

Table of Contents

Abstract	2
Introduction	2
General workflow	5
Data preparation	6
Model training	8
Case specific usage: Focusing a prior via Reinforcement Learning	8
Methods	9
Model architecture	9
Validation set	10
Pretraining the prior via teacher forcing	12
Focusing the prior via reinforcement learning	13
Motivation	13
Mathematical Background	14
Policy Iteration Rewards	15
Experiments	18
Evaluation metrics	19
Results	20
Comparison of the learning strategies	20
Comparison of slicing strategies	23
Following specific reactions	28
Scaffolds with varied numbers of attachment points	32
Discussions	34
Conclusions	36
Acknowledgements	37
Associated content	37
Supporting information	37
References	37

Lib-INVENT: Reaction Based Generative Scaffold Decoration for in silico Library Design

Vendy Fialková[§], Jiayi Zhao^{§,⊥}, Kostas Papadopoulos[§], Ola Engkvist[§], Esben Jannik Bjerrum[§], Thierry Kogej[§], Atanas Patronov^{*§}

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

[⊥] Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

* Corresponding author: atanas.patronov@astrazeneca.com

Abstract

Due to the strong relationship between desired molecular activity to its structural core, screening of focused, core sharing chemical libraries is a key step in lead optimisation. Despite the plethora of current research focused on *in silico* methods for molecule generation, to our knowledge, no tool capable of designing such libraries has been proposed. In this work, we present a novel tool for *de novo* drug design called Lib-INVENT. This is capable of rapidly proposing chemical libraries of compounds sharing the same core while maximising a range of desirable properties. To further help the process of designing focused libraries, the user can list specific chemical reactions that can be used for the library creation. Lib-INVENT is therefore a flexible tool for generating virtual chemical libraries for lead optimisation in a broad range of scenarios. Additionally, the shared core ensures that the compounds in the library are similar, possessing desirable properties and can be also synthesized under the same or similar conditions. The Lib-INVENT code is freely available in our public repository: <https://github.com/MolecularAI/Lib-INVENT>. The code necessary for data preprocessing is further available at: <https://github.com/MolecularAI/Lib-INVENT-dataset>.

Introduction

With the recent advances in deep learning techniques, such techniques are becoming increasingly popular tools in a range of areas – from automated vehicles to medicinal chemistry^{1,2}. This is especially true for drug discovery where the symbiosis between machine learning models and human experts has the potential to significantly speed up the process of early drug discovery³. Due to their generalisation

abilities, deep generative models have become the core engine in most recent *de novo* design tools^{4,5}. Despite the progress in the field of deep learning such tools are still in the early stages of development as they are adapting to satisfy the more specific needs of drug design⁶.

One of these specific requirements is in the lead optimization stage when aiming to use focused libraries of small molecules to identify a promising lead compound^{7,8}. Generally, the purpose of lead optimization is to retain the favourable properties of the compound while optimizing properties which still prevents the compound from becoming a drug candidate⁹. Since the desired activity is normally tied up to a given scaffold¹⁰, this use case boils down to retaining a certain molecular core and varying only specific moieties to satisfy the complex demands on the properties of the candidate molecule⁷. In practice this can be addressed by screening very focused libraries that share the same core¹¹. As an ideal scenario when creating such a library in the lab, it should be possible to introduce the proposed moieties via the same or similar reactions to ensure that they can be carried out under the same conditions. Related investigations have been conducted on a much smaller scale in the works on matched molecular pairs¹² and fragment linking⁷; the explorations have however not been previously extended to library generation nor considered chemical reactions.

For the purpose of this paper, we define a chemical library as follows:

Definition 1: Given a scaffold s , library is a set of molecules with conditions:

1. All include substructure s
2. All molecules are accessible by the same sequence of synthetically relevant chemical transformations.

In this paper, we propose a solution based on *de novo* generative model capable of addressing the use cases outlined above. Building on the REINVENT framework¹³, we extend the objective from a single compound design to a library design. Specifically, the model can suggest moieties to decorate an input scaffold with a variable number of attachment points for these decorations. In addition, the model can be put in a reinforcement learning (RL) scenario in order to learn to maximise a user defined set of

objectives. The resulting ideas will therefore be focused according to specific lead optimization goals. In contrast with prior work on scaffold decoration, these goals may include a set of reactions assigned to each attachment point of the scaffold so that the model learns to produce moieties attachable to that specific attachment point in agreement with the given reactions. This way of generating chemical libraries gives the user a significant level of control over the output, enabling them to focus the model's creativity and leverage prior knowledge¹⁴. Satisfying condition 2 of the library definition further means that the generated library is more suitable for automation in the design and execution stage by reducing the number of reagents and reactions required in synthesis. It further allows the chemist to optimally select reactions with a desired profile, which includes but is not limited to considerations of efficiency, literature coverage, or safety. Thus, the number of DMTA (design-make-test-analyse) cycles required in the drug discovery process decreases, improving the productivity of the incorporation of a generative model in the lead optimisation pipeline.

The original REINVENT algorithm¹⁵ proposes optimal compounds solving a specific user-defined objective and the recent GraphINVENT extends this to work to molecular graphs¹⁶. The algorithm introduced here, called Lib-INVENT, takes the work further and closer towards utilization of chemistry automation platforms by building focused, easy synthesisable libraries. Related models have appeared in the literature over the recent years, focusing both on scaffold decoration itself or on the usage of reinforcement learning to guide the decorative process^{14,17,18}. The major enhancement Lib-INVENT brings to these methods lies in the volume and diversity of its output within a focused chemical space. Crucially, the fact that the generated libraries can be produced from the same starting scaffold using specific chemical reactions facilitates the uptake of the ideas in a wet lab environment and contributes to the possibility of automation of the drug design process. By focusing on designing and synthesizing libraries instead of single molecules the learning in each Design-Make-Test-Analyse cycle can be increased and accordingly there is a need for fewer cycles to reach a clinical candidate.

General workflow

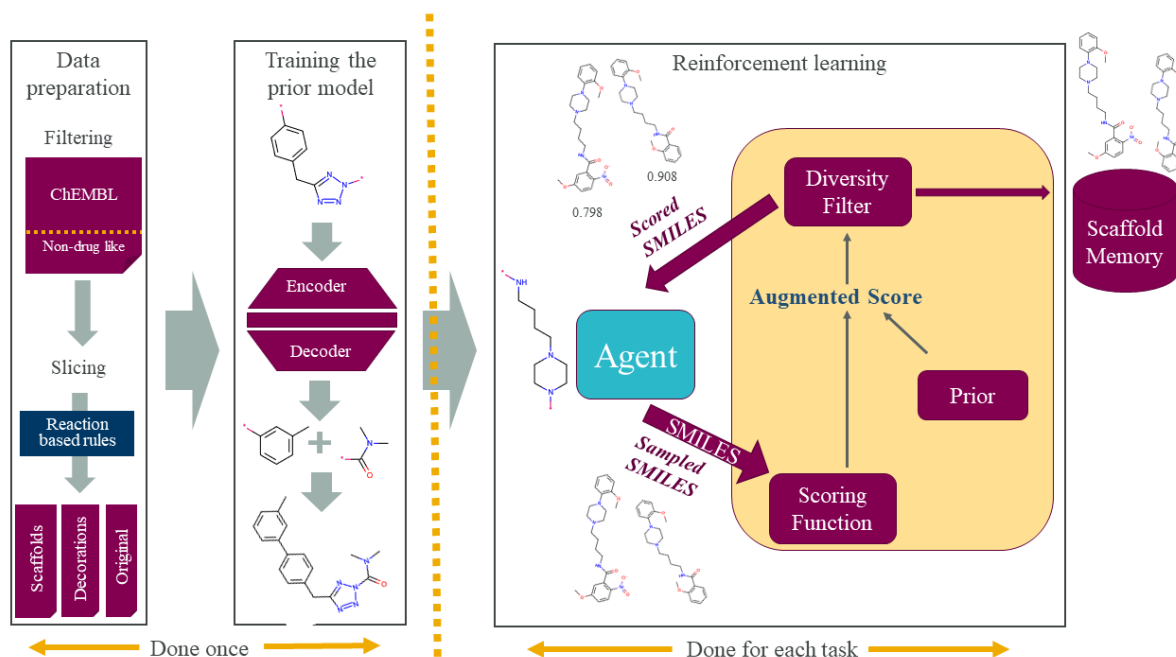


Figure 1: The general workflow of the model.

This section gives a high level overview of the individual steps of data preparation, model training and the usage of the algorithm to optimise various user-defined objectives. Figure 1 shows an overview of the workflow. Specific technical details will be further discussed in the section Methods. The motivation for the choices made in both data preparation and model design is further explained in the Discussion.

The model is a recurrent neural network (RNN) which takes a scaffold as an input and returns complete compounds obtained by attaching decorations to the input scaffold. There are two stages to the training: firstly, a general model is trained to learn the syntax of the SMILES language. We shall henceforth refer to this model as the prior and stress that the training of the prior is not specific to a particular task and thus only occurs once. The second step corresponds to the general usage of this model and is analogous to the usage of the past REINVENT models: the prior is focused to solve a user defined objective. In the case of Lib-INVENT, this is achieved through reinforcement learning and may involve a requirement of fulfilment of specific chemical reaction types. Starting from a scaffold of interest, the general prior thus rapidly adapts to propose a focused chemical library consisting of thousands of compounds sharing

a scaffold and chemical properties. Importantly, the generated compounds are collected during the RL process and not after, meaning that the model is typically no longer used after completing the RL run.

Data preparation

Compounds from the publicly available ChEMBL Database, version 27¹⁹, represented by SMILES strings, were used to train the prior model. This choice of representing the chemical compounds by sequences of characters has several advantages and is common in the cheminformatics literature²⁰. Firstly, despite losing a certain level of chemical information¹¹, this representation is significantly more memory efficient than the use of molecular graph data while implicitly retaining the molecular graph structure. Moreover, the SMILES strings are compatible with the chemical reactions expressed using the SMARTS language. This is crucial in the context of this work which focuses on incorporating knowledge of chemical pathways directly into the generative model.

As is standard for computational applications in drug discovery, the first step of the data preparation process involves data purging and sanitisation²¹. The purpose of purging is to remove undesirable compounds and outliers from the dataset⁶. This among others includes molecules containing rare SMILES tokens or elements which the model is unlikely to be able to learn and thus merely pollute the model's vocabulary, molecules with extremely large or low molecular weights or salts, which are neither drug-like nor chemically friendly. Approximately 25 % of the compounds present in the database have been removed at this stage. For details on the implementation and filter criteria, see the supporting information.

The second pre-processing step necessary to train a scaffold decorating model is compound slicing. There are many ways of slicing a molecule to obtain scaffold-decoration pairs for training a decorator model²². Recently, exhaustive slicing of single-bonds according to RECAP²³ rules has been explored^{17,18}. While this approach appears natural at a glance, it is not always effective for a wet lab chemist attempting to synthesise the proposed compounds²⁴. The ability to follow real chemical reactions when decorating the scaffold is crucial; our experiments demonstrate that training on data sliced according to RECAP rules does not teach the prior to understand these chemical principles. This means that the model is unable to satisfy reaction requirements when designing chemical libraries.

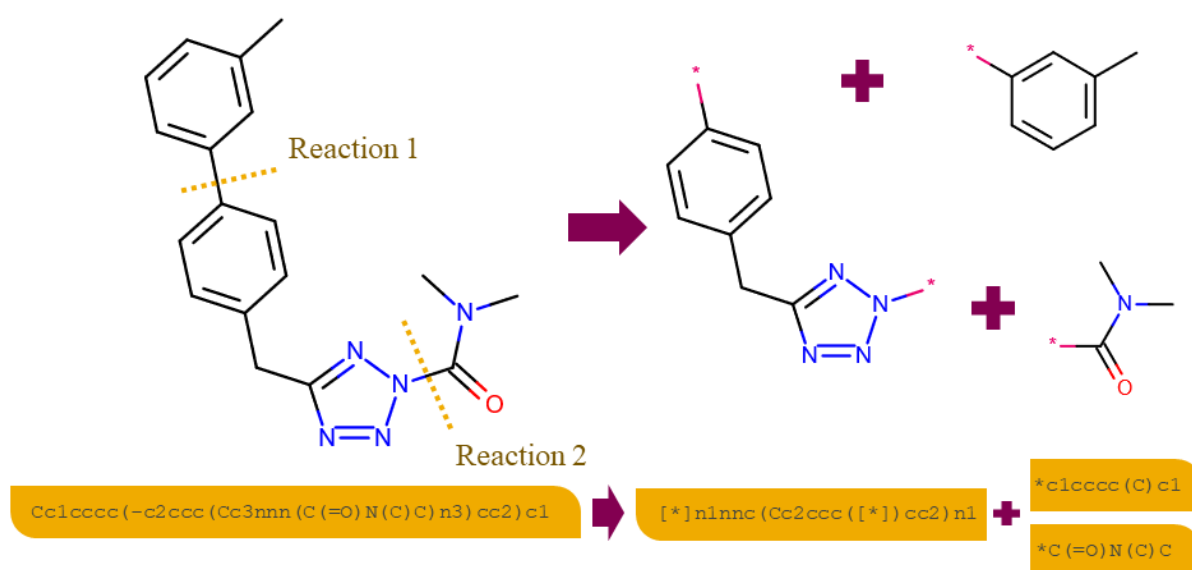


Figure 2: An example of a sliced molecule resulting in a scaffold with two attachment points and the corresponding decorations.

Practical synthesis and chemical considerations should thus be taken into account when slicing the molecules to ensure that the reverse process (forward synthesis) is synthetically valid²⁵. In their recent paper, Horwood and Noutahi²⁶ propose incorporating chemical synthesis routes directly into the model by designing a *de novo* generator based on chemical reactions. Given starting reactants, their model proposes drug-like molecules by selecting other appropriate reactants as well as specific reactions used to transform and connect the molecules into a resulting compound. This novel approach significantly improves the synthesizability of the proposed molecules; it however still lacks the ability to design libraries as well as the degree of flexibility and generality desirable in *de novo* generators. Specifically,

training has to occur on a dataset relevant to the final task at hand and there is limited capacity for knowledge transfer and extension to more specific tasks without retraining.

In this work, a novel data pre-processing approach is proposed to build a knowledge of chemical reactions directly on the training dataset comprised of the filtered ChEMBL database. Reaction based rules are used to slice the training compounds into scaffolds and decorations so that each split is a result of a known, easily implementable chemical reaction. We demonstrate that this method enables the generative model to propose decorations according to the chemical reactions used in training. The output therefore benefits from high validity and better likelihood of being synthetically feasible. An illustration of the process is provided in Figure 2.

Model training

The prior model is trained using the teacher forcing algorithm²⁷ to maximise the conditional likelihoods of the generated compounds given the scaffolds. Even a single pass through the dataset teaches the model to generate chemically valid SMILES strings; the optimal state balancing the coverage of chemical space and overfitting is however reached after approximately eight epochs. At each epoch, a different randomised representation of the training and validations SMILES is used as this further prevents overfitting²⁸. As mentioned previously, it is crucial to note that this training only needs to happen once since the prior can be reused for a wide range of tasks without additional transfer learning stages often required in previously introduced models²⁹.

Case specific usage: Focusing a prior via Reinforcement Learning

Case specific usage of the model involves focusing the prior on a specific task. This finetuning is efficiently achieved by setting up a reinforcement learning loop in which the prior iteratively proposes compounds receives task-specific rewards for its output. During the run, all high scoring compounds are stored in a virtual chemical library called scaffold memory; the production of the library therefore begins instantaneously once a RL run is set up and continues throughout the training of the RL agent.

The rewards are shaped by a scoring function defining desirable chemical or structural properties and guide the model towards producing compounds of interest⁸. However, since the objective is to explore

a rather narrow space of solutions (molecules) designed for a given scaffold, this may lead to a mode collapse³⁰. To achieve a stable RL process we introduce a mechanism that relies on Diversity filters (DF) previously described by Blaschke *et al.*¹⁵. Diversity filters and prior likelihoods of the proposed compounds can be included when calculating the reward. Diversity filters penalise the RL agent for repeatedly generating the same compound, which significantly reduces the risk of mode collapse towards a single high scoring solution (molecule). The prior likelihood serves as an additional regulariser, anchoring the agent to the previously learnt chemical space and ensuring that the SMILES syntax is not forgotten¹³.

Another reward modifying factor are the reaction filters (RF). The introduction of reaction filters to the learning process means that the proposed libraries can be synthesised using selected reactions, facilitating the creation of focused libraries. RFs are designed to be selective, so that a reaction or a set of reactions can be specified for each attachment point of the scaffold. This gives the user significant control over the output of the model and enables leveraging prior chemical knowledge. The full practical implementation of the RL procedure is described along with its mathematical background in the section Focusing the prior via reinforcement learning. A number of reaction definitions is further published in our public repository.

We emphasize that focusing the pretrained prior using reinforcement learning makes the Lib-INVENT decorator model widely applicable in a variety of real world scenarios with a range of reactants. Libraries containing thousands of high scoring, synthesisable molecules can be obtained within minutes or tens of minutes while the more expensive training of the prior model does not need to be repeated for new libraries.

Methods

Model architecture

The architecture of the model is analogous to the scaffold decorator introduced by Arús-Pous *et al.*¹⁷. The decorator model uses an encoder-decoder architecture where both the encoder and decoder are

RNNs with three hidden layers of dimension 512 and the embedding is of size 256. During training, dropout at rate 0.1 has been used.

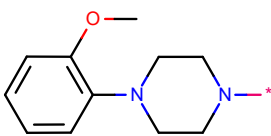
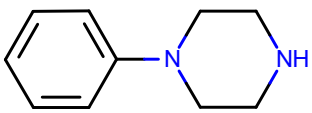
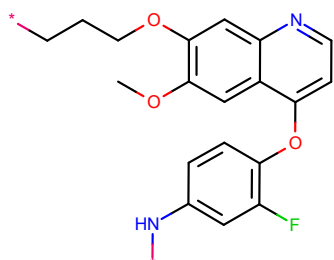
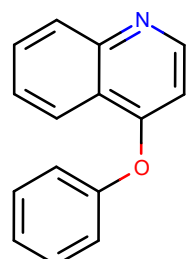
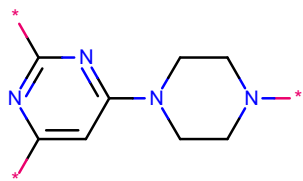
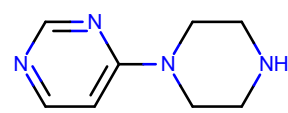
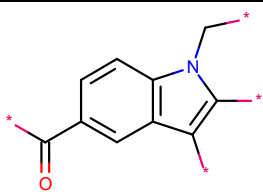
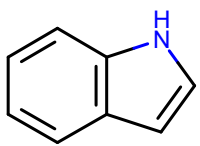
We refer to the collection of tokens recognized and used by the model as the vocabulary. This is composed of all the SMILES characters present in the pruned training dataset and enriched by the special ‘END’ and ‘START’ tokens determining the beginning and ending of a SMILES string. The length of the vocabulary corresponds to the dimension of the multinomial distribution over which the tokens are sampled. For details of the tokens included, see supporting information.

Validation set

As in any machine learning model, a good validation set is required in order to fairly evaluate the performance³¹. The objective of the prior model training is to learn to decorate scaffolds so that the resulting compounds lie in the drug-like chemical space spanned by the training dataset. This nature of the modelling objective affects the choices made when preparing the validation set. While it is common to randomly hold out a portion of the training data and use these for evaluations³², the sliced dataset used here does not lend itself well to this approach. To be able to fairly judge the generalisation ability of the model on previously unseen scaffolds, it is necessary to ensure that the validation scaffolds are not present in the training dataset. Optimally, even compounds structurally similar to these need to be removed from the training set to validate the performance of the model fairly³³. At the same time, the general distribution of the validation scaffold properties should mimic that of the training set to evaluate how well the model learns to follow the data distribution.

With these considerations in mind, the training-validation split was handcrafted by selecting one scaffold with each number of attachment points at random. Then, all scaffolds sharing a Murcko scaffold³⁴ with one of the four validation ones have been removed from the training set and added to validation. The choice to consider only the molecular cores consisting of ring structures stripped of side chains is motivated by the fact that the removed sets of compounds resemble the concept of ‘chemical series’ as used by medicinal chemists¹⁰. Removing entire chemical series based on a specific Murcko scaffold thus naturally reduces the bias in model evaluation and objectively tests its generalisation ability.

Table 1: The held out validation scaffolds.

Library scaffold	Bemis-Murcko scaffold
 <chem>[*]N1CCN(c2ccccc2OC)CC1</chem>	
 <chem>[*]CCOC1CC2NCCC(OC3CCC(N[*])CC3F)C2CC1OC</chem>	
 <chem>[*]N1CCN(c2cc([*])nc([*])n2)CC1</chem>	
 <chem>[*]Cn1c([*])c([*])c2cc(C(=O)[*])ccc21</chem>	

Because the aim of our experiments is to showcase some of the capabilities of the Lib-INVENT generative model against publicly known and commonly discussed DRD2 target^{17,18}, we remove all compounds sharing a scaffold with compounds found in the dataset obtained by slicing the DRD2 scaffolds according to the set of reactions previously used to slice ChEMBL³⁵. This way, the training

and validation sets are kept independent and the subsequent validation on the DRD2 target remains unbiased.

Representative scaffolds for the held out compounds used for validation are shown in Table 1 along with their Bemis-Murcko representations. The resulting validation set excluding the DRD2 data contained 241,137 unique entries. The training set contained the remaining 23,080,572 entries. These numbers show that the consideration of Bemis-Murcko scaffolds filters out a non-negligible number of compounds very similar to the original held out scaffold. At the same time, the size of this dataset means that despite the validation representing only about 1 % of the data, sufficient information is still included in order to assess the generative ability of the decorator.

Up until now, all SMILES have been canonicalized to ensure uniqueness. However, using different SMILES representations during training of deep learning models, leads to improvements of generalizability, both in activity modelling³⁶, representation learning³⁷ and SMILES generation. Before training the generative model, a different randomised representation of the training dataset is obtained for each epoch of teacher forcing training. The same is applied to the validation set.

Pretraining the prior via teacher forcing

As mentioned before, the training process of Lib-INVENT resembles the training of REINVENT 2.0¹⁵. First, teacher's forcing³⁸ is used to train the prior model capable of creating chemically valid compounds containing a given scaffold. In our case, the prior is an RNN taking a scaffold as an input and returning relevant decorations to connect to all available attachment points of the scaffold, much like the model recently introduced by Arús-Pous *et al*¹⁷. The number of outputted molecules corresponds to the batch size.

The generation process can be seen as a sequential conditional likelihood maximisation problem. The output of the model represents a probability distribution over the token space containing all the possible SMILES tokens present in the training dataset enriched by the 'START' and 'END' tokens, given the scaffold and previously generated tokens in the decoration. The objective function to be maximised can this be written as:

$$J(\theta|S = s) = \prod_{\text{decoration points}} \left\{ P(X_1|S = s, \theta) \times \prod_{i=2}^T P(X_i = x_i|X_{i-1} = x_{i-1}, \dots, X_1 = x_1, S = s, \theta) \right\} \quad (1)$$

Here, θ represents the network parameters to be determined, $X_i, i = 1, \dots, T$ are the random variables corresponding to the tokens while the x_i are the observed (or in this case previously generated) tokens. Analogously, S and s refer respectively to the random variable corresponding to the input scaffold and the specific scaffold itself. In this work, the scaffold is given *a priori* and the distribution is therefore deterministic. Finally, T is another random variable determining the length of the decoration SMILES string. In practice, we do not sample its distribution. Instead, the process ends when the ‘END’ token is sampled.

The implementation and training details are described in the supporting information.

Focusing the prior via reinforcement learning

Motivation

Due to the vastness of chemical space, it is typically not sufficient to be able to produce drug-like molecules; indeed, depending on the specific design objective, exploration of a narrower chemical subspace is many times desirable, especially in lead optimization³⁹. Specific focusing is thus a crucial step in developing a generative model capable of proposing compounds useful in a context like lead optimization. To achieve this, the parameters of the pretrained prior network need to be modified to target a narrower chemical subspace. At the same time, it has been observed that deviating too far from the prior can have catastrophic consequences where the model loses its knowledge of valid SMILES syntax^{13,39}.

In order to focus the model, a RL agent is initialised as a network with weights and architecture identical to those of the pretrained prior. To define the task, a reward function is constructed to guide the agent’s learning, taking SMILES strings as input and returning scores in the range [0, 1]. The function rewards compounds with desirable properties, promotes varied output through diversity filters and specifies desired reactions to be used via reaction filters. Then, standard policy iteration RL is applied: In successive iterations, the agent proposes decorations for the scaffold and updates its parameters in a

gradient ascent fashion based on the rewards these decorations receive. During the training, all syntactically valid compounds (SMILES strings) with a score exceeding a user defined threshold are stored in the scaffold memory and made available to the user at the end of the run. A successful run results in a large and diverse scaffold memory since the model produces new relevant output at each step during the run. In an optimal scenario, the scaffold memory increases linearly with the number of steps, with the gradient corresponding the batch size. This motivates the following definition of a yield metric used to evaluate the degree of success of the runs:

$$\text{yield} = \frac{|\text{Scaffold memory}|}{\text{Batch size} \times \text{Number of steps}} \quad (2)$$

The consideration of yield as opposed to the raw number of molecules produced is important since the produced numbers ultimately depend on the selected batch size. A model trained with a batch size of 32 returns twice as many compounds at each epoch as one with batch size 16. The important question, however, is how many of the 32 compounds are relevant and unique.

Mathematical Background

The starting point for a mathematical description of the RL procedure is defining a state space S_t and the corresponding action space $A_t(s_t)$ as well as rewards $r_t := R(a_t)$ for all $s_t \in S_t, a_t \in A_t$. In the context of molecule decoration, an action is a proposed decoration (or decorations) for the scaffold while the state contains information about all previously proposed decorations and the rewards assigned to these, i.e. $s_t = \sum_{\tau=1}^{t-1} r_\tau$. Note the reward function R is fixed throughout the training.

At each step, the RL agent randomly samples an action (i.e. proposes a decoration) according to its policy π_θ . The aim is to find the value of the parameters θ leading to an optimal policy π_{θ^*} maximising the expected cumulative rewards across the whole run. In other words:

$$\theta^* = \operatorname{argmax}_{\theta} \sum_{t=0}^T \mathbb{E}_{A \sim \pi_{\theta}} (R(A) | S_t = s_t) \quad (3)$$

The expected value is maximised at each time step in a greedy manner. The RL objective function at each step can therefore be written as:

$$J(\theta) = \mathbb{E}_{A \sim \pi_{\theta}} (R(A) | S_t, \theta), \quad (4)$$

where the expectation is taken over the distribution of the actions.

Gradient ascent methods are typically used to maximise the objective. Exploiting the fact that $\nabla \log f(x) = \frac{\nabla f(x)}{f(x)}$, the gradient of eq 4 at step $t + 1$ can be written as:

$$\nabla_{\theta} J(\theta) = \sum_{a \in A_{t+1}} R(a) \nabla_{\theta} \log \pi(A = a | S_t = s_t, \theta). \quad (5)$$

Equation 5 is the basis of many popular RL algorithms such as REINFORCE⁴⁰. If the goal is to maximise cumulative rewards across N training epochs, it suffices to add an extra summation over all the timesteps, which results in a similar expression – the key feature of which is the fact that it is sufficient to compute the gradients of the log likelihoods to obtain a gradient ascent update step.

Without further regularisation or adjustments, these methods are known to suffer from high variance and instability⁴¹. In the case of molecular generation, however, the aim is to produce a large number of varied, interesting molecules²⁶. This means that a certain level of variance is desirable to promote exploration of the chemical space and prevent mode collapse towards a single, high scoring solution. Our experiments show that with an appropriate choice of the reward function, high variance does not hinder the models from producing relevant output.

Policy Iteration Rewards

A crucial requirement for a successful set-up of a RL run is a good definition of the reward. In our case, this has to guide the agent in the right direction to solve the specific practical task and promote diversity. Similar to Blaschke *et al.*¹⁵, we investigate rewards assembled from a combination of two elements: A scoring function $S(a) \in [0, 1]$ quantifying how well the proposed compound solves the task and prior

likelihood $\pi_{\theta_{prior}}(a) = \pi(a, \theta_{prior})$. Since the agent and prior share architecture, their likelihood functions differ only in the values of the parameters θ .

The Scoring Function

The $S(a)$ itself is composed of multiple weighed elements which are summed or multiplied; the final score is then normalised to lie in the unit interval $[0, 1]$. A range of components is supported, from molecular descriptors such as molecular weight, topological polar surface area (TPSA), pretrained predictive models, docking⁴² and ROCS similarity⁴³. As mentioned previously, diversity and reaction filters may be imposed to further restrict the space of relevant output and promote diversity.

Diversity filter works by penalising the model for producing an identical compound multiple times in a single batch. This is beneficial in preventing the agent from repeatedly proposing the same, high scoring compound multiple times, which can lead to mode collapse⁴⁴. A well selected diversity filter therefore balances exploration and exploitation of the chemical space.

Finally, reaction filters are a tool giving the user greater control over the generated compounds. Two types of reaction filters are implemented: general filter determining what reactions should occur to decorate the scaffold, and a selective filter assigning the specific reactions to the individual attachment points. This requires chemical understanding of the nature of the problem to avoid the situation where a non-feasible reaction is required for a given attachment point; it is nevertheless a novel and efficient way of generating libraries of similar drug like compounds which are readily synthesisable.

Different Reward Strategies

The motivation for the use of the prior likelihood in the reward function is identical to the one of Olivecrona *et al.*¹³. The pretraining ensures that the model is capable of generating valid SMILES of drug-like molecules. This serves as an anchor; it is desirable to discourage the agent from deviating too far from its prior state since a strong focus on maximising the score alone can lead to either a mode collapse or to a loss of the generative ability altogether. Once the agent moves to a parameter space which does not lead to valid SMILES syntax, it does not receive any rewards at all and cannot continue learning through gradient ascent.

Based on the discussion above, we follow previous work in defining the augmented log likelihood $\log \pi_A(a) = \log \pi_{\theta_{prior}}(a) + \sigma S(a)$. Here, σ is a constant hyperparameter scaling the output in the same range. We note that log likelihood is a monotonic increasing function taking values in $(-\infty, 0)$, which means that the reward is a monotonic increasing function in $(-\infty, \sigma)$. In experimental setups, this likelihood is shown to serve well as a directional guide, leading the agent to more focused and interesting chemical spaces. The intuitive rationale for this is that the augmented likelihood balances the prior anchor with the task-specific objective.

Finally, four different RL learning strategies are proposed based on four different reward functions:

1. $R(a) = S(a)$. This method, henceforth referred to as **MASCOF (Maximise Scoring Function)**, is a simple implementation of the standard REINFORCE algorithm where the scoring function directly serves as the reward⁴⁰. This standard approach to solving a RL problem by maximising the scoring function without anchoring it to the prior is a natural first step and can be seen as a baseline for the other methods. Our experiments however demonstrate that the RL agent struggles to remain in the valid chemical space without the anchor. Similar observations have been made in the past, typically arguing that the initial sparseness of rewards leads to the model struggling to begin learning⁴⁴.
2. $R(a) = \log \pi_A(a)$. Since the augmented likelihood attempts to balance the prior likelihood and the scoring function, it can be seen as a reward itself. We call this method **MAULI (Maximise Augmented Likelihood)**.
3. $R(a) = \log \pi_A(a) - \log \pi_{\theta}(a)$. This approach, dubbed **DAP (Difference between Augmented and Posterior)**, can be shown to be equivalent to the strategy introduced by¹³. Their work frames the RL slightly differently, focusing on loss minimisation of the square loss between the augmented and posterior log likelihoods: $\mathcal{L}(\theta) = (\log \pi_A(a) - \log \pi_{\theta}(a))^2$. While not a standard policy iteration approach, it does perform well in focusing the agent. For a full derivation of the equivalence of these two approaches, we refer the interested reader to the appendix.

4. $R(a) = -(\log \pi_A(a) - \log \pi_\theta(a))^2$. Noting that the rationale behind the DAP strategy is minimization of the difference between the two likelihoods, the fact that the likelihoods are unbounded means that with the formulation in 3., the reward may in theory keep increasing infinitely as the posterior probability approaches zero. In practice, this is rarely observed. For mathematical rigour, however, we define a final strategy called **SDAP** (Squared **D**ifference between **A**ugmented and **P**osterior). The negative of the squared loss is used directly as a reward function here, meaning that the agent is always encouraged to approach the augmented likelihood; maximizing the reward is equivalent to minimizing the square loss.

Experiments

Lib-INVENT represents a novel approach to drug discovery through scaffold decoration based on specific chemical reactions. This approach was designed to improve the productivity in the DMTA cycle through proposing a library of compounds that can be synthesized through the same chemical reactions. Thus more compounds can be synthesized with the same effort in an DMTA cycle and accordingly each DMTA cycle will be more informative⁴⁵. We therefore introduce a range of experiments with the aim to demonstrate the potential of our proposed models to improve the productivity. Specifically, we focus on promoting diversity of output, generating molecules readily synthesizable by a given reaction and determining R-group substitutions for lead optimization projects.

The objectives of the experiments are the following:

- Determine the optimal learning strategy for the reinforcement learning loop.
- Demonstrate the ability to follow specified reactions to decorate a given scaffold and contrast this with a model trained on a dataset obtained using RECAP rules as opposed to reaction based slicing.
- Demonstrate the ability to decorate scaffolds with various numbers of attachment points.

A baseline objective for the experiments is the creation of novel ligands for the DRD2 target. Two sets of tasks, based on two approaches to steer the model towards the desirable chemical space, have been

executed. In the first set of experiments, a QSAR predictive model for the activity of the generated compounds is used as a component of the scoring function. This model is subsequently replaced by a 3D shape and pharmacophore similarity ROCS scoring component to promote 3D similarity of the output to haloperidol, a known DRD2 active ligand. The details of the implementations can be found in our public repository. In all the experiments, a diversity filter is further added to the scoring function to promote variation in the output, along with custom alerts preventing the agent from proposing compounds with too large rings and non-drug-like groups.

Figure 3 displays the testing scaffold. The choice is motivated by it being a good starting scaffold for generating DRD2 actives. Further, it has two attachment points, which is common in real world applications. While we demonstrate the ability of the Library Design decorator to work with scaffolds with up to four attachment points, library synthesis is most commonly executed on fewer attachment as this gives a better balance between the flexibility and complexity of the library production step.

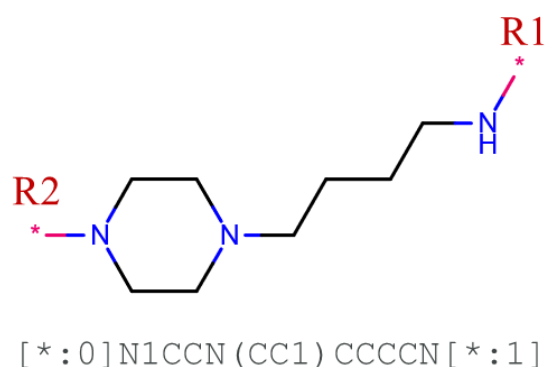


Figure 3: The testing scaffold. We note that in SMILES syntax, the decoration points are labelled by [*:0] and [*:1], which correspond to R2 and R1, respectively, in the molecular graph.

Evaluation metrics

The complexity of the task of molecule generation means that the choice of a metric for model evaluation needs to be considered carefully to ensure that all relevant issues are addressed. For library generation, it is desirable to produce libraries based on user defined criteria. Within these criteria, however, the libraries still differ in size, diversity and scores achieved according to the scoring function. We have previously defined the yield metric which helps evaluating what fraction of the generated ideas are

scored above a given threshold. This alone is nevertheless not sufficient to give a fair comparison of the libraries.

We address the question of diversity of the output via two approaches as appropriate in the given scenario. To determine the effect of a change in the scoring function, the overlap between the outputs is evaluated. This would show whether the methods produce significantly different results. When testing the effect of specific reaction filters, it may be more interesting to analyse the variation in the chemical properties of the proposed decorations for each attachment point. This smaller-scale view offers a more fine grained picture of the level of control the user has over the design of their specific library.

Results

Comparison of the learning strategies

For each of the four learning strategies, two experiments are set up to contrast their abilities of proposing molecules according to a given set of criteria. In the first experiment, only a QSAR predictive property is used. The motivation for this experiment is to benchmark the abilities of the models to generate compounds when unconstrained by chemical reactions. The results of the experiments are displayed in Table 2, which gives an overview of the average results over three individual runs. For a detailed breakdown over the runs, refer to the supporting information.

Table 2: Comparison of the four learning strategies for a QSAR model with no reaction filters. The uncertainty boundaries correspond to the largest deviation from the mean observed over the three runs. We note these are very low, showing a strong consistency between the trials. The Yield metric is calculated as previously defined in eq 2.

	Number of compounds found	Yield	Average mean score in scaffold memory
DAP	10510 ±69	0,821±0,005	0,722±0,005
MAULI	8573±271	0,670±0,021	0,658±0,015
MASCOF	4432±50	0,346±0,004	0,657±0,022
SDAP	4755±153	0,372±0,012	0,695±0,011

In a second experiment, a selective reaction filter is added to the scoring function. Attachment point R1 should be decorated using amide coupling while the second attachment follows the Buchwald reaction. The exact implementation and SMIRKS definitions of the corresponding equations can be found in our public repository. The results of the experiment are shown in Table 3.

The numerical results show that, in agreement with past observations¹³, the DAP learning strategy is the most successful one on multiple counts. Firstly, the high average score of the compounds in the scaffold memory for both runs indicates the ability of this model to produce high scoring molecules consistently throughout the run. This is further supported by the size of the output and a correspondingly high yield: even when selective reaction filters are applied, over 80 % of the proposed compounds had a score higher than the threshold of 0.4 chosen as the condition for inclusion in the scaffold memory. Moreover, nearly 90 % of the scaffold memory compounds satisfy both of the reaction filters, which gives strong support for using this strategy in virtual chemical library creation.

Table 3: Results for the four learning strategies when a QSAR predictive model and a selective reaction filter are employed.

	Number of compounds found	Yield	Average mean score in output	Ratio of fully satisfied reaction filters
DAP	10454±192	0,817±0,015	0,729±0,008	0,892±0,032
MAULI	5179±518	0,405±0,012	0,564±0,009	0,230±0,027
MASCOF	2846±854	0,222±0,067	0,574±0,030	0,297±0,076
SDAP	4033±302	0,315±0,024	0,622±0,019	0,457±0,136

Finally, to understand the training of the four respective strategies, we plot the average scores achieved at each step. It is crucial to note that thanks to the pretraining of the prior, it is not expected to observe a steeply increasing training curve since the choice of the starting scaffold is task specific and typically leads to high scores from the first iteration. The comparison is nevertheless a useful aid in the comparison as it explains the process further.

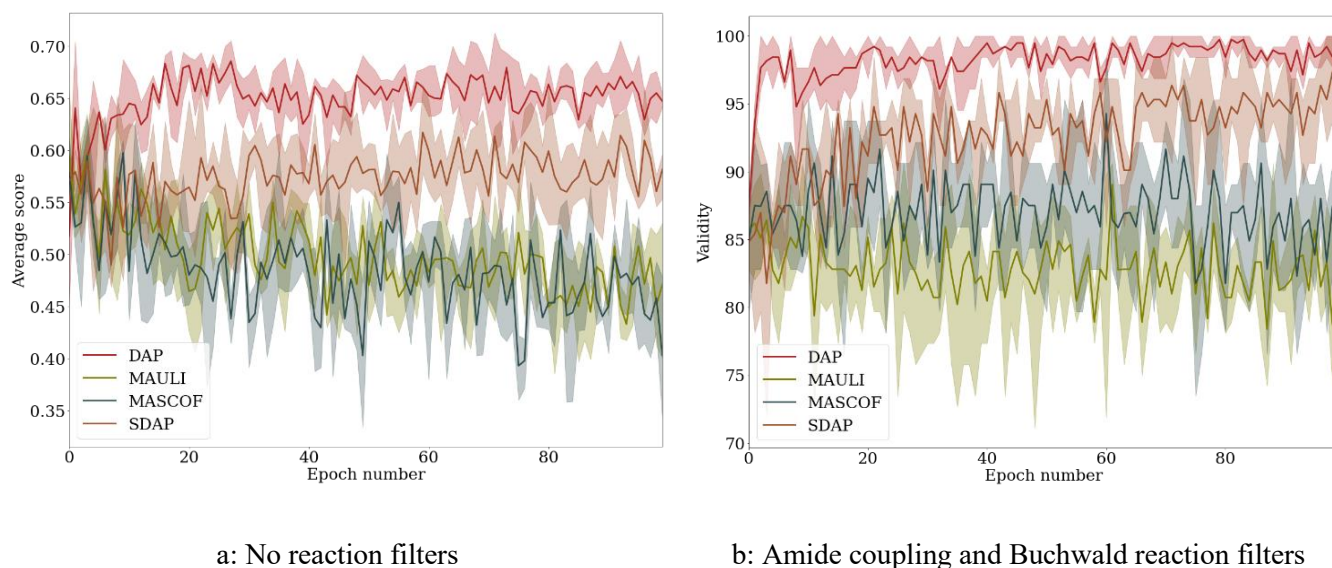


Figure 4: The average score across a generated batch of compounds per epoch for each of the running strategies

The evolution of the average scores across the runs is displayed in Figure 4. In both scenarios, the DAP strategy clearly outperforms the remaining optimisation methods, quickly increasing from the starting point and then plateauing at the highest level. When reaction filters are introduced, this dominance becomes even more significant. As displayed in Figure 5, the DAP strategy is the only strategy capable of rapidly adapting to this requirement and satisfying these filters. The SDAP strategy also demonstrates learning, but is much slower in adapting to the specific task. Both MASCOF and MAULI, on the other hand, decline slightly from the initial point as they struggle to retain the prior knowledge of the chemical space, which is demonstrated by the dropping validity. No evidence of learning to follow the required reactions is apparent.

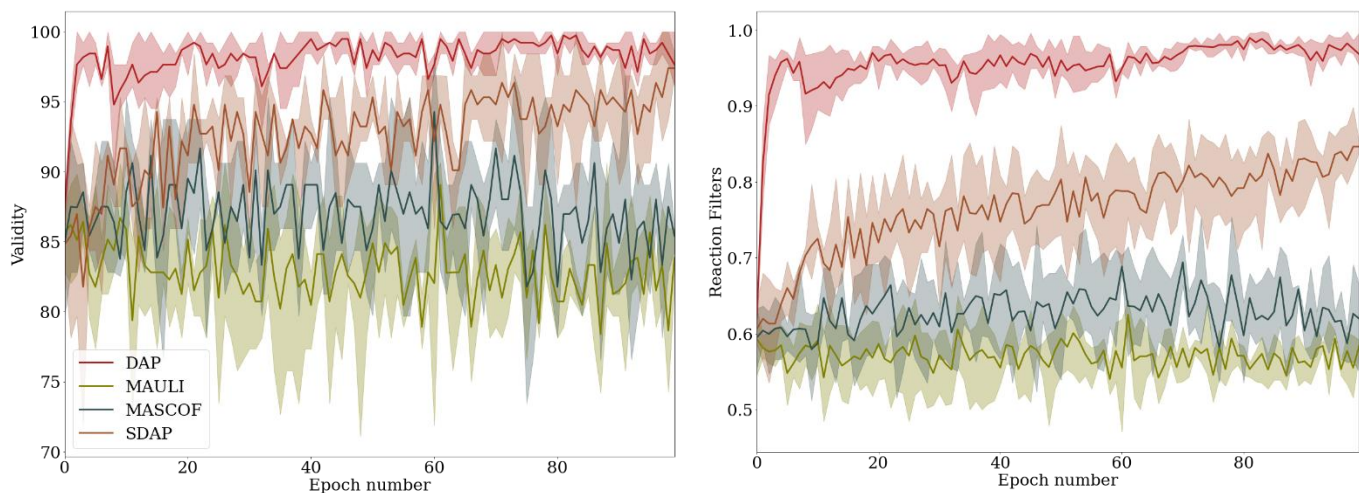


Figure 5: The validity of the output and reaction filter scores per learning strategy.

In both of the figures above, the shaded area corresponds to minimum and maximum values achieved over the three runs while the solid lines represent the mean. Besides the expected stochasticity arising from the randomness in the optimisation procedure, the plots indicate that the general behaviour of the strategies is consistent across runs. This observation is in agreement with the previous analysis of numerical data. In the subsequent experiments, we therefore restrict all in depth analysis to a single run per model only as the stochasticity does not significantly affect the output, justifying the low levels of variance between runs by numerical tables. Moreover, since the analysis above shows a clear dominance of the DAP learning strategy, this is our method of choice in all the subsequent experiments.

Comparison of slicing strategies

Reaction based slicing used to pre-process the dataset is one of the key novel contribution of this work. We therefore design a second set of experiments aimed at evaluating the effect of pretraining on data sliced according to chemical reactions as opposed to RECAP rules when tackling reaction filters. To this end, we use the model of Arús-Pous *et al.* as an alternative to benchmark against¹⁷. This model provides a fair comparison since it’s architecture and training procedure are exactly equivalent to our Lib-INVENT prior, with the crucial difference in data preparation.

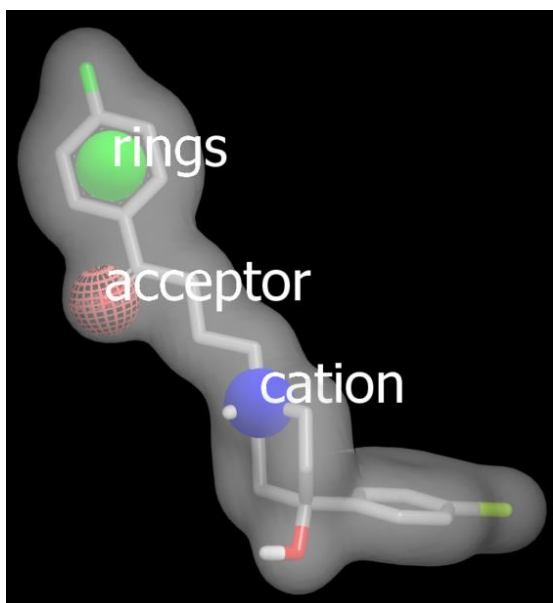


Figure 6: The ROCS shape and pharmacophore query definition for haloperidol

Two experiments are conducted with these two priors. In both, the same reaction filters as before are imposed, decorating attachment point 1 by amide coupling and attachment point 2 through the Buchwald reaction. The difference lies in the scoring function component, which remains the QSAR predictive model in the first experiment and is replaced by ROCS 3D similarity scoring in the second task. The purpose of this change is to uncouple the effect of the scoring function from the reaction filter and evaluate the effect of the pre-processing method as accurately as possible. The definition of the shape and pharmacophore ROCS query based on haloperidol is displayed in Figure 6.

Table 4: Comparison of reaction based slicing and RECAP slicing rules.

Pre-processing method	Model	Number of compounds found	Yield	Average mean score in scaffold memory	Ratio of fully satisfied reaction filters
Reaction	QSAR	10454	0,817	0,729	0,892
Based Slicing	ROCS	10326	0,807	0,597	0,890
RECAP	QSAR	8388	0,655	0,506	0,154
Slicing Rules	ROCS	8339	0,651	0,462	0,000

Numerical comparison of the experiments is displayed in Table 4. The key difference in the results is the ratio of high scoring molecules capable of satisfying the imposed reaction filters. While the model trained on data sliced according to reaction rules consistently scores very highly and therefore produces

libraries synthesisable via these two reactions, the model trained on data pre-processed using RECAP rules struggles to fulfil these criteria. With the exception of one run, the model fails to learn to follow the reaction routes. This gives clear evidence for the positive effect of this novel data slicing method for applications involving specific chemical reactions.

It can be further noted that the ROCS task seems to be more difficult for the models to learn. Interestingly, both the yield and the ratios of compounds satisfying the reaction filters are not significantly changed by the change between QSAR and ROCS scoring components; the difference lies in the average scores achieved by compounds in the scaffold memory. As the training plots in Figure 9 show, this is due to the fact that the scores start relatively low and gradually increase throughout the runs as the agents learn to satisfy the ROCS input.

To understand the diversity of the compounds proposed by agents trained with these two different scoring function components, we further contrast molecular properties of the decorations proposed by the respective methods when trained on a dataset obtained using reaction based slicing. Figure 7 demonstrates that the change in a scoring function component guiding the training affects the proposed decorations. On the example of attachment point 2, we see that while the groups proposed by an agent trained using ROCS are generally lighter, they tend to contain more rings and have more hydrogen bond acceptors. In some cases, we can further note that the reaction filter have not been satisfied for a share of the output – for example in the cases where the Buchwald reaction fails to propose a compound containing an aromatic ring. This demonstrates the need for a careful consideration of the scoring function design along with the reaction filters to achieve optimal results for a given task. Sample compounds proposed by each of the methods are plotted in Figure 8 for comparison. We can observe

the formation of the amide bonds as required by the reaction filters as well as the previously noted tendency of the ROCS guided model to propose decorations with multiple rings.

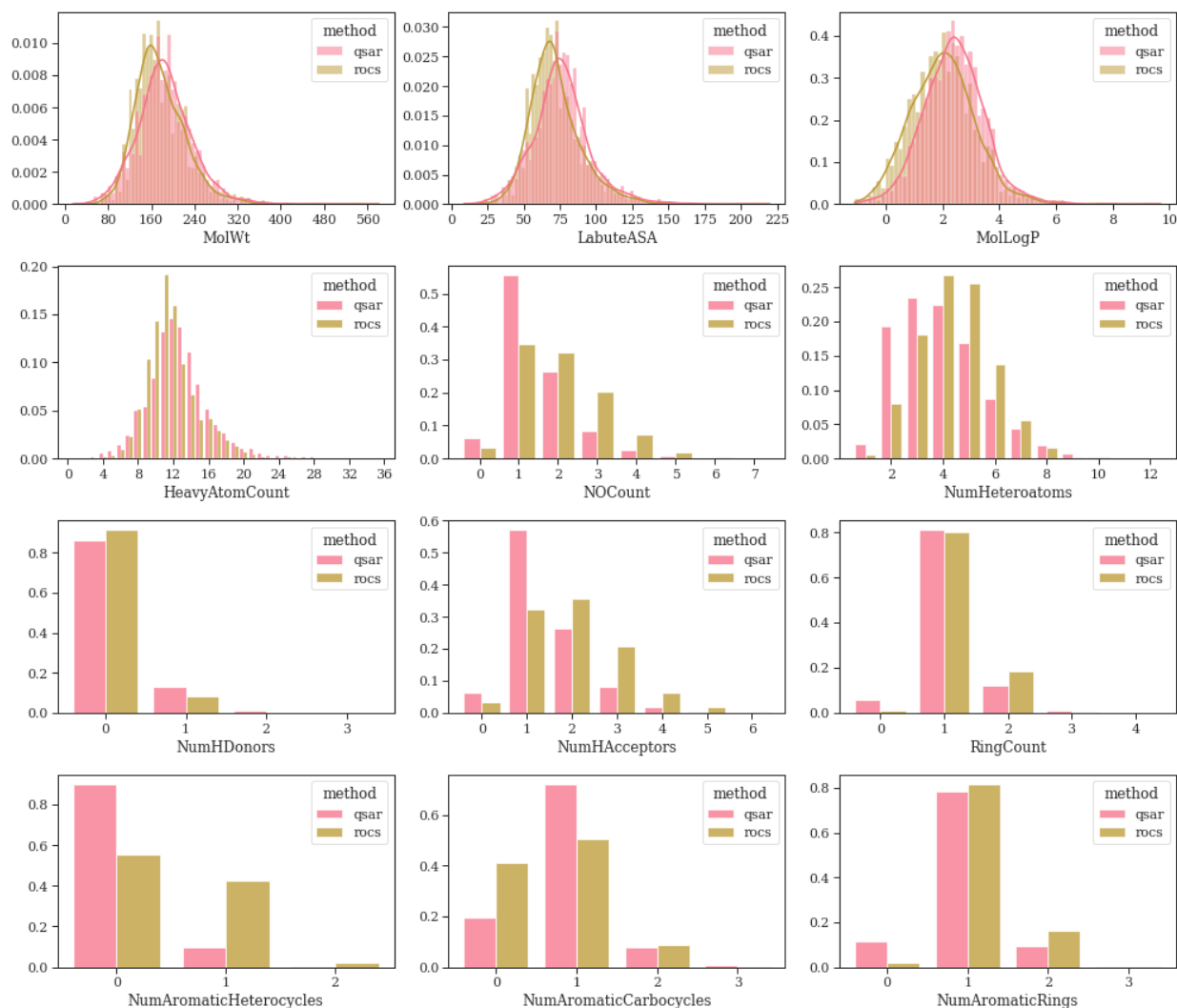


Figure 7: Example molecular properties of decorations for attachment point 2 when the Buchwald reaction filter is imposed.

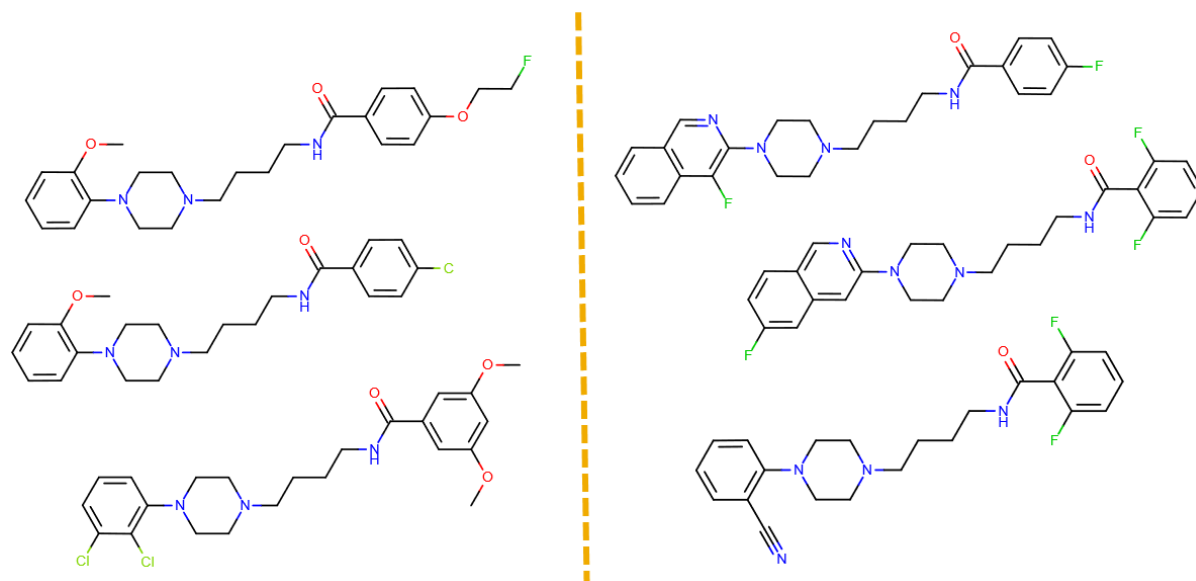


Figure 8: High scoring compounds proposed by a model guided by a QSAR predictive property (left) and by a ROCS scoring component (right). All of these compounds satisfy both the amide coupling and Buchwald reaction filters.

As mentioned previously, when training the model on a dataset obtained using RECAP slicing rules, we observe one successful run of the model optimising QSAR. This likely reason for this success is that the model has learnt to satisfy the reaction filters after randomly producing a compound which scored high on them. The experiments demonstrate that while the reaction based prior satisfies reaction filters systematically and consistently, successful runs for the RECAP based prior occur with a lower probability. This is further illustrated in Figure 9 and gives a compelling argument for the use of reaction based slicing in data pre-processing.

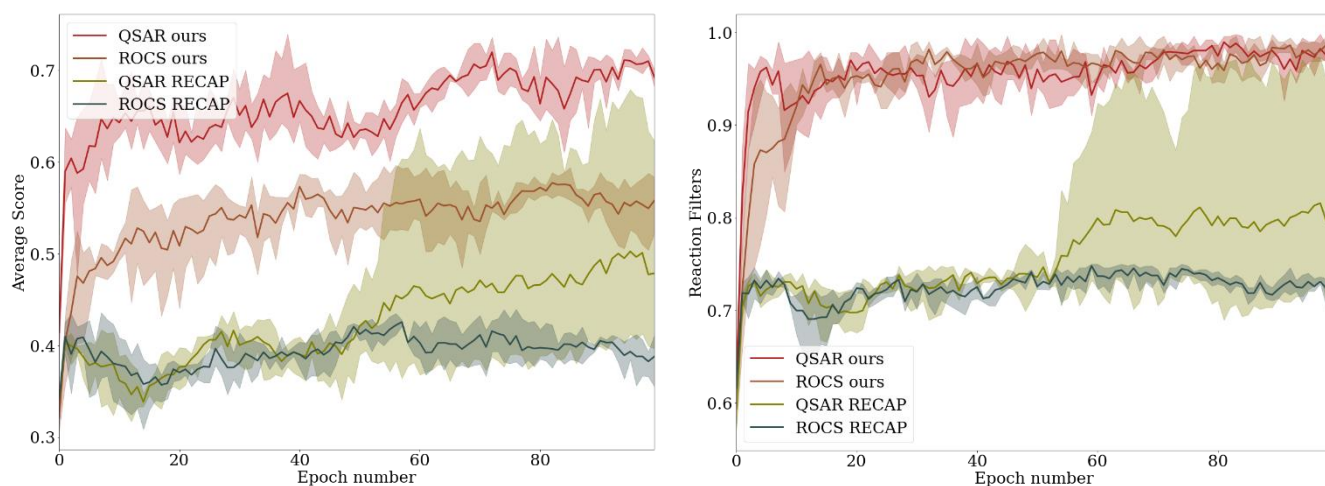


Figure 9: Comparison of the two learning strategies.

Following specific reactions

Table 5: Comparison of varying reaction filters for a QSAR and ROCS model.

Pre-processing method	Model	Number of compounds found	Yield	Average mean score in scaffold memory	Ratio of fully satisfied reaction filters
QSAR model	Buchwald-Amide	10454	0,817	0,734	0,892
	Buchwald-Sulphonamide	10083	0,788	0,688	0,847
	S _N Ar--Amide	9809	0,766	0,585	0,359
	S _N Ar--Sulphonamide	9228	0,721	0,641	0,577
ROCS model	Buchwald-Amide	10326	0,807	0,596	0,890
	Buchwald-Sulphonamide	10207	0,797	0,592	0,871
	S _N Ar--Amide	9837	0,768	0,545	0,551
	S _N Ar--Sulphonamide	9560	0,747	0,552	0,541

Using the optimal learning strategy and the prior model pretrained on data sliced using reaction rules, we propose a new set of experiments to demonstrate the effect of selective reaction filters on the produced libraries. For each of the attachment points, we select a relevant plausible chemical reaction that can serve for introducing desirable moieties. Specifically, sulphonamide coupling is used as an alternative to amide coupling for attachment point 1 and the Buchwald reaction of attachment point 2 may be replaced by a nucleophilic heteroaromatic substitution (S_NAr). We experiment with each of the

four possible combinations of these reaction filters to demonstrate the effect of these filters on the produced compounds. For illustrative purposes, a high scoring compound discovered for each of these combinations of reaction filters by a QSAR-guided predictive model is plotted in Figure 10. The reaction filters have a clear effect on the proposed molecules, enforcing the formations of appropriate bonds.

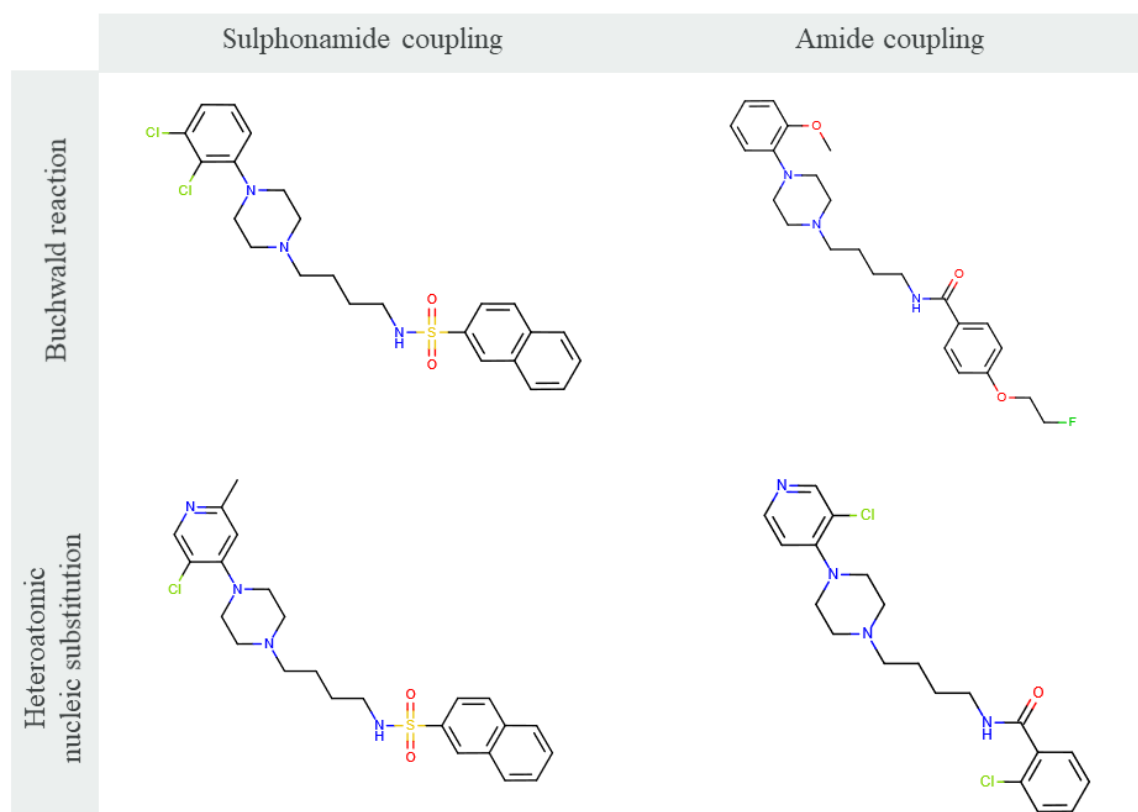


Figure 10: Comparison of compounds proposed by models optimising for various reaction filters.

All of the sets of reaction filters are applied to the two different setups of the scoring function as in the previous experiments – using either the QSAR predictive model or ROCS similarity component to direct the model towards the target chemical subspace of compounds active on the DRD2 dataset. The reason for using different scoring functions in this experiment is to demonstrate the effect the scoring function has on the output and decouple this with the effect of reaction filters. The numerical results of these experiments are displayed in Table 5.

The performance of the ROCS model when reaction filters are exchanged is more stable than for QSAR, showing similar patterns and an ability to learn to follow different reaction routes. This is presumably caused by a lower degree of inductive bias built into the model through this scoring component. The consistently lower average scores in the scaffold memory can be attributed to the greater difficulty to learn this component in general; it is more difficult to score highly the structural similarity (and match) requirements of a ROCS component. This does nevertheless not mean that the model performs badly; on the contrary, the high yields show that it is an effective guide towards a desirable area of the chemical space. Moreover, the highest achievable scores of the ROCS component are typically lower than for a QSAR model and commonly lie around the value 0.8.

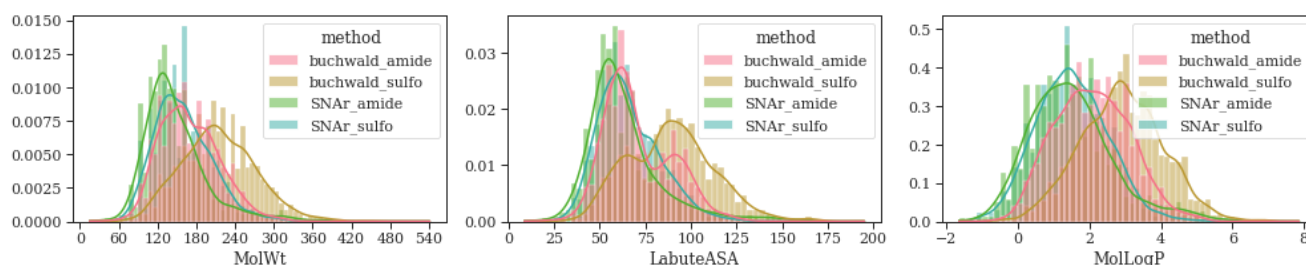


Figure 11: Continuous molecular descriptors of decorations of attachment point 1 for each of the four reaction filters applied, trained using a QSAR model. This attachment point is decorated by either amide or sulphonamide coupling.

To quantify the differences in the properties of the decorations arising from various reaction filters, we examine the distributions of key molecular properties of the proposed functional groups for each attachment point based on the applied reaction filters. A selection of molecular descriptors of decorations generated for attachment point 1, decorated via amide or sulphonamide coupling, is

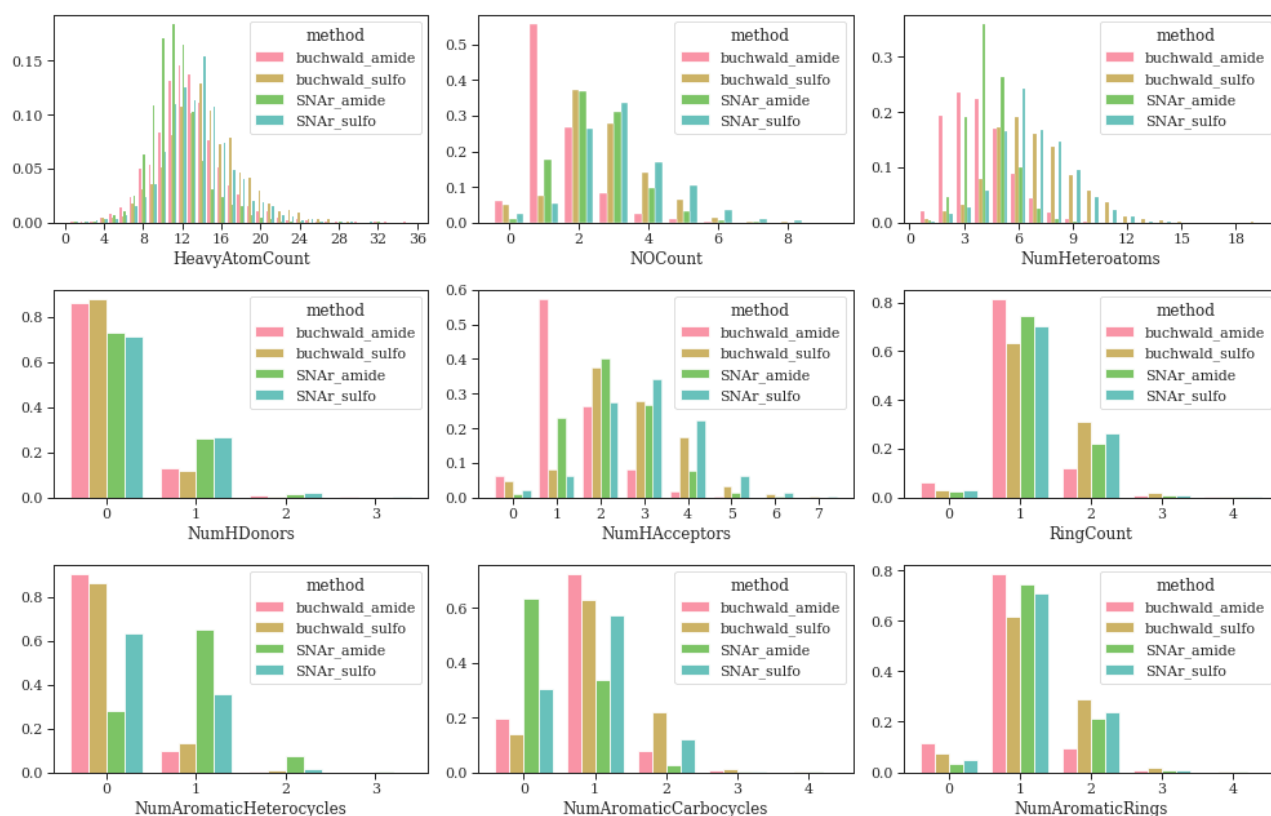


Figure 12: Discrete molecular descriptors of decorations of attachment point 2 for each of the four reaction filters applied. Note that this attachment point is decorated using either via the Buchwald reaction of the S_NAr substitution; the differences observed in this plot therefore primarily arise as a result of this reaction filter.

displayed in Figure 11. A significant increase in the weight of the proposed decorations, caused by the presence of more heavy atoms, occurs when the sulphonamide coupling is introduced. Figure 12 further shows selected discrete properties of the decorations proposed for the second attachment point. In both plots, variation in the distributions can be observed across all four combinations of reaction filters; the effect of the reaction filters imposed for the given attachment point is nevertheless clearly notable. This is to be expected since the agent receives rewards based on the entire compounds proposed but each attachment point is strongly influenced by the prescribed reaction.

As a final point of comparison of the reaction filters, Figure 13 displays the distribution of selected molecular properties for each of the two attachment points using the original reaction filter composed of amide coupling and the Buchwald reaction. In general, amide coupling produces somewhat smaller and lighter decorations. The distributions moreover tend to be less peaked and centred around the mode,

which is to be expected for this reaction as it is more general. Once again, we further note that not all proposed compounds satisfy the Buchwald reaction filter since decorations missing an aromatic ring are proposed for attachment point 2.

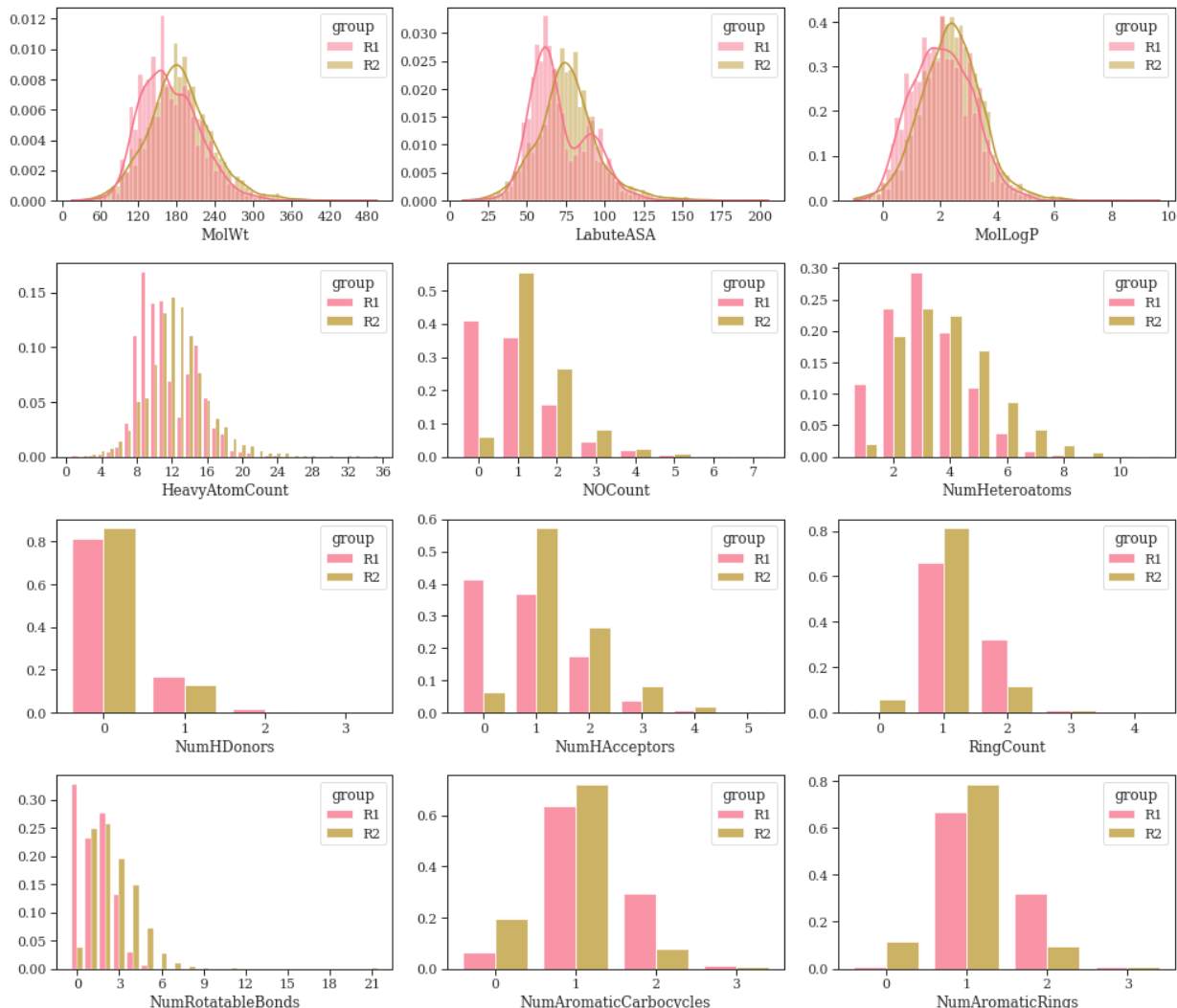


Figure 13: Comparison of the properties of the decorations proposed for each attachment point. The first attachment is decorated using amide coupling, the second via the Buchwald reaction.

Scaffolds with varied numbers of attachment points

So far, all experiments focused on a two-attachment point scaffold. To give a fair picture of the decorator’s abilities, we additionally introduce tasks working with scaffolds containing one to four decoration points. Since the purpose of this section is primarily proof of concept, we restrict our attention to simple experiments aiming to force the model to start growing large enough decorations to satisfy molecular weight requirements. Reaction filter is not implemented here for simplicity. The experiments

nevertheless demonstrate that the decorator is capable of working with these scaffolds to produce unique and valid compounds.

For the purpose of these simple experiments, we use two scaffolds from the DRD2 dataset with one and three attachment points. In both cases, the weight requirement on the final compound is for it to lie between 450 and 650. These values have been chosen to force the original scaffolds to grow significantly without leaving the domain of chemically reasonable compounds in the output since there are no other constraints to guide the model. The scaffolds are displayed in Figure 14.

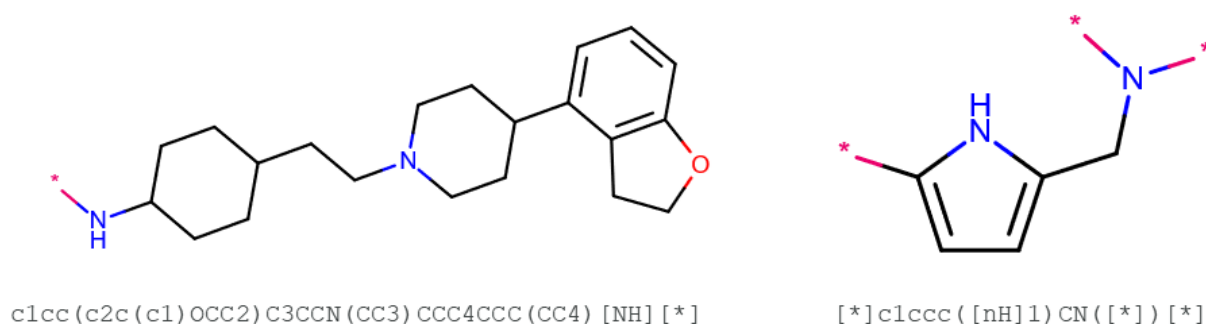


Figure 14: The scaffolds with one and three attachment points used in the final experiment.

As demonstrated in the plots below, the model does not struggle with any of these tasks, rapidly adjusting to the requirement and starting to generate compounds in the appropriate molecular weight range. Similarly, the validity of the proposed output is consistently over 90 %. These experiments clearly show that the use cases of the decorator model include working with scaffolds of varying numbers of attachment points.

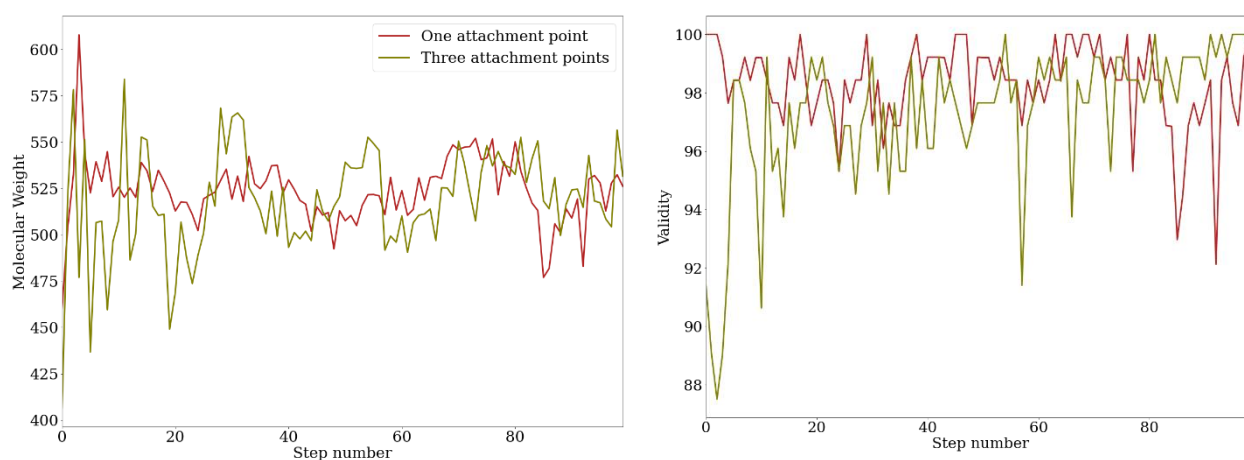


Figure 15: Satisfying weight requirements for varying number of attachment points.

Discussions

We have designed and executed a range of experiments to establish the abilities of the newly proposed Lib-INVENT model. Most importantly, the results clearly demonstrate the model’s superiority in learning to follow specific chemical reactions, which is achieved by the introduction of a novel compound slicing strategy. The decorator model has proven to be capable of rapidly designing libraries of molecules synthesisable from a given scaffold by following a set of reactions as defined by the user, giving the user fine control over the output and making the model widely and readily applicable in a multitude of scenarios. This expands the capabilities of the REINVENT family of generative models from string and graph based compound design to library design.

In contrast with prior work on scaffold decoration, which trained the models on data obtained using RECAP rules for slicing^{17,18}, Lib-INVENT has been trained on compounds sliced according to chemical reactions. This crucially affects the speed and reliability with which the model adjusts to reaction filter restrictions and is a significant step towards the automation of the symbiosis between *in silico* and *in vitro* library generation. The design of the reinforcement learning loop further introduces a rapid way to focus the model to a desirable part of the chemical space. As the experiments demonstrate, the learning is instantaneous and results in the design of varied and focused chemical libraries.

The first task was to determine an optimal learning strategy for setting up the reinforcement learning rewards. Four different strategies have been proposed based on arguments discussed in the literature. While not immediately intuitive, the DAP learning strategy has proven to be the most successful one. The motivation for this reward setup is a “carrot on a stick” scenario. A combination of the prior likelihood and the scoring function is used to guide the agent towards a desirable subspace of the chemical space while ensuring that underlying chemical syntax is not forgotten. Two of the remaining strategies, on the other hand, attempted to maximise the score or a sum of the score and prior likelihood directly. Despite appearing more natural at a glance, this approach does not work as well since the models struggle to retain the ability to propose valid molecules as they start focusing on the score too

much. A possible rationale for this is the notoriously high variance typically observed for policy iteration RL; while we note that the generative model requires this variance to explore the chemical space, too much variation combined with a lack of anchor to the prior knowledge is detrimental to the performance. The final method explored in the paper minimises the square of the loss used for the optimal DAP strategy. This is more mathematically sound as the reward is bounded but does not appear beneficial in practice since the edge scenarios where unboundedness of the DAP reward could be an issue rarely arise. We therefore confirm the observations of Olivecrona *et al.* in selecting the DAP strategy as the method of choice¹³.

Two different scoring components have been used to guide the model to propose new ligands for the DRD2 receptor. A simple QSAR property prediction model has the advantage of a faster execution and overall higher score but its stronger inductive bias restricts the model to a narrower domain⁴⁶. As a result, a QSAR based model strongly favours certain decorations and therefore struggles to fulfil some reaction routes incompatible with these functional groups. In the second use case scenario, ROCS similarity measure was used to demonstrate that various scoring function components may be used to guide the model to a desirable chemical space. A certain degree of experimentation or user intuition is often required to determine the optimal combination of guidance for the model and freedom to explore to obtain the best possible libraries as each of the scoring components introduces its own biases and benefits. The results nevertheless confirm that Lib-INVENT is a flexible tool admitting a wide range of inputs and able to return appropriate output.

An important note regarding the selective reaction filter is that the user is responsible for providing correct and valid reactions for correct attachment points in order to get a good result. While a range of reaction definitions is provided in the public repository, the reactions prescribed to a given attachment point have to be feasible. If an infeasible reaction is required, the model is not going to be able to fulfil this reaction filter and always receive a score of 0. Depending on the setup, this can lead to a failed run with a very small and irrelevant outputs as the low scores do not guide the agent in the right direction. It is therefore essential that the user is aware of this potential pitfall. A consistently low reaction filter score, plateauing at a value lower than one, is often an indicator of a wrong reaction requirement.

The observations from the experiments confirm the ability of Lib-INVENT to generate readily synthesisable virtual chemical libraries. The model offers a great level of flexibility, giving the user the option to determine not only the molecular properties of their output, but also the shape and chemical pathways to be followed in library synthesis. This functionality has a potential to bridge many of the current problems with incorporation of *in silico* design in drug discovery as the process to determine ways to produce the proposed molecules has been laborious in the past. Moreover, the existence of the pretrained prior along with the use of reinforcement learning enables rapid and efficient library generation by eliminating the need to use transfer learning to refocus on a new task. Besides the speed with which the model focuses on a new problem, reinforcement learning moreover offers the possibility to set up flexible objectives and combine various scoring components.

Conclusions

In order to achieve efficient and natural symbiosis between computation and traditional wet lab methods in drug discovery, it is essential to overcome a few prevailing bottlenecks. One of the key issues is the low efficiency in incorporating deep learning into the production pipeline caused by complicated lead synthesis and an unfocused output of generative models. The objective of this work has been to provide a method to start bridging this gap between *in silico* and *in vitro* drug design by developing a tool taking the needs of real life synthesis into consideration and increasing the productivity by reducing the number of DMTA cycles performed.

In this work, we have introduced a flexible generative model capable of proposing optimal decorations given a scaffold and a set of user-specified objectives. Thanks to a novel compound slicing method based on chemical reactions, these objectives can moreover include reaction filters. Lib-INVENT therefore enables rapid generation of focused virtual chemical libraries which can be used for lead optimisation and are readily synthesisable *in vitro*. Even when these filters are not specified, however, the output of the model benefits from high synthesisability.

To the best of our knowledge, our model is the first one to be capable of following specific reaction constraints in designing entire chemical libraries within which the diversity is narrowly focused to a

domain determined by the user. This makes the decorator readily applicable in a broad range of scenarios. The model is released in our public repository along with the corresponding code.

Acknowledgements

We would like to thank Panagiotis-Christos Kotsias for discussions about model benchmarking and validation.

Associated content

Supporting information

The document contains technical and implementation details such as hyperparameters used in training and the exact model vocabulary. A breakdown of results of the reruns of the experiments is provided as well as a more complete mathematical background of the project.

References

- (1) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug Discovery with Explainable Artificial Intelligence. *Nat. Mach. Intell.* **2020**, 2 (10), 573–584. <https://doi.org/10.1038/s42256-020-00236-4>.
- (2) Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R. K. Artificial Intelligence in Drug Discovery and Development. *Drug Discov. Today* **2021**, 26 (1), 80–93. <https://doi.org/10.1016/j.drudis.2020.10.010>.
- (3) Bush, J. T.; Pogany, P.; Pickett, S. D.; Barker, M.; Baxter, A.; Campos, S.; Cooper, A. W. J.; Hirst, D.; Inglis, G.; Nadin, A.; Patel, V. K.; Poole, D.; Pritchard, J.; Washio, Y.; White, G.; Green, D. V. S. A Turing Test for Molecular Generators. *J. Med. Chem* **2020**, 63, 11964–11971. <https://doi.org/10.1021/acs.jmedchem.0c01148>.
- (4) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in *De Novo* Molecular Design. *Mol. Inform.* **2018**, 37 (1–2), 1700123.

<https://doi.org/10.1002/minf.201700123>.

- (5) Xue, D.; Gong, Y.; Yang, Z.; Chuai, G.; Qu, S.; Shen, A.; Yu, J.; Liu, Q. Advances and Challenges in Deep Generative Models for de Novo Molecule Generation. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2019**, *9* (3), e1395. <https://doi.org/10.1002/wcms.1395>.
- (6) Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noé, F.; Clevert, D. A. Efficient Multi-Objective Molecular Optimization in a Continuous Latent Space. *Chem. Sci.* **2019**, *10* (34), 8016–8024. <https://doi.org/10.1039/c9sc01928f>.
- (7) Yang, Y.; Zheng, S.; Su, S.; Zhao, C.; Xu, J.; Chen, H. SyntaLinker: Automatic Fragment Linking with Deep Conditional Transformer Neural Networks. *Chem. Sci.* **2020**, *11* (31), 8312–8322. <https://doi.org/10.1039/d0sc03126g>.
- (8) Grebner, C.; Matter, H.; Plowright, A. T.; Hessler, G. Automated de Novo Design in Medicinal Chemistry: Which Types of Chemistry Does a Generative Neural Network Learn? *J. Med. Chem.* **2020**, *63* (16), 8809–8823. <https://doi.org/10.1021/acs.jmedchem.9b02044>.
- (9) Hughes, J. P.; Rees, S. S.; Kalindjian, S. B.; Philpott, K. L. Principles of Early Drug Discovery. *British Journal of Pharmacology*. Wiley-Blackwell March 2011, pp 1239–1249. <https://doi.org/10.1111/j.1476-5381.2010.01127.x>.
- (10) Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization. *Mol. Inform.* **2010**, *29* (5), 366–385. <https://doi.org/10.1002/minf.201000019>.
- (11) Maziarz, K.; Jackson-Flux, H.; Cameron, P.; Sirockin, F.; Schneider, N.; Stiefl, N.; Brockschmidt, M. Learning to Extend Molecular Scaffolds with Structural Motifs. *arXiv* **2021**. <https://doi.org/arXiv:2103.03864>.
- (12) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50* (3), 339–348. <https://doi.org/10.1021/ci900450m>.

- (13) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminform.* **2017**.
<https://doi.org/https://doi.org/10.1186/s13321-017-0235-x>.
- (14) 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>.
- (15) 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**. <https://doi.org/10.1021/acs.jcim.0c00915>.
- (16) Mercado, R.; Rastemo, T.; Lindelöf, E.; Klambauer, G.; Engkvist, O.; Chen, H.; Jannik Bjerrum, E. Graph Networks for Molecular Design. *Mach. Learn. Sci. Technol.* **2021**, 2 (2), 025023. <https://doi.org/10.1088/2632-2153/abcf91>.
- (17) Arús-Pous, J.; Patronov, A.; Bjerrum, E. J.; Tyrchan, C.; Raymond, J. L.; Chen, H.; Engkvist, O. SMILES-Based Deep Generative Scaffold Decorator for de-Novo Drug Design. *J. Cheminform.* **2020**, 12 (1), 38. <https://doi.org/10.1186/s13321-020-00441-8>.
- (18) Langevin, M.; Minoux, H.; Levesque, M.; Bianciotto, M. Scaffold-Constrained Molecular Generation. *J. Chem. Inf. Model.* **2020**, 12, acs.jcim.0c01015.
<https://doi.org/10.1021/acs.jcim.0c01015>.
- (19) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, 47 (D1), D930–D940. <https://doi.org/10.1093/nar/gky1075>.
- (20) Göller, A. H.; Kuhnke, L.; Montanari, F.; Bonin, A.; Schneckener, S.; ter Laak, A.; Wichard, J.; Lobell, M.; Hillisch, A. Bayer's in Silico ADMET Platform: A Journey of Machine Learning over the Past Two Decades. *Drug Discov. Today* **2020**, 25 (9), 1702–1709.

<https://doi.org/10.1016/j.drudis.2020.07.001>.

- (21) Göller, A. H.; Kuhnke, L.; Montanari, F.; Bonin, A.; Schneckener, S.; ter Laak, A.; Wichard, J.; Lobell, M.; Hillisch, A. Bayer's in Silico ADMET Platform: A Journey of Machine Learning over the Past Two Decades. *Drug Discovery Today*. Elsevier Ltd September 2020, pp 1702–1709. <https://doi.org/10.1016/j.drudis.2020.07.001>.
- (22) Li, Y.; Hu, J.; Wang, Y.; Zhou, J.; Zhang, L.; Liu, Z. DeepScaffold: A Comprehensive Tool for Scaffold-Based De Novo Drug Discovery Using Deep Learning. **2019**. <https://doi.org/10.1021/acs.jcim.9b00727>.
- (23) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. ChemInform Abstract: RECAP - Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *ChemInform* **2010**, 29 (36). <https://doi.org/10.1002/chin.199836303>.
- (24) Heikamp, K.; Zuccotto, F.; Kiczun, M.; Ray, P.; Gilbert, I. H. Exhaustive Sampling of the Fragment Space Associated to a Molecule Leading to the Generation of Conserved Fragments. *Chem. Biol. Drug Des.* **2018**, 91 (3), 655–667. <https://doi.org/10.1111/cbdd.13129>.
- (25) Bradshaw, J.; Kusner, M. J.; Paige, B.; Benevolentai, M. H. S. S.; Miguel Hernández-Lobato, J. Generating Molecules via Chemical Reactions. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)*.; 2019.
- (26) Horwood, J.; Noutahi, E. Molecular Design in Synthetically Accessible Chemical Space via Deep Reinforcement Learning. *ACS Omega* **2021**. <https://doi.org/10.1021/acsomega.0c04153>.
- (27) Kolen, J. F.; Kremer, S. C. A Field Guide to Dynamical Recurrent Networks. In *A Field Guide to Dynamical Recurrent Networks*; Kolen, J. F., Kremer, S. C., Eds.; Wiley-IEEE Press, 2010; pp 200–203. <https://doi.org/10.1109/9780470544037>.
- (28) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J. L.; Chen, H.; Engkvist, O. Randomized SMILES Strings Improve the Quality of Molecular

- Generative Models. *J. Cheminform.* **2019**, *11* (1), 71. <https://doi.org/10.1186/s13321-019-0393-0>.
- (29) Krishnan, S. R.; Bung, N.; Bulusu, G.; Roy, A. Accelerating *De Novo* Drug Design against Novel Proteins Using Deep Learning. *J. Chem. Inf. Model.* **2021**, *acs.jcim.0c01060*. <https://doi.org/10.1021/acs.jcim.0c01060>.
- (30) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today*. Elsevier Ltd June 1, 2018, pp 1241–1250. <https://doi.org/10.1016/j.drudis.2018.01.039>.
- (31) Bender, A.; Cortés-Ciriano, I. Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 1: Ways to Make an Impact, and Why We Are Not There Yet. *Drug Discovery Today*. Elsevier Ltd February 1, 2021, pp 511–524. <https://doi.org/10.1016/j.drudis.2020.12.009>.
- (32) 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.* **2020**, *13* (1), 26. <https://doi.org/10.26434/chemrxiv.12941744.v1>.
- (33) Rifaioğlu, A. S.; Nalbat, E.; Atalay, V.; Martin, M. J.; Cetin-Atalay, R.; Doğan, T. DEEPScreen: High Performance Drug-Target Interaction Prediction with Convolutional Neural Networks Using 2-D Structural Compound Representations. *Chem. Sci.* **2020**, *11* (9), 2531–2557. <https://doi.org/10.1039/c9sc03414e>.
- (34) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893. <https://doi.org/10.1021/jm9602928>.
- (35) 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* (1), 17. <https://doi.org/10.1186/s13321-017-0203-5>.

- (36) Bjerrum, E. J. Smiles Enumeration as Data Augmentation for Neural Network Modeling of Molecules. *arXiv*. 2017.
- (37) Bjerrum, E.; Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules* **2018**, *8* (4), 131. <https://doi.org/10.3390/biom8040131>.
- (38) Goyal, A.; Lamb, A.; Zhang, Y.; Zhang, S.; Courville, A.; Bengio, Y. Professor Forcing: A New Algorithm for Training Recurrent Networks. In *Advances in Neural Information Processing Systems*; Neural information processing systems foundation, 2016; pp 4608–4616.
- (39) Skinnider, M. A.; Stacey, R. G.; Wishart, D.; S.; Foster, L. J. Deep Generative Models Enable Navigation in Sparsely Populated Chemical Space. **2021**. <https://doi.org/10.26434/CHEMRXIV.13638347.V1>.
- (40) Williams, R. J. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.* **1992**, *8* (3–4), 229–256. <https://doi.org/10.1007/bf00992696>.
- (41) Wu, C.; Rajeswaran, A.; Duan, Y.; Kumar, V.; Bayen, A. M.; Kakade, S.; Mordatch, I.; Abbeel, P. Variance Reduction for Policy Gradient with Action-Dependent Factorized Baselines. *arXiv* **2018**. <https://doi.org/arXiv:1803.07246>.
- (42) Morris, G. M.; Lim-Wilby, M. Molecular Docking. *Methods Mol. Biol.* **2008**, *443*, 365–382. https://doi.org/10.1007/978-1-59745-177-2_19.
- (43) Kumar, A.; Zhang, K. Y. J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Front. Chem.* **2018**, *6*, 315. <https://doi.org/10.3389/fchem.2018.00315>.
- (44) Korshunova, M.; Huang, N.; Capuzzi, S.; Radchenko, D. S.; Savych, O.; Moroz, Y. S.; Wells, C.; Willson, T. M.; Tropsha, A.; Isayev, O. A Bag of Tricks for Automated De Novo Design of Molecules with the Desired Properties: Application to EGFR Inhibitor Discovery. *chemRxiv*

2021. <https://doi.org/10.26434/chemrxiv.14045072.v1>.

- (45) de Souza Neto, L. R.; Moreira-Filho, J. T.; Neves, B. J.; Maidana, R. L. B. R.; Guimarães, A. C. R.; Furnham, N.; Andrade, C. H.; Silva, F. P. In Silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Front. Chem.* **2020**, *8*, 93.
<https://doi.org/10.3389/fchem.2020.00093>.
- (46) Baxter, J. A Model of Inductive Bias Learning. *J. Artif. Intell. Res.* **2000**, *12*, 149.
<https://doi.org/10.1613/jair.731>.