Improving *De Novo* Molecular Design with Curriculum Learning

Jeff Guo[§,⊥], *Vendy Fialková[§,⊥], Juan Diego Arango[§], Christian Margreitter[§], Jon Paul Janet[§], Kostas Papadopoulos[§], Ola Engkvist[§,€], Atanas Patronov[*§]*

§ Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg 43183, Sweden

€ Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg 41756, Sweden

* Corresponding author: atanas.patronov@astrazeneca.com

⊥ These authors have contributed equally

Reinforcement learning (RL) is a powerful paradigm that has gained popularity across multiple domains. However, applying RL may come at a cost of multiple interactions between the agent and the environment. This cost can be especially pronounced when the single feedback from the environment is slow or computationally expensive, causing extensive periods of nonproductivity. Curriculum learning (CL) provides a suitable alternative by arranging a sequence of tasks of increasing complexity with the aim of reducing the overall cost of learning. Here, we demonstrate the application of CL for drug discovery. We implement CL in the *de novo* design platform, REINVENT, and apply it on illustrative *de novo* molecular design problems of different complexity. The results show both accelerated learning and a positive impact on the quality of the output when compared to standard policy based RL. To our knowledge, this is the first application of CL for the purposes of *de novo* molecular design. The code is freely available at https://github.com/MolecularAI/Reinvent.

Introduction

The application of deep learning for drug discovery provides potential to accelerate therapeutics development. One fundamental challenge in any drug discovery campaign is *de novo* molecular design, involving the design and prioritization of candidate molecules for experimental validation.[1,2] *De novo* molecular design entails a multi-parameter optimization (MPO) search in chemical space, estimated to be in the range of 10^{23} to 10^{60} molecules.[3] Recently, deep learning has been applied towards more efficient methods of sampling chemical space such that it is possible to identify promising candidate molecules faster. Deep generative models using policy based reinforcement learning (RL)[4–10], value based RL[11], learning a molecular latent space[12], and other methods including tree search[13] and genetic algorithms[14–16] have been proposed to generate molecules that possess a desired set of properties. In the policy based RL paradigm, an agent (a generative model) learns a policy (series of actions to take at given states) to generate molecules that maximize a reward which is typically computed based on a pre-defined reward function.[4–10] Often, physics-based approximations of binding affinity such as molecular docking are included as a component in the reward function in order to design candidate molecules with enhanced predicted potency. Given sufficiently long training time, these models can learn to generate molecules which satisfy the desired MPO objective. However, in cases with complex reward functions where minima are difficult to find, the resulting small gradients elicit minimal change to the agent policy. Consequently, the agent may spend many epochs sampling from areas in chemical space that are far away from the desired objective. The issue is exacerbated when computationally demanding components are included in the reward function, such as molecular docking. Thus, policy based RL can be infeasible for complex MPO objectives, leading to suboptimal allocation of computational resources and eventually suboptimal molecules identified for synthesis.

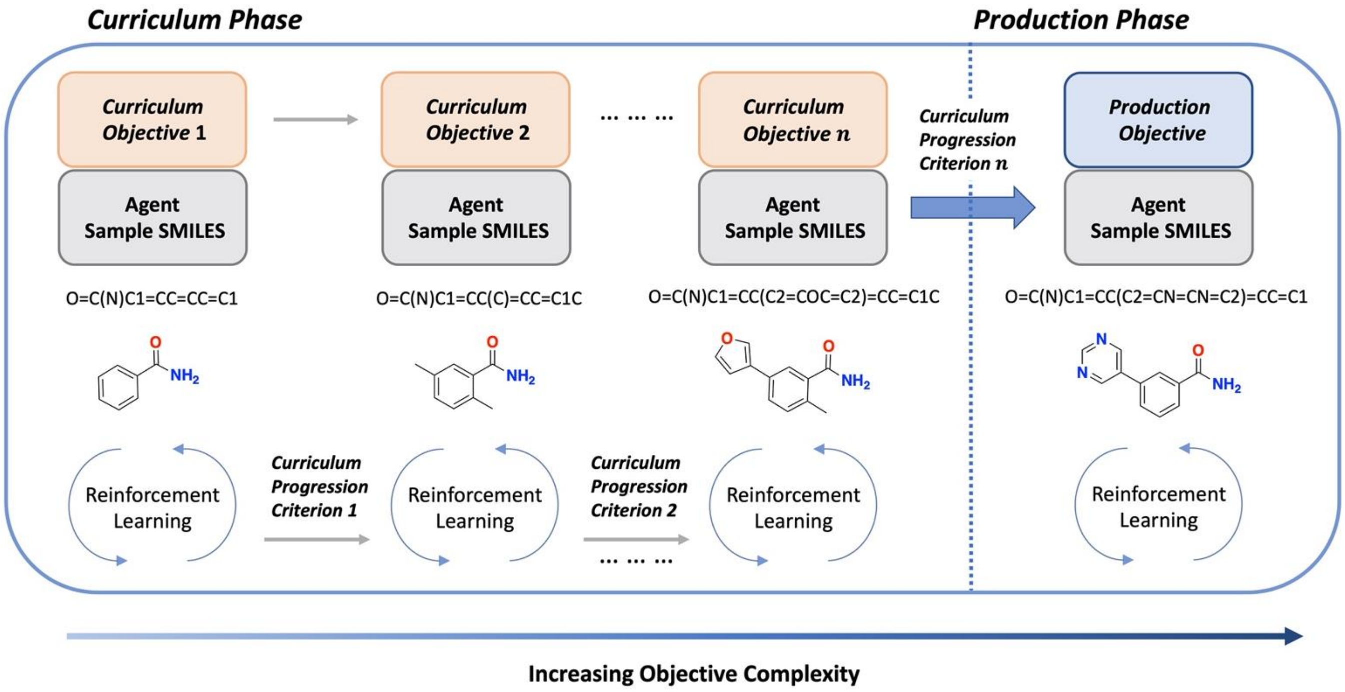

Fig. 1 Curriculum learning overview. In the *Curriculum Phase*, the agent progresses through successive *Curriculum Objective*s that gradually increase in complexity. The agent samples compounds in the SMILES format through a RL cycle such that the conditional probabilities are updated to maximize the reward obtained based on a scoring function comprised of the current *Curriculum Objective*.[17] *Curriculum Progression Criterions* check for sufficient learning of each *Curriculum Objective* based on a threshold that the agent must achieve. If and only if the final *Curriculum Progression Criterion* is satisfied does the agent progress to the *Production Phase* in which a scoring function comprised of the *Production Objective* is applied.

Curriculum learning (CL) has been proposed as a training strategy to overcome difficulties in learning complex tasks.[18] The basis of CL is to decompose complex objectives into simpler constituent objectives that are sequentially learned, guiding training towards successful convergence of the final objective. Provided a curriculum where constituent objectives are strongly correlated with the final objective, corresponding gradients from sequential simpler tasks are more effective at traversing the optimization landscape and can accelerate convergence.[19,20] Similarly, CL can be applied to non-gradient based objectives, e.g., presence of a target structural motif, by devising a curriculum with gradually increasing complexity, e.g., decomposing the target structural motif into simpler constituents. *De novo* molecular design often requires optimizing correlated properties that cumulatively define favourable chemical space, e.g., generating known active scaffolds and improving binding affinity.[21] By applying concepts from CL, existing limitations of policy based RL for *de novo* molecular design can be circumvented. CL provides a strategy to lower the learning barrier of complex MPO objectives, reaching a state of *productivity* within a reasonable timeframe.

In this work, we build on the *de novo* molecular design platform, REINVENT, and introduce a CL implementation that can address complex tasks where policy based RL has difficulties to identify suitable molecules due to the complexities of the reward function.[5] The use of CL extends REINVENT's applicability to complex reward functions that were previously infeasible with standard policy based RL. We demonstrate the use of CL in REINVENT by formulating a complex task to design 3-phosphoinositide-dependent protein kinase-1 (PDK1) inhibitors, adopting a structure-based optimization approach.[22] We show that immediate states of *productivity* can be achieved by assembling a curriculum compared to standard policy based RL, which can circumvent high computational costs associated with reward components such as molecular docking. Moreover, we show that CL provides a natural method for agent policy regularization, such that minor changes in the curriculum can steer *de novo* molecular design, enabling control over the quality and diversity of the results in a predictable manner and leading to high quality molecules proposed for synthesis.

Curriculum Learning Setup

Curriculum Learning Formulation. The implementation of CL builds on the REINVENT platform described by Blaschke et al. (see Methods).[5] In CL, a complex task is decomposed into simpler constituent tasks to accelerate training and convergence. The goal is to guide the agent to learn tasks with increasing complexity before ultimately providing the *Production Objective*. Agent Learning progresses through the *Curriculum Phase* to the *Production Phase* (Fig. 1). In the former, the agent is trained on simpler sequential tasks with gradually increasing complexity. In the latter, the agent reaches a state of *productivity*, whereby the agent samples compounds in favourable areas of chemical space that satisfy the *Production Objective*. Agent policy update is maintained in the *Production Phase* to ensure the agent continues to sample favourable compounds from diverse minima. The CL strategy for *de novo* molecular design is formally defined below:

Definition 1: A scoring function, $S\colon \mathrm{SMILES} \rightarrow [0,1]$ is formulated as a weighted geometric mean: $S(x) = \left(\prod_{i=1}^{n} c_i(x)^{w_i}\right)^{1/\sum_{i=1}^{n} w_i}$, where x is a sampled compound in the SMILES format and $c_i\colon \mathrm{SMILES} \rightarrow [0,1]$ and w_i are the ith component and its corresponding weighting, respectively. $S(x)$ computes the *desirability* of the sampled compound, x, and its corresponding gradient is used to update the agent policy.

Definition 2: A *Curriculum*, C consists of a sequence of Objectives, $O = \{O_{C_1}, \dots, O_{C_{n-1}}, O_{C_n}, O_P\}$, where subscripts C and P denote *Curriculum* and *Production Objectives*, respectively. For each Objective, O, there is a corresponding scoring function S to compute the *desirability* of a sampled compound based on the current Objective, e.g., possessing a specific structural motif. Progression through a *Curriculum* is controlled by *Curriculum Progression Criterions*, $P = \{P_1, \dots, P_{n-1}, P_n\}$, such that the *Curriculum*, $C = \{O, P\}$.

Curriculum Phase. In the *Curriculum Phase*, the goal is for the agent to learn to generate compounds that satisfy sequential *Curriculum Objective*s with increasing complexity that guide the agent towards the *Production Phase*. $O_{C_1}, \dots, O_{C_{n-1}}, O_{C_n}$ are designated C*urriculum Objectives* with corresponding *Curriculum Progression Criterions* $P_1, \dots, P_{n-1}, P_n$, that enforces sufficient agent learning of each sequential *Curriculum Objective* based on a score threshold. If the score threshold is met, the agent progresses to the sequential *Curriculum Objective*, otherwise, the agent continues learning the current *Curriculum Objective*. This process collectively constitutes the *Curriculum Phase*.

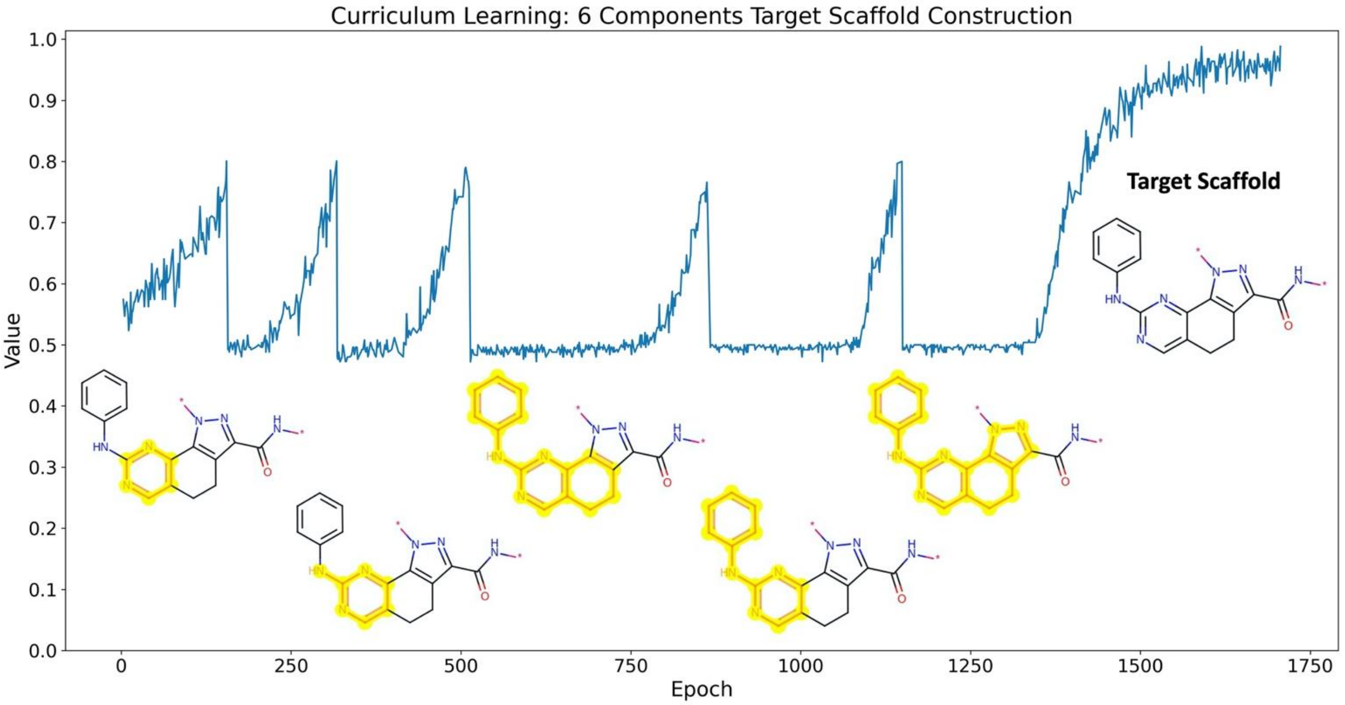

Fig. 2 Curriculum learning target scaffold construction. We define a curriculum where the target dihydro-pyrazoloquinazoline scaffold is decomposed into sequential simpler substructures (highlighted) to guide the agent. The score drops momentarily when a successive substructure objective is introduced as it is unlikely that currently sampled compounds will also possess the new substructure, by chance. By using curriculum learning, the agent is able to find the target scaffold within 1750 epochs while standard policy based RL is unsuccessful in the same number of epochs (see Supporting Information Fig. S1).

Production Phase. If and only if the final *Curriculum Progression Criterion,* P_n is satisfied, the *Production Objective*, O_P, is activated. Presumably, the agent is in a state of *productivity* and samples compounds that satisfy the *Production Objective*. Balance between chemical space exploration and exploitation can be achieved by tuning hyperparameters (see Methods). The agent samples for a pre-defined number of epochs and all compounds that score above a minimum threshold are stored and outputted at the end.

Results

In this section, we devise three experiments to demonstrate the enhanced capability of CL to satisfy complex objectives relative to standard policy based RL:

1. ***Production Objective*****:** Generate compounds that possess a target scaffold. ***Curriculum*****:** Achieve a state of *productivity* by decomposing the target scaffold into simpler sequential substructures with gradually increasing structural complexity.

2. ***Production Objective*****:** Generate compounds that are drug-like and satisfy a molecular docking constraint.[23] ***Curriculum*****:** Achieve a state of *productivity* by first teaching the agent to sample compounds with Tanimoto (2D) similarity to a reference ligand.

3. ***Production Objective*****:** Generate compounds that are drug-like satisfy a molecular docking constraint.[23] ***Curriculum*****:** Achieve a state of *productivity* by first teaching the agent to sample compounds with 3D shape-based similarity to a reference ligand (see Methods for more details on 3D shape-based similarity).[24,25]

For experiments 2 and 3, we further define a "Low" (0.5) and "High" (0.8 for Tanimoto (2D) and 0.75 for the 3D shape-based similarity) scenario denoting the minimum score the agent must achieve with the *Curriculum Objective* activated before proceeding to the *Production Objective*. The purpose of these scenarios is to investigate the effect of variable degrees of agent *Curriculum Objective* knowledge on compound sampling in the *Production Phase* and how it impacts the state of *productivity*.

Experiment 1: Target Scaffold Construction. As an initial example, we show that CL can guide the agent to generate compounds possessing a relatively complex scaffold that is not present in the training set for the prior (Fig. 2). The dihydro-pyrazoloquinazoline scaffold was identified as a promising starting point for PDK1 inhibitor design owing to good cell permeability and low promiscuity.[22] The goal is to generate compounds that possess the scaffold, mimicking an analogue series generation. We first demonstrate that the task is too complex for standard policy based RL and denote this as baseline RL (see Supporting Information Fig. S1). In the baseline experiment, the only component in the scoring function is the dihydro-pyrazoloquinazoline scaffold. Each generated compound scores either 1.0 or 0.5, denoting whether the scaffold is present or not, respectively. The average score of the baseline experiment does not exceed 0.5 across 2000 epochs, indicating the scaffold is not found. Given that the scaffold is not present in the training set, the likelihood of sampling a compound possessing the scaffold is much lower and the inability to do so prevents meaningful agent learning. It is worth noting that provided unlimited time, baseline RL will almost surely find the scaffold due to sampling stochasticity. On the other hand, CL can accelerate convergence by decomposing the target scaffold into simpler substructures with gradually increasing structural complexity (Fig. 2). There are 5 *Curriculum Objective*s, each assigned to successively more complex substructures with *Curriculum Progression Criterion* thresholds of 0.8. The agent is tasked to generate compounds possessing each substructure until the average score is 0.8. When a *Curriculum Progression Criterion* is satisfied, the successive and more complex *Curriculum Objective* is activated. A sharp decrease in average score accompanies each *Curriculum Objective* update, e.g., at approximately epoch 150 (Fig. 2), as it is unlikely that currently sampled compounds will also possess the successive substructure, by chance. Over the course of training, the agent learns to generate compounds possessing increasingly complex substructures until the target scaffold is constructed.

Experiments 2 and 3: Satisfying a Molecular Docking Constraint. Often, one does not only want to generate compounds with a specific target scaffold but is also interested in applying a physics-based approximation of binding affinity as a design criterion, such as molecular docking.[6,7,13–15] By enforcing docking constraints, experimentally validated interactions can be retained, bolstering confidence in the plausibility of potency of the generated compounds. However, it is unlikely a random sampling of molecules will satisfy a docking configuration, especially one that enforces constraints. Consequently, in baseline RL with a complex objective, the agent may rely strictly on stochastic sampling to generate favourable compounds and leverages experience replay to achieve convergence.[5] Problematically, generated compounds that poorly satisfy the objective yield small gradients that elicit minimal agent policy update. If this period of nonproductivity is extensive, the baseline RL experiment can be computationally prohibitive.

In this section, we demonstrate that simple curricula, utilizing a single *Curriculum Objective* can accelerate agent *productivity* and generate compounds that satisfy a docking constraint (see Methods for experiment hyperparameters). Simulating a real-world application where one must allocate limited computational resources, baseline RL and CL performances are compared, given a maximum number of permitted *production* epochs (300), i.e., epochs that involve docking, as these are relatively computationally demanding. For CL, *Curriculum Objective*s are first applied to guide the agent and the number of permitted *curriculum* epochs is not limited, as these are computationally inexpensive (see Supporting Information Table S2). Angiolini et al. design PDK1 inhibitors by leveraging the dihydro-pyrazoloquinazoline scaffold which forms two hydrogen-bonding interactions with Ala 162 (Fig. 3a) that are crucial for potency.[22] The structure-based optimization is mimicked by defining the following *Production Objective*:

***Production Objective*:** Generate compounds that retain the two hydrogen-bonding interactions with Ala 162, possess enhanced predicted potency compared to the reference ligand (as assessed by docking score) and are drug-like, as measured by the Quantitative Estimate of Druglikeness (QED).[23]

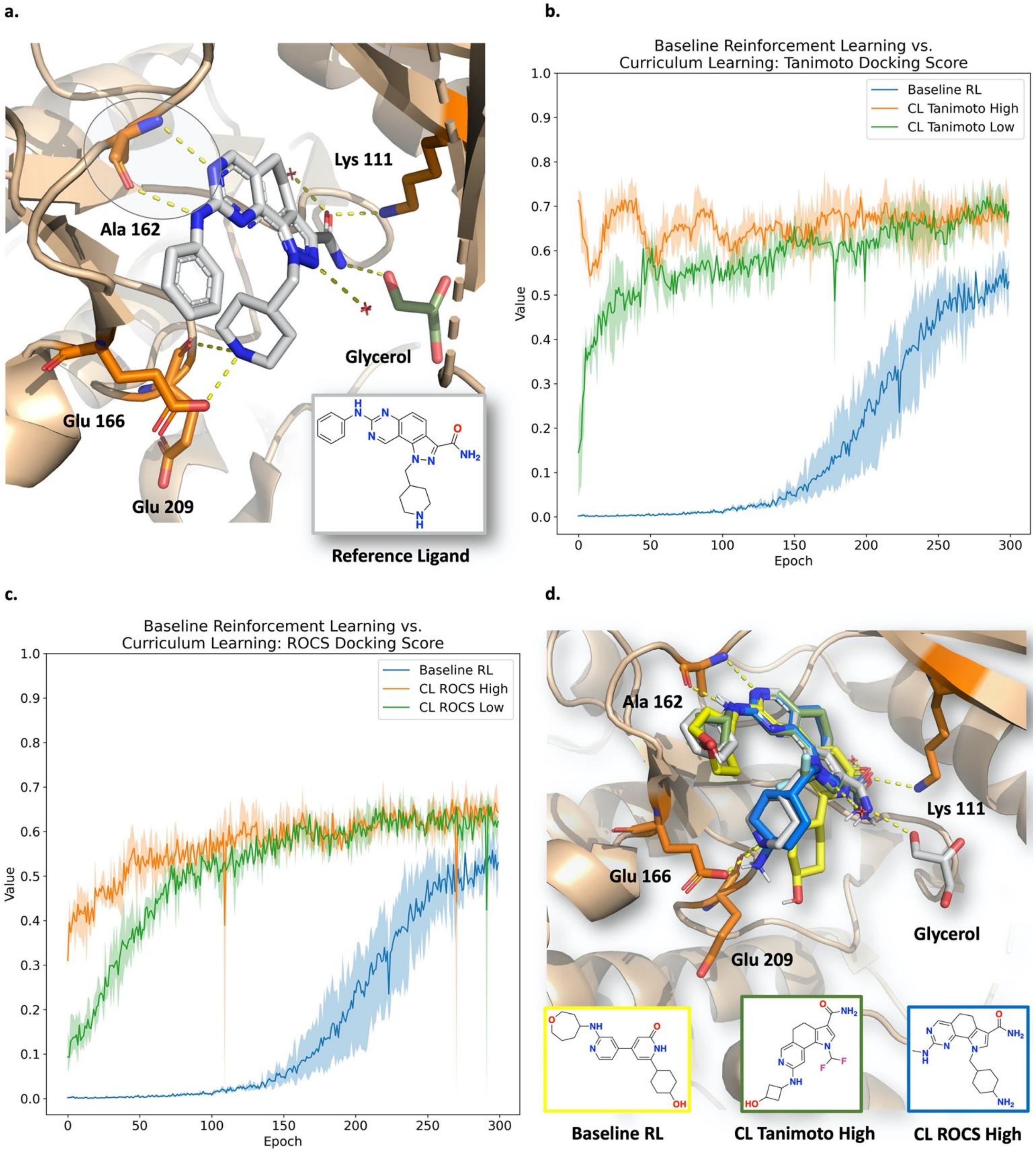

Fig. 3 Baseline reinforcement learning vs. curriculum learning to design PDK1 inhibitors. Values in the plots represent the average over triplicate experiments and the shaded regions are the minimum and maximum values observed. **a.** Reference ligand binding pose (**PDB ID: 2XCH**). Waters and ligand-protein interactions are shown in red and as yellow dotted lines, respectively. The two hydrogen-bonding interactions with Ala 162 are highlighted. The objective is to design compounds that retain the hydrogen-bonding interactions and possess enhanced predicted binding affinity relative to the reference ligand. **b.** Baseline reinforcement learning vs. curriculum learning (Tanimoto *Curriculum Objective*) *Production Phase* docking score. Docking struggles significantly in the baseline RL experiments and only reaches a state of *productivity* after approximately 300 epochs. Curriculum learning using Tanimoto (2D) similarity guides the agent to immediately generate compounds that satisfy the docking constraint. **c.** Baseline reinforcement learning vs. curriculum learning (ROCS *Curriculum Objective*) *Production Phase* docking score. The same observations as **b.** are made. **d.** Top generated compounds from selected experiments

that exceed a total score encompassing docking and QED above a threshold. The predicted poses are superimposed with the reference ligand (grey). The binding poses retain the two hydrogen bonding interactions with Ala 162, as enforced by the docking constraint.

First, we show that the *Production Objective* is challenging for baseline RL (Fig. 3b and Fig. 3c). The docking score is approximately 0 for the first 100 epochs, indicating essentially no compounds sampled satisfy the docking constraint. From epochs 100-200, some compounds satisfy the docking constraint but the score (averaged over all compounds sampled) is still low. It is only from epoch 200 onward that the docking score begins a steep improvement and indicates the point at which the agent starts entering a state of *productivity*. It is evident that baseline RL is not optimal as the agent spends a significant amount of time generating compounds that do not satisfy the *Production Objective*. It is worth noting, however, that the agent eventually converges given enough time (see Supporting Information Fig. S4).

To circumvent the limitations of baseline RL, we devise curricula and introduce 2 *Curriculum Objective*s to guide the agent to *productivity*: Tanimoto (2D) and ROCS (3D) shape-based similarity to the reference ligand.[24,25] In the former, the rationale is that by teaching the agent to first generate compounds with 2D similarity to the reference ligand, subsequently generated compounds will have a greater likelihood of satisfying the docking constraint. The rationale for ROCS is identical except with 3D similarity to match the shape and electrostatics of the reference ligand, providing an opportunity to modify the central core, termed 'scaffold hopping'.[21] Triplicate baseline RL experiments with Tanimoto and ROCS components (using a scoring function comprised of Tanimoto/ROCS, docking, and QED together, respectively) were conducted for a thorough comparison with CL. These baseline experiments did not improve agent *productivity* and similar training progress as the baseline shown in Fig. 3b and Fig. 3c is observed (see Supporting Information Fig. S5 and S6). For the "Low" and "High" Tanimoto scenarios, the agent is immediately capable of generating compounds that satisfy the docking constraint (Fig. 3b). More specifically, although docking starts at a relatively low value (but higher than baseline RL) for the "Low" Tanimoto experiment, the agent quickly improves over the first 50 epochs and continues to do so for the remainder of the experiment. In the "High" Tanimoto scenario, docking starts at a score that exceeds the maximum score achieved by the baseline RL agent over 300 epochs and maintains *productivity*. The results are intuitive as enforcing the agent to first learn to generate compounds with higher 2D similarity to the reference ligand should increase the likelihood of satisfying the docking constraint. Similar observations are made when using ROCS as a *Curriculum Objective* (Fig. 3c). In both the “Low” and “High” scenarios, docking starts more favourably than baseline RL but unlike the Tanimoto experiments, the ROCS experiments start at a less favourable docking score. Firstly, these results are not completely surprising as training the agent to satisfy a 3D shape similarity objective will decrease the likelihood, relative to 2D similarity, in satisfying the docking constraint owing to more potential conformational discrepancies of the generated compounds compared to the reference ligand, and is not without precedent.[26] Secondly, the agent still improves significantly over 100 and 50 epochs for the “Low” and “High” ROCS scenarios, respectively. These results convincingly demonstrate that the improvement in CL performance over baseline RL is attributed to the sequential nature of the CL objectives as opposed to the presence of the additional *Curriculum Objectives* only.

To visualize the quality of the results, the binding pose of the top generated compound (based on total score: docking and QED) from selected experiments is superimposed with the reference ligand (Fig. 3d). The binding poses retain the two hydrogen bond interactions with Ala 162, as enforced by the docking constraint. Furthermore, the superimposed binding poses demonstrate excellent agreement with the reference ligand, supporting plausibility. Thus, we show that using Tanimoto (2D) and ROCS (3D) shape-based similarities to the reference ligand as *Curriculum Objectives* can guide the agent to satisfy a complex *Production Objective* and the results demonstrate CL outperforms baseline RL given the same number of *production* epochs. Moreover, tuning the degree of *Curriculum Objective* optimization, as shown in the “Low” and “High” scenarios, provides direct control in guiding the agent to *productivity*.

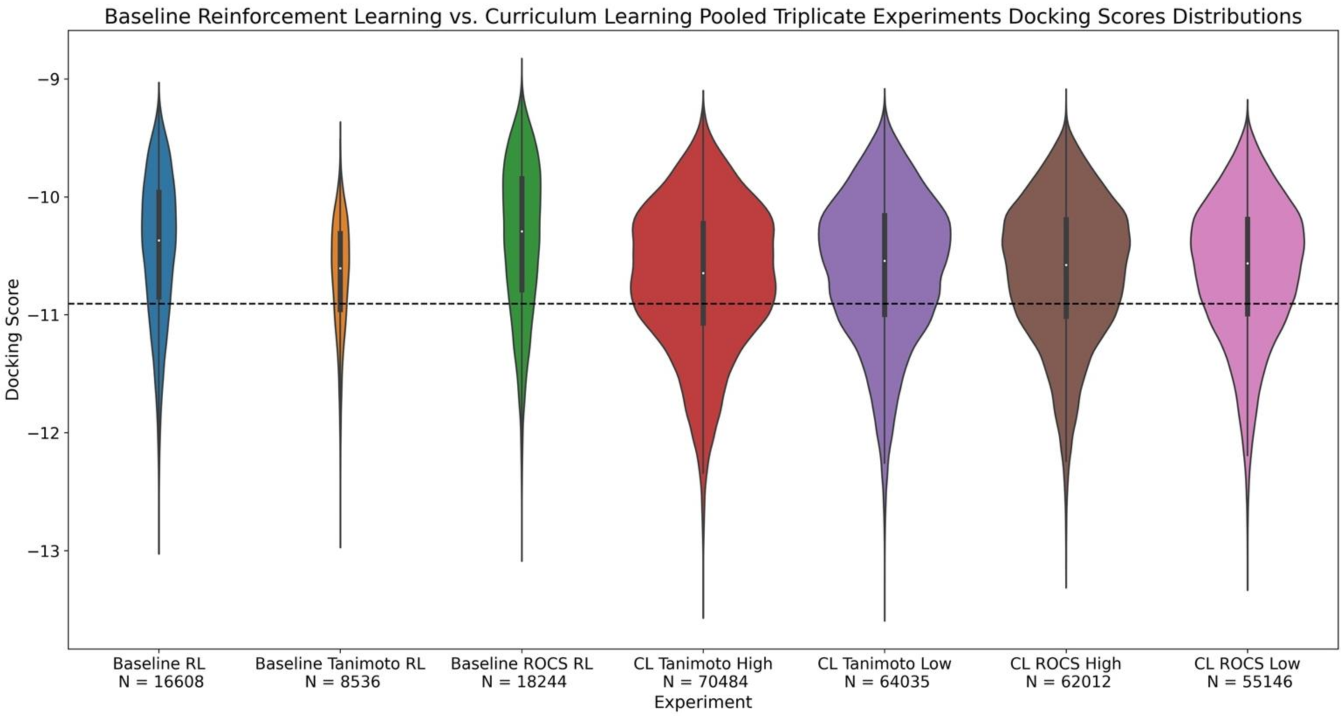

Fig. 4 Baseline reinforcement learning vs. curriculum learning docking scores distribution. RL: Reinforcement Learning and CL: Curriculum Learning. Each individual violin plot represents pooled triplicate experiments. The results shown consist of all the stored compounds from the 300 permitted *production* epochs with the *Production Objective*: docking and QED. 'N' in the x-axis labels is the number of compounds collected (those that exceed a total score encompassing docking and QED above a threshold) in each pooled violin plot. 'Baseline Tanimoto RL' and 'Baseline ROCS RL' refers to baseline reinforcement learning using a scoring function composed of docking, QED, and Tanimoto/ROCS together, respectively. Lower Glide docking scores denote a greater predicted binding affinity. The docking score for the reference ligand is -10.907 kcal/mol and is shown by the horizontal black dotted line. Curriculum learning not only collects more compounds than baseline reinforcement learning but the compounds also possess more favourable docking scores, on average.

***Curriculum Objectives* Enhance Objective Optimization.** To further investigate the output of the baseline RL and CL experiments, all docking scores of the collected compounds were pooled from the triplicate experiments and the resulting distributions are illustrated in Fig. 4. Firstly, CL generates a significantly greater quantity of favourable compounds compared to baseline RL, as only those that pass a minimum score based on docking and QED are stored. This is consistent with Fig. 3b-d where the baseline RL agent struggles for the first 150 epochs, predominantly sampling compounds that do not satisfy the *Production Objective*. Secondly, compounds generated by CL exhibit more favourable docking scores than baseline RL, on average. Thirdly, between the *Curriculum Objectives* (Tanimoto and ROCS), the "High" scenario has a greater density of favourable docking scores (around -11 kcal/mol) compared

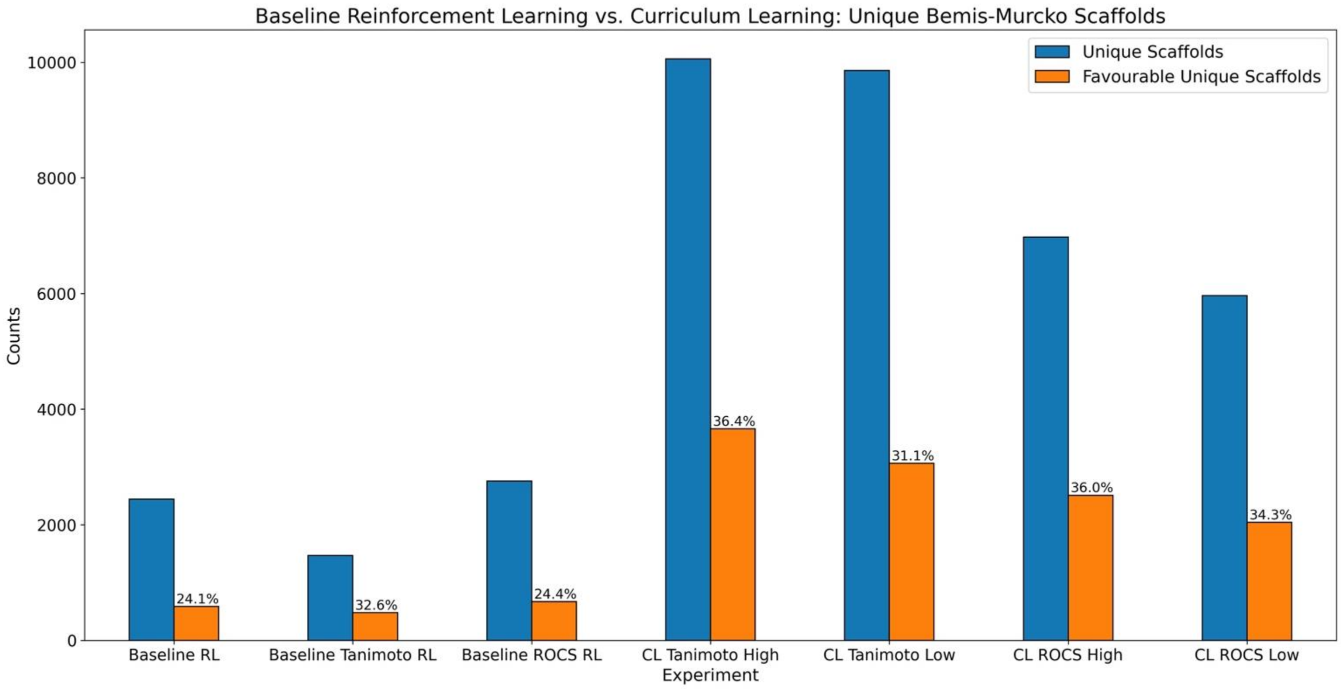

Fig. 5 Baseline reinforcement learning vs. curriculum learning unique Bemis-Murcko scaffolds. RL: Reinforcement Learning and CL: Curriculum Learning. Number of unique Bemis-Murcko scaffolds in the collected compounds (those that exceed a total score encompassing docking and QED above a threshold). Values in the plot represent the average over triplicate experiments (see Supporting Information Table S4 for individual experiment quantities). 'Favourable Unique Scaffolds' denotes the scaffolds that possess a more favourable docking score than the reference ligand. The fraction of 'favourable' scaffolds generated is shown as an annotated percentage.

to the "Low" scenario. To quantify this, the fraction of compounds collected that possess a docking score better than the reference ligand (-10.907 kcal/mol) was calculated for each experiment (see Supporting Information Table S3). The task chosen resembles a potential real-world application as the reference ligand is an experimentally validated nanomolar (nM) inhibitor.[22] In all cases, CL generates between 2941-9068 and between 12.42%-23.79% more compounds that dock more favourably than the reference ligand by absolute counts and percentage, respectively, compared to baseline RL. Furthermore, between the *Curriculum Objectives* Tanimoto and ROCS, the "High" scenario outperforms the "Low" scenario (between 316-3415 and between -0.4%-10.57%) at the same task. Thus, a single *Curriculum Objective* provides a tunable parameter that can enhance and control the degree in which the agent is able to satisfy a *Production Objective*.

***Curriculum Objectives* Maintain Scaffold Exploration.** Scaffold diversity was investigated by extracting and averaging the number of unique Bemis-Murcko scaffolds from the triplicate experiments, shown in Fig. 5. (see Supporting Information Table S4 for individual experiments).[27] It is evident that the CL experiments generate more unique scaffolds than baseline RL. This is expected from the training plots observed in Fig. 3b and 3c where the baseline RL experiments generate essentially no favourable compounds in the first 100 epochs. Between the *Curriculum Objectives*, Tanimoto generates more unique scaffolds than ROCS. Similarly, "High" scenarios generate more unique scaffolds than "Low" scenarios for both Tanimoto and ROCS. To assess the quality of the generated scaffolds, we denote scaffolds 'favourable' if the corresponding compound exhibits a more favourable docking score than the reference ligand. CL generates more unique 'favourable' scaffolds than baseline RL by absolute counts *and* percentage (Fig. 5). This is in agreement with the docking scores distributions in Fig. 4 that illustrate clear enrichment in docking scores for the CL experiments. The results show that using *Curriculum Objectives* increases the number of 'favourable' scaffold ideas generated and maintains agent exploration as enforced by the diversity filter (DF, see Methods).

To further investigate scaffold diversity, the overlap between the unique Bemis-Murcko scaffolds of the pooled triplicate experiments is quantified. In general, replicate experiments result in different datasets with low scaffold overlap (see Supporting Information Fig. S9-11). Interestingly, however, there is no overlap between the pooled scaffolds in both the "Low" and "High" scenarios in the CL Tanimoto and ROCS experiments (see Supporting Information Fig. S10 and S11, respectively). This suggests that tuning the *Curriculum Objective* optimization can guide the agent to different areas of chemical space. In addition, no overlap is observed between the baseline RL experiments and the CL Tanimoto and ROCS experiments (see Supporting Information Fig. S10 and S11, respectively). Taken together, these observations show that CL and variable degrees of *Curriculum Objective* optimization guides the agent to different areas of chemical space compared to baseline RL.

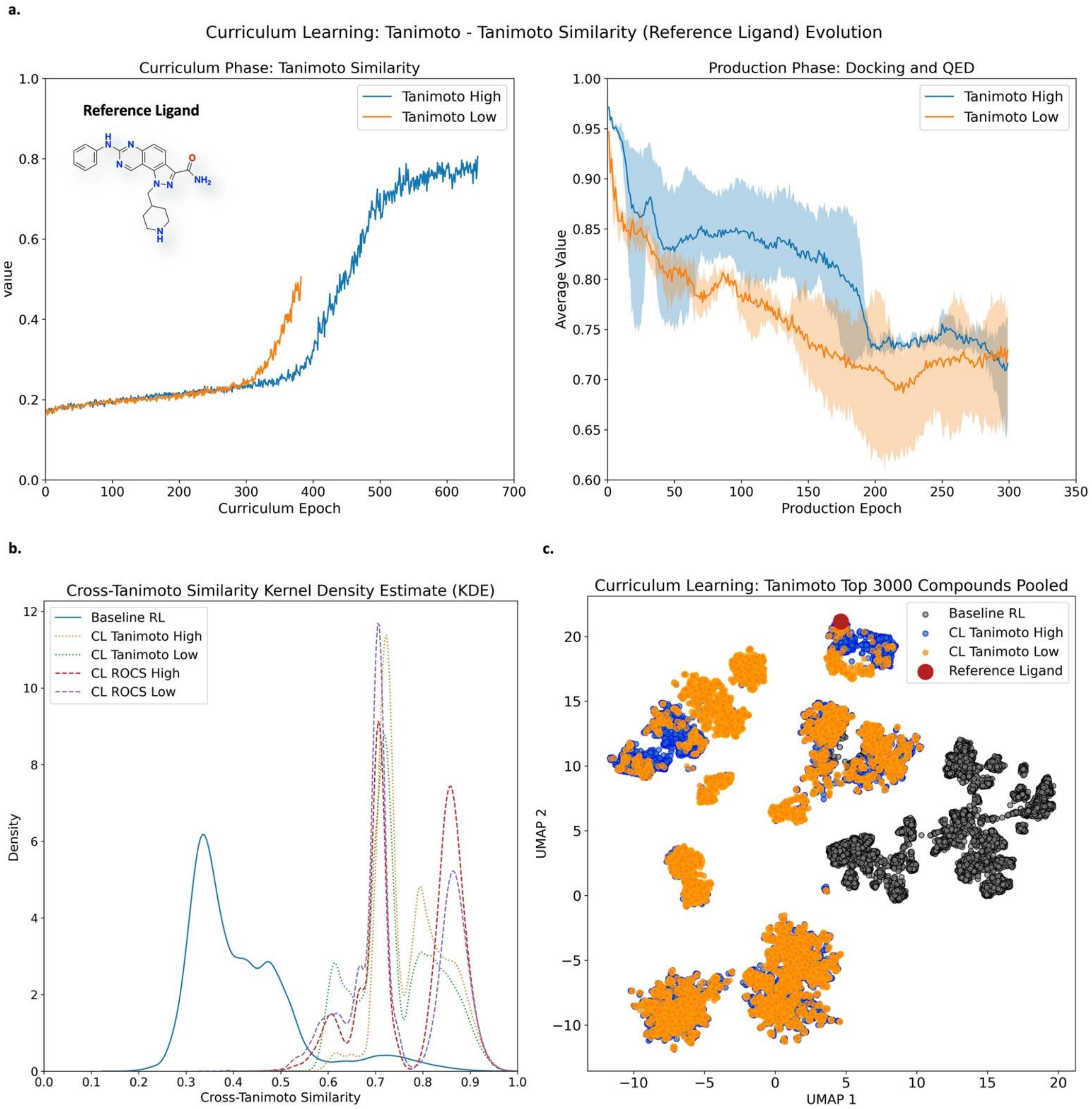

Fig. 6 Agent knowledge retention and effects of *Curriculum Objectives* on the solution space diversity. Values in the plots represent the average over triplicate experiments and the shaded regions are the minimum and maximum values observed. **a.** Tanimoto similarity to the reference ligand evolution. Left subplot depicts the *Curriculum Phase* where the agent is taught to sample compounds with Tanimoto (2D) similarity to the reference ligand. The right subplot depicts the *Production Phase*. In general, "High" Tanimoto experiments sample more compounds that possess greater similarity to the reference ligand. **b.** Cross-Tanimoto similarity for intra-set diversity. The plot shows the pooled collected compounds (those that exceed a total score encompassing docking and QED above a threshold) from the triplicate experiments, in which the overall dataset was reduced in size by a factor of 10 to decrease computation time. Relative to the baseline RL experiments, CL generates compounds with notably greater intra-set similarity. The effect is more pronounced in the "High" scenarios compared to the "Low" scenarios. **c.** Curriculum learning (Tanimoto *Curriculum Objective*) UMAP. The top 3000 compounds were extracted from each triplicate experiment. Overall, the "Low" and "High" scenarios sample from areas 'close' in chemical space, but generally distinct from baseline RL.

Direct Steering of Agent Policy: Trade-off Between *Production Objective* Optimization and Solution Space Diversity. To further elucidate the role of *Curriculum Objectives* and the extent to which the agent retains acquired knowledge in downstream production tasks, the generated compounds from the CL Tanimoto experiments were pooled and the average Tanimoto similarity to the reference ligand calculated for each epoch (Fig. 6a). The left subplot shows the gradual optimization of Tanimoto similarity for the "Low" and "High" scenarios, representing the *Curriculum Phase*. The right subplot shows the Tanimoto similarities for all the compounds that are collected (those that exceed a total score encompassing docking and QED above a threshold) in the *Production Phase*. In general, the compounds generated from the "High" Tanimoto experiments possess a greater Tanimoto similarity to the reference ligand than the "Low" Tanimoto experiments, as expected (see Supporting Information Fig. S13 for distribution of Tanimoto similarities). Interestingly, however, the difference is not drastic and can be explained by cross-referencing the training plots shown in Fig. 3b. The "Low" Tanimoto experiments start at notably lower docking scores than the "High" Tanimoto scenario and suggests that the compounds collected at the beginning are those that happen to exhibit high Tanimoto similarity to the reference ligand. This is further supported by extracting the number of compounds collected at each epoch (see Supporting Information Fig. S14). The "Low" Tanimoto experiments generate less favourable compounds in the first 50 epochs when the *Production Objective* is activated. In contrast, the "High" Tanimoto experiments generate more favourable compounds at every *production* epoch, on average, as the decreased selective pressure encourages the agent to continue sampling from similar areas of chemical space to maintain reward. However, without incentive, beyond the applied DF that penalizes the agent for repeated sampling of the same Bemis-Murcko scaffold to explore more diverse areas of chemical space, the question arises whether the agent becomes overwhelmingly focused on a narrow solution space.

To investigate the effect of *Curriculum Objectives* on the sampled solution space, cross-Tanimoto similarities between each unique compound pair in the pooled datasets were calculated to quantify *how* *different* the collected compounds are to each other (see Methods). Relative to the baseline RL experiments, collected compounds in the CL experiments exhibit greater intra-set similarity, interpreted as the agent sampling compounds from 'closer' areas in chemical space (Fig. 6b). Moreover, the "High" scenarios have a greater density of high cross-Tanimoto similarities than the "Low" scenarios. Uniform Manifold Approximation and Projection (UMAP) was used as a dimension reduction technique to visualize the sampled solution space from the CL Tanimoto experiments.[28] There is notable similarity, without overlap (as there is no scaffold overlap, see Supporting Information Fig. S10), between the compounds sampled from the "Low" and "High" scenarios, with the former spanning some separate and distinct areas (Fig. 6c). The results suggest that moderate optimization of *Curriculum Objectives* (as in the "Low" scenarios) already significantly narrows the agent perceived solution space, in agreement with the cross-Tanimoto similarity distributions shown in Fig. 6b. The similarity between the generated compounds from the "Low" and "High" experiments was quantified by calculating the cross-Tanimoto similarity between the two datasets (see Supporting Information Fig. S16). The majority of cross-Tanimoto similarities is > 0.7, confirming that the generated compounds from both scenarios were sampled from areas 'close' in chemical space (Fig. 6c). Taken together, the observations in this section suggest that devising a curriculum and using *Curriculum Objectives* to guide the agent to a *Production Objective* facilitates knowledge retention that is exploited to achieve a state of *productivity*. However, there is an inverse relationship between using similarity-based *Curriculum Objectives* to enhance *Production Objective* optimization and intra-set diversity, imposing a trade-off when using CL over baseline RL.

Conclusions

In this work, we build on the *de novo* molecular design platform, REINVENT, by adapting curriculum learning (CL) to accelerate agent convergence on complex multi-parameter optimization (MPO) objectives.[5] Relative to baseline reinforcement learning (RL) which may issue many non-productive calls to expensive physics-based descriptors, simple curricula consisting of even one *Curriculum Objective* can successfully guide the agent to achieve *productivity* in substantially reduced time. We demonstrate the application of CL on two *Production Objectives*: Constructing a relatively complex scaffold and satisfying a molecular docking constraint. In the former, given the same number of epochs, CL successfully constructs the complex structure from simpler constituents while baseline RL is unsuccessful. In the second application example, using Tanimoto (2D) or ROCS (3D) shape similarity to the reference ligand as *Curriculum Objectives* guides the agent to areas of chemical space that satisfies the docking constraint.[24,25] In contrast, baseline RL significantly struggles, spending many epochs generating unfavourable compounds. CL facilitates direct steering of agent policy towards a *Production Objective* by providing the ability to teach the agent specific knowledge. The results show that teaching the agent to optimize *Curriculum Objectives* to a greater degree can enhance the ability to satisfy a complex *Production Objective*, relative to baseline RL. However, optimizing similarity-based *Curriculum Objectives* to a greater degree leads to lower intra-set diversity, as the agent generates compounds that are 'closer' in chemical space. Thus, devising appropriate curricula allows one to accelerate agent convergence and steer agent policy update for bespoke applications.

Methods

REINVENT Curriculum Learning Extension. The implementation of CL builds on the REINVENT generative model, which uses a recurrent neural network (RNN) architecture.[5,29] The *de novo* molecular design task is formulated as a natural language processing (NLP) problem where compounds are sampled in the SMILES format based on conditional probabilities.[17,30] The RNN in this work features three hidden layers of 512 long short-term memory (LSTM) cells with an embedding size of 256 and a linear layer with softmax activation.[5,31] A prior generative model is first trained on the ChEMBL dataset to learn the

SMILES syntax before focusing the model towards a MPO task.[5,17,32] For further details on REINVENT, see the work by Blaschke et al.[5]

REINVENT's Learning Hyperparameters. The same hyperparameters were used for the baseline RL and CL experiments: batch size of 128, learning rate of 0.0001, sigma scalar factor of 128, and using the Adam optimizer.[33]

Agent Exploration and Exploitation. Balance between agent chemical space exploration and exploitation was achieved by using a diversity filter (DF) and inception. A DF enforces diverse results by defining *buckets* with limited size that track the number of compounds sampled possessing the same scaffold. Once a bucket is full, further sampling of compounds with the same scaffold will be penalized.[5,34] Inception is a form of experience replay to mitigate catastrophic forgetting and can speed up convergence by replaying previously sampled favourable compounds to the agent.[5,35] For further details on REINVENT, see the work by Blaschke et al.[5] In the baseline RL experiments, an Identical Murcko Scaffold DF (penalizes the agent if the same Bemis-Murcko scaffold is sampled beyond the *bucket* size) and inception were applied. In contrast, the implementation of CL in REINVENT allows one to initialize separate DFs and inception for the *Curriculum Phase* and *Production Phase*. During the *Curriculum Phase*, the goal is for the agent to acquire intermediate knowledge. Thus, no DF was applied as it can be counterproductive to guiding the agent to favourable areas of chemical space. In the *Production Phase*, a new inception (previous favourable compounds during the *Curriculum Phase* cleared) was initialized. Presumably, the agent is in a state of *productivity* and samples compounds that satisfy the *Production Objective*.[5,34] Thus, an Identical Murcko Scaffold DF was applied to encourage exploration, such that the agent samples from different local minima.[5,27]

ROCS 3D Shape Similarity. ROCS is a 3D shape similarity metric, comprised of two components: ‘shape’ and ‘color’. The components are quantified by the match, if at all, between the volumes occupied and the defined pharmacophoric features between the two ligands, respectively.[24,25] Compounds with similar “shape” and “color” are more likely to exhibit similar properties. The implementation of ROCS in REINVENT is described in detail by Papadopoulos et al.[36] In the CL experiments, the hyperparameters used for ROCS were 1:1 shape:color, giving equal weighting to each component in the final ROCS similarity score.

Molecular Docking Constraint Experiments. The PDK1 receptor crystal structure was obtained from the Protein Data Bank (PDB) with **PDB ID: 2XCH**.[22] A receptor grid was generated in the Maestro GUI with two hydrogen-bonding constraints specified between the reference ligand and Ala 162.[37] Ligand preparation and docking was performed using DockStream, which is integrated with REINVENT, facilitating parallelization over numerous CPU cores.[38] 3D coordinates for all agent sampled compounds from the baseline RL and CL experiments were generated using LigPrep. Default parameters were used except for the pH tolerance range set to 7.0 ± 1.0 with Epik and a maximum of two stereoisomers kept per compound.[39] Glide docking used Standard Precision (SP) with the followings settings: allow only amide trans isomers, allow up to 25 poses for post-docking minimization, apply strain correction, and apply enhanced sampling with a factor of 2.[40–43] All baseline RL and CL experiments were allowed 300 *production* epochs, i.e., epochs that involve docking, for a reasonable allocation of computational resources and for a fair comparison between baseline RL and CL. The docking score transformation was chosen to encourage agent sampling of compounds that possess a more favourable docking score than the reference ligand (see Supporting Information Fig. S2).

Cross-Tanimoto Similarity.

The cross-Tanimoto similarity is calculated as the Tanimoto similarity for each unique compound pair in a dataset. Note that the compound pairs 'AB' and 'BA' are the same, and hence only calculated once.

References

(1) Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. Artificial Intelligence in Drug Discovery: Recent Advances and Future Perspectives. *Expert Opin. Drug Discov.* **2021**, 1–11. https://doi.org/10.1080/17460441.2021.1909567.

(2) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A.; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; Zentgraf, M.; Hill, J. E.; Krutoholow, E.; Kohler, M.; Blaney, J.; Funatsu, K.; Luebkemann, C.; Schneider, G. Rethinking Drug Design in the Artificial Intelligence Era. *Nat. Rev. Drug Discov.* **2020**, *19* (5), 353–364. https://doi.org/10.1038/s41573-019-0050-3.

(3) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided Mol. Des.* **2013**, *27* (8), 675–679. https://doi.org/10.1007/s10822-013-9672-4.

(4) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. *Sci. Adv.* **2018**, *4* (7), eaap7885. https://doi.org/10.1126/sciadv.aap7885.

(5) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model.* **2020**, *60* (12), 5918–5922. https://doi.org/10.1021/acs.jcim.0c00915.

(6) Thomas, M.; Smith, R. T.; O'Boyle, N. M.; de Graaf, C.; Bender, A. Comparison of Structure- and Ligand-Based Scoring Functions for Deep Generative Models: A GPCR Case Study. *J. Cheminformatics* **2021**, *13* (1), 39. https://doi.org/10.1186/s13321-021-00516-0.

(7) Goel, M.; Raghunathan, S.; Laghuvarapu, S.; Priyakumar, U. D. MoleGuLAR: Molecule Generation Using Reinforcement Learning with Alternating Rewards. **2021**. https://doi.org/10.33774/chemrxiv-2021-cg9p8.

(8) Ståhl, N.; Falkman, G.; Karlsson, A.; Mathiason, G.; Boström, J. Deep Reinforcement Learning for Multiparameter Optimization in de Novo Drug Design. *J. Chem. Inf. Model.* **2019**, *59* (7), 3166–3176. https://doi.org/10.1021/acs.jcim.9b00325.

(9) Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *ArXiv170510843 Cs Stat* **2018**.

(10) Sanchez-Lengeling, B.; Outeiral, C.; Guimaraes, G. L. Optimizing Distributions over Molecular Space. An Objective-Reinforced Generative Adversarial Network for Inverse-Design Chemistry (ORGANIC). 18.

(11) Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P. Optimization of Molecules via Deep Reinforcement Learning. *Sci. Rep.* **2019**, *9* (1), 10752. https://doi.org/10.1038/s41598-019-47148-x.

(12) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276. https://doi.org/10.1021/acscentsci.7b00572.

(13) Ma, B.; Terayama, K.; Matsumoto, S.; Isaka, Y.; Sasakura, Y.; Iwata, H.; Araki, M.; Okuno, Y. Structure-Based de Novo Molecular Generator Combined with Artificial Intelligence and Docking Simulations. *J. Chem. Inf. Model.* **2021**. https://doi.org/10.1021/acs.jcim.1c00679.

(14) Bai, Q.; Tan, S.; Xu, T.; Liu, H.; Huang, J.; Yao, X. MolAICal: A Soft Tool for 3D Drug Design of Protein Targets by Artificial Intelligence and Classical Algorithm. *Brief. Bioinform.* **2021**, *22* (3). https://doi.org/10.1093/bib/bbaa161.

(15) Choi, J.; Lee, J. V-Dock: Fast Generation of Novel Drug-like Molecules Using Machine-Learning-Based Docking Score and Molecular Optimization. **2021**. https://doi.org/10.33774/chemrxiv-2021-75t5k.

(16) Nigam, A.; Pollice, R.; Aspuru-Guzik, A. JANUS: Parallel Tempered Genetic Algorithm Guided by Deep Neural Networks for Inverse Molecular Design. *ArXiv210604011 Cs* **2021**.

(17) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(18) Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*; ACM Press: Montreal, Quebec, Canada, 2009; pp 1–8. https://doi.org/10.1145/1553374.1553380.

(19) Weinshall, D.; Cohen, G.; Amir, D. Curriculum Learning by Transfer Learning: Theory and Experiments with Deep Networks. *ArXiv180203796 Cs* **2018**.

(20) Hacohen, G.; Weinshall, D. On The Power of Curriculum Learning in Training Deep Networks. *ArXiv190403626 Cs Stat* **2019**.

(21) Zhao, H. Scaffold Selection and Scaffold Hopping in Lead Generation: A Medicinal Chemistry Perspective. *Drug Discov. Today* **2007**, *12* (3), 149–155. https://doi.org/10.1016/j.drudis.2006.12.003.

(22) Angiolini, M.; Banfi, P.; Casale, E.; Casuscelli, F.; Fiorelli, C.; Saccardo, M. B.; Silvagni, M.; Zuccotto, F. Structure-Based Optimization of Potent PDK1 Inhibitors. *Bioorg. Med. Chem. Lett.* **2010**, *20* (14), 4095–4099. https://doi.org/10.1016/j.bmcl.2010.05.070.

(23) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4* (2), 90–98. https://doi.org/10.1038/nchem.1243.

(24) *ROCS 3.4.2.1: OpenEye Scientific Software, Santa Fe, NM. Http://Www.Eyesopen.Com.*

(25) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shapematching and Docking as Virtual Screening Tools.

(26) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging Chemical and Biological Space: "Target Fishing" Using 2D and 3D Molecular Descriptors. *J. Med. Chem.* **2006**, *49* (23), 6802–6810. https://doi.org/10.1021/jm060902w.

(27) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893. https://doi.org/10.1021/jm9602928.

(28) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* **2020**.

(29) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES Strings Improve the Quality of Molecular Generative Models. *J. Cheminformatics* **2019**, *11* (1), 71. https://doi.org/10.1186/s13321-019-0393-0.

(30) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminformatics* **2017**, *9* (1), 48. https://doi.org/10.1186/s13321-017-0235-x.

(31) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9* (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

(32) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (Database issue), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(33) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* **2017**.

(34) Blaschke, T.; Engkvist, O.; Bajorath, J.; Chen, H. Memory-Assisted Reinforcement Learning for Diverse Molecular de Novo Design. *J. Cheminformatics* **2020**, *12* (1), 68. https://doi.org/10.1186/s13321-020-00473-0.

(35) Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T. P.; Wayne, G. Experience Replay for Continual Learning. *ArXiv181111682 Cs Stat* **2019**.

(36) Papadopoulos, K.; Giblin, K. A.; Janet, J. P.; Patronov, A.; Engkvist, O. De Novo Design with Deep Generative Models Based on 3D Similarity Scoring. *Bioorg. Med. Chem.* **2021**, *44*, 116308. https://doi.org/10.1016/j.bmc.2021.116308.

(37) *Schrödinger Release 2021-2: Maestro, Schrödinger, LLC, New York, NY, 2021.*

(38) Guo, J.; Janet, J. P.; Bauer, M. R.; Nittinger, E.; Giblin, K. A.; Papadopoulos, K.; Voronov, A.; Patronov, A.; Engkvist, O.; Margreitter, C. DockStream: A Docking Wrapper to Enhance De Novo Molecular Design. **2021**. https://doi.org/10.33774/chemrxiv-2021-qvhml.

(39) *Schrödinger Release 2019-4: LigPrep, Schrödinger, LLC, New York, NY, 2019.*

(40) *Schrödinger Release 2019-4: Glide, Schrödinger, LLC, New York, NY, 2019.*

(41) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749. https://doi.org/10.1021/jm0306430.

(42) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47* (7), 1750–1759. https://doi.org/10.1021/jm030644s.

(43) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, *49* (21), 6177–6196. https://doi.org/10.1021/jm051256o.