BLIP2MOL: PHENOTYPE–TARGET FUSION FOR MOLECULAR GENERATION VIA MULTIMODAL LATENT DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

De novo molecular design is commonly driven either by protein structure, which favors target affinity, or by phenotypic readouts, which better reflect system-level biological activity. However, existing methods rarely integrate both modalities in a unified generative framework, limiting their ability to produce molecules that are simultaneously bioactive and target-compatible. We propose Blip2Mol, a multimodal molecular generation framework that bridges chemical-induced transcriptomic responses, protein binding pockets, and molecular structures through a shared latent space. Blip2Mol first pretrains a molecular variational autoencoder to obtain a continuous chemical representation. It then aligns transcriptome and pocket embeddings to the molecular latent space via BLIP-style dual cross-modal alignment with contrastive and matching objectives, enabling semantic consistency across phenotype, target, and molecule modalities. Next, we develop a conditional latent diffusion model with a learnable gated fusion mechanism to generate molecules jointly conditioned on transcriptomic response and pocket context. Experiments on 10 cancer-related targets show that Blip2Mol consistently outperforms representative phenotype-guided and docking-based baselines, achieving stronger docking affinity while maintaining superior drug-likeness, synthetic accessibility, validity, uniqueness, and novelty. Ablation studies further confirm the effectiveness of multimodal conditioning and the critical role of the diffusion module. These results demonstrate the promise of our unified model for phenotype-aware and target-aware drug design.

1 INTRODUCTION

Historically, Phenotypic Drug Discovery (PDD) dominated the pharmaceutical field (Al-Ali, 2016; Moffat et al., 2017). By directly evaluating whether candidate compounds induce desired physiological effects at the cellular or organism level, PDD has yielded numerous milestone "first-in-class" drugs (Moffat et al., 2017; Swinney & Anthony, 2011; Vincent et al., 2022). However, propelled by rapid advancements in structural biology, genomics, and computational chemistry, Target-Based Drug Design (TBDD) has gradually superseded PDD as the prevailing paradigm. By leveraging the resolved 3D structures of pathogenic proteins (Burley, 2021), TBDD enables the precise design of ligands that strongly interact with specific binding pockets (Vost et al., 2025), thereby significantly enhancing the specificity of drug discovery (Yan et al., 2025).

Despite the tremendous success of TBDD, relying solely on target-level structural affinity fails to fully capture a drug's behavior within complex biological systems. As a result, many high-affinity molecules fail in clinical trials due to insufficient systemic efficacy or unexpected toxicity (Lin et al., 2019). Conversely, while traditional PDD intuitively reflects system-level efficacy, it is severely bottlenecked by low screening throughput, challenges in target deconvolution (mechanism elucidation), and prolonged development cycles (Takács et al., 2025; Calado, 2024). Fortunately, the maturation of high-throughput sequencing technologies has introduced a novel perspective: chemical-induced transcriptomic profiles (Szalai & Veres, 2023). As an RNA-level response to external perturbations, transcriptomic profiles systematically capture the global impact of compounds on cellular states. Thus, large-scale transcriptomic data (e.g., the LINCS L1000 dataset) can serve as an effective and scalable "data proxy" for phenotypic screening, paving the way for computationally

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

driven, high-throughput phenotypic drug design (Szalai & Veres, 2023; DeMeo et al., 2025). Given the central role of transcriptional profiles in this study, we consistently use the term "phenotypic profile" to refer to transcriptomic profiles.

Current computational solutions for de novo drug design have largely evolved along two parallel tracks. On one hand, target-based generative models (e.g., Pocket2Mol (Peng et al., 2022), SampleDock (Xu et al., 2021)) leverage 3D structural information to design molecules highly complementary to protein pockets, achieving remarkable docking scores but often lacking explicit guarantees of systemic biological efficacy (Guo & Liu, 2025). On the other hand, an emerging class of phenotype-driven generative methods (e.g., GxVAEs (Li & Yamanishi, 2024), SmilesGEN (Liu et al., 2025)) decodes molecular structures directly from compound-induced transcriptomic profiles. While these methods ensure biological relevance, the absence of explicit 3D spatial constraints often results in suboptimal binding affinities to specific target proteins (Liu et al., 2025). Despite significant advancements in both paradigms (Calado, 2024), there remains a critical dearth of computational methods capable of seamlessly integrating system-level phenotypic responses with protein-level target structures. Bridging this cross-modal gap to generate drug-like molecules that possess both ideal biological activity (phenotypic response) and high target affinity (structural complementarity) persists as a paramount challenge in generative AI for drug discovery.

To bridge the gap between phenotype-based and target-based drug design, we propose Blip2Mol, a multimodal molecular generation framework motivated by the idea that effective molecular design should be guided not only by target structure but also by the biological functions reflected in phenotypic readouts. To this end, Blip2Mol jointly aligns representations of transcriptomic signatures, protein pockets, and molecules in a unified semantic space, capturing cross-modal relationships among phenotypes, targets, and candidate molecules. It then employs a conditional diffusion model with gated multimodal fusion to generate molecules conditioned on both phenotype and target information. Extensive experiments demonstrate that Blip2Mol enables the discovery of candidate molecules with desired phenotypic effects and potential target-binding capability.

The main contributions of this study are summarized as follows:

- A cross-modal alignment mechanism built upon the BLIP paradigm with contrastive and matching objectives to enforce semantic consistency among phenotypic, structural, and molecular representations.
- A multimodal conditional diffusion generator that leverages aligned latent representations to improve the biological activity and target-binding affinity of generated molecules.
- Experiments on 10 cancer-related targets demonstrate that Blip2Mol consistently outperforms existing unimodal baselines in binding affinity, drug-likeness, and synthetic accessibility, highlighting its potential as an effective computational framework for multimodal drug design.

2 METHOD

2.1 OVERVIEW OF BLIP2MOL

Blip2Mol is a molecular design framework that integrates multimodal representation alignment with conditional diffusion generation. The framework consists of three sequential modules. First, we pretrain a GRU-based molecular variational autoencoder to map SMILES strings into a chemical latent space (Gómez-Bombarelli et al., 2018; Blaschke et al., 2018; Sicho et al., 2023). Next, following the BLIP paradigm (Li et al., 2022), we align both phenotypic profiles and protein binding pockets with the latent space through contrastive learning and cross-modal matching, thereby enforcing cross-modal semantic consistency. Finally, we train a diffusion-based generator conditioned on the aligned latent representations to generate molecules under multimodal guidance (Ragoza et al., 2022).

More specifically, Blip2Mol uses a fused latent representation that jointly encodes phenotypic responses and protein pocket information as the conditional signal. Starting from Gaussian noise, the model iteratively denoises latent variables and then reconstructs molecular structures using the pretrained molecular decoder. In this way, Blip2Mol can generate drug-like molecules that are more likely to exhibit both the desired biological activity and strong target affinity (Vincent et al., 2022).

2.2 Molecular Variational Autoencoder

We represent each molecule as a SMILES sequence $\mathbf{x} = (x_1, x_2, \ldots, x_L)$, where L is the sequence length and each token x_i is drawn from the SMILES vocabulary. To obtain a continuous molecular latent space, we construct a GRU-based variational autoencoder (MolVAE) that encodes SMILES strings into latent Gaussian variables (Gómez-Bombarelli et al., 2018). Given a molecule $\mathbf{x}$, the encoder outputs an approximate posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$ that is regularized toward the standard normal prior, while the decoder $p_\varphi(\mathbf{x} \mid \mathbf{z})$ autoregressively reconstructs the original SMILES sequence conditioned on the latent variable $\mathbf{z}$. Using the reparameterization trick, the model is pretrained by maximizing the evidence lower bound:

$$\mathcal{J}_{\text{mol}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\varphi(\mathbf{x}|\mathbf{z})\right] - D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}))\,, \tag{1}$$

where ϕ and φ are the parameters of the encoder and decoder, respectively, and the prior is defined as $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

We pretrain MolVAE on a million-scale dataset of SMILES sequences until convergence. After pretraining, the molecular encoder maps arbitrary molecules into a chemical structural space, while the decoder reconstructs molecular structures with high fidelity (Kusner et al., 2017). This molecular representation serves as the anchor for subsequent cross-modal alignment and conditional generation.

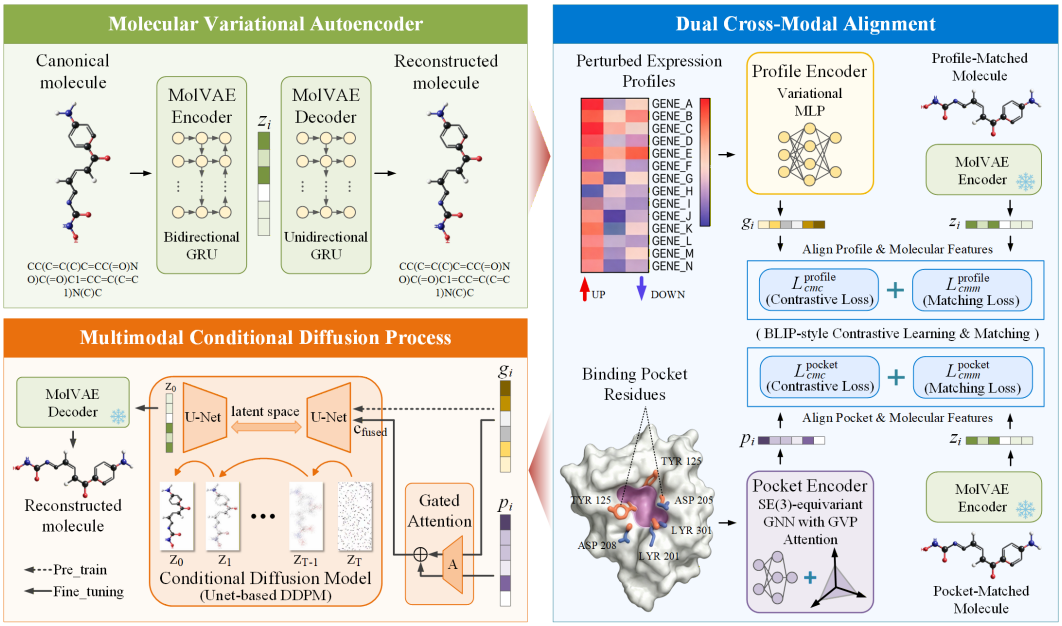


Figure 1: Overview of Blip2Mol framework. It employs a molecular variational autoencoder to learn drug-molecule embeddings, aligns phenotypic profiles and target-pocket representations with matched molecular embeddings using BLIP-style objectives, and fuses multimodal conditions to guide diffusion-based molecule generation.

2.3 Dual Cross-Modal Alignment

To establish a unified semantic representation across different modalities, we adopt a BLIP-style cross-modal alignment framework. With the pretrained molecular encoder frozen (Wang et al., 2025), we map phenotypic profiles and protein pocket structures into the molecular latent space, and train the profile encoder and pocket encoder to produce semantically consistent representations in the shared latent space.

The profile encoder is implemented as a multilayer feed-forward network that takes the perturbed expression profiles as input and maps them into a low-dimensional representation $\mathbf{g}$. For a mini-batch of B profile-molecule pairs, let $\{\mathbf{g}_i\}_{i=1}^B$ and $\{\mathbf{z}_i\}_{i=1}^B$ denote the transcriptomic and molecular latent

representations, respectively. Define the similarity matrix $S \in \mathbb{R}^{B\times B}$ with entries $s_{ij} = \frac{g_i^\top z_j}{\tau \|g_i\| \|z_j\|}$, where τ is a temperature parameter. We optimize a cross-modal contrastive (CMC) objective:

$$\mathcal{L}_{cmc}^{\text{profile}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ij})} + \log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ji})}\right]. \tag{2}$$

To further strengthen alignment, we introduce a cross-modal matching loss (CMM) with hard negative mining. Given a phenotypic profile $\mathbf{g}_i$, the molecular latent representation $\mathbf{z}_i$ of the corresponding molecule is used as the positive sample. In contrast, the latent representation of the most similar mismatched molecule within the mini-batch is used as the hard negative sample $\bar{\mathbf{z}}_i$. We define the positive matching score as $a_i = \mathbf{g}_i^\top \mathbf{z}_i$ and the negative matching score as $b_i = \mathbf{g}_i^\top \bar{\mathbf{z}}_i$. We then concatenate these scores into a vector $\hat{\mathbf{y}} = [a_1, \ldots, a_B, b_1, \ldots, b_B]$, with the corresponding labels $\mathbf{y} = [1, \ldots, 1, 0, \ldots, 0]$. The CMM loss is defined as a binary cross-entropy loss:

$$\mathcal{L}_{cmm}^{\text{profile}} = -\frac{1}{2B}\sum_{k=1}^{2B}[y_k \log\sigma(\hat{y}_k) + (1-y_k)\log(1-\sigma(\hat{y}_k))] \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The overall profile-molecule objective alignment is $\mathcal{L}^{\text{profile}} = \mathcal{L}_{cmc}^{\text{profile}} + \mathcal{L}_{cmm}^{\text{profile}}$, which encourages the profile encoder to map phenotypic responses into the molecular latent space.

Similarly, to align protein pockets with molecules, we construct a pocket encoder based on an SE(3)-equivariant GNN. For each residue in the pocket, we use scalar and vector features that capture residue identity, local backbone conformation, ligand context, and backbone/side-chain geometry. These features are processed by a geometric vector perceptron (GVP (Jing et al., 2020)) embedding layer, multiple equivariant Transformer message-passing layers (Jing et al., 2021), and an attention pooling module, yielding a fixed-dimensional pocket representation $\mathbf{p}$. Given the molecular representation $\mathbf{z}$, we replace the phenotypic feature $\mathbf{g}$ with the pocket representation $\mathbf{p}$ and train the pocket encoder using similar alignment objectives $\mathcal{L}_{\text{cmc}}^{\text{pocket}}$ and $\mathcal{L}_{\text{cmm}}^{\text{pocket}}$. This results in a shared latent space in which phenotypic profiles, protein pockets, and molecules are semantically comparable.

2.4 Conditional Diffusion Process

To use phenotypic response and protein pocket information to guide molecular generation, we build a multi-conditional diffusion model in the molecular latent space (Jang & Ye, 2025). The diffusion model is trained in two stages. In the first stage, we condition only on phenotypic profiles to train a base diffusion model, enabling it to capture the semantic association between phenotypic responses and molecular structures. In the second stage, we integrate protein pocket features and fine-tune the model with a learnable gated fusion mechanism, such that the denoising process is jointly guided by biological function and structural compatibility.

Let $\mathbf{z}_0$ denote the latent representation of a molecule of interest. The forward diffusion process is defined as a Markov chain of length T that gradually perturbs the latent variable with Gaussian noise:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{4}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, and β_s is a predefined noise variance schedule.

Given a profile $\mathbf{g}$ induced by this molecule, we use a 1D U-Net in the latent space as the noise prediction network $\epsilon_\theta(\cdot)$. The training objective is the standard denoising mean squared error (MSE):

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon,\ \mathbf{g},\ t\right)\right\|_2^2\right] \tag{5}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and t is uniformly sampled from $\{1, \ldots, T\}$. This stage enables the model to learn the mapping between phenotypic responses and molecular structures.

To further improve chemical compatibility with the target protein pocket, we also consider the pocket latent representation $\mathbf{p}$, and adaptively fuse them using a learnable gating mechanism:

$$\lambda = \sigma\left(\mathbf{W}_2\,\mathrm{ReLU}\big(\mathbf{W}_1[\mathbf{g};\mathbf{p}] + \mathbf{b}_1\big) + \mathbf{b}_2\right) \tag{6}$$

where $[\cdot;\cdot]$ denotes vector concatenation, $\sigma(\cdot)$ is the sigmoid function, and $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters. As a result, the two modalities are fused as

$$\mathbf{c}_{\text{fused}} = \lambda \odot \mathbf{g} + (1 - \lambda) \odot \mathbf{p} \tag{7}$$

in which $\odot$ denotes element-wise multiplication. Accordingly, we replace the single-modality condition $\mathbf{g}$ with the fused condition $\mathbf{c}_{\text{fused}}$ in Equation (5).

During inference, we start from Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoise it to obtain the molecular latent variable $\mathbf{z}_0$, which is fed into the pretrained molecular decoder to reconstruct the SMILES sequence, thereby yielding molecules that are expected to satisfy both biological activity and target-binding affinity.

In this study, the molecular encoder uses a bidirectional GRU as backbone, whereas the molecular decoder uses a unidirectional GRU for autoregressive generation. The encoder and decoder each contain three layers with 256 hidden units and a dropout rate of 0.1. The profile encoder is a three-layer feed-forward network with hidden sizes of 512, 256, and 128, using ReLU activations and a dropout rate of 0.2. The pocket encoder is based on an SE(3)-equivariant graph neural network, where the GVP scalar hidden dimension is 256, the number of vector channels is 64, the number of equivariant Transformer layers is 6, the k-NN graph parameter is set to $k = 48$, and the output pocket latent dimension is 64. The diffusion model adopts a 1D U-Net backbone with input dimension 64 and hidden dimension 128, together with a noise schedule and a total of T = 1000 diffusion steps.

3 Experiments

3.1 Data Resources

We extracted 1,048,575 SMILES strings from the ZINC database (Irwin & Shoichet, 2005) to pretrain the molecular variational autoencoder. During modality alignment, we used 13,755 Level 5 perturbed expression profiles from the MCF7 cell line in LINCS L1000 (Subramanian et al., 2017) to align phenotypic and molecular representations, and 396,383 human protein-ligand complexes from the CrossDocked2020 dataset (Francoeur et al., 2020) to align pocket and molecular representations. The diffusion model was first pretrained using MCF7 perturbed expression profiles as conditioning input, and was then fine-tuned on multimodal samples constructed via InChIKey matching across molecules, phenotypic profiles, and protein pocket information.

For evaluation, we used the expression profiles induced by genetic perturbations of 10 cancer-related target genes. Eight targets—AKT1, AKT2, AURKB, CTSK, EGFR, HDAC1, MTOR, and PIK3CA—were knockdown, while SMAD3 and TP53 were overexpressed. The corresponding profiles were obtained from LINCS L1000, and these protein structures were collected from the RCSB PDB (Burley et al., 2025). These two modalities were jointly used as conditioning inputs during testing.

3.2 Performance Metrics

To comprehensively evaluate the generated molecules, we considered multiple metrics, including chemical validity, uniqueness, novelty, drug-likeness (QED), synthetic accessibility (SA), and target-binding affinity. Validity is defined as the proportion of generated molecules that can be successfully parsed as chemically valid structures. Uniqueness measures the proportion of distinct structures among valid molecules. Novelty evaluates the model's generalizability by determining whether a generated molecule is absent from the training set. Drug-likeness is quantified by the QED score, where higher values indicate better drug-like properties. Synthetic accessibility is quantified by the SA score, where lower values indicate easier synthesis. Target affinity is estimated using LeDock docking scores, with lower scores indicating stronger predicted binding affinity.

3.3 Comparison to phenotype-guided methods

We first compared Blip2Mol with three representative phenotype-driven molecular generation methods, including SmilesGEN (Liu et al., 2025), GxVAEs (Li & Yamanishi, 2024), and TRIOMPHE (Kaitoh & Yamanishi, 2021). For objective evaluation, we ran each method independently to generate 100 molecules for each target protein and then computed the performance metrics. Blip2Mol

adopts dual-modal input so that phenotypic response and protein pocket structures are integrated to generate molecules.

As shown in Figures 2 and A1, Blip2Mol achieved the best or near-best docking scores on the vast majority of targets, indicating consistently stronger binding affinity than SmilesGEN and GxVAEs overall, while the latter two exhibited more variable performance across different proteins. For drug-likeness (QED), the molecules generated by Blip2Mol consistently attained the highest scores across all targets, substantially surpassing the other baseline methods, indicating superior drug-like properties. For SA, Blip2Mol also demonstrated competitive performance, with SA scores generally lower than those of GxVAEs and TRIOMPHE and largely comparable to those of SmilesGEN. Taken together, these results show that Blip2Mol substantially improves target-binding affinity while maintaining strong drug-likeness and synthetic accessibility.

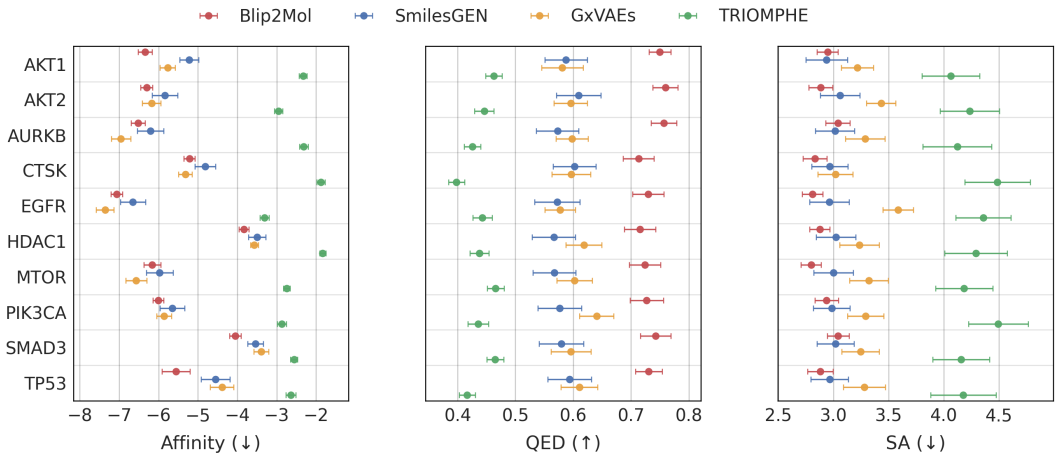


Figure 2: Phenotype-based model comparison on Affinity, QED, and SA metrics

Moreover, Blip2Mol achieves 100% uniqueness and 100% novelty across all 10 targets, while maintaining validity consistently above 90%. In contrast, SmilesGEN, GxVAEs, and TRIOMPHE exhibit varying degrees of degradation in validity, uniqueness, and novelty (Figure 3). For example, TRIOMPHE attains validity below 30% and novelty below 20% on most targets. These results further demonstrate that the proposed multimodal conditional fusion and diffusion-based generation framework enables Blip2Mol to consistently generate chemically valid, structurally unique, and training-set-novel molecules, highlighting its superior generation fidelity and generalization capability.

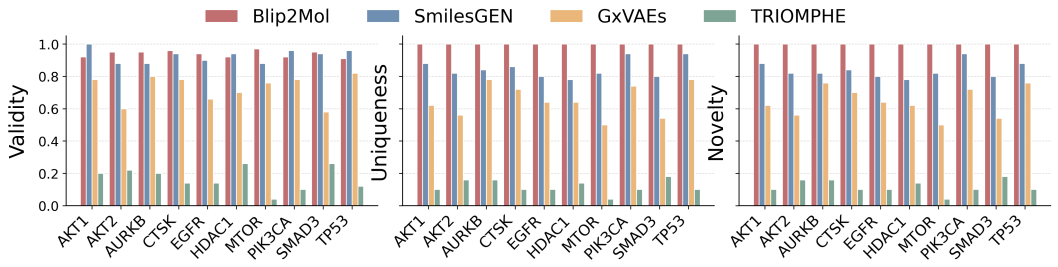


Figure 3: Comparison to phenotype-based methods in terms of validity, uniqueness, and novelty

3.4 Comparison to Docking-Based Method

We further compared Blip2Mol with four docking-based molecular generation methods, including Pocket2Mol (Peng et al., 2022), SBDD-3D (Luo et al., 2021), SBMolGen (Ma et al., 2021), and DiffSBDD (Schneuing et al., 2024). Each method was independently run to generate 100 molecules per target protein for performance metric evaluation.

As shown in Figure 4 and A2, Blip2Mol achieved the best or near-best docking scores on most targets, showing particularly strong mean binding affinity on EGFR (-7.05), AKT1 (-6.34), and AURKB (-6.52). Regarding drug-likeness (QED), Blip2Mol consistently outperformed Pocket2Mol, SBDD-3D, and DiffSBDD, with performance on par with SBMolGen. In terms of synthetic accessibility (SA), Blip2Mol also demonstrated clear overall advantages over Pocket2Mol, SBDD-3D, and DiffSBDD, while achieving results close to those of SBMolGen, despite the latter showing slightly better performance on a few specific targets. These results indicate that, by integrating phenotypic and structural multimodal conditioning, Blip2Mol is able to generate molecules with strong binding affinity, improved drug-likeness, and favorable synthetic feasibility, leading to superior overall performance compared with existing docking-based molecular generation methods.

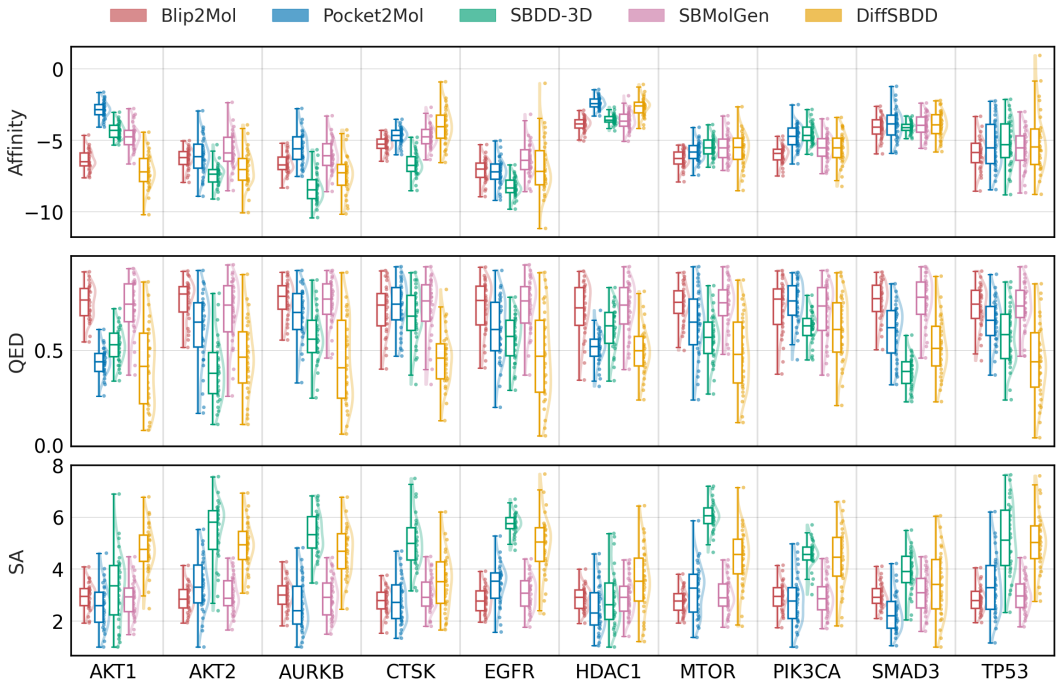


Figure 4: Comparison to target-based methods in terms of affinity, QED, and SA metrics

3.5 Ablation Experiments

To evaluate the contribution of the diffusion module to molecular generation, we compared the full Blip2Mol model with its ablated variant, Blip2Mol-NoDiff, in which the diffusion module was removed. The experimental results show that Blip2Mol consistently achieved substantially higher validity across all 10 targets (Table A1). For instance, validity increased from 81% to 92% for AKT1 and from 91% to 97% for MTOR. In terms of uniqueness, Blip2Mol consistently reached 100%, whereas Blip2Mol-NoDiff exhibited overall inferior performance. These results indicate that the stochastic denoising process introduced by the diffusion module not only improves the chemical validity of the generated molecules but also enhances their structural diversity and uniqueness, making it a key component for high-quality molecular generation.

To evaluate the complementary effects of phenotypic and pocket information, we performed an ablation study by comparing the full Blip2Mol model with two single-modality variants: Blip2Mol-onlyProfile and Blip2Mol-onlyPocket. As shown in Table 1, the pocket-only variant achieved superior affinity on several proteins. For example, it obtained an affinity of -6.67 on AURKB, slightly better than that of the full model (-6.52). This suggests that explicit protein pocket information can enhance the target-binding affinity of generated molecules. However, its overall performance on QED and SA was slightly worse, implying that pocket-only conditioning may improve binding at the expense of drug-likeness and synthetic accessibility. In contrast, the profile-only variant showed advantages in QED and SA. For example, it achieved a QED of 0.76 on HDAC1, indicating that phenotypic profiles

Table 1: Comparison of Blip2Mol to ablated variants on target-specific molecule generation

Target	Blip2Mol			Blip2Mol-onlyProfile			Blip2Mol-onlyPocket		
Protein	Affinity	QED	SA	Affinity	QED	SA	Affinity	QED	SA
AKT1	-6.36±0.17	**0.74**±0.02	2.92±0.09	-6.14±0.29	**0.74**±0.03	2.91±0.10	**-6.41**±0.26	0.72±0.02	**2.85**±0.10
AKT2	**-6.28**±0.16	0.73±0.02	2.84±0.10	-6.14±0.16	**0.74**±0.03	**2.83**±0.09	-6.26±0.19	0.72±0.02	3.00±0.09
AURKB	-6.57±0.21	0.73±0.03	3.00±0.10	-6.40±0.31	0.73±0.02	**2.91**±0.10	**-6.67**±0.14	**0.74**±0.02	3.01±0.11
CTSK	-5.25±0.15	0.72±0.03	2.93±0.11	-5.14±0.15	**0.74**±0.02	**2.89**±0.09	**-5.27**±0.13	0.70±0.03	2.93±0.10
EGFR	-7.01±0.23	0.72±0.02	**2.87**±0.09	**-7.05**±0.16	**0.73**±0.03	2.88±0.09	-7.01±0.28	0.69±0.03	2.96±0.12
HDAC1	-3.81±0.11	0.71±0.02	2.91±0.11	-3.75±0.12	**0.76**±0.02	**2.88**±0.11	**-3.84**±0.09	0.71±0.02	2.93±0.10
MTOR	**-6.17**±0.14	**0.73**±0.02	**2.86**±0.10	-6.08±0.23	**0.73**±0.02	2.90±0.10	**-6.17**±0.13	0.72±0.02	2.91±0.11
PIK3CA	**-5.99**±0.19	0.72±0.03	**2.89**±0.09	-5.84±0.19	**0.75**±0.02	2.96±0.11	-5.92±0.20	0.71±0.02	2.95±0.10
SMAD3	-4.09±0.16	0.74±0.03	2.97±0.11	-3.98±0.22	**0.75**±0.02	**2.93**±0.10	**-4.14**±0.19	0.74±0.02	2.97±0.10
TP53	-5.64±0.34	0.72±0.02	2.94±0.11	**-5.75**±0.27	**0.74**±0.02	**2.89**±0.10	-5.68±0.37	0.72±0.03	2.91±0.10

contribute to the generation of molecules with more favorable medicinal chemistry properties and more reasonable chemical structures. However, without explicit target constraints, its affinity was generally weaker than that of the full model on most targets. By effectively integrating both modalities, the full Blip2Mol model achieved a better trade-off among affinity, QED, and SA across multiple targets. For instance, on EGFR, Blip2Mol achieved the best affinity (-7.05) while still maintaining favorable molecular properties. Overall, the full model delivered the most balanced performance across all three metrics, demonstrating a significant synergistic effect between phenotypic and pocket-based conditioning. These findings confirm the advantage of multimodal modeling for generating high-quality candidate molecules.

3.6 Therapeutic Molecules Generation

To assess Blip2Mol's ability to generate therapeutic molecules, we generated 100 novel molecules for each of eight targets: AKT1, AKT2, AURKB, CTSK, EGFR, HDAC1, PIK3CA, and TP53. Seven DrugBank-approved drugs were used as references: Capivasertib, Barasertib, Odanacatib, Erlotinib, Vorinostat, Alpelisib, and Eprenetapopt. MTOR and SMAD3 were excluded due to the lack of approved reference drugs.

As shown in Figure 5, Blip2Mol generated molecules with better docking scores than the corresponding approved drugs for AKT1, PIK3CA, and TP53, achieving -9.03, -7.48, and -9.23 kcal/mol compared with -8.28, -7.10, and -3.54 kcal/mol, respectively. Moreover, the generated molecules also showed favorable molecular properties. For example, the SA scores on AKT1 and CTSK were 2.57 and 2.91, lower than those of Capivasertib and Odanacatib, respectively. Their QED scores on AURKB and HDAC1 reached 0.73 and 0.84, substantially higher than those of Barasertib and Vorinostat. These results suggest that Blip2Mol can generate target-specific molecules with favorable binding affinity and physicochemical profiles. We also visualized the representative generated molecules for all ten targets in Figure A3.

4 Related Work

Phenotype-guided molecular generation aims to design molecules with desired bioactivities from transcriptomic profiles. Existing methods can be broadly grouped into three categories: phenotype-only generative models, phenotype-based screening models, and phenotype-target joint generation models. Representative phenotype-only methods, including GxVAEs (Li & Yamanishi, 2024), SmilesGEN (Liu et al., 2025), and GexMolGen (Cheng et al., 2024), generate molecules by aligning or conditioning molecular representations on transcriptomic profiles. TRIOMPHE (Kaitoh & Yamanishi, 2021) instead adopts a multitask framework to predict activities across multiple cell lines for candidate screening. Despite encouraging progress, phenotype-only methods often overlook target binding and may therefore yield molecules with suboptimal affinity, XMolRL (Guo & Liu, 2025) incorporates both phenotypic and target information for molecular generation through reinforcement learning.

Target-driven molecular generation leverages protein structures or binding pocket information to design molecules with high predicted binding affinity (Danel et al., 2023; Méndez-Lucio et al., 2020; Spiegel & Durrant, 2020). Representative approaches include autoregressive pocket-based generation (Pocket2Mol (Peng et al., 2022)), docking-guided reinforcement learning (SampleDock (Xu

et al., 2021)), fragment-based assembly (SBMolGen (Ma et al., 2021)), and diffusion models (DiffSBDD (Schneuing et al., 2024)). Although these methods can improve structural complementarity or docking scores, they often suffer from high computational cost, limited efficiency, or reduced structural novelty (Tang et al., 2024; Qin et al., 2026). More importantly, these target-driven methods are effective at generating molecules with strong predicted binding affinity, but they do not explicitly constrain biological activity. As a result, the generated molecules may bind well to the target protein without necessarily eliciting the desired phenotypic effects. DrugCLIP (Jia et al., 2026) performs target-aware virtual screening via CLIP-style alignment, whereas we adopt a BLIP-style alignment strategy (Li et al., 2022) to enable stronger multimodal interaction for molecular generation.

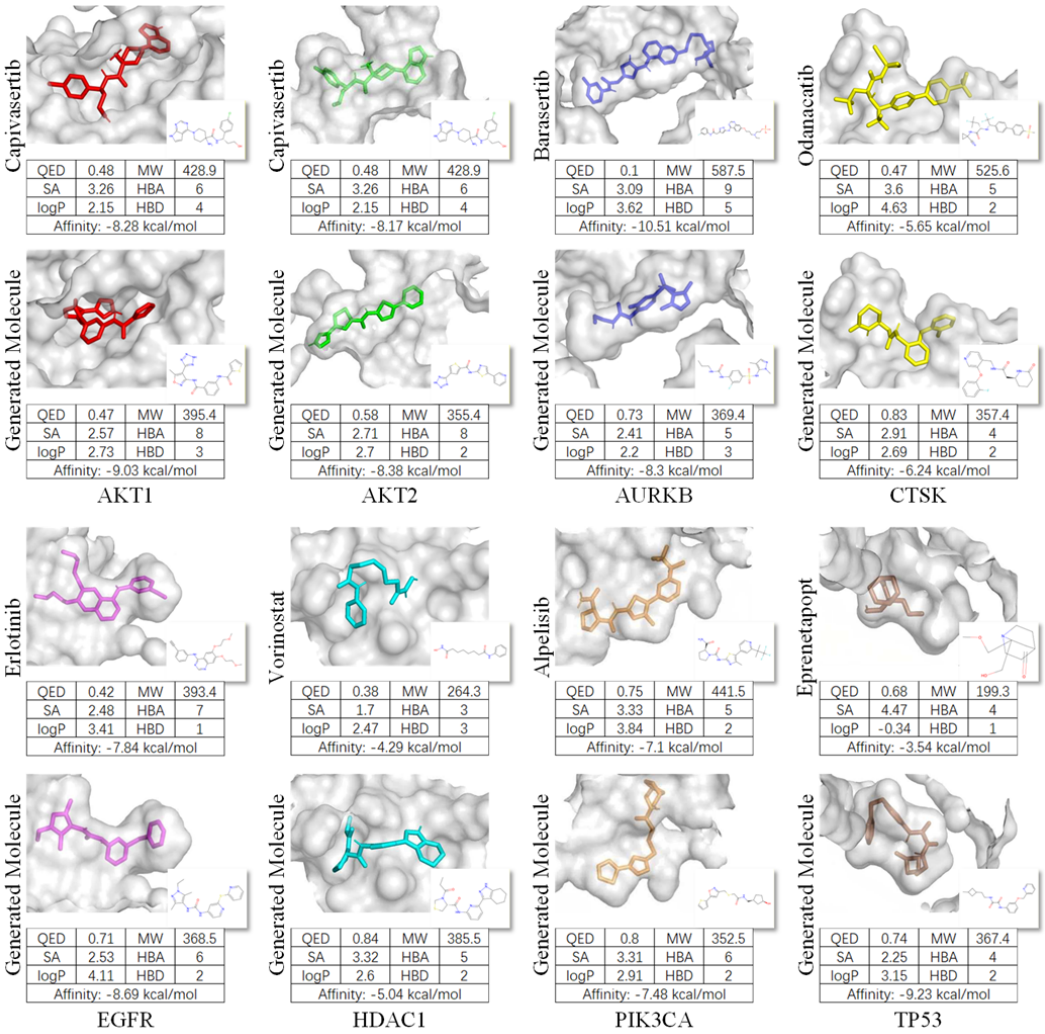


Figure 5: Docking poses and molecular properties of Blip2Mol-generated molecules compared to approved drugs

5 Conclusion

We propose Blip2Mol, a multimodal molecular generation framework that integrates phenotypic responses with protein pocket structures. By aligning both modalities with the molecular latent space and employing conditional diffusion generation, Blip2Mol generates molecules with phenotypic relevance and target compatibility. Extensive experiments show that Blip2Mol outperforms existing baselines in docking affinity, drug-likeness, and synthetic accessibility. Ablation studies further confirm the benefits of multimodal conditioning and diffusion modeling, highlighting Blip2Mol as a promising approach for phenotype- and target-aware drug design.

AI USE STATEMENT

We used generative AI tools only for writing manuscript text. We did not use generative AI tools for generating research ideas, experimental results, and scientific claims. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects or personal data, and we are not aware of any ethical concerns.

REPRODUCIBILITY STATEMENT

We provide the source code and datasets at `https://anonymous.4open.science/r/Blip2Mol-C013`, and Section 3 describes the data processing steps.

REFERENCES

Hassan Al-Ali. The evolution of drug discovery: from phenotypes to targets, and back. *MedChemComm*, 7(5):788–798, 2016.

Thomas Blaschke, Marcus Olivecrona, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Application of generative autoencoder in de novo molecular design. *Molecular informatics*, 37(1-2): 1700123, 2018.

Stephen K Burley. Impact of structural biologists and the protein data bank on small-molecule drug discovery and development. *Journal of Biological Chemistry*, 296:100559, 2021.

Stephen K Burley, Rusham Bhatt, Charmi Bhikadiya, Chunxiao Bi, Alison Biester, Pratyoy Biswas, Sebastian Bittrich, Santiago Blaumann, Ronald Brown, Henry Chao, et al. Updated resources for exploring experimentally-determined pdb structures and computed structure models at the rcsb protein data bank. *Nucleic acids research*, 53(D1):D564–D574, 2025.

Cecília RC Calado. Bridging the gap between target-based and phenotypic-based drug discovery. *Expert Opinion on Drug Discovery*, 19(7):789–798, 2024.

Jiabei Cheng, Xiaoyong Pan, Yi Fang, Kaiyuan Yang, Yiming Xue, Qingran Yan, and Ye Yuan. Gexmolgen: cross-modal generation of hit-like molecules via large language model encoding of gene expression signatures. *Briefings in bioinformatics*, 25(6):bbae525, 2024.

Tomasz Danel, Jan Łęski, Sabina Podlewska, and Igor T Podolak. Docking-based generative approaches in the search for new drug candidates. *Drug Discovery Today*, 28(2):103439, 2023.

Benjamin DeMeo, Charlotte Nesbitt, Samuel A Miller, Daniel B Burkhardt, Inna Lipchina, Doris Fu, Peter Holderrieth, David Kim, Sergey Kolchenko, Artur Szalata, et al. Active learning framework leveraging transcriptomics identifies modulators of disease phenotypes. *Science*, 390(6776): eadi8577, 2025.

Paul G Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B Iovanisci, Ian Snyder, and David R Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling*, 60(9):4200–4215, 2020.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Haotian Guo and Hui Liu. Exmolrl: Phenotype-target joint generation of de novo molecules via multi-objective reinforcement learning. *arXiv preprint arXiv:2509.21010*, 2025.

John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

Jinho Jang and Jong Chul Ye. Ldmol: Text-to-molecule diffusion model with structurally informative latent space space surpass the ar models. In *The International Conference on Machine Learning (ICML), 2025*. International Machine Learning Society (IMLS), 2025.

Yinjun Jia, Bowen Gao, Jiaxin Tan, Jiqing Zheng, Xin Hong, Wenyu Zhu, Haichuan Tan, Yuan Xiao, Liping Tan, Hongyi Cai, et al. Deep contrastive learning enables genome-wide virtual screening. *Science*, 391(6781):eads9530, 2026.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

Bowen Jing, Stephan Eismann, Pratham N Soni, and Ron O Dror. Equivariant graph neural networks for 3d macromolecular structure. *arXiv preprint arXiv:2106.03843*, 2021.

Kazuma Kaitoh and Yoshihiro Yamanishi. Triomphe: Transcriptome-based inference and generation of molecules with desired phenotypes by machine learning. *Journal of Chemical Information and Modeling*, 61(9):4303–4320, 2021.

Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International conference on machine learning*, pp. 1945–1954. PMLR, 2017.

Chen Li and Yoshihiro Yamanishi. Gxvaes: Two joint vaes generate hit molecules from gene expression profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13455–13463, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Ann Lin, Christopher J Giuliano, Ann Palladino, Kristen M John, Connor Abramowicz, Monet Lou Yuan, Erin L Sausville, Devon A Lukow, Luwei Liu, Alexander R Chait, et al. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science translational medicine*, 11(509):eaaw8412, 2019.

Hui Liu, Shiye Tian, and Xuejun Liu. Phenotypic profile-informed generation of drug-like molecules via dual-channel variational autoencoders. *arXiv preprint arXiv:2506.02051*, 2025.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021.

Biao Ma, Kei Terayama, Shigeyuki Matsumoto, Yuta Isaka, Yoko Sasakura, Hiroaki Iwata, Mitsugu Araki, and Yasushi Okuno. Structure-based de novo molecular generator combined with artificial intelligence and docking simulations. *Journal of Chemical Information and Modeling*, 61(7): 3304–3313, 2021.

Oscar Méndez-Lucio, Benoit Baillif, Djork-Arné Clevert, David Rouquié, and Joerg Wichard. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature communications*, 11(1):10, 2020.

John G Moffat, Fabien Vincent, Jonathan A Lee, Jörg Eder, and Marco Prunotto. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature reviews Drug discovery*, 16(8):531–543, 2017.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International conference on machine learning*, pp. 17644–17655. PMLR, 2022.

Rui Qin, Zijie Chen, Yurong Li, Meijing Fang, Longji Shen, Yilong Su, Odin Zhang, Qinghan Wang, Qun Su, Jike Wang, et al. Revisiting target-aware de novo molecular generation with tarpass: Between rational design and texas sharpshooter. *Advanced Science*, pp. e75411, 2026.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical science*, 13(9):2701–2713, 2022.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.

Martin Sicho, Sohvi Luukkonen, Helle W van den Maagdenberg, Linde Schoenmaker, Olivier JM Béquignon, and Gerard JP van Westen. Drugex: deep learning models and tools for exploration of drug-like chemical space. *Journal of chemical information and modeling*, 63(12):3629–3636, 2023.

Jacob O Spiegel and Jacob D Durrant. Autogrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *Journal of cheminformatics*, 12(1):25, 2020.

Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507–519, 2011.

Bence Szalai and Dániel V Veres. Application of perturbation gene expression profiles in drug discovery—from mechanism of action to quantitative modelling. *Frontiers in Systems Biology*, 3: 1126044, 2023.

Gergely Takács, György T Balogh, and Róbert Kiss. A data-driven journey using results from target-based drug discovery for target deconvolution in phenotypic screening. *RSC Medicinal Chemistry*, 16(7):3190–3196, 2025.

Xiangru Tang, Howard Dai, Elizabeth Knight, Fang Wu, Yunyang Li, Tianxiao Li, and Mark Gerstein. A survey of generative ai for de novo drug design: new frontiers in molecule and protein generation. *Briefings in Bioinformatics*, 25(4):bbae338, 2024.

Fabien Vincent, Arsenio Nueda, Jonathan Lee, Monica Schenone, Marco Prunotto, and Mark Mercola. Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nature Reviews Drug Discovery*, 21(12):899–914, 2022.

Lucy Vost, Yael Ziv, and Charlotte Mary Deane. Incorporating targeted protein structure in deep learning methods for molecule generation in computational drug design. *Chemical Science*, 2025.

Yifei Wang, Yunrui Li, Lin Liu, Pengyu Hong, and Hao Xu. Advancing drug discovery with enhanced chemical understanding via asymmetric contrastive multimodal learning. *Journal of chemical information and modeling*, 65(13):6547–6557, 2025.

Ziqiao Xu, Orrette R Wauchope, and Aaron T Frank. Navigating chemical space by interfacing generative artificial intelligence and molecular docking. *Journal of Chemical Information and Modeling*, 61(11):5589–5600, 2021.

Zhiqiang Yan, Yuqing Li, Ying Cao, Xuetao Tao, Jin Wang, and Yongsheng Jiang. Binding specificity and local frustration in structure-based drug discovery. *Current Medicinal Chemistry*, 2025.

A Additional Experimental Results

Table A1: Ablation study on the effect of the diffusion module.

Target Protein	Blip2Mol			Blip2Mol-NoDiff		
	Validity	Uniqueness	Novelty	Validity	Uniqueness	Novelty
AKT1	**92.0**	**100.0**	**100.0**	81.0	96.3	**100.0**
AKT2	**95.0**	**100.0**	**100.0**	82.0	98.8	**100.0**
AURKB	**95.0**	**100.0**	**100.0**	85.0	98.8	**100.0**
CTSK	**96.0**	**100.0**	**100.0**	87.0	97.7	**100.0**
EGFR	**94.0**	**100.0**	**100.0**	82.0	**100.0**	**100.0**
HDAC1	**92.0**	**100.0**	**100.0**	87.0	95.4	**100.0**
MTOR	**97.0**	**100.0**	**100.0**	91.0	**100.0**	**100.0**
PIK3CA	**92.0**	**100.0**	**100.0**	**92.0**	97.8	**100.0**
SMAD3	**95.0**	**100.0**	**100.0**	81.0	97.5	**100.0**
TP53	**91.0**	**100.0**	**100.0**	89.0	98.9	**100.0**



Figure A1: Comparison to phenotype-guided molecular generation models on Affinity, QED, and SA scores presented as box plots.

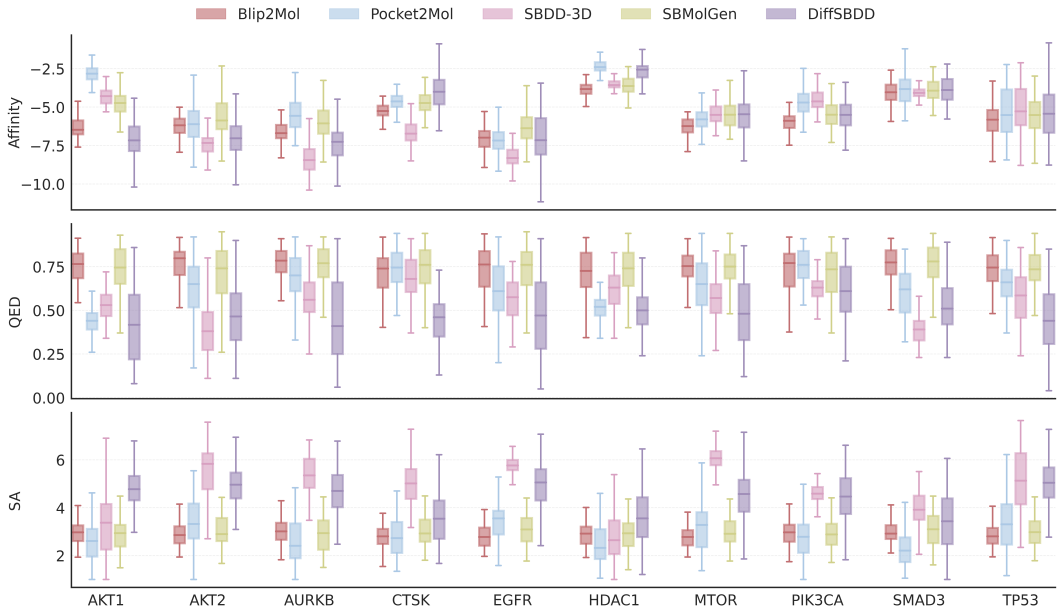


Figure A2: Comparison to target-based molecular generation models on Affinity, QED, and SA scores presented as box plots.

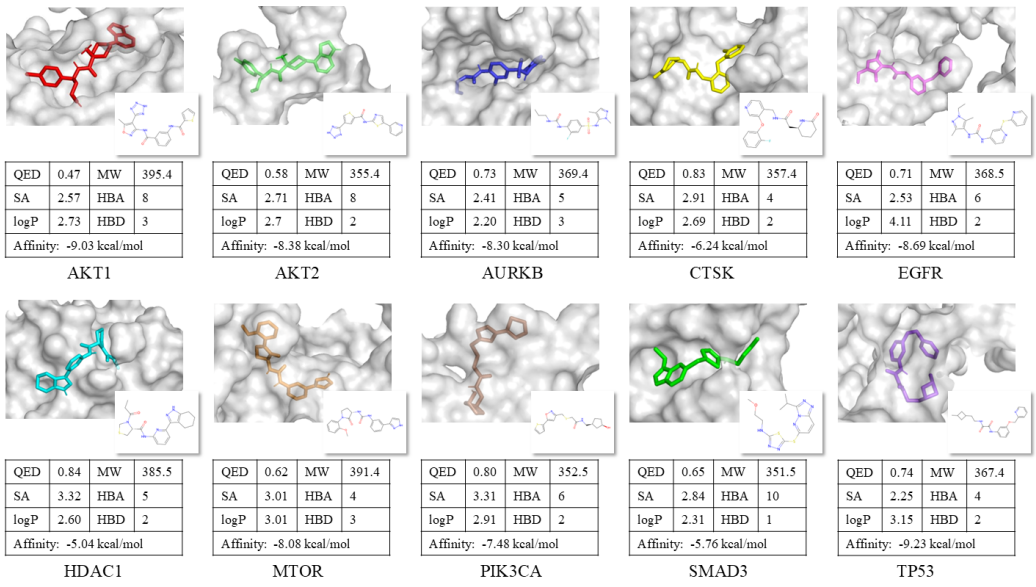


Figure A3: Local 3D visualization of molecular docking poses between Blip2Mol-generated molecules and binding pockets across 10 target proteins.