From In Silico to In Vitro: Evaluating Molecule Generative Models for Hit Generation

Nagham Osman
University College London
London, UK
nagham.osman.21@ucl.ac.uk

Vittorio Lembo
University of Urbino Carlo Bo
Urbino, Italy

Giovanni Bottegoni
University of Urbino Carlo Bo
Urbino, Italy

Laura Toni
University College London
London, UK

Abstract

Hit identification is a critical yet resource-intensive step in the drug discovery pipeline, traditionally relying on high-throughput screening of large compound libraries. Despite advancements in virtual screening, these methods remain time-consuming and costly. Recent progress in deep learning has enabled the development of generative models capable of learning complex molecular representations and generating novel compounds *de novo*. However, using ML to replace the entire drug-discovery pipeline is highly challenging. In this work, we rather investigate whether generative models can replace one step of the pipeline: *hit-like* molecule generation. To the best of our knowledge, this is the first study to explicitly frame hit-like molecule generation as a standalone task and empirically test whether generative models can directly support this stage of the drug discovery pipeline. Specifically, we investigate if such models can be trained to generate *hit-like* molecules, enabling direct incorporation into, or even substitution of, traditional hit identification workflows. We propose an evaluation framework tailored to this task, integrating physicochemical, structural, and bioactivity-related criteria within a multi-stage filtering pipeline that defines the *hit-like* chemical space. Two autoregressive and one diffusion-based generative models were benchmarked across various datasets and training settings, with outputs assessed using standard metrics and target-specific docking scores. Our results show that these models can generate valid, diverse, and biologically relevant compounds across multiple targets, with a few selected GSK-3β hits synthesized and confirmed active in vitro. We also identify key limitations in current evaluation metrics and available training data.

1 Introduction

Traditionally, drug discovery begins with target validation, followed by hit identification which is the first stage introducing chemical matter and novelty. A *hit* is a small molecule with reproducible activity, acceptable synthetic accessibility, and physicochemical properties [Goodnow, 2006]. Identifying high-quality hits is critical, as it initiates the hit-to-lead process to improve potency, selectivity, and pharmacokinetic properties. To guide this process, medicinal chemists often rely on heuristics such as Lipinski's Rule of Five, which defines drug-likeness based on molecular weight, lipophilicity, hydrogen-bonding capacity, and related efficiency metrics that balance potency with physicochemical properties. Despite its importance, hit identification remains a resource-intensive task. Various experimental strategies have been developed, including high-throughput screening (HTS) of libraries of commercially available compounds, in-house collections, or natural products, aimed at identifying

NeurIPS 2025 AI for Science Workshop.

biologically active hits. However, the main bottleneck lies in the fact that these approaches are both time-consuming (months to years) [Paul et al., 2010] and financially demanding [Hughes et al., 2011]. To partially overcome these limitations, computational methods such as structure-based virtual screening have been adopted to efficiently prioritize promising molecules. Nevertheless, even with these tools, identifying high-quality hits from vast chemical libraries continues to represent a major bottleneck in drug discovery, as these methods also entail considerable time and resource demands.

In recent years, deep learning (DL) has gained interest in molecular design due to its potential to learn complex structures and property relationships from large chemical datasets. To address persistent challenges in early drug discovery for *de novo* molecule generation, research has increasingly explored generative models that are trained either to replicate the distribution of known compounds or to optimize specific constraints, such as molecular properties or binding affinity. Common strategies include one-shot generation [De Cao and Kipf, 2018, Vignac et al., 2023, Zang and Wang, 2020], sequential atom- or bond-level construction [Gebauer et al., 2022, Segler et al., 2018, Zhou et al., 2019], and fragment-based assembly [Podda et al., 2020, Seo et al., 2021, Jin et al., 2020, Gupta et al., 2018]. However, prior research on generative models has predominantly focused on generating molecules in general, without considering how such models perform when applied to a specific step in the drug discovery process. This work investigates their applicability to a more specific and demanding task: generating hit molecules. It reframes the scientific question from the general inquiry: *Can generative models produce valid, drug-like molecules?* to a more targeted one: *Can generative models accelerate the hit identification process by directly generating compounds with hit-like characteristics?* Accordingly, our focus shifts from evaluating generative models based solely on chemical validity to assessing their ability to produce molecules suitable for entry into physics-based screening workflows. The ultimate goal is to identify biologically viable hits that can progress into hit-to-lead and lead optimization stages, either via specialized ML models or through conventional medicinal chemistry carried out by pharmaceutical experts. As the drug discovery process remains largely sequential and devoid of true shortcuts, we ask whether generative models can effectively replace one of its early steps, hit identification. If successful, this approach could substantially reduce the time and cost associated with early-stage drug discovery, potentially compressing a multi-year process into just a few months [Zhavoronkov et al., 2019, Xu et al., 2025].

Several benchmarking and scoring frameworks for de novo molecular generation have been proposed in recent years [Thomas et al., 2024, Cerveira et al., 2024] which provide flexible, target-aware scoring functions and standardized benchmarking protocols for drug-like molecule generation. These frameworks are primarily designed to compare optimization strategies or reward functions across diverse molecular design objectives.

In contrast, the main contribution of this work lies in defining, operationalizing, and empirically stress-testing hit generation as a realistic downstream task for molecular generative models. Our work focuses on a distinct and more narrowly defined task: hit generation as a standalone stage of the drug discovery pipeline. Rather than proposing new scoring functions or optimization objectives, we explicitly operationalize hit-likeness using medicinal-chemistry filters, distributional alignment to known ligands, and target-specific docking distributions, and evaluate whether off-the-shelf generative models can directly produce molecules suitable for downstream screening and experimental validation. To our knowledge, no prior work has reported a task-oriented, end-to-end evaluation of off-the-shelf generative models explicitly focused on hit generation, combining medicinal-chemistry constraints, target-specific docking distributions, and prospective in vitro validation.

To this aim, our contributions are the following:

- We define an evaluation framework tailored to the hit-generation task by combining structural, physicochemical, and bioactivity metrics with a new multi-stage filtering pipeline;
- We construct a comprehensive filtering pipeline by integrating established drug-likeness and medicinal chemistry constraints to define and evaluate the hit-like chemical space;
- We benchmark representative models from autoregressive and diffusion-based paradigms across diverse datasets and training settings, and evaluate their outputs using standard and task-specific criteria, such as docking scores across a panel of biological targets;
- We demonstrate that generative models, when evaluated under explicit hit-like constraints, can produce not only chemically valid but also biologically meaningful hit candidates, including compounds that were prospectively validated through *in vitro* assays against GSK-3β;

- We identify key limitations in current generative models and publicly available datasets that lack training data that are explicitly tailored to early-stage hit identification.

Our results demonstrate that deep generative models, when trained and fine-tuned appropriately, can generate hit-like molecules suitable for early-stage screening, including compounds with confirmed in vitro activity. These findings highlight both the potential and current limitations of generative models in practical hit discovery.

2 Method: Benchmarks, Training, and Evaluation Pipelines

We evaluate three graph-based generative models, two autoregressive and one diffusion-based, that have shown strong performance in general molecule generation but have not been assessed for the specific task of hit generation. This section outlines the models, datasets, training settings, evaluation pipelines, and in vitro validation of GSK-3β hits.

2.1 Studied Models

The three graph-based models considered are MolRNN [Li et al., 2018], GraphINVENT [Mercado et al., 2020], and DiGress [Vignac et al., 2023]. MolRNN and GraphINVENT are autoregressive, sequentially building molecular graphs step by step. They differ in their graph encoders: MolRNN uses a gated recurrent unit (GRU), whereas GraphINVENT employs a message-passing neural network (MPNN). DiGress, in contrast, uses a diffusion-based, one-shot generation approach with a graph transformer. This selection was based on the models' performance in preliminary evaluations and allows comparison of sequential versus non-sequential generation strategies, and the impact of different encoder architectures (GRU, MPNN, transformer) on scalability, and chemical validity.

MolRNN MolRNN [Li et al., 2018] is a sequential generative model that constructs molecular graphs by leveraging graph neural networks (GNNs) and recurrent neural networks (RNNs). It starts with an empty graph and iteratively guides the generation process using a probabilistic decoding policy to apply one of three actions: adding atoms, connecting atoms, or terminating the process. Each step involves computing atom-level embeddings through stacked GNN layers and refining graph-level representations with a GRU unit. A multi-layer perceptron is then used to predict action probabilities. The model is trained via maximum likelihood estimation using importance sampling. Full technical details are available in [Li et al., 2018].

GraphINVENT GraphINVENT [Mercado et al., 2020] is a graph-based generative model that constructs molecules by autoregressively adding atoms and bonds. GraphINVENT's model architecture is similar to that of MolRNN; however, it employs MPNNs, which offer improved scalability and more fine-grained edge handling compared to GRUs. Molecular generation is guided by an Action Probability Distribution (APD) that restricts to chemically valid actions. The architecture features an MPNN encoder and a readout block to produce APD logits, trained by minimizing KL divergence against ground-truth actions derived from BFS-based preprocessing. See [Mercado et al., 2020] for a comprehensive description.

DiGress DiGress [Vignac et al., 2023] is a discrete denoising diffusion model for graph generation. Instead of continuous noise, it introduces categorical perturbations to node and edge features via discrete Markov transitions to ensure discreteness of its representation. For the denoising phase, a graph transformer network is employed that learns to reverse the noise process, using cross-entropy loss over node and edge predictions. The model incorporates structural and spectral features, such as cycle counts and Laplacian eigenvalues, to enhance its generative capabilities. Full methodology and results are detailed in [Vignac et al., 2023].

2.2 Datasets

To train our models, we utilize three types of datasets in this work: a general-purpose dataset (REINVENT [Blaschke et al., 2020]), a filtered subset we term the hit-like dataset, and a target-specific ligand dataset.

REINVENT Dataset The REINVENT dataset [Olivecrona et al., 2017] is a widely used subset of ChEMBL [Mendez et al., 2019], originally introduced by Olivecrona et al. It contains 1,086,248 unique compounds after filtering for molecules with 10–50 heavy atoms and restricting elements to H, B, C, N, O, F, Si, P, S, Cl, Br, and I. Duplicate compounds were removed to ensure uniqueness.

Hit-like Dataset The hit-like dataset, constructed for this study, is derived from the same ChEMBL source but filtered using the hit-like criteria such as molecular weight, synthetic accessibility score, ring constraints, and others described in Sec. 2.4.3. This process yielded 58,837 molecules, approximately 2.68% of the original database, representing structures with desirable drug-like properties.

Target-Specific Ligand Sets While our main focus is unconditional generation, target-specific ligand sets were also constructed to explore whether such models could be refined toward activity against defined biological targets, providing a complementary perspective on their downstream applicability. To construct these datasets, we retrieved ligands from ChEMBL annotated with activity against a panel of seven protein targets: dopamine D3 receptor (D3R, class A GPCR), adenosine A2A receptor (ADORA2A, class A GPCR), heat shock protein HSP90α (HSP90α, molecular chaperone), glycogen synthase kinase-3β (GSK-3β, serine/threonine kinase), proto-oncogene tyrosine-protein kinase Src (SRC, tyrosine kinase), coagulation factor II (thrombin, protease), and peroxisome proliferator-activated receptor α (PPARα, nuclear receptor). Ligands were selected based on two criteria: (i) pChEMBL value ≥ 5, and (ii) assay confidence score $= 9$.

2.3 Training Pipeline

As shown in the left hand side of Figure 1, each model was trained under three settings: (1) using the full REINVENT dataset, (2) using the hit-like dataset, and (3) fine-tuning a REINVENT-trained model with the hit-like dataset. Additionally, MolRNN and DiGress models that were initially trained on the REINVENT dataset were further fine-tuned in a fourth setting on seven target-specific ligand sets (see 2.2 for details). This choice was made to enable comparison between one autoregressive and one diffusion-based model. MolRNN was selected over GraphINVENT for this step due to its superior performance in preliminary benchmarks.

All models followed their original training objectives, with early stopping on validation loss. For fine-tuning experiments, models were initialized from pretrained REINVENT weights and continued training for up to 20 epochs, using smaller batch sizes to avoid overfitting. For generation, the best-performing checkpoint for each trained model (lowest validation loss) was selected. To ensure fair comparison across models during downstream analysis, we continued generation until each model produced the same number of VUN and hit-like filtered molecules. This controlled sampling allowed consistent evaluation on docking and drug-likeness metrics. After training and sampling, each model was passed to the evaluation pipeline described in Sec. 2.4.

2.4 Evaluation Pipeline

To assess model performance, we adopt a three-layer evaluation framework of increasing drug relevance. Our goal is to construct a realistic evaluation pipeline (Figure 1, right), aiming to determine whether generative models can replace hand-crafted hit-like molecule design.

2.4.1 Basic Structural Validity Metrics

The first layer of evaluation focuses on foundational checks to ensure that generated molecules are structurally and chemically sound. Validity measures the percentage of generated molecules that obey valence rules and contain no structural errors. Uniqueness assesses the proportion of valid molecules that are structurally distinct within the generated set. Novelty quantifies the fraction of valid and unique molecules that are not present in the training dataset. These three metrics (VUN) serve as the minimal baseline for generative model performance and are standard across the field.

2.4.2 Distributional and Structural Similarity Metrics

The second evaluation layer measures how closely generated molecules resemble the training chemical space. We employ a set of metrics that assess both molecular properties and structural patterns that align with standard benchmarking protocols such as GuacaMol [Brown et al., 2019] and

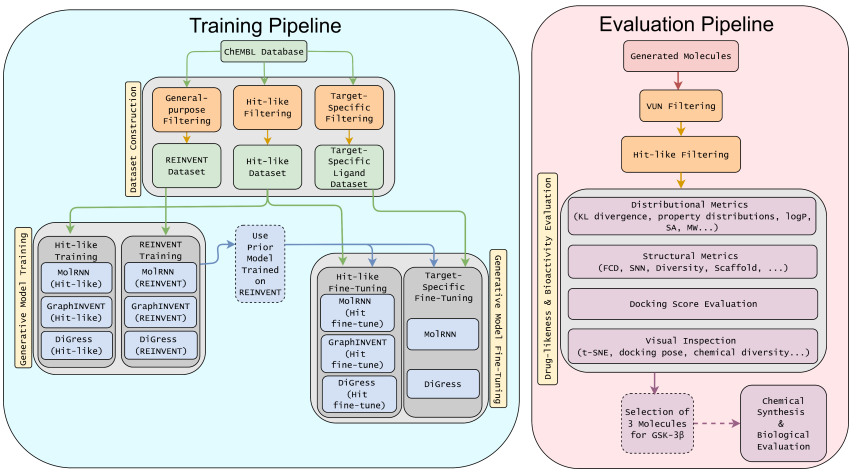

Figure 1: *Integrated Training and Evaluation Framework.* **Left:** Dataset construction and model training pipeline, showing the filtering strategies applied to ChEMBL to obtain general-purpose, hit-like, and target-specific datasets, their use in training MolRNN, GraphINVENT, and DiGress models, and subsequent fine-tuning in hit-like and target-specific settings. **Right:** Evaluation pipeline for generated molecules, outlining the multi-stage workflow from VUN and hit-like filtering through distributional, structural, and docking-based metrics, visual inspection, and final selection of top GSK-3β candidates for synthesis and in vitro testing.

MOSES [Polykovskiy et al., 2018]. We use Fréchet ChemNet Distance (FCD) Preuer et al. [2018] for overall distributional similarity based on chemical and biological features of generated and reference molecules, BRICS fragments [Degen et al., 2008] and Bemis–Murcko scaffolds [Bemis and Murcko, 1996] similarities for structural patterns, Similarity to Nearest Neighbor (SNN) for proximity to known compounds based on Tanimoto similarity, and Jaccard-based internal diversity on Morgan fingerprints [Morgan, 1965] for variation within the generated set.

2.4.3 Drug-likeness and Bioactivity-Relevant Metrics

This final evaluation tier assesses whether generative models produce molecules with realistic hit-like properties and potential biological activity. This aspect is one of the key contributions of our work where we emphasize evaluation by drug-likeness and biological relevance which are critical factors in determining the practical value of generated compounds.

Hit-like Filters To align with early-stage discovery goals, we applied a multi-stage filtering pipeline integrating established drug-likeness and medicinal chemistry criteria. Each molecule was required to satisfy: (1) a Novartis-inspired [Schuffenhauer et al., 2020] severity score (Sev.) ≤ 10, which penalizes structural liabilities such as PAINS motifs [Baell and Holloway, 2010], reactive groups, toxicophores, and unstable chemotypes; (2) molecular weight (MW) between 150–350 Da; (3) logP between 1–3 as a proxy for solubility; (4) a minimum pChEMBL value of 5 at any known target to ensure meaningful bioactivity; (5) a synthetic accessibility score (SAS) ≤ 5; (6) between one (NoR) and four rings (< 4R), with no ring exceeding eight atoms (< 8t), a maximum of two rings larger than six atoms (< 2R6t), and exclusion of fused (FusedR) or small aromatic (AromR) ring systems; and (7) elemental composition restricted to C, N, O, F, P, S, Cl, Br, and I. This integrated approach provides a consistent and stringent definition of hit-likeness across datasets and generative outputs.

Docking Calculation To evaluate the biological relevance of the generated molecules, we computed docking scores estimating their binding affinity to the seven protein targets described in Sec. 2.2. Docking simulations were performed only on molecules that satisfied both the VUN criteria and the complete set of hit-like filters, ensuring that only the most plausible candidates were assessed for target binding. Docking scores, with lower values indicating stronger predicted interactions, serve as an approximate measure of bioactivity and complement chemical and structural evaluations. Details of protein and ligand preparation, docking settings, and software are provided in Appendix A.

Table 1: Molecular generation performance metrics for MolRNN, GraphINVENT, and DiGress across training settings.

Model	Valid (%)	VUN (%)	Filters (%)	FCD	Frag.	Scaff.	SNN	Div.
MolRNN (Hit-like)	99.98	64.60	**76.87**	1.74	0.93	**0.87**	**0.61**	0.87
MolRNN (REINVENT)	99.98	**96.74**	11.46	2.05	0.92	0.79	0.50	0.87
MolRNN (Hit fine-tune)	99.91	94.64	67.57	1.91	0.92	0.84	0.52	0.87
GraphINVENT (Hit-like)	**100.00**	**96.42**	58.73	2.10	**0.99**	0.77	0.47	**0.88**
GraphINVENT (REINVENT)	**100.00**	**96.98**	12.85	1.70	**0.99**	0.78	0.48	**0.88**
GraphINVENT (Hit fine-tune)	**100.00**	93.06	50.80	**1.13**	**0.99**	0.86	0.53	**0.88**
DiGress (Hit-like)	70.00	69.15	65.59	3.01	**0.99**	0.68	0.44	**0.88**
DiGress (REINVENT)	75.81	75.26	10.27	2.63	**0.99**	0.65	0.45	**0.88**
DiGress (Hit fine-tune)	71.62	71.51	14.11	2.91	0.98	0.63	0.45	**0.88**

Table 2: Percentage of generated molecules failing each filtering criterion.

Model	Sev.	SAS	MW	logP	NoR	<4R	<8t	<2R6t	AromR	FusedR	All
MolRNN (Hit-like)	**0.80**	**0.30**	9.97	**14.39**	0.16	**0.27**	**0.37**	0.03	0.03	**0.02**	**23.13**
MolRNN (REINVENT)	7.29	2.60	77.69	66.01	0.47	14.35	1.14	0.17	0.10	0.41	88.54
MolRNN (Hit fine-tune)	1.96	0.37	15.14	19.63	0.25	0.40	0.26	**0.01**	**0.02**	0.06	32.43
GraphINVENT (Hit-like)	7.51	6.59	11.26	21.90	1.85	1.49	8.45	1.75	1.69	1.12	41.27
GraphINVENT (REINVENT)	9.06	8.59	73.25	63.59	1.43	16.30	8.98	3.28	1.23	1.80	87.15
GraphINVENT (Hit fine-tune)	5.00	3.49	22.94	31.32	1.04	1.86	4.24	0.96	0.65	0.68	49.20
DiGress (Hit-like)	4.58	3.19	**4.57**	16.36	0.33	2.22	10.30	1.44	2.56	0.86	34.41
DiGress (REINVENT)	7.15	10.18	78.03	62.39	0.24	22.09	13.70	3.62	0.87	2.31	89.73
DiGress (Hit fine-tune)	4.79	10.42	74.80	54.40	**0.04**	25.29	15.25	4.05	0.71	1.96	85.89

Biological Evaluation To assess biological relevance of generated molecules, we performed hit selection for GSK-3β, a target where our models showed low KL divergence and many compounds outperforming the median docking score of known ligands. Candidate molecules were filtered to meet two criteria: (1) structural novelty (Tanimoto distance ≥ 0.5 from known binders), and (2) strong predicted binding (docking scores $<$ mean -2 SD). From this subset, compounds were prioritized based on strong docking scores, visual inspection of binding poses, and synthetic feasibility. This process led to the selection of three top compounds subjected to further experimental testing in vitro to validate their potency. This procedure reflects a realistic medicinal-chemistry triage rather than a retrospective enrichment analysis.

3 Experimental Results and Discussions

This section presents our experimental results, including general-purpose molecule generation, docking score calculations, and the biological evaluation of selected hits against the GSK-3β target.

3.1 Molecular Generation Performance and Filter Compliance

The first step is to evaluate whether the models could reliably generate chemically valid, unique, and novel (VUN) molecules. As shown in Table 1, MolRNN and GraphINVENT achieved near-perfect validity ($> 99.9\%$) across all settings, while DiGress reached only 70–76%. Uniqueness was uniformly high, but novelty varied: GraphINVENT consistently exceeded 93%, MolRNN was sensitive to training data (dropping to 64.6% in the Hit-like regime), and DiGress remained moderate ($\sim$ 70–75%) across all cases. We next assessed how many VUN molecules passed the hit-like property filters. Direct training on the Hit-like dataset yielded the highest compliance (MolRNN: 76.9%), while REINVENT-trained models passed at $< 12\%$, underscoring the mismatch between broad chemical priors and focused hit-like requirements. Fine-tuning substantially improved MolRNN and GraphINVENT compliance, but DiGress saw only modest gains. Distributional metrics in Table 1 highlight trade-offs between novelty and hit-like resemblance. GraphINVENT (Hit fine-tune) achieved the lowest FCD (1.13), indicating the closest match to the reference, while MolRNN showed the highest scaffold similarity, and DiGress exhibited generally higher FCD and scaffold deviation.

Table 3: Average KL divergence (mean $\pm$ standard deviation) between the docking score distributions of generated molecules and the reference ligand sets across all targets: (a) grouped by combination of algorithm and training set; (b) grouped by algorithm.

Model	KL divergence
DiGress (REINVENT)	0.099 ± 0.061
DiGress (Hit-like)	0.009 ± 0.012
DiGress (Hit fine-tune)	0.089 ± 0.056
GraphINVENT (REINVENT)	0.100 ± 0.067
GraphINVENT (Hit-like)	0.003 ± 0.002
GraphINVENT (Hit fine-tune)	0.087 ± 0.062
MolRNN (REINVENT)	0.083 ± 0.056
MolRNN (Hit-like)	0.004 ± 0.003
MolRNN (Hit fine-tune)	0.095 ± 0.064

(a) By model and training set

Algorithm	KL divergence
DiGress	0.066 ± 0.062
GraphINVENT	0.064 ± 0.067
MolRNN	0.061 ± 0.062

(b) By algorithm

To pinpoint limiting factors, Table 2 reports per-filter failure rates. MolRNN (Hit-like) showed minimal violations across all criteria, while REINVENT-trained MolRNN and DiGress frequently exceeded MW and logP thresholds. GraphINVENT tended toward higher ring-complexity and logP failures, partially mitigated by fine-tuning. DiGress, while rarely producing ringless molecules, overrepresented large and fused ring systems, with high MW violations persisting even after fine-tuning. DiGress consistently underperformed autoregressive models in validity and hit-like filter compliance. We attribute this primarily to the challenges of discrete diffusion in low-data, highly constrained regimes such as hit-like chemical space. Unlike autoregressive models, which enforce chemical validity at each generation step, diffusion-based approaches rely on global denoising and appear more sensitive to distributional sparsity and strict post hoc filters. This suggests that, in their current form, diffusion graph models may be less well suited to direct hit generation without additional constraint mechanisms. Overall, autoregressive models, when trained or fine-tuned on hit-like data, achieve the best balance between generative breadth and property compliance, whereas diffusion-based generation shows lower baseline validity and weaker adaptation to narrow physicochemical constraints.

3.2 Docking Analysis

Docking performance was evaluated across seven targets (Sec. 2.2) using KL divergence, quantifying the similarity between docking score distributions of generated molecules and known active ligands for each target. As shown in Table 3a, models trained directly on the Hit-like set achieved the lowest divergences (all < 0.01), reflecting strong physicochemical alignment with the reference ligands. Unfiltered REINVENT-trained models exhibited higher divergence (0.083–0.100), consistent with the broader and less constrained chemical space. Fine-tuned models achieved intermediate values, indicating partial adaptation toward the hit-like profile.

When averaged across targets, MolRNN achieved lowest KL divergence (0.061 ± 0.062), with GraphINVENT and DiGress performing similarly (Table 3b). Per-target analysis in Appendix C revealed highest divergences for SRC and PPARα, reflecting their higher MW and logP of their known ligands compared to the generated set. In contrast, GSK-3β and ADORA2A showed the closest alignment with the hit-like constraints.

While KL divergence captures distributional similarity, it does not assess binding efficacy. We therefore compared docking scores of generated compounds to known ligands for each target, calculating the fraction outperforming (lower) the ligand median (Table 4). Across 63 model–training combinations, HSP90α showed the highest success rates (38–52%), PPARα the lowest ($< 1\%$), and GSK-3β (moderate 10–16%), with similar trends for D3R and SRC. Given that many ChEMBL ligands are optimized molecules occupying broader chemical space, we repeated the analysis using only ligands passing hit-like filters (see full table details in Table 7 in Appendix C. This raised success rates for five targets, most notably for Thrombin ($< 2\%$ to 35–45%), while GSK-3β and HSP90α observed modest drops. MolRNN (Hit fine-tune) achieved the highest success in 6 of 7 targets, with

Table 4: Median docking score values of the ligand sets for PPARα, HSP90α, and GSK-3β, and the percentage of generated molecules achieving a lower (better) docking score than the respective median ligand score, grouped by generative model.

Model	Target	Median	% Better
Hit-like	PPARα	-9.15	0.38
	HSP90α	-6.03	39.97
	GSK-3β	-7.27	10.13
REINVENT	PPARα	-9.15	0.35
	HSP90α	-6.03	42.69
	GSK-3β	-7.27	12.18
Hit fine-tune	PPARα	-9.15	0.54
	HSP90α	-6.03	40.22
	GSK-3β	-7.27	11.60

(a) DiGress

Model	Target	Median	% Better
Hit-like	PPARα	-9.15	0.38
	HSP90α	-6.03	42.07
	GSK-3β	-7.27	15.58
REINVENT	PPARα	-9.15	0.46
	HSP90α	-6.03	43.05
	GSK-3β	-7.27	15.60
Hit fine-tune	PPARα	-9.15	0.52
	HSP90α	-6.03	51.82
	GSK-3β	-7.27	20.59

(b) MolRNN

Model	Target	Median	% Better
Hit-like	PPARα	-9.15	0.38
	HSP90α	-6.03	38.35
	GSK-3β	-7.27	14.00
REINVENT	PPARα	-9.15	0.28
	HSP90α	-6.03	39.78
	GSK-3β	-7.27	13.38
Hit fine-tune	PPARα	-9.15	0.34
	HSP90α	-6.03	39.07
	GSK-3β	-7.27	14.86

(c) GraphINVENT

DiGress (Hit fine-tune) narrowly leading for PPARα. The weakest performances came from DiGress (Hit-like) in four targets, GraphINVENT (REINVENT) in two, and GraphINVENT (Hit-like) for HSP90α. Overall, fine-tuned models on hit-like data generated a substantial fraction of compounds with docking scores comparable to or better than known ligands, though performance varied by target.

3.3 Hit Selection and Biological Evaluation

Given its strong generative performance, GSK-3β was chosen for hit prioritization, yielding three candidate molecules that were synthetized and tested in vitro. All showed high structural similarity (TD > 0.5) to known kinase inhibitors, consistent with the conserved nature of kinase binding pockets. Compound 1 displayed strong GSK-3β inhibition (IC_{50} = 314 $\pm$ 15.3 nM; Table 5 in Appendix B; Figure 2A), while compounds 2 and 3 showed weaker activity. Such low-nanomolar activity, as in compound 1, is highly promising at the hit stage, particularly for a small, structurally novel scaffold. Notably, the training dataset contained few ligands of comparable potency and size, making compound 1 a clear example of successful extrapolation by the generative model. Docking analysis confirmed the hinge-binding of compound 1 through two key hydrogen bonds (Asp133, Val135) and a Lys85 interaction, explaining its high potency despite the rarity of its scaffold ($< 1.1\%$ among kinase binders present in ChEMBL).

As shown in Figure 2, chemical space analysis further supported its relevance: t-SNE projections revealed that compound 1 clusters with known GSK-3β inhibitors, while activity benchmarking demonstrated that it outperforms both the full ligand set and the subset of hit-like inhibitors (–log(activity) = 6.50 vs. median values of 6.39 and 5.89, respectively). These results highlight that generative models can design chemically novel yet biologically potent compounds, even without explicit fine-tuning on a specific target, by leveraging kinase features embedded in the training data.

4 Conclusion

This work presents a task-focused evaluation of generative models explicitly framed around hit generation, combining medicinal-chemistry filtering, docking-based distributional analysis, and prospective experimental in vitro validation. By isolating hit generation as a standalone objective, our study complements existing benchmarking and scoring frameworks by assessing whether generative models can directly produce candidates suitable for early-stage experimental screening. Our results demonstrate that such models can generate chemically novel compounds meeting hit-like criteria and displaying measurable biological activity, confirming their potential to accelerate early-stage drug discovery. Nonetheless, limitations remain. Diffusion-based models, despite strong performance in distribution learning, struggled with target-specific fine-tuning, due to their higher data demands and sensitivity to dataset size. Performance on targets such as PPARα and SRC was constrained by the scarce high-quality hit-like data highlighting the broader issue of limited availability even for well-studied proteins. Moreover, widely used evaluation metrics such as VUN, FCD, and scaffold similarity did not consistently correlate with predicted bioactivity, underscoring the need for benchmarks that better capture biological relevance.

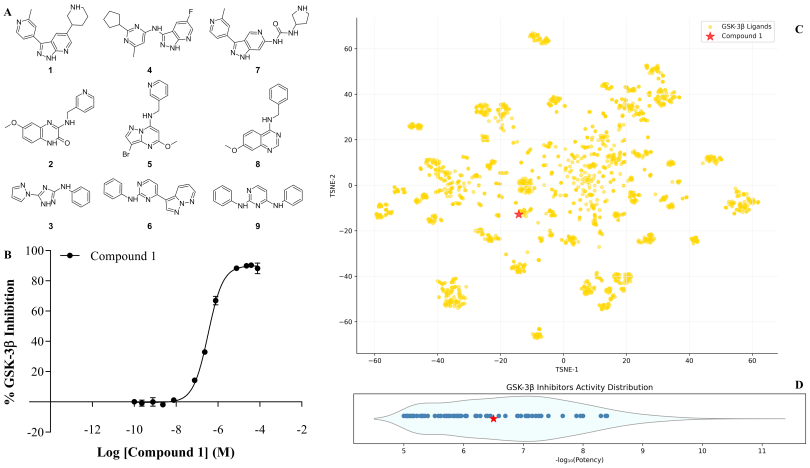

Figure 2: (A) Structures of selected compounds (1–3), their most similar known GSK-3β inhibitors (4–6), and closest kinase inhibitors (7–9) by TD. (B) Dose–response curve for GSK-3β inhibition by compound 1, shown as mean $\pm$ SD of three replicates, representative of three experiments. (C) t-SNE projection of compound 1 (red star) within the chemical space of known GSK-3β ligands (gold dots). (D) Comparison of pChEMBL value of compound 1 with known inhibitors, shown as a violin plot (all inhibitors, light blue) and subset with hit-like properties (blue dots).

Future progress will require richer, better-curated hit-like datasets to support both general and target-specific training, as shown by constrained performance on PPARα and SRC due to scarce ligands with suitable physicochemical properties. In parallel, evaluation frameworks must integrate bioactivity-relevant endpoints alongside distributional and structural metrics. Advances in architectures that learn effectively from limited target-specific data through improved transfer learning or data augmentation strategies will also be key. Incorporating these improvements into a modular AI-driven pipeline, where generative models focus on high-quality hit generation followed by ML-based or traditional hit-to-lead optimization, offers a promising path to accelerate early-stage drug discovery while addressing current gaps in predictive power and applicability.

References

Jonathan B. Baell and Georgina A. Holloway. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, 2010.

Guy W. Bemis and Mark A. Murcko. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996.

J. A. Bertrand, S. Thieffine, A. Vulpetti, C. Cristiani, B. Valsasina, S. Knapp, H. M. Kalisz, and M. Flocco. Structural characterization of the GSK-3β active site using selective and non-selective ATP-mimetic inhibitors. *Journal of Molecular Biology*, 333(2):393–407, 2003.

Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov. REINVENT 2.0: An AI Tool for De Novo Drug Design. *Journal of Chemical Information and Modeling*, 60(12):5918–5922, 2020.

Paul A. Brough, Wynne Aherne, Xavier Barril, Jenifer Borgognoni, Kathy Boxall, Julie E. Cansfield, Kwai-Ming J. Cheung, Ian Collins, Nicholas G. M. Davies, Martin J. Drysdale, Brian Dymock, Suzanne A. Eccles, Harry Finch, Alexandra Fink, Angela Hayes, Robert Howes, Roderick E. Hubbard, Karen James, Allan M. Jordan, Andrea Lockie, Vanessa Martins, Andrew Massey, Thomas P. Matthews, Edward McDonald, Christopher J. Northfield, Laurence H. Pearl, Chrisostomos Prodromou, Stuart Ray, Florence I. Raynaud, Stephen D. Roughley, Swee Y. Sharp, Allan Surgenor, D. Lee Walmsley, Paul Webb, Mike Wood, Paul Workman, and Lisa Wright. 4,5-Diarylisoxazole Hsp90 chaperone inhibitors: Potential therapeutic agents for the treatment of cancer. *Journal of Medicinal Chemistry*, 51(2):196–218, 2008.

Nathan Brown, Marco Fiscato, Marwin H.S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling*, 59(3): 1096–1108, 2019.

Arthur Cerveira, Frederico Kremer, Darling Lourenço, and Ulisses B. Corrêa. Evaluation framework for ai-driven molecular design of multi-target drugs: Brain diseases as a case study. In *2024 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, June 2024. doi: 10.1109/CEC60901.2024.10611839. URL `https://ieeexplore.ieee.org/document/10611839`.

Ellen Y. T. Chien, Wei Liu, Qiang Zhao, Vsevolod Katritch, Gye Won Han, Michael A. Hanson, Lei Shi, Amy Hauck Newman, Jonathan A. Javitch, Vadim Cherezov, and Raymond C. Stevens. Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. *Science (New York, N.Y.)*, 330(6007):1091–1095, 2010.

Hipólito Nicolás Cuesta-Hernández, Julia Contreras, Pablo Soriano-Maldonado, Jana Sánchez-Wandelmer, Wayland Yeung, Ana Martín-Hurtado, Inés G. Muñoz, Natarajan Kannan, Marta Llimargas, Javier Muñoz, and Iván Plaza-Menacho. An allosteric switch between the activation loop and a c-terminal palindromic phospho-motif controls c-Src function. *Nature Communications*, 14(1):6548, 2023.

Nicola De Cao and Thomas Kipf. MolGAN: an implicit generative model for small molecular graphs. In *ICML 2018 Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem*, 3(10):1503–1507, 2008.

R. M. C. Di Martino, D. Russo, I. Penna, S. Demuro, A. Dalle Vedove, R. Spagnuolo, G. Ottonello, S. M. Bertozzi, M. Summa, J. Desantis, A. Valeri, L. Pruccoli, S. K. Tripathi, A. Tarozzi, P. Storici, S. Girotto, R. Bertorelli, A. Armirotti, G. Cruciani, T. Bandiera, A. Cavalli, and G. Bottegoni. Rational optimization of D3R/GSK-3β dual target-directed ligands as potential treatment for bipolar disorder: Design, synthesis, X-ray crystallography, molecular dynamics simulations, *in vitro* ADME, and *in vivo* pharmacokinetic studies. *European Journal of Medicinal Chemistry*, 297: 117899, 2025.

Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, 2004.

Richard A. Friesner, Robert B. Murphy, Matthew P. Repasky, Leah L. Frye, Jeremy R. Greenwood, Thomas A. Halgren, Paul C. Sanschagrin, and Daniel T. Mainz. Extra precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. *Journal of Medicinal Chemistry*, 49(21):6177–6196, 2006.

Niklas W. A. Gebauer, Michael Gastegger, Stefaan S. P. Hessmann, Klaus-Robert Müller, and Kristof T. Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature Communications*, 13(1):973, 2022.

Robert A. Goodnow. Hit and lead identification: integrated technology-based approaches. *Drug Discovery Today: Technologies*, 3(4):367–375, 2006.

Anvita Gupta, Alex T. Müller, Berend J. H. Huisman, Jens A. Fuchs, Petra Schneider, and Gisbert Schneider. Generative recurrent networks for *de novo* drug design. *Molecular Informatics*, 37(1-2): 1700111, 2018.

Thomas A. Halgren, Robert B. Murphy, Richard A. Friesner, Hege S. Beard, Leah L. Frye, W. Thomas Pollard, and Jay L. Banks. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of Medicinal Chemistry*, 47(7):1750–1759, 2004.

J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott. Principles of early drug discovery. *British Journal of Pharmacology*, 162:1239–1249, 2011.

Veli-Pekka Jaakola, Mark T. Griffith, Michael A. Hanson, Vadim Cherezov, Ellen Y. T. Chien, J. Robert Lane, Adriaan P. IJzerman, and Raymond C. Stevens. The 2.6 å crystal structure of a human A_{2A} adenosine receptor bound to an antagonist. *Science*, 322(5905):1211–1217, 2008.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In *Proceedings of the 37th International Conference on Machine Learning 2020*, 2020.

Shotaro Kamata, Takuji Oyama, Kenta Saito, Akihiro Honda, Yume Yamamoto, Keisuke Suda, Ryo Ishikawa, Toshimasa Itoh, Yasuo Watanabe, Takahiro Shibata, Koji Uchida, Makoto Suematsu, and Isao Ishii. PPARα ligand-binding domain structures with endogenous fatty acids and fibrates. *iScience*, 23(11):101727, 2020.

Yibo Li, Liangren Zhang, and Zhenming Liu. Multi-objective de novo drug design with conditional graph generative model. *Journal of Cheminformatics*, 10(1):33, 2018.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1): D930–D940, 2019.

Rocío Mercado, Tobias Rastemo, Edvard Lindelöf, Günter Klambauer, Ola Engkvist, Hongming Chen, and Esben Jannik Bjerrum. Graph Networks for Molecular Design. *Machine Learning: Science and Technology*, 2(2):025023, 2020.

H. L. Morgan. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113, 1965.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, 2017.

Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg, and Aaron L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, 2010.

Donald J. P. Pinto, Michael J. Orwat, Mimi L. Quan, Qi Han, Robert A. Galemmo, Eugene Amparo, Brian Wells, Christopher Ellis, Ming Y. He, Richard S. Alexander, Karen A. Rossi, Angela Smallwood, Pancras C. Wong, Joseph M. Luettgen, Alan R. Rendina, Robert M. Knabb, Lawrence Mersinger, Charles Kettner, Steven Bai, Kan He, Ruth R. Wexler, and Patrick Y. S. Lam. 1-[3-Aminobenzisoxazol-5′-yl]-3-trifluoromethyl-6-[2′-(3-(R)-hydroxy-N-pyrrolidinyl)methyl-[1,1′]-biphen-4-yl]-1,4,5,6-tetrahydropyrazolo-[3,4-c]-pyridin-7-one (BMS-740808): A highly potent, selective, efficacious, and orally bioavailable inhibitor of blood coagulation factor Xa. *Bioorganic & Medicinal Chemistry Letters*, 16(15):4141–4147, 2006.

Marco Podda, Davide Bacciu, and Alessio Micheli. A deep generative model for fragment-based molecule generation. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020*, 2020.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models, 2018.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. Fréchet ChemNet Distance: A metric for generative models for molecules in drug discovery, 2018.

Ansgar Schuffenhauer, Nadine Schneider, Samuel Hintermann, Douglas Auld, Jutta Blank, Simona Cotesta, Caroline Engeloch, Nikolas Fechner, Christoph Gaul, Jerome Giovannoni, Johanna Jansen, John Joslin, Philipp Krastel, Eugen Lounkine, John Manchester, Lauren G. Monovich, Anna Paola Pelliccioli, Manuel Schwarze, Michael D. Shultz, Nikolaus Stiefl, and Daniel K. Baeschlin. Evolution of Novartis' Small Molecule Screening Deck Design. *Journal of Medicinal Chemistry*, 63(23):14425–14447, 2020.

Marwin H. S. Segler, Thierry Kogej, Christian Tyrchan, and Mark P. Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4(1): 120–131, 2018.

Seonghwan Seo, Jaechang Lim, and Woo Youn Kim. Fragment-based molecular generative model with high generalization ability and synthetic accessibility, 2021.

Morgan Thomas, Noel M. O'Boyle, Andreas Bender, and Chris De Graaf. Molscore: a scoring, evaluation and benchmarking framework for generative models in de novo drug design. *Journal of Cheminformatics*, 16:64, May 2024. ISSN 1758-2946. doi: 10.1186/s13321-024-00861-w. URL `https://jcheminf.biomedcentral.com/articles/10.1186/s13321-024-00861-w`.

Clément Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. DiGress: Discrete denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Zuojun Xu, Feng Ren, Ping Wang, Jie Cao, Chunting Tan, Dedong Ma, Li Zhao, Jinghong Dai, Yipeng Ding, Haohui Fang, Huiping Li, Hong Liu, Fengming Luo, Ying Meng, Pinhua Pan, Pingchao Xiang, Zuke Xiao, Sujata Rao, Carol Satler, Sang Liu, Yuan Lv, Heng Zhao, Shan Chen, Hui Cui, Mikhail Korzinkin, David Gennert, and Alex Zhavoronkov. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. *Nature Medicine*, 31(8):2602–2610, 2025.

Chengxi Zang and Fei Wang. Moflow: an invertible flow model for generating molecular graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 617–626, 2020.

Alex Zhavoronkov, Yan A. Ivanenkov, Alex Aliper, Mark S. Veselov, Vladimir A. Aladinskiy, Anastasiya V. Aladinskaya, Victor A. Terentiev, Daniil A. Polykovskiy, Maksim D. Kuznetsov, Arip Asadulaev, Yury Volkov, Artem Zholus, Rim R. Shayakhmetov, Alexander Zhebrak, Lidiya I. Minaeva, Bogdan A. Zagribelnyy, Lennart H. Lee, Richard Soll, David Madge, Li Xing, Tao Guo, and Alán Aspuru-Guzik. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, 37(9):1038–1040, 2019.

Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N. Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific Reports*, 9(1):10752, 2019.

A Docking Studies Details

Ligand docking calculations were carried out using Glide [Halgren et al., 2004, Friesner et al., 2004] as implemented in Schrödinger 2023–04 (Schrödinger LLC, New York, USA). All ligands were prepared using the LigPrep tool. For each compound, stereoisomers were generated for all chiral centers, up to a maximum of 32. Ionization states were generated using Epik Classic, considering a pH range of 7.4 ± 1.0. All other parameters were kept at their default settings. All receptors were prepared using the Protein Preparation tool implemented in Schrödinger 2023–04. Default settings were used. Water molecules, ions, and other non-protein atoms were removed from protein structures.

The position and size of the binding boxes of GSK-3β (PDB code: 1Q41 [Bertrand et al., 2003], chain A), thrombin (PDB code: 2FZZ [Pinto et al., 2006], chain H), and SRC (PDB code: 7OTE [Cuesta-Hernández et al., 2023], chain A) systems were defined by the coordinates of their co-crystallized ligand. To account for receptor flexibility, the van der Waals radii for atoms with a partial charge ≤ 0.25 were scaled to 0.9 for each system. For the HSP90α system (PDB code: 2VCI [Brough et al., 2008], chain A), the binding box was defined in the same manner as described above, except

that three water molecules near the key residue Asp93—reported to form a hydrogen-bond network between the ligand and the receptor—were retained.

The binding boxes of the PPARα system (PDB code: 6KB4 [Kamata et al., 2020], chain A), D3R (PDB code: 3PBL [Chien et al., 2010], chain A), and ADORA2A (PDB code: 3EML [Jaakola et al., 2008], chain A) systems were created following the same procedure described above, with the additional inclusion of the following atomic constraints during docking:

1. PPARα: three out of four hydrogen-bond constraints between the R groups of Ser280, Tyr314, His440, and Tyr464 and the hydrogen-bond acceptor groups of the compounds to be docked must be satisfied,
2. ADORA2A: one hydrogen-bond constraint between the R group of Asn253 and the hydrogen-bond acceptor groups of the ligands,
3. D3R: one hydrogen-bond constraint between the R group of Asp110 and the hydrogen-bond donor group of a protonated nitrogen in the ligand.

The pattern was defined using the SMARTS string: `"[#1][#7+]"`.

The Glide Standard Precision (SP) [Friesner et al., 2006] protocol was used for docking calculations. Each ligand was docked with the van der Waals radii of atoms having partial charges ≤ 0.15 scaled by a factor of 0.9. For each ligand, only the best-scoring pose was retained. All calculations were carried out using the OPLS4 force field.

B Biological Evaluation

B.1 GSK-3β kinase assay

The GSK-3β kinase assay was performed as described in [Di Martino et al., 2025]. Briefly, the inhibitory potency against human recombinant GSK-3β (Carna Biosciences) was evaluated using the LANCE® Ultra time-resolved fluorescence resonance energy transfer (TR-FRET) assay (PerkinElmer), by measuring phosphorylation of the specific substrate human muscle glycogen synthase (ULight-GS (Ser641/pSer657)), according to the manufacturer's instructions. Test compounds, staurosporine (reference compound), or DMSO (control) were mixed with the enzyme (2 nM) in a buffer containing 50 mM HEPES (pH 7.5), 1 mM EGTA, 10 mM $MgCl_2$, 2 mM DTT, and 0.01% Tween-20. The reaction was initiated by adding 50 nM of the substrate ULight-PASVPPSPSLSRHSSPHQ(pS)ED and 3 μM ATP, followed by incubation for 1 hour at 23 °C.

Following incubation, the reaction was stopped by adding 8 mM EDTA. After 5 min, the anti-phospho-GS antibody labeled with europium chelate was added. After 1 more hour, the kinase reaction was monitored by irradiation at 320 nm, and fluorescence was measured at 615 nm and 665 nm using the EnVision 2014 Multilabel Reader (PerkinElmer). The calculated signal ratio at 665/615 nm was proportional to the extent of ULight-GS phosphorylation.

Compounds were screened at three concentrations (1, 10, and 50 μM). The most potent compound was further tested in a dose-response format at 11 concentrations ranging from 300 pM to 100 μM in technical triplicates. Results were expressed as percent inhibition of control enzyme activity.

B.2 Analysis of the Biological Data

Three-concentration screening and dose-response curves were run in at least three independent experiments, each performed in technical triplicates. IC_{50} values were determined by non-linear regression analysis of log[concentration]–response curves, generated from mean replicate values using a four-parameter Hill equation curve fit in GraphPad Prism 8 (GraphPad Software Inc., CA, USA).

Compounds 1, 2, and 3 present high TD values of 0.64 to ChEMBL3652544 (4), 0.61 to ChEMBL250386 (5), and 0.65 to ChEMBL359554 (6), respectively. In addition, we extended the chemical similarity analysis to all kinase inhibitors present in the ChEMBL database (see Supplemental Materials & Methods section for details). This step was motivated by the conserved nature of kinase binding pockets, where chemically similar scaffolds can often exhibit non-selective interactions. Remarkably, all selected compounds show a minimum TD value above 0.5, even when compared against the full kinase inhibitor set. Specifically, compound 1 presents a minimum TD of

0.51 to ChEMBL3663033 (7), compound 2 to ChEMBL326596 (TD = 0.58, 8), and compound 3 to ChEMBL578061 (TD = 0.63, 9). Representative structures of the most similar GSK-3β and kinase inhibitors to our selected compounds are shown in Figure 2A.

To evaluate the ability of the selected compounds to modulate GSK-3β kinase activity, their inhibitory effects were assessed at three concentrations (1, 10, and 50 μM) using the LANCE® Ultra time-resolved fluorescence resonance energy transfer (TR-FRET) assay, as described in the Biological Evaluation section. Compounds 3 and 2 exhibited weak (5 $\pm$ 0.7%) to moderate (49.6 $\pm$ 0.9%) inhibition of GSK-3β at the highest tested concentration (50 μM), respectively. In contrast, compound 1 showed strong inhibition already at 1 μM and was therefore selected for IC_{50} determination (Table 5). The IC_{50} of compound 1 was calculated to be 314 $\pm$ 15.3 nM (Figure 2B).

Table 5: Percentage of inhibition of the selected compounds at three different concentrations performed through a TR-FRET assay.

	1 μM	10 μM	50 μM
1	71.3 $\pm$ 1.7	88.2 $\pm$ 0.1	89.9 $\pm$ 0.2
2	1 $\pm$ 3.6	12.4 $\pm$ 0.8	49.6 $\pm$ 0.9
3	n.i.	n.i.	5 $\pm$ 0.7

The 1H-pyrazolo[3,4-b]pyridine moiety of the potent hit compound 1 anchors the molecule to the hinge region of GSK-3β through two key hydrogen bonds with the backbone of Asp133 and Val135. Notably, this scaffold is rare among GSK-3β inhibitors, with only 19 instances out of 1734 (1.1%), and similarly underrepresented among all kinase inhibitors in ChEMBL (727 out of 100365, 0.72%). The high potency of compound 1 may also be explained by an additional hydrogen bond between the nitrogen atom of its 2-methylpyridine ring and the side-chain amino group of Lys85, a conserved residue known to enhance inhibitor affinity. These findings highlight the model's ability to generate chemically novel yet functionally relevant hits. Despite not being explicitly fine-tuned on GSK-3β ligands, MolRNN successfully designed a potent hit by leveraging general kinase features embedded in the REINVENT training dataset. This is further supported by the t-SNE analysis (Figure 2C), where compound 1 (red star), although structurally novel, clusters within the dense chemical space of known GSK-3β binders (gold dots).

In light of its strong inhibitory activity, we benchmarked compound 1 against known GSK-3β inhibitors to better assess its relative potency. Specifically, we compared its activity value, expressed as –log(activity), with those of known GSK-3β inhibitors. As shown in the violin plot (Figure 2D), compound 1 (red star) outperforms both the full ligand set (light blue shape) and the subset of hit-like GSK-3β inhibitors (blue dots), with a –log(activity) value of 6.50 compared to median values of 6.39 and 5.89, respectively.

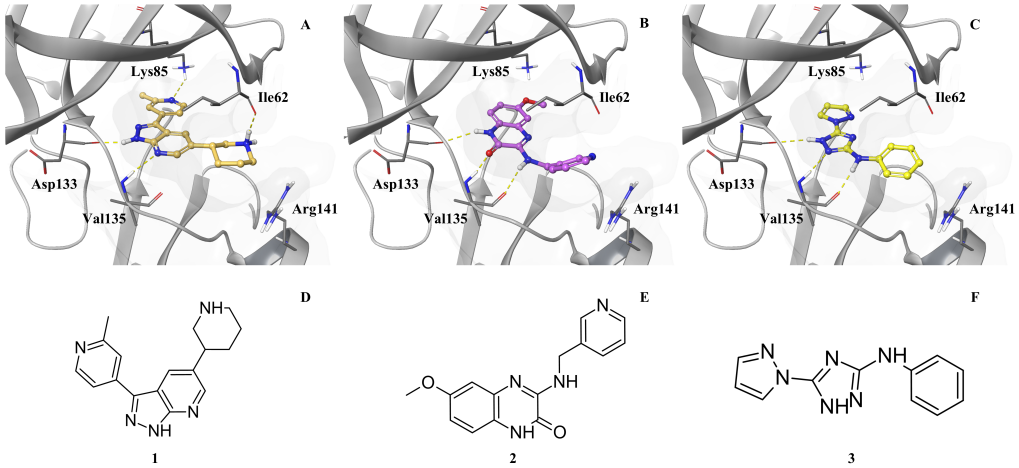

Figure 3: Predicted binding conformations of compounds 1 (carbon atoms in gold, A), 2 (carbon atoms in purple, B), and 3 (carbon atoms in yellow, C) at the binding site of GSK-3β. In panels A–C, the protein structure is shown as a thin grey ribbon. Residues interacting with the docked compound are displayed in stick representation with light grey carbons and explicitly labelled. Hydrogen bonds are depicted as yellow dashed lines, while a grey mesh highlights the boundaries of the binding pocket within 5 Å of each ligand. The chemical structures of compounds 1, 2, and 3 are shown in panels D–F, respectively.

C Per-target Docking Results and Analysis and Physicochemical Profiles

Table 6: KL divergence between the docking score distributions of generated molecules and target ligand sets for each of the seven investigated targets. Lower values indicate higher similarity between generated and reference distributions.

Target	KL divergence
ADORA2A	0.027 ± 0.015
D3R	0.055 ± 0.040
GSK-3β	0.019 ± 0.012
HSP90α	0.039 ± 0.030
PPARα	0.098 ± 0.074
SRC	0.122 ± 0.089
Thrombin	0.084 ± 0.063

Table 7: For each target, the median docking scores of the full and Hit-like ligand sets, along with the percentage of generated molecules (across all trained models) scoring better than each respective median.

Experiment	Target	Median (Full)	Median (Hit-like)	% Better (Full)	% Better (Hit-like)
DiGress (Hit-like)	ADORA2A	−8.89	−8.86	1.63	1.79
	PPARα	−9.15	−6.92	0.38	23.28
	HSP90α	−6.03	−6.55	39.97	17.78
	GSK-3β	−7.27	−7.83	10.13	3.26
	Thrombin	−7.59	−6.10	0.98	35.38
	SRC	−7.71	−7.55	5.23	7.34
	D3R	−6.81	−6.46	13.34	23.61
DiGress (REINVENT)	ADORA2A	−8.89	−8.86	2.12	2.29
	PPARα	−9.15	−6.92	0.35	21.42
	HSP90α	−6.03	−6.55	42.69	19.74
	GSK-3β	−7.27	−7.83	12.18	4.38
	Thrombin	−7.59	−6.10	1.22	40.20
	SRC	−7.71	−7.55	6.74	9.32
	D3R	−6.81	−6.46	12.45	21.72
DiGress (Hit fine-tune)	ADORA2A	−8.89	−8.86	2.23	2.47
	PPARα	−9.15	−6.92	0.54	25.34
	HSP90α	−6.03	−6.55	40.22	17.57
	GSK-3β	−7.27	−7.83	11.60	3.94
	Thrombin	−7.59	−6.10	1.31	41.16
	SRC	−7.71	−7.55	6.44	8.96
	D3R	−6.81	−6.46	14.12	24.44
MolRNN (Hit-like)	ADORA2A	−8.89	−8.86	2.07	2.27
	PPARα	−9.15	−6.92	0.38	22.86
	HSP90α	−6.03	−6.55	42.07	19.50
	GSK-3β	−7.27	−7.83	15.58	6.35
	Thrombin	−7.59	−6.10	1.16	36.95
	SRC	−7.71	−7.55	7.01	9.71
	D3R	−6.81	−6.46	13.05	22.53
MolRNN (REINVENT)	ADORA2A	−8.89	−8.86	2.64	2.86
	PPARα	−9.15	−6.92	0.46	22.40
	HSP90α	−6.03	−6.55	43.05	20.32
	GSK-3β	−7.27	−7.83	15.60	6.39
	Thrombin	−7.59	−6.10	1.52	41.89
	SRC	−7.71	−7.55	7.75	10.62
	D3R	−6.81	−6.46	13.48	22.81
MolRNN (Hit fine-tune)	ADORA2A	−8.89	−8.86	2.90	3.17
	PPARα	−9.15	−6.92	0.52	24.91
	HSP90α	−6.03	−6.55	51.82	25.16
	GSK-3β	−7.27	−7.83	20.59	8.83
	Thrombin	−7.59	−6.10	1.55	45.29
	SRC	−7.71	−7.55	9.46	12.89
	D3R	−6.81	−6.46	15.89	27.28
GraphINVENT (Hit-like)	ADORA2A	−8.89	−8.86	2.15	2.35
	PPARα	−9.15	−6.92	0.38	22.47
	HSP90α	−6.03	−6.55	38.35	17.13
	GSK-3β	−7.27	−7.83	14.00	5.44
	Thrombin	−7.59	−6.10	1.36	39.10
	SRC	−7.71	−7.55	6.67	9.33
	D3R	−6.81	−6.46	12.45	22.25
GraphINVENT (REINVENT)	ADORA2A	−8.89	−8.86	1.95	2.13
	PPARα	−9.15	−6.92	0.28	19.37
	HSP90α	−6.03	−6.55	39.78	17.95
	GSK-3β	−7.27	−7.83	13.38	5.38
	Thrombin	−7.59	−6.10	1.19	37.98
	SRC	−7.71	−7.55	6.43	8.90
	D3R	−6.81	−6.46	11.08	20.15
GraphINVENT (Hit fine-tune)	ADORA2A	−8.89	−8.86	2.34	2.54
	PPARα	−9.15	−6.92	0.34	21.21
	HSP90α	−6.03	−6.55	39.07	17.55
	GSK-3β	−7.27	−7.83	14.86	6.04
	Thrombin	−7.59	−6.10	1.42	39.79
	SRC	−7.71	−7.55	7.18	9.93
	D3R	−6.81	−6.46	11.92	20.97

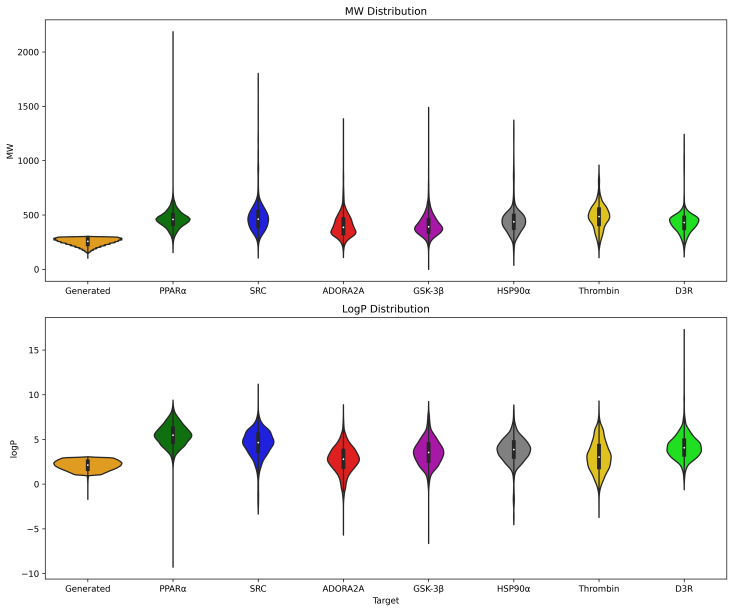

Figure 4: Comparison of MW (upper panel) and logP (lower panel) of the generated molecules with respect to each target ligand set.