Improving de novo molecular design with curriculum learning

Jeff Guo1,4, Vendy Fialková1,4, Juan Diego Arango1, Christian Margreitter1, Jon Paul Janet2, Kostas Papadopoulos1, Ola Engkvist1,3 and Atanas Patronov1✉

Reinforcement learning is a powerful paradigm that has gained popularity across multiple domains. However, applying reinforcement learning may come at the 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 non-productivity. Curriculum learning 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 curriculum learning for drug discovery. We implement curriculum learning in the de novo design platform REINVENT, and apply it to illustrative molecular design problems of different complexities. The results show both accelerated learning and a positive impact on the quality of the output when compared with standard policy-based reinforcement learning.

The application of deep learning for drug discovery provides potential to accelerate therapeutics development. One fundamental challenge is molecular design, involving the design and prioritization of candidate molecules for experimental validation. Molecular design entails a multiparameter optimization (MPO) search in chemical space, estimated to be in the range of $10^{13}$–$10^{20}$ molecules. Existing methods for molecular design include virtual screening (VS), which continues to be a valuable paradigm in identifying candidate molecules. Correspondingly, recent work by Sadybekov et al. introduced V-SYNTHES, a synthon-based screening approach capable of handling ultralarge compound libraries and yielding synthetically tractable molecules that were experimentally validated.

Recently, deep learning methods have emerged as an alternative to VS. An advantage with these methods is that a much larger chemical space can be sampled when compared with methods relying on enumerating molecules. However, care needs to be taken that the generated molecules are synthetically feasible. Deep generative models using policy-based reinforcement learning (RL), value-based RL, learning a molecular latent space, and other methods including tree search and genetic algorithms have been proposed to design molecules possessing desired 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 on the basis of a predefined reward function. Often, physics-based approximations of binding affinity such as molecular docking are included as a component in the reward function to design molecules with enhanced predicted potency. Given sufficiently long training time, these models can learn to generate molecules that satisfy the desired MPO objective. However, in cases with complex reward functions where minima are difficult to find, 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 intensive 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.

Curriculum learning (CL) has been proposed as a training strategy to overcome difficulties in learning complex tasks. 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 with 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. Molecular design often requires optimization of correlated properties that cumulatively define favourable chemical space—for example, generating known active scaffolds and improving binding affinity. By applying concepts from CL, existing limitations of policy-based RL for 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 tasks where policy-based RL has difficulties. 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 case study to design 3-phosphoinositide-dependent protein kinase-1 (PDK1) inhibitors. We show that immediate states of productivity can be achieved by assembling a curriculum compared with 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 molecular design, enabling control over the quality and diversity of the results in a predictable manner.

CL set-up

CL formulation. The implementation of CL builds on the REINVENT platform described by Blaschke et al. (Methods).
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 providing the production objective. Agent learning progresses through the curriculum phase to the production phase and is controlled by curriculum progression criteria, checking that the agent achieves an adequate score threshold for each objective (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 (see Methods for a formal definition of the CL strategy). Agent policy update is maintained in the production phase to ensure that the agent samples from diverse minima.

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. Scoring function: matching substructure to target scaffold.
   - Curriculum objective: achieve a state of productivity by decomposing the target scaffold into simpler substructures with gradually increasing structural complexity. Scoring function: matching sequential substructures.
2. Production objective: generate compounds that are drug-like and satisfy a molecular docking constraint. Scoring function: docking constraint and quantitative estimate of drug-likeness (QED).
   - Curriculum objective: guide the agent to sample compounds with Tanimoto (two-dimensional, 2D) similarity to a reference ligand. Scoring function: Tanimoto (2D) similarity to the reference ligand.
3. Curriculum objective: guide the agent to sample compounds with three-dimensional (3D) shape-based similarity (ROCS) to a reference ligand (Methods). Scoring function: ROCS (3D) similarity to the reference ligand.

For experiments 2 and 3, we further define a ‘low’ (0.5) and a ‘high’ (0.8 for Tanimoto and 0.75 for ROCS) 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.

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, and studies kept the phenyl substituent constant. The goal is to generate compounds that possess this target scaffold. We first demonstrate that the task is too complex for standard policy-based RL and denote this as baseline RL (Supplementary Fig. 1). In the baseline experiment, the only component in the scoring function is the target 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 2,000 epochs, indicating that 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 low 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 five curriculum objectives, 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—for example, at approximately epoch 150 (Fig. 2), as it is unlikely that currently
Satisfying a molecular docking constraint. 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, that is, predicted to retain experimentally validated interactions (see Methods for experiment hyperparameters). Simulating a real-world application where limited computational resources must be allocated, baseline RL and CL performances are compared, given a maximum number (300) of permitted production epochs, that is, epochs that involve docking, as these are relatively computationally demanding. For CL, curriculum objectives are first applied to guide the agent and the number of permitted curriculum epochs is not limited, as these are computationally inexpensive (Supplementary Table 2). 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. 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 with the reference ligand (as assessed by LigPrep and Glide docking score) and are drug-like, as measured by the QED score. The QED score also prevents compounds from exploiting the weaknesses of docking algorithms to achieve favourable scores—for example, presence of excessive hydrogen-bond donors would substantially restrict membrane permeability.

First, we show that the production objective is challenging for baseline RL (Fig. 3b,c). The docking score is approximately 0 for the first 100 epochs, indicating that essentially no compounds sampled satisfy the docking constraint. For epochs 100–200, some compounds satisfy the docking constraint but the score is still low. It is only from epoch 200 onwards that the docking score begins to ascend. This improvement indicates the point at which the agent starts entering a state of productivity. It is evident that baseline RL is suboptimal, as the agent spends a substantial 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, and achieves scores similar to those of CL (Supplementary Fig. 5).

To circumvent the limitations of baseline RL, we devise curricula and introduce two curriculum objectives to guide the agent to productivity: Tanimoto (2D) and ROCS (3D) shape-based similarity to the reference ligand. 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. 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 fair comparison with CL. These baseline experiments did not improve agent productivity, and training progress similar to that of the baseline shown in Fig. 3b,c is observed (Supplementary Figs. 6 and 7). 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 for baseline RL, but unlike the Tanimoto experiments the ROCS experiments start at a less favourable docking score. First, 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, of satisfying the docking constraint owing to more potential conformational...
discrepancies of the generated compounds compared with the reference ligand. Second, 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 attributable 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 poses of selected generated compounds are superimposed with the reference ligand (Fig. 3d). The binding poses retain the two hydrogen-bonding interactions with Ala162, 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 that CL outperforms baseline RL given the same number of production epochs. Moreover, tuning
Fig. 4 | Baseline RL versus CL docking score distribution. 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’ refer to baseline RL using a scoring function composed of docking, QED and Tanimoto/ROCS together, respectively. ‘CL Tanimoto Low’ and ‘CL Tanimoto High’ refer to using Tanimoto similarity as a curriculum objective that requires achieving corresponding thresholds of ‘low’ (0.5) or ‘high’ (0.8) score before progressing to the production objective. Analogously, ‘CL ROCS low’ and ‘CL ROCS high’ denote the results based on using ROCS as a curriculum objective where the low score is 0.5 and the high is 0.75. Lower Glide docking scores denote a greater predicted binding affinity. The docking score for the reference ligand is $-10.907 \text{ kcal mol}^{-1}$ and is shown by the horizontal black dashed line. Not only does CL collect more compounds than baseline RL, but the compounds also possess more favourable docking scores, on average.

**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 Supplementary Table 4 for individual experiments). 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,c, where the baseline RL experiments generate essentially no favourable compounds in the first 100 epochs. Of 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 as ‘favourable’ if the corresponding compound possesses 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 score distributions in Fig. 4, which 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, Methods). Importantly, the unique scaffold count demonstrates the capability of CL to perform scaffold hopping (see Supplementary Figs. 10–12 for example compounds generated in the high Tanimoto experiment and Supplementary Figs. 13–15 for unique scaffold statistics for each replicate experiment).

**Direct steering of agent policy allows 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 collected compounds from the CL Tanimoto experiments were pooled and the average Tanimoto similarity (compound to scaffold) to the reference ligand calculated for each epoch.
Fig. 5 | Baseline RL versus CL unique Bemis–Murcko scaffolds. 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 Supplementary Table 4 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.

(Fig. 6a). The left-hand subplot shows the gradual optimization of Tanimoto similarity for the low and high scenarios, representing the curriculum phase. The right-hand subplot shows the Tanimoto similarities for all the compounds collected in the production phase. In general, the compounds generated from the high Tanimoto experiments possess a greater Tanimoto similarity to the reference ligand than do those from the low Tanimoto experiments, as expected (see Supplementary Fig. 17 for the distribution of Tanimoto similarities). Moreover, the gradual decrease in Tanimoto similarity at the scaffold level further supports the capability of CL to perform scaffold hopping (Fig. 6a). Interestingly, however, the difference is not marked and can be explained by cross-referencing the training plots shown in Fig. 3b. The low Tanimoto experiments start at docking scores notably lower than those in the high Tanimoto scenario, which suggests that the compounds collected at the beginning are those that happen to exhibit high Tanimoto similarity to the reference ligand. Consequently, the low Tanimoto experiments generate less favourable compounds in the first 50 epochs when the production objective is activated (Supplementary Fig. 18).

To investigate the effect of similarity-based curriculum objectives on solution space diversity, cross-Tanimoto similarities between each unique compound pair in the pooled datasets were calculated to quantify how different the collected compounds are from each other (Methods). Relative to the baseline RL experiments, collected compounds in the CL experiments exhibit greater intraset 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 do the low scenarios. Uniform Manifold Approximation and Projection (UMAP) was used as a dimension reduction technique to visualize the solution space diversity from the CL Tanimoto experiments\(^5\). There is notable similarity, without overlap (as there is no scaffold overlap, Supplementary Fig. 14), between the compounds sampled from the low and high scenarios (Fig. 6c). The results suggest that moderate optimization of similarity-based curriculum objectives (as in the low scenarios) already narrows substantially the agent perceived solution space, in agreement with the cross-Tanimoto similarity distributions shown in Fig. 6b. The similarity between the compounds generated from the low and high experiments was quantified by calculating the cross-Tanimoto similarity between the two datasets (Supplementary Fig. 20). The majority of cross-Tanimoto similarities are \(>0.7\), confirming that the compounds generated from the two 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 CL to accelerate agent convergence on complex MPO objectives\(^5\). Relative to baseline RL, which may issue many non-productive calls to expensive descriptors, 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 scaffold 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 a curriculum objective guides the agent to areas of chemical space that satisfy the docking constraint\(^5\). In contrast, baseline RL visibly 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 intraset diversity, as the agent generates compounds that are closer in chemical space. Future work will apply CL towards other pertinent drug discovery tasks. For example, Nicolaou et al. apply
evolutionary algorithm to design compounds with good selectivity and present an interesting case study20.

Methods

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

Definition 2. A curriculum, \( C \), consists of a sequence of objectives, 
\( O = (O_{C1}, \ldots, O_{Cn}, O_P) \), where subscripts \( C \) and \( P \) denote curriculum and production objectives, respectively. For each objective \( O \), there is a corresponding \( S \) to compute the desirability of a sampled compound based on the current objective—for example, possessing a specific structural motif. Progression through a curriculum is controlled by curriculum progression criteria, \( P = (P_1, \ldots, P_{n-1}, P_n) \), such that \( C = (O, P) \).

Curriculum phase. In the curriculum phase, the goal is for the agent to learn to generate compounds that satisfy sequential curriculum objectives with increasing complexity that guide the agent towards the production phase. \( O_{C1}, \ldots, O_{Cn}, O_P \) are designated curriculum objectives with corresponding curriculum progression criteria \( P_1, \ldots, P_{n-1}, P_n \), which enforce sufficient agent learning of each sequential curriculum objective on the basis of a score threshold. If the score threshold is met, the agent progresses to the next curriculum objective; otherwise, the agent continues learning the current curriculum objective. This process collectively constitutes the curriculum phase.
Production phase. If and only if the final curriculum progression criterion is satisfied, the production objective $O_2$ 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 (Methods). The agent samples for a predefined number of epochs and all compounds that score above a minimum threshold are stored and outputted at the end.

REINVENT CL extension. The implementation of CL builds on the REINVENT generative model, which uses a recurrent neural network architecture. The molecular design task is formulated as a natural language processing problem, where compounds are sampled in the SMILES format on the basis of conditional probabilities. The recurrent neural network in this work features three hidden layers of 512 long short-term memory cells with an embedding size of 256 and a linear output layer with softmax activation. A prior generative model is first trained on the ChEMBL database to learn the SMILES syntax. The agent is initialized as the prior and is then focused towards an MPO task via RL. For further details on REINVENT, see the work by Blaschke et al.

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, all scoring function components’ weights set to 1 and using the Adam optimizer. Further details see on REINVENT; see the work by Blaschke et al. 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. Thus, an identical Murcko scaffold DF with a bucket size of 25 was applied to encourage exploration, such that the agent samples from different local minima. Finally, different curriculum progression criteria were applied in experiments 2 and 3 via the low and high standards to investigate the effect of curriculum objective optimization and agent exploration and exploitation. The results show that optimizing similarity-based curriculum objectives to a greater degree can decrease intrinsic diversity. Thus, curriculum progression criteria provide an additional control mechanism for result diversity.

ROCS 3D shape similarity. ROCS is a 3D shape similarity metric, comprised of two components: ‘shape’ and ‘colour’. The components are quantified by the match, if any, between the volumes occupied and the defined pharmacophoric features between the two ligands, respectively. Compounds with similar ‘shape’ and ‘colour’ are more likely to exhibit similar properties. The implementation of ROCS in REINVENT is described in detail by Papadopoulos et al. In the CL experiments, the hyperparameters used for ROCS were 1:1 shape:colour, 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 with Protein Data Bank identifier 2XCH (ref. 10). A receptor grid was generated in the Maestro graphical user interface with two hydrogen-bonding constraints specified between the reference ligand and Ala 162 (ref. 10). Ligand preparation and docking was performed using DockStream, which is integrated with REINVENT, facilitating parallelization over numerous CPU cores. 3D coordinates for all agent-sampled compounds 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. Glide docking used standard precision with the following 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 (refs. 10, 11). All baseline RL and CL experiments were allowed 300 production epochs, that is, epochs that involve scoring function evaluations. Amazon Web Services (AWS) p3.8xlarge instance and parallelized over 8 and 48 CPU cores using DockStream, respectively.

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.

Data availability
The trained generative model to reproduce the experiments in this work is provided at https://github.com/MolecularAI/ReinventCommunity/blob/master/notebooks/models/random.prior.new. The raw data that support the findings of this study are available from the corresponding author upon request.

Code availability
The code used in this study is available at https://github.com/MolecularAI/Reinvent. A corresponding tutorial for the code is available at https://github.com/MolecularAI/ReinventCommunity/blob/master/notebooks/Automated_Curriculum_Learning_Demo.ipynb. The specific frozen version of the code is available at https://zenodo.org/badge/486692949.svg. The DOI badge is provided at https://zenodo.org/badge/48669249.svg.
19. Nigam, A., Pollice, R. & Aspuru-Guzik, A. JANUS: parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. Preprint at https://arxiv.org/abs/2106.04011 (2021).

20. Nicolaou, C. A., Apostolakis, J. & Pattichis, C. S. De novo drug design using multiobjective evolutionary graphs. J. Chem. Inf. Model. 49, 295–307 (2009).

21. Bengio, Y., Louradour, J., Collobert, R. & Weston, J. Curriculum learning. In ICML’09: Proc. 26th Annual International Conference on Machine Learning 41–48 (ACM, 2009); https://doi.org/10.1145/1553374.1553380

22. Weinshall, D., Cohen, G. & Amir, D. Curriculum learning by transfer learning: theory and experiments with deep networks. Preprint at https://arxiv.org/abs/1802.03796 (2018).

23. Hacohen, G. & Weinshall, D. On the power of curriculum learning in training deep networks. Proc. 36th International Conference on Machine Learning 2535–2544 (PMLR, 2019).

24. Zhao, H. Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. Drug Discov. Today 12, 149–155 (2007).

25. Angiolini, M. et al. Structure-based optimization of potent PDK1 inhibitors. Bioorg. Med. Chem. Lett. 20, 4095–4099 (2010).

26. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. Nat. Chem. 4, 90–98 (2012).

27. ROCS 3.4.2.1 (OpenEye Scientific Software, 2021).

28. Hawkins, P. C. D., Skillman, A. G. & Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. J. Med. Chem. 50, 74–82 (2007).

29. Schrödinger Release 2019-4: LigPrep (Schrödinger, 2019).

30. Schrödinger Release 2019-4: Glide (Schrödinger, 2019).

31. Friesner, R. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. 47, 1739–1749 (2004).

32. Halgren, T. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J. Med. Chem. 47, 1750–1759 (2004).

33. Friesner, R. A. et al. Extra Precision Glide: docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. J. Med. Chem. 49, 6177–6196 (2006).

34. Alex, A., Millan, D. S., Perez, M., Wakenhut, F. & Whitlock, G. A. Intramolecular hydrogen bonding to improve membrane permeability and absorption in beyond rule of five chemical space. MedChemComm 2, 669–674 (2011).

35. Nettles, J. H. et al. Bridging chemical and biological space: ‘target fishing’ using 2D and 3D molecular descriptors. J. Med. Chem. 49, 6802–6810 (2006).

36. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. J. Med. Chem. 39, 2887–2893 (1996).

37. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).

38. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31–36 (1988).

39. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. J. Cheminformaticas 9, 48 (2017).

40. Hochreiter, S. & Schmidhuber, J. Long short-term memory. Neural Comput. 9, 1735–1780 (1997).

41. Gaulton, A. et al. The ChEMBL database in 2017. Nucleic Acids Res. 45, D945–D954 (2017).

42. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2017).

43. Blaschke, T., Engkvist, O., Bajorath, J. & Chen, H. Memory-assisted reinforcement learning for diverse molecular de novo design. J. Cheminformaticas 12, 68 (2020).

44. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T. P. & Wayne, G. Experience replay for continual learning. Preprint at https://arxiv.org/abs/1811.11682 (2019).

45. 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. 44, 116308 (2021).

46. Schrödinger Release 2021-2: Maestro (Schrödinger, 2021).

47. Guo, J. et al. DockStream: a docking wrapper to enhance de novo molecular design. J. Cheminformaticas 13, 89 (2021).

48. Patronov, A., Margreitter, C., Guo, J. & Blaschke T. patronov/Reinvent: REINVENT 3.2 (v3.2). Zenodo https://doi.org/10.5281/zenodo.6502363 (2022).

Acknowledgements
We thank K. Giblin, A. Tomberg and E. Nittinger for constructive user feedback that helped us develop the concepts presented in work.

Author contributions
V.E., J.G., J.D.A. and A.P. developed the code. J.G., A.P., C.M. and K.P. designed the experiments. J.G. performed the experiments and analyses. J.G. wrote the manuscript and all other authors revised it. A.P. supervised the work. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00494-4.

Correspondence and requests for materials should be addressed to Atanas Patronov.

Peer review information Nature Machine Intelligence thanks Christos Nicolaou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2022, corrected publication 2022.