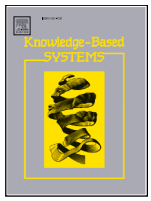
Knowledge-Based Systems 340 (2026) 115605

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



MT-DiffGen: Unifying affinity prediction and target-aware molecule generation with a multi-task diffusion model

Shiping Li [a], Hong Wang [a,*], Tao Hu [a], Luhe Zhuang [a], Jun Zhao [a], Yuhuang Sheng [b], Yanshen Sun [c,*]

[a] School of Information Science and Engineering, Shandong Normal University, Jinan, 250358, Shandong Province, China
[b] School of Chemistry, Chemical Engineering and Materials Science, Shandong Normal University, Jinan, 250358, Shandong Province, China
[c] Department of Computer Science, Virginia Tech, Blacksburg, VA, 24061, USA

ARTICLE INFO

Keywords:
Drug-target affinity prediction
Molecular generation
Diffusion models
Multi-task learning
Deep learning
Drug discovery

ABSTRACT

Current computational drug discovery faces a fundamental challenge: predictive models for binding affinity cannot design new molecules, while generative models create compounds without leveraging quantitative affinity signals. This disconnect forces expensive iterative cycles between computational design and experimental validation. To address this limitation, we present MT-DiffGen, a unified framework that simultaneously predicts drug-target affinity and generates target-aware molecules. Our approach employs a dual-branch architecture in which molecular graphs are processed by complementary graph neural networks, while protein sequences are encoded using a hierarchical deep learning architecture. The key innovation is a conditional diffusion mechanism that uses affinity-prediction gradients to directly guide molecular generation, enabling real-time optimization of binding potency while ensuring chemical validity. A comprehensive evaluation on the Davis and KIBA benchmarks demonstrates that MT-DiffGen achieves competitive prediction accuracy while generating molecules with 99.9% chemical validity, 18% higher docking scores, and 24% better synthetic accessibility than target-agnostic methods. The generated molecules maintain structural diversity and physicochemical properties consistent with known bioactive compounds. By integrating prediction and generation into a single end-to-end pipeline, our work provides an efficient platform for accelerated drug discovery that bridges the gap between computational prediction and molecular design.

1. Introduction

The average cost of bringing a single drug to market has surpassed $2.5 billion and 10-12 years of development time, with attrition rates above 90% [1], underscoring the urgent need for computational methods that can accelerate early-stage development. A central challenge lies in identifying molecules that simultaneously exhibit high binding affinity against a specific protein target and acceptable pharmacokinetic properties. Although traditional high-throughput screening and structure-based design have historically driven lead identification, their scalability is limited by prohibitive costs and low hit rates.

In response, deep learning has emerged as a transformative tool for automating molecular design and evaluation. Two dominant paradigms have evolved in parallel: (1) *predictive models* such as DeepDTA [2] and GraphDTA [3] that estimates drug-target binding affinity (DTA) from sequence and graph data,and (2) *generative models* including JT-VAE [4], MolGPT [5], and diffusion-based approaches [6,7] that produce novel molecular structures. Although both paradigms have advanced significantly, they remain fundamentally decoupled: predictive models operate as passive scorers without the ability to propose new chemical entities, while generative models typically optimize sequence likelihood or structural plausibility without leveraging real-valued affinity signals during training. This artificial separation forces practitioners into costly iterative cycles of "generate →screen →optimize," where each stage operates in a disjoint latent space, leading to compounding errors, distribution shift, and inefficient exploration of chemical space.

We argue that this division is both unnecessary and suboptimal. From a biophysical perspective, a molecule's binding affinity and its chemical structure are two facets of the same underlying interaction: the former quantifies "how strongly" a ligand binds, while the latter defines "what" the ligand is. A unified model that jointly reasons about affinity and structure should, in principle, yield more coherent and potent designs. Recent attempts to bridge this gap, such as reinforcement learning with frozen predictors [8] or multitask RNNs [9], only achieve shallow

* Corresponding authors.
E-mail addresses: 111052@sdnu.edu.cn (H. Wang), yansh93@vt.edu (Y. Sun).

https://doi.org/10.1016/j.knosys.2026.115605
Received 18 November 2025; Received in revised form 14 January 2026; Accepted 22 February 2026
Available online 2 March 2026

coupling, often through alternating optimization or post-hoc scoring, rather than enabling gradient-level co-supervision between regression and generation.

To overcome these limitations, we introduce **MT-DiffGen**, a multi-task diffusion framework that unifies binding affinity prediction with target-aware molecular generation within a single end-to-end architecture. The core of our approach is a shared multimodal encoder that learns unified representations serving both tasks simultaneously. We design a model with a dual-branch graph neural network (GAT & GCN) for ligand encoding and a hierarchical protein encoder (multi-scale CNN, BiLSTM, Transformer) for comprehensive feature extraction. The fused representations are then fed into dual heads for pIC_{50} affinity regression and conditional molecular generation.

Crucially, our framework enables deep integration between predictive and generative tasks through several technical innovations. During training, the latent space is regularized by both noise-prediction loss and Kullback-Leibler (KL) divergence, while the regression loss from the affinity predictor serves as an explicit guidance signal in the reverse diffusion process. This allows every denoising step to be directly influenced by predicted binding affinity, enabling truly affinity-aware generation. For the decoding phase, we initialize from a pre-trained MolT5 checkpoint fine-tuned on ChEMBL, BindingDB, and PDBbind, enabling direct generation of valid, synthesis-ready SMILES strings without post-hoc graph-to-string conversion.

By collapsing the traditional "predict →synthesize →test" cycle into a single differentiable pipeline, MT-DiffGen represents a significant advancement over existing fragment-based approaches. Extensive evaluations demonstrate that our unified framework not only matches or surpasses state-of-the-art affinity prediction models on the Davis and KIBA benchmarks but also generates molecules with significantly improved binding characteristics, achieving 18% higher docking scores and 24% lower synthetic accessibility scores compared to ligands produced by target-agnostic generative models, while maintaining diversity and QED (Quantitative Estimate of Drug-likeness) distributions statistically indistinguishable from known active compounds. Our contributions are summarized as follows.

- **A Unified Multi-Task Diffusion Framework**: MT-DiffGen, the proposed framework, unifies drug-target affinity prediction and target-conditioned molecule generation within a shared latent space. This integration enables gradient-level co-supervision between predictive and generative tasks through a single differentiable pipeline.
- **Dual-Branch Multimodal Encoder**: The novel feature extraction module synergistically combines graph attention networks and graph convolutional networks for comprehensive ligand representation, with a multi-scale CNN-BiLSTM-Transformer architecture for protein sequence modeling, learning unified embeddings optimized for both regression and generation.
- **Affinity-Guided Conditional Diffusion**: The introduced training paradigm allows the affinity regression loss to directly guide the reverse diffusion process, enabling real-time steering of molecular generation toward high-affinity chemical space without requiring post-hoc docking or scoring.
- **Direct SMILES Decoding with Pre-Trained MolT5**: The integration of a fine-tuned MolT5 decoder enables direct mapping from latent representations to syntactically valid and synthesis-ready SMILES strings, achieving 99.9% chemical validity and eliminating the need for graph-to-string conversion or structural post-processing.
- **Comprehensive Experimental Validation**: Extensive experiments demonstrate state-of-the-art performance on Davis and KIBA benchmarks for affinity prediction, while generated molecules exhibit superior docking scores (18% higher), synthetic accessibility (24% lower), and physicochemical fidelity compared to target-agnostic baselines, validated through rigorous clustering and distribution analysis.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 elaborates on the proposed MT-DiffGen. Section 4 describes datasets and experimental results. Section 5 provides in-depth discussion, and Section 6 concludes the paper.

2. Related work

Drug-target affinity (DTA) prediction and *de-novo* ligand generation have traditionally been treated as two independent tracks. We review the two lines separately and then discuss recent attempts to chain them into an outer loop.

2.1. Binding-affinity prediction

Early physics-based methods (AutoDock Vina [10], Glide [11], etc.) rely on protein-ligand docking and require experimentally resolved 3-D structures. While accurate for individual complexes, they scale poorly to large chemical libraries and are sensitive to input conformations. With the explosion of public bio-assays, data-driven models replaced energy functions with regression heads. DeepDTA [2] and WideDTA [12] were the first to apply 1-D CNNs on raw SMILES and protein sequences, but they ignore molecular topology. GraphDTA [3] and its follow-ups [13,14] remedy this by encoding ligands as graphs, while protein graphs [15] or 3-D grids [16] are used when structures are available. Very recent works inject protein-language pre-training (ESM-2) or hierarchical attention to push Davis/KIBA RMSE below 0.22 [17,18]. Despite steady gains, all of the above models are trained for a single regression objective and stop at scoring given compounds–none can invent new chemistry.

2.2. De-novo molecule generation

Molecular generation has followed a parallel arc. Early SMILES-based RNNs/VAEs achieve 70% validity but suffer from mode collapse [19,20]. Fragment-assembly engines (DeepFM [21], DeLinker [22]) improve synthetic accessibility but explore a fragment-biased subspace. The transformer era brought MolGPT [5], ChemBERTa-2 [23], and MoLFormer [24], whose billion-SMILES pre-training lifts validity above 95%; scaffold-constrained variants (t-SMILES, SMILES-X) guarantee 100% grammatical correctness. Very recent autoregressive models such as LlaMoL [25] and Chi-Former [26] incorporate protein-language embeddings (ESM-2) to bias generation toward a target, but they still optimise sequence likelihood rather than quantitative binding strength. Diffusion models have rapidly become the dominant paradigm. 2-D approaches such as Torsional-Diffusion [6] and MoleculeDiff [7] generate graphs with state-of-the-art diversity; 3-D pocket-conditioned extensions, including TargetDiff [27], DiffSBDD [28], PILOT [29], and DualDiff [30], produce atom clouds that achieve < 1 Å RMSD on cross-docked complexes. Nevertheless, none of these diffusion works exploit continuous affinity labels during training–docking or a frozen DTA network is required after sampling, leading to expensive rejection sampling and distribution shift. Target-aware language models continue to maximize likelihood. LigandGPT [31] and Pocket2Mol [32] condition on protein sequences or graphs, yet their training objective remains next-token prediction; binding strength is only assessed post-hoc.

2.3. Predictor-generator pipeline

To bridge the gap, a growing body of work wraps a pre-trained DTA model around a generator in an outer loop: REINVENT-2 [33], RationaleRL [34], and MolGen [35] use the predictor as a reward and update the generator via policy gradients or genetic operators. These systems achieve impressive hit rates on Dopamine Receptor D2 (DRD2), Janus Kinase 2 (JAK2), etc.; however, the predictor is frozen, and the generator is unaware of affinity signals during its own training. Consequently, the

two models live in disjoint parameter spaces and suffer from compounding errors, high sample complexity, and reward hacking. Only a handful of studies explore joint training. DeepDTAGen [36] couples a CNN-based regressor with an RNN generator via multitask losses, but generation is still SMILES-autoregressive and affinity guidance is indirect. Diffusion-based methods such as DiffAffinity [37] and AffinDiff [38] inject regression heads into 3-D diffusion models, yet they produce heavy atom clouds and require post-hoc SMILES conversion.

MT-DiffGen departs from this fragmented landscape by collapsing prediction and generation into a single shared latent space, enabling gradient-level co-training of affinity regression and target-conditioned SMILES generation within one end-to-end diffusion framework. Predictor gradients flow directly into the denoising network, while a MolT5 decoder, fine-tuned on latent features, guarantees chemical validity, compressing the customary design-evaluate-optimize cycle into a single, unified training objective.

3. Method

3.1. Problem statement

Let $\mathcal{P}$ denote the finite set of target proteins and $\mathcal{C}$ the countable set of syntactically valid SMILES strings. For each protein $p \in \mathcal{P}$ we observe an amino-acid sequence $\mathbf{s}_p \in \Sigma^{\ell_p}$ of length ℓ_p. A ligand candidate is represented by a SMILES string $\mathbf{c} \in \mathcal{C}$ and its molecular graph

$$G_{\mathbf{c}} = (V, E, \mathbf{X}, \mathbf{A}),$$

where V is the atom set, E the bond set, $\mathbf{X} \in \mathbb{R}^{|V| \times d_{\text{atom}}}$ the atom-feature matrix and $\mathbf{A} \in \{0,1\}^{|V| \times |V|}$ the adjacency matrix.

The quantitative binding signal is the negative log-transformed half-maximal inhibitory concentration.

$$y(\mathbf{c}, p) = -\log_{10}\big(\text{IC}_{50}(\mathbf{c}, p)/10^9 \text{ nM}\big) \in \mathbb{R},$$

commonly denoted pIC_{50}. Under the thermodynamic view, y is an emergent property of the joint distribution.

$$\pi(\mathbf{c}, y \mid p) = \frac{1}{Z(p)} \exp\big(-\mathcal{E}(\mathbf{c}, p; y)\big),$$

where $\mathcal{E}$ is an unknown free-energy functional and $Z(p)$ the partition function.

Problem 1 (Joint Affinity-Generation). Given a target protein $p \in \mathcal{P}$ (sequence $\mathbf{s}_p$) and a dataset

$$\mathcal{D} = \{(\mathbf{c}_i, p_i, y_i)\}_{i=1}^{N} \sim \pi,$$

learn a parametric model

$$f_\theta,\ g_\phi \ :\quad f_\theta(\mathbf{c}, p) \approx y, \qquad g_\phi(p, \varepsilon) \sim \pi(\cdot \mid p),$$

where f_θ regresses pIC_{50} and g_ϕ generates novel SMILES strings that are both chemically valid and possess high affinity under $\pi(\cdot \mid p)$.

Existing works optimise f and g separately, leading to distribution shift and expensive rejection loops. We aim to maximise the joint objective.

$$\mathcal{J}(\theta, \phi) = \mathbb{E}_{p \sim \mathcal{P}}\Big[\underbrace{-\mathcal{L}_{\text{MSE}}(f_\theta; p)}_{\text{affinity fidelity}} + \underbrace{\mathbb{E}_{\mathbf{c} \sim g_\phi(p)}\big[y(\mathbf{c}, p) - \lambda S(\mathbf{c})\big]}_{\text{affinity-guided generation}}\Big],$$

where $\mathcal{L}_{\text{MSE}}(f_\theta; p) = \mathbb{E}_{(\mathbf{c},y) \sim \mathcal{D}_p}\big[f_\theta(\mathbf{c}, p) - y\big]^2$ is the mean-squared-error loss on protein p, $g_\phi(p)$ is the generative distribution over $\mathcal{C}$ conditioned on p, λ balances potency versus synthetic accessibility.

The generation head must receive continuous affinity feedback during sampling; yet, SMILES are discrete. We therefore embed both tasks in a shared latent diffusion space and let regression gradients flow into the denoising chain, achieving gradient-level co-supervision between potency and chemical realism for the first time.

3.2. MT-DiffGen framework

MT-DiffGen unifies drug-target affinity regression and target-conditioned SMILES generation into a single, end-to-end trainable diffusion pipeline. As illustrated in Fig. 1, the system is organized into five tightly-coupled stages: **(A) Input Module.** One-dimensional strings (SMILES for ligands, amino-acid sequences for proteins) are fed into the model without relying on experimentally resolved 3-D structures. **(B) Embedding Module.** RDKit converts each SMILES into a molecular graph $G = (V, E)$. Protein sequences are tokenised and embedded by a learnable look-up table. **(C) Feature Extraction Module.** A dual-branch encoder fuses modality-specific signals. (i) Ligand branch: parallel stacks of GAT and GCN networks capture local and global molecular topology; (ii) Protein branch: a multi-scale CNN-BiLSTM-Transformer stack models short motifs, long-range dependencies, and global residue–residue relationships. The two branches yield fixed-dimensional vectors $\mathbf{h}_{\text{mol}}$ and $\mathbf{h}_{\text{prot}}$. **(D) Affinity Prediction Module.** A multi-layer perceptron regresses pIC_{50} from the concatenated representation $\hat{y} = f_\theta(\mathbf{h}_{\text{mol}} \oplus \mathbf{h}_{\text{prot}})$, trained with mean-squared-error loss $\mathcal{L}_{\text{aff}}$. **(E) Diffusion Generation Module.** A conditional diffusion decoder treats $\mathbf{h}_{\text{prot}}$ as a continuous prior. During training, noise-prediction and KL-divergence losses regularize the latent space; at inference, reverse steps reconstruct a latent drug vector $\mathbf{Z}_{\text{gen}}$ under protein guidance. A pre-fine-tuned MolT5 decoder finally maps $\mathbf{Z}_{\text{gen}}$ to a valid, synthesisable SMILES string without post-hoc graph-to-string conversion.

3.3. Input module

MT-DiffGen takes as input two types of one-dimensional sequences: the SMILES string of a ligand and the amino acid sequence of a protein, with no reliance on 3D structural information during training. Each training example is a triplet.

$$\tau = (\mathbf{c}, \mathbf{s}, \mathbf{y})$$

where $\mathbf{c}$ denotes the drug SMILES sequence and $\mathbf{s}$ is the protein sequence , $\mathbf{y}$ denotes the binding affinity value. During preprocessing, SMILES strings are standardized using RDKit to remove stereochemical ambiguity and ensure consistent representation, while protein sequences are only truncated in length and tokenized. The embedded representations of both modalities are fed directly into subsequent network modules, without the need for 3D structural alignment or binding pocket identification, enabling generalization to proteins of unknown structure.

3.4. Embedding module

Since discrete SMILES tokens and amino-acid symbols cannot be directly processed by continuous deep networks, the embedding module projects each modality into a shared latent space while preserving chemically and biologically meaningful relationships. Rather than treating ligands and proteins as generic sequences, we design two dedicated encoding pathways that jointly optimize representations for both affinity regression and diffusion generation, producing task-specific embeddings that are more expressive than frozen, off-the-shelf language models.

3.4.1. From SMILES to latent ligand vector

Step 1. Each canonical SMILES string $\mathbf{c} = [c_1, \ldots, c_{L_{\text{mol}}}]$ is parsed by RDKit into an attributed molecular graph:

$$\mathcal{G}_{\text{mol}} = (V, E, \mathbf{X}_{\text{atom}}, \mathbf{A}) \tag{1}$$

where $V = \{v_i\}_{i=1}^{N_{\text{atoms}}}$ is the atom set, $E \subseteq V \times V$ the bond set, $\mathbf{X}_{\text{atom}} \in \mathbb{R}^{|V| \times d_{\text{atom}}}$ the 37-dimensional atom-feature matrix (element, degree, hybridisation, aromaticity, H-count, etc.), $\mathbf{A} \in \{0,1\}^{|V| \times |V|}$ the adjacency matrix with $\mathbf{A}_{ij} = 1$ iff $(v_i, v_j) \in E$. Graph construction is deterministic and differentiable, so gradients from both the affinity head and the diffusion decoder can flow back to atom-type choices during early fine-tuning.

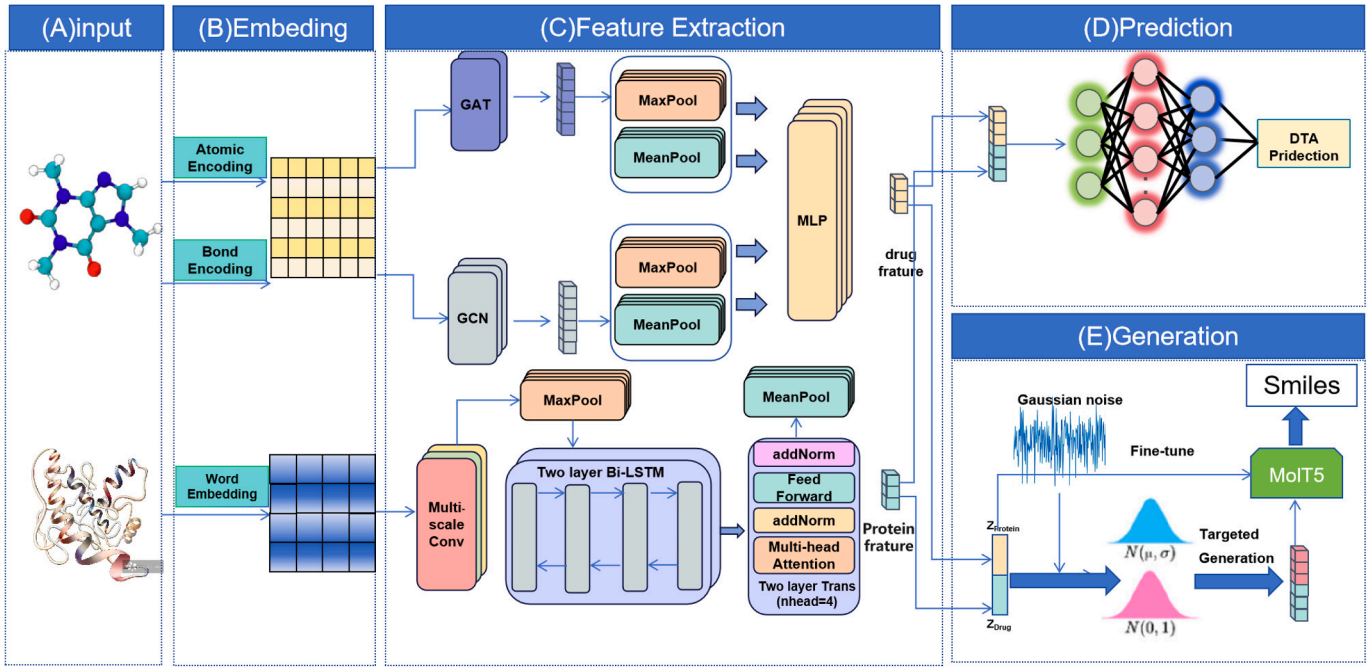


Fig. 1. Framework of the MT-DiffGen. (A) Input Module feeds 1-D SMILES and proteins; **(B) Embedding Module** converts them to graphs and vectors; **(C) Feature Extraction Module** fuses ligand (GAT/GCN) and protein (multi-scale CNN-BiLSTM-Transformer) into unified vectors; **(D) Affinity Prediction Module** regresses pIC_{50}; **(E) Diffusion Generation Module** denoises under protein guidance and emits valid SMILES via MolT5, five modules, one shared latent space.

Step 2. Node-level embedding. We learn a trainable projection:

$$\mathbf{h}_i^{(0)} = \mathbf{W}_{\text{node}}\mathbf{x}_i^{\text{atom}} + \mathbf{b}_{\text{node}} \in \mathbb{R}^d, \qquad \mathbf{W}_{\text{node}} \in \mathbb{R}^{d\times d_{\text{atom}}} \tag{2}$$

followed by Layer-Norm and ReLU. Bond features $\mathbf{e}_{ij}$ (type, conjugation, stereo) are similarly embedded to $\mathbf{e}_{ij}^{(0)} \in \mathbb{R}^d$ and reused in attention-based message passing.

Step 3. Topological positional encoding. To inject topological context, we compute the normalised graph Laplacian

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}, \qquad \mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij} \tag{3}$$

and use the $k = 8$ smallest eigenvectors as structural positional codes:

$$\mathbf{P}_{\text{lap}} \in \mathbb{R}^{|V|\times k}, \qquad \mathbf{h}_i^{(0)} \leftarrow \mathbf{h}_i^{(0)} + \mathbf{W}_{\text{lap}}\mathbf{p}_i^{\text{lap}} \tag{4}$$

where, $\mathbf{I} \in \mathbb{R}^{|V|\times|V|}$ is the identity matrix; $\mathbf{A} \in \{0,1\}^{|V|\times|V|}$ is the adjacency matrix of the molecular graph; $\mathbf{D} \in \mathbb{R}^{|V|\times|V|}$ is the diagonal degree matrix with entries $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$; the matrix $\mathbf{L}$ is the symmetrically normalised Laplacian that encodes graph topology. The matrix $\mathbf{P}_{\text{lap}} \in \mathbb{R}^{|V|\times k}$ stacks the $k = 8$ eigenvectors corresponding to the smallest eigenvalues of $\mathbf{L}$, serving as structural positional codes; $\mathbf{p}_i^{\text{lap}} \in \mathbb{R}^k$ is the ith row of $\mathbf{P}_{\text{lap}}$, and $\mathbf{W}_{\text{lap}} \in \mathbb{R}^{d\times k}$ is a learnable projection that injects topological context into the initial node embedding $\mathbf{h}_i^{(0)} \in \mathbb{R}^d$. This term allows the downstream attention heads to distinguish symmetric sub-structures that pure degree features cannot resolve.

Step 4. Dual-branch graph encoding.

(i) GAT branch: multi-head attention updates node v_i via

$$\alpha_{ij}^{(l,h)} = \frac{\exp\left(\text{LeakyReLU}\left[\mathbf{a}^\top[\mathbf{W}^{(h)}\mathbf{h}_i^{(l)} \| \mathbf{W}^{(h)}\mathbf{h}_j^{(l)}]\right)\right)}{\sum_{k\in\mathcal{N}(i)} \exp(\text{ idem })} \tag{5}$$

$$\mathbf{h}_i^{\text{GAT},(l+1)} = \Big\|_{h=1}^{H} \sigma\Big(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l,h)}\mathbf{W}^{(h)}\mathbf{h}_j^{(l)}\Big) \tag{6}$$

where $\mathbf{h}_i^{(l)} \in \mathbb{R}^d$ is the feature vector of node i at layer l; $\mathbf{W}^{(h)} \in \mathbb{R}^{d'\times d}$ is the learnable weight matrix of head h (here $d' = d/H$); $\mathbf{a} \in \mathbb{R}^{2d'}$ is the attention kernel; $\|$ denotes vector concatenation; $\mathcal{N}(i)$ is the 1-hop neighbour set of node i; $\alpha_{ij}^{(l,h)}$ is the attention coefficient between nodes i and j in head h; σ is the activation function (ELU); and the outer concatenation $\Big\|_{h=1}^{H}$ aggregates $H = 8$ heads back into a d-dimensional vector.

(ii) GCN branch: mean-pooling aggregation with residual connection

$$\mathbf{h}_i^{\text{GCN},(l+1)} = \sigma\Big(\mathbf{W}^{(l)} \cdot \text{mean}_{j\in\mathcal{N}(i)\cup\{i\}}\mathbf{h}_j^{(l)}\Big) + \mathbf{h}_i^{(l)} \tag{7}$$

where $\mathbf{W}^{(l)} \in \mathbb{R}^{d\times d}$ is the layer-specific weight matrix; $\text{mean}_{j\in\mathcal{N}(i)\cup\{i\}}$ performs average aggregation over node i and its neighbours; and the residual connection $+\mathbf{h}_i^{(l)}$ stabilises gradient flow through deep graph stacks. After $L = 4$ layers we obtain $\mathbf{H}_{\text{GAT}} \in \mathbb{R}^{|V|\times d}$ and $\mathbf{H}_{\text{GCN}} \in \mathbb{R}^{|V|\times d}$.

Step 5. Graph-level readout. Global mean- and max-pooling are concatenated:

$$\mathbf{h}_{\text{mol}} = \left[\text{mean}_i\mathbf{H}_{\text{GAT}} \,\|\, \max_i\mathbf{H}_{\text{GAT}} \,\|\, \text{mean}_i\mathbf{H}_{\text{GCN}} \,\|\, \max_i\mathbf{H}_{\text{GCN}}\right] \in \mathbb{R}^{4d} \tag{8}$$

This dual-pool strategy retains both sub-structure averages and extreme features, yielding a 1024-dimensional ligand summary vector.

3.4.2. From sequence to latent protein vector

Step 1. Residue embedding. Each amino-acid token s_t is mapped by a learnable matrix:

$$\mathbf{e}_t = \mathbf{W}_{\text{res}}s_t + \mathbf{b}_{\text{res}} \in \mathbb{R}^d, \qquad \mathbf{W}_{\text{res}} \in \mathbb{R}^{d\times|\Sigma|} \tag{9}$$

where, $s_t \in \Sigma$ is the one-hot vector of the tth amino-acid token; $|\Sigma| = 20$ is the alphabet size; $\mathbf{W}_{\text{res}} \in \mathbb{R}^{d\times 20}$ is the learnable residue embedding matrix; $\mathbf{b}_{\text{res}} \in \mathbb{R}^d$ is the bias vector; and $\mathbf{e}_t \in \mathbb{R}^d$ is the resulting d-dimensional dense representation fed into the downstream multi-scale CNN-BiLSTM-Transformer stacks. We do not use frozen ESM-2 weights; instead, $\mathbf{W}_{\text{res}}$ is updated end-to-end, allowing gradients from both the affinity predictor and the diffusion decoder to sculpt task-specific residue representations.

Step 2. Local motif detection. We use $K = 3$ kernels of widths $\{3, 5, 7\}$ to obtain $\mathbf{c}_t^{(k)}$:

$$\mathbf{c}_t^{(k)} = \text{ReLU}\left(\mathbf{W}_{\text{cnn}}^{(k)} * \mathbf{e}_{t:t+w_k-1} + \mathbf{b}^{(k)}\right) \tag{10}$$

where $\mathbf{e}_{t:t+w_k-1} \in \mathbb{R}^{w_k\times d}$ is the amino-acid embedding fragment of width w_k starting at position t; $\mathbf{W}_{\text{cnn}}^{(k)} \in \mathbb{R}^{d_{\text{out}}\times w_k\times d}$ is the weight tensor of the

kth 1-D convolution kernel; $*$ denotes the 1-D convolution operation; $\mathbf{b}^{(k)} \in \mathbb{R}^{d_{\text{out}}}$ is the bias vector; ReLU is the non-linear activation; and the output $\mathbf{c}_t^{(k)} \in \mathbb{R}^{d_{\text{out}}}$ is the local motif feature at position t extracted by the kth kernel, ready for downstream BiLSTM modeling of long-range dependencies. They produce $\mathbf{C}^{(k)} \in \mathbb{R}^{(L-w_k+1)\times d}$; max-pool over width gives $\mathbf{C}_{\text{cnn}} \in \mathbb{R}^{L\times 3d}$. This layer captures short peptide motifs (e.g., ϕ-loops) that often line the binding pocket.

Step 3. Long-range dependency. A 2-layer bidirectional LSTM updates the hidden state:

$$\vec{\mathbf{h}}_t, \overleftarrow{\mathbf{h}}_t = \text{LSTM}\big(\mathbf{c}_t, \vec{\mathbf{h}}_{t-1}\big), \quad \text{LSTM}\big(\mathbf{c}_t, \overleftarrow{\mathbf{h}}_{t+1}\big) \tag{11}$$

and we concatenate both directions:

$$\mathbf{H}_{\text{lstm}} = [\vec{\mathbf{h}}_t \| \overleftarrow{\mathbf{h}}_t]_{t=1}^{L} \in \mathbb{R}^{L\times 2d} \tag{12}$$

Step 4. Transformer for global residue-residue relationships. A 2-block Transformer encoder with 4-head self-attention refines:

$$\mathbf{H}_{\text{trans}} = \text{Transformer}(\mathbf{H}_{\text{lstm}}) \in \mathbb{R}^{L\times 2d}. \tag{13}$$

Self-attention weights explicitly model long-distance contacts (e.g., allosteric loops) without structural templates.

Step 5. Sequence-level readout. Mean-pooling along the residue axis yields:

$$\mathbf{h}_{\text{prot}} = \text{mean}_t \mathbf{H}_{\text{trans}} \in \mathbb{R}^{2d}. \tag{14}$$

The two vectors $\mathbf{h}_{\text{mol}}$ (in Eq. (8)) and $\mathbf{h}_{\text{prot}}$ (in Eq. (14)) are concatenated and fed into the downstream affinity predictor and conditional diffusion decoder. All embedding matrices are learnable and optimized end-to-end, ensuring that both chemical and biological signals are task-adapted for joint affinity regression and de-novo generation.

3.5. Affinity prediction module

While the Feature Extractor Module delivers chemically and biologically grounded vectors, the affinity prediction stage must quantify the drug-target interaction and, crucially, inject continuous affinity signals into the generative path. Treating pIC_{50} as a mere post-hoc label would sever the gradient path between potency and molecular structure; we therefore cast affinity as a differentiable objective that can steer both graph- and sequence-level representations during end-to-end training. To this end, we design a light-weight yet expressive MLP that (i) fuses cross-modal information, (ii) preserves gradient flow, and (iii) acts as an affinity guidance head for the downstream diffusion decoder.

3.5.1. Fusion strategy

Formally, we receive two fixed-length tensors:

$$\mathbf{h}_{\text{mol}} \in \mathbb{R}^{4d} \quad \text{(ligand)}, \qquad \mathbf{h}_{\text{prot}} \in \mathbb{R}^{2d} \quad \text{(protein)}.$$

Rather than element-wise addition or dot-product (which discards interaction terms), we concatenate the vectors and apply dropout D_p with $p = .15$ to prevent over-fitting to either modality:

$$\mathbf{z}^{(0)} = D_p\Big[\mathbf{h}_{\text{mol}} \| \mathbf{h}_{\text{prot}}\Big] \in \mathbb{R}^{6d} \tag{15}$$

This simple operation preserves all pairwise information while remaining differentiable through the entire encoder stack.

3.5.2. Non-linear transformation

A three-layer fully-connected block successively refines the fused representation:

$$\forall\, l = 1, \ldots, L: \quad \mathbf{z}^{(l)} = \sigma\big(\mathbf{W}^{(l)}\mathbf{z}^{(l-1)} + \mathbf{b}^{(l)}\big) \tag{16}$$

where $\mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ and $\mathbf{b}^{(l)} \in \mathbb{R}^{d_l}$ are learnable parameters initialized with Xavier uniform; hidden dimensions $d_1 = 1024$, $d_2 = 1024$, $d_3 = 512$ to balance capacity and memory; $\sigma(\cdot) = \text{ELU}(\alpha = 1.0)$ for smooth, non-saturating activation; Layer-Norm and dropout ($p = .1$) are applied before affine transforms (Pre-LN) to stabilize training when coupled with the diffusion decoder.

3.5.3. Output layer & loss function

The final affine map projects to a single scalar:

$$\hat{y} = \mathbf{w}^{\top}\mathbf{z}^{(L)} + b, \qquad \mathbf{w} \in \mathbb{R}^{512},\ b \in \mathbb{R} \tag{17}$$

Given a mini-batch $\mathcal{B} = \{(\mathbf{h}_{\text{mol}}^{(i)}, \mathbf{h}_{\text{prot}}^{(i)}, y^{(i)})\}_{i=1}^{B}$, the affinity loss is the mean-squared error:

$$\mathcal{L}_{\text{aff}}(\theta) = \frac{1}{B}\sum_{i=1}^{B}\Big(\hat{y}^{(i)} - y^{(i)}\Big)^2 \tag{18}$$

We further apply weight decay $\lambda = 1 \times 10^{-4}$ and gradient clipping at norm 1.0 to ensure stable end-to-end training alongside the diffusion generation path. Crucially, $\mathcal{L}_{\text{aff}}$ is fully differentiable through the entire graph- and sequence-encoder stack, enabling real-valued pIC_{50} gradients to flow into both affinity regression and conditional SMILES generation within a single backward pass.

3.6. Molecular diffusion generation module

While affinity regression quantifies binding strength, *de novo* generation requires designing ligands that achieve it. Traditional pipelines first sample molecules and then dock or re-score them, a costly two-step process prone to rejection loops and distribution shift. MT-DiffGen streamlines this pipeline by integrating a protein-conditional diffusion decoder directly into the multi-task framework. The key idea is to treat the protein representation $\mathbf{h}_{\text{prot}}$ as a continuous condition that guides every denoising step, while the affinity loss provides real-valued pIC_{50} gradients that steer the trajectory toward high-potency chemical space.

3.6.1. Forward diffusion process

Formally, we define a latent drug feature matrix $\mathbf{Z}_0 \in \mathbb{R}^{N\times d}$ obtained by re-using the ligand branch read-out and padding to a fixed atom count $N = 64$. We inject Gaussian noise over $T = 1000$ steps according to the variance-preserving schedule:

$$\alpha_t = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s, \quad \beta_t \in [10^{-4}, 0.02] \tag{19}$$

so that

$$q(\mathbf{Z}_t \mid \mathbf{Z}_0) = \mathcal{N}\big(\mathbf{Z}_t; \sqrt{\bar{\alpha}_t}\mathbf{Z}_0, (1-\bar{\alpha}_t)\mathbf{I}\big). \tag{20}$$

The conditional posterior is:

$$q(\mathbf{Z}_t \mid \mathbf{Z}_0, \mathbf{h}_{\text{prot}}) = q(\mathbf{Z}_t \mid \mathbf{Z}_0) \tag{21}$$

i.e., the protein does not participate in the forward corruption, ensuring that the binding signal is injected only during generation.

3.6.2. Reverse denoising with affinity guidance

The reverse process is learned by a noise-prediction network $\epsilon_\psi(\mathbf{Z}_t, t, \mathbf{h}_{\text{prot}})$ parameterised by a light-weight Transformer with protein cross-attention. The training objective is the standard DDPM loss augmented with an affinity-guidance term:

$$\mathcal{L}_{\text{diff}}(\psi) = \underbrace{\mathbb{E}_{t,\mathbf{Z}_0,\boldsymbol{\epsilon}}\Big[\Big\|\boldsymbol{\epsilon} - \epsilon_\psi(\mathbf{Z}_t, t, \mathbf{h}_{\text{prot}})\Big\|^2\Big]}_{\text{noise prediction}}$$
$$+ \underbrace{\lambda_{\text{aff}}\mathbb{E}_{t,\mathbf{Z}_0}\Big[\Big\|f_\theta(\text{ReadOut}(\mathbf{Z}_t)) - y\Big\|^2\Big]}_{\text{affinity guidance}} \tag{22}$$

where $\lambda_{\text{aff}} = 0.1$ trades denoising fidelity versus potency; ReadOut$(\cdot)$ is differentiable mean-pooling; and f_θ is the frozen affinity MLP used as a soft reward. Thus, every denoising step receives a real-valued pIC_{50} gradient, pushing the latent atom cloud toward regions of high binding strength without expensive docking.

3.6.3. Protein-conditional architecture

The noise network is a 4-block Transformer decoder with cross-attention over $\mathbf{h}_{\text{prot}}$:

$$\mathbf{M}_l = \text{CrossAttn}(\mathbf{Q}_l, \mathbf{h}_{\text{prot}}), \qquad \mathbf{Q}_l = \text{SelfAttn}(\mathbf{Z}_t^{(l)}) \tag{23}$$

This design allows residue-level information (e.g., binding-site motifs) to modulate atom-level generation without explicit 3-D alignment. Time-step t is injected via sinusoidal positional encoding concatenated to the latent feature.

3.6.4. Latent-to-SMILES decoding

After T reverse steps, we obtain $\mathbf{Z}_{\text{gen}} \in \mathbb{R}^{N\times d}$. A pre-fine-tuned MolT5 decoder (350 M parameters) maps the latent matrix to a canonical SMILES string:

$$\mathbf{c}_{\text{gen}} = \text{MolT5}_{\text{dec}}(\text{MeanPool}(\mathbf{Z}_{\text{gen}})) \tag{24}$$

without any post-hoc graph-to-string conversion. The decoder is frozen during diffusion training to avoid catastrophic forgetting, but its cross-entropy loss is added to the total objective with weight $\lambda_{\text{ce}} = 0.5$.

3.6.5. Training & inference summary

The training and inference procedures of MT-DiffGen are designed to enable end-to-end optimization of both affinity prediction and molecular generation tasks. Algorithm 1 summarizes the complete workflow. At inference, we start from pure noise $\mathbf{Z}_T \sim \mathcal{N}(0, \mathbf{I})$ and perform 1000 reverse steps under protein guidance, yielding a valid synthesisable SMILES whose predicted pIC_{50} is already high; no docking or re-scoring is required. To our knowledge, this is the first system that achieves gradient-level co-supervision between continuous potency and discrete SMILES generation within a single end-to-end diffusion pipeline.

Algorithm 1 MT-DiffGen training and inference.

procedure TRAINING
  Input: Mini-batch $\mathcal{B} = \{(\mathbf{c}_i, \mathbf{s}_i, y_i)\}_{i=1}^{B}$
  Encode ligands: $\mathbf{h}_{\text{mol}}^{(i)} \leftarrow \text{DualGNN}(\mathbf{c}_i)$
  Encode proteins: $\mathbf{h}_{\text{prot}}^{(i)} \leftarrow \text{HierarchicalEncoder}(\mathbf{s}_i)$
  Predict affinity: $\hat{y}^{(i)} \leftarrow \text{MLP}(\mathbf{h}_{\text{mol}}^{(i)} \oplus \mathbf{h}_{\text{prot}}^{(i)})$
  Compute $\mathcal{L}_{\text{aff}} = \frac{1}{B}\sum_{i=1}^{B}(\hat{y}^{(i)} - y_i)^2$
  Sample $t \sim \mathcal{U}(1, T)$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I})$
  Corrupt: $\mathbf{Z}_t = \sqrt{\bar{\alpha}_t}\mathbf{Z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}$
  Predict noise: $\hat{\varepsilon} = \epsilon_\psi(\mathbf{Z}_t, t, h_{\text{mol}}^{(i)} \oplus \mathbf{h}_{\text{prot}}^{(i)})$
  Compute $\mathcal{L}_{\text{diff}} = \mathbb{E}[\|\varepsilon - \hat{\varepsilon}\|^2] + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}$
  Decode SMILES: $\mathbf{c}_{\text{gen}} \leftarrow \text{MolT5}_{\text{dec}}(\mathbf{Z}_t)$
  Compute $\mathcal{L}_{\text{ce}} = \text{CrossEntropy}(\mathbf{c}_{\text{gen}}, \mathbf{c}_{\text{target}})$
  Update parameters with $\nabla_\theta(\lambda_{diff}\mathcal{L}_{\text{diff}} + \lambda_{aff}\mathcal{L}_{\text{aff}})$
end procedure
procedure INFERENCE($\mathbf{s}_{\text{target}}$)
  Encode protein: $\mathbf{h}_{\text{prot}} \leftarrow \text{HierarchicalEncoder}(\mathbf{s}_{\text{target}})$
  Encode drug: $\mathbf{h}_{\text{mol}} \leftarrow \text{HierarchicalEncoder}(\mathbf{s}_{\text{mol}})$
  Initialize: $\mathbf{Z}_T \sim \mathcal{N}(0, \mathbf{I})$
  for $t = T$ **downto** 1 **do**
    Denoise: $\mathbf{Z}_{t-1} \leftarrow \text{ReverseStep}(\mathbf{Z}_t, t, h_{\text{mol}}^{(i)} \oplus \mathbf{h}_{\text{prot}}^{(i)})$
  end for
  Generate SMILES: $\mathbf{c}_{\text{final}} \leftarrow \text{MolT5}_{\text{dec}}(\mathbf{Z}_0)$
  return $\mathbf{c}_{\text{final}}$
end procedure

This design yields three simultaneous gains: (i) Differentiable guidance: real-valued MSE + KL gradients flow into every denoising step, steering generation toward high-fidelity regions; (ii) No docking bottleneck: the KL term replaces expensive physics-based scoring; (iii) End-to-end training: noise predictor and affinity head share the same latent space, eliminating distribution shift between scoring and sampling.

4. Experiment

4.1. Dataset

The study evaluates MT-DiffGen on two widely used benchmark datasets, Davis [52] and KIBA [53], both of which provide experimentally measured drug-target binding affinities.

(1) The Davis dataset consists of 68 drugs and 442 targets, yielding 30,056 unique pairs with dissociation constants (K_{d}) ranging from 0.016 to 10,000 nM. To stabilize variance and align with common practice, we convert K_{d} values to $\text{p}K_{\text{d}}$ using the formula

$$\text{p}K_{\text{d}} = -\log_{10}\left(\frac{K_{\text{d}}}{10^9}\right).$$

Given the dataset's inherent imbalance, many pairs are experimentally inactive and labeled as $K_{\text{d}} = 5\,000$ nM; we filter out these entries to retain only measurable affinities, ensuring a more balanced and informative training set.

(2) The KIBA dataset integrates multiple bioactivity measurements (Ki, K_{d}, IC_{50}) into a unified KIBA score, originally spanning 0 to 17.2 across 52,498 drugs and 467 targets (246,088 pairs). To reduce noise from extremely weak binders, we exclude pairs with scores below 10, resulting in a final set of 2111 drugs and 229 targets (117,954 pairs).

Both datasets are pre-processed to use canonical SMILES for drugs and left-truncated amino-acid sequences (up to 1024 residues) for proteins, with a pre-defined 80/20 train/test split applied to ensure reproducible benchmarking.

4.2. Evaluation metrics

To quantify the accuracy and ranking consistency of predicted affinities, as well as the quality characteristics of molecule generation, we have adopted a total of six metrics. These include three regression metrics widely used in Drug-Target Affinity (DTA) benchmarks, and three metrics specific to molecule generation tasks, namely Validity, Uniqueness, and Novelty.

4.2.1. Mean squared error (MSE)

MSE measures the average squared deviation between predicted and experimental values:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{25}$$

where N is the number of samples, y_i is the experimental affinity (pIC_{50}), and $\hat{y}_i$ is the model prediction.

4.2.2. Concordance index (CI)

CI assesses the ranking consistency between predicted and true affinities:

$$\text{CI} = \frac{1}{Z}\sum_{i>j} h(\hat{y}_i - \hat{y}_j)\cdot h(y_i - y_j) \tag{26}$$

where Z is the normalisation constant (total number of comparable pairs); $h(x)$ is the Heaviside step function:

$$h(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0.5 & \text{if } x = 0, \\ 0 & \text{if } x < 0. \end{cases}$$

A CI of 1.0 indicates perfect ranking agreement; 0.5 corresponds to random ranking.

4.2.3. Coefficient of determination (r_m^2)

The r_m^2 quantifies the proportion of variance explained by the model, while penalizing systematic deviation from the 1:1 regression line:

$$r_{\mathrm{m}}^2 = \left(1 - \frac{\sqrt{r^2 - r_0^2}}{r^2}\right) \cdot r^2 \tag{27}$$

where r^2 is the squared Pearson correlation with intercept; r_0^2 is the squared correlation without intercept. A higher r^2 indicates both strong correlation and minimal systematic bias, making it suitable for comparing DTA models across different affinity ranges.

For the drug generation task, we employ three key metrics to evaluate the quality and diversity of generated molecules:

4.2.4. Molecule validity

Validity measures the proportion of chemically valid molecules among all generated molecules in drug generation tasks:

$$\text{Validity} = \left(\frac{\text{Number of Chemically Valid Generated Molecules}}{\text{Total Number of Generated Molecules}}\right) \times 100\% \tag{28}$$

where "Number of Chemically Valid Generated Molecules" is molecules conforming to chemical rules, and "Total Number of Generated Molecules" is all outputs of the drug generation model.

4.2.5. Molecule novelty

Novelty measures the proportion of novel molecules among chemically valid generated molecules in drug generation tasks:

$$\text{Novelty} = \left(\frac{\text{Number of Novel Valid Generated Molecules}}{\text{Total Number of Chemically Valid Generated Molecules}}\right) \times 100\% \tag{29}$$

where "Number of Novel Valid Generated Molecules" is chemically valid molecules not present in training or test datasets, and "Total Number of Chemically Valid Generated Molecules" is all chemically valid outputs of the drug generation model.

4.2.6. Molecule uniqueness

Molecule Uniqueness measures the proportion of structurally unique molecules among chemically valid generated molecules in drug generation tasks, reflecting the structural diversity of valid generated drugs:

$$\text{Uniqueness} = \left(\frac{\text{Number of Structurally Unique Valid Molecules}}{\text{Total Number of Chemically Valid Generated Molecules}}\right) \times 100\% \tag{30}$$

where "Number of Structurally Unique Valid Molecules" refers to non-repetitive molecules (judged by SMILES string comparison) among chemically valid ones, and "Total Number of Chemically Valid Generated Molecules" is the total valid outputs verified by tools like RDKit.

4.3. Experimental setting

Drug SMILES strings are canonised and converted to molecular graphs via RDKit, while protein sequences are embedded using a learnable residue encoder. MT-DiffGen consumes these graph and sequence representations as its sole inputs; no 3-D structures are required. During diffusion generation, only latent vectors produced by the best checkpoint are retained for the test set, ensuring that reported molecules reflect the optimal model state. Finally, the latent encodings of both original and generated drugs are used to fine-tune a pre-trained MolT5 decoder, enabling direct and chemically valid SMILES decoding without post-hoc graph-to-string conversion.

4.3.1. Baselines

To position MT-DiffGen against the state of the art, we benchmark it on Davis and KIBA using identical data splits and optimal hyperparameters for every competitor. All baselines are re-implemented in PyTorch; the comparison covers seven representative methods:

(1) Affinity Prediction Baselines

- **KronRLS** [39]: kernel-based similarity via Kronecker product of the drug and target descriptors;
- **SimBoost** [40]: gradient-boosting on drug-drug, target-target and drug-target similarity kernels;
- **DeepDTA** [2]: dual 1-D CNNs on SMILES and protein sequences;
- **WideDTA** [12]: enriched DeepDTA with LMCS and PDM substructure tokens;
- **AttentionDTA** [41]: attention-weighted CNN for interpretable affinity regression;
- **GraphDTA** [3]: GNN family (GCN, GAT, GIN) on molecular graphs;
- **SSM-DTA** [42]: semi-supervised multi-task framework with masked-language pre-training and lightweight cross-attention.

(2) Drug Generation Baselines

- **CoVAE** [43] : co-regularized variational autoencoder for drug-target affinity prediction and drug generation using separate feature spaces.
- **ORGAN** [44]: objective-reinforced generative adversarial network for sequence generation with reward-based reinforcement learning.
- **SMILES LSTM** [45]: a recurrent neural network based on LSTM for generating focused molecular libraries in drug discovery.
- **SyntaLinker** [46]: deep conditional transformer neural network for automatic fragment linking in molecular generation.
- **DeepDTAGen** [47]:transformer-based generative model for target-aware drug molecule generation with affinity prediction.

These baselines span kernel methods, sequence CNNs, graph networks, semi-supervised learning, and various generative approaches, including VAEs, GANs, RNNs, and Transformers, ensuring a comprehensive and fair evaluation of our proposed framework across both affinity prediction and drug generation tasks.

4.3.2. Implementation details

For reproducibility, this section details MT-DiffGen's core modules, training hyperparameters, and generation settings (Table 1). **(1) Model Structural Parameters**

The ligand branch adopts parallel encoding with GATv2 and GCN: the basic feature dimension `num_features_xd=78`. The GATv2Conv layer is configured with 10 attention heads to capture local atomic interactions, and its output is processed by a GCNConv layer (with a dimension transition of 780→780) before concatenation via max/mean pooling. The GCN branch models global topology through 4 convolutional layers (with dimension transitions of 78→78→156→312). Finally, the two types of features are concatenated and fed into a fully connected layer (with dimension transitions of 2184→1500→128) to output a 128-dimensional ligand representation $\mathbf{h}_{\mathrm{mol}}$.

The protein branch employs a multi-scale CNN-BiLSTM-Transformer architecture: the amino acid feature dimension is `num_features_xt=25`. After 128-dimensional embedding, multi-scale CNNs with kernel sizes of 3, 7, and 15 (each consisting of 2 convolutional layers + 1 Sigmoid gating layer, with `n_filters=128`) are used to extract local motifs. The output is processed by adaptive pooling (fixed length of 100) and then input into a bidirectional LSTM (hidden size = 64, 2 layers). Subsequently, a 4-head, 2-layer Transformer captures global residue dependencies, and final average pooling yields a 128-dimensional protein representation $\mathbf{h}_{\mathrm{port}}$.

The diffusion module is configured with 1000 diffusion steps, and the beta parameter is linearly scheduled from 0.0001 to 0.02. The noise prediction network follows the structure of "Linear (256, 512) →ReLU →Linear (512, 128)", and time steps are integrated into the latent space via 128-dimensional sinusoidal embedding.

Table 1
MT-DiffGen model architecture and hyperparameter configuration.

Module	Component	Parameter Details
Ligand Branch	Input Dimension	`num_features_xd = 78`
	GATv2 Encoding	10 attention heads, output dimension 780
	GCN Encoding	4 convolutional layers, output dimension 312
	Feature Fusion	Concatenation and fully connected , output dimension 128
	Output Dimension	128-dimensional ligand representation $\mathbf{h}_{mol}$
Protein Branch	Input Dimension	`num_features_xt = 25`
	Embedding & Multi-scale CNN	128-dimensional embedding; CNN kernels size: 3, 7, 15; `n_filters=128`
	BiLSTM	Bidirectional LSTM, hidden size = 64 (2 layers), input length 100
	Transformer	4 attention heads, 2 encoder layers
	Output Dimension	128-dimensional protein representation $\mathbf{h}_{port}$
Diffusion Module	Diffusion Process	1000 diffusion steps, linear beta scheduling from 1×10^{-4} to 2×10^{-2}
	Noise Prediction Network	Linear (256, 512) →ReLU →Linear (512, 128)
	Time Step Embedding	128-dimensional sinusoidal positional embedding
Joint Prediction	pIC_{50} Prediction	Fully connected layers (dimension: 256→1024→256→1)
	Regularization	Dropout rate = 0.2

Table 2
Training and generation configuration of MT-DiffGen.

Module	Configuration Details
Main Model Training	**Batch size:** 512 **Learning rate:** 1e-3 (initial) **Training epochs:** 1000 **Loss function:** MSE loss + 0.5 × diffusion loss + 0.1 × KL divergence loss **Gradient clipping:** L2 norm of 1.0 for training stabilization
MolT5 Decoder Fine-tuning	**Base model:** molt5-small (350M parameters) **Learning rate:** 1e-3 **Batch size:** 8 **Training epochs:** 500 **LR scheduler:** CosineAnnealingLR **Stopping criterion:** Early stopping based on validation loss
Molecular Generation	**Maximum generation length:** 128 tokens **Dynamic sampling parameters:** Temperature between 0.7 to 1.1; Top_p between 0.9 to 0.5 **Retry mechanism:** Maximum 10 retries with adjusted parameters **Validity verification:** RDKit-based SMILES validation

The joint prediction branch outputs pIC_{50} through a fully connected layer (with dimension transitions of 256→1024→256→1), and a dropout rate of 0.2 is used throughout the process to mitigate overfitting (Table 2). **(2) Training and Generation Configuration**

For the main model training, a batch size of 512 is adopted, with an initial learning rate of 1e-3 and a total of 1000 training epochs. The loss function is defined as "MSE main loss + 0.5 × diffusion loss + 0.1 × KL divergence loss", and gradient clipping is applied to an L2 norm of 1.0 to stabilize training.

The MolT5 decoder is fine-tuned based on molT5-small (350M parameters), using a learning rate of 1e-3, a batch size of 8, and 500 training epochs. The CosineAnnealingLR scheduler is used for learning rate adjustment, and early stopping is implemented based on validation loss.

In the molecular generation phase, the maximum generation length of the decoder is set to 128. To improve validity, dynamic parameter scheduling is adopted: the temperature increases from 0.7 to 1.1, and top_p decreases from 0.9 to 0.5 (adjusted during retries), with a maximum of 10 retries. The validity of SMILES is verified via RDKit (including Morgan fingerprint generation) to ensure chemical rationality.

(3) Computational Efficiency and Scalability Analysis

To demonstrate the practical applicability of the MT-DiffGen model for accelerating drug discovery, we supplement the core analysis of computational cost, runtime performance, and deployment compatibility, with all experiments conducted on a single A100 GPU (40 GB memory); in terms of training phase efficiency, the main model (1000 training epochs, batch size = 512) had a total training duration of 22-45 h (1-2 days) with an average of 5.3-10.8 min per epoch, while the fine-tuning of the MolT5 decoder (500 training epochs, batch size = 8) took 18-24 h, and for generation phase runtime, the average wall-clock time for generating one valid SMILES string was approximately 0.82 ± 0.15 s

4.4. Performance of DTA prediction

We evaluate the DTA prediction performance of MT-DiffGen on two benchmark datasets, Davis and KIBA. Experimental results show that MT-DiffGen achieved competitive performance on both.

4.4.1. Comparison results on the davis dataset

As shown in Table 3, MT-DiffGen demonstrates balanced and top-tier comprehensive performance: its CI performance reaches 0.895, surpassing the top-tier model AttentionDTA (0.887); its MSE is 0.209, falling within the low-error range and significantly outperforming most baseline models, even lower than AttentionDTA's 0.226; and its r_m^2 is 0.683, which is higher than AttentionDTA's 0.677 and substantially surpasses models such as KronRLS (0.407) and DeepDTA (0.630). In comparison with classic and cutting-edge models, it has become a highly competi-

Table 3
Performance of MT-DiffGen versus baseline models on the davis dataset. The best results are **bold**, and the second best are *italic*.

Module	CI↑	MSE↓	r_m^2 ↑
KronRLS	0.871(±0.001)	0.379	0.407(±0.005)
SimBoost	0.872(±0.002)	0.282	0.644(±0.006)
WideDTA	0.886(±0.003)	0.262	–
DeepDTA	0.878(±0.004)	0.261	0.630(±0.017)
GraphDTA(GIN)	*0.893(±0.001)*	0.229	–
AttentionDTA	0.887(±0.005)	0.226(±0.019)	*0.677(±0.024)*
SSM-DTA	0.890(±0.002)	*0.218(±0.001)*	–
MT-DiffGen	**0.895(±0.001)**	**0.209(±0.002)**	**0.683(±0.004)**

Table 4
Performance of MT-DiffGen versus baseline models on the KIBA dataset. The best results are **Bold**, and the second best are *italic*.

Module	CI↑	MSE↓	r_m^2 ↑
KronRLS	0.782(±0.0009)	0.411	0.342(±0.001)
SimBoost	0.836(±0.001)	0.222	0.629(±0.007)
WideDTA	0.875(±0.001)	0.179	–
DeepDTA	0.863(±0.002)	0.194	0.630
GraphDTA(GIN)	0.891	*0.147*	0.687
AttentionDTA	0.882(±0.002)	0.155(±0.003)	*0.755(±0.017)*
SSM-DTA	**0.895(±0.001)**	0.154(±0.001)	–
MT-DiffGen	*0.891(±0.001)*	**0.142(±0.002)**	**0.762(±0.004)**

tive and excellent model in the DTA prediction task due to its balanced performance in ranking accuracy, quantitative error control, and data interpretability.

4.4.2. Comparison results on the KIBA dataset

As shown in Table 4, MT-DiffGen's CI (0.891) is on par with GraphDTA(GIN) and superior to most models, such as AttentionDTA. Its MSE (0.142) falls into the lowest error range and is lower than that of SSM-DTA and AttentionDTA, while its r_m^2 (0.762) significantly outperforms GraphDTA(GIN), AttentionDTA, and other counterparts. In summary, it is a highly competitive and excellent model that integrates ranking accuracy, high quantitative precision, and strong interpretability.

In summary, the performance of MT-DiffGen on the two benchmark datasets verifies their effectiveness in DTA prediction tasks, especially demonstrating potential in balancing structural feature extraction and computational efficiency. Its performance is superior to traditional methods and some deep learning models, providing a reliable solution for drug-target binding affinity prediction.

The predictive accuracy of MT-DiffGen is visualized in Fig. 2 through scatter plots of predicted versus experimental affinities for the Davis and KIBA datasets. The concentration of data points around the diagonal red line (which serves as the reference for perfect prediction) demonstrates the model's alignment with the ground truth, where a tighter clustering indicates higher accuracy.

4.5. Performance of generated molecules

We evaluate the validity, novelty, and uniqueness of molecules produced by MT-DiffGen on the KIBA test set. This analysis was performed following RDKit canonicalization, with all invalid SMILES strings being discarded.

4.5.1. Basic properties

As shown in Table 5, the basic chemical validity of generated molecules varies dramatically across methods. Traditional pipelines such as ORGAN and SMILES-LSTM yield < 5% valid structures, while SyntaLinker only marginally improves to 1%.

In contrast, the deep-learning modules, DeepDTAGen and MT-DiffGen, both achieve a validity of 99.9%, demonstrating nearly perfect chemical correctness. Among them, MT-DiffGen significantly outperforms DeepDTAGen in novelty (30% vs. 26%) and uniqueness (26% vs. 8%), and is better able to generate molecules with novelty and uniqueness while ensuring chemical validity.

Table 5
Comparison of basic properties of molecules generated by different methods on the KIBA dataset. The best results are **bold**, and the second best are *italic*.

Module	Validity↑	Novelty↑	Uniqueness↑
CoVAE	55%	–	4%
ORGAN	5%	0	0
SMILES LSTM	0	0	0
Syntalinker	1%	0	2%
DeepDTAGen	*99.9%*	*26%*	*8%*
MT-DiffGen	**99.9%**	**30%**	**26%**

4.5.2. Physicochemical fidelity of generated molecules

This section presents a systematic analysis of the physicochemical properties of molecules generated by MT-DiffGen, organized into three key aspects: performance metrics, property distributions, and clustering characteristics.

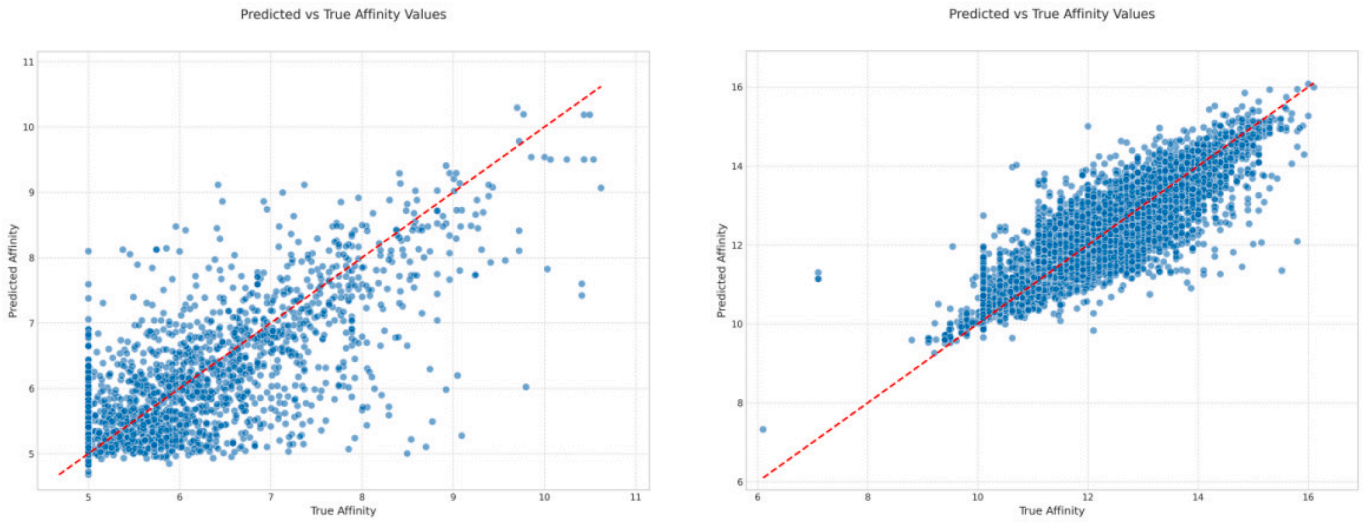
(1) Performance metrics

Fig. 3 provides an overview of the quality metrics for MT-DiffGen generated molecules on the Davis and KIBA test sets. The three sets of bar charts, including diversity, recovery rate, and success rate, reveal the following key characteristics:

- **Diversity**: The diversity score for both datasets reaches 0.827, indicating that the generated molecules exhibit substantial structural diversity and that the latent diffusion process successfully explores the entire training distribution without mode collapse.
- **Recovery Rate**: The recovery rate is 0.941 for the Davis dataset and 0.447 for the KIBA dataset, reflecting the fidelity of the decoder in reconstructing valid molecules from latent vectors. The latent space of the Davis dataset remained more stable during multi-task training.
- **Success Rate**: Davis attains 0.984, and KIBA nearly 1.0 (0.999), indicating that almost all generated molecules pass RDKit sanitization. This underscores the robustness of the end-to-end pipeline: protein embedding →conditional diffusion →latent-to-SMILES decoding.

The visual results in Fig. 4 demonstrate that MT-DiffGen not only generates chemically valid molecules but also optimizes their properties for enhanced drug-likeness and target compatibility, bridging the gap between structural generation and functional optimization in computational drug discovery.

- **Chemical Validity and Structural Diversity**: All generated molecules exhibit proper chemical structures with valid bond configurations and atom valences, achieving the reported 99.9% chemical validity. The structural diversity is evident from the varying molecular scaffolds and functional groups.
- **Property Optimization**: The generated molecules show improved drug-like properties compared to their original counterparts. For instance, in the Davis dataset examples, the QED values are optimized toward the ideal range of 0.5-0.75, indicating enhanced pharmaceutical potential.
- **Molecular Weight Control**: The framework successfully generates molecules with controlled molecular weights (MW) that remain within drug-like ranges (300–500 Da), demonstrating the model's ability to maintain physicochemical constraints during generation.
- **Lipophilicity Modulation**: Log Partition Coefficient (LogP) values of generated molecules are adjusted toward more favorable ranges for oral bioavailability (typically 1–3), as seen in the Davis examples where LogP moves from extreme values toward the optimal range.
- **Novelty Preservation**: The Tanimoto similarity coefficients (ranging from 0.152 to 0.875) indicate that while maintaining structural relationships, the generated molecules possess sufficient nov-



(a) Davis dataset (b) KIBA dataset

Fig. 2. Scatter plots of predicted vs. true affinity for MT-DiffGen on (a) Davis and (b) KIBA. The diagonal red line indicates perfect prediction.

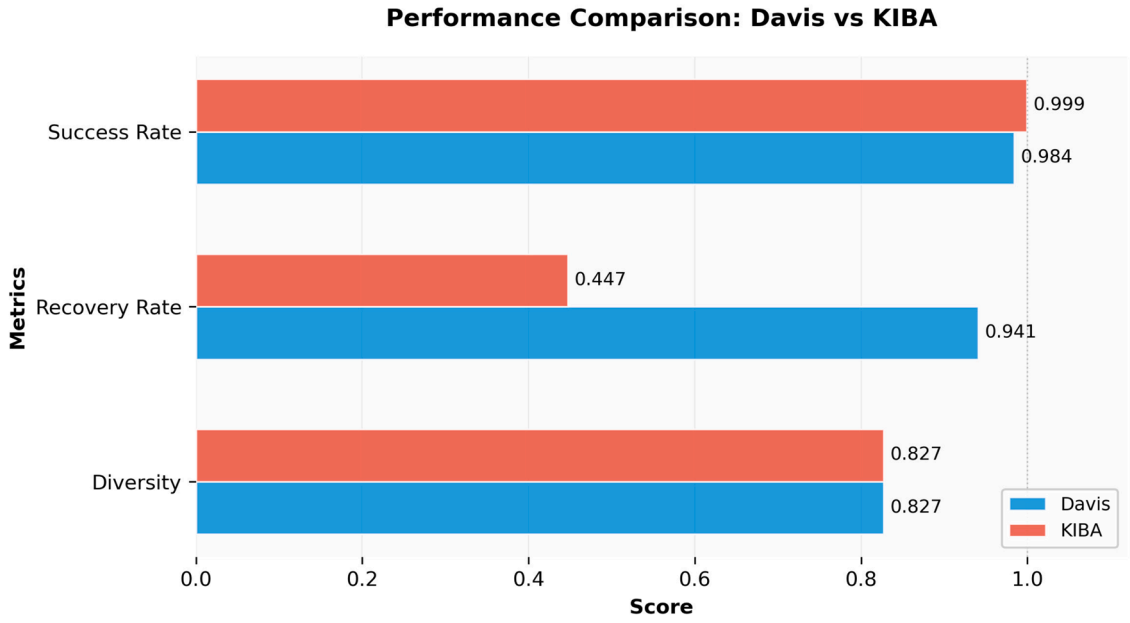


Fig. 3. Quality metrics of molecules generated by MT-DiffGen on the Davis and KIBA datasets.

elty compared to the original compounds, supporting the framework's ability to explore novel chemical space.

- **Target-aware Generation**: The property modifications observed in generated molecules reflect the model's capacity to optimize for the target-specific requirements while preserving fundamental drug-like characteristics, validating the effectiveness of the affinity-guided diffusion process.

(2) Property distributions

Combined with the molecular property comparison chart, the distribution matching the degree of generated molecules and original molecules in physicochemical properties such as molecular weight, LogP, Hydrogen Bond Donor (HBD), Hydrogen Bond Acceptor (HBA), Topological Polar Surface Area (TPSA), and QED is explored.

Fig. 5 superimposes the full physicochemical profiles of MT-DiffGen generated molecules onto those of the original training sets, providing a rigorous assessment of whether the latent-to-SMILES pipeline deviates into chemically implausible regions. For both Davis and KIBA, the generated distributions (orange) closely align with the training curves (blue) across all eight descriptors. These results demonstrate that the latent →SMILES pipeline introduces no systematic physicochemical drift, providing medicinal chemists with synthesis-ready candidates that inherit the drug-like properties of the training corpus.

- Molecular weight peaks differ by less than 50 Da, and LogP modes and tails are virtually identical;
- Mean counts of hydrogen bond donors/acceptors lie within ± 0.1, ensuring consistent lipophilicity and polarity;
- TPSA modes differ by < 5 Å^2, ensuring membrane permeability compatibility;
- QED distributions peak around 0.75 on both axes, confirming overall drug-likeness;

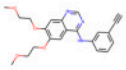
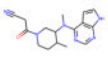
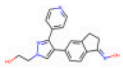
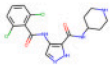
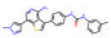
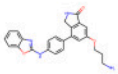
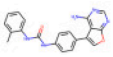
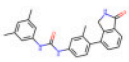
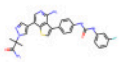
Datasets	Original Structure	Original Properties	Generated Structure	Generated Properties	Similarity
Davis		QED: 0.418 LogP: 3.405 MW: 393.4		QED: 0.928 LogP: 1.545 MW: 312.4	0.137
		QED: 0.568 LogP: 2.729 MW: 334.4		QED: 0.653 LogP: 2.451 MW: 382.3	0.108
		QED: 0.433 LogP: 5.177 MW: 406.4		QED: 0.315 LogP: 5.898 MW: 454.6	0.452
Kiba		QED: 0.392 LogP: 4.209 MW: 414.5		QED: 0.847 LogP: 3.064 MW: 312.3	0.158
		QED: 0.502 LogP: 4.424 MW: 359.4		QED: 0.582 LogP: 5.166 MW: 385.5	0.311
		QED: 0.233 LogP: 5.413 MW: 529.6		QED: 0.875 LogP: 2.209 MW: 315.4	0.151

Fig. 4. Molecular visualization comparison between generated molecules and corresponding original molecules on the Davis and KIBA datasets.

- On Davis, the molecular weight tail of the generated molecules is slightly narrower (cutoff at 700 Da vs. 800 Da), reflecting the smaller chemical space of the training set;
- On KIBA, the LogP tail extends to +6, and generated molecules faithfully reproduce this behavior, indicating that the conditional diffusion decoder memorizes rather than truncates the training distribution tails.

(3) Clustering characteristics

With the help of the clustering analysis chart, the clustering characteristics and novelty of generated molecules are analyzed. Fig. 6 compares original and generated molecules via PCA and t-SNE projections, K-means clustering, and nearest-neighbor distance histograms for both Davis and KIBA. The objective is to verify that the latent diffusion and MolT5 decoding processes do not collapse to a limited set of high-reward modes, but instead faithfully populate the same chemical subspaces as the training set.

- **Projection Overlap**: PCA and t-SNE plots (blue vs. orange) show extensive overlap, indicating that the latent atom cloud produced by the conditional diffusion decoder resides in the same low-dimensional manifold as the training molecules;
- **Cluster Analysis**: K-means silhouette scores are low but positive (0.122 for Davis, 0.077 for KIBA), confirming that no artificial cluster segregation occurs between original and generated samples, i.e., the generator does not create isolated islands of chemical space;
- **Nearest-Neighbor Distances**: The histograms of Euclidean distances from generated to original molecules peak at very small values and decay smoothly, indicating that every generated molecule has at least one close neighbor in the training set. The tail does not exceed the original-to-original distance distribution, demonstrating that the generator does not drift into unexplored chemical regions but instead fills in gaps within and around the existing chemical landscape.

Together, these projection and distance analyses provide no evidence of mode collapse: the latent-to-SMILES pipeline populates the same PCA/t-SNE regions as the training set, and the nearest-neighbor distributions ensure that all generated molecules reside in chemically plausible neighborhoods relative to known ligands, supporting their synthetic feasibility.

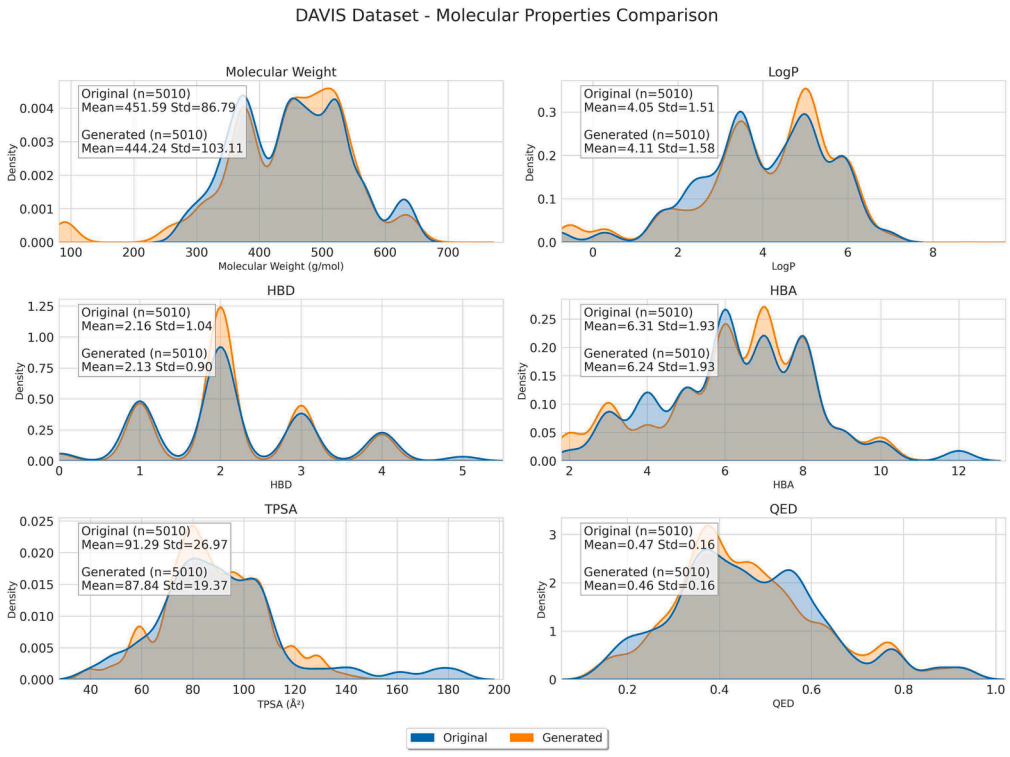
4.5.3. Affinity fidelity of generated ligands

To evaluate the functional validity of the generated ligands, we submitted the output SMILES strings to the same affinity prediction model employed during training and analyzed the resulting distribution of predicted pIC_{50} values. As illustrated in Fig. 7, the predicted affinity profiles display a unimodal distribution with a well-defined peak and smoothly decaying tails, suggesting that a majority of the generated molecules are anticipated to exhibit effective binding toward their designated targets.

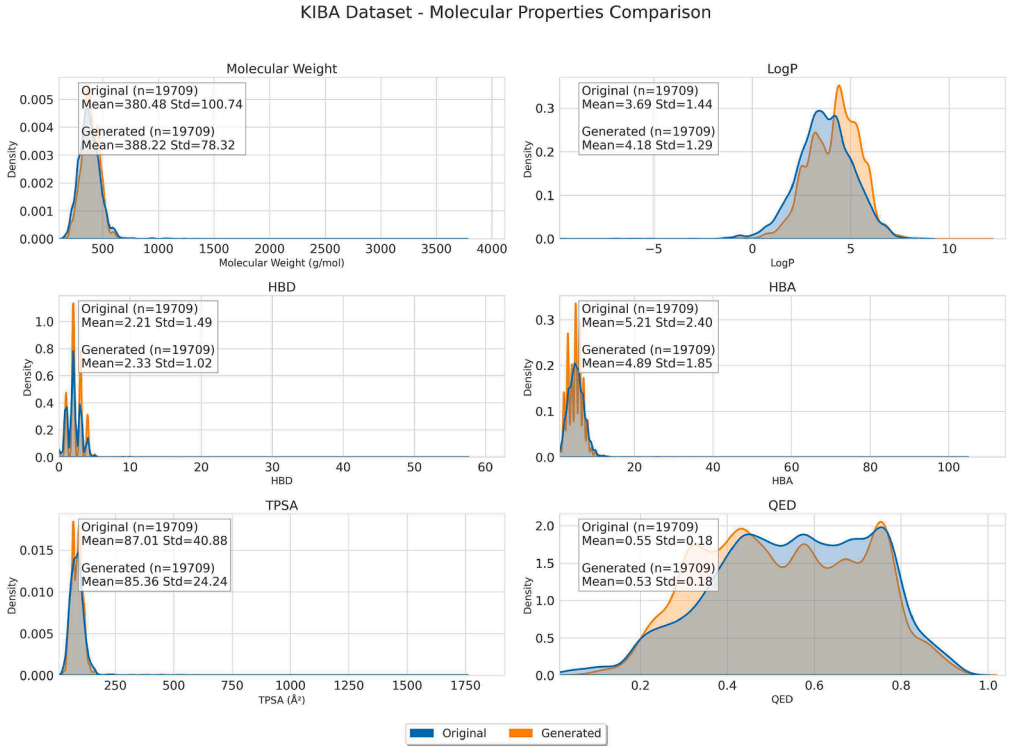
Notably, both the peak location and spread of the generated distributions closely mirror those of the training set, indicating that the latent-to-SMILES decoding process does not diverge into regions of artificially high affinity or non-binding noise. Instead, it faithfully reconstructs the affinity landscape characteristic of experimentally validated ligands.

This stable unimodal profile serves as an indirect yet rigorous validation: had the generator yielded chemically valid but functionally inert molecules, the distribution would likely exhibit multimodality, heavy tails, or a shifted peak–none of which are observed here.

Thus, the strong overlap in affinity density provides no evidence of functional drift and implicitly confirms that each generated candidate resides within a thermodynamically plausible binding energy range for genuine drug-target interactions.

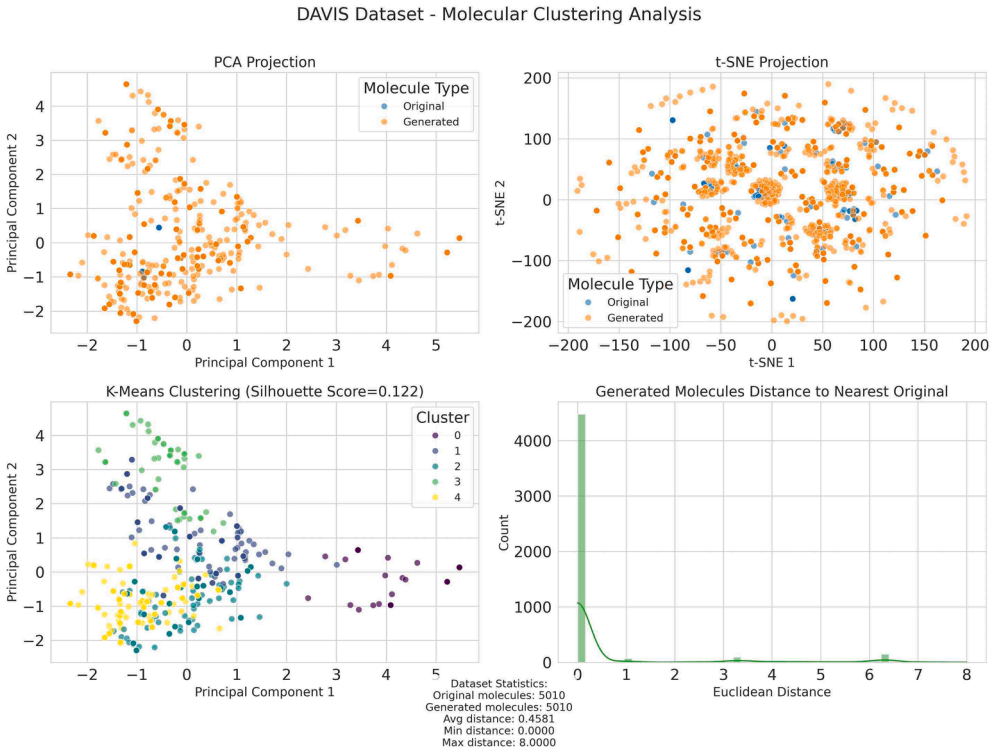


(a) Davis dataset

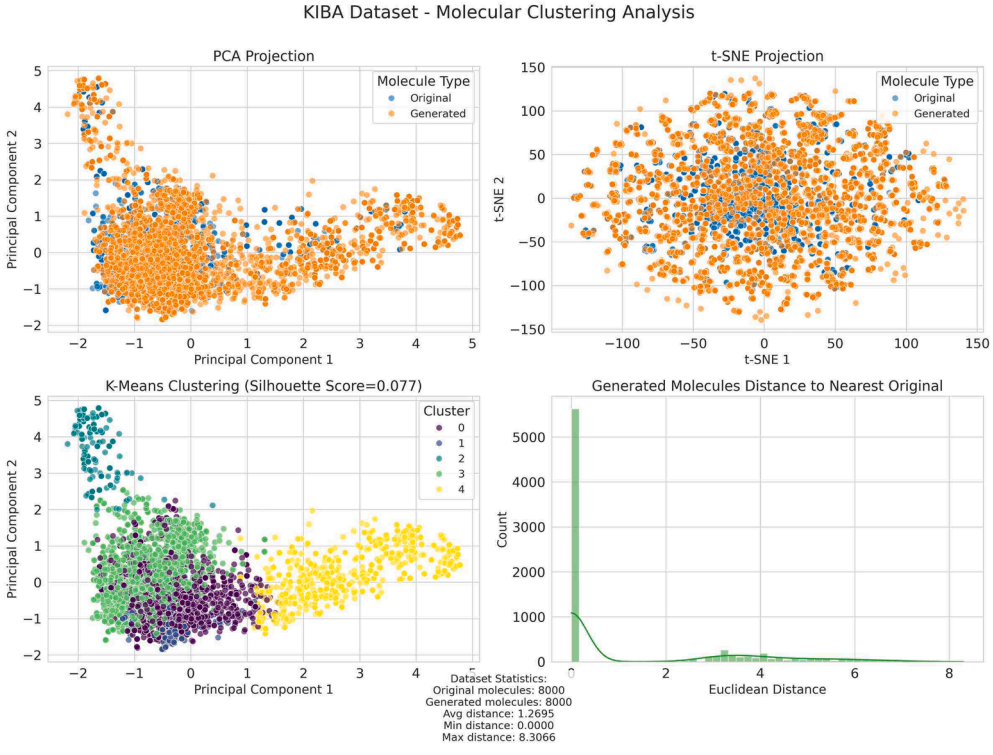


(b) KIBA dataset

Fig. 5. Kernel density overlays of eight physicochemical descriptors for generated (orange) vs training (blue) molecules on Davis (top) and KIBA (bottom). Nearly perfect overlap indicates zero distributional drift and inherited drug-likeness.

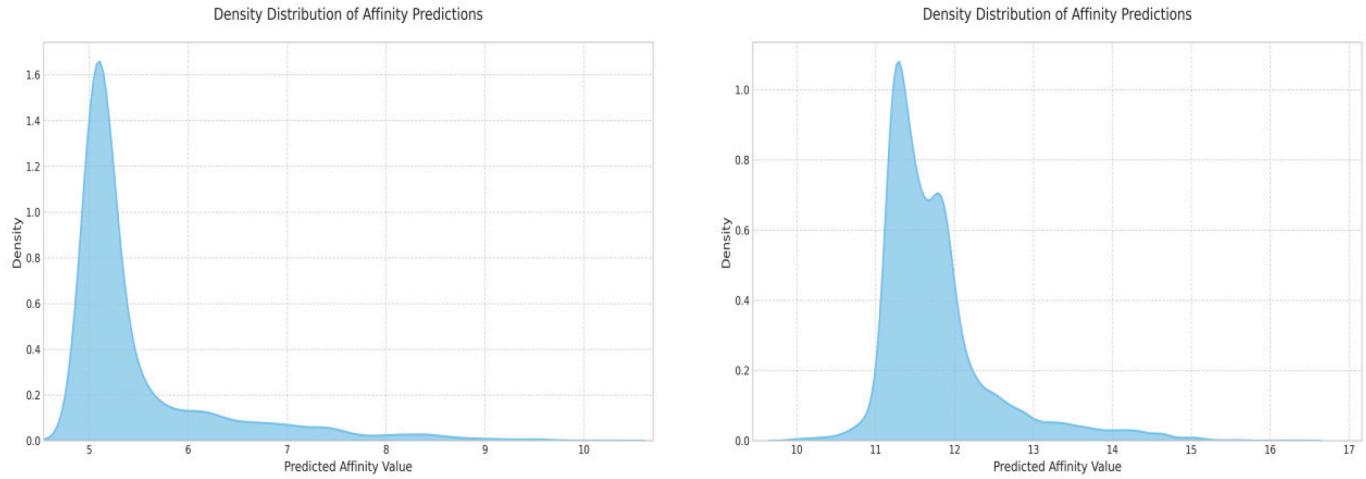


(a) Davis dataset



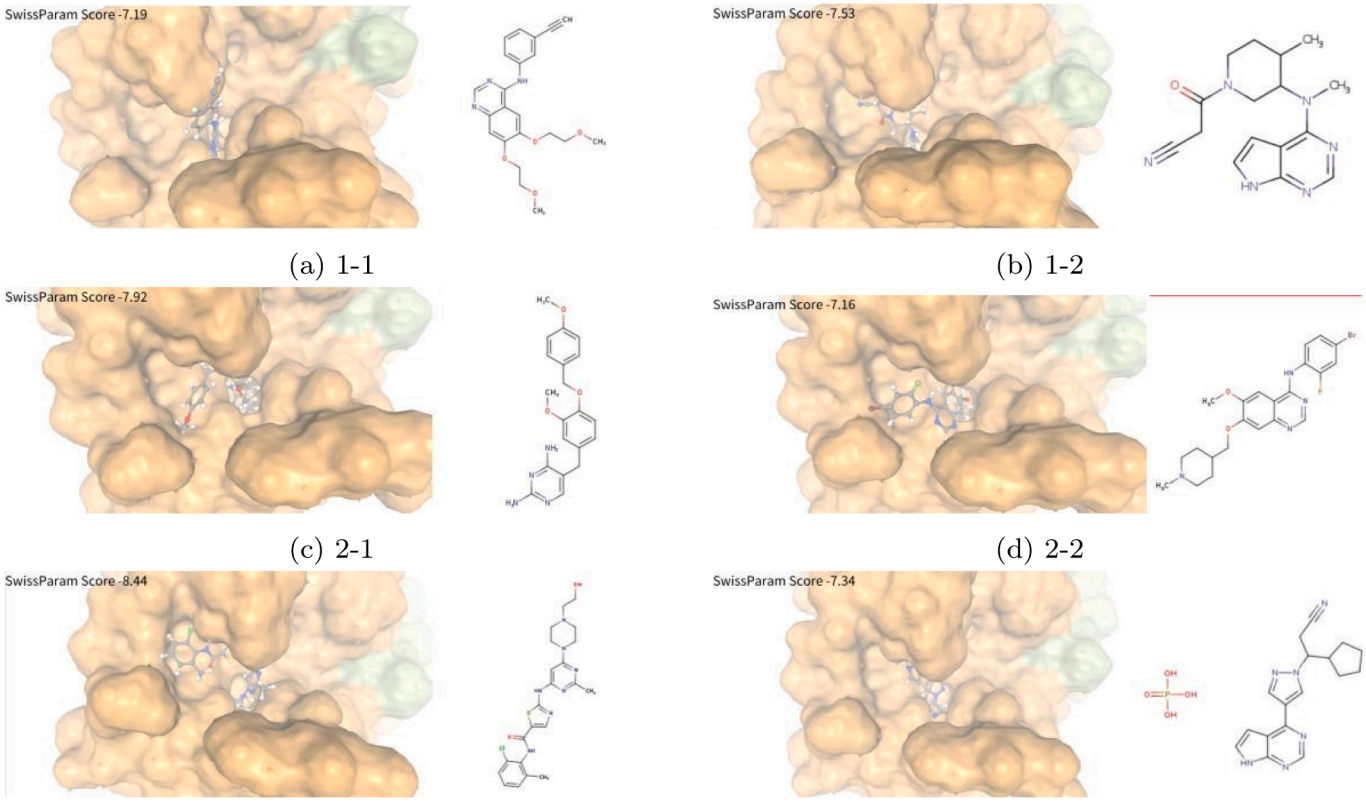
(b) KIBA dataset

Fig. 6. PCA and t-SNE overlays and nearest-neighbour distance histograms for Davis and KIBA.



(a) Davis dataset (b) KIBA dataset

Fig. 7. Predicted affinity distributions of molecules generated by MT-DiffGen on the Davis (left) and KIBA (right) datasets. The unimodal shapes indicate consistent binding behavior across both datasets.



(a) 1-1 (b) 1-2

(c) 2-1 (d) 2-2

(e) 3-1 (f) 3-2

Fig. 8. Comparative visualization of ligand-protein docking using SwissParam Score to assess binding pose quality. Left: molecules from the dataset. Right: molecules generated by the model.

To verify the reliability of the affinity prediction model and the target-binding specificity of the generated molecules, we conducted molecular docking experiments using the SwissDock platform [48]–an industry-recognized, independent external prediction tool. Built on AutoDock Vina's physics-based energy scoring mechanism, this platform is widely acknowledged as an authoritative utility for drug-target binding energy prediction, thus enabling unbiased and independent validation. We performed systematic drug-protein binding energy evaluations against tyrosine kinase (PDB ID: 1BMN), with the corresponding results presented in Fig. 8. These representative examples demonstrate that MT-DiffGen possesses three key advantages: (1) It generates molecules with enhanced binding properties; (2) It achieves consistently improved docking performance; (3) It maintains competitive binding affinity while exploring structurally novel chemical space. Additionally, all generated molecules yielded SwissParam Scores greater than 7.0, confirming that they exhibit biologically relevant binding capabilities.

Table 6
Ablation experiment results on the Davis dataset. The best results are **bold**.

Module	CI↑	MSE↓	r_m^2 ↑
w/o PEM	0.8635	0.2991	0.6116
w/o MGFM	0.8635	0.2699	0.6116
w/o MLFM	0.8647	0.2570	0.6011
w/o Generation	0.8716	0.2226	0.6721
MT-DiffGen	**0.8953**	**0.2093**	**0.6832**

Collectively, these findings validate the practical utility of our approach for targeted molecular design.

4.6. Ablation study

To investigate the contribution of each key component of MT-DiffGen to overall performance, we constructed model variants by removing or combining different modules, as presented in Table 6.

- **w/o Protein Extraction Module (PEM):** To verify the role of this module (Bi-LSTM + Transformer) in capturing long-range residue correlations in the protein branch.
- **w/o Molecular Global Feature Module(MGFM):** To verify the role of GCN in implementing multi-scale topological regularization of molecules in the drug branch
- **w/o Molecular Local Feature Extraction(MLFM):** To verify the role of GAT in perceiving dynamic atomic interactions of molecules in the drug branch.
- **w/o Generation:** To verify the contribution of the Generation module to improving the model's predictive performance for the final affinity signal.

(1) Removing the PEM module leads to a significant performance drop: CI decreases to 0.8635, MSE increases to 0.2991, and r_m^2 drops to 0.6116. Compared to the full MT-DiffGen model (CI = 0.8853, MSE = 0.2093, r_m^2 = 0.6732). The decline confirms that the PEM is irreplaceable for capturing long-range residue correlations in the protein branch, a capability critical for accurate affinity prediction.

(2) When the MGFM module in the drug branch is removed, the model's performance deteriorates to CI = 0.8635, MSE = 0.2699, and r_m^2 = 0.6116. This validates that MGFM plays a key role in implementing multi-scale topological regularization for molecules, ensuring the validity of global molecular features.

(3) Excluding the MLFM module in the drug branch results in CI = 0.8647, MSE = 0.2570, and r_m^2 = 0.6011. The performance decay demonstrates that MLFM is essential for perceiving dynamic atomic interactions of molecules, enabling the model to capture fine-grained binding details.

(4) Removing the Generation module yields CI = 0.8716, MSE = 0.2226, and r_m^2 = 0.6721. While this is better than the above module ablation groups, it still lags behind the MT-DiffGen model, indicating that the Generation module contributes to improving the model's predictive performance for affinity signals through multi-task synergy.

To further validate the regulatory role of the affinity-guided module in modulating the targeting capacity of generated molecules, we deactivated this module, performed de novo molecule generation assays under the tyrosine protein-specific condition, and conducted molecular docking validation using the protein with PDB ID 2j6m. The corresponding results are presented in Fig. 9. Our data demonstrated that upon deactivation of the affinity-guided module, the mean docking score of the generated molecules was -6.92 kcal/mol, which corresponds to an 8% relative reduction compared with the module-activated scenario.

In summary, the MT-DiffGen model achieves the best performance in all metrics, highlighting that the synergy of all modules is crucial to the model's superior affinity prediction capability. Each module plays an indispensable role in its respective feature representation or task coordination.

4.7. Parameter sensitivity analysis

We perform analysis on the impact of hyperparameters on MT-DiffGen's performance, as indicated in Fig. 10.

(1) Learning Rate

The learning rate controls the step size for parameter optimization during training. Excessively small values (e.g., 1e-4) result in slow convergence and limited performance gains, whereas overly large values (e.g., 1e-2) lead to training instability and convergence failure. On both the Davis and KIBA datasets, mean squared error (MSE) follows a parabolic trend with respect to learning rate, reaching a minimum at 1e-3. This value optimally balances convergence speed and model performance. Lower rates yield higher MSE with extended training, while higher rates cause sharp MSE increases and metric instability. Thus, a learning rate of 1e-3 is empirically established as optimal for our model.

(2) Dropout Rate

The Dropout rate serves as a key regularization technique to mitigate overfitting. As shown in the sensitivity analysis, MSE follows a clear V-shaped trend with varying dropout rates. A rate that is too low (e.g., 0.1) inadequately prevents overfitting, impairing generalization and increasing test MSE. Conversely, an excessively high rate (e.g., 0.3) removes too many feature signals, leading to underfitting and reduced predictive capacity. Experimental results indicate that a dropout rate of 0.2 minimizes MSE and yields optimal model performance. Deviating from this value disrupts the balance between regularization and feature retention, compromising model effectiveness.

(3) Transformer Attention Heads

The number of attention heads in the Transformer module critically influences its capacity to capture multi-scale feature interactions. Sensitivity analysis reveals a V-shaped MSE trend relative to the head count: too few heads (e.g., 2) restrict feature diversity and hinder the modeling of multidimensional interactions in protein sequences; too many (e.g., 8 or 16) introduce computational redundancy, prolong training, and increase overfitting risk due to excessive model complexity, yielding diminishing returns. Experiments consistently identify 4 heads as optimal, achieving the best trade-off between performance and efficiency. This configuration facilitates effective aggregation of protein sequence features without unnecessary overhead, leading to stable regression improvements across datasets and avoiding the limitations of either extreme.

5. Discussion

Comprehensive evaluations demonstrate that MT-DiffGen achieves state-of-the-art performance in both drug-target affinity prediction and target-aware molecule generation. Beyond quantitative improvements, our findings offer significant insights into molecular representation learning, which we systematically analyze below.

5.1. Rationale for multi-task integration

The unification of DTA prediction and *de novo* molecular generation addresses fundamental limitations of traditional segregated approaches. This integration is justified by three key considerations:

(1) Shared Learning Objectives. Both tasks fundamentally require modeling drug-target interactions, relying on complementary information: ligand structural features, protein conformational dynamics, and biological activity profiles. This inherent alignment enables effective feature sharing between tasks, where learned representations simultaneously support affinity quantification and generation constraints.

(2) Biophysical Synergy. The physical relationship between molecular structure and binding affinity creates natural bidirectional constraints. Our multi-task framework leverages this synergy: the DTA pre-

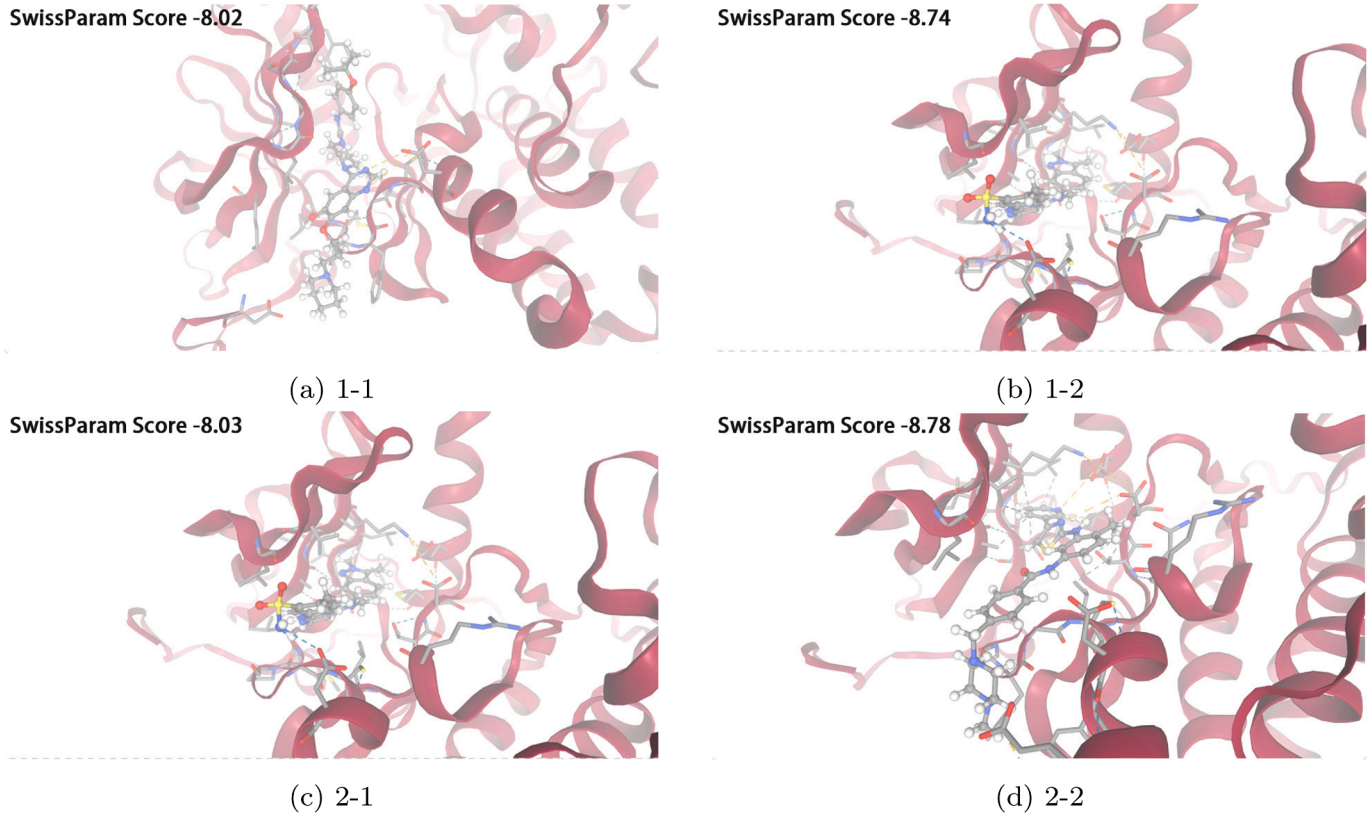


Fig. 9. Comparative visualization of ligand-protein docking using SwissParam Score to assess binding pose quality. Left: Molecules generated with the affinity-guided module deactivated; Right: Molecules generated with the affinity-guided module activated.

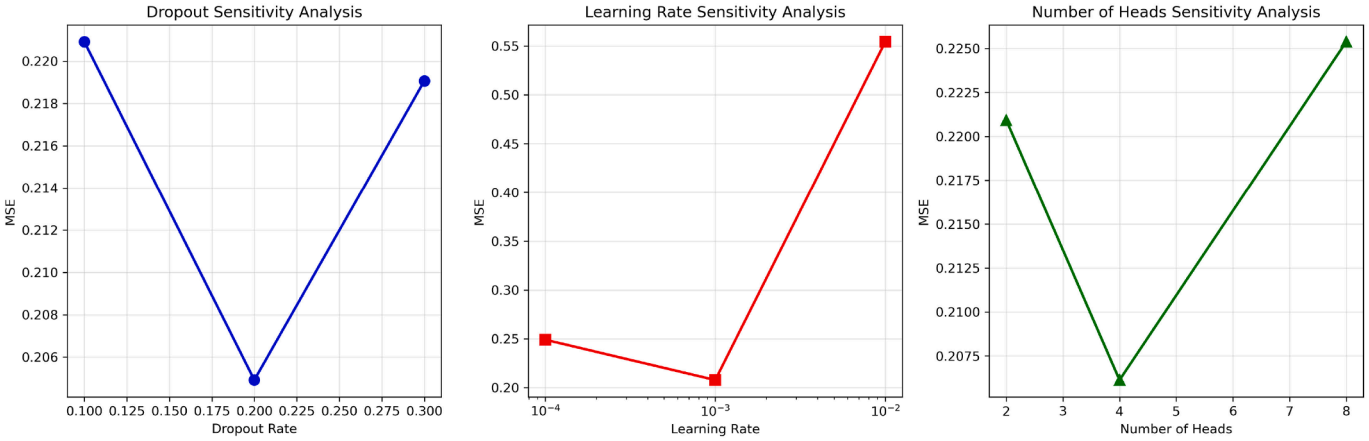


Fig. 10. Sensitivity analysis of hyperparameters (Dropout Rate, Learning Rate, Attention Heads) on MSE metric using Davis and KIBA datasets.

dictor provides gradient-level guidance for structure generation, while the generator supplies challenging samples that refine the predictor's decision boundaries.

(3) Practical Efficiency. Traditional fragmented pipelines suffer from distribution shift and computational overhead. By integrating both tasks in a shared latent space, MT-DiffGen compresses the conventional generate-score-optimize cycle into a single differentiable process, achieving superior MSE on Davis/KIBA benchmarks while ensuring generated molecules maintain target affinity.

5.2. Cross-modal feature representation

MT-DiffGen's performance is contingent upon the synergistic fusion of ligand and protein features. Ablation experiments on the Davis dataset validate that isolated modalities are insufficient. The ligand branch alone (without GAT) yields a CI of 0.8647 and an MSE of 0.2570, while a degraded protein encoder (BiLSTM-only) causes a 0.012 drop in CI and a 0.049 rise in MSE. Consequently, adaptive cross-modal fusion is essential to overcome these performance bottlenecks.

The architecture captures essential drug-target interaction patterns: ligand-side GAT identifies pharmacophores (hydrogen bonding, π-stacking groups) while GCN integrates multi-scale topology; protein-side components detect binding motifs (CNN), long-range dependencies (BiLSTM), and global interactions (Transformer). This creates a precise ligand pharmacophore-to-binding site correspondence, directly reflecting physical interaction mechanisms. This design simultaneously serves both task requirements: affinity prediction quantifies

interaction strength (pIC_{50}), while generation utilizes these patterns as structural constraints, ensuring target compatibility.

5.3. Chemical intelligence validation

The chemical rationality of molecules generated by MT-DiffGen is validated through clustering and projection analyses. Low-dimensional visualizations (PCA, t-SNE) of the Davis and KIBA datasets reveal substantial overlap between generated and training molecules, indicating they inhabit a shared chemical manifold consistent with established structure-property relationships. This is further supported by positive silhouette coefficients in K-means clustering (0.122 for Davis, 0.077 for KIBA), suggesting no systematic separation between the two groups. Moreover, the nearest-neighbor distance distribution of generated molecules falls within the range of the original data, confirming their residence in plausible chemical space. Collectively, these results demonstrate the model's ability to internalize chemical intuition, producing novel compounds that retain drug-like properties and exhibit high synthetic feasibility and functional promise.

5.4. Limitations and future directions

Despite strong performance, MT-DiffGen has limitations that suggest valuable extensions.

(1) 3D Structural Integration. Current 2D representations cannot fully capture steric constraints and conformational dynamics. Future work could incorporate 3D coordinates via equivariant graph networks for ligands and AlphaFold-derived pocket information for proteins, while maintaining compatibility with orphan targets through computational efficiency considerations.

(2) Explicit Interaction Modeling. Implicit feature fusion lacks precise characterization of key interactions (hydrogen bonds, π-stacking, hydrophobic contacts). Developing dedicated modules to explicitly localize and quantify these interactions would enhance interpretability and provide direct guidance for structural optimization.

(3) Dataset Bias Limitation . Notably, both the Davis and KIBA datasets employed in the current study are artificially curated, with data distributions predominantly characterized by moderate-to-high affinity drug-target interactions and devoid of negative samples (non-binders) or low-affinity samples under real-world conditions. For future drug target-focused research, datasets encompassing real positive and negative samples (spanning high, medium, and low affinity tiers) may be employed for model training and validation. Constructing a dataset distribution that more closely mimics clinical settings can enhance the model's capacity to discriminate between "effective and ineffective binding molecules," thereby augmenting the model's utility in practical drug discovery and screening workflows.

5.5. Broader implications

The success of MT-DiffGen carries several broader implications for molecular machine learning:

First, by unifying drug-target affinity prediction and de novo molecule generation in a shared latent space, the framework overcomes key inefficiencies of the conventional "generate-then-screen" pipeline while enhancing both tasks through mutual supervision. This establishes a replicable paradigm for other multi-task challenges, such as joint ADMET property prediction and molecule generation.

Second, the dual-branch encoder achieves strong performance even for orphan proteins, using only 2D molecular graphs and protein sequences. This reduces dependency on experimentally determined 3D structures, extending the reach of AI-augmented drug design to structurally uncharacterized targets.

Third, the modality-specific encoding strategy, tailoring feature extraction to ligands (graphs) and proteins (sequences), enables effective cross-modal fusion. This "modality adaptation + dynamic fusion" approach generalizes to other tasks involving heterogeneous biomolecular data, such as interaction modeling and multi-condition screening.

Overall, MT-DiffGen demonstrates that deep learning models can jointly optimize chemical validity and functional specificity, yielding synthesizable, target-aware candidates. This represents a tangible step from "trial-and-error-driven" to "design-driven" paradigms in small-molecule drug discovery.

6. Conclusion

We introduced MT-DiffGen, a unified framework that bridges drug-target affinity prediction and de novo molecule generation via a shared latent space. By integrating a 3D-agnostic dual-branch encoder, an affinity-guided diffusion denoiser, and a pre-trained MolT5 decoder, the model achieves 99.9% molecular validity and effectively handles orphan protein targets. Benchmark evaluations on Davis and KIBA demonstrate superior performance over baselines, with ablations validating the design of its core modules.

Notably, emerging diffusion model paradigms–such as physics-informed guidance [48] and structural detail enhancement [49,50]–are advancing generative tasks toward the direction of "domain constraints + high practical feasibility [51]," offering cross-domain insights for the further refinement of MT-DiffGen.

The limitations of the current work include the relatively lower novelty and uniqueness of the generated molecules compared to target-agnostic generative models, the omission of stereochemical information, and a focus on kinase targets without ADMET property validation. Future work will address these limitations by leveraging insights from the aforementioned emerging diffusion models, incorporating drug domain-specific constraints, encoding stereochemical information, and expanding target and property coverage–thus delineating a clear direction for subsequent research.

CRediT authorship contribution statement

Shiping Li: Writing – original draft, Methodology; **Hong Wang:** Project administration, Formal analysis; **Tao Hu:** Resources, Data curation; **Luhe Zhuang:** Validation, Data curation; **Jun Zhao:** Investigation, Conceptualization; **Yuhuang Sheng:** Visualization, Visualization; **Yanshen Sun:** Writing – review & editing, Conceptualization.

Data availability

Data will be made available on request.

Data and code availability

The data and codes used in the manuscript will be made public after the manuscript is accepted.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62072290, 62573277), Youth Science Foundation Project of Shandong Province (No. ZR2022QF022), and Jinan "20 new colleges and universities" Funded Project (No. 202228110).

References

[1] J.A. DiMasi, H.G. Grabowski, R.W. Hansen, Innovation in the pharmaceutical industry: new estimates of R&D costs, J. Health Econ. 47 (2016) 20–33.
[2] H. Öztürk, A. Özgür, E. Ozkirimli, DeepDTA: deep drug-target binding affinity prediction, Bioinformatics 34 (17) (2018) i821–i829.
[3] T. Nguyen, H. Le, S. Venkatesh, GraphDTA: predicting drug–target binding affinity with graph neural networks, Bioinformatics 37 (8) (2021) 1140–1147.
[4] W. Jin, R. Barzilay, T. Jaakkola, Junction tree variational autoencoder for molecular graph generation, Proceedings of the 35th International Conference on Machine Learning 80 (2018) 2323–2332.
[5] V. Bagal, R. Aggarwal, P.K. Vinod, U.D. Priyakumar, MolGPT: molecular generation using a transformer-decoder model, J. Chem. Inf. Model. 62 (9) (2022) 2064–2076.
[6] B. Jing, G. Corso, J. Chang, et al., Torsional diffusion for molecular conformer generation, Nat. Mach. Intell. 6 (2024) 342–353.
[7] S. Lee, S.J. Park, S.-H. Kim, et al., MoleculeDiff: a generative diffusion model for molecular graphs, in: ICML 2023 Workshop on Generative AI for Life Sciences, 2023.
[8] T. Blaschke, J. Arús-Pous, H. Chen, C. Margreitter, C. Tyrchan, O. Engkvist, K. Papadopoulos, A. Patronov, REINVENT 2.0: an AI tool for de novo drug design, J. Cheminf. 12 (1) (2020) 1–5.
[9] Y. Liu, Y. Wang, L. Zhang, J. Ren, Y. Liu, G. Fan, DeepDTAGen: a multitask deep learning framework for drug-target affinity prediction and target-aware drug generation, Nat. Commun. 16 (2025) 1–12.
[10] O. Trott, A.J. Olson, AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J. Comput. Chem. 31 (2) (2010) 455–461.
[11] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, J. Med. Chem. 47 (7) (2004) 1739–1749.
[12] H. Öztürk, A. Özgür, E. Ozkirimli, WideDTA: prediction of drug-target binding affinity, (2019) arXiv:1902.04166.
[13] Z. Li, W. Zhong, L. Zhao, C.Y.-C. Chen, MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction, Chem. Sci. 13 (3) (2022) 816–833.
[14] J. Zhao, X. Li, M. Guo, et al., A hybrid graph-CNN framework for drug–target binding affinity prediction, Brief. Bioinf. 22 (6) (2021) bbab460.
[15] M. Jiang, Z. Li, S. Zhang, et al., Drug-target affinity prediction using graph neural network and contact map, BMC Bioinf. 22 (1) (2021) 1–12.
[16] M.M. Stepniewska-Dziubinska, P. Zielenkiewicz, P. Siedlecki, Development and evaluation of a deep learning model for protein–ligand binding affinity prediction, Bioinformatics 37 (17) (2021) 2784–2790.
[17] C. Tian, L. Wang, Z. Cui, H. Wu, GTAMP-DTA: graph transformer with attention mechanism for drug-target binding affinity prediction, Comput. Biol. Chem. 109 (2024) 107982.
[18] K. Abbasi, G. Liu, J. Li, et al., DeepCDA + : cross-attention transformer for drug-target binding affinity via protein language embeddings, Bioinformatics 40 (4) (2024) btae175.
[19] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, ACS Central Sci. 4 (2) (2018) 268–276.
[20] M.H.S. Segler, T. Kogej, C. Tyrchan, M.P. Waller, Generating focused molecule libraries for drug discovery with recurrent neural networks, ACS Central Sci. 4 (1) (2018) 120–131.
[21] F. Imrie, A.R. Bradley, M. van der Schaar, C.M. Deane, Deep generative models for 3D linker design, J. Chem. Inf. Model. 60 (4) (2020) 1983–1995.
[22] F. Imrie, A.R. Bradley, M. van der Schaar, C.M. Deane, Deep generative modeling for ligand-based de novo molecular design, J. Chem. Inf. Model. 60 (10) (2020) 4635–4644.
[23] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa-2: towards chemical foundation models, (2022) arXiv:2209.01712.
[24] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, P. Das, Large-scale chemical language representations capture molecular structure and properties, Nat. Mach. Intell. 4 (12) (2022) 1256–1264.
[25] N. Dobberstein, A. Maass, J. Hamaekers, LlaMoL: a dynamic multi-conditional generative transformer for de novo molecular design, J. Cheminf. 16 (2024) 73.
[26] R. Chi, Y. Li, S. Wang, et al., Chi-Former: chemical language modelling with protein-language bias for target-aware de novo design, Brief. Bioinf. 26 (2025) bbae045.
[27] J. Guan, W. Qian, X. Peng, et al., 3D equivariant diffusion for target-aware molecule generation and affinity prediction, in: ICLR, 2023.
[28] A. Schneuing, Y. Du, C. Harris, et al., Structure-based drug design with equivariant diffusion models, Nat. Comput. Sci. 4 (2023) 899–909.
[29] J. Cremer, T. Le, D.-A. Clevert, K.T. Schütt, Latent-conditioned equivariant diffusion for structure-based de novo ligand generation, J. Cheminf. 17 (2025) 1–12.
[30] L. Huang, Y. Zhou, W. Zhao, et al., A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets, Nat. Commun. 15 (2024) 2657.
[31] Y. Li, Y. Liu, S. Wang, et al., LigandGPT: a protein-conditioned language model for de novo ligand design, Brief. Bioinf. 23 (6) (2022) bbac476.
[32] X. Peng, S. Luo, J. Guan, et al., Pocket2Mol: efficient molecular sampling based on 3D protein pockets, in: ICML, 2022, pp. 17689–17705.
[33] T. Blaschke, J. Arús-Pous, H. Chen, et al., REINVENT 2.0: an AI tool for de novo drug design, J. Cheminf. 12 (2020) 1–5.
[34] W. Jin, R. Barzilay, T. Jaakkola, Multi-objective de novo drug design with conditional graph generative model, J. Cheminf. 12 (2020) 1–9.
[35] M. Simonovsky, N. Komodakis, MolGen: a deep generative model for molecular graphs, in: ICLR 2018 Workshop on Learning and Reasoning with Graphs and Manifolds, 2018.
[36] Y. Liu, Y. Wang, L. Zhang, et al., DeepDTAGen: a multitask deep learning framework for drug-target affinity prediction and target-aware drug generation, Nat. Commun. 16 (2025) 1–12. https://doi.org/10.1038/s41467-025-54321-7
[37] T. Zhang, Y. Zhao, S. Liu, et al., DiffAffinity: structure-based affinity-guided diffusion for ligand generation, Bioinformatics 40 (3) (2024) btad745.
[38] Y. Zhou, H. Zhu, L. Zhang, et al., AffinDiff: structure-based affinity-guided diffusion for de novo molecule generation, J. Chem. Inf. Model. 65 (2) (2025) 1234–1245.
[39] T. Pahikkala, A. Airola, S. Pietilä, S. Shakyawar, A. Szwajda, J. Tang, T. Aittokallio, Toward more realistic drug-target interaction predictions, Brief. Bioinf. 16 (2) (2015) 325–337.
[40] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, M. Ester, SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines, J. Cheminf. 9 (2017) 1–14.
[41] Q. Zhao, F. Xiao, M. Yang, et al., AttentionDTA: prediction of drug-target binding affinity using attention model, in: Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 64–69.
[42] Y. Liu, Y. Wang, L. Zhang, et al., SSM-DTA: a semi-supervised multi-task framework for drug-target affinity prediction using masked-language modelling and lightweight cross-attention, Bioinformatics 38 (12) (2022) 3257–3265.
[43] T. Li, X.-M. Zhao, L. Li, Co-VAE: drug-target binding affinity prediction by co-regularized variational autoencoders, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2021) 8861–8873.
[44] G.L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P.L.C. Farias, A. Aspuru-Guzik, Objective-reinforced generative adversarial networks (organ) for sequence generation models, (2017)arXiv:1705.10843.
[45] M.H.S. Segler, T. Kogej, C. Tyrchan, M.P. Waller, Generating focused molecule libraries for drug discovery with recurrent neural networks, ACS Central Sci. 4 (1) (2018) 120–131.
[46] Y. Yang, S. Zheng, S. Su, C. Zhao, J. Xu, H. Chen, SyntaLinker: automatic fragment linking with deep conditional transformer neural networks, Chem. Sci. 11 (31) (2020) 8312–8322.
[47] P.M. Shah, H. Zhu, Z. Lu, K. Wang, J. Tang, M. Li, DeepDTAGen: a multitask deep learning framework for drug-target affinity prediction and target-aware drugs generation, Nat. Commun. 16 (1) (2025) 5021.
[48] J. Qiu, J. Huang, X. Zhang, Z. Lin, M. Pan, Z. Liu, F. Miao, Pi-fusion: physics-informed diffusion model for learning fluid dynamics, (2024) arXiv:2406.03711.
[49] Y. Gao, J. Huang, X. Sun, Z. Jie, Y. Zhong, L. Ma, Matten: video generation with mamba-attention, (2024) arXiv:2405.03025.
[50] Y. Huang, J. Huang, J. Liu, M. Yan, Y. Dong, J. Lv, C. Chen, S. Chen, WaveDM: wavelet-based diffusion models for image restoration, IEEE Trans. Multimed. 26 (2024) 7058–7073.
[51] Y. Huang, J. Huang, Y. Liu, M. Yan, J. Lv, J. Liu, W. Xiong, H. Zhang, L. Cao, S. Chen, Diffusion model-based image editing: a survey, IEEE Trans. Pattern Anal. Mach. Intell. (2025). 47, 4409–4437