

On The Difficulty of Validating Molecular Generative Models Realistically: A Case Study on Public and Proprietary Data

Koichi Handa^{1,2}, Morgan C. Thomas¹, Michiharu Kageyama², Takeshi Iijima², and Andreas Bender^{1*}*

¹Centre for Molecular Informatics, Department of Chemistry, University of Cambridge,
Lensfield Road, Cambridge, CB2 1EW, UK

²Toxicology & DMPK Research Department, Teijin Institute for Bio-medical Research,
Teijin Pharma Limited, 4-3-2 Asahigaoka, Hino-shi, Tokyo 191-8512, Japan

Abstract

While a multitude of deep generative models have recently emerged there exists no best practice for their *practically relevant* validation. On the one hand, novel *de novo*-generated molecules cannot be refuted by retrospective validation (so that this type of validation is biased); but on the other hand prospective validation is expensive and then often biased by the human selection process. In this case study, we frame retrospective validation as the ability to mimic human drug design, by answering the following question: *Can a generative model trained on early-stage project compounds generate middle/late-stage compounds de novo?* To this end, we used experimental data that contains the elapsed time of a synthetic expansion following hit identification from five public (where the time series was pre-processed to better reflect realistic synthetic expansions) and six in-house project datasets, and used REINVENT as a widely adopted RNN-based generative model. After splitting the dataset and training REINVENT on early-stage compounds, we found that rediscovery of middle/late-stage compounds was much higher in public projects (at 1.60%, 0.64%, and 0.21% of the top 100, 500, and 5,000 scored generated compounds) than in in-house projects (where the values were 0.00%, 0.03%, and 0.04%, respectively). Similarly, average single nearest neighbour similarity between early- and middle/late-stage compounds in public projects was higher between active compounds than inactive compounds; however, for in-house projects the converse was true, which makes rediscovery (if so desired) more difficult. We hence show that the generative model recovers very few middle/late-stage compounds from real-world drug discovery projects, highlighting the fundamental difference between purely algorithmic design and drug discovery

as a real-world process. Evaluating *de novo* compound design approaches appears, based on the current study, difficult or even impossible to do retrospectively.

Introduction

De novo generative drug design is a current technique of interest^{1,2}, not least due to cost pressures³ and current endeavours to integrate computational and experimental work into Design-Make-Test-Analyze (DMTA) cycles⁴. Looking back, *de novo* design algorithms have been developed since at least the 1980s⁵. For some time, the mainstream method was the combination of fragment-like building blocks with genetic algorithms^{6,7}. Nowadays, due to the rapid growth of computer hardware including GPU computing, machine learning and deep neural networks applied to molecular generative models have become tractable^{8,9,10}.

As with any method, validation – and how to perform validation in a practically relevant manner – has been discussed actively¹¹. In the early stages of deep generative models, many researchers only concentrated on how the model produced novel compounds efficiently by copying the distribution of the training dataset, referred to as distribution-learning. Therefore, the principle performance metrics developed were validity, uniqueness, novelty, and diversity which are included in benchmarks such as MOSES and Fréchet ChemNet Distance^{12,13}. However, in a practical drug discovery process, goal-directed optimization is much more important. The gold standard of measuring model performance would be to synthesize and test *de novo* molecules experimentally (and compare to a baseline control¹⁴); however, this is intractable for all models considering the experimental resource requirement, given the number of models available and the number of *de novo* molecules proposed². Recently the CACHE initiative started, whose aim is to validate computationally suggested or generated compounds by experimental testing; however, this activity is limited in scope due to the cost of synthesizing

novel structures¹⁵. In order to fill the need for goal-directed benchmarking, Guacamol¹⁶ has been developed, which contains benchmarks such as rediscovery of and similarity to known active compounds. Although this benchmark is very practical for generative model utilized in lead optimization stage, the dataset is retrieved from ChEMBL¹⁷ and just removes the target compound from the training dataset, where then the task is to rediscover those removed compounds computationally. However, analogues may still remain in the training dataset (of which there are often many, given ChEMBL is constructed from publications which often contain SAR of related compounds), and suggested novel molecules may well be active although *not* being contained in the dataset, and hence also this type of validation has its shortcomings.

A real-world interpretation of generative models in the drug discovery context remains difficult, and the current work attempts to better understand this by retrospectively applying performance measures to generative models applied to public and private drug discovery data sources. The objective of the task is hence to achieve late-stage project compounds, given information from early-stage compounds, in a limited number of steps, and hence in a sample-efficient way (for a more detailed recent evaluation of the sample efficiency of different methods see a recent study¹⁸). This early/late data split strategy is in analogy to ‘time-split’ validation in the QSAR area, where splitting data into training and test sets along the time domain has been proposed before¹⁹.

However, drug discovery is not ligand discovery, and drug discovery does not only consist of optimizing a single objective in a proxy assay system²⁰. More specifically, during the lead

optimization stage of a drug discovery project, multiple-parameter optimization (MPO), for parameters such as primary target activity, activity against off-targets, and also physicochemical and ADME properties such as permeability, intrinsic clearance, solubility etc. need to be optimized simultaneously²¹, an area which has found consideration only in few computational studies^{22,23}. In reality the MPO process is very complicated in a drug discovery project, because the target profile could be easily changed (and even multiple times) during the course of a project, where new problems appear every step along project progress (Figure 1)²⁴. In this work, we attempted to see whether generative models can be validated retrospectively, on the one hand with public data mapped onto a pseudo-time axis, and on the other hand with real-world project data from different projects in a pharmaceutical company.

Regarding the architecture of the deep generative model, we decided to use one of the widely used approaches in the field, namely REINVENT^{25,26}. Recently, many architectures of generative models for *de novo* design have been published such as recurrent neural network (RNNs)²⁷, convolutional neural network (CNNs)^{28,29} and graph convolutional neural network (GCNN)³⁰. In the drug discovery field, although there are many models including variational auto encoder (VAE)¹, and Generative Adversarial Networks (GAN)³¹, due to the success of NLP, which was driven by many techniques like long-short time memory (LSTM)³², gated recurrent unit (GRU)³³ and attention mechanism³⁴, language models have found great resonance^{22,23,35,36}. Of those language models we here chose REINVENT, an RNN-type language model with the ability to perform goal-directed optimization through fine-tuning and reinforcement learning, due to its availability and wide use^{37,38,39,40,41,42,43}.

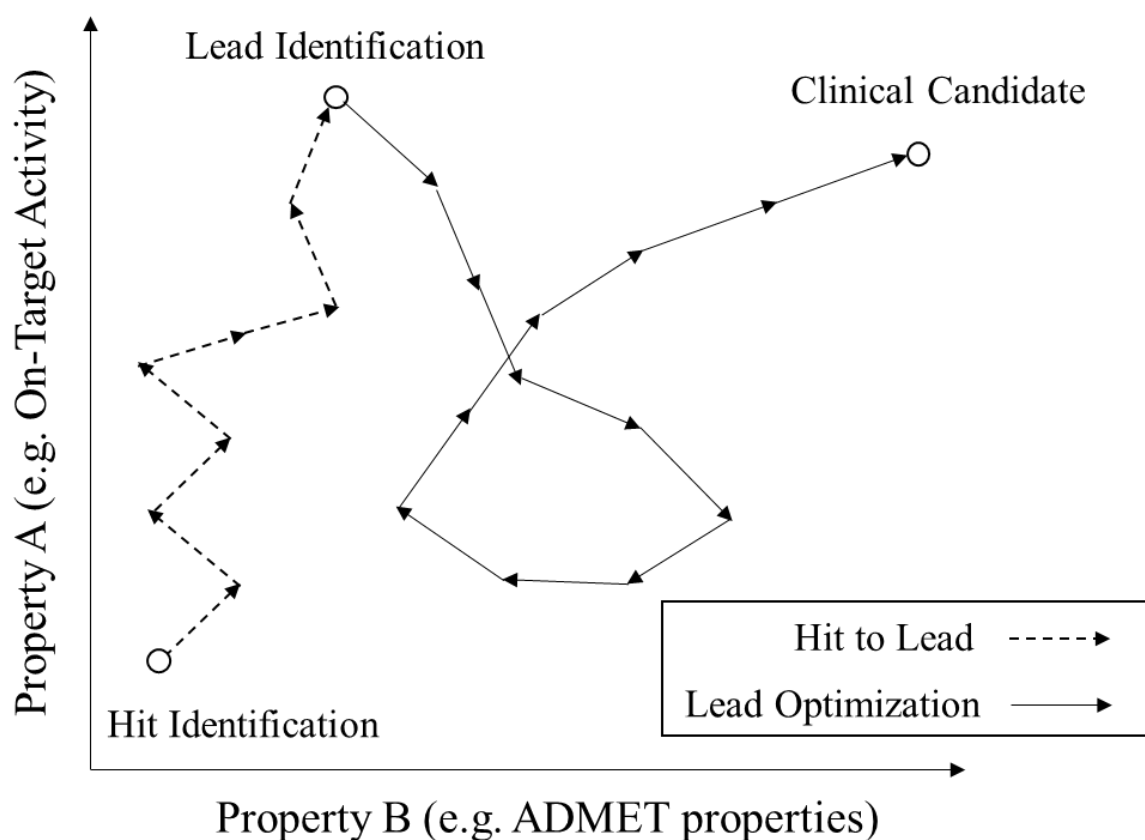


Figure 1 An example of a trajectory of compounds from hit identification to clinical candidate.

It can be seen that multiple properties matter in optimization (where in particular the X-axis subsumes a large number of additional properties), and that optimization is usually not linear in practice.

Materials and Methods

Dataset

For the public dataset, data for five targets were selected from Excape-DB⁴⁴, namely for DRD2 (Dopamine Receptor D2, 4,341 active compounds), GSK3 (Glycogen synthase kinase 3, 4,646 active compounds), CDK2 (Cyclin-dependent kinase 2, 2,065 active compounds), EGFR (Epidermal Growth Factor Receptor, 4,777 active compounds), and ADRB2 (Adrenergic receptor β 2, 2,616 active compounds). The rationale was to select datasets which have been well studied in previous publications and which include more than 1,000 compounds individually with pXC50 values. For the in-house dataset, six projects were collected from TEIJIN Pharma's in-house database which also include more than 1,000 compounds individually. These are named here as A, B, C, D, E, and F. Figure 2 and Table S1 show the number of compounds for each dataset, separated by activity values and 'early', 'middle' and 'late' stage annotations, further details of which are explained in the following.

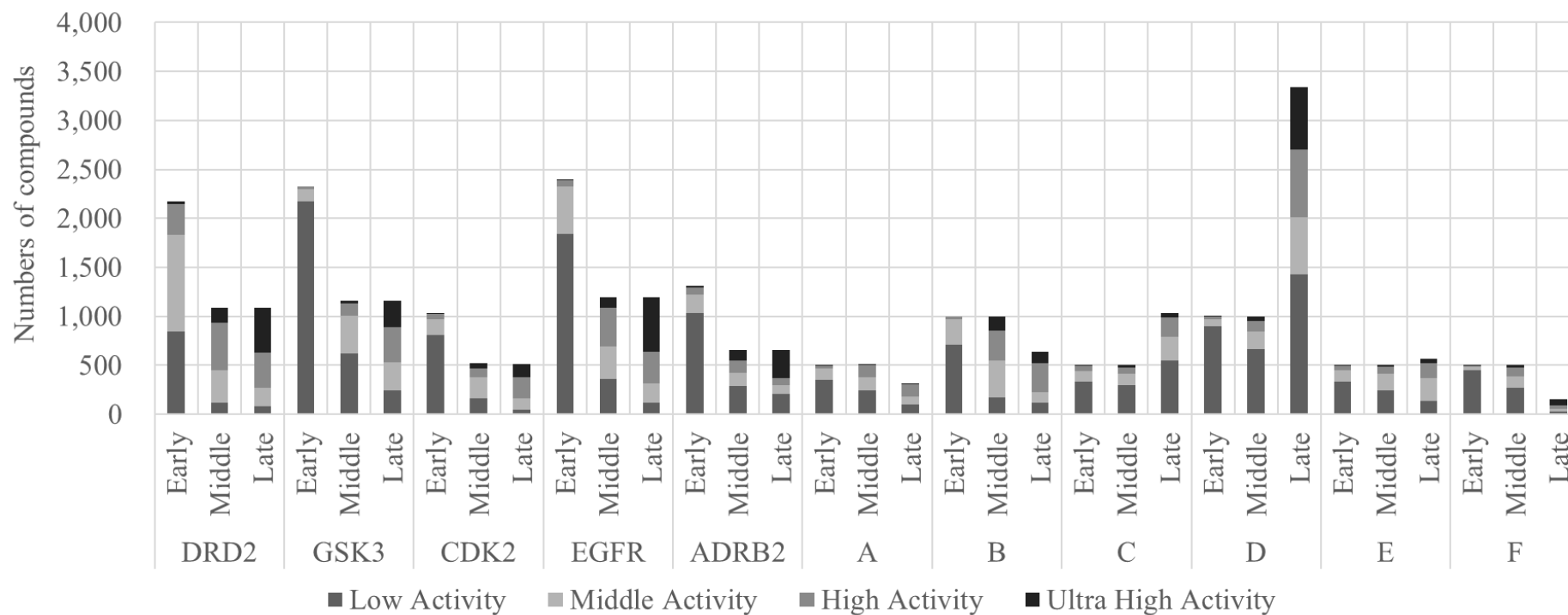


Figure 2 The datasets used in this study include wide range of activity.

The thresholds for activity classes generally are pXC50 values of less than 6 for low, over 6 to less than 7 for middle, over 7 to less than 8 for high, and over 8 for ultra-high compound activity.

Time Series Pre-Processing

Public Dataset

The public dataset utilized in this study was derived from ExCAPE-DB. All targets are well-studied and include more than 1,000 bioactivity data points with pXC50 values. The simplified molecular-input line-entry system (SMILES) strings were obtained from ExCAPE-DB for all molecules, and canonicalized using the RDKit (version 2020.09.01) component “RDKit Canon SMILES” and “Speedy SMILES De-salt” in the KNIME (version 4.3.4). However, this public database contains no ‘project registration date’ and the compounds deposited in the underlying databases (ChEMBL¹⁷ and PubChem⁴⁵) are usually done by publication or grouped upload, not reflecting realistic project time series optimization. Therefore, in order to mimic the time series of a practical drug discovery process that increases the activity with time elapsed, data was mapped onto a ‘pseudo-time axis’ as follows. We transformed the data by principle component analysis (PCA) using Datawarrior (ver 5.2.1)⁴⁶, and then the following three steps. (1) The canonical SMILES of the public dataset were input to calculate the FragFp⁴⁷ fingerprints. The FragFP fingerprints were used to calculate the normalized PCA scores of 3 components. (2) Then, these scores and pXC50 value of each compound was used to obtain another PCA score of 3 components. These 3 final PCA scores hence include information on both similarity of compounds in fingerprint space as well as bioactivity. (3) Finally, the Euclidean distance of all compounds in each dataset to the compound that has lowest pXC50 value was calculated using the final 3 PCA scores. This process introduces an ordering of compounds in bioactivity space (from low to high potency), as well as chemical space (from a low potency starting point, to high potency compounds with increasing dissimilarity to the starting point). We are aware that this process does not necessarily resemble a real-world drug discovery project, but it at least represents compound progression towards higher-potency compounds, which, given the

limited availability of public domain timestamped project data is the only practically feasible option we were able to identify for a public dataset. Then, the datasets were divided by both the activities and pseudo-stages.

To categorize the activities, pXC50 thresholds for activity classes in most projects are less than 6 for “low”, over 6 to less than 7 for “middle”, over 7 to less than 8 for “high”, over 8 for “ultra-high”. To categorize the stages for public projects already transformed into a pseudo time-series, those from the beginning of the compounds to 50%, 25% (accumulated from 50% to 75%), and 25% (accumulated from 75% to 100%) were classified into early, middle and late stage, respectively. As an example, Figure 3 shows the DRD2 dataset, with compounds classified across the different stages of the drug discovery ‘project’ mapped onto the pseudo-time axis, as well as the different bioactivity ranges used in this work.

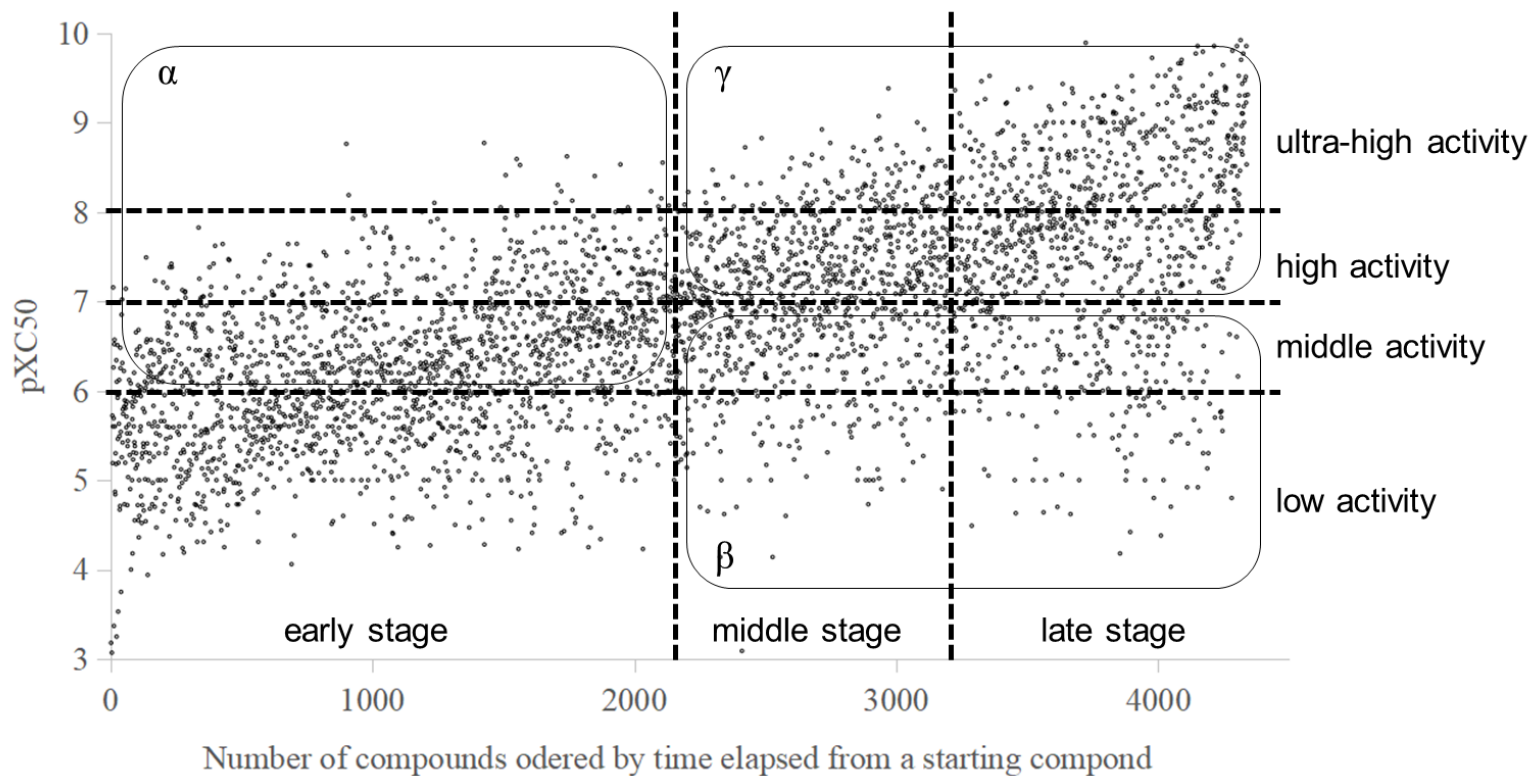


Figure 3 An example of data division according to stages and bioactivities.

The region of α that consists of more than middle activity compounds in the stage of early corresponds to the training dataset for fine-tuning to produce focused agent. The region of β consists of low and middle activity compounds in the middle and late stage, and the region of γ consists of more than high activity compounds in the middle and late stage. The X-axis is unitless.

In-house Dataset

The in-house dataset was retrieved from TEIJIN Pharma Ltd's Database, and we selected 6 projects (A to F) which have more than 1,000 bioactivity datapoints of pXC50 values. The date of completed synthesis for in-house compounds was recorded in the TEIJIN database for a given project. Consequently, being different from public dataset, we directly used the date for the time-series of the in-house dataset. An additional difference from the public dataset was that we know there was at least one additional property to be improved for each project that was not on-target activity, such as metabolism, physicochemical properties, etc., and this can even change at different timelines of the project²¹. Hence a second objective of using this dataset, apart from benchmarking generative models, was to evaluate to what extent the structures generated by *de novo* generative models actually follow the optimization trajectory across a set of real-world drug discovery projects. Regarding data classification along the bioactivity axis, for projects A, B, C, D, the setting of activity classification was the same as public dataset. However, for projects of E and F, given the bioactivity distribution in those cases, thresholds for activity classes have been set to less than 7 for "low", over 7 to less than 8 for "middle", over 8 to less than 9 for "high", over 9 for "ultra-high". As for the classification of stages of in-house projects, 500 or 1,000 was selected based on the progression of bioactivity in order to evenly split activity groups.

Regional Classification and Similarity Analysis

In order to establish 'project progress' with respect to potency, we next defined intervals α , β and γ (as shown in Figure 3 for the DRD2 dataset) in bioactivity space as follows. The region α is compounds with over middle activity in the early stage, while the region β is compounds

with low and middle activity in the middle and late stage and region γ is compounds with high and ultra-high activity in the middle and late stage. We then analyzed progression in the bioactivity domain across different project stages (*i.e.*, to which extent potency could be optimized by the model) by calculating the average similarity of generated molecules to the single nearest neighbour (aSNN) present in a given part of the dataset, based on the Tanimoto similarity of Morgan fingerprints using RDKit^{12,48} and with the aSNN calculations performed as implemented in MolScore³⁹.

Model Training

We used REINVENT²⁶ as a *de novo* generative design strategy, given its wide use in the field⁴⁹. Figure 4 shows the workflow of this study using the REINVENT framework, which is described in more detail as follows:

(i) Pretraining of Prior Model

Compounds were prepared in accordance with the REINVENT pipeline²⁶ as standardized non-isomeric SMILES. The Prior network was pre-trained on a dataset of 1,442,368 compounds derived from ChEMBL where the molecules were restrained to containing between 10 and 50 heavy atoms and elements {H, B, C, N, O, F, Si, P, S, Cl, Br, I}¹⁷. Only for public dataset, the compounds included in the ChEMBL dataset were omitted. In the pre-training, the Prior network was trained for a total of 10 epochs with a batch size of 128 with an adaptive learning rate starting from 0.0005. All other settings were set to default²⁶. All neural network training was conducted on an NVIDIA GeForce GTX 1650 Ti.

(ii) Data Preparation and Transformation

We next utilized the dataset of each project (either public or in-house) for fine-tuning, whose compounds were also processed as described in (i). Since the purpose of fine-tuning is to focus on the higher activity compounds, the compounds chosen for this step were early-stage compounds with above-average activity (region α).

(iii) Model Generation and Training

For focused agent network

The Pre-trained prior network obtained in (i) was fine-tuned by the compounds prepared in (ii), and the model obtained here was called the Focused Agent network. During fine-tuning, the pre-trained prior network was trained for a total of 10 epochs with a batch size of 128 with an adaptive learning rate starting from 0.0005. The other settings were adopted as default²⁶.

Random Forest model for the reinforcement learning scoring function

All compounds in the early stage were used to build a classification model which was used as the scoring function for reinforcement learning (RL) to optimize. Compounds possessing above average activity were classified as active and those below average activity classified as inactive. The dataset was divided into 70% training and 30% test, and ECFP6 descriptors⁵⁰ (1,024 bit, radius: 3) were generated using RDKit (version 2020.09.01) Chem functions⁴⁸ while a Random Forest (RF) (Python (ver. 3.7.10), scikit-learn (ver 0.24.2) library RandomForest ensemble.RandomForestClassifier function was used for machine learning⁵¹(Figure S1). The parameters of RF were set as follows; max_depth: 20, n_estimator: 100, others: default setting. RL was performed for 500 steps with a sigma value of 128 and learning rate of 0.0001.

(iv) Compound Generation

The compounds were generated from the Focused Agent network from (iii) and scored by the *in silico* classification model from (iii) repeatedly as the RL framework. 5,000 *de novo* molecule were sampled in total from the final network. The highest-ranked 100 and 500 compounds were selected according to the *in silico* classification score for subsequent analysis.

(v) Evaluations

As basic metrics, validity, uniqueness and novelty were calculated for all runs which have also been used in previous work^{12,16}. Validity is the fraction of correctness that a SMILES string translates to a real structure. Low validity is indicative of a poorly behaving model that has struggled to learn the SMILES grammar. Uniqueness is the fraction of unique molecules, where non-unique molecules are defined as having canonical SMILES that match those previously sampled or in the same batch. Low uniqueness is indicative of a poorly behaving model that is ‘stuck’ in a particular region of chemical space. Novelty is the ratio of valid, unique canonical SMILES not present in the training dataset (pre-training: ChEMBL, fine-tuning: above average activity compound in the early stage of each project which locates in region α), and low novelty indicates the model cannot generalize beyond the training data, which is precisely the aim of *de novo* design. All these calculations were implemented in Python 3.7.10 using the original code following the equations below, Eq. 2 to 4, where N_{gen} represents the number of generated compounds, N_{val} represents the number of valid compounds, N_{uni} represents the number of unique compounds in generated compounds, and N_{unk} represents the number of unknown compounds in generated compounds⁵².

$$Validity (\%) = \frac{N_{val}}{N_{gen}} \times 100 \quad (\text{Eq. 1})$$

$$Uniqueness (\%) = \frac{N_{uni}}{N_{val}} \times 100 \quad (\text{Eq. 2})$$

$$\text{Novelty (\%)} = \frac{N_{unk}}{N_{uni}} \times 100 \quad (\text{Eq. 3})$$

Finally, the highest-scored compounds selected from (iv) were evaluated by the following metrics. The calculation of these metrics was implemented in Python 3.7.10.

1. Rediscovery ratio, defined as Eq.4, in order to assess whether experimentally confirmed highly- or ultra-highly active compounds were generated, where N_{redis} represents the number of generated compound which agree with the real high or ultra-high activity compound in the middle or late stage.

$$\text{Rediscovery (\%)} = \frac{N_{redis}}{N_{gen}} \times 100 \quad (\text{Eq. 4})$$

2. aSNN in the middle stage, to evaluate whether generated compounds were similar to compounds from the middle stage of a given project with high or ultra-high activity.
3. aSNN in the late stage, to evaluate whether generated compounds were similar to compounds from the late stage of a given project with high or ultra-high activity (which means that the generative model behaves similarly to ‘real world’ projects, to the extent captured by the data used in this work and at that stage).

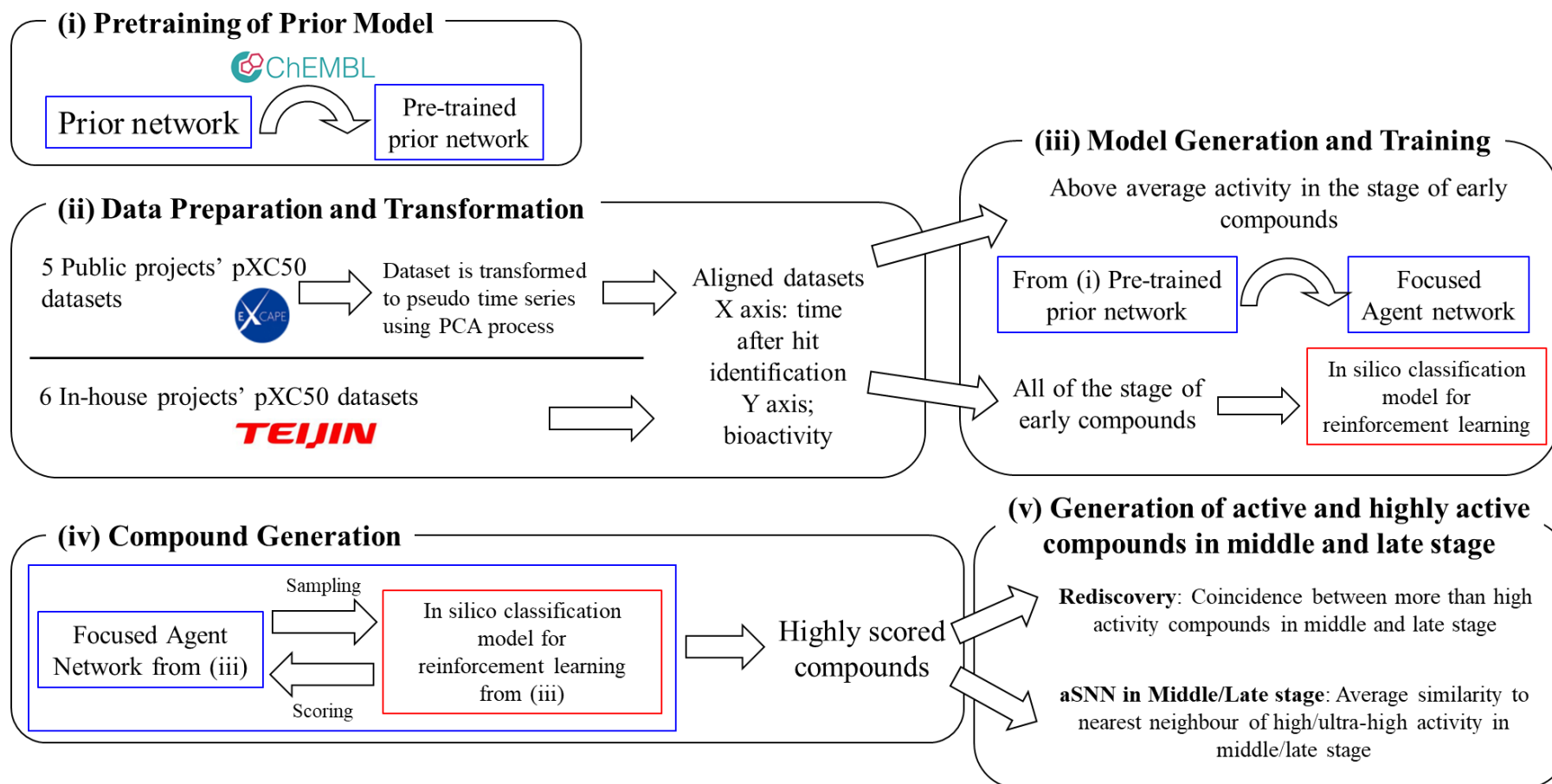


Figure 4 Workflow of this study (for details see main text).

As options, Inception and diversity filter (DF) could be used in the sampling process of (iv).

Generative Models and Options

Control experiment: Prior Network Only

As a baseline to compare to RL we generated compounds from the pre-trained prior network only, called “Control” in the following. The aim of this baseline was to investigate the effect of fine-tuning and reinforcement learning using the dataset prepared in this study.

Reinforcement Learning (RL)

This represents the ‘vanilla’ approach of this work, employing only Reinforcement Learning.

Diversity Filters (DF)

The next variation was the use of a “diversity filter (DF)”²⁶, which has been shown before to give an increase in the structural diversity of compounds generated^{42,53}. The parameters of DF were set to default as follows; name: IdenticalMurckoScaffold, nbmax: 25, minscore: 0.4, minsimilarity: 0.4. This run is called “RL-DF”.

Inception

The purpose of “Inception”²⁶ is to keep track of previously well scored compounds and to randomly expose a subset of them to the agent, thus helping to direct the learning. The parameters of Inception were set as follows; memory_size: 20, sample_size: 5. In this study, 30 compounds that were at least of ‘high’ activity in early stage were used.

Consequently, there are five different ways the generative model was run, which were Pre-trained prior network (Control), RL, RL-DF, RL-Inception, RL-DF-Inception.

Compound Clustering

To investigate the profiles of the activity of real (public and in-house) compounds according to the time elapsed quantitatively, we used compound k-means clustering as implemented in `sklearn.cluster` using ECFP6 fingerprints calculated using RDKit⁴⁸ and cluster size: 10. Then, we counted the number of compounds in each cluster and in each region (α , β and γ in Figure 3).

Furthermore, to understand the chemistry of generated compounds, we examined it by visual inspection, using DRD2 compounds as an example. From each cluster the centroid structure of each cluster was selected as a representative, and the structure which has the highest pXC50 value was selected as the highest-scoring structure of its cluster.

Results and Discussion

Dataset Characterization across The Bioactivity and Time Domain

We firstly aimed to understand the distribution of our datasets across the time and bioactivity domain, the results of which are shown in Figure 5. For the public projects, the aSNN between α and γ are much higher (by around 0.1) than those between α and β . However, for the in-house projects, the aSNN between α and γ were mostly similar to, or lower than, values between α and β except for project C. The underlying reason is likely that chemical series from publications including high-activity ligands were quite different from those with lower activities (hence giving area β a different composition), which is the result of different ligands (which different activity) being reported in different publications, w.r.t. both chemistry and publication date, given that those were the criteria used for dataset assembly here. On the other hand, for the in-house dataset this wasn't really the case, meaning that in relatively more cases late-stage high-activity space was still in a chemical area similar to that occupied at project start (although the situation is quite different for different projects). It can clearly be seen that both classes of datasets hence behave differently, which is entirely expected from the way they were constructed (see methods section for details).

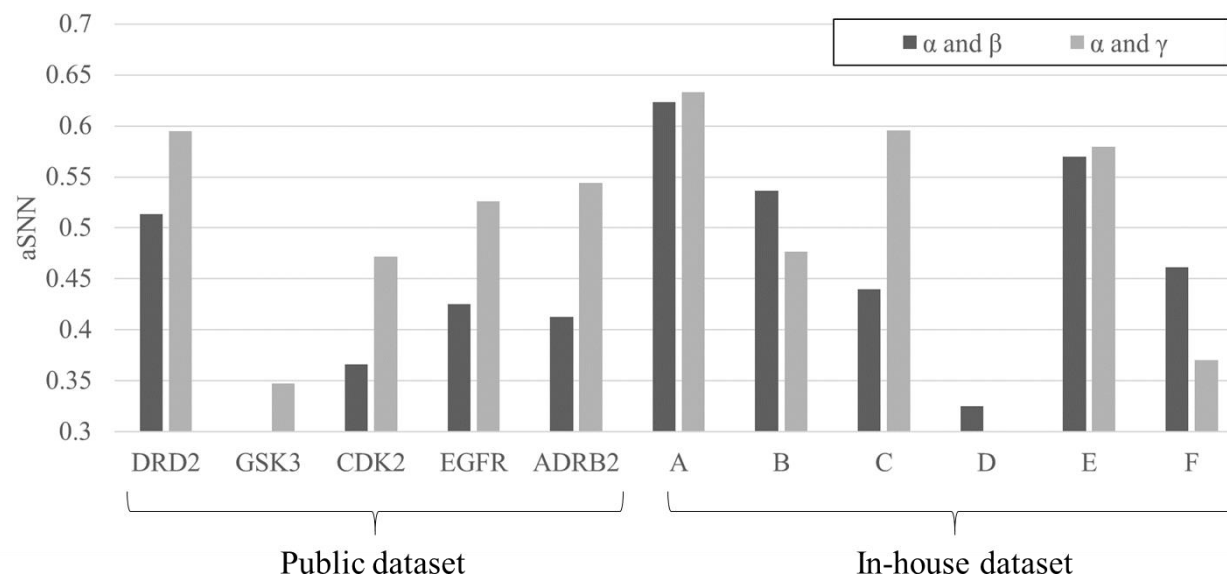


Figure 5 Average of single nearest neighbour similarity (aSNN) between training and test compounds for all projects for low or high activity real compounds were largely different from public and in-house projects.

It can be seen that the profiles in Public dataset (aSNN of α - β < α - γ) was different from in-house (mostly, aSNN of α - β > α - γ). The cut-off values of aSNN considered similar was set to be 0.3.

Metrics of Generated Compounds (Validity, Uniqueness, and Novelty)

Next, we calculated the validity, uniqueness, and novelty of the generated compounds. The results for RL are shown in Table S2. The validity for each target was over 98%. The uniqueness of generated compounds for the public dataset was relatively high, from 39.4% (GSK3) to 82.2% (DRD2), while the corresponding value for the in-house datasets was much lower, ranging from 15.1% (project F) to 50.9% (project E). The novelty of each target was over 70% (for detailed results see Table S3-S5). Through all the runs, regardless of targets, the validity and novelty were high enough, over 95% and 70%, respectively, which are appropriate in practice. The lower uniqueness values in in-house datasets ranging from 15.1% (project F) to 50.9% (project E) might reflect more congeneric compounds used in focused learning compared to the combination of different publications in public datasets. Across projects, the uniqueness of the RL-inception runs were lower than the other runs, from 36.4% (GSK3) to 59.8% (DRD2) for the public dataset, and from 19.4% (project F) to 40.8% (project E) for the in-house dataset which is lower than for the original RL runs. However, if the DF was used as an option, the low uniqueness was completely recovered, both for the RL-DF run, as well as the combination with Inception, with values ranging from 99.0% (project B) to 99.8% (CDK2) for the RL runs, while values for RL-DF-inception ranged from 96.8% (project B) to 98.5% (DRD2 and ADRB2). This underlines the importance of using diversity filters to ensure uniqueness of generated structures across the different situations considered here⁴⁰.

Rediscovery

We next analyzed the rediscovery rates of generated compounds using RL alone, the results of which are shown in Figure 6. For public projects, other than GSK3, we could find compounds

identical to real high activity compounds. The percentage of rediscovery for DRD2, CDK2, EGFR, and ADRB2 were 0.30%, 0.13%, 0.52%, and 0.09% for all 5,000 generated compounds, respectively; when using the *in silico* classification score 1.0%, 0.8%, 1.0%, and 0.4% of the 500 highest-scored generated compounds represent to known actives, while this was the case for 2%, 3%, 2%, and 1% for the top 100 scored generated compounds, respectively. For in-house projects, only the generative models for project A and B could find identical compounds to the real high activity compounds. The percentage of rediscovery in project A and B were 0.10%, and 0.15% for all 5,000 generated compounds; when using *in silico* classification scores for further selection 0.20% and 0.00% of the top 500 scored generated compounds represent known actives, while the top 100 scored generated compounds had no rediscovery (more details are shown in Table S6). This decrease of rediscovery shows the prospective performance of QSAR models is too poor to achieve enrichment in this case, performing marginally better than random (Figure S1). Hence, we consistently find that rediscovery was much higher for public projects than in-house projects. Rediscovery was less than 1% for all generated compounds, and less than 3% for the top 100 scored compounds (Figure 6), which is significantly lower than in a previous study: In work by Segler et. al.⁵⁴ where the rediscovery ratio was around 10% for two bioactivity endpoints (which were growth inhibition endpoints though, namely inhibitory activity for *Plasmodium falciparum* and *Staphylococcus aureus*). However, methodological differences exist: In this previous study the test dataset was selected randomly and removed from the training dataset, which means that congeneric compounds might still exist in the training dataset, and then the generative models were fine-tuned. This explains the high rediscovery rates; however, this situation doesn't really resemble a real-world drug discovery situation. On the other hand, in a study performed by Atance et. al.⁵⁵, which removed a test dataset for DRD2 completely from the training dataset, the percentage of rediscovery was less than 1%; the condition of this study was more similar to ours with respect

to conditions and results obtained.

We found that rediscovery (the percentage of known actives present in the *de novo*-generated compounds), was greater in public projects (1.60%, 0.64%, and 0.21% of the top 100, 500, and all 5,000 generated compounds, respectively) than that in in-house projects (where the values were 0.00%, 0.03%, and 0.04%, respectively). This shows that the public dataset which was mapped on a pseudo-time axis behaves fundamentally different from a real-world drug discovery project, leading to very different numerical results.

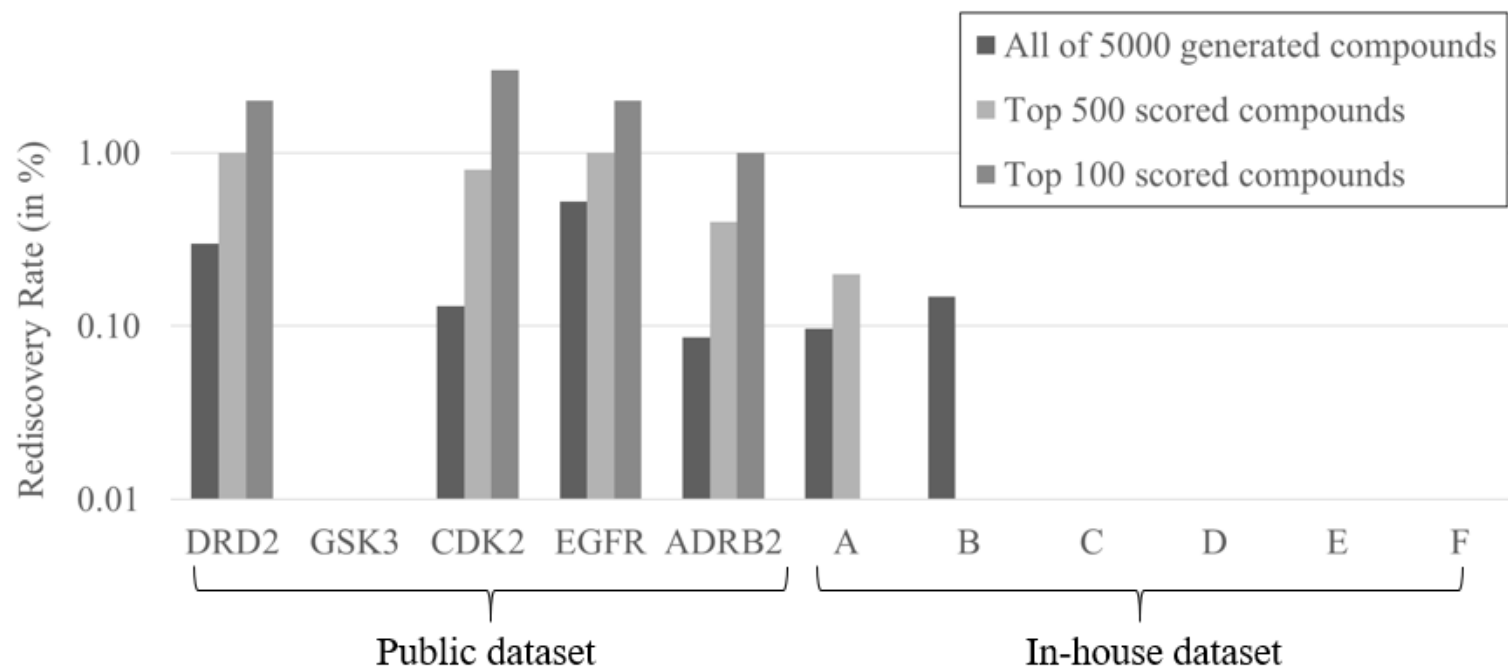


Figure 6 Rediscovery of compounds was significantly higher for public projects than in-house projects in the reinforcement learning (RL) setting. For further details see Table S6.

Similarity Analysis of Generated Compounds to Middle Stage Compounds

To investigate whether the generative model can produce compounds similar to known actives, we next calculated the aSNN (average Similarity of the Nearest Neighbour) between the generated compounds and known active compounds. The aSNN between generated compounds and the real compounds which belong to the middle stage are shown in Figure 7. For the public projects, given all of 5,000 generated compounds, aSNN through the projects of high/ultra-high activity compounds were much higher than that of low/middle activity compounds (the average of aSNN across projects for low/middle/high/ultra-high activity was 0.304/0.367/0.420/0.408, respectively)⁵⁶. Hence, for the public dataset, and given the particular way this dataset was constructed, optimization towards the single objective of primary target activity was possible. For the in-house projects those trends were inconsistent (the average of aSNN across projects for low/middle/high/ultra-high activity was 0.431/0.425/0.427/0.348, respectively). For project A and B the aSNN to high activity compounds was higher than the corresponding value for low/middle activity compounds; however, for projects C to F the aSNN of generated compounds to high/ultra-high compounds was conversely lower than that to low/middle activity compounds (Figure 7A). We can hence conclude that both datasets behaved very differently: While for the public dataset evolution towards the chemical space of higher-potency compounds was generally possible, this was not the case for the in-house projects analyzed.

Next we analyzed the compounds selected by the *in silico* classification model to investigate the effect of this data processing step. For the public dataset we found that the aSNN of the 500 compounds highest ranked by the *in silico* classification model was much higher than when all generated compounds were used, namely the compounds generated were more similar to the ultra-high activity compounds (the average of aSNN through projects for

low/middle/high/ultra-high activity was 0.329/0.424/0.531/0.540, respectively). For the in-house projects the compounds generated were more similar to the high activity compounds as well, but this is not the case for ultra-high activity compounds (the average of aSNN through projects for low/middle/high/ultra-high activity was 0.527/0.534/0.543/0.442, respectively). This trend held for the 100 compounds highest ranked by the *in silico* classification model, where the aSNN across projects for low/middle/high/ultra-high activity was 0.343/0.443/0.531/0.540 (for public data) and 0.563/0.579/0.602/0.489 (for in-house data), respectively. Especially for GSK3 and CDK2, the aSNN of high/ultra-high compounds was more than two times higher than when using all generated compounds (Figure 7B, C). Furthermore, in project C and F when using top 100 scoring, the aSNN of high activity compounds were higher than that of low/middle activity compounds (Figure 7C). Hence we can conclude that filtering by an *in silico* classification model has an overall beneficial effect to select higher activity compounds across most of the public and in-house datasets, with the magnitude of the effect widely varying.

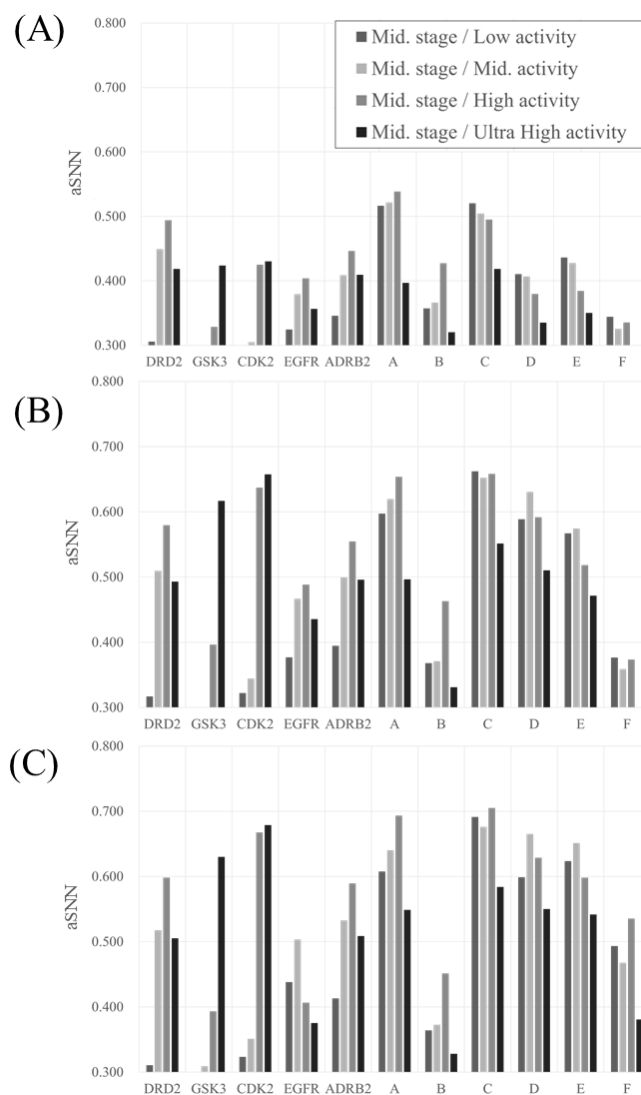


Figure 7 Average of the single nearest neighbour similarity (aSNN) between generated compounds and test compounds in the middle stage for all projects in reinforcement learning (RL) for (A) all 5,000 compounds generated, for (B) the highest-scored 500 compounds by an *in silico* classification model, and the (C) highest-scored 100 scored compounds by an *in silico* classification model.

It can be seen that activity model selection generally increases aSNN, with the magnitude of the effect widely varying across projects. The cut-off values of aSNN considered similar was set to be 0.3.

Similarity Analysis of Generated Compounds to Late Stage Compounds

Next, we analysed the aSNN between generated compounds and the real compounds which belong to the late project stage, which usually means both greater chemical evolution, and more bioactive (or generally optimized, with respect to the objective properties) compounds. Generally speaking, the assumption was that the more time elapsed, the more difficult it will be for the model to generate compounds similar to real late-stage project compounds. The results of this analysis are shown in Figure 8. It can be seen that the value of most aSNNs was lower than those in the middle stage (Figure 7). In the public projects, given all of 5,000 generated compounds, the aSNN of generated compounds to high/ultra-high activity compounds was higher than that to low/middle activity compounds (the average of aSNN values across projects for low/middle/high/ultra-high activity being 0.259/0.297/0.341/0.404, respectively). However, this was not the case with the in-house projects, with the average aSNN values across projects for low/middle/high/ultra-high activity being 0.376/0.357/0.361/0.311, respectively (Figure 8A). Hence, we can conclude that again both types of projects behave differently; for the public dataset evolution towards the chemical space of higher-potency compounds was generally possible, but not so for the in-house projects.

Then, to investigate the effect of score filtering, we analyzed the top scored of the generated compounds. Although the absolute value of aSNN was higher when we using top 500 (with the average aSNN across projects for low/middle/high/ultra-high activity being 0.281/0.336/0.402/0.516 for the public datasets and 0.455/0.429/0.441/0.370 for the in-house datasets) as well as the top 100 compounds (with the average of aSNN across projects for low/middle/high/ultra-high activity being 0.289/0.349/0.414/0.517 for the public datasets and 0.490/0.464/0.479/0.394 for the in-house datasets), there were not as drastic changes compared to those in the middle stage when performing the same analysis (Figure 7B, C and Figure 8B,

C). We can conclude that score filtering in the late stage is more difficult compared to the middle stage, and this might be derived from the greater time elapsed (and hence chemical space evolving) since the generation of the models.

In absolute terms, a similarity above ca. 0.3 in ECFP4/Tanimoto space (as a broad rule of thumb) often indicates similar bioactivity⁵⁷, which was in many cases reached by the current projects. Hence, in absolute terms, it seems that the chemistry generated should be suitable for drug discovery.

However, public and in-house projects behaved vastly different throughout the current analysis, which is understandable given the differences in how both datasets were constructed. For in-house datasets, the aSNN were mostly higher than 0.3; however, the aSNN to the real high or ultra-high active compounds was across projects consistently lower (Figure 7 and Figure 8) than to the real low or middle active compounds. This could be influenced by the somewhat artificial setup of this study, where we focused on a single objective, namely on-target activity; however, during any practical drug discovery project the consideration of multiple (and often competing) objectives is inevitable²⁰. This is supported by the analysis shown in Figure 5, where it is clear that in real projects compound evolution does not simply follow an optimization of on-target activity. Consequently, it is more difficult to reproduce a compound trajectory from real-world project data, compared to that of just optimizing on-target activity, which is what we consistently also observe from our results in this study.

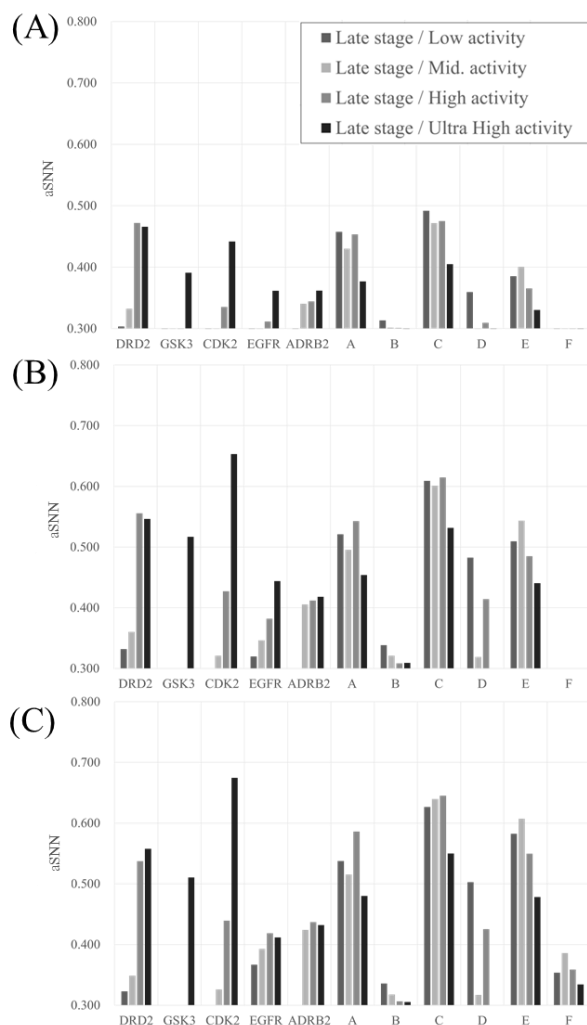


Figure 8 Average of single nearest neighbour similarity (aSNN) between generated compounds and test compounds in the late stage for all projects using reinforcement learning (RL) for (A) all 5,000 compounds generated, (B) the top 500 scored compounds by an *in silico* classification model, and (C) the top 100 scored compounds by an *in silico* classification model.

It can be seen that, generally speaking, values are lower than in Figure 6 for middle-stage compounds, and hence long-term compound evolution is much more difficult to model than short-term compound evolution. The cut-off values of aSNN considered similar was set to be 0.3.

Accuracy of *In Silico* Classification Models Did Not Affect The Result of Rediscovery and aSNN

Next, we investigated the relationship between *in silico* classification model predictivity and rediscovery/aSNN, to evaluate whether better bioactivity models also lead to higher values across these performance measures. The results of this analysis are shown in Figure S1A. It can be seen that balanced accuracies for the public dataset ranged from 0.660 (EGFR) to 0.747 (ADBR2), while those for the in-house dataset ranged from 0.708 (project F) to 0.828 (project A). Consequently, the balanced accuracy for the in-house dataset was a bit of higher than for the public dataset; however, in absolute terms, all of the balanced accuracies were high and all the *in silico* classification models obtained were considered to be worth being applied. However, converse to model performance, the rediscovery and aSNN obtained were actually higher in public projects than in-house ones (Figure 6, Figure 7, and Figure 8). We think this is inevitable since we evaluated the predictive models using the *early stage* compounds selected randomly as the external test set, given that using middle or late stage data as training or validation dataset would lead to information leakage during model generation and evaluation. So, finally the predictivity of *in silico* classification models could be worse for the later stage compounds than when we investigated it as a post-hoc analysis in Figure S1B. Our results suggest that solely based on the performance metrics of the *in silico* model on the external dataset we are not able to predict the success of RL generative models, because there was no relationship between balanced accuracy (on early-stage compound information available) and rediscovery/aSNN.

The Effect of Diversity Filter and Inception on Performance for Rediscovery and aSNN

We next investigated whether variations of the protocol, namely the use of a diversity filter and

inception, were able to improve performance metrics. Rediscovery obtained when using those options is shown in Table S6. Compared with the result of RL, we could not find any beneficial effect of a DF and inception when evaluating rediscovery. The aSNN results across model options in the middle stage are shown in Figure S2 and those for late stage are shown in Figure S3. We also could not find any effect of DF and inception which contribute to an increase in the similarity of generated compounds to high or ultra-high activity compounds when evaluating aSNN. However, especially for DF, this option surely contributed to avoidance of mode collapse when evaluating uniqueness and novelty (Table S4 and S5)⁵⁸. So, the effect of DF should be thought as fitting for purpose, such as scaffold hopping. Regarding inception, it has also shown before that it could contribute to the lead optimization process²⁶; however, in the current study also no beneficial effect could be observed.

Predictive Cluster Analysis

According to the investigation we performed so far, the generated compounds for the public dataset had better rediscovery ratio and aSNN than the in-house dataset (Figure 6, Figure 7, Figure 8). Firstly, we investigated the real activity of each compound, and concluded that the public dataset which was transformed on a pseudo-time axis has an explicit relationship between activity and pseudo time; however, for the in-house dataset this relationship was much less profound (Figure 5). In order to investigate this difference from the viewpoint of each compound's topology and predictivity for late-stage active compounds, we further analysed the public and in-house dataset by counting the compound's number in each region α , β , and γ after clustering by k-means. This analysis was meant to show whether each project had predictive or unproductive clusters for the generative model. Specifically, if the number of compounds

coming from α is not zero (i.e., active early stage compounds are contained in a given cluster) and the number of compounds from γ (i.e., late-stage active compounds) is higher than the number of compounds from β (i.e. late-stage inactive compounds), this is a ‘predictive’ cluster for the generative model, in the sense that the chemical space of the late-stage, desirable compounds is present in the early stage for the generative model (and vice versa for ‘unpredictive’ clusters). Table 1 shows the results of this analysis. It can be seen that for the public projects, most projects had more predictive clusters than unpredictable (the numbers of clusters which predictive/unpredictive in DRD2 was 7/1, for GSK this was: 4/4, for CDK2 4/3, for EGFR 5/4, and for ADRB2 5/5). On the other hand, in in-house projects, all of the models had more unpredictable clusters than predictive (the numbers of cluster predictive/unpredictive in project A was 0/6, in project B 1/5, in project C 1/6, in project D 0/5, in project E 0 /7, and in project F 1/3; Table 1). Based on this result, we conclude that the difference of rediscovery ratio and aSNN between the public and in-house projects might be based on whether projects have predictive clusters (and to which extent this is the case), or not. In this sense, we can formulate as a success criterion to utilize generative models in the drug discovery process is that we require seeds of promising compounds in the training dataset. Active learning might be a stepwise approach here. It should also be mentioned that the evaluation performed here is based on *known* actives only – hence, we were not able to evaluate the quality of potential ‘false positive’ compounds.

Table 1 Clustering of real compounds (public and in-house dataset) by k-means (k=10) and classification into α , β , and γ .

If the number of α is not zero and that of γ is more than that of β , this seemed to be predictive for the generative model (light grey column). On the other hand, if the number of α is not zero and that of γ is less than that of β , this seemed to be unproductive (dark grey column).

Public	DRD2			GSK3			CDK2			EGFR			ADRB2		
Cluster ID	α	β	γ	α	β	γ	α	β	γ	α	β	γ	α	β	γ
1	57	0	0	1	11	35	0	10	11	5	50	75	13	4	28
2	160	123	127	24	47	48	1	38	21	44	7	1	7	188	13
3	131	35	155	5	80	35	7	1	1	6	58	76	99	266	33
4	150	53	238	0	72	0	64	167	27	63	56	460	15	7	143
5	60	11	132	24	538	226	21	27	75	72	35	5	45	154	10
6	170	60	139	7	47	71	85	203	144	5	77	211	8	9	0
7	387	200	298	0	100	133	6	16	30	37	83	60	1	7	43
8	123	145	272	70	605	219	7	12	47	262	470	174	19	54	30
9	4	8	68	18	9	19	30	63	140	27	171	319	22	15	170
10	82	90	16	2	23	5	1	0	0	31	0	0	43	14	120
	Predictive: 7 clusters Unproductive: 1 cluster			Predictive: 4 clusters Unproductive: 4 clusters			Predictive: 4 clusters Unproductive: 3 clusters			Predictive: 5 clusters Unproductive: 4 clusters			Predictive: 5 clusters Unproductive: 5 clusters		

In-house	A			B			C			D			E			F		
Cluster ID	α	β	γ	α	β	γ	α	β	γ	α	β	γ	α	β	γ	α	β	γ
1	0	0	0	11	386	145	97	101	24	1	475	147	8	6	2	0	125	32
2	25	41	7	57	37	7	0	59	36	25	266	33	0	43	38	26	0	1
3	53	11	0	0	94	115	21	154	42	0	256	123	2	67	14	10	16	1
4	0	59	102	0	23	195	0	214	21	0	337	69	4	85	3	0	5	30
5	5	155	34	0	108	163	7	114	102	11	160	17	28	17	0	0	90	8
6	41	129	94	112	0	2	22	28	0	0	433	568	40	97	8	2	16	5
7	12	10	5	60	16	0	0	107	0	0	418	241	0	133	137	0	0	0
8	0	57	9	46	37	7	1	20	82	64	63	26	28	255	78	14	97	11
9	0	0	0	0	65	227	10	324	21	0	284	230	53	9	1	0	35	67
10	16	96	4	3	11	0	12	85	0	1	165	32	0	66	9	0	62	51
	Predictive: 0 cluster Unproductive: 6 clusters			Predictive: 1 cluster Unproductive: 5 clusters			Predictive: 1 cluster Unproductive: 6 clusters			Predictive: 0 clusters Unproductive: 5 clusters			Predictive: 0 cluster Unproductive: 7 clusters			Predictive: 1 cluster Unproductive: 3 clusters		

Case study of generated compounds' structure (DRD2)

In order to understand the chemistry of generated compounds better, we next examined it by visual inspections *via* clustering, using DRD2 ligands as an example. Representatives (most common structures (MCS) and highest scored structures (HSS)) from k-means clustering are depicted in Figure 9. Compared to known ligands (Figure 9A), there were some compounds that may require more careful consideration with regards to 'drug like' properties or synthetic ease. Regarding molecules from the pre-trained prior model (Figure 9B) there were, (1) two small, fragment-like molecules present in the set (MCS b-3, MCS b-9), (2) many oxygen atoms present (MCS b-7, HSS b-3), (3) connections like hydrazine (MCS b-2 and b-8), and cycle to cycle connection by one attachment point like tetrahydropyran to pyrrolidine (HSS b-8). Regarding the generated compounds by RL (Figure 9C) there were, (1) a few with long flexible chains (MCS c-1, MCS c-3), (2) sulfinyl groups present that are more commonly seen in antibiotics (HSS c-2, HSS c-4); however, there were no functional groups or idiosyncratic topologies without precedent in ChEMBL (i.e., the training data set). Therefore, the process of fine-tuning and RL using real chemical datasets had a beneficial effect on the generation of practical chemical structures.

(A)

Most Common					
	pXC50: 7.28 (a-0, CS: 605)	pXC50: 7.69 (a-1, CS: 479)	pXC50: 8.27 (a-2, CS: 188)	pXC50: 8.47 (a-3, CS: 224)	pXC50: 8.07 (a-4, CS: 388)
	pXC50: 6.66 (a-5, CS: 159)	pXC50: 8.15 (a-6, CS: 109)	pXC50: 6.80 (a-7, CS: 1292)	pXC50: 8.34 (a-8, CS: 692)	pXC50: 5.90 (a-9, CS: 388)
Highest Score					
	pXC50: 9.80 (a-0)	pXC50: 10.2 (a-1)	pXC50: 10.2 (a-2)	pXC50: 10.5 (a-3)	pXC50: 15.0 (a-4)
	pXC50: 11.5 (a-5)	pXC50: 8.15, identical (a-6)	pXC50: 12.5 (a-7)	pXC50: 10.4 (a-8)	pXC50: 8.10 (a-9)

(B)

Most Common					
	Score: 0.420 (b-0, CS: 379)	Score: 0.439 (b-1, CS: 400)	Score: 0.493 (b-2, CS: 577)	Score: 0.339 (b-3, CS: 382)	Score: 0.362 (b-4, CS: 467)
	Score: 0.380 (b-5, CS: 619)	Score: 0.370 (b-6, CS: 363)	Score: 0.323 (b-7, CS: 662)	Score: 0.362 (b-8, CS: 365)	Score: 0.352 (b-9, CS: 477)
Highest Score					
	Score: 0.772 (b-0)	Score: 0.753 (b-1)	Score: 0.743 (b-2)	Score: 0.682 (b-3)	Score: 0.836 (b-4)
	Score: 0.710 (b-5)	Score: 0.766 (b-6)	Score: 0.908 (b-7)	Score: 0.877 (b-8)	Score: 0.745 (b-9)

Figure 9 Example of DRD2 for the comparison of real (A) and generated compounds (B: from pre-trained prior model, C: from RL model) by visual inspection.

The number after CS is the number of compounds included in the same cluster.

(C)

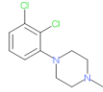
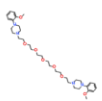
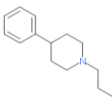
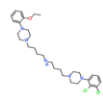
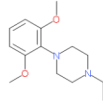
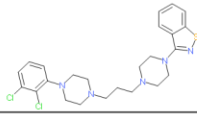
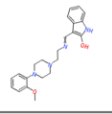
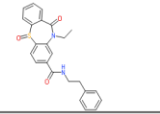

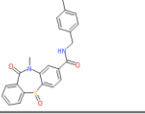
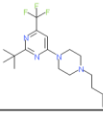
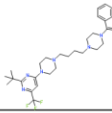
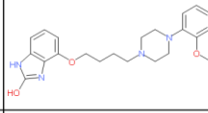
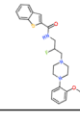
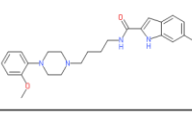
Most Common					
	Score: 0.526 (c-0, CS: 72)	Score: 0.838 (c-1, CS: 356)	Score: 0.458 (c-2, CS: 201)	Score: 0.890 (c-3, CS: 451)	Score: 0.480 (c-4, CS: 576)
Highest Score					
	Score: 0.954 (c-0)	Score: 0.978 (c-1)	Score: 0.848 (c-2)	Score: 0.993 (c-3)	Score: 0.969 (c-4)
					
	Score: 0.929 (c-5)	Score: 0.999 (c-6)	Score: 0.980 (c-7)	Score: 0.976 (c-8)	Score: 0.986 (c-9)

Figure 9 Example of DRD2 for the comparison of real (A) and generated compounds (B: from pre-trained prior model, C: from RL model) by visual inspection. (Continued)

Conclusion

In this research we asked the question “*Can a generative model trained on early-stage project compounds generate middle/late-stage compounds de novo?*” To this end, we used experimental data from five public and six in-house project datasets to model the elapsed time of a synthetic expansion following hit identification using REINVENT as a widely adopted RNN-based generative model. For the public dataset, data was mapped on a pseudo-time axis to reflect project progress. As a result, we found that rediscovery was much greater for public projects than that for in-house projects. The aSNN between early- and middle/late-stage compounds in public projects was higher between active compounds than inactive compounds; however, for in-house projects the converse was true. We next analyzed whether project compounds presented in predictive clusters, and we found compounds from in-house dataset did have fewer predictive clusters than public dataset. This criterion we found most predictive of the success of generative models (as measured here); while the performance of bioactivity models was not found to be predictive for this measure. Considering the difference in result between public and in-house dataset, objectively evaluating *de novo* compound design is hence, based on the current study, difficult or even impossible retrospectively, with large variation between projects. At the same time, we have shown that the generative model recovers very few middle/late-stage compounds from *real-world* drug discovery projects, highlighting the fundamental difference between human and automated design, as well as the difference between single-objective and multi-parameter optimization, with the latter being the norm in real-world drug discovery projects.

AUTHOR INFORMATION

Corresponding Authors

*Andreas Bender - Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK. E-mail: ab454@cam.ac.uk

*Koichi Handa - Toxicology & DMPK Research Department, Teijin Institute for Bio-medical Research, Teijin Pharma Limited, 4-3-2 Asahigaoka, Hino-shi, Tokyo 191-8512, Japan. E-mail: ko.handa@teijin.co.jp

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

The authors declare no conflicts of interest associated with this manuscript.

Data and Software Availability

The data that support the findings of this study are available on request from the corresponding author. All software used in this study was freely available.

ACKNOWLEDGMENT

The authors thank Hongbin Yang, Benoit Baillif at the university of Cambridge for help with the advice of research structure at some points, and Yohei Matsueda in TEIJIN Pharma Ltd. for help with the understanding of datasets.

ABBREVIATIONS

average of single nearest neighbour, aSNN; natural language process, NLP; multiple parameters optimization, MPO; recurrent neural network, RNNs; fully connected neural network, FCNNs; convolutional neural network, CNNs; variational auto encoder, VAE; Generative Adversarial Networks, GAN; long-short time memory, LSTM; Dopamine Receptor D2, DRD2; Glycogen synthase kinase 3, GSK3; Cyclin-dependent kinase 2, CDK2; Epidermal Growth Factor Receptor, EGFR; Anrenergic receptor beta2, 2616 active compounds, ADRB2; reinforcement learning, RL; random forest, RF; diversity filter, DF;

References

- 1 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>.
- 2 Thomas, M.; Boardman, A.; Garcia-Ortegon, M.; Yang, H.; de Graaf, C.; Bender, A. Applications of Artificial Intelligence in Drug Design: Opportunities and Challenges. *Methods Mol. Biol.* **2022**, *2390*, 1–59. https://doi.org/10.1007/978-1-0716-1787-8_1.
- 3 Scannell, J. W.; Bosley, J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. *PLoS One* **2016**, *11* (2), e0147215. <https://doi.org/10.1371/journal.pone.0147215>.
- 4 Plowright, A. T.; Johnstone, C.; Kihlberg, J.; Pettersson, J.; Robb, G.; Thompson, R. A. Hypothesis Driven Drug Design: Improving Quality and Effectiveness of the Design-Make-Test-Analyse Cycle. *Drug Discov. Today* **2012**, *17* (1–2), 56–62. <https://doi.org/10.1016/j.drudis.2011.09.012>.
- 5 Danziger, D. J.; Dean, P. M. Automated Site-Directed Drug Design: A General Algorithm for Knowledge Acquisition about Hydrogen-Bonding Regions at Protein Surfaces. *Proc. R. Soc. London. Ser. B, Biol. Sci.* **1989**, *236* (1283), 101–113. <https://doi.org/10.1098/rspb.1989.0015>.
- 6 Douguet, D.; Thoreau, E.; Grassy, G. A Genetic Algorithm for the Automated Generation of Small Organic Molecules: Drug Design Using an Evolutionary Algorithm. *J. Comput. Aided. Mol. Des.* **2000**, *14* (5), 449–466. <https://doi.org/10.1023/a:1008108423895>.
- 7 Schneider, G.; Lee, M. L.; Stahl, M.; Schneider, P. De Novo Design of Molecular Architectures by Evolutionary Assembly of Drug-Derived Building Blocks. *J. Comput. Aided. Mol. Des.* **2000**, *14* (5), 487–494. <https://doi.org/10.1023/a:1008184403558>.
- 8 Pandey, M.; Fernandez, M.; Gentile, F.; Isayev, O.; Tropsha, A.; Stern, A. C.; Cherkasov, A. The Transformational Role of GPU Computing and Deep Learning in Drug Discovery. *Nat. Mach. Intell.* **2022**, *4* (3), 211–221. <https://doi.org/10.1038/s42256-022-00463-x>.
- 9 Gawehn, E.; Hiss, J. A.; Brown, J. B.; Schneider, G. Advancing Drug Discovery via GPU-Based Deep Learning. *Expert Opin. Drug Discov.* **2018**, *13* (7), 579–582. <https://doi.org/10.1080/17460441.2018.1465407>.
- 10 Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18* (6), 463–477. <https://doi.org/10.1038/s41573-019-0024-5>.
- 11 Vogt, M. Exploring Chemical Space — Generative Models and Their Evaluation. *Artif. Intell. Life Sci.* **2023**, *3*, 100064. <https://doi.org/10.1016/j.aillsi.2023.100064>.

- 12 Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*, 565644. <https://doi.org/10.3389/fphar.2020.565644>.
- 13 Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; Klambauer, G. Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J. Chem. Inf. Model.* **2018**, *58* (9), 1736–1741. <https://doi.org/10.1021/acs.jcim.8b00234>.
- 14 Bender, A.; Schneider, N.; Segler, M.; Patrick Walters, W.; Engkvist, O.; Rodrigues, T. Evaluation Guidelines for Machine Learning Tools in the Chemical Sciences. *Nat. Rev. Chem.* **2022**, *6* (6), 428–442. <https://doi.org/10.1038/s41570-022-00391-9>.
- 15 <https://cache-challenge.org/> (access date: December 2nd, 2022)
- 16 Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59* (3), 1096–1108. <https://doi.org/10.1021/acs.jcim.8b00839>.
- 17 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* (D1), D945–D954. <https://doi.org/10.1093/nar/gkw1074>.
- 18 Thomas, M.; O’Boyle, N. M.; Bender, A.; De Graaf, C. Re-Evaluating Sample Efficiency in de Novo Molecule Generation. **2022**. <https://doi.org/https://arxiv.org/abs/2212.01385>.
- 19 Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53* (4), 783–790. <https://doi.org/10.1021/ci400084k>.
- 20 Bender, A.; Cortes-Ciriano, I. Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 2: A Discussion of Chemical and Biological Data. *Drug Discov. Today* **2021**, *26* (4), 1040–1052. <https://doi.org/10.1016/j.drudis.2020.11.037>.
- 21 Beckers, M.; Fechner, N.; Stiefl, N. 25 Years of Small-Molecule Optimization at Novartis: A Retrospective Analysis of Chemical Series Evolution. *J. Chem. Inf. Model.* **2022**. <https://doi.org/10.1021/acs.jcim.2c00785>.
- 22 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>.
- 23 He, J.; You, H.; Sandström, E.; Nittinger, E.; Bjerrum, E. J.; Tyrchan, C.; Czechtizky, W.; Engkvist, O. Molecular Optimization by Capturing Chemist’s Intuition Using Deep Neural Networks. *J. Cheminform.* **2021**, *13* (1), 26. <https://doi.org/10.1186/s13321-021-00497-0>.

- 24 Delaney, J. Modelling Iterative Compound Optimisation Using a Self-Avoiding Walk. *Drug Discov. Today* **2009**, *14* (3–4), 198–207. <https://doi.org/10.1016/j.drudis.2008.10.007>.
- 25 Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminform.* **2017**, *9* (1), 48. <https://doi.org/10.1186/s13321-017-0235-x>.
- 26 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>.
- 27 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>.
- 28 Sewak, M.; Sahay, S. K.; Rathore, H. An Overview of Deep Learning Architecture of Deep Neural Networks and Autoencoders. *J. Comput. Theor. Nanosci.* **2020**, *17* (1), 182–188. <https://doi.org/10.1166/jctn.2020.8648>.
- 29 Bouwmans, T.; Javed, S.; Sultana, M.; Jung, S. K. Deep Neural Network Concepts for Background Subtraction: A Systematic Review and Comparative Evaluation. *Neural Netw.* **2019**, *117*, 8–66. <https://doi.org/10.1016/j.neunet.2019.04.024>.
- 30 Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput. Aided. Mol. Des.* **2016**, *30* (8), 595–608. <https://doi.org/10.1007/s10822-016-9938-8>.
- 31 De Cao, N.; Kipf, T. MolGAN: An Implicit Generative Model for Small Molecular Graphs. **2018**.
- 32 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>.
- 33 Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. **2014**, 1–9.
- 34 Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. **2017**.
- 35 Ertl, P.; Lewis, R.; Martin, E.; Polyakov, V. In Silico Generation of Novel, Drug-like Chemical Matter Using the LSTM Neural Network. **2017**.
- 36 He, J.; Nittinger, E.; Tyrchan, C.; Czechtizky, W.; Patronov, A.; Bjerrum, E. J.; Engkvist, O. Transformer-Based Molecular Optimization beyond Matched Molecular Pairs. *J. Cheminform.* **2022**, *14* (1), 18. <https://doi.org/10.1186/s13321-022-00599-3>.
- 37 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. *J. Cheminform.* **2021**, *13* (1), 89. <https://doi.org/10.1186/s13321-021-00563-7>.

- 38 Marques, G.; Leswing, K.; Robertson, T.; Giesen, D.; Halls, M. D.; Goldberg, A.; Marshall, K.; Staker, J.; Morisato, T.; Maeshima, H.; Arai, H.; Sasago, M.; Fujii, E.; Matsuzawa, N. N. De Novo Design of Molecules with Low Hole Reorganization Energy Based on a Quarter-Million Molecule DFT Screen. *J. Phys. Chem. A* **2021**, *125* (33), 7331–7343. <https://doi.org/10.1021/acs.jpca.1c04587>.
- 39 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. Cheminform.* **2021**, *13* (1), 39. <https://doi.org/10.1186/s13321-021-00516-0>.
- 40 Thomas, M.; O’Boyle, N. M.; Bender, A.; de Graaf, C. Augmented Hill-Climb Increases Reinforcement Learning Efficiency for Language-Based de Novo Molecule Generation. *J. Cheminform.* **2022**, *14* (1), 68. <https://doi.org/10.1186/s13321-022-00646-z>.
- 41 Blaschke, T.; Bajorath, J. Fine-Tuning of a Generative Neural Network for Designing Multi-Target Compounds. *J. Comput. Aided. Mol. Des.* **2022**, *36* (5), 363–371. <https://doi.org/10.1007/s10822-021-00392-8>.
- 42 Blaschke, T.; Engkvist, O.; Bajorath, J.; Chen, H. Memory-Assisted Reinforcement Learning for Diverse Molecular de Novo Design. *J. Cheminform.* **2020**, *12* (1), 68. <https://doi.org/10.1186/s13321-020-00473-0>.
- 43 Yoshimori, A.; Kawasaki, E.; Kanai, C.; Tasaka, T. Strategies for Design of Molecular Structures with a Desired Pharmacophore Using Deep Reinforcement Learning. *Chem. Pharm. Bull. (Tokyo)*. **2020**, *68* (3), 227–233. <https://doi.org/10.1248/cpb.c19-00625>.
- 44 Sun, J.; Jeliaskova, N.; Chupakin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliaskov, V.; Kochev, N.; Ashby, T. J.; Chen, H. ExCAPE-DB: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. *J. Cheminform.* **2017**, *9*, 17. <https://doi.org/10.1186/s13321-017-0203-5>.
- 45 Sayers, E. W.; Beck, J.; Bolton, E. E.; Bourexis, D.; Brister, J. R.; Canese, K.; Comeau, D. C.; Funk, K.; Kim, S.; Klimke, W.; Marchler-Bauer, A.; Landrum, M.; Lathrop, S.; Lu, Z.; Madden, T. L.; O’Leary, N.; Phan, L.; Rangwala, S. H.; Schneider, V. A.; Skripchenko, Y.; Wang, J.; Ye, J.; Trawick, B. W.; Pruitt, K. D.; Sherry, S. T. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2021**, *49* (D1), D10–D17. <https://doi.org/10.1093/nar/gkaa892>.
- 46 Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460–473. <https://doi.org/10.1021/ci500588j>.
- 47 Ertl, P.; Patiny, L.; Sander, T.; Rufener, C.; Zasso, M. Wikipedia Chemical Structure Explorer: Substructure and Similarity Searching of Molecules from Wikipedia. *J. Cheminform.* **2015**, *7*, 10. <https://doi.org/10.1186/s13321-015-0061-y>.
- 48 RD-kit: <https://www.rdkit.org/docs/index.html#> (access date: June 5th, 2023)

- 49 Sousa, T.; Correia, J.; Pereira, V.; Rocha, M. Generative Deep Learning for Targeted Compound Design. *J. Chem. Inf. Model.* **2021**, *61* (11), 5343–5361. <https://doi.org/10.1021/acs.jcim.0c01496>.
- 50 Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. <https://doi.org/10.1021/ci100050t>.
- 51 Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (2), 5–32. <https://doi.org/https://doi.org/10.1023/A:1010933404324>.
- 52 Du, Y.; Fu, T.; Sun, J.; Liu, S. MolGenSurvey: A Systematic Survey in Machine Learning Models for Molecule Design. **2022**.
- 53 Bjerrum, E. J.; Margreitter, C.; Blaschke, T.; de Castro, R. L.-R. Faster and More Diverse de Novo Molecular Optimization with Double-Loop Reinforcement Learning Using Augmented SMILES. **2022**.
- 54 Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4* (1), 120–131. <https://doi.org/10.1021/acscentsci.7b00512>.
- 55 Atance, S. R.; Diez, J. V.; Engkvist, O.; Olsson, S.; Mercado, R. De Novo Drug Design Using Reinforcement Learning with Graph-Based Deep Generative Models. *J. Chem. Inf. Model.* **2022**, *62* (20), 4863–4872. <https://doi.org/10.1021/acs.jcim.2c00838>.
- 56 Jasial, S.; Hu, Y.; Vogt, M.; Bajorath, J. Activity-Relevant Similarity Values for Fingerprints and Implications for Similarity Searching. *F1000Research* **2016**, *5*. <https://doi.org/10.12688/f1000research.8357.2>.
- 57 Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. *Org. Biomol. Chem.* **2004**, *2* (22), 3256–3266. <https://doi.org/10.1039/B409865J>.
- 58 Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model.* **2018**, *58* (6), 1194–1204. <https://doi.org/10.1021/acs.jcim.7b00690>.