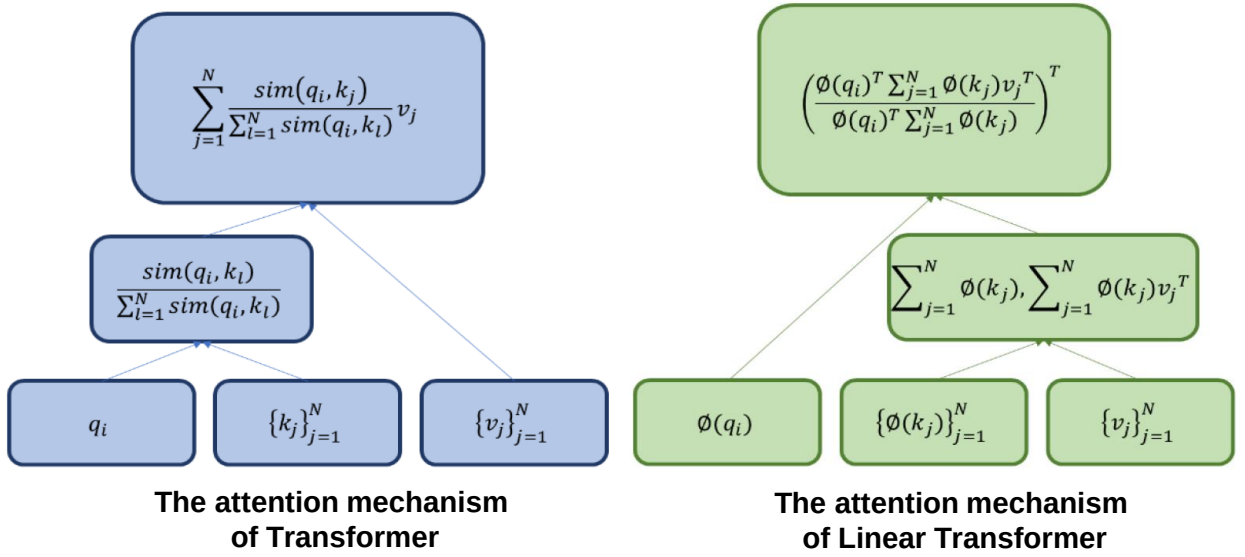
The Vanilla Transformer employs a similarity function for query and key defined as

$$\mathrm{sim}(q_i, k_j) = e^{\frac{q_i^T k_j}{\sqrt{d}}} \qquad V_i' = \frac{\sum_{j=1}^{N} \mathrm{sim}(Q_i, K_j) V_j}{\sum_{j=1}^{N} \mathrm{sim}(Q_i, K_j)}$$

the Linear Transformer uses a kernel function $k(x, y): \mathbb{R}^{2\times F} \to \mathbb{R}_+$ to define similarity

$$\mathrm{sim}(\boldsymbol{q}_i, \boldsymbol{k}_j) = \phi(\boldsymbol{q}_i)^{\top} \varphi(\boldsymbol{k}_j)$$

$$\phi(x) = \varphi(x) = \mathrm{elu}(x) + 1$$



By leveraging the associative property of matrix multiplication

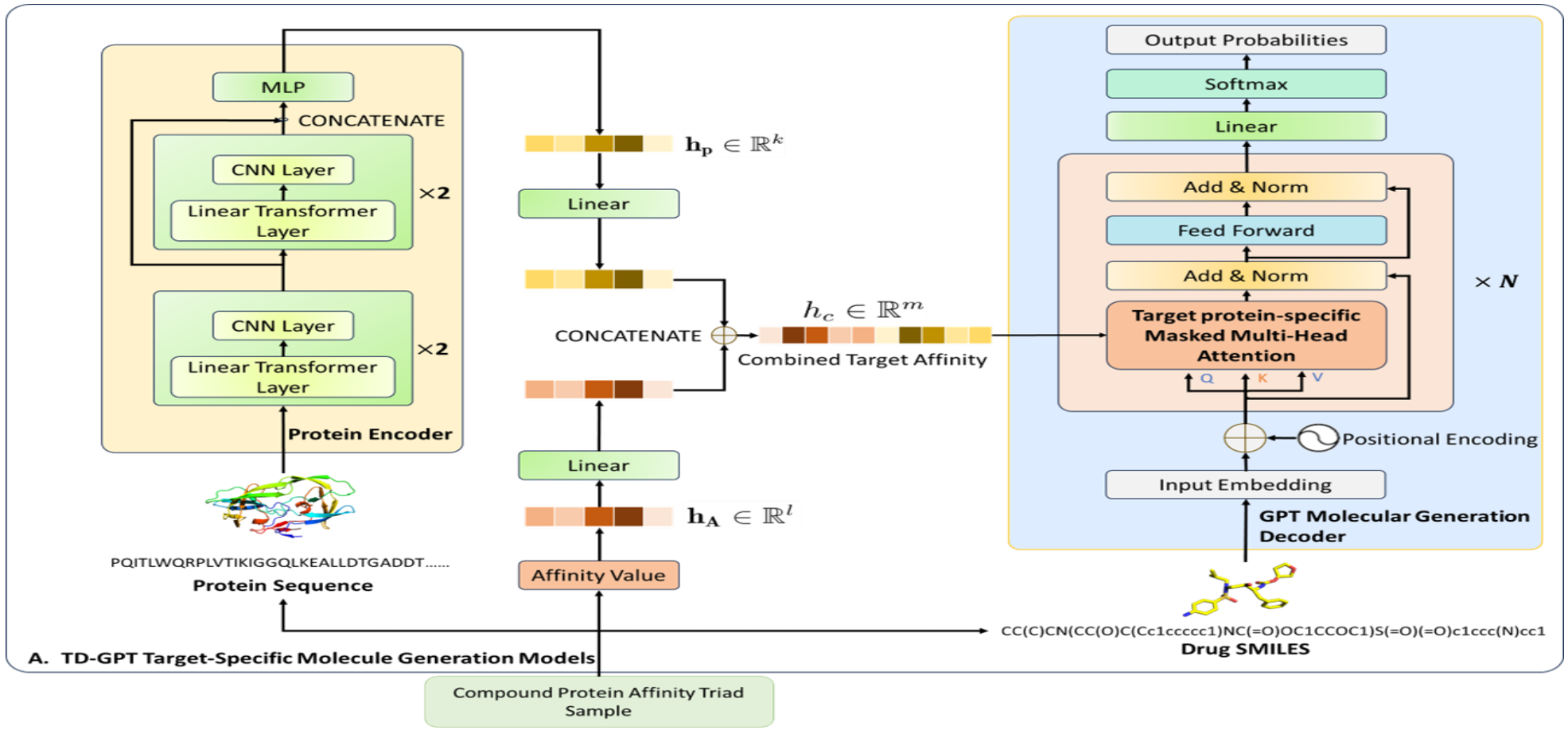
$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})_i = \frac{\sum_{j=1}^{N} \phi(Q_i)^T \phi(K_j) V_j}{\sum_{j=1}^{N} \phi(Q_i)^T \phi(K_j)} = \frac{\phi(Q_i)^T \sum_{j=1}^{N} \phi(K_j) V_j}{\phi(Q_i)^T \sum_{j=1}^{N} \phi(K_j)}$$

The Linear Transformer reduces the computational complexity to $O(N)$.

2 METHOD

- ## 2.3 TD-GPT Targeted Molecular Generation Model

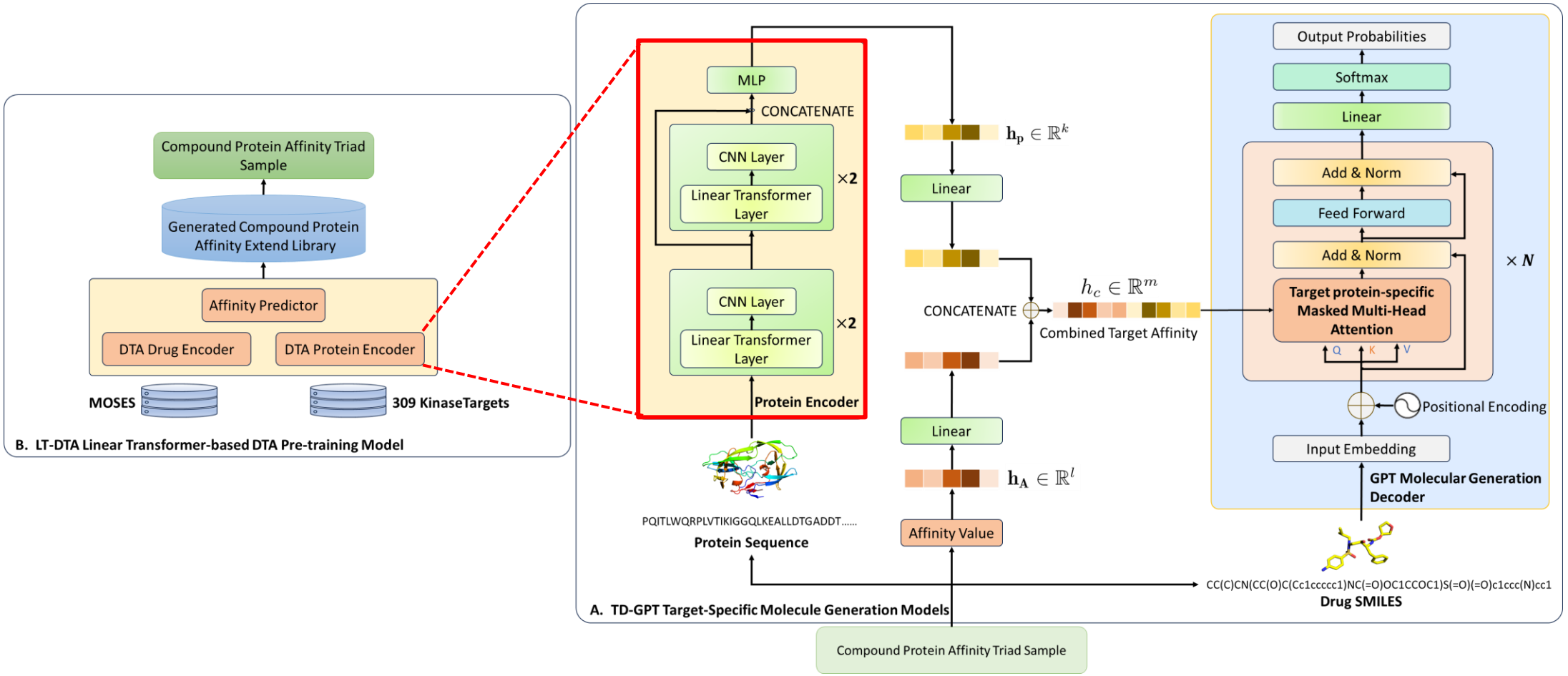
The Molecular GPT Model



2 METHOD

- ## 2.3 TD-GPT Targeted Molecular Generation Model

The Affinity-Enhanced Protein Feature Encoder



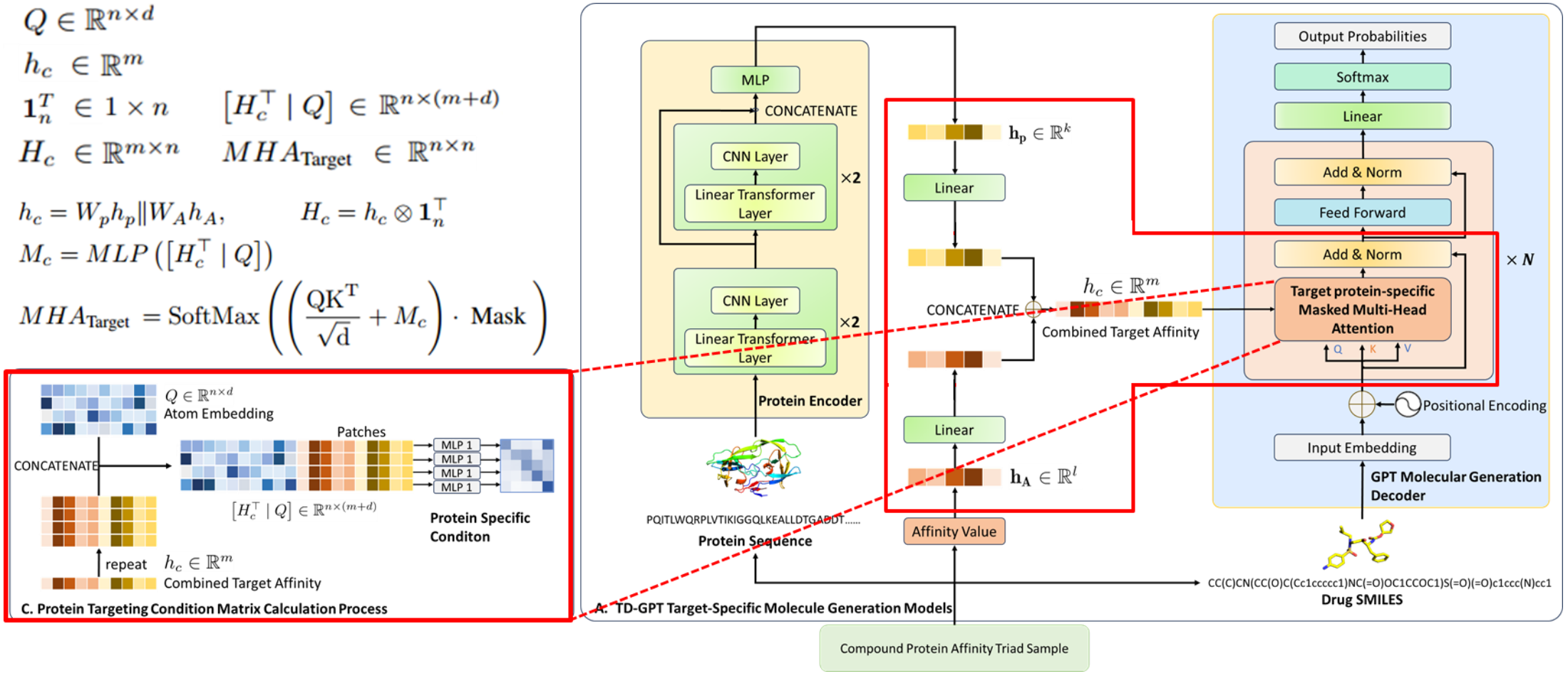
2 METHOD

- ## 2.3 TD-GPT Targeted Molecular Generation Model

Target Protein-Specific Attention Module

$$Q \in \mathbb{R}^{n\times d}$$
$$h_c \in \mathbb{R}^m$$
$$\mathbf{1}_n^T \in 1\times n \qquad [H_c^\top \mid Q] \in \mathbb{R}^{n\times(m+d)}$$
$$H_c \in \mathbb{R}^{m\times n} \qquad MHA_{\text{Target}} \in \mathbb{R}^{n\times n}$$

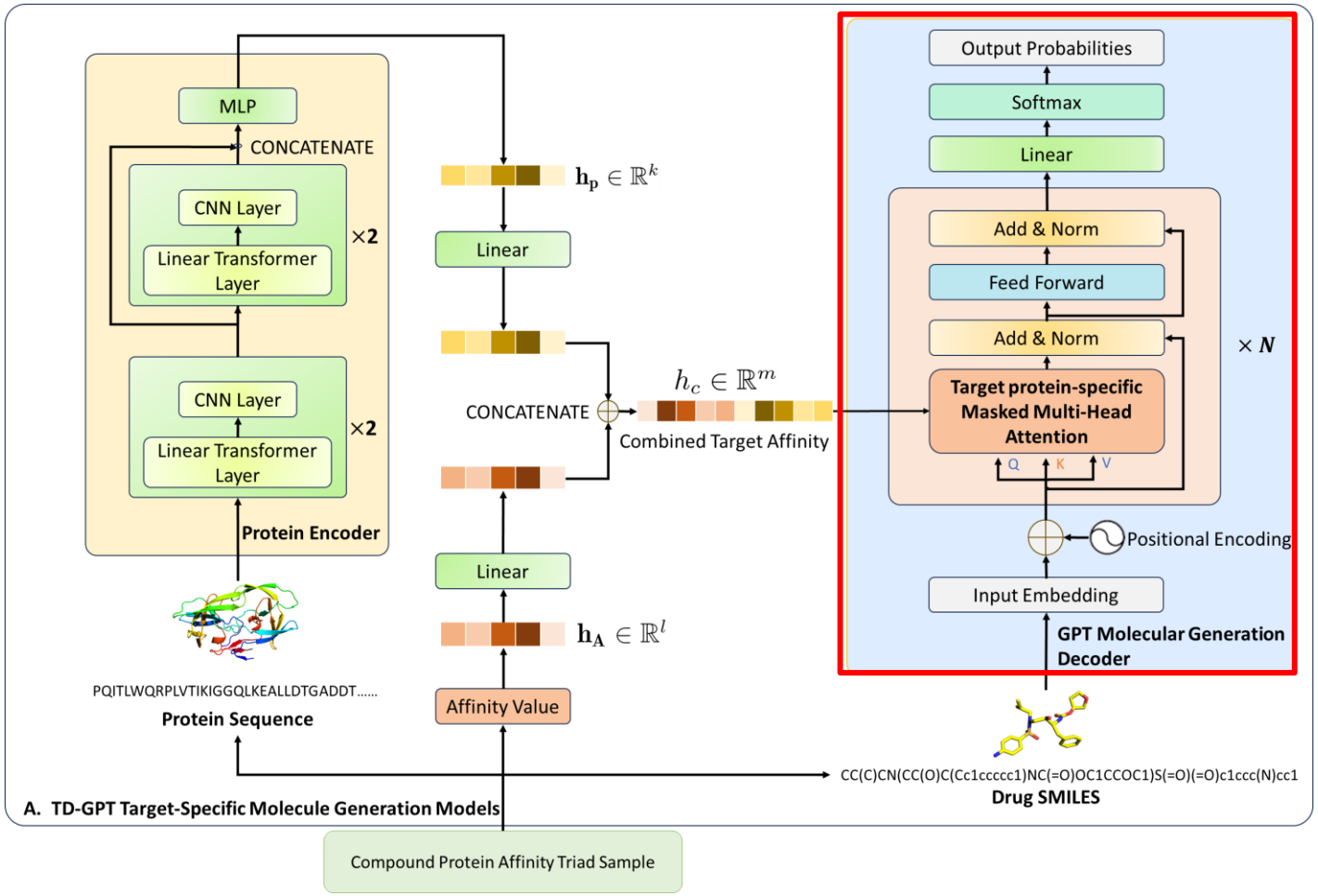
$$h_c = W_p h_p \| W_A h_A, \qquad H_c = h_c \otimes \mathbf{1}_n^\top$$
$$M_c = MLP\left([H_c^\top \mid Q]\right)$$
$$MHA_{\text{Target}} = \text{SoftMax}\left(\left(\frac{\text{QK}^\text{T}}{\sqrt{\text{d}}} + M_c\right)\cdot \text{Mask}\right)$$



2 METHOD

- ## 2.3 TD-GPT Targeted Molecular Generation Model

GPT Molecular Generation Decoder



- **input**

$$s = \{s_1, s_2, \ldots, s_N\}$$

- **GPT Transformer**

$$h^{(0)} = sW_e + W_p$$

$$\bar{h}^{(l)} = \mathrm{LN}\left(h^{(l-1)} + \mathrm{MHA}_{\mathrm{Target}}\left(h^{(l-1)}, h_c\right)\right)$$

$$h^{(l)} = \mathrm{LN}\left(\bar{h}^{(l)} + \mathrm{FFN}\left(\bar{h}^{(l)}\right)\right)$$

- **probability detector**

$$P\left(s_{i+1} \mid s_1, \ldots, s_i\right) = \mathrm{softmax}\left(h_i^{(n)} W_o\right)$$

- **log-likelihood (NLL) loss of the decoded SMILES string**

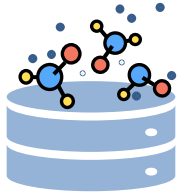
$$NLL(S \mid c) = -\left[\ln P\left(s_1 \mid c\right) + \sum_{i=2}^{N} \ln P\left(s_i \mid s_{1:i-1}, c\right)\right]$$

3 EXPERIMENTS

- **Experimental Design and Baselines**

1. Non-target-specific Generation

by setting the protein target condition vector of TD-GPT to 0, assessing if the model grasped drug molecule properties

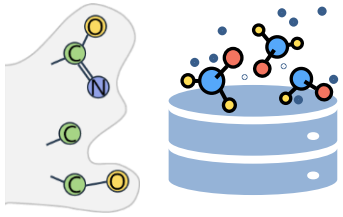


- Baselines:

  NGram, VAE, AAE, LatenGAN

2. Target-specific Generation

checking its ability to produce molecules with target-binding capabilities



- Baselines:

  cRNN, cMolGPT, PETrans

3 EXPERIMENTS

- **Measures**

  - **Measures related to model generation efficiency include:**
    - Validity： Molecules that adhere to basic chemical rules
    - Uniqueness： The proportion of unique molecules
    - Novelty： Molecules not in the training set
    - SNN： Shortest Novelty-Normalized

  - **Measures related to the drug-likeness of generated molecules include:**
    - QED： Quantitative Estimate of Druglikeness
    - SA： Synthetic Accessibility Score
    - Activity

3 EXPERIMENTS

- **Datasets**

  - **BindingDB dataset**

    309 human kinase targets and 95,921 molecules, totaling 182,311 affinity data entries

  - **MOSES dataset**

    consists of 1.9 million lead-like molecules from the ZINC dataset with a molecular weight of 250–350 Da

  - **Extended (drug, target, affinity) triplet database**

    used the LD-DTA model to predict drug-target affinity on the MOSES dataset for the 309 kinase targets, forming an extended (drug, target, affinity) triplet database, which served as the training dataset for the TD-GPT targeted molecular generation model.

4 RESULTS

- **Non-target-specific Molecular Generation**

Table 1. Comparison of TD-GPT with Non-target-specific Generation Models. Bold font indicates the best.

Models	Valid	Unique @ 1k	Unique @ 10k	Novelty	SNN	Measures Product
HMM	0.076	0.623	0.567	-	0.388	-
NGram	0.238	0.974	0.922	-	0.521	-
VAE	0.977	1	**0.998**	0.695	0.608	0.412
AAE	0.937	1	0.997	0.695	**0.626**	0.406
LatentGAN	0.897	1	0.997	**0.949**	0.538	0.457
TD-GPT (Ours)	**0.993**	**1**	0.994	0.781	0.619	**0.477**

4 RESULTS

- **Target-specific Molecular Generation**

Table 2. Comparison of TD-GPT with Other Models for Target-specific Generation. Bold font indicates the best.

Target	Model	Valid	Unique @10k	Novelty	QED	SA
EGFR	cRNN	0.921	0.861	0.662	-	-
	cMolGPT	0.885	0.940	0.898	-	-
	PETrans	0.895	0.719	**1**	0.452	2.736
	TD-GPT (Ours)	**0.934**	**0.978**	0.962	**0.742**	**2.672**
HTR1A	cRNN	0.922	0.844	0.498	-	-
	cMolGPT	0.905	0.896	0.787	-	-
	PETrans	0.905	0.624	**1**	0.529	2.971
	TD-GPT (Ours)	**0.952**	**0.979**	0.926	**0.755**	**2.657**
S1PR1	cRNN	0.926	0.861	0.514	-	-
	cMolGPT	0.926	0.838	0.684	-	-
	PETrans	0.815	0.420	**1**	0.459	**2.559**
	TD-GPT (Ours)	**0.995**	**0.980**	0.931	**0.751**	2.666

4 RESULTS

- **Target-specificity Experiment of TD-GPT**

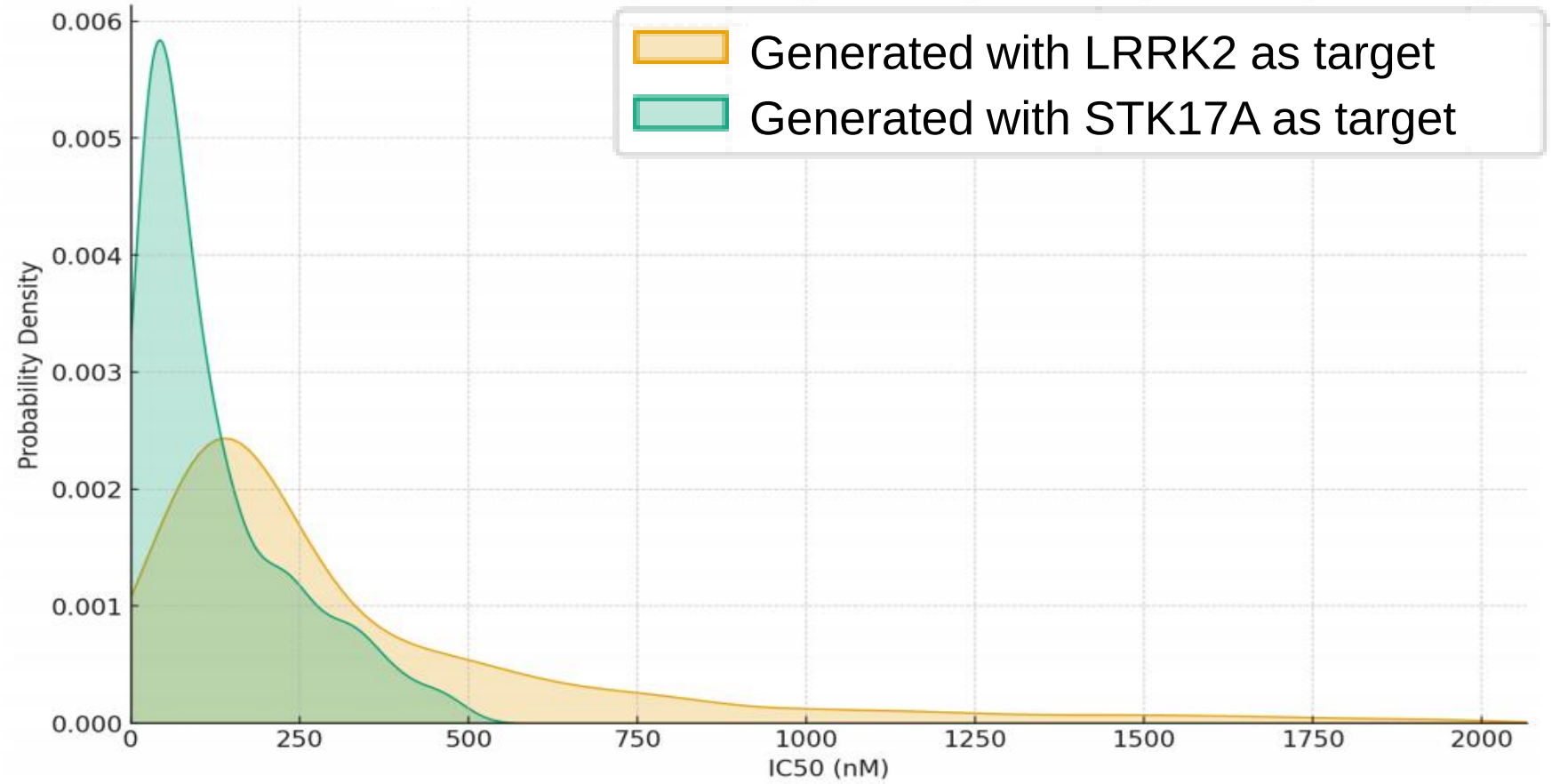


Fig. 2. Activity Distribution for Target STK17A / LRRK2 of Molecules Generated with Different Targets

4 RESULTS

- **Affinity(Activity) Controllability of TD-GPT**

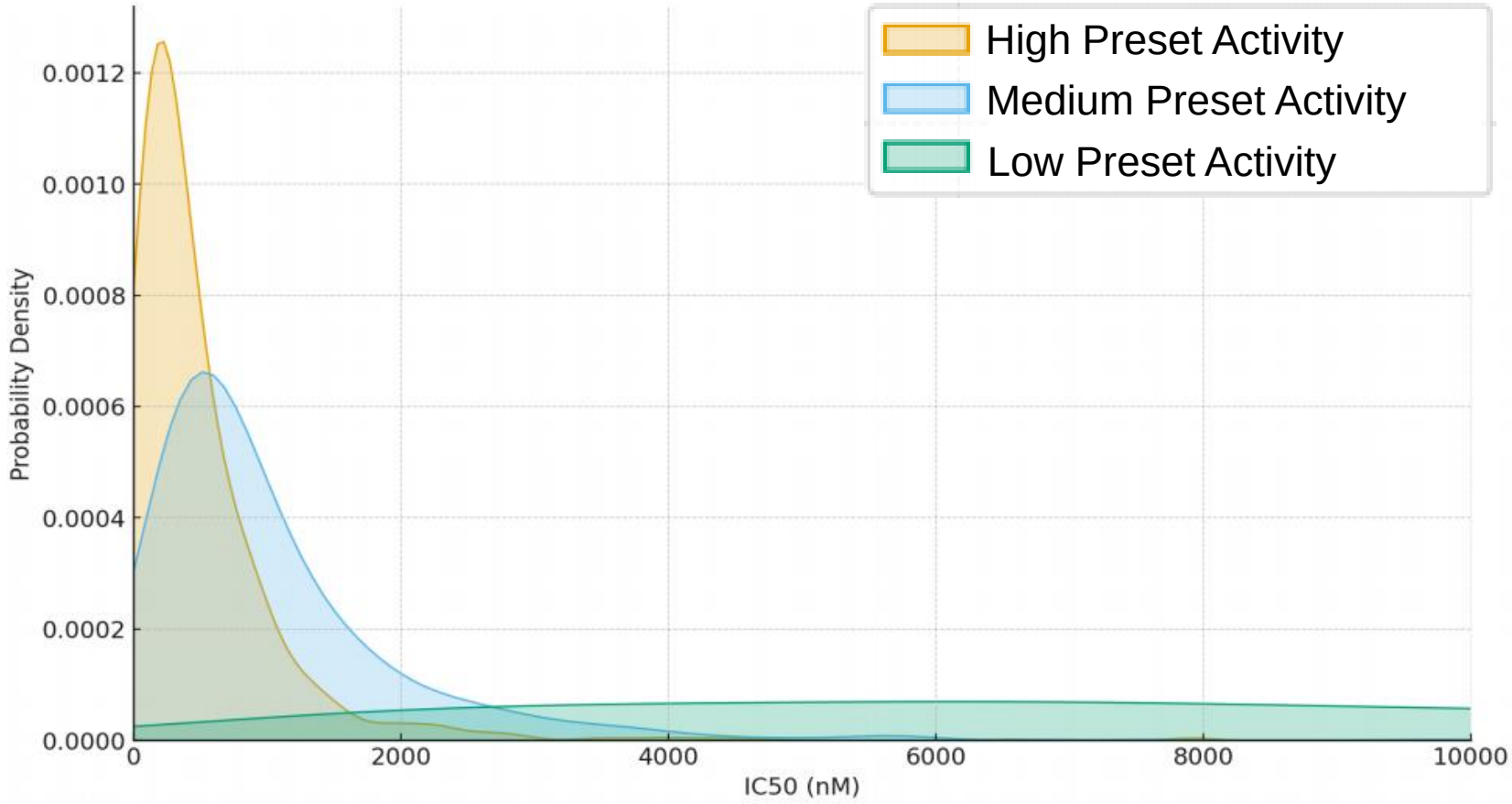


Fig. 3. Activity Distribution of Molecules Generated for LRRK2 under Different Preset Affinity Conditions.

4 RESULTS

- **Example of Generating Target-Specific Candidate Drug Molecules**

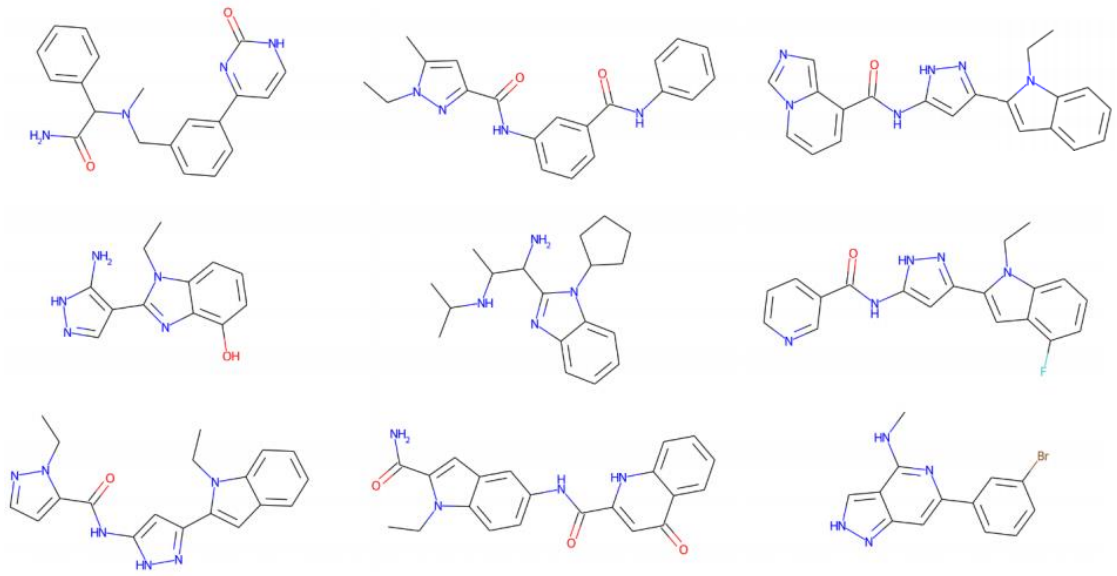


Fig. 4. High-affinity (active) candidate drug molecules generated using the TD-GPT Framework.

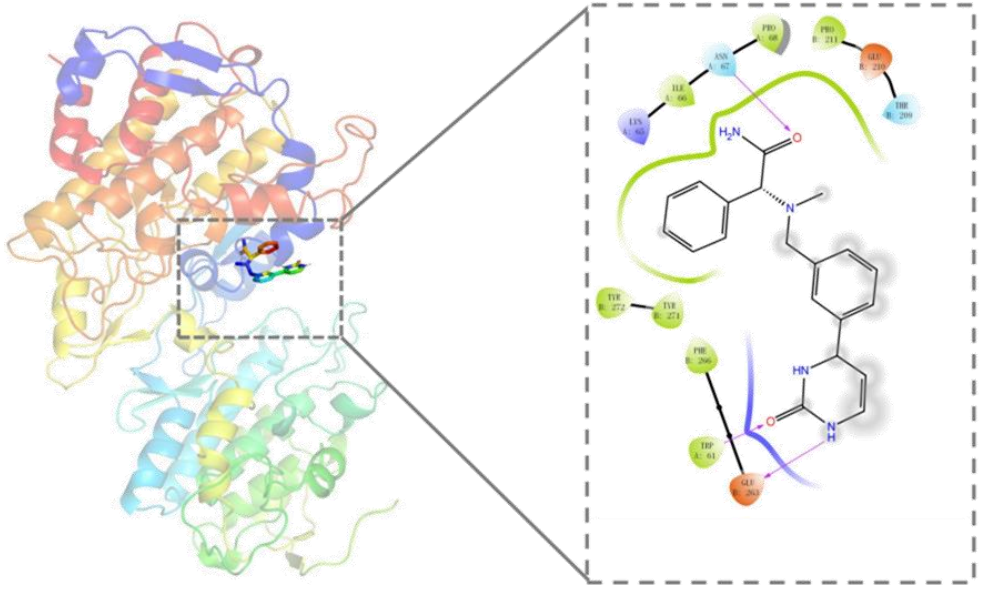


Fig. 5. Representative candidate drug molecule docking results with the PDZ-binding kinase target.

5 SUMMARY

1. TD-GPT, a novel deep learning framework for targeted drug molecule generation, integrates LT-DTA for affinity prediction and database expansion,

2. and a target-specific attention module for optimized molecule generation,

3. demonstrating superior performance in generating high-affinity, target-specific molecules.

THANKS!