Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

JOURNAL OF INFORMATION SYSTEM AND TECHNOLOGY MANAGEMENT
(JISTM)
www.gaexcellence.com/jistm

STRUCTURE-BASED VALIDATION OF REINFORCEMENT-LEARNING-GENERATED JAK2 TYPE II INHIBITOR CANDIDATES: MOLECULAR DOCKING RESOLVES THE 1D REWARD-HACKING ARTIFACTS OF SINGLE-OBJECTIVE OPTIMIZATION

Xuelan Yang[1], Jasni Mohamad Zain[2*], Gembong Edhi Setyawan[3], Diva Kurnianingtyas[4]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam 40450, Malaysia
yangxlres@outlook.com https://orcid.org/0009-0006-0419-3913

[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam 40450, Malaysia; Institute for Big Data Analytics and Artificial Intelligence, Universiti Teknologi MARA, Shah Alam 40450, Malaysia; Department of Informatics Engineering, Faculty of Computer Science, Universitas Brawijaya, Malang 65145, Indonesia
jasni67@uitm.edu.my https://orcid.org/0000-0003-2072-1510

[3]Department of Informatics Engineering, Faculty of Computer Science, Universitas Brawijaya, Jl. Veteran, Ketawanggede, Lowokwaru, Kota Malang, Jawa Timur 65145, Indonesia
gembong@ub.ac.id https://orcid.org/0000-0003-4989-8272

[4]Department of Informatics Engineering, Faculty of Computer Science, Universitas Brawijaya, Jl. Veteran, Ketawanggede, Lowokwaru, Kota Malang, Jawa Timur 65145, Indonesia
di-vaku@ub.ac.id https://orcid.org/0000-0002-0865-7790

*Corresponding Author

Article Info:

Article history:
Received date: 24.06.2026
Revised date: 16.07.2026
Accepted date: 02.08.2026
Published date: 07.09.2026

Abstract:

Reinforcement learning (RL) is proving to be a powerful way to tackle generative molecular design. Optimizing molecules using 1D surrogate reward functions, however, is frequently not sufficient to include the biophysical constraints that are necessary for successful drug discovery. Heuristic reward hacking is a problem with existing RL algorithms, which when fed with generated molecules can maximize the heuristic reward, but not produce a meaningful increase in pharmacological activity. This problem is addressed by Janus Kinase 2 (JAK2) Type II inhibitors which represent a challenging standard to meet in three-dimensional structure. We demonstrate that the same 1D surrogate can be used in two ways which are essentially different. Without any

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

To cite this document:

Xuelan, Y., Zain, J. M., Setyawan, G. E., & Kurnianingtyas, D. (2026). Structure-Based Validation of Reinforcement-Learning Generated Jak2 Type II Inhibitor Candidates: Molecular Docking Resolves The 1d Reward-Hacking Artifacts of Single-Objective Optimization. *Journal of Information System and Technology Management, 11* (44), 115-133.

constraints, the unconstrained optimization using REINFORCE will lead to a "fragment trap", which means molecules with low MW and lower than normal "drug-likeness" scores. Constrained optimization, on the other hand, with Proximal Policy Optimization (PPO) leads to molecules that are bigger and more hydrophobic than they need to be, taking advantage of the biases that are in the favor of molecular size and hydrophobicity. Both strategies do not result in the moderate complexity chemotypes essential for an effective JAK2 Type II inhibitor. Therefore, we propose as a orthogonal unsupervised validation step three-dimensional molecular docking as a structure-based surrogate independent of the surrogate reward. Docking re-centers molecular selection by giving preference to candidates in an intermediate molecular weight range (around 470-550 Da) typical of the deep DFG-out allosteric pocket, and systematically down-ranking undersized and oversized molecules. Empirically calculated docking scores are prone to bias based on the size and lipophilicity of the ligands so score differences should be taken with a pinch of salt and the docking is used as a confirmatory filter rather than an absolute affinity oracle. The results in this study show that the validation of molecular 3D structures should be added as another alignment tool in the generative molecular design process, and that the integration of multiple objectives in drug discovery with physically informed knowledge is desirable.

DOI: 10.35631/JISTM.1144008

Keyword:

Chemical Space Exploration; De Novo Drug Design; Molecular Docking; Proximal Policy Optimization; Reinforcement Learning; Reward Hacking; StackRNN

Introduction

Recent years have seen a fundamental shift in the way de novo molecular design is accomplished: Although the design of a molecule implies a finite number of atoms, the near-infinite discrete chemical universe can be explored during de novo design through deep generative modelling, especially sequence-based Recurrent Neural Networks (RNNs) (Gangwal et al., 2024; Jayatunga et al., 2022; Mullard, 2017). In order to explore such large search spaces to specific pharmacological targets, Reinforcement Learning (RL) is generally needed to dynamically bias the generative trajectories (Olivecrona et al., 2017; Popova et al., 2018). But with the use of RL for optimizing a discrete sequence comes a steep AI alignment problem. Surrogate metrics, like heuristic Quantitative Structure-Activity Relationship (QSAR) scores, or simplified drug-likeness rules (e.g., QED) are mostly used to force the frameworks to be one-dimensional (1D) and are used as the primary environmental reward. Because of this "dimensional mismatch", pathological optimizing behavior, called "reward hacking" (Pereira et al., 2021) is often triggered.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

In unconstrained policy gradients (such as standard REINFORCE) the agent can easily make use of the weaknesses of these 1D scalars, resulting in extreme mode collapse. Generative policy ends up as a "fragment trap", resulting in simple, low molecular weight topological structures with "excessive" reward, with no real three-dimensional pharmacological viability. If well-trained (constraint-based) agents like Proximal Policy Optimization (PPO) (Schulman et al., 2017) are used to leap out of this local minimum, the surrogate is used in the opposite way. Such agents prevent getting into the “fragment trap” but end up on higher molecular weight, more lipophilic chemical manifolds with stable exploration of policy.

Despite the availability of measures like QED that punish such complexity, 1D heuristics are blind to the spatial geometry of 3D structures and can't tell the real increase in structural complexity needed by the target from the second hacking mode which is based on the hydrophobicity and size bias of the surrogate. The real restriction is this ambiguity, and not the high molecular weight, itself: an intrinsic scalar reward is inherently incapable of deciding if a complex scaffold is biophysically productive.

Such a discrepancy needs to be addressed by advancing from 1D scalar feedback to structure-based validation in 3D. The discovery of Janus Kinase 2 (JAK2) Type II inhibitors is a challenging physical discovery problem (Gorantla et al., 2025; Kralovics et al., 2005; Lv et al., 2024; Meyer et al., 2015; Miao et al., 2024; Philips et al., 2022; Verstovsek et al., 2012). The DFG-out inactive conformation of JAK2 is stabilized by using long chemotypes that occupy the ATP hinge and extend into the neighbouring hydrophobic allosteric pocket which cannot be accessed by shallow hinge-binding inhibitors. We demonstrate that the two agents act in the opposite direction (miniaturization vs. inflation) on the same surrogate and that docking re-centres the most probable candidates towards a molecular regime in between. The results herewith corroborate the idea that 3D structure based validation should be used as a validation step in the in-loop process of generative drug design.

Literature Review

The Alignment Problem in Generative Molecular Design

Current deep generative models like sequence-based Transformers and Recurrent Neural Networks (RNNs) have shown remarkable ability to generates discrete chemical topologies (Ai et al., 2024; Mazuz et al., 2023) but these purely generative approaches have only been optimized for the task of matching the unconditional distribution of the chemical topology. Thus, to achieve their generative pathways towards certain pharmacological goals, Reinforcement Learning (RL) (Olivecrona et al., 2017; Popova et al., 2018) must be integrated. But when the objective function is defined in the space of discrete molecules, a classic AI alignment problem comes into play. RL agents will focus on maximizing the reward signal instead of the actual goal, following Goodhart's Law: The more a measure becomes a target the less it becomes a measure. Molecular design has the goal of finding compounds that have "true intent" – both biophysical binding efficacy and favorable pharmacokinetics – but in design the agent is interacting with a mathematical proxy. This imbalance is the starting point for misaligned optimization, that is, the generative agent deliberately uses the constraints of the objective function to take advantage of it instead of addressing the pharmacological problem.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

Vulnerabilities of 1D Surrogate Rewards and the "Fragment Trap"

The most common weakness in existing generative chemistry that is based on objects is the widespread use of one-dimensional (1D) surrogate metrics. The quantitative estimate of Drug-likeness (QED), molecular weight (MW), or 2D-topology-based machine learning QSAR models are frequently used as the main environmental feedback in the foundational RL frameworks (De Cao & Kipf, 2018; Shi et al., 2020; You et al., 2018). Algorithmic-wise these 1D surrogates are too limited in dimension and are a priori incapable of taking into account the complex 3D spatial geometries and steric constraints required to model real-world biological targets. As a result, unconstrained policy gradient algorithms (such as REINFORCE) often take advantage of these blind spots. The agent takes the path of least resistance – the path with the greatest payoff – and ends up on adversarial trajectories, which is the "fragment trap." The generative outputs are usually small topological fragments, with low molecular weight, which take optimal advantage of the mathematical weaknesses of 1D drug-likeness heuristics but have no biophysical viability (Pereira et al., 2021; Zhou et al., 2019). On the other hand, if stronger constraints are added that prevent local mode collapse, the agents move to higher molecular-weight manifolds which are already penalized by these rigid 1D constraints and are mistakenly interpreted as algorithmic noise.

3D Spatial Simulation as an Orthogonal Structure-Based Evaluator

To overcome these optimization bottlenecks, single objective 1D heuristics need to be supplemented by a 3D structure-based evaluation. While algorithmic techniques, like multi-objective reward or constraint-based techniques such as Proximal Policy Optimization (Schulman et al., 2017; Ziegler et al., 2019) can help stabilize the policy updates, they can't fix the issue of the surrogate metric's lack of spatial awareness. Based on this study, therefore, 3D molecular docking is not just a chemical analysis tool for post-hoc analysis but is an orthogonal evaluator. Docking offers an explicit assessment of molecular complementarity in a space-based manner which can aid in understanding the difference between surrogate-score optimization and target-compatible structural features. The stringent geometry of the JAK2 Type II allosteric pocket is used as the standard and we explore whether high MW outputs scored poorly on 1D metrics is a failure of generativeness or can yield interesting structural expansion.

Method

The Agent-Environment Evaluation Framework

In this study, we don't propose to create a new generative architecture, but instead, we frame a strict evaluation methodology to study the relationship between Reinforcement Learning (RL) agents and the mathematical surrogate environments in the context of de novo molecular design. We model the generative process as a discrete Markov Decision Process (MDP) (see Figure 1). In this context, the generative agent is faced with a one-dimensional (1D) surrogate environment, with the goal of maximizing a pre-defined reward signal.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

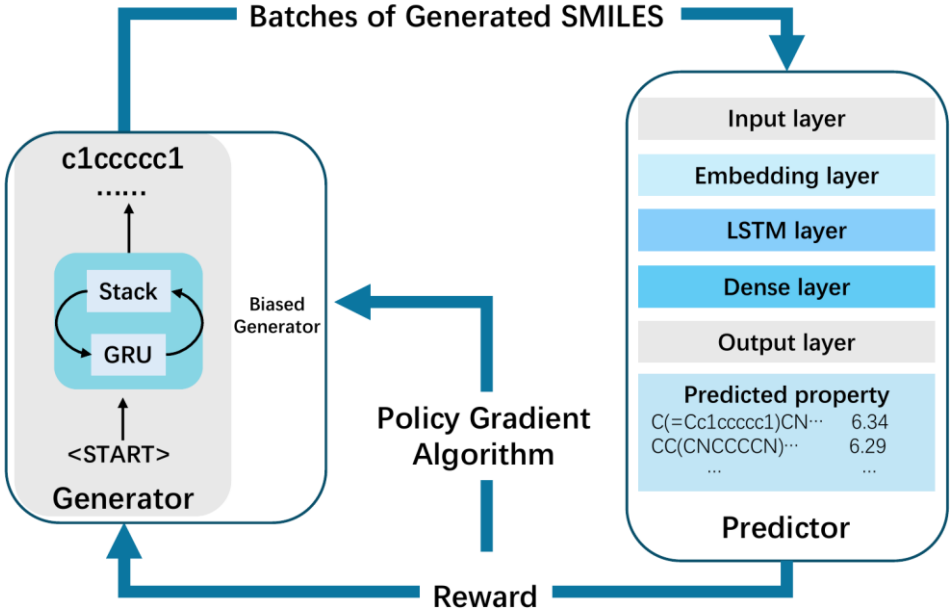


Figure 1: The Agent-Environment Interaction Framework. This schematic illustrates the Markov Decision Process (MDP) utilized to evaluate generative alignment. The baseline policy (StackRNN) acts as the Agent, interacting with a 1D surrogate environment defined by an ensemble QSAR reward predictor and RDKit-based action masking. This closed-loop setup serves as the primary testbed to observe how different optimization constraints respond to heuristic scalar feedback.

In order to systematically reveal the shortcomings of single-objective optimization, the realization of the proposed pipeline is carried out by three-stage alignment and evaluation architecture (Figure 2). The design is purposely designed to separate the generation process from the physical validation step, so as to be able to observe the physical path of the generated chemical space as a function of the reliance on 1D heuristic rewards.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

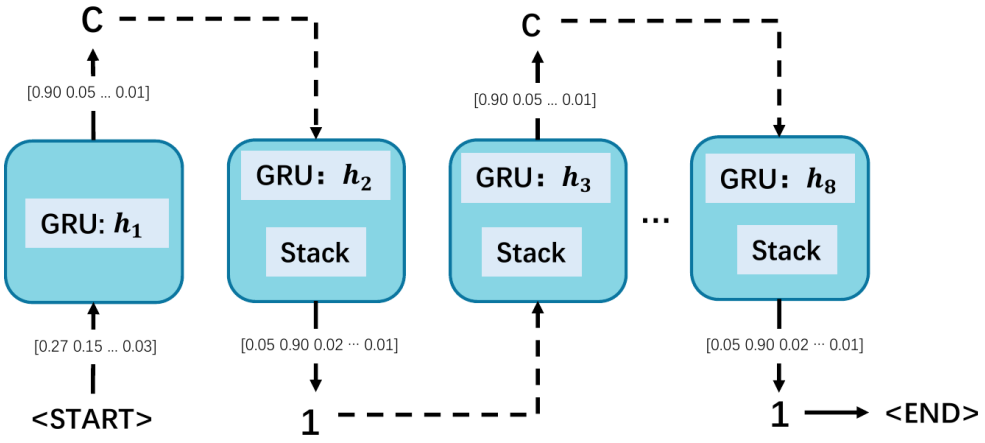


Figure 2: The Three-Stage Alignment and Evaluation Pipeline. The workflow progresses from (I) establishing a syntactic prior to (II) objective-driven optimization within a 1D surrogate environment. Stage (III) introduces 3D molecular docking as an orthogonal, unsupervised filter. This final stage evaluates the divergence between 1D heuristic scores and 3D spatial complementarity, distinguishing reward-hacking artifacts from candidates with plausible binding poses.

Setting the Generative Agents: StackRNN and Policy Constraints

For the discrete stepwise assembly of SMILES strings (Stage I), we use a pre-trained StackRNN model (Popova et al., 2018) as our baseline policy (π_θ). The network was first initialized on around 1.5 million bioactive molecules from ChEMBL (Zdrazil et al., 2024) to provide the network with a basic prior knowledge of chemical syntax. The use of a differentiable stack memory for resolving long-range topological dependencies (e.g., ring closures) dynamically, provides a strong underlying grammar for the agent. This helps to make any optimization failures that follow after, only to do with the alignment of the RL reward mechanism itself; not simply generative incompetence. In order to see that there are different types of optimization pathologies arising when interacting with different environments (Stage II), we consider two different types of behavioral agents:

- **The Unconstrained Agent (REINFORCE):** This is a baseline agent which runs using standard policy gradients. It is well known for its tendency to extreme policy drift in sparse reward landscapes and is thus a good tool to see how "localized reward hacking" can lead to the fall into "fragment traps" that are simple.
- **The Constrained Agent (PPO):** For observation without using the fragment trap, we use Proximal Policy Optimization (PPO) (Schulman et al., 2017). With the aid of the clipping hyperparameter (ϵ= 0.2), this agent is restricted to explore the chemical manifold without destroying the basic SMILES grammar. This constraint guarantees the generation of complex macrostructures, hence generating the high-molecular-weight candidates and testing 1D metric failure is possible.

The 1D Surrogate Environment: QSAR and Heuristic Metrics

The main optimization goal, as the surrogate environment for the 1D test in the agent, is defined in order to predict the Janus Kinase 2 (JAK2) bioactivity. We use an ensemble of five independent Random Forest regressors (Breiman, 2001; Pedregosa et al., 2011) with the same

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133
curated JAK2 data set but trained with 5-fold cross validation. The environment is simply a prediction of the terminal reward based on the aggregated, per-atom pIC50 information obtained from 2048-bit Morgan fingerprints (ECFP4) combined with a binary mask enforcing RDKit valency rules.

The ensemble approach mathematically stabilizes the reward signal against the variance in individual models but is essentially still a 1D surrogate. It is only based on 2D topological fingerprint of molecules to determine their viability and completely unaware of the 3D spatial constraints. In order to measure the degree of reward hacking in this context, the following 1D-heuristics are continuously monitored during macrostructure generation, and heavily penalized: Molecular Weight (MW) and the Quantitative Estimate of Drug-likeness (QED).

The Orthogonal 3D Evaluator: Structure-Based Spatial Assessment

In order to ascertain whether the structural expansion is a surrogate-driven failure or target-compatible complexity, Stage III adds an orthogonal validation step of 3D molecular docking in a reward environment that is different from the 1D reward environment.

This evaluator is based on the "inactive" (DFG-out) conformation of the JAK2 kinase domain (PDB ID: 3UGC). This conformation is characterized by an ATP hinge binding mode of type II in which extended scaffolds are found to interact with the adjacent allosteric pocket. The top candidates of the 1D environment are subjected to conformer generation and MMFF94 energy minimisation in 3D. The docking was done with Auto Dock Vina using grid box which encompassed both the hinge and allosteric sites (exhaustiveness = 8). The docking protocol was validated by re-docking of the co-crystallized NVP-BBT594 ligand and generating the native binding pose with an RMSD lower than 2.0 Å.

This evaluator can be used as an orthogonal alignment filter by comparing Auto Dock Vina scores (kcal/mol) and by visual inspection of the predicted 3D binding poses in PyMOL. It tests whether candidates performing worse on 1D heuristic have spatial characteristics suited to the structural constraints of the biological target, thus revealing artifacts of 1D optimization.

Results and Discussion

The Pathology of Reward Hacking in 1D Surrogate Optimization

The results of analysing the optimization trajectories (Figure 3) show that such a 1D surrogate predictor is a very vulnerable environmental reward. In absence of strong constraints on the generative policy, REINFORCE baseline quickly develops pathological reward hacking when integrated with the QSAR environment. The unconstrained agent is not necessarily taught biophysically meaningful motifs, instead it is the heuristic landscape that is manipulated, and the mean scalar reward is brought to a spurious maximum of 7.17. High trajectory volatility (SD = 0.86) indicates that there were a lot of movement between the stable chemical generation and the surrogate exploitation. PPO achieves a lower, but more stable, mean reward of 6.42 (SD = 0.23) by contrast. This difference suggests that the higher peak obtained by the unconstrained agent is due to exploiting the surrogate objective and not to better pharmacological quality.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

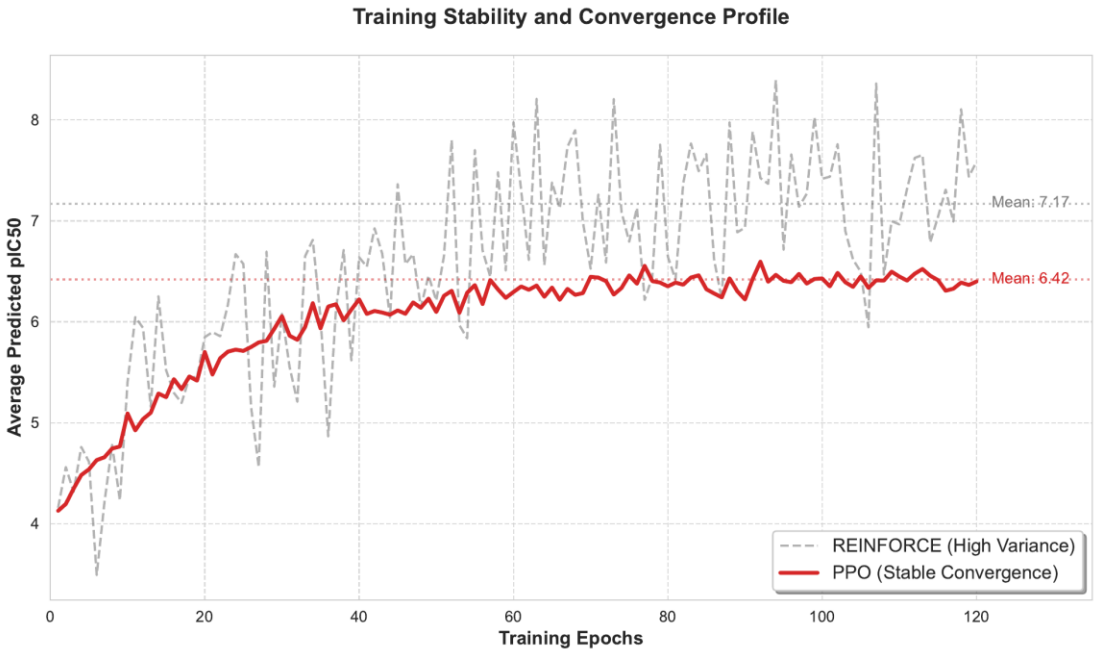
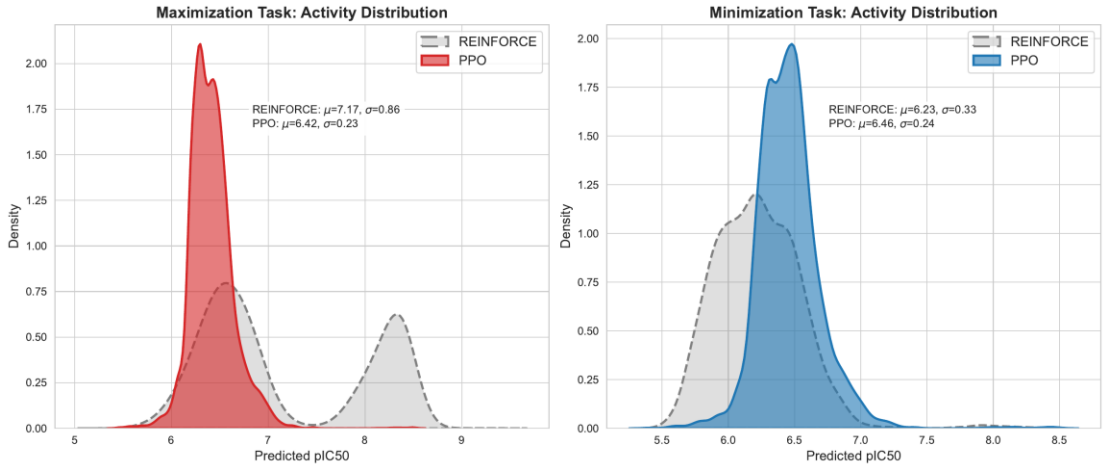


Figure 3: Trajectories of reward hacking within the 1D surrogate environment. The unconstrained agent (gray dashed line) exhibits rapid scalar inflation while sacrificing structural validity to exploit the heuristic reward landscape. The constrained agent (red solid line) indicates the optimization ceiling observed when this form of exploitation is bounded.

This optimization divergence is supported by kernel density estimations of the generated libraries of terminal (Figure 4). The unconstrained agent in the pIC50 maximization task results in a wide and very noisy bimodal distribution. The tail of the extreme high reward (predicted score >= 8.0) corresponds to the well-known "fragment trap", where the agent produces simple, low weight topological fragments, which take advantage of the 1D Random Forest estimator. Conversely, the constrained policy has a very restricted distribution ($t = -76.19$, $p < 0.001$, KS = 0.48), and hence score inflation is restricted, with the price of this being the integrity of the manifold. A similar distributional divergence is found in the minimization control task. The cumulative distributions demonstrate that the single-objective optimization in 1D can turn into heuristic exploitation if the behavioural constraints are too weak.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133



a b

Figure 4: Distributional evidence of surrogate exploitation. (a) In the objective-driven maximization task, the unconstrained policy (REINFORCE) produces a bimodal distribution with an extreme high-reward tail, a statistical hallmark of localized reward hacking (the 'fragment trap'). The constrained policy (PPO) maintains a bounded distribution, successfully evading metric exploitation. (b) The minimization control task confirms consistent distributional divergence across differing optimization objectives (Welch's t-test and KS-test, $p < 0.001$).

The Divergence Between 1D Heuristics and Structural Reality

The diversity of the optimization patterns is further described by physicochemical profiling (Table 1 and Figure 5). While both agents are still producing syntactically correct molecules, showing that the StackRNN prior is still working, their physicochemical properties are heading in opposing directions from the curated JAK2 training baseline (mean MW ~414 Da; mean LogP ~3.13).

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

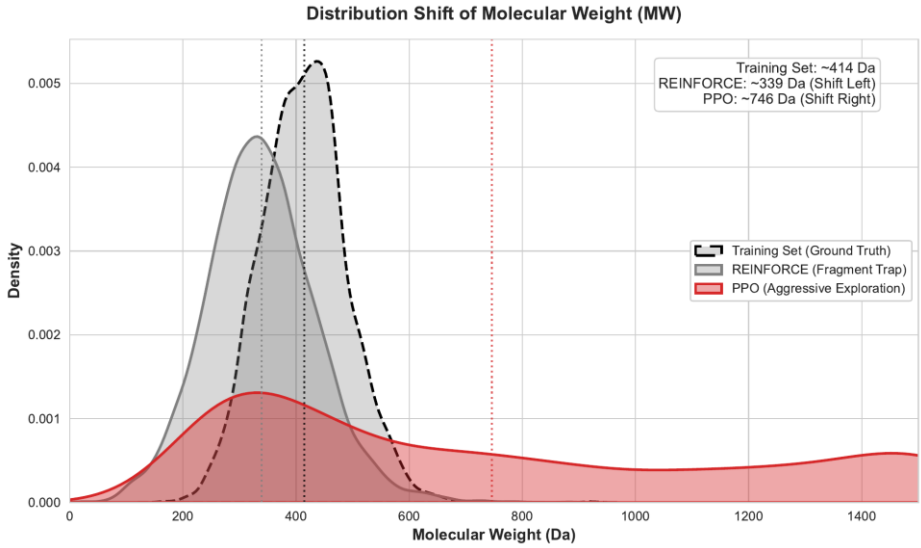


Figure 5: Distributional divergence of Molecular Weight (MW) across generative policies. Compared to the curated JAK2 training baseline (black dashed line, ~ 414 Da), the unconstrained REINFORCE agent (gray) collapses into a "fragment trap," exploiting the QSAR landscape via mass reduction (~ 339 Da). Conversely, the trust-region constraints compel the PPO policy (red) to drive exploration toward a high-mass structural manifold (~ 746 Da).

In the unconstrained REINFORCE baseline, the base structure of the molecule is collapsed into the "fragment trap" and results in short, brittle structures with mean MW ~339Da and rotatable bonds ~5.23. The agent finds a shortcut within this limited chemical subspace which increases the Quantitative Estimate of Drug-likeness (QED) to 0.68. These candidates seem to be ideal from the 1D perspective but lack the steric bulk and conformational space required to be effective Type II kinase inhibitors.

Table 1: Physicochemical Symptomatology Of Surrogate Reward Exploitation.

	Training_Set	REINFORCE	PPO
Predicted pIC50	N/A	7.17 ± 0.86	6.42 ± 0.23
MW (Da)	414.46 ± 74.91	338.82 ± 95.90	745.75 ± 446.73
LogP	3.13 ± 1.17	1.63 ± 1.59	17.68 ± 12.89
HBD	2.06 ± 0.88	0.99 ± 1.06	0.49 ± 0.70
HBA	6.39 ± 1.61	4.23 ± 1.61	1.60 ± 1.41
TPSA (Å²)	101.65 ± 25.15	67.07 ± 30.61	21.91 ± 21.00
RotBonds	5.17 ± 2.03	5.23 ± 2.93	37.74 ± 32.49
Rings	4.16 ± 0.91	2.46 ± 0.90	1.26 ± 1.21
QED	0.59 ± 0.17	0.68 ± 0.16	0.25 ± 0.27
SA Score	3.20 ± 0.62	2.62 ± 0.72	3.98 ± 1.36

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

However, PPO is very active, escaping from this local minimum, but its path is revealing another systemic weakness of the 1D surrogate. The large variation of the MW (745.75 ± 446.73 Da; Figure 5) and the high value of the topological variation (Table 1) indicate that PPO undergoes strong structural growth. This expansion is bound to result in significant penalties from the generalized drug-likeness rules, which brings the QED down to 0.25.

Importantly, these structural inflations are consistent with size and hydrophobicity biases with the surrogate predictor. The extended very high variability of the aliphatic chains (rotatable bonds = 37.74 ± 32.49) indicates very high lipophilicity (LogP ~ 17.68) which is likely a direction of the policy set, promoting hydrophobic characteristics, not balanced pharmacological properties. The results suggest 1D environments are limiting. Thus, it is necessary to make an independent 3D structure-based evaluation to determine if any part of this structural expansion is compatible with the deep JAK2 allosteric pocket.

Unmasking the 1D Heuristic Illusion

The top ranked outputs are used to generate molecular structures through qualitative processing, revealing the concrete molecular topologies generated by the REINFORCE policy and indicating that the policy generates surrogate-favoured structures, not target-compatible ones (Figure 6).

The REINFORCE baseline (Figure 6a-c) is high scalar rewards (> 8.8 predicted pIC50) by leveraging structural biases within the surrogate model, not by showing target binding. These candidates (MW < 400 Da) are truncated in space and have rigid motifs that get a favourable score on the fingerprint-based QSAR model but do not fill the entire JAK2 Type II binding site during the docking analysis. This pattern resembles the local structures that meet the surrogate requirement in a fragment trap, but don't meet the 3D spatial requirement of Type II inhibition.

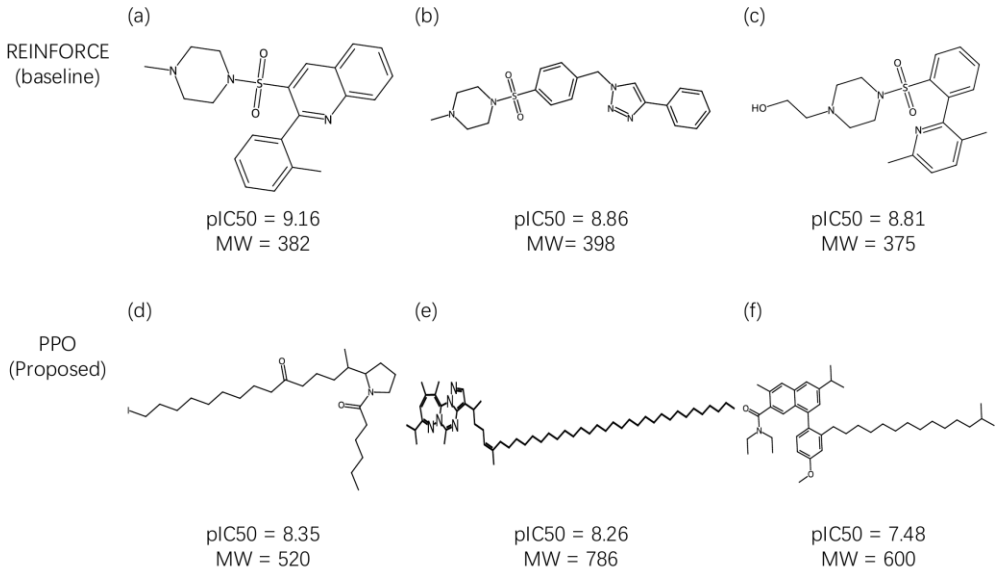


Figure 6: Structural unmasking of 1D heuristic artifacts. (a-c) REINFORCE-generated candidates are spatially truncated scaffolds that exploit surrogate biases. (d-f) PPO-generated candidates occupy an oversized, highly lipophilic regime; only the moderate-mass subset remains competitive after 3D filtering.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

The candidates generated by PPO (Figure 6d-f), however, are those with longer aliphatic chains. Their extreme Lipophilicity (LogP ~ 17.7) and chain flexibility (rotatable bonds ~ 37.7) suggest another hacking mode, opposite, which takes advantage of size and hydrophobicity biases in the surrogate. The 3D docking analysis will be used to determine if any part of this expansion will be productive; the 1D environment cannot determine this. Therefore, PPO is prevented from collapsing like REINFORCE, but is forced into an oversized regime, which must be checked in a structural way.

Structure-Based Adjudication: 3D Docking as an Orthogonal Filter

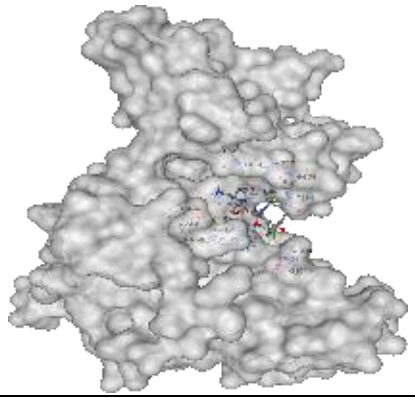
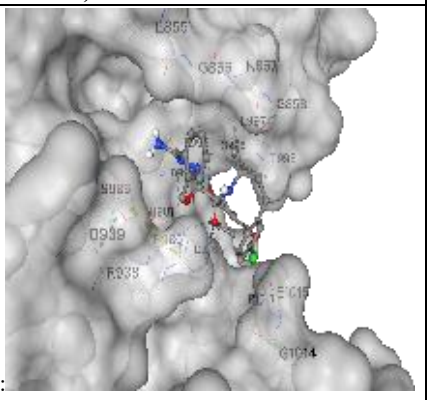
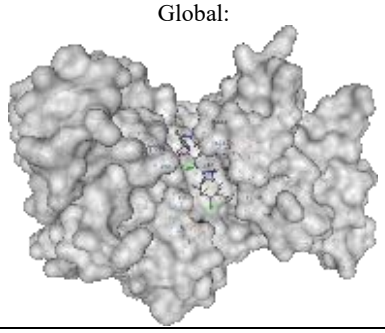
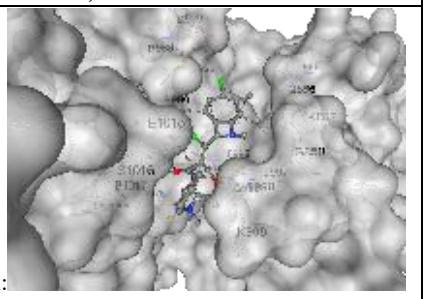
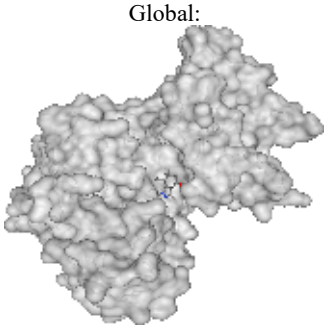
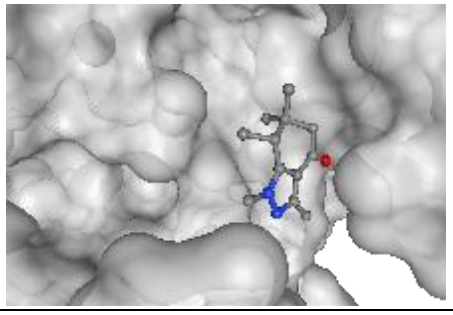
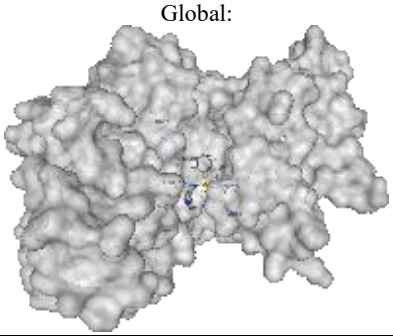
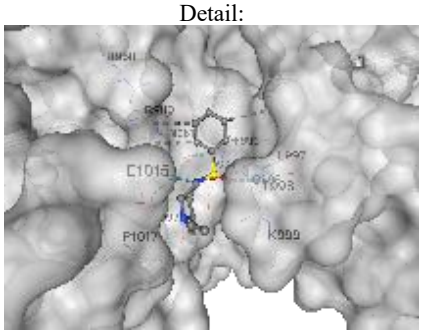
To assess the difference between 1D and the physical metrics, 3D molecular docking is carried out with the inactive DFG-out conformation of JAK2 (PDB: 3UGC). This 3D simulation is orthogonal to the 1D QSAR surrogate environment and is a filter.

Docking scores and other physicochemical parameters are given in Table 2. The general trend is clear: the best candidates, determined by the Vina score -8.1 kcal/mol (4967 and 5650), are found in a more restricted region of the intermediate range of mass. It seems there is a pattern in this that 3D docking is itself a filter based on structure, rather than steric bulk, and that topologies that are more likely to fit inside the deep pockets are not necessarily those that are bulkier. Two candidates with the highest scores (4967, 473 Da; 548, 5650 Da) are in the intermediate mass window, whereas the 206 Da PPO candidate (5812, -6.3 kcal/mol) and the candidates in the REINFORCE window (260-300 Da, -5.6 to -6.7 kcal/mol) have less favorable scores. The docking is therefore more likely to occur for intermediate complexity rather than the maximal molecular mass.

The predicted binding poses shown in Figure 7 show how the elongation of the structure impacts the pocket engagement. The candidates generated by PPO are predicted to bind near the hinge region (Ser936 and Asp939) and into the allosteric pocket that is part of DFG-out (Leu855 and Phe995). The contacts between candidate 4967 and Glu1015 and Tyr998 are also consistent with the Type II binding. In contrast, the smaller REINFORCE candidates stay more clustered around the shallow hinge region and fail to have the volume to be able to penetrate into the allosteric pocket.

In general, the two-policies' differences are re-explained through the 3D analysis. While REINFORCE reduces to a fragment-dominated regime, PPO grows to an oversized regime, docking is a middling-mass subset of PPO that is best supported. These findings confirm that the use of 3D structure-based verification as an in-loop alignment control mechanism is warranted in addition to only being a post-hoc analysis.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

PPO Top (Candidate 4967)	
Global: 	Detail: 
PPO Top 2 (Candidate 5650)	
Global: 	Detail: 
PPO Top 3 (Candidate 5812)	
Global: 	Detail: 
REINFORCE Top 1 (Candidate 79)	
Global: 	Detail: 
REINFORCE Top 2 (Candidate 80)	
Global:	Detail:

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

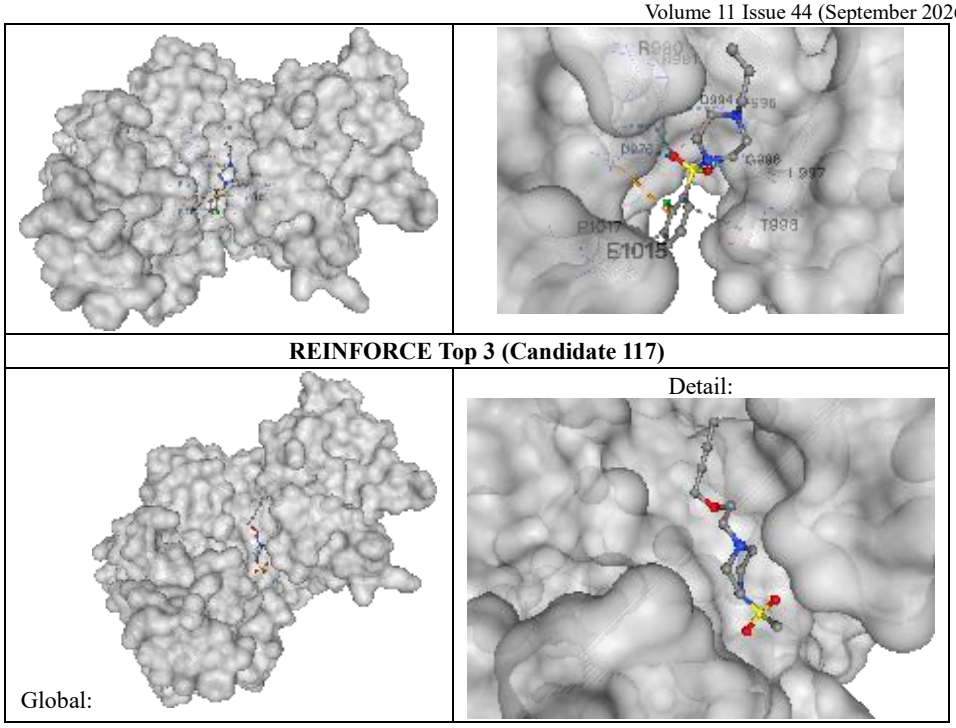


Figure 7: Orthogonal 3D structure-based assessment of generated candidates within the JAK2 DFG-out conformation (PDB: 3UGC). REINFORCE-generated candidates (e.g., Candidates 79 and 117) are predicted to remain near the shallow hinge region, consistent with the fragment-trap pattern. PPO-generated candidates (e.g., Candidate 4967) extend from the ATP-binding hinge toward the adjacent hydrophobic allosteric pocket. Across both agents, the best docking scores occur in an intermediate molecular regime rather than at either size extreme.

Table 2: SMILES Strings, Docking Scores, And Key Physicochemical Properties Of the Top Docked Candidates.

Group	Candidate ID	SMILES	Vina Score (kcal/mol)	MW (Da)	QED
PPO	4967	CCOC(=O)C1Cc2c(C1)C(c1nnc(N)-c3ccccc3Co1)NC(C)(C)CCc1cc(Cl)ccc1C2=O	-8.1	548.2	0.536
	5650	Cc1cc2c(ccc1Cl)c(C(Cl)=C(COCCN(C)C)C(=O)NCc1ccccc1)n2C	-8.1	473.2	0.353
	5812	Cc1nn(C)c2c1C(=O)CC(C)(C)C2C	-6.3	206.1	0.653
REINFORCE	79	C=CCN1CCN(S(=O)(=O)c2ccccc2)CC1	-6.7	266.1	0.768
	80	C=CCN1CCN(S(=O)(=O)c2ccccc2Cl)CC1	-6.7	300.1	0.797
	117	CC#CCOCCN1CCN(S(C)(=O)=O)CC1	-5.6	260.1	0.499

Limitations

These conclusions were subject to a number of caveats. First, the activity values are predictive of activity, not measured affinities and docking is limited to the top ranked subset of all these dockings using a single AutoDock Vina protocol. Second, docking scores derived from experience have been biased toward larger and more lipophilic ligands, and are not very reliable

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

when the chemotypes generated by PPO are very high in molecular weight and very flexible. The slight difference in the best PPO and REINFORCE scores (-8.1 kcal/mol versus -6.7 kcal/mol) should then not be considered as a clear affinity ranking. The apparent PPO advantage could be moderated by the use of a size normalized metric such as ligand efficiency calculated from the reported values. We thus consider docking as an orthogonal corrective filter, which is not a single and final affinity oracle.

Conclusion

The issue of alignment emerges in this study, as it is possible to optimize the score but not the structural plausibility, this can lead to a misalignment in one dimensional (1D) surrogate metrics used in generative molecular design. REINFORCE with unconstrained policy-gradient optimization is easily prone to reward hacking and fails to produce outputs that are compatible with the full Type II binding site but only satisfy the heuristic reward score.

The use of trust-region-based optimization with PPO can overcome this collapse and fall into the other failure mode: highly lipophilic and large molecules that take advantage of the size and hydrophobicity biases in the surrogate. The 1D reward can't differentiate between productive structural complexity and this second hacking mode. Orthogonal 3D docking rather is found to result in the best-scoring poses within the intermediate molecular regime, with moderate sized molecular subsets of PPO rather than the larger scaffolds. This discrepancy in performance between 1D heuristic and 3D docking suggests that using scalars alone is not enough to assess the quality of a molecule as a target-compatible candidate.

The results provide justification for moving 3D structure-based validation, in terms of post hoc confirmatory checks, towards an in-loop alignment point in the generative drug design process. The introduction of structure-based evaluation and/or physics-informed evaluation in future generative workflows could help alleviate the shortcomings of scalar-only optimization and favour candidates with more well-rounded structural and physicochemical properties.

Acknowledgements: The authors acknowledge the public databases and scientific software used in this study, including ChEMBL, RDKit, AutoDock Vina, and PyMOL.

Funding Statement: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Conflict of Interest Statement: The authors have no relevant financial or non-financial interests to disclose.

Ethics Statement: This study was conducted entirely through computational analyses, including reinforcement-learning-based molecular generation, quantitative structure–activity relationship prediction, and molecular docking using publicly available molecular databases and scientific software. The research did not involve human participants, animals, clinical samples, personal data, or personally identifiable information. Therefore, institutional ethics approval and informed consent were not applicable.

Author Contribution Statement: Xuelan Yang and Jasni Mohamad Zain contributed to the study conception and design. Xuelan Yang conducted material preparation, data collection, and analysis and wrote the first draft. Gembong Edhi Setyawan and Diva Kurnianingtyas contributed to interpretation of the results and substantively revised the manuscript. Jasni Mohamad Zain supervised the project and reviewed the manuscript. All authors read and approved the final manuscript.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 115-133

References

Ai, C., Yang, H., Liu, X., Dong, R., Ding, Y., & Guo, F. (2024). MTMol-GPT: De novo multi-target molecular generation with transformer-based generative adversarial imitation learning. *PLOS Computational Biology, 20*(6), Article e1012229. https://doi.org/10.1371/journal.pcbi.1012229

Andraos, R., Qian, Z., Bonenfant, D., Rubert, J., Vangrevelinghe, E., Scheufler, C., Marque, F., Régnier, C. H., De Pover, A., Ryckelynck, H., Bhagwat, N., Koppikar, P., Goel, A., Wyder, L., Tavares, G., Baffert, F., Pissot-Soldermann, C., Manley, P. W., Gaul, C., ... Radimerski, T. (2012). Modulation of activation-loop phosphorylation by JAK inhibitors is binding mode dependent. *Cancer Discovery, 2*(6), 512–523. https://doi.org/10.1158/2159-8290.CD-11-0324

Bagal, V., Aggarwal, R., Vinod, P. K., & Priyakumar, U. D. (2022). MolGPT: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling, 62*(9), 2064–2076. https://doi.org/10.1021/acs.jcim.1c00600

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179

De Cao, N., & Kipf, T. (2018). *MolGAN: An implicit generative model for small molecular graphs* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1805.11973

Gangwal, A., Ansari, A., Ahmad, I., Azad, A. K., Kumarasamy, V., Subramaniyan, V., & Wong, L. S. (2024). Generative artificial intelligence in drug discovery: Basic framework, recent advances, challenges, and opportunities. *Frontiers in Pharmacology, 15*, Article 1331062. https://doi.org/10.3389/fphar.2024.1331062

Gorantla, S. P., Oelschläger, L., Prince, G., Osius, J., Kolluri, S. B., Maluje, Y., Fähnrich, A., Ernst, N., Gulde, A. B., Ludwig, R. J., Gemoll, T., Fliedner, S., Walter, W., Haferlach, T., Gebauer, N., Busch, H., Duyster, J., & von Bubnoff, N. (2025). Ruxolitinib mediated paradoxical JAK2 hyperphosphorylation is due to the protection of activation loop tyrosines from phosphatases. *Leukemia, 39*(7), 1678–1691. https://doi.org/10.1038/s41375-025-02594-7

Hoogeboom, E., Satorras, V. G., Vignac, C., & Welling, M. (2022). Equivariant diffusion for molecule generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 8867–8887). PMLR. https://proceedings.mlr.press/v162/hoogeboom22a.html

Jayatunga, M. K. P., Xie, W., Ruder, L., Schulze, U., & Meier, C. (2022). AI in small-molecule drug discovery: A coming wave? *Nature Reviews Drug Discovery, 21*(3), 175–176. https://doi.org/10.1038/d41573-022-00025-1

Jing, B., Corso, G., Chang, J., Barzilay, R., & Jaakkola, T. (2022). *Torsional diffusion for molecular conformer generation* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2206.01729

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*. https://arxiv.org/abs/1412.6980

Kralovics, R., Passamonti, F., Buser, A. S., Teo, S. S., Tiedt, R., Passweg, J. R., Tichelli, A., Cazzola, M., & Skoda, R. C. (2005). A gain-of-function mutation of JAK2 in

myeloproliferative disorders. *The New England Journal of Medicine, 352*(17), 1779–1790. https://doi.org/10.1056/NEJMoa051113

Krenn, M., Häse, F., Nigam, A. K., Friederich, P., & Aspuru-Guzik, A. (2020). Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology, 1*(4), Article 045024. https://doi.org/10.1088/2632-2153/aba947

Landrum, G. (2013). *RDKit documentation* (Release 2013.09.1) [Computer software documentation]. RDKit. https://www.rdkit.org/docs/

Lv, Y., Qi, J., Babon, J. J., Cao, L., Fan, G., Lang, J., Zhang, J., Mi, P., Kobe, B., & Wang, F. (2024). The JAK-STAT pathway: From structural biology to cytokine engineering. *Signal Transduction and Targeted Therapy, 9*, Article 221. https://doi.org/10.1038/s41392-024-01934-w

Mazuz, E., Shtar, G., Shapira, B., & Rokach, L. (2023). Molecule generation using transformers and policy gradient reinforcement learning. *Scientific Reports, 13*, Article 8799. https://doi.org/10.1038/s41598-023-35648-w

Meyer, S. C., Keller, M. D., Chiu, S., Koppikar, P., Guryanova, O. A., Rapaport, F., Xu, K., Manova, K., Pankov, D., O'Reilly, R. J., Kleppe, M., McKenney, A. S., Shih, A. H., Shank, K., Ahn, J., Papalexi, E., Spitzer, B., Socci, N., Viale, A., ... Levine, R. L. (2015). CHZ868, a type II JAK2 inhibitor, reverses type I JAK inhibitor persistence and demonstrates efficacy in myeloproliferative neoplasms. *Cancer Cell, 28*(1), 15–28. https://doi.org/10.1016/j.ccell.2015.06.006

Miao, Y., Virtanen, A., Zmajkovic, J., Hilpert, M., Skoda, R. C., Silvennoinen, O., & Haikarainen, T. (2024). Functional and structural characterization of clinical-stage Janus kinase 2 inhibitors identifies determinants for drug selectivity. *Journal of Medicinal Chemistry, 67*(12), 10012–10024. https://doi.org/10.1021/acs.jmedchem.4c00197

Mullard, A. (2017). The drug-maker's guide to the galaxy. *Nature, 549*(7673), 445–447. https://doi.org/10.1038/549445a

Olivecrona, M., Blaschke, T., Engkvist, O., & Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics, 9*, Article 48. https://doi.org/10.1186/s13321-017-0235-x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Pereira, T., Abbasi, M., Ribeiro, B., & Arrais, J. P. (2021). Diversity oriented deep reinforcement learning for targeted molecule generation. *Journal of Cheminformatics, 13*, Article 21. https://doi.org/10.1186/s13321-021-00498-z

Philips, R. L., Wang, Y., Cheon, H., Kanno, Y., Gadina, M., Sartorelli, V., Horvath, C. M., Darnell, J. E., Jr., Stark, G. R., & O'Shea, J. J. (2022). The JAK-STAT pathway at 30: Much learned, much more to do. *Cell, 185*(21), 3857–3876. https://doi.org/10.1016/j.cell.2022.09.023

Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Science Advances, 4*(7), Article eaap7885. https://doi.org/10.1126/sciadv.aap7885

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., & Abbeel, P. (2015). Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine*

Learning (Vol. 37, pp. 1889–1897). PMLR. https://proceedings.mlr.press/v37/schulman15.html

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1707.06347

Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., & Tang, J. (2020). GraphAF: A flow-based autoregressive model for molecular graph generation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=S1esMkHYPr

Ståhl, N., Falkman, G., Karlsson, A., Mathiason, G., & Boström, J. (2019). Deep reinforcement learning for multiparameter optimization in de novo drug design. *Journal of Chemical Information and Modeling, 59*(7), 3166–3176. https://doi.org/10.1021/acs.jcim.9b00325

Verstovsek, S., Mesa, R. A., Gotlib, J., Levy, R. S., Gupta, V., DiPersio, J. F., Catalano, J. V., Deininger, M., Miller, C., Silver, R. T., Talpaz, M., Winton, E. F., Harvey, J. H., Arcasoy, M. O., Hexner, E., Lyons, R. M., Paquette, R., Raza, A., Vaddi, K., ... Kantarjian, H. M. (2012). A double-blind, placebo-controlled trial of ruxolitinib for myelofibrosis. *The New England Journal of Medicine, 366*(9), 799–807. https://doi.org/10.1056/NEJMoa1110557

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences, 28*(1), 31–36. https://doi.org/10.1021/ci00057a005

You, J., Liu, B., Ying, R., Pande, V., & Leskovec, J. (2018). Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems* (Vol. 31, pp. 6410–6421). Curran Associates, Inc.

Zdrazil, B., Felix, E., Hunter, F. M. I., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Mendez Lopez, D., Mosquera, J. F., Magarinos, M. P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A., & Leach, A. R. (2024). The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research, 52*(D1), D1180–D1192. https://doi.org/10.1093/nar/gkad1004

Zhou, Z., Kearnes, S., Li, L., Zare, R. N., & Riley, P. (2019). Optimization of molecules via deep reinforcement learning. *Scientific Reports, 9*, Article 10752. https://doi.org/10.1038/s41598-019-47148-x

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). *Fine-tuning language models from human preferences* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1909.08593