


UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

CS/IT Honours Project Final Paper 2023

Title: Improving Simulated Molecular Docking With Curriculum Learning

Author: Claire Fielden (FLDCLA001)

Project Abbreviation: ADD

Supervisor(s): Associate Professor Geoff Nitschke

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	
Theoretical Analysis	*0*	*25*	
Experiment Design and Execution	*0*	*20*	20
System Development and Implementation	*0*	*20*	5
Results, Findings and Conclusions	*10*	*20*	20
Aim Formulation and Background Work	*10*	*15*	15
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	
Total marks		**80**	

Improving Simulated Molecular Docking with Curriculum Learning

Claire Fielden
FLDCLA001@myuct.ac.za
University of Cape Town

ABSTRACT

Automated Drug design (ADD) attempts to design novel ligands for biological targets using ML algorithms. Drug design is extremely costly, taking on average 10 years, with a clinical failure rate of 90%. ML-based drug discovery is a problem of exploration and exploitation of chemical space. Measurement of ligand selectivity is impeded by molecular docking simulations requiring a vast amount of computational resources. Reinforcement Learning (RL) algorithms suffer as a result. Computational models receive insufficient rewards to learn an optimal policy to generate adequate scaffolds. This paper serves as an investigation into Curriculum Learning (CL), a reward shaping mechanism to alleviate the Sparse Rewards Problem.

CCS CONCEPTS

• **Applied computing → Molecular structural biology**; • **Computing methodologies → Reinforcement learning**;

KEYWORDS

Automated Drug Design, Curriculum Learning, Molecular Docking, Reinforcement Learning

1 INTRODUCTION

The main goal of ADD is to identify novel active compounds that can satisfy optimization goals such as bioactivity, selectivity and physico-chemical properties [35]. In recent work, AI-based generative models have been proposed to search chemical space for promising small molecules [74]. The major obstacle is the sheer number of possible solutions, estimated to be on the order of 10^{60} [43], effectively eliminating brute force search as a possibility [30]. *Hit-to-lead optimization* can be performed though Ligand Based Drug Design (LBDD) and Structure Based Drug Design (SBDD) methods [15]. This paper will investigate the applicability of CL to the identification of antagonists for the PfPI4K enzyme [33]. This target receptor is essential to the malaria parasite's development within an infected host [19].

The efficacy of an antagonist drug is measured by its IC50 value, predicted computationally through molecular docking simulations. SBDD is computationally expensive because pharmacodynamics are mechanistic [66]. Consequently, RL algorithms are hindered by the Sparse Rewards Problem, with the quality of proposed drug candidates suffering as a result.

A curriculum centred on a LBDD approach will be evaluated by the productivity of the agent in it's *Production Phase*. Additionally, the ligands produced by CL will be assessed in terms of Synthetic Accessibility (SA) and Binding Affinity (BA) to determine whether both of these properties can be optimized simultaneously [52].

1.1 Research Questions

Currently, the Sparse Rewards Problem problem in ADD is being mitigated by reward-shaping mechanisms such as Potential Based Reward Shaping (PBRS) [4]. However, previous work utilizing this reward shaping mechanism, as discussed in Section 2.1, has shown to produce infeasible candidate ligands. This is due to the agent being rewarded for considering a singular metric in isolation. Bootstrapping [60] and heuristic biasing [36] have also been proposed, as discussed in Section 2.2. However, these methods do little to combat reward sparseness during molecular docking simulations. Experience replay and diversity filters show great promise in balancing exploration and exploitation, as discussed in Section 2.3, however are still limited by the training data used.

The improvement of learning efficiency [28] due to CL will be measured by productivity. An ML model converges when additional training will not improve its prediction capabilities [34]. Productivity is measured based on Average Convergence Rate (ACR) [16] and the average Docking Score (DS) during the *Production Phase*. This research aims to identify a curriculum that can accelerate convergence of an RL agent using sparse DSs as reward, whilst retaining preceding optimizations to SA.

1.1.1 Can CL improve productivity? Provided a curriculum where constituent objectives are strongly correlated to the final, overarching objective, corresponding gradients from sequential simpler tasks are more effective at traversing chemical space [63]. *Curriculum Progression Criteria* train an agent to optimize a specific objective in its curriculum, with higher criteria corresponding to a more sophisticated agent [6]. Diversity filters penalize the score achieved for repetitively sampled compounds [7], thus forcing an agent to become more explorative.

The optimal combination of sophistication and exploration will be assessed in terms of productivity. The result will be compared to that of the standard RL agent to determine if CL can be applied as a reward shaping strategy for SBDD.

1.1.2 Can CL optimize more than one chemical property simultaneously? The trade-off between optimality and diversity, as addressed by Gao and Coley [24] in Section 2.3, is alleviated by Diversity Oriented Synthesis (DOS). Galloway et al. [23] identifies the need to access biologically relevent chemical space in terms of novelty, without sacrificing the ligand's capability to interact with, and thus modulate, biological targets. Currently, baseline RL *hit-to-lead optimization* methods fail on two accounts. Firstly, agents generate unfavourable compounds repetitively, and are unable to propose novel scaffolds [29]. Secondly, structural objectives are considered with little value placed on additional physico-chemical optimization, such as toxicity [61]. The importance of MPO is further dicussed in Section 2.2.

CL aims to alleviate these failures by guiding an agent towards novel regions of chemical space whilst maintaining synthetic feasibility. The SA Scores of the ligands produced by CL agents will be compared to that of an agent without reward shaping. This will evaluate whether a CL agent can fulfill multiple pharmaceutical objectives simultaneously.

1.2 Contributions

In the context of RL-based *hit-to-lead optimization*, an optimal model cannot prioritize DS in isolation. Alternatively, an agent should explore chemical space to sample novel ligands, whilst exploiting rewards for drug-like and synthetically accessible compounds. This work will fulfill the deficiencies in current ADD applications by:

(1) Evaluating Curriculum Learning as a reward shaping mechanism to improve the productivity of an RL agent.
(2) Providing an optimal curriculum for the generation of an antagonist with high BA and synthesizability.
(3) Employing Schrödinger[1] software to perform High-Throughput Virtual Screening (HTVS) for the PfPI4K target receptor.
(4) Adding an additional scoring function to REINVENT's framework in the form of Ertl and Schuffenhauer's [21] SA Score.

2 BACKGROUND AND RELATED WORK

2.1 Reinforcement Learning Approaches

A DDQN to maximize Quantitative Estimate of Drug Likeness (QED) was implemented by Zhou et al. [81] in order to perform molecular optimization. Starting from an empty molecule, the modification process was presented as a MDP, in which an atom or bond was added at each step. The deep neural network implemented a scalarized MPO strategy by maintaining a high Tanimoto Similarity to the specified base scaffold. However, this strategy identified a trade-off between optimality and diversity. In the DDQN, a lack of randomness resulted in a policy that generated the same molecule repetitively, despite the use of a bootstrapped DQN [57]. Additionally, as the weight of the Tanimoto Similarity objective increased, a higher priority was placed on that objective. As a result, QED of the generated scaffolds suffered.

Similarly, RLMOL [45] makes use of a DDQN to generate ligands with a maximal BA for the PfPI4K target receptor. In this framework, the agent was given three different chemotypes as starting molecules. However, as the calculation of DS is computationally demanding, the agent could only receive a reward in the final state of its rollout trajectory. A reward shaping mechanism similar to PBRS [54] was implemented to proliferate reward sparseness. By linearly extrapolating the terminal state reward backwards, the agent was conditioned to select features in the ECFPs [64] that resulted in higher DSs. However, a substantial number of impossible atomic arrangements were generated due to the reward function considering DS in isolation. Additionally, this work made use of *Autodock GPU* [37], a simulator that has shown an inability to capture the specific interactions between ligand and receptor atoms [73]. This is due to the prioritization of computation time over accuracy. Backbone flexibility and movement of several key secondary elements of the receptor involved during ligand binding are neglected as a result [49].

2.2 Multi-Parameter Optimization

Drug-like ligands are optimized to possess multiple pharmaceutical properties by making use of SMILES strings [74]. MolAICal [5] uses a DL model trained using *Autodock Vina* to perform HTVS. Based on the resulting BAs, ligands are filtered according to SA Scores. However, this pocket-based molecular generation method [17] requires extensive computational resources. Gao and Coley [24] suggests the elimination of *post-hoc* filtering, as computational resources are at risk of being misused on ligands that are not synthetically feasible to begin with. Instead, heuristic biasing is used to modify the main objective function to penalize the generation of unsynthesizable compounds.

However, the incorporation of heuristic biasing into the scoring function of a molecular docking simulator causes discrepancies between DS and IC50. Alternatively, Pu et al. suggests eToxPred [61], a DBN to estimate SA Score directly from binary ECFPs. As a significant amount of pre-processing is required to obtain reasonable performance from this predictor, RDKit[2] supplies a computationally inexpensive algorithm for synthesizability calculations. SA Score is further described Appendix B.3 [24].

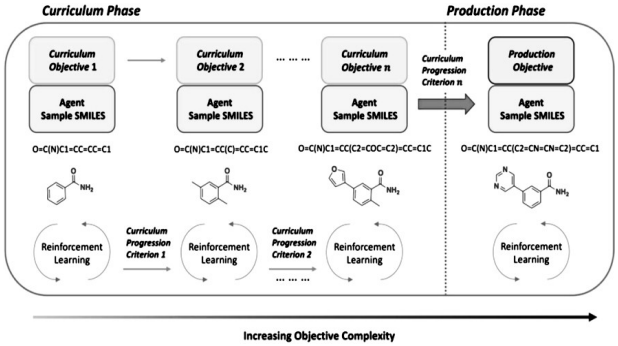


Figure 1: Curriculum Learning Overview [29]

2.3 Reward Shaping Mechanisms

DL approaches often employ LBDD methods, such as Tanimoto Similarity, to optimize the physico-chemical properties of generated ligands. Such properties can be calculated almost instantaneously, providing immediate reward to an MDP. Specific bioactivity, however, can only be measured using SBDD approaches, which are characterized by long computation times and a tendency to failure [39]. In order to extract informative feedback from a sparse reward signal, fine-tuning by transfer learning can be applied. This involves training a model by minimizing loss [50].

Convergence can be improved through reward shaping mechanisms that balance the trade-off between exploration and exploitation during RL. Diversity filters [9] can be used to induce exploratory behaviour by penalizing repeatedly generated scaffolds. Experience replay improves exploitation by retaining highest-scoring compounds in the agent's replay memory buffer [48]. At each RL step, a fraction of the agent's previously generated compounds are randomly sampled to guide the agent towards better-scoring regions of chemical space [78].

2.4 Curriculum Learning

Through the use of GDRL [2], REINVENT[3] makes use of LBDD methods to maximize the scoring functions for generated scaffolds. CL arranges an RL agent's tasks in order of increasing complexity, as displayed in Figure 1.

[1]Schrödinger, a computational platform for the exploration of chemical space and prediction of molecular behavior: https://github.com/Schrödinger

[2]RDKit, a collection of cheminformatics and machine-learning software written in C++ and Python: https://github.com/rdkit/rdkit.git

[3]REINVENT, software from the Molecular AI department at AstraZeneca Research and Development for the purposes of De Novo Drug Design (DNDD): https://github.com/MolecularAI/ReinventCommunity.git

The CL strategy is split into the *Curriculum Phase* and the *Production Phase*. In the *Curriculum Phase*, the agent progresses through successive *Curriculum Objectives* that become more complex. *Curriculum Progression Criterion* decipher whether the agent can progress to the next *Curriculum Objective*, based on the average score for the minibatch of generated compounds. The knowledge obtained during the *Curriculum Phase* is retained when the agent progresses into the *Production Phase*. Previous work by Guo et al. [29] implemented CL to generate PDK-1 inhibitors. The scoring functions were QED and Tanimoto Similarity [8]. The *Production Objective* was to possess enhanced predicted BA to PDK-1. The agent successfully learned a policy to generate optimal binding poses through the use of DockStream [71]. This docking wrapper assessed the BAs between the target receptor crystal structure and generated ligand.

3 DESIGN AND IMPLEMENTATION

3.1 Model Architecture

This model consists of an agent and a prior. The prior achieved its policy as described in Section 3.1.1. When the CL algorithm begins [13], the agent has the same architecture and vocabulary as the prior. This policy is modified using transfer learning, as described in Section 3.1.2.

3.1.1 Prior The prior is a sequence generator that makes use of a trained RNN [25]. The model has been fully trained through the use of teacher forcing on the ChEMBL dataset [79]. The RNN implements three stacked LSTM layers, as introduced by Hochreiter and Schmidhuber [32]. The information contained in each LSTM cell is regulated by neural network gates [1]. The use of LSTMs during training resolves the problem of vanishing and exploding gradients that may occur during backpropagation as a result of long sequences [77].

The prior formulates the task as a Natural Language Problem, upon which it is trained using Maximum Likelihood Estimation [25]. Each hidden layer in the architecture has an embedding size of 256. The output of the RNN first passes through a linear layer and softmax activation function, which provides a token probability distribution.

For each step in the prior's episode, tokens in a sampled SMILES string are one-hot encoded to produce input vectors [27]. Each LSTM cell receives a token, T_i, and outputs a probability of what the next token is likely to be, $P(T_{i+1})$. The RNN uses BPTT to produce an optimal conditional probability distribution [56].

Sequence generation may occur once the prior has learned an optimal policy. The prior's policy is the enumeration of chemical space, resulting in a vocabulary of molecules. The prior used in this work was based on that provided by REINVENT [29]. An exhaustive list of hyperparameters can be found in Table A1 of Appendix A.

3.1.2 Agent The RL agent employs the same RNN architecture as the prior, along with its parameters [13]. The agent is trained in an on-policy fashion. In the *Curriculum Phase*, fine-tuning is used to shift the probability distribution from that of the prior towards a distribution that optimizes Tanimoto Similarity, QED and SA. When CL begins, the agent has the same policy as the prior and an empty replay memory buffer of size s_{incept}.

For a single episode in the agent's curriculum task, the agent performs a sequence of actions, $A = a_1, a_2, ..., a_n$. Each action, a_n, corresponds to the generation of a single SMILES string. Each sample receives a reward, $R(\mathbf{T})$, based on the scoring functions listed in Appendix B. The long-term return $G(A, R = \sum_i^n R(\mathbf{T}_i))$ is the cumulative reward for an episode. An episode starts at $i = 0$ and ends at $i = n$, where n is the batch size.

Table 1: Curriculum Objective Thresholds

Task	Scoring Function	Low	High
1	Tanimoto Similarity	0.1875	0.25
2	QED	0.375	0.5
3	SA	3.75	5

The prior serves as a reference point for the likelihood of sampling a given SMILES string. For every batch of SMILES strings generated by the agent, the prior calculates the Negative Log-Likelihood (NLL). Equation 1 denotes the probability of sampling a token, T_i at step X_i, given the previously sampled tokens.

$$NLL(\mathbf{T})_{prior} = -\sum_{i=1}^{n} \ln P(X_i = T_i | X_{i-1} = T_{i-1} ... X_i = x_1) \quad (1)$$

The new policy, π, must be anchored to that of the prior, π_{prior}, in order to maintain the pre-acquired knowledge of chemical space. Therefore, an augmented likelihood, $NLL(\mathbf{T})_{\mathbb{U}}$, is introduced, as per Equation 2.

$$NLL(\mathbf{T})_{\mathbb{U}} = NLL(\mathbf{T})_{prior} - \sigma R(\mathbf{T})) \quad (2)$$

The reward for the sampled scaffold is multiplied by σ, a scalar coefficient used to adjust the scoring function [51]. Sampled compounds with a $R(\mathbf{T})$ surpassing a threshold of $\vec{i}_{thresh}$ are added to the agent's replay memory buffer. The $NLL(\mathbf{T})_{agent}$ is calculated in the same way as Equation 1, however the fine-tuned parameters of the agent are used instead. The loss function for an episode, $NLL(\mathcal{E})$, is the agreement between the agent likelihood and augmented likelihood, as per Equation 3.

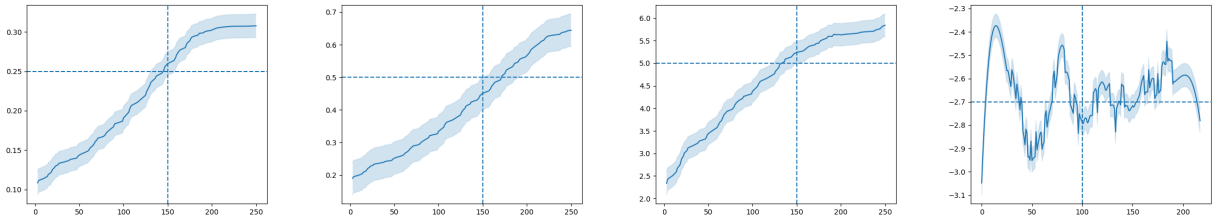
$$NLL(\mathcal{E}) = [NLL(\mathbf{T})_{\mathbb{U}}) - NLL(\mathbf{T})_{agent})]^2 \quad (3)$$

At the end of each episode, the agent's replay memory buffer is sampled for a random subset of 10 high-scoring compounds. These scaffolds are added to the agent's minibatch. The agent's parameters are updated based on the backpropogation [10] of $NLL(\mathcal{E})$. The prior, however, is not altered.

3.2 Curriculum Design

The score for a *Curriculum Objective* is defined by the cumulative reward, $\overline{\mathcal{G}}$. When the cumulative reward surpasses the *Curriculum Progression Criteria*, the agent may continue to the next task in its curriculum. The *Curriculum Phase* is given a maximum of 500 episodes. If the agent fails to meet all *Curriculum Progression Criteria* in this limit, the agent does not progress to the *Production Phase*. The inability to generate optimally-scoring scaffolds may be due to unreasonably high *Curriculum Progression Criteria*, or the desired substructure may not yet exist in the vocabulary of the prior. When the agent achieves a score that satisfies the final task in its curriculum, it proceeds to the *Production Phase*.

3.2.1 Diversity Filter A diversity filter is used to evaluate whether the Bemis-Murcko scaffold [7] has been sampled before. The agent makes use of buckets, each with a maximum size, b. Each bucket contains a scaffold that has been sampled before. When an agent samples a compound, the diversity filter strips away all side chains and analyzes the remaining scaffold. The



(a) **(b)** **(c)** **(d)**

Figure 2: Baseline RL Scoring Functions per Training Episode (a) Tanimoto Similarity in the first 250 episodes of the explorative curriculum **(b)** The explorative agent's QED in the first 250 episodes of the second curriculum task **(c)** SA Scores in the explorative agent's final *Curriculum Objective* **(d)** GlideScores during the first 200 episodes of the benchmarking agent's *Production Phase*

stripped scaffold is allocated a score, T_c. T_c is compared to a similarity threshold, ξ. If a bucket exists for this scaffold, the scaffold is added to the bucket. If the bucket is full, the score for the sampled compound, $R(\mathbf{T})$, is penalized. If the compound's T_c falls below ξ, a new bucket is created.

Agents using a diversity filter will be referred to as explorative hereinafter. The curriculum without a diversity filter will be referred to as exploitative.

3.2.2 Curriculum Progression Criteria Scoring functions used to reward the agent in the *Curriculum Phase* are provided by RDKit [41], the equations for which are listed in Appendix B. *Curriculum Objectives* were split into *low* and *high* scenarios, as per Table 1. The effect of these thresholds was analogous to variable degrees of knowledge. The idea of an *informed* and *uninformed* agent was first introduced by the DeGroot model [18] and further expanded upon by Banerjee et al. [6]. Henceforth, the agent with high *Curriculum Progression Criteria* will be referred to as sophisticated and its uninformed counterpart as naïve.

3.2.3 Sequence As the explorative agent penalizes the reward for a generated scaffold more readily, it is intuitive to define the threshold for the sophisticated agent based on the results in Figure 2a, 2b and 2c. The average $\overline{\mathcal{G}}$ achieved at episode 150 was used as the *high* threshold for each curriculum task. The *low* threshold was set to 75% of the *high* threshold.

The first *Curriculum Objective* was Tanimoto Similarity, as per Equation 9. This constrained the search of chemical space to a region known to possess the PfPI4K inhibitors [80] in Table 2. Thus, the first task was to generate structural analogues to pre-existing molecular backbones [45].

The second *Curriculum Objective* was to produce structures with high QED, measured by Equation 10. Based on the knowledge obtained from exploring high-BA regions of chemical space, the agent was led to learn a policy for constructing scaffolds with desirable physico-chemical properties.

Lastly, in order for the agent to progress into the *Production Phase*, it was trained to predict synthesizability [47]. Equation 11 is the combination of a molecular fragment score [76] and complexity penalty. Ligands are rewarded for possessing previously synthesized structural features [46] and penalized for undesirable features such as stereogenic centers [44].

3.3 Production Phase

The experience replay memory gained from the *Curriculum Phase* was carried over into the *Production Phase* once the agent had satisfied all *Curriculum Progression Criteria*. The agent was given 100 production episodes. No diversity filter was used, and thus scaffolds generated were simply stored without penalty applied to the scoring function.

Schrödinger's software [67] was used to simulate protein-ligand docking [75] in the hopes achieving more reliable results than *Autodock* [73]. Glide [62] has shown to improve scoring accuracy by conducting a complete systematic search of the conformational, orientational, and positional space of a docked ligand [22]. Prerequisite data required to make use of this software suite is supplied in Appendix C.

DockStream and LigPrep [67] pre-processed the ligands generated during the *Production Phase* [69] into PDB structures [40]. Ligand embedding took place at a pH of tolerance of 7.0 ± 1.0, before minimizing the structures using the OPLS3e force field [26]. Glide allowed a maximum of 2 stereoisomers [42] and 3 docking poses to be generated per ligand. The sampling of torsions was biased, forcing ligands to conform to the target receptor within an angle of 20 degrees. Ligands that did not meet the docking constraints received a GlideScore of zero. Results were generated using Glide's HTVS scoring precision, which is intended for the rapid screening of very large numbers of ligands.

3.4 Experiments

The list of hyperparameters applied to the *Curriculum* and *Production* phases can be found in Table A2 and of A3 of Appendix A. In order to find the curriculum that enhanced productivity, four agents were compared, as outlined in Table 3.

3.4.1 Productivity Research Question 1 aims to determine an optimal curriculum to enhance the productivity of an RL agent in its *Production Phase*. The agent's *Curriculum Progression Criteria* are used to compare sophistication and naïvety. The diversity filter [9] penalizes redundancy to enhance exploration. Each curriculum will be analyzed to determine whether it enhances productivity, as compared to the baseline RL algorithm outlined in Section 4.

3.4.2 Multi-Parameter Optimization LBDD methods are used during the *Curriculum Phase*. The question remains whether the use of these methods as *Curriculum Objectives* provide a suitable reward shaping mechanism for the SBDD task. In the case that the agent successfully learns a policy to optimize DS in the *Production Phase*, the resulting ligands will be assessed in terms of their SA Score. Research Question 2 is addressed by assessing whether physico-chemical properties and BAs can be maximized simultaneously in the *Production Phase*.

Table 2: Existing PI4KIIIβ Inhibitors

Ligand	GlideScore	SA Score
PIK-93	$-3.308\,\text{kcal mol}^{-1}$	3.045
Aminopyridine	$-5.116\,\text{kcal mol}^{-1}$	1.799
Napthyridine	$-4.237\,\text{kcal mol}^{-1}$	1.606
Imidazopyridazine	$-4.003\,\text{kcal mol}^{-1}$	2.467

4 EVALUATION AND BENCHMARKING

A benchmarking algorithm was used as the point of reference for productivity and MPO. The benchmark did not make use of CL [29], but instead began SBDD from the prior described in Section 3.1.1. Starting from this policy, the agent was initialized with the *Production Phase* hyperparameters, as listed in Table A3 of Appendix A, and trained as described in Section 3.3.

Due to the stochasticity with which each RL agent sampled chemical space, results could not be compared directly. Therefore each experiment was performed 8 times. Discussions in Section 5 are therefore an overview of 40 experiments, each run on a single node, making use of 3 GPUs and 30 CPU cores.

4.1 Productivity

For each episode in the *Production Phase*, the agent's NLL loss, as per Equation 3, was calculated. Over the 8 experiments, the benchmarking agent achieved the loss curve displayed in Figure 3a. The experiment was initially run for 200 episodes, at the end of which loss was still improving, and thus did not converge.

As molecular docking is extremely costly, it was infeasible to run each experiment's *Production Phase* for 200 episodes. Therefore, each agent was limited to 100 production episodes and $ACR(x)$ was used as a metric for comparison. Average convergence rate, as described by Chen and He [16], was measured as per Equation 4. The average change in loss denotes the speed at which the agent improves upon its prediction capabilities. The first half of the *Production Phase* was compared to the second half to determine the gradient of the loss curve.

$$\Delta\overline{NLL} = \frac{1}{100}\left(\sum_{i=51}^{100} NLL_i\right) - \left(\sum_{i=1}^{50} NLL_i\right) \tag{4}$$

The ACR for each agent was quantified over 8 experiments as per Equation 5.

$$ACR(\overline{x}) = \frac{1}{8}\sum_{i=1}^{8} \Delta\overline{NLL}_i = \frac{1}{n}\left(\Delta\overline{NLL}_1 + \cdots + \Delta\overline{NLL}_8\right) \tag{5}$$

Equation 6 was used to measure the average GlideScore for each agent's *Production Phase*. The resulting molecular docking scores for baseline RL are displayed in Figure 2d.

$$\overline{DS}_{agent} = \frac{1}{8}\sum_{i=1}^{8} \overline{DS}_i = \frac{1}{8}\left(\overline{DS}_1 + \cdots + \overline{DS}_8\right) \tag{6}$$

Productivity is calculated as per Equation 7. The benchmarking agent achieved a $\overline{DS}_{agent}$ of $-2.671\,\text{kcal mol}^{-1}$ and $ACR(\overline{x})$ of $99.146 \times 10^{-2}\,\text{epoch}^{-1}$, resulting in a productivity of 2.648.

$$Productivity = ACR(\overline{x}) \times |\overline{DS}_{agent}| \tag{7}$$

4.2 Multi-Parameter Optimization

The ability to perform MPO was based on the fulfillment of both LBDD and SBDD objectives simultaneously. To perform this evaluation, it was assumed that the agent with the highest $\overline{DS}$ had the most efficient policy for SBDD. The best-performing agent for each experiment category was assessed in terms of Equation 8.

$$SA_{best} = \frac{1}{n}\sum_{i=1}^{n} SA(lig_i) = \frac{1}{n}\left(SA(lig_i) + \cdots + SA(lig_n)\right) \tag{8}$$

The SA Scores for the ligands generated in the 100^{th} episode were averaged to draw correlation to BA. The SA_{best} produced by the benchmarking agent was 3.176. The relationship between DS and SA will quantify an agent's ability to optimize LBDD and SBDD objectives simultaneously. This will determine whether CL can be used to produce ligands that possess high BA and *in vitro* feasibility.

4.3 Comparison to Known Ligands

The results for DS and SA obtained for these experiments were compared to that of pre-existing ligands for the PfPI4K target receptor. More specifically, work done by Ibezim et al. [33] has categorized Type III Beta *phosphatidylinositol 4-kinase* (PI4KIIIβ) to be the only clinically validated drug target in *Plasmodium* kinases [70]. The interaction between the ATP-binding pocket of PI4KIIIβ in complex with *Rab11a-GTP gammaS* [12] was processed for HTVS using the Schrödinger software suite [67].

The antimalarials in Table 2 were chosen based on work by Delves et al. [19] and Maccallum [45], based on the promising inhibitory activity they propose against various *P. falciparum* strains. PIK-93 [11], napthyridine [53], imidazopyridazine [68] and aminopyradine [55] were docked to the PI4KIIIβ receptor using Glide [26]. Currently, the SA and DS achieved by the baseline RL algorithm does not compare to pre-existing substructures [41]. The lowest conformations from the CL experiments will be compared to these compounds in Section 5.3.

5 RESULTS AND DISCUSSION

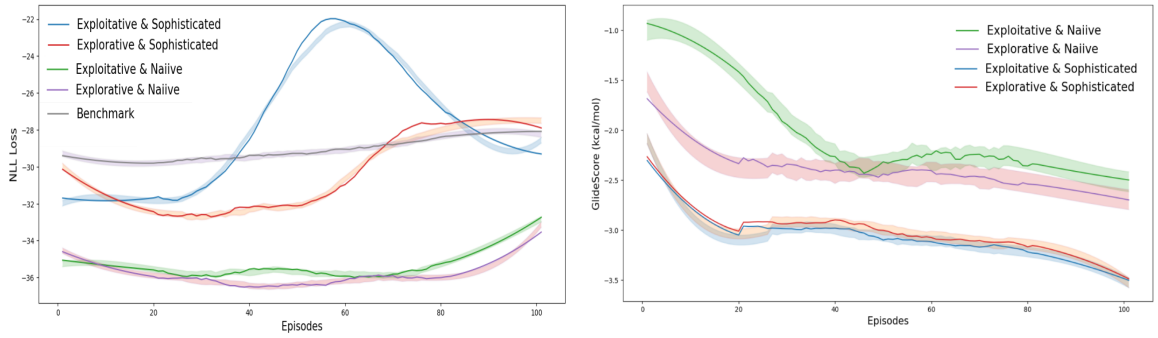
The NLL loss curves for each agent is displayed in Figure 3a, with an increase in slope indicating an improvement to prediction capabilities. Figure 3b compares the GlideScores amongst the CL experiments. More negative GlideScores correspond to higher BA. The impact of CL on MPO is displayed in Figure 5.

5.1 Productivity

5.1.1 Agent 1 The naïve and exploitative agent achieved the worst GlideScores on average, as indicated by the green line in Figure 3b, with a $\overline{DS}$ of $-2.010\,\text{kcal mol}^{-1}$. This is due to the agent's experience replay buffer containing more low-scoring compounds. The cumulative reward, $\overline{\mathcal{G}}$, denotes the desirability of the sampled compound and determines when an agent may progress to its next task. As this was very low for the naïve agent, it began its *Production Phase* with an inclination towards generating low-BA ligands. Moreover, repetitive scaffolds were not penalized. Therefore, the agent was not guided towards sampling alternative regions of chemical space in neither the *Curriculum Phase* nor *Production Phase*. This led to a low $ACR(\overline{x})$ of $39.849 \times 10^{-2}\,\text{epoch}^{-1}$. The agent began training with a NLL of approximately -34 and displayed little improvement within 100 episodes. This agent, indicated by the green line in Figure 3a, is thus slower than the baseline RL experiment, whose grey line has larger NLL values but a smaller gradient. Agent 1's $NLL(\mathcal{E})$ is half that of baseline RL, which implies that the policy updates made during the *Curriculum Phase* were detrimental

Table 3: Table of Experiments

Experiment Number	Description	Curriculum Progression Criteria	Diversity Filter
1	Exploitative and Naïve	Low Threshold	No Filter
2	Exploitative and Sophisticated	High Threshold	No Filter
3	Explorative and Naïve	Low Threshold	Bemis-Murcko scaffold
4	Explorative and Sophisticated	High Threshold	Bemis-Murcko scaffold



(a) Training Loss Values Per Episode **(b) Docking Scores Per Episode**

Figure 3: Performance Curves for the *Production Phase*

Agent 1: Exploitative and Naïve, *Curriculum Phase Length:* ±125 episodes, **Agent 2:** Exploitative and Sophisticated, *Curriculum Phase Length:* ±145 episodes, **Agent 3:** Explorative and Naïve, *Curriculum Phase Length:* ±150 episodes, **Agent 4:** Explorative and Sophisticated, *Curriculum Phase Length:* ±350 episodes, **Benchmark:** *No Curriculum Phase*

to the initial policy of the prior. Therefore, it is probable that the lower *Curriculum Progression Criteria* are detrimental as it generalizes the prediction capabilities of the prior towards an irrelevent and undesirable area of chemical space.

Agent 1 generates scaffolds with a lower probability of binding to the target receptor. The productivity of the naïve and exploitative agent was 0.801, and thus does not provide an improvement to the learning efficiency [28] of baseline RL.

5.1.2 Agent 2 The sophisticated and exploitative agent displayed the best GlideScores, averaging $-3.070\,\text{kcal}\,\text{mol}^{-1}$, as displayed by the blue line in Figure 3b. Additionally, it was able to achieve the best overall $NLL(\mathcal{E})$ amongst all the agents. Training began with a loss of -32 and outperformed all other experiments within 60 episodes. The $ACR(\overline{x})$ achieved was $485.460 \times 10^{-2}\,\text{epoch}^{-1}$, revealing attractive acceleration capabilities. This is due to the agent's ability to exploit high-reward regions of chemical space sampled during its *Curriculum Phase* early in the *Production Phase*. Learning is directed by the compounds in the agent's replay memory buffer, an abundance of which exist due to the lack of penalization applied to repetitively sampled compounds.

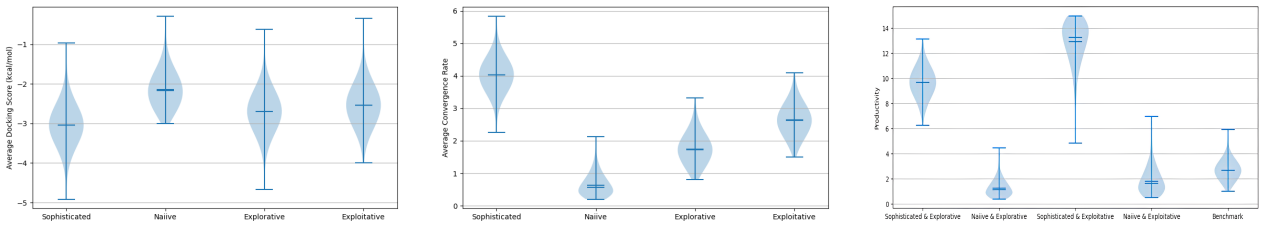
The sophisticated agents began their *Production Phase* with notably better docking scores than the naïve agents. This is due to the compounds generated at the beginning of the *Production Phase* exhibiting higher Tanimoto Similarity to previous high-BA functional analogues. Agent 2 achieved an average productivity of 14.903. This is the highest productivity amongst all the experiments, and outperforms the standard RL algorithm.

This demonstrates the use of CL as a reward shaping mechanism, as a sophisticated strategy will provide the prefocusing of the prior required to accelarate *hit-to-lead optimization*. The experience replay buffer allows the agent to focus on relevant regions of chemical space, thus enhancing productivity.

5.1.3 Agent 3 The naïve and explorative agent achieved a similar training loss curve to Agent 1. Agent 3's $NLL(\mathcal{E})$ is denoted by the purple curve in Figure 3a. The NLL tends toward zero, however does not achieve the same magnitude as baseline RL. The $ACR(\overline{x})$ for both naïve agents was well below 0.4, and thus less than half that of the benchmarking algorithm. This implies that a naïve curriculum is detrimental to the learning efficiency of an RL algorithm. Agent 3 achieved an $ACR(\overline{x})$ of $20.244 \times 10^{-2}\,\text{epoch}^{-1}$, which is the worst pace amongst all experiments. The violin plots in Figure 4b reiterate that the incorrect use of a diversity filter can stagnate an RL agent's convergence rate. Prioritizing exploitation improves $ACR(\overline{x})$, resulting in an agent that is able to quickly improve upon its prediction capabilities.

The effect of naïvety is clear in Figure 4a. The mean DS for naïve agents tend toward zero, indicating an inability to produce ligands that successfully dock to the target receptor. On the contrary, the sophisticated agent's DS tends towards -5, indicating better predicted BA. However, these violin plots also emphasize the explorative agent's tendency to produce ligands with higher DS than the exploitative agent. This is pronounced amongst the naïve agents, as Agent 3 achieved slightly better GlideScores than Agent 1. This is accentuated by the purple line in Figure 3b, which remains below the green line throughout the *Production Phase*. The $\overline{DS}$ for Agent 3 was $-2.392\,\text{kcal}\,\text{mol}^{-1}$, which is marginally worse than the average of -2.671 achieved by the benchmarking agent.

It is interesting to consider that, although the explorative and naïve agent is less capable of improving its prediction capabilities than its exploitative counterpart, it is able to produce higher-BA ligands. This is due to the fact that a larger range of chemical space was explored during the *Curriculum Phase*. As a result, the agent was able to respond to poor rewards from Glide [26] in the *Production Phase* by generating more diverse scaffolds. The resulting policy was more likely to satisfy molecular docking



(a) Average Docking Scores **(b) Average Convergence Rates** **(c) Average Productivity**

Figure 4: Agent 1: Exploitative and Naïve, *Average Productivity:* ±0.801, **Agent 2:** Exploitative and Sophisticated, *Productivity:* ±14.903, **Agent 3:** Explorative and Naïve, *Average Productivity:* ±0.484, **Agent 4:** Explorative and Sophisticated, *Average Productivity:* ±9.657, **Benchmark:** *Average Productivity:* ±2.648

constraints, despite the fact that that these predictions did not make much improvement as training progressed.

Agent 3 outperformed Agent 1 in terms of GlideScore, but proved to be the least efficient experiment overall, quantified by a productivity of 0.484. This is worse than the benchmarking agent and approximately half that of Agent 1, as displayed in Figure 4c. It can be concluded that a diversity filter has a negative effect on productivity when paired with naïvety. This is a consequence of the scaffolds in Agent 3's replay memory being the most sparse. The *Curriculum Phase* was brief due to the low *Curriculum Progression Criteria*. Few scaffolds were added to the replay memory buffer during this time, and they were unlikely to possess high Tanimoto Similarity to pre-existing antimalarials [19]. Moreover, these scaffolds were penalized for redundancy during the *Curriculum Phase*, decreasing the likelihood of a compound surpassing the $\vec{i}_{thresh}$ required to be added to the agent's replay memory.

Agent 3's curriculum was detrimental to the productivity of baseline RL. Exploration significantly narrowed the perceived solution space when combined with naïvety. During the *Curriculum Phase*, insufficient requirements for reward from the environment resulted in the generation of suboptimal ligands. Therefore, the prior's parameters were misadjusted to produce a worsened policy. The combination of an unrefined policy and sparse reward signals from the *Curriculum Phase* produced the least performant agent in the *Production Phase*.

5.1.4 Agent 4 Figure 3b reveals that the sophisticated and explorative agent achieved a similar trend in GlideScore to that of Agent 2, with a slightly worse average of $-3.016\,\text{kcal}\,\text{mol}^{-1}$. Similarly, in Figure 3a, the sophisticated and explorative agent is denoted by the orange curve with a gradient, $ACR(\overline{x})$, of $320.236 \times 10^{-2}\,\text{epoch}^{-1}$. Both sophisticated agents more than tripled the $ACR(\overline{x})$ of baseline RL. This displays the benefit of prioritizing knowledge attainment during the *Curriculum Phase*, as higher *Curriculum Progression Criteria* result in a policy that has been focussed towards high-BA chemical space. Both sophisticated agents improved upon the $\overline{DS}$ achieved by the benchmarking agent.

Agent 4 is slower than its exploitative counterpart, a phenomenon that is highlighted in Figure 4b. Exploration during the *Curriculum Phase* results in the penalization of repetitive scaffolds. This is extremely frequent when considering a sophisticated agent, as larger *Curriculum Progression Criteria* result in longer *Curriculum Phases*. Compounds were less likely to be added to Agent 4's replay memory buffer during CL as they seldomly surpassed $\vec{i}_{thresh}$. Thus, the *Production Phase* began with considerably less knowledge of chemical space than Agent 2. This slowed the improvement upon Agent 4's prediction capabilities in response to molecular docking simulations. The negative effect of diversity filters on $ACR(\overline{x})$, however, was not as pronounced for the sophisticated agents as it was for the naïve agents because of the high *Curriculum Progression Criteria*. The higher threshold allowed sufficient fine-tuning of the prior's policy to cultivate improvement to baseline RL.

The average DS achieved by the explorative agent was approximately equal to that of the exploitative agent, as per Figure 4a, under a sophisticated threshold. This reveals little correlation between the use of a diversity filter and GlideScore when sufficient training has taken place before the *Production Phase*. Agent 4's productivity was 9.657, showing only a slight decrease in learning efficiency compared to Agent 2. Figure 4c highlights that both sophisticated curricula outperform the benchmarking algorithm.

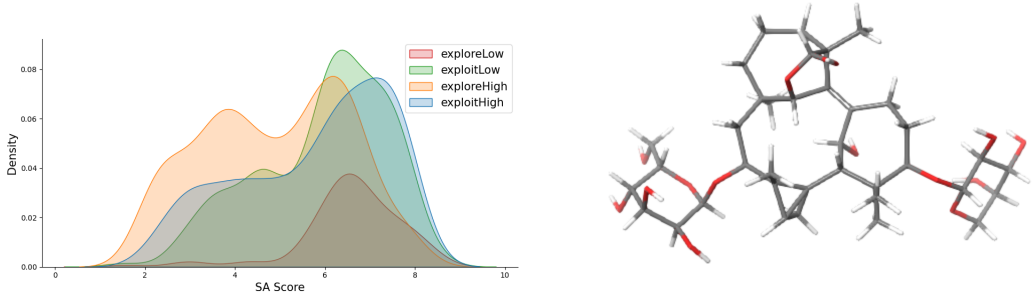
Therefore, it can be concluded that the use of a sophisticated and explorative curriculum can improve upon the productivity of standard RL-based *hit-to-lead optimization*. Exploration may be a hindrance to convergence rate if not enough time is spent in the *Curriculum Phase*. The benefit of diversity, such as the case of Agent 3, is that the agent is less likely to repetitively generate low-scoring compounds with similar scaffolds.

5.2 Analysis of Multi-Parameter Optimization

The probability density functions for synthesizability are displayed in Figure 5a. All of the CL experiments outperformed the benchmarking agent, with Agent 1 achieving the worst SA Score of 4.887. This indicates that although a naïve and exploitative curriculum results in an unproductive RL agent, it is a relevant reward shaping mechanism when applied to the MPO task of ADD.

Agent 2 achieved the best average SA Score of 6.622. This is nearly double the synthesizability score achieved by the benchmarking experiment. The average SA Scores for the sophisticated agents were better than that of the naïve agents. This exhibits the advantage of high *Curriculum Progression Criteria* when generating ligands with better SA and BA, concurrently, during the *Production Phase*.

The explorative experiments both achieved similar averages, ranging between 5.5 and 6. Therefore, CL clearly allows the simultaneous optimization of physico-chemical properties in both cases. The probability density function in Figure 5a was used to further assess the correlation between GlideScore and SA. The distribution is left-tailed for all agents, so it is more likely for SA Scores to fall below the mean. However, the bell curve for the explorative and sophisticated agent is more evenly



(a) Probability Density Function of Synthesizability **(b) Top-Scoring Ligand**

Figure 5: Agent 1 (green): Exploitative and Naïve, *Average SA:* ±4.887, **Agent 2 (blue):** Exploitative and Sophisticated, *Average SA:* ±6.623, **Agent 3 (red):** Explorative and Naïve, *Average SA:* ±5.759, **Agent 4 (orange):** Explorative and Sophisticated, *Average SA:* ±5.939, **Benchmark:** *Average Productivity:* ±3.176**, Ligand:** GlideScore: ±−5.125 kcal mol^{-1}, SA Score: ±6.961

distributed. This indicates a lower degree of asymmetry amongst SA Scores and a greater density of high-BA ligands compared to the other agents.

In conclusion, sophisticated *Curriculum Progression Criteria* are able to leverage LBDD and SBDD objectives during the *Production Phase*. Naïve agents struggle to produce ligands with high SA, however even a brief amount of fine-tuning of the prior produces a policy that outperforms that of the baseline RL experiment. Agent 2 achieved the highest SA Score on average. However, Agent 4 presents a greater density across the x axis and thus a greater likelihood of optimizing selectivity and synthesizability [23] simultaneously in the generated ligands.

5.3 Summary

The value of a Bemis-Murcko scaffold [9] diversity filter was not pronounced throughout the experiments. This lack of performance is due to the meagre amount of *Curriculum Phase* episodes provided. At the beginning of RL, the agent will automatically prioritize exploration as the learned probability distribution is approximately uniform. Only once training progresses will the agent become more exploitative, after which the effect of a diversity filter would become apparent, and thus the application thereof was meaningless for a 100-episode *Production Phase*.

Curriculum Progression Criteria directly impacted the number of episodes spent in the *Curriculum Phase*, with more sophisticated and explorative agents taking the longest to meet their curriculum thresholds. Although LBDD objectives are inexpensive, naïvety proposed better computational complexity. However, mode collapse occurred in both naïve agents, and productivity declined as a result. Therefore transfer learning through the use of a naïve curriculum, as compared to baseline RL, produces a less focused and unproductive agent.

Sophisticated agents improved upon the $ACR(\overline{x})$ and $\overline{DS}$ of the RL benchmarking algorithm. There was also an exponential improvement to productivity. The best-performing agent was Agent 3. Research Question 1 has therefore been satisfied, as the use of a curriculum that is sophisticated and exploitative can accelarate RL-based SBDD in the *Production Phase*.

Moreover, the exploitative and sophisticated agent displayed an improvement to baseline RL's MPO capabilities. It was successfully able to produce scaffolds that optimize synthesizability in addition to BA. The best-performing ligand from Agent 3 is displayed in Figure 5b. This GlideScore and SA Score are comparable to the antimalarials [19] listed in Table 2. Therefore, in answer to Research Question 2, CL is an improvement to current RL algorithms as it improves upon the synthesizability and binding affinity [23] of proposed ligands concurrently.

6 CONCLUSIONS

Curriculum Learning is a useful reward shaping mechanism for RL-based *hit-to-lead optimization*. The use of a sophisticated curriculum improves productivity by focusing a policy toward high-BA regions of chemical space. Corresponding gradients attained during the *Curriculum Phase* accelerate convergence on high DS ligands in the *Production Phase*.

Curriculum Learning can perform Multi-Parameter Optimization by producing more chemically viable ligands with higher DSs than baseline RL. The sophisticated agent was able to optimize physico-chemical properties and bioactivity simultaneously during the *Production Phase*.

Further investigation is required on the use of diversity filters. In this work, the *Production Phase* consisted of 100 episodes, and thus the true benefit of DOS was not exhibited. For exploration to improve standard RL *hit-to-lead optimization*, sophisticated *Curriculum Progression Criteria* should be made use of to avoid mode collapse. Moreover, a longer *Curriculum Phase* is required to accommodate the combination of sophistication and exploration. This shows promise in proposing novel regions of chemical space. Such propositions would be worth the *Curriculum Phase* training time, which is computationally inexpensive, due to expected improvement in *Production Phase* convergence time and ligand quality.

In this work, the optimal curriculum for productivity was sophisticated and exploitative. This same curriculum proposed high-BA conformations with better predicted synthesizability. Therefore, CL is a framework that can improve RL-based *hit-to-lead optimization*, and perhaps be extended to other applications impeded by the Sparse Rewards Problem.

REFERENCES

[1] Santhosh Amilpur and Raju Bhukya. 2022. Predicting novel drug candidates against Covid-19 using generative deep neural networks. *Journal of Molecular Graphics and Modelling* 110 (Jan. 2022), 108045. https://doi.org/10.1016/j.jmgm.2021.108045

[2] Per-Arne Andersen, Morten Goodwin, and Ole-Christoffer Granmo. 2020. CostNet: An End-to-End Framework for Goal-Directed Reinforcement Learning. (2020), 94–107. https://doi.org/10.1007/978-3-030-63799-6_7

[3] Amy C. Anderson. 2011. Structure-Based Functional Design of Drugs: From Target to Lead Compound. (Oct. 2011), 359–366. https://doi.org/10.1007/978-1-60327-216-2_23

[4] Babak Badnava, Mona Esmaeili, Nasser Mozayani, and Payman Zarkesh-Ha. 2023. A new Potential-Based Reward Shaping for Reinforcement Learning Agent. (2023). arXiv:cs.AI/1902.06239

[5] Qifeng Bai, Shuoyan Tan, Tingyang Xu, Huanxiang Liu, Junzhou Huang, and Xiaojun Yao. 2020. MolAICal: A soft tool for 3D drug design of protein targets by Artificial Intelligence and classical algorithm. *Briefings in Bioinformatics* 22, 3 (2020). https://doi.org/10.1093/bib/bbaa161

[6] Abhijit Banerjee, Emily Breza, Arun Chandrasekhar, and Markus Mobius. 2019. *Naive Learning with Uninformed Agents.* Technical Report. https://doi.org/10.3386/w25497

[7] Guy W. Bemis and Mark A. Murcko. 1996. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* 39, 15 (Jan. 1996), 2887–2893. https://doi.org/10.1021/jm9602928

[8] G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. 2012. Quantifying the chemical beauty of drugs. *Nature Chemistry* 4, 2 (Jan. 2012), 90–98. https://doi.org/10.1038/nchem.1243

[9] Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov. 2020. REINVENT 2.0: An AI Tool for De Novo Drug Design. *Journal of Chemical Information and Modeling* 60, 12 (Oct. 2020), 5918–5922. https://doi.org/10.1021/acs.jcim.0c00915

[10] Mikael Boden. 2002. A guide to recurrent neural networks and backpropagation. *the Dallas project* 2, 2 (2002), 1–10.

[11] J.E. Burke, A.J. Inglis, O. Perisic, G.R. Masson, S.H. McLaughin, F. Rutaganira, K.M. Shokat, and R.L. Williams. 2014. Phosphatidylinositol 4-kinase III beta-PIK93 in a complex with Rab11a- GTP gammaS. (May 2014). https://doi.org/10.2210/pdb4d0l/pdb

[12] John E. Burke, Alison J. Inglis, Olga Perisic, Glenn R. Masson, Stephen H. McLaughlin, Florentine Rutaganira, Kevan M. Shokat, and Roger L. Williams. 2014. Structures of PI4KIIIβ complexes show simultaneous recruitment of Rab11 and its effectors. *Science* 344, 6187 (May 2014), 1035–1038. https://doi.org/10.1126/science.1253397

[13] Víctor Campos, Pablo Sprechmann, Steven Hansen, Andre Barreto, Steven Kapturowski, Alex Vitvitskyi, Adrià Puigdomènech Badia, and Charles Blundell. 2021. Beyond Fine-Tuning: Transferring Behavior in Reinforcement Learning. (2021). arXiv:cs.LG/2102.13515

[14] Thomas Cauchy, Jules Leguy, and Benoit Da Mota. 2022. Definition and exploration of realistic chemical spaces using the connectivity and cyclic features of ChEMBL and ZINC. (Dec. 2022). https://doi.org/10.26434/chemrxiv-2022-2b41l

[15] H.C. Stephen Chan, Hanbin Shan, Thamani Dahoun, Horst Vogel, and Shuguang Yuan. 2019. Advancing Drug Discovery via Artificial Intelligence. *Trends in Pharmacological Sciences* 40, 8 (Aug. 2019), 592–604. https://doi.org/10.1016/j.tips.2019.06.004

[16] Yu Chen and Jun He. 2021. Average convergence rate of evolutionary algorithms in continuous optimization. *Information Sciences* 562 (2021), 200–219. https://doi.org/10.1016/j.ins.2020.12.076

[17] Yangyang Chen, Zixu Wang, Lei Wang, Jianmin Wang, Pengyong Li, Dongsheng Cao, Xiangxiang Zeng, Xiucai Ye, and Tetsuya Sakurai. 2023. Deep generative model for drug design from protein target sequence. *Journal of Cheminformatics* 15, 1 (March 2023). https://doi.org/10.1186/s13321-023-00702-2

[18] Morris H. Degroot. 1974. Reaching a Consensus. *J. Amer. Statist. Assoc.* 69 (1974), 118–121. https://api.semanticscholar.org/CorpusID:120077293

[19] Michael Delves, David Plouffe, Christian Scheurer, Stephan Meister, Sergio Wittlin, Elizabeth A. Winzeler, Robert E. Sinden, and Didier Leroy. 2012. The Activities of Current Antimalarial Drugs on the Life Cycle Stages of Plasmodium: A Comparative Study with Human and Rodent Parasites. *PLoS Medicine* 9, 2 (Feb. 2012), e1001169. https://doi.org/10.1371/journal.pmed.1001169

[20] Matthew D. Eldridge, Christopher W. Murray, Timothy R. Auton, Gaia V. Paolini, and Roger P. Mee. 1997. *Journal of Computer-Aided Molecular Design* 11, 5 (1997), 425–445. https://doi.org/10.1023/a:1007996124545

[21] Peter Ertl and Ansgar Schuffenhauer. 2009. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* 1, 1 (June 2009). https://doi.org/10.1186/1758-2946-1-8

[22] 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. 2004. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry* 47, 7 (Feb. 2004), 1739–1749. https://doi.org/10.1021/jm0306430

[23] Warren R.J.D. Galloway, Albert Isidro-Llobet, and David R. Spring. 2010. Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules. *Nature Communications* 1, 1 (Sept. 2010). https://doi.org/10.1038/ncomms1081

[24] Wenhao Gao and Connor W. Coley. 2020. The Synthesizability of Molecules Proposed by Generative Models. *Journal of Chemical Information and Modeling* 60, 12 (2020), 5714–5723. https://doi.org/10.1021/acs.jcim.0c00174 arXiv:https://doi.org/10.1021/acs.jcim.0c00174 PMID: 32250616.

[25] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington. 2011. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* 40, D1 (Sept. 2011), D1100–D1107. https://doi.org/10.1093/nar/gkr777

[26] Schrödinger Glide. 2017. LLC, New York, NY, 2019. *Glide, Schrödinger* (2017).

[27] Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850* (2013).

[28] Alex Graves, Marc G. Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. 2017. Automated Curriculum Learning for Neural Networks. (2017). arXiv:cs.NE/1704.03003

[29] Jeff Guo, Vendy Fialková, Juan Arango, Christian Margreitter, Jon Paul Janet, Kostas Papadopoulos, Ola Engkvist, and Atanas Patronov. 2022. Improving de novo molecular design with curriculum learning. *Nature Machine Intelligence* 4 (06 2022), 1–9. https://doi.org/10.1038/s42256-022-00494-4

[30] Jeff Guo, Jon Paul Janet, Matthias R. Bauer, Eva Nittinger, Kathryn A. Giblin, Kostas Papadopoulos, Alexey Voronov, Atanas Patronov, Ola Engkvist, and Christian Margreitter. 2021. DockStream: a docking wrapper to enhance de novo molecular design. *Journal of Cheminformatics* 13, 1 (Nov. 2021). https://doi.org/10.1186/s13321-021-00563-7

[31] Jeff Guo, Jon Paul Janet, Matthias R. Bauer, Eva Nittinger, Kathryn A. Giblin, Kostas Papadopoulos, Alexey Voronov, Atanas Patronov, Ola Engkvist, and Christian Margreitter. 2021. DockStream: a docking wrapper to enhance de novo molecular design. *Journal of Cheminformatics* 13, 1 (Nov. 2021). https://doi.org/10.1186/s13321-021-00563-7

[32] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9 (12 1997), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

[33] Akachukwu Ibezim, Mbanefo S. Madukaife, Sochi C. Osigwe, Nadja Engel, Ramanathan Karuppasamy, and Fidele Ntie-Kang. 2022. Fragment-based virtual screening discovers potential new Plasmodium PI4KIIIβ ligands. *BMC Chemistry* 16, 1 (March 2022). https://doi.org/10.1186/s13065-022-00812-2

[34] Jiang Jiawei, Fangcheng Fu, Tong Yang, Yingxia Shao, and Bin Cui. 2020. SKCompress: compressing sparse and nonuniform gradient in distributed machine learning. *The VLDB Journal* 29 (09 2020). https://doi.org/10.1007/s00778-019-00596-3

[35] Yuko Kawasaki and Ernesto Freire. 2011. Finding a better path to drug selectivity. *Drug Discovery Today* 16, 21-22 (Nov. 2011), 985–990. https://doi.org/10.1016/j.drudis.2011.07.010

[36] Alan Kerstjens and Hans De Winter. 2022. LEADD: Lamarckian evolutionary algorithm for de novo drug design. *Journal of Cheminformatics* 14, 1 (Jan. 2022). https://doi.org/10.1186/s13321-022-00582-y

[37] Ryangguk Kim and Jeffrey Skolnick. 2008. Assessment of programs for ligand binding affinity prediction. *Journal of Computational Chemistry* 29, 8 (2008), 1316–1331. https://doi.org/10.1002/jcc.20893 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20893

[38] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. 2015. PubChem Substance and Compound databases. *Nucleic Acids Research* 44, D1 (Sept. 2015), D1202–D1213. https://doi.org/10.1093/nar/gkv951

[39] Maria Korshunova, Niles Huang, Stephen Capuzzi, Dmytro S. Radchenko, Olena Savych, Yuriy S. Moroz, Carrow Wells, Timothy M. Willson, Alexander Tropsha, Olexandr Isayev, and et al. 2021. A bag of tricks for automated de novo design of molecules with the desired properties: Application to EGFR inhibitor discovery. (2021). https://doi.org/10.26434/chemrxiv.14045072

[40] A. Kouranov. 2006. The RCSB PDB information portal for structural genomics. *Nucleic Acids Research* 34, 90001 (Jan. 2006), D302–D305. https://doi.org/10.1093/nar/gkj120

[41] Greg Landrum et al. 2013. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 8 (2013), 31.

[42] Arthur A. Levin, Laurie J. Sturzenbecker, Sonja Kazmer, Thomas Bosakowski, Christine Huselton, Gary Allenby, Jeffrey Speck, Cl. ratzeisen, Michael Rosenberger, Allen Lovey, and Joseph F. Grippo. 1992. 9-Cis retinoic acid stereoisomer binds and activates the nuclear receptor RXRα. *Nature* 355, 6358 (Jan. 1992), 359–361. https://doi.org/10.1038/355359a0

[43] Hongming Li and Yong Fan. 2019. Early prediction of alzheimer's disease dementia based on baseline hippocampal MRI and 1-year follow-up cognitive measures using deep recurrent neural networks. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (Jul 2019). https://doi.org/10.1109/isbi.2019.8759397

[44] Chun Ma, Yue Sun, Junfeng Yang, Hao Guo, and Junliang Zhang. 2023. Catalytic Asymmetric Synthesis of Troger's Base Analogues with Nitrogen Stereocenter. *ACS Central Science* 9, 1 (2023), 64–71.

[45] Rob Maccallum. 2021. Automatic Hit-to-Lead Optimization. (2021).

[46] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath. 2013. Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry* 57, 8 (Nov. 2013), 3186–3204. https://doi.org/10.1021/jm401411z

[47] A. K. Mandagere. 2002. Strategies in Lead Selection and Optimization: Application of a Graphical Model and Automated In Vitro ADME Screening. (2002), 185–202.

[48] Binyamin Manela and Armin Biess. 2019. Bias-Reduced Hindsight Experience Replay with Virtual Goal Prioritization. *CoRR* abs/1905.05498 (2019). arXiv:1905.05498 http://arxiv.org/abs/1905.05498

[49] Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. 2011. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer Aided-Drug Design* 7, 2 (June 2011), 146–157. https://doi.org/10.2174/157340911795677602

[50] Daniel Merk, Francesca Grisoni, Lukas Friedrich, and Gisbert Schneider. 2018. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Communications Chemistry* 1, 1 (Oct. 2018). https://doi.org/10.1038/s42004-018-0068-1

[51] Paul-Henri Michel and Vincent Heiries. 2015. An Adaptive Sigma Point Kalman Filter Hybridized by Support Vector Machine Algorithm for Battery SoC and SoH Estimation. (2015), 1–7. https://doi.org/10.1109/VTCSpring.2015.7145678

[52] Varnavas D. Mouchlis, Antreas Afantitis, Angela Serra, Michele Fratello, Anastasios G. Papadiamantis, Vassilis Aidinis, Iseult Lynch, Dario Greco, and Georgia Melagraki. 2021. Advances in De Novo Drug Design: From Conventional to Machine Learning Methods. *International Journal of Molecular Sciences* 22, 4 (Feb. 2021), 1676. https://doi.org/10.3390/ijms22041676

[53] K. Nakatani, S. Hagihara, Y. Goto, A. Kobori, M. Hagihara, G. Hayashi, M. Kyo, M. Nomura, M. Mishima, and C. Kojima. 2006. Solution structure of the AA-mismatch DNA complexed with naphthyridine-azaquinolone. (April 2006). https://doi.org/10.2210/pdb1x26/pdb

[54] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. 1999. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. (1999), 278–287.

[55] S. Oh and R.K. Hite. 2022. Human TMEM175 in complex with 4-aminopyridine. (Nov. 2022). https://doi.org/10.2210/pdb8dhm/pdb

[56] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. 2017. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics* 9, 1 (Sept. 2017). https://doi.org/10.1186/s13321-017-0235-x

[57] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep Exploration via Bootstrapped DQN. *CoRR* abs/1602.04621 (2016). arXiv:1602.04621 http://arxiv.org/abs/1602.04621

[58] John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. 2006. How many drug targets are there? *Nature reviews Drug discovery* 5, 12 (2006), 993–996.

[59] Rosy Pradhan, Manabendra Patra, Ajaya K Behera, Bijay K Mishra, and Rajani K Behera. 2006. A synthon approach to spiro compounds. *Tetrahedron* 62, 5 (2006), 779–828.

[60] Augustin Prodan and Remus Campean. 2005. Bootstrapping methods applied for simulating laboratory works. *Campus-Wide Information Systems* 22, 3 (July 2005), 168–175. https://doi.org/10.1108/10650740510606171

[61] Limeng Pu, Misagh Naderi, Tairan Liu, Hsiao-Chun Wu, Supratik Mukhopadhyay, and Michal Brylinski. 2019. EToxPred: A machine learning-based approach to estimate the toxicity of drug candidates. *BMC Pharmacology and Toxicology* 20, 1 (2019). https://doi.org/10.1186/s40360-018-0282-6

[62] Matthew P. Repasky, Mee Shelley, and Richard A. Friesner. 2007. Flexible Ligand Docking with Glide. *Current Protocols in Bioinformatics* 18, 1 (June 2007). https://doi.org/10.1002/0471250953.bi0812s18

[63] Jean-Louis Reymond, Lars Ruddigkeit, Lorenz Blum, and Ruud van Deursen. 2012. The enumeration of chemical space. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2, 5 (April 2012), 717–733. https://doi.org/10.1002/wcms.1104

[64] David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* 50, 5 (2010), 742–754. https://doi.org/10.1021/ci100050t

[65] Magnus Rydén, Patrick Moldenhauer, Simon Lindqvist, Tobias Mattisson, and Anders Lyngfelt. 2014. Measuring attrition resistance of oxygen carrier particles for chemical looping combustion with a customized jet cup. *Powder Technology* 256 (2014), 75–86.

[66] Mohammed Saji Salahudeen and Prasad S. Nishtala. 2017. An overview of pharmacodynamic modelling, ligand-binding approach and its application in clinical practice. *Saudi Pharmaceutical Journal* 25, 2 (Feb. 2017), 165–175. https://doi.org/10.1016/j.jsps.2016.07.002

[67] LLC Schrödinger. 2019. Schrödinger release 2019-4: protein preparation wizard; Epik; impact; prime; glide; LigPrep; induced fit docking protocol. *New York, NY* (2019).

[68] P.L. Shaffer, J. Tang, and P. Yakowec. 2013. Crystal Structure of PI3K-gamma in Complex with Imidazopyridazine 19e. (April 2013). https://doi.org/10.2210/pdb4fhk/pdb

[69] John C Shelley, Anuradha Cholleti, Leah L Frye, Jeremy R Greenwood, Mathew R Timlin, and Makoto Uchimaya. 2007. Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *Journal of computer-aided molecular design* 21 (2007), 681–691.

[70] A Shukla, A Singh, LP Pathak, N Shrivastava, PK Tripathi, MP Singh, and K Singh. 2012. Inhibition of P. falciparum pfATP6 by curcumin and its derivatives: A bioinformatic Study. *Cellular and molecular biology* 58, 1 (2012), 182–186.

[71] Shalini Singh and Pradeep Srivastava. 2015. Molecular Docking Studies of Myricetin and Its Analogues against Human PDK-1 Kinase as Candidate Drugs for Cancer. *Computational Molecular Bioscience* 05, 02 (2015), 20–33. https://doi.org/10.4236/cmb.2015.52004

[72] Abigail N. Smith, Daniel J. Blackwell, Bjorn C. Knollmann, and Jeffrey N. Johnston. 2021. Ring Size as an Independent Variable in Cyclooligomeric Depsipeptide Antiarrhythmic Activity. *ACS Medicinal Chemistry Letters* 12, 12 (Nov. 2021), 1942–1947. https://doi.org/10.1021/acsmedchemlett.1c00508

[73] Leonardo Solis-Vasquez, Andreas F. Tillack, Diogo Santos-Martins, Andreas Koch, Scott LeGrand, and Stefano Forli. 2022. Benchmarking the performance of irregular computations in AutoDock-GPU molecular docking. *Parallel Comput.* 109 (2022), 102861. https://doi.org/10.1016/j.parco.2021.102861

[74] Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. 2022. Why 90to improve it? *Acta Pharmaceutica Sinica B* 12, 7 (2022), 3049–3062. https://doi.org/10.1016/j.apsb.2022.02.002

[75] Pedro H. M. Torres, Ana C. R. Sodero, Paula Jofily, and Floriano P. Silva-Jr. 2019. Key Topics in Molecular Docking for Drug Design. *International Journal of Molecular Sciences* 20, 18 (Sept. 2019), 4574. https://doi.org/10.3390/ijms20184574

[76] Andreas Truszkowski, Mirco Daniel, Hubert Kuhn, Stefan Neumann, Christoph Steinbeck, Achim Zielesny, and Matthias Epple. 2014. A molecular fragment cheminformatics roadmap for mesoscopic simulation. *Journal of Cheminformatics* 6, 1 (Oct. 2014). https://doi.org/10.1186/s13321-014-0045-3

[77] Jianyong Wang, Lei Zhang, Quan Guo, and Zhang Yi. 2017. Recurrent neural networks with auxiliary memory units. *IEEE transactions on neural networks and learning systems* 29, 5 (2017), 1652–1661.

[78] Zhe Wang, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. 2016. Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics* (2016). https://doi.org/10.1039/c6cp01555g

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

[80] Haizhen A. Zhong and Suliman Almahmoud. 2023. Docking and Selectivity Studies of Covalently Bound Janus Kinase 3 Inhibitors. *International Journal of Molecular Sciences* 24, 7 (March 2023), 6023. https://doi.org/10.3390/ijms24076023

[81] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N. Zare, and Patrick Riley. 2019. Optimization of Molecules via Deep Reinforcement Learning. *Scientific reports* 9, 1 (2019), 10752–10752.

APPENDICES

A PARAMETERS AND HYPERPARAMETERS

The Reinforcement Learning agent used in this work was initialized to have the same network architecture and policy as the prior. The prior achieved its policy as described in Section 3.1.1. This policy is modified using transfer learning, as described in Section 3.1.2. The models produced by this research can be found on this paper's GitHub Repository.

A.1 Prior

The prior is trained using MLE on the ChEMBL dataset [25]. This training set consists of 1.5 million molecules constrained to having between 10 and 50 heavy atoms. These ECFPs contain only atoms and elements in the set $S = H, B, C, N, O, F, Si, P, S, Cl, Br, I$.

Table A1: Prior Framework

	Hyperparameter	**Values**
Architecture	RNN with 3 hidden layers	512 LSTM nodes per layer
	Regularization	Droupout in each layer
	Dense output layer	Softmax activation function
Optimization	Adam [14]	$\beta 1 = 0.9$ $\beta 2 = 0.999$ $\epsilon = 10^{-8}$
Training	Episodes	50 000
	Batch Size	128
	Learning Rate	0.001
	Learning Rate Decay	0.02 every 100 steps
	Forget Gate Bias	0.5
	Gradient Clipping	Range=$[-3, 3]$

A.2 Agent

The agent leverages the prior's pre-trained policy during CL, which it fine-tunes using the parameters in Tables A2 and A3. The reward for a scaffold, $R(\mathbf{T})$), ranges between 0 and 1, a value that is too minuscule to modulate the likelihood according to the desirability of the SMILES string sequence. Therefore, it is multiplied by a σ scalar factor. The score for a sampled scaffold is defined by the scoring functions in Appendix B. In the case of a diversity filter, the sampled molecule may be penalized by its Bemis-Murcko scaffold [7].

Table A2: Curriculum Phase

Hyperparameter	**Description**	**Values**
Curriculum Tasks	Scoring Function	Thresholds
	Tanimoto Similarity (Equation 9)	$[0.1875, 0.25]$
	QED (Equation 10)	$[0.375, 0.5]$
	Synthetic Accessibility (Equation 11)	$[3.75, 5]$
Optimization	Adam [14]	$\beta 1 = 0.9$ $\beta 2 = 0.999$ $\epsilon = 10^{-8}$
Training	Maximum Episodes	500
	Batch Size (n)	128
	Learning Rate	0.0001
	σ	128
Diversity Filter	Bemis-Murcko scaffold [7]	Default values as per REINVENT[29]
	Bucket Size (b)	25
	Minimum Similarity (ξ)	0.4
Experience replay	Initialization	Empty
	Memory Size (s_{incept})	100
	Sample Size	10
	Minimum Score (α)	0.4
	Inception Threshold ($\vec{i}_{thresh}$)	Determined by minimum score of scaffolds in the replay memory buffer

Table A3: Production Phase

Hyperparameter	Description	Values
Optimization	Adam [14]	$\beta 1 = 0.9$ $\beta 2 = 0.999$ $\epsilon = 10^{-8}$
Training	Maximum Episodes	100
	Batch Size (n)	128
	Learning Rate	0.0001
	σ	128
	Diversity Filter	None
Scoring Function	DockStream	Estimation of DS
	Ligand embedding	LigPrep
	Docking Backend	Glide
Scoring Function Transformation	Reverse sigmoid	$1/(1+e^{kx})$
	Range of transformation	$[-11, -5]$
	Steepness (k)	0.25
Experience replay	Initialization	Copied from *Curriculum Phase*
	Memory Size (s_{incept})	100
	Sample Size	10
	Minimum Score (α)	0.4
	Inception Threshold ($\vec{i}_{thresh}$)	Determined by minimum score of scaffolds in the replay memory buffer

B SCORING FUNCTIONS

The following Scoring Functions were used to calculate the reward, $R(\mathbf{T})$, for a SMILES string sampled by the agent during CL. These equations are provided by RDKit and can be accessed via their GitHub Repository. Each scoring function maximizes a particular LBDD property, and can be calculated robustly at high-throughput.

B.1 Tanimoto Similarity

The calculation of Tanimoto Similarity is typically based on SMILES strings, represented as molecular fingerprints [64]. These ECFPs are encoded as binary vectors whose elements have values of 1 or 0, corresponding, respectively, to the presence or absence of molecular fragments [76]. The SMILES string feature set represents molecular structure and properties. These molecular representations are compared computationally using 2D similarity as a metric [46], as per Equation 9. In order to calculate the Tanimoto Coefficient, T_c between two scaffolds, A is defined as the number of features present in molecule $\mathbf{M}_A$. Similarly, B is defined as the number of features present in molecule $\mathbf{M}_B$. The variable C is the number of features shared by molecules $\mathbf{M}_A$ and $\mathbf{M}_B$.

$$T_c(\mathbf{M}_A, \mathbf{M}_B) = \frac{C}{A + B + C} \tag{9}$$

B.2 Quantitive Estimate of Drug-Likeness

QED allows ECFPs to be ranked in terms of drug-likeness [8] according to the following desirability functions:

- molecular weight
- octanol-water partition coefficient
- number of hydrogen bond donors
- number of hydrogen bond acceptors
- molecular polar surface area
- number of rotatable bonds
- number of aromatic rings
- number of structural alerts

These functions are derived empirically by describing the underlying property distributions of 771 approved drugs, obtained from the ChEMBL DrugStore database [58]. These molecular properties were based on their ability to influence the likelihood of attrition [65]. The geometric mean of the ADS functions are calculated as per Equation 10, where d_i is a desirability function corresponding to molecule $\mathbf{M}$.

$$QED(\mathbf{M}) = exp(\frac{1}{n}\sum_{i=1}^{n} \ln d_i) \tag{10}$$

B.3 Synthetic Accessibility

An estimation of the synthesizability of a molecule based on the combination of molecule complexity and fragment contributions obtained by analyzing structures 934,046 already-synthesized molecules from the PubChem database [38]. Fragment frequency represents ease of synthesizability, based on the assumption that more frequently-occurring substructures in the database are more easy to synthesize [46]. A frequency distribution of fragments is calculated as a logarithm of the ratio between the actual fragment count and the number of fragments forming 80% of all fragments in the database [21]. Therefore, the $fragmentScore$ in Equation 11 is defined by how frequently the substructures in molecule **M** occur in the frequency distribution, where less frequent fragments have negative scores.

$$SAScore(\mathbf{M}) = fragmentScore - complexityPenalty \tag{11}$$

Similarly, the $complexityPenalty$ is calculated based on Equation 12. $ringComplexity$ penalizes complex ring systems found in molecule **M**, based on the detection of spiro rings and ring fusions [59]. $stereoComplexity$ is based on whether there exists stereogenic centers within the molecule [44]. The penalty for presence of macrocycles is determined when number of aromatic rings within the molecule exceeds 8 [72]. Lastly, the **M** receives a $sizePenalty$ as the length of its SMILES string increases.

$$complexityPenalty(\mathbf{M}) = ringComplexity + stereoComplexity + macrocyclePenalty + sizePenalty \tag{12}$$

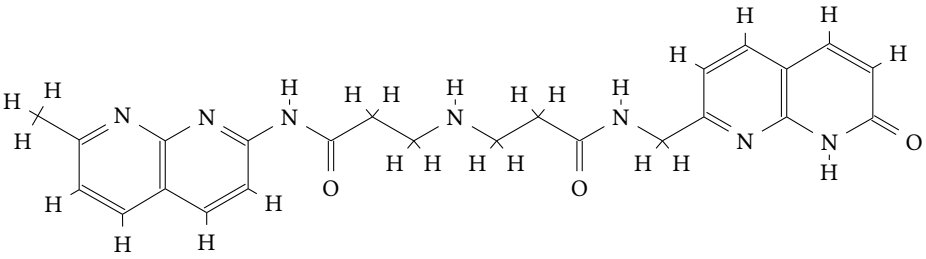
The *SAScore* for molecule **M** is ranked between 1 (difficult to synthesize) and 10 (easy to synthesize), and efficiently discriminates feasible ligands from infeasible ones.

C SCHRÖDINGER SOFTWARE SUITE

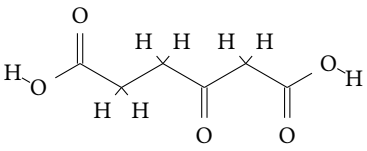
The license for the Schrödinger Software Suite, as well as the computational resources required such as GPUs, were provided by the CSIR's Centre for High Performance Computing (CHPC).

C.1 Preparation of Receptor Grid

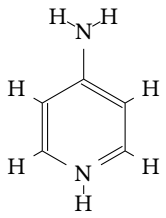
The interaction between the ATP-binding pocket of PI4KIIIβ in complex with *Rab11a-GTP gammaS* [12] was assessed using Maestro's *Protein Preparation Wizard* [70]. The PDB file (**PDB ID: 4d0l** [11]) was accessed from the RCSB Protein Data Bank [40]. The receptor-ligand complex was pre-processed with PROPKA hydrogen bond optimization at pH 7.4 and minimized using OPSL3e force-field [80]. A grid file was generated for the PI4KIIIβ receptor using the Glide Grid Generation protocol. The bound PIK-93 (**PDB ID: 093**) ligand was used as the centroid of the protein binding pocket. The docking grid, along with the empty PI4KIIIβ binding cavity, was saved in order to be used by Glide [62].



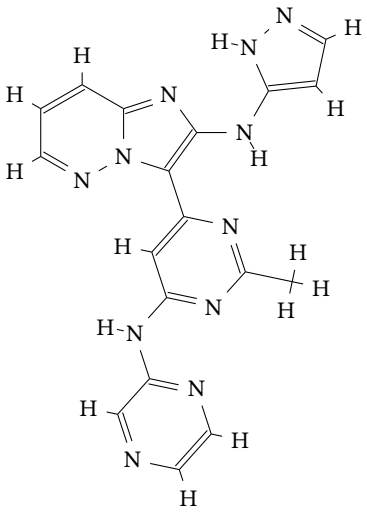
Napthyridine (NAZ)



PIK-93 (093)



Aminopyridine (4AP)



Imidazopyridazine (0U0)

Figure 6: Pre-existing PfPI4K Inhibitors

C.2 Glide Configuration

After target preparation [3] and receptor grid generation, ligand embedding [78] occurred using console mode, allowing SMILES strings generated by REINVENT [29] to be input to DockStream [31]. DSs were returned for each low-energy conformer in the ligand embedding pool. DSs are based on the ChemScore function of Eldridge et al. [20], in which lower GlideScores denote a greater predicted binding affinity.

Table A4: Molecular Docking Simulation Parameters

Hyperparameter	Description	Values
Ligand Preperation	Embedding Pool	LigPrep [67]
	Target pH	7.0 +/- 1.0
	Force Field	OPLS3e
Glide Configuration		
	Allow Amide Distortion	True
	Conformational Sampling	2
	Poses Per Ligand	3
	Apply Strain Correction	True
	REward Intra-Hydrogen Bonds	True

GLOSSARY

Curriculum Progression Criteria For each task in an agent's curriculum, a scoring function is provided as an evaluation metric to decipher the degree to which the agent has learned a successful policy for completing the task. When the average score for a given Curriculum Objective surpasses a pre-defined threshold, the agent has met the *progression criteria* and can begin a new, independent RL loop to learn the next task.

Production Phase When a trained agent has completed all its *Curriculum Progression Criteria*, it enters a new RL loop, in which it must produce novel compound ideas and as well as undergo further training to complete the overarching *Production Objective*.

A priori Latin: "from what comes before" / "from first principles, before experience". Refers to reasoning and deductions which follow from theory rather than empirical data.

Hit-to-lead optimization Hits identified from screening are modified to increase their Binding Affinity (BA) for the target receptor. This process is completed in order to identify promising lead compounds.

In vitro Latin: "in the glass". Refers to experiments conducted outside of a biological context, usually in some form of assay.

Antagonist A drug that binds to a target receptor and makes it unable for the agonists to bind to the cell receptor appropriately.

Antimalarial A compound with properties that cause the inhibition of the *Plasmodium falciparum* parasite growth within infected erythrocytes.

Average Convergence Rate The rate at which the approximation error of an algorithm approaches zero, measured per generation. It is defined as the geometric average of the reduction rate of the approximation error over consecutive generations.

Backpropogation An algorithm for calculating the gradient of a loss function with respect to each modifiable weight, then adjusting the said parameter accordingly.

Bemis-Murcko scaffold The hierarchy of a molecule, defined by its ring structures and linker fragments connecting rings.

Binding affinity The extent to which a drug binds to a target receptor at any given concentration, otherwise known as the "firmness" with which the drug binds to the receptor.

Binding cavity The final orientation of amino acid residues around a reference ligand defines the binding cavity and the docking search space.

Bioactivity The ability for a substance to influence an organism, tissue or cell. The bioactivity of ligands is assessed through Docking Score (DS) scores to determine how well a ligand can bind to a target receptor, and thus the effectiveness of the ligand.

Bucket A collection of scaffolds. A given compound is added to the bucket if its scaffold possesses a similarity above a user-defined threshold. If the similarity is below the threshold, a new bucket is created to contain the scaffold.

Chemical space The ensemble of all organic molecules to be considered when searching for new drugs.

Chemotype A particular class of molecules characterized by their common molecular scaffold.

Conformation The orientation of a ligand to the binding cavity such that the "best fit" is achieved, in order to reduce the free energy of the overall system.

Convergence When the loss of a machine learning model settles within a near-zero range during training. At this point, there are little to no changes in the model's learning rate.

Curriculum A list of *Curriculum Objectives*, each with their own *Curriculum Progression Criterion*, that the agent must meet in order to progress to the next *Curriculum Objective*. The goal of the curriculum is to accelerate the *Production Phase*.

Curriculum Learning A reward shaping algorithm that optimizes the training process of an RL agent, specifically by providing an agent with training examples that become more difficult as the agent progresses through its curriculum. The ultimate goal is to minimize the loss on the final goal, the *Production Objective*.

Deep Belief Network Deep Belief Networks combine unsupervised learning principles and neural networks. The architecture consists of layers of Restricted Boltzman Machine (RBM)s, which are trained one at a time in an unsupervised manner. The output of one RBM is used as the input to the next RBM, and the final output is used for supervised learning tasks such as classification or regression.

Deep neural network A neural network that selects the best action for an agent to take based on the maximum Q-value of the next state. The Q-Network is optimized towards a target network that is periodically updated with the latest weights every *(*k) steps, where *(*k) is a hyperparameter.

Diversity filter A bucket is used to collect all unique SMILES string that have a score above a user-defined threshold. At the end of the RL run, all collected SMILES in the bucket are outputted to a file. The score for recurring scaffolds are penalized, thus forcing the agent to become explorative.

Diversity Oriented Synthesis The generation of a small-molecule collection with a high degree of structural, and thus functional, diversity that interrogates large areas of chemical space simultaneously.

Docking Score The scalar value returned from a docking simulator which provides an estimate of the Binding Affinity (BA) between a ligand and target receptor.

Double Deep Q Network A deep neural network that implements .

Drug design Consisting of SBDD and LBDD, it is the identification and design of new inhibitors or the optimization thereof.

Episode A sequence of actions taken from an initial state, ending in a terminal state, in which the agent receives a reward. In a single episode, the agent interacts with the environment according to a particular policy.

Experience replay A replay memory technique used inRL where the agent's experiences are stored at each

time-step, pooled over many episodes into a replay memory buffer. This replay memory is sampled randomly for a minibatch of experience, and use this to learn off-policy, as with Double Deep Q Networks and Deep Belief Networks.

Exploitative When an agent takes the greedy approach when exploring its RL environment. The agent tries to maximize its immediate reward by capitalizing on knowledge already gained.

Explorative An agent that is more inclined towards exploring its environment, thus having the primary focus of improving knowledge about each action instead of achieving immediate reward. To maximize long-term benefits in an RL framework, the agent is more likely to discover new features in chemical space.

Fine-tuning An approach to fine-tuning, in which a pre-existing neural network, with pre-defined weights, is exposed to new data. The parameters in the neural network are thus adjusted to perform well on a new task.

Fingerprint The bitstring representation of a chemical structure.

Functional analogues Molecules which share a structure that is similar enough that they can be expected to exhibit similar chemistry.

Generative model A ML model that is is capable of automatically discovering regularities in input data through the use of unsupervised learning, in order to generate new examples.

GlideScore The Docking Score returned by Schrodinger's Glide molecular docking simulator after post-docking minimization and strain correction have been calculated for the binding pose of the ligand-receptor complex.

Ground truth Actual, expected output from the training dataset at current time step.

Half maximal inhibitory concentration An indication of how much drug is needed to inhibit a biological process by half. A measure of ligand potency and efficacy.

Heuristic biasing In the context of an optimization problem, biasing is the modification of the objective function to prioritize certain computational paths over others in order to address a problem in an acceptable time frame.

Hyperparameter Parameters that are explicitly defined to control the learning process in order to improve the learning of the model. These values are set before starting the learning process of the model.

Inhibitor A molecule that binds to an enzyme and blocks its activity.

Ligand In the context of drug design, this is a molecule which binds to a target receptor. Note that this is slightly different to the way the term is used in coordination chemistry, where it refers to a molecule which binds to a metal atom forming a coordination complex.

Ligand Based Drug Design The ADD approach used in the absence of 3D target receptor information, relying on knowledge of pre-existing ligands known to bind to the biological target of interest.

Ligand embedding The generation of 3D molecular configurations for the ligands that are to be docked, thus defining an initial ligand conformational state that can affect docking search space traversal.

LogP The partition coefficient of a molecule between aqueous and lipophilic phases usually considered as octanol and water, defining a drug's solubility.

Loss A loss function used when optimizing an Machine Learning (ML) model's prediction capabilities, commonly used to measure the difference between two probability distributions.

Markov Decision Process The way an agent makes decisions whilst inhabiting an environment that changes state randomly in response to action choices made by the agent. The state of the environment affects the immediate reward obtained by the agent, as well as the probabilities of future state transitions. The agent's objective is to select actions to maximize a long-term measure of total reward.

Minibatch Minibatches are equally-sized subsets of the dataset being used to train a model. The gradient, or error in prediction, is calculated over the mini-batch and weights updated.

Mode collapse When the model distribution collapses to cover only a few samples from the training data. Therefore, repetitive scaffolds are sampled from chemical space.

Molecular docking An *a priori* simulation of binding between a target macromolecule and a ligand, using a Docking Score to rank a compound's specificity against a particular target.

Molecular fragment A molecular fragment structure for a chemical compound consists of molecular fragments which are connected by harmonic springs where a single fragment may have multiple connections to other fragments.

Naïve An uninformed agent has a reasoning process governed by lower *Curriculum Progression Criteria* than that of a sophisticated agent. It will select the best present action based on a scarce number of initial beliefs, as it has not been constrained to a policy that requires the production of high-scoring scaffolds.

Parameter The coefficients of the ML model, chosen by the model itself. These coefficients are optimized by the model itself during the learning algorithm, in order to minimize prediction error.

Pharmaceutical Relating to the preparation of medicinal drugs

Pharmacodynamics The study of the concentration of the drug at the receptor site upon administration, its intensity, and duration of action.

Physico-chemical These are properties that are both physical and chemical in nature

Policy A policy π(s) comprises the suggested actions that the agent should take for every possible state s$\in$ S. It is thus a strategy the agent uses to pursue goals.

Pose The position and orientation of the ligand relative to the target receptor, including the core conformation and rotamer group conformations.

Post-docking minimization Minimization optimizes bond lengths and angles, as well as torsional angles, and re-scores the top-ranked poses using the scaled Coulomb-van der Waals term and the GlideScore.

Prior A pre-trained RNN model with a vocabulary of manually curated bioactive molecules with drug-like properties. This network architecture is used as a starting point for the RL agent.

Productivity By exploiting knowledge retention, the agent is able to find favourable areas of chemical space at a faster rate. Thus, more favourable compounds can be sampled in a reduced amount of time.

Quantitative Estimate of Drug Likeness The qualitative measure of drug-likeness, directly proportional to the desirability of a molecule. It reflects the underlying distribution of molecular properties such as molecular weight, LogP, and the presence of unwanted chemical functionalities.

Reinforcement Learning A feedback-based ML technique in which an agent learns to behave in an environment by taking actions that will maximize a pre-defined reward.

Replay memory buffer A data structure that temporarily saves the agent's observations, allowing experiences to be updated multiple times whilst the agent is trained.

Reward shaping In order to ameliorate the Sparse Rewards Problem, small intermediate rewards are created in order to accelerate algorithm convergence.

Rollout trajectory The sequence of episodes that provide the RL agent with experience, at the end of which the agent either satisfies the criteria for convergence or its training is terminated.

Scaffold The core structure of a compound or series, defining its molecular topology.

Selectivity The property of a lead compound that improves its binding affinity towards the target receptor and lowers its affinity towards off-target molecules.

SMILES string Simplified molecular-input line-entry system (SMILES) is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings.

Sophisticated An informed agent selects the best present action based on an extensive amount of initial beliefs, based on having been exposed to high *Curriculum Progression Criteria* during its *Curriculum Phase*. The opposite of a naïve agent.

Sparse Rewards Problem An issue usually associated with RL algorithms, in which an agent does not receive a sufficient amount of feedback from its environment when taking an action.

Stereoisomer Oan organic molecule containing different types of isomers within itself, each with distinct characteristics that further separate each other as different chemical entities having different properties.

Stochasticity When probability and randomness are used to produce predictions.

Strain correction These terms are evaluated by optimizing each ligand pose as a free ligand, first with constraints on all torsions, then without these constraints. The difference is used to compute a penalty term for unreasonably high strain. The strain correction is only added if it is above a threshold, and the excess strain above this threshold is scaled before adding it to the GlideScore.

Structural analogue Molecules that share similar scaffolds but with different substituents and a sufficiently similar molecular backbone.

Structure Based Drug Design The ADD approach in which the structural information of the drug target is exploited for the development of its inhibitor.

Synthesizability The measure Synthetic Accessibility based on structure complexity and similarity, as well as synthetic pathways.

Synthetic Accessibility The measure of the ease of synthesis of a chemical compound.

Tanimoto Similarity The measure of how similar two molecules are, given their molecular fingerprints, based on common bits and the resulting value of the Tanomoto coefficient.

Target preparation The refinement of a protein crystal structure, often involving adding missing hydrogen atoms, defining side chain ionization and tautomeric states, and minimizing the conformational energy.

Target receptor The region within a macromolecule where another, smaller molecule will bind thus operating as a transducer of biological signals.

Teacher forcing A method for training Recurrent Neural Network (RNN)s that make use of the ground truth from a prior time step as input. Therefore, the output from the previous time step in the sequence-to-sequence model is replaced, thus stabilizing the training process by preventing error accumulation in the generated sequence, which can occur when the model's previous output is incorrect.

Terminal state The final step in an agent's rollout trajectory.

Transfer learning The reuse of an existing model to solve a new task. Pre-trained layers are frozen to avoid loss of knowledge and new, trainable layers are added. This is unlike fine-tuning, where only certain layers are frozen for knowledge retention.

ACRONYMS

ACR Average Convergence Rate
ADD Automated Drug design
ADS Asymmetric Double Sigmoidal
AI Artificial Intelligence

BA Binding Affinity
BPTT Back Propogation Through Time

CL Curriculum Learning

DBN Deep Belief Network
DDQN Double Deep Q Network
DL Deep Learning
DNDD De Novo Drug Design
DOS Diversity Oriented Synthesis
DS Docking Score

ECFP Extended-Connectivity Fingerprints

GDRL Goal-Directed Reinforcement Learning

HTVS High-Throughput Virtual Screening

IC50 Half maximal inhibitory concentration

K Steepness of the reverse sigmoid function

LBDD Ligand Based Drug Design
LSTM Long Short-Term Memory

MDP Markov Decision Process
ML Machine Learning
MLE Maximum Likelihood Estimation
MPO Multi-Parameter Optimization

NLL Negative Log-Likelihood
NLP Natural Language Problem

PBRS Potential Based Reward Shaping
PDK-1 *Phosphoinositide-dependent kinase 1*
PfPI4K *Plasmodium falciparum phosphatidylinositol 4-kinase* enzyme
PI4KIIIβ Type III Beta *phosphatidylinositol 4-kinase*

QED Quantitative Estimate of Drug Likeness

RBM Restricted Boltzman Machine
RL Reinforcement Learning
RNN Recurrent Neural Network

SA Synthetic Accessibility
SBDD Structure Based Drug Design