How to Reward Your Drug Agent?

Andrea Karlova
Department of Computer Science
University College London
London, United Kingdom
akarlova@cs.ucl.ac.uk

Wim Dehaen
Department of Informatics and Chemistry
University of Chemistry and Technology
Prague, Czech Republic
dehaeni@vscht.cz

Andrei Penciu
April19 Discovery Ltd
London, UK
andrei@april19.ai

Abstract

Constructing novel molecules from scratch using deep generative models provides a valuable alternative to traditional virtual screening methods, which are limited to searching the already discovered chemicals. In particular, molecular optimisation combined with sampling guided by reinforcement learning seems like a promising path for discovering novel molecular designs and allows for domain-specific customisation of the desired solutions. The choice of a chemically relevant reward function and the exhaustive assessment of its properties remains a challenging task. We introduce the reward function which gives enough flexibility to quantify the biological activity with respect to a selected protein target, drug-likeness, synthesizability and incorporates the custom index of penalised physico-chemical properties. In order to customise the hyper-parameters influencing the RL agent performance, we propose the methodology that helps quantify the chemical relevance of the reward function by quantifying the chemical significance of the samples. We assess the performance of the reward function by docking the molecules with relevant protein targets and quantify the difference with the ground truth samples using Wasserstein distance.

1 Introduction

The chemical space of accessible and stable small molecules from which potential candidates for drug design can be selected consists of about 10^{60} compounds [Bohacek et al., 1996]. As the search is computationally intractable [Sanchez-Lengeling B, 2018], it is critical for the molecular discovery to generate a pool of candidate molecules that satisfy problem-specific properties. For early detection of bad leads, scores based on physico-chemical properties [Lipinski et al., 1997, Bickerton et al., 2012] and comparisons to already discovered topological structures provide a useful starting point and benchmark. Nevertheless, these approaches introduce bias and inefficiencies into the inevitable automation of the drug discovery process.

Molecular optimisation uses computational methods to construct compounds that can be sufficiently small and show bio-activity with protein targets of interest [Sanchez-Lengeling B, 2018]. The drug activity is deduced by its affinity to a target. Structure-based screening utilises in-silico modelling approaches that encompass structural information of the drug target. A classic modelling technique in this category is molecular docking the biologically active molecule (ligand) is scored based on the ligand conformation at the protein binding pocket. Optimal placement of the ligand is generated by optimising the conformation of the ligand to minimize the scoring function [Halperin et al., 2002].

NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications

Compared to other early-stage virtual screening methods, molecular docking is computationally expensive [McGaughey et al., 2007], which indicates that for a truly high-throughput virtual screening system, candidate molecules need to be smartly pre-selected so that only a tiny fraction of the potential chemicals is being assessed.

Directed generative models, such as VAEs [Kingma and Welling, 2014, Rezende et al., 2014] have been successfully used in inverse designs [Sanchez-Lengeling B, 2018] to generate novel chemical structures [Gómez-Bombarelli et al., 2018, 2016]. VAEs typically learn to sample the structured molecular data in the form of a SMILE string [Weininger, 1988], which describes the schematic graph topology of the molecule using the underlying grammar. The original pitfalls associated with this approach such as generating semantically correct SMILE string which lacks the chemical meaning has been addressed by various researches leading to improvements of reconstruction quality of the generated SMILE strings by learning the underlying grammar of the strings [Kusner et al., 2017, Jin et al., 2018]. Because VAE compresses the information from the data with the help of KL-divergence regulariser, we can expect that the learnt latent representations captures the similarities based on the underlying grammar rules used for the construction of the SMILE strings rather than on physico-chemical properties of the molecules [Karlova et al., 2021].

For molecular optimisation, it is desirable that the molecules are embedded into the latent space which is smoothly organised according to the chemical properties and consequently allows for numerical optimisation [Griffiths and Hernández-Lobato, 2020, Zhavoronkov et al., 2019]. Also using the hierarchical prior allows for improved reconstruction and has sufficient capacity to better capture the relationships between latent representations.

The sampling from the latent space can be guided by reinforcement learning [Popova M, 2018, You et al., 2018, Zhavoronkov et al., 2019]. The generation of the new molecule is decomposed as a sequence of actions while biasing the generation towards the compounds that optimise the selected reward function. The choice of reward function proves to be equally challenging. Ideally, the task requires finding the reward function which in a comprehensive way quantifies the bonding probability with the protein target, drug-likeness and synthesize-ability. Popova M [2018] considered quantitative structure-activity relationships models (QSAR)[Hansch et al., 2001] rewards, while [Zhavoronkov et al., 2019] suggests using a combination of self-organising maps (SOMs) [Kohonen, 1990] fitted to target affinities.

The benchmarking platforms [Polykovskiy et al., 2020, Grant LL, 2021, Brown, 2019] were introduced to compare the drug generative pipelines in a standardised manner. These platforms provide metrics which help to detect over-fitting, imbalance of frequent structures or mode collapse, among others. In particular, the metrics quantify the validity and uniqueness of the generated SMILEs, novelty, fragment and scaffold similarity, similarity to the nearest neighbours, internal diversity and Frechet ChemNet distance [Brown, 2019]. Nevertheless, these measures are insufficient to assess the chemical relevance of the reward function.

In this paper we address the problem of how to realistically choose the reward function in the molecular optimisation guided by reinforcement learning. We propose the methodology which helps to quantify the chemical relevance of the reward function. At the core of our method is the quantification of the distances between probability distributions of the biologically active candidates and the candidate samples. In particular, we investigate the joint probability distributions of the docking scores and other physico-chemical properties of the generated samples from the RL agent. We compare these to the same probability distributions of the benchmark bio-active molecules and the decoys. We use optimal transport metrics to quantify this distance. We also propose a tuneable reward function which is a linear combination of the predicted bio-activity and higher level metrics of drug-likeness and chemical and synthetic feasibility. We examine its performance with our methodology to demonstrate its usefulness.

The paper is organised as follows: in the next section we discuss factors important for the selection of preferable drug-like molecules such that it is possible to capture typical benchmarks and medicinal chemistry requirements. In section 3 we introduce the DAFT reward function and discuss how to tune its hyper-parameters. In the section Experiments we give an example of the DAFT reward setting and evaluate its performance on the DUD-E dataset [Mysinger et al., 2012] for GRIA2, DRD3, ESR1 and ABL1 protein receptors.

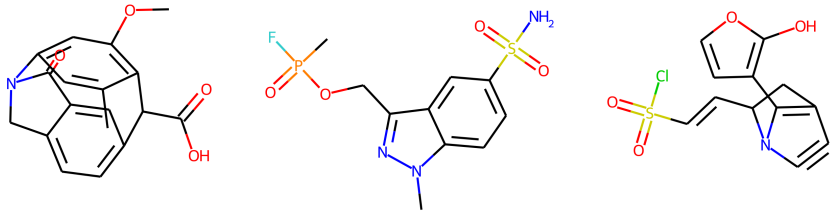

Figure 1: Three selected examples of bad molecules that perform very well in the QED drug-likeness estimate but fail for real-life wet lab applications for assorted reasons. From left to right: 1. molecule is synthetically intractable and is very unstable because of ring strain 2. molecule has methylphosphonofluoridate functionality which is highly toxic and present in organophosphate chemical weapons 3. molecule has assorted problems, including exotic Michael acceptor, reactive sulfonyl chloride group, very strained ring triple bond, and a bridgehead double bond, violating Bredt's rule.

2 Drug-likeliness Feasibility Criteria

For a medical chemist, a set of de novo molecules produced by various ML generative models is a shamefully often low starting point as the predicted compounds may contain problematic reactive groups, the structures are hard to synthesize or possess suboptimal physiochemical properties [Renz, 2019]. Therefore, the post hoc assessment by experienced wet lab personnel is required, with their time often being wasted on ineffective, unstable, not synthesizable and otherwise problematic non-druglike compounds.

In a drug design project, a specific macromolecular target such as a particular receptor or certain enzyme is targeted with the disease process implying this macromolecular target. The goal is to find safe, tolerable small molecules that modulate the functioning of the macromolecular target and modulate the role it plays in the disease process. In order to assess how strongly a small molecule modulates its intended target, we measure the affinity. Affinity is measured and expressed in units such as Ki, Kd, IC50, depending on the measuring assay used and the type of activity. Predicting the affinity of a molecule at its target is a challenging undertaking. Even very advanced modelling techniques do not capture subtle effects like long term dynamics and solvent effects. Nonetheless, by finding patterns and recurring structural features in molecules with known affinity, it is possible to estimate potential activity and filter compounds in a way that enriches the active compounds of the remaining molecular set.

The druglikeness of a compound is determined by a wide balance of factors, including chemical stability, lack of problematic substructures (such as PAINS or reactive functional groups), advantageous physiochemical properties that will affect pharmacokinetics and a good spatial distribution of functional groups corresponding to good ligand efficiency and binding site complementarity. Specific rules of thumb exist to assess druglikeness, most famously Lipinski's rule of five, which sets maximal ranges for hydrogen bond acceptors, hydrogen bond donors, lipophilicity and molecular weight. The QED-score, quantitative estimate of druglikeness [Bickerton et al., 2012] is a single weighted score of 8 calculated properties of molecules, concretely molecular weight, partition coefficient, hydrogen bond donor count, hydrogen bond acceptor count, polar surface area, rotatable bonds, number of aromatic rings and the presence of any structural alerts. These rules are not strict: there are classes of drugs that strongly violate these rules and yet are safe marketed drugs, such as the group of macrolide antibiotics that includes WHO essential medicine Erythromycin. Conversely, there are also compounds which have near optimal QED-scores, because highly unstable and exotic molecules can also meet all the formal criteria, see Figure 1. These more crude druglikeness metrics are often not enough.

Synthesizability, in the case of de novo molecular generation, and commercial availability, in the case of an extensive pre-existing screening library, are also important for practical aspects. Compounds

need to be procured with minimal effort to use available resources efficiently. A standard metric for synthesizability is the SAScore [Ertl and Schuffenhauer, 2009], which determines synthetic accessibility in a rule based way and correlates quite well with real life synthesizability, addressing some of the issues with non-druglike molecules that have good QED scores. The SAScore distribution of commercially available compounds, druglike compounds and natural product-like compounds is also significantly different. The amount of sp^3 hybridized carbons has been used as a crude proxy for 3-dimensionality of compounds, although this does not take the complexity of the framework into account, whereas MCE-18 [Ivanenkov et al., 2019] will assign a score for non-trivial tetrahedral carbons that are not just long linear hydrocarbon segments.

More subtle qualities like sphericity, non-flatness and flexibility have also been suggested as important. In fact, when a molecule is very flexible (estimated by the number of rotatable bonds) structure-based drug design methods based on pose sampling, such as pharmacophore modelling and molecular docking, will encounter problems. When a set of molecules is considered the coverage of chemical space, i.e. the diversity of the set is also important, [Feher and Schmidt, 2002, Ajay et al., 1998].

3 Methods

We aim to quantify the performance of the distribution learning models. Consider a sample $\mathcal{D}$ from the oracle $p_{data}(x)$ generating the molecules. The generative model $q(x)$ approximates the true generative process. We consider a ground truth sample of biologically active molecules and decoys available in the chemical databases such as CHEMLb [Mendez et al., 2018].

The sampling of the SMILE string has been formulated as a reinforcement learning problem. The state space $\mathcal{S}$ is the collection of partially or fully constructed SMILE strings, state s_t corresponds to the string constructed up till character t. The action space $\mathcal{A}$ consists of the dictionary of the characters used by the SMILE grammar and so action a_t is a character which will be appended to the string s_t. The transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \longrightarrow \mathcal{S}$ is learnt by VAE. Next we denote the reward function as $R : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$. Each episode corresponds to appending new character to the existing SMILE string, so the maximal number of episodes T is the maximal length of the SMILE string we want to generate. The environment is the SMILE string generating process modelled by the decision Markov process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mu_0, \gamma, R, T)$, where μ_0 is the initial state distribution and γ the discount factor. Construction of the new molecule is a task for stochastic optimisation, where we search for the policy π_θ that maximizes the state-action trajectories in the environment, i.e. $V^\pi(s_t^*) = \mathbb{E}_\pi \sum_{t=t^*}^{T} \gamma^t R(s_t, a_t | s_t^*)$.

The choice of the reward function determines the overall value and selected trajectories in action-state space. We propose to use the reward function which gives enough flexibility to quantify the biological activity with respect to selected protein target $A(s_t, a_t; \rho)$, drug-likeness $QED(s_t, a_t)$, synthesizability $SA(s_t, a_t)$ and mainly the incorporate the custom index of penalised physico-chemical properties $CP(s_t, a_t)$. We define the drug-likeliness and predicted ligand affinity to a protein target (DAFT) reward function as:

$$DAFT(s_t, a_t; \rho) = \begin{cases} \kappa_1 A(s_t, a_t; \rho) + \kappa_2 QED(s_t, a_t) + \kappa_3 CP(s_t, a_t) + \kappa_4 SA(s_t, a_t), \\ \tau_{\text{filter}}, \end{cases} \tag{1}$$

where $m \equiv m(s_t, a_t)$ denotes the molecule corresponding to SMILE s_t with concatenated a_t, ρ is the protein target and τ is a penalties vector corresponding to vector of filters from the space of hard constrains $\mathbb{F}$. For example if the molecule $m(s_t, a_t)$ constructed from the SMILE (s_t, a_t) is not chemically valid, we set the penalty corresponding to the chemical non-validity. Alternatively, if the molecule is chemically valid but not biologically active, we can penalise for this too. The hyper-parameters $\kappa_i, i = 1 \ldots, 4$ allow for customising the importance of particular component of the reward function and also for absorbing the different scales of the used factors. Term $A(s_t, a_t; \rho)$ corresponds to the functional of the affinity and quantifies the experimental protocol of the binding to a target ρ, e.g. PKi, EC50, docking score or models predicting these quantities, such as SOMs or QSARs type predictors. The custom property allows for hand-crafting rules which can incorporate various functional relationship of underlying physico-chemical variables, e.g. charge, docking score, weighted average of measures describing the binding flexibility of the molecule, measures of molecular similarity or any other desirable feature. The selection of the hyper-parameters κ_i, filters and construction of custom property allows for targeting various illnesses, e.g. targeting distribution to a specific tissue type or organ to modulate a cancer involved kinase tissue-specifically.

The high tunability of hyper-parameters κ_i allows for sufficient flexibility but requires monitoring of the quality of the sample. The standard measures which monitor the sample diversity, novelty or FrechetChemNet distance are insufficient for this task.

We use the following method: consider the sample of the ground truth data, which has relevant labels describing bio-activity, docking score, chemical properties and other scoring metrics. Split the control sample into categories based on the desirable property, e.g. actives and decoys. Compute docking scores of the generated sample and the protein target. Molecular docking is a technique used to estimate non-covalent complexes, usually either between two proteins or between a protein and a small molecule. In the context of a ligand-receptor complex, docking has two main intended outcomes, predicting the pose of the ligand in the receptor pocket and estimating the binding activity. A energy-like scoring function is used to assess a certain pose of the ligand, the docking algorithm will try optimize the scoring function and return the best-scoring pose.

Consider joint probability distribution $q(d_i, \mathbf{c}_i; m, \rho)$ of the docking scores d_i between the molecule m and protein target ρ and a vector of chemical properties $\mathbf{c}_i$ of the molecule m. We would like to compare the distance between our molecules sampled by the RL agent and ground truth distribution $p_{data}(\bar{d}_i, \bar{\mathbf{c}}_i; \bar{m}, \rho)$. The samples from distribution q can have a low variance due to constrains in the reward function and therefore be concentrated on the small regions, whereas the true distribution p_{data} can spread above manifolds of various shapes. Choosing KL-divergence would lead to mis-leading results because KL-divergence goes to infinity when the distributions approaches Dirac measure. To overcome this obstacle we use the Wasserstein distance to quantify the aforementioned distance. In particular we compute the optimal transport between the empirical distributions using the Sinkhor algorithm. [Chizat et al., 2020].

4 Experiments

To validate our proposed methodology, we use two benchmark datasets: DUD-E, the Directory of Useful Decoys [Mysinger et al., 2012], and MOSES, the Molecular Sets [Polykovskiy et al., 2020]. DUD-E dataset is a collection of 102 protein targets and related 22886 ligands from CHEMBl [Mendez et al., 2018], manually selected database of bioactive molecules with drug-like properties. In DUD-E, each ligand is complemented with 50 decoys from ZINC database [Irwin et al., 2012] and has chemical properties similar to the particular ligand. Otherwise the decoys are selected to be topologically dissimilar to the ligands and only the most dissimilar decoys are included. The ligands for each target are selected such that the chemotype diversity is also guaranteed. MOSES is a benchmarking platform introduced to help to standartise the training and comparisons of generative models of the molecular designs. The dataset available with the platform contains 1.9 million compounds hand-selected from ZINC such that the internal diversity is guaranteed. We use the train and test splits as provided with training set of 1.5 million compounds. We examine the modification of the architecture inspired by Zhavoronkov et al. [2019]. To learn the distribution of the SMILE string from MOSES training set, we use supervised VAE with 16-dimensional latents, hierarchical prior represented as Gaussian-mixture and RNN based encoder and decoder architecture. See Appendix A for details on the architecture and training.

From 102 DUD-E targets we select 4 protein targets ρ: Dopamine D3 receptor (DRD3), Estrogen receptor alpha (ESR1), AMPA 2 ionotropic glutamate receptor (GRIA2) and Tyrosine-protein kinase ABL (ABL1). These targets were selected because they have a large amount of experimental biological data deposited and because they correspond to 4 typical protein families targeted in drug development campaigns: a membrane receptor, a nuclear receptor, an ion channel and a kinase enzyme, respectively.

Activity at the four targets of interest was estimated using a random forest classifier and a random forest regressor trained on bio-activity data and ECFP6 fingerprints generated from the structures of active compounds and decoys deposited in DUD-E. For typical QSAR tasks, including prediction of activity at DUD-E targets, random forests using folded circular fingerprints are a standard technique and generally tends to have a good performance on par with other methods [Polishchuk, 2017, Chen et al., 2012].

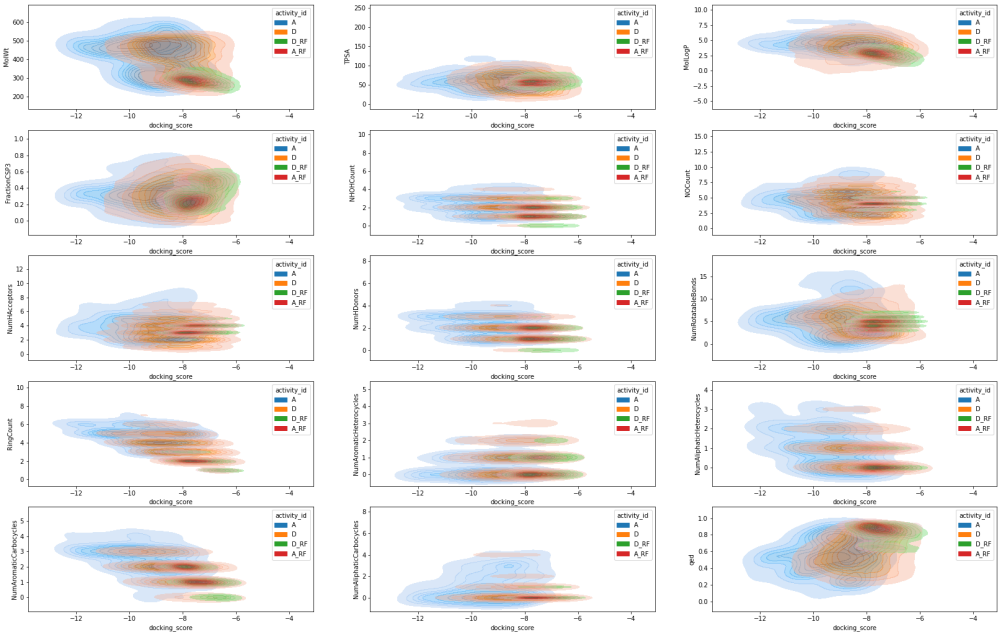

Figure 2: Model Specific Biases for Samples Docked to ESR1 Receptor: to assess the quality of the sample provided by the agent, we depict the joint distribution of the docking score and various physico-chemical molecular descriptors. The plotted descriptors are molecular weight, partition coefficient of the molecule (log P), fraction of sp^3-hybridized carbon atoms and number of rotatable bonds, a measure of flexibility, count of NH and OH, total ring count, and specifically number of aromatic heterocycles, number of aromatic carbocycles, number of aliphatic heterocycles, number of aliphatic carbocycles, total polar surface area, quantitative estimation of drug-likeness (QED-score). To quantify the biological feasibility of the sampled molecules for ESR1 receptor, the docking scores were computed using Vina software. We depicted four samples: the ground truth ligands, decoys, respectively, for ESR1 receptor as available in DUD-E (label A, label D, respectively and the samples sampled by RL agent which we label using the QSAR-RF model. The promising bioactive candidates should be located in the region of the active compounds of the DUD-E benchmark sample. See Table 1 for Wasserstein distances between different groups of samples.

We explored the performance of the agent with hand-picked DAFT reward of the following form:

$$DAFT(m;\rho) = \begin{cases} \frac{pKi(m,\rho)}{10} + QED(m) + \frac{CP(m)}{5} - \frac{SA(m)}{10} & \text{if } m \text{ is a valid ligand,} \\ -2 & \text{if } m \text{ is a valid decoy,} \\ -5 & \text{if } m \text{ is chemically invalid.} \end{cases} \tag{2}$$

The goal of this function is to encourage valid, MCF and PAINS Filter passing molecules, that are also drug-like and synthesizable, resemble CNS penetrating compounds and present non-decoy like interaction with specific targets (pKi). We consider two penalties filtering for chemically valid molecules and biologically active ones. The details on the CP constructions are in Appendix A.2. To explain the motivation behind the choices of parameters $\kappa_i = 1, \ldots, 4$ corresponding to vector $(0.1, 1., 0.2, -0.1)$: value SA(m) is in range between 1 to 10 with value 1 indicating the best synthesize-ability, therefore the choice of -0.1. We search for pKi values which are between 6 and 10, therefor we scale down the term $A(m, \rho)$ with choice 0.1.

Next we train the RL-agent for 8000 epochs using REINFORCE algorithm. The exploration-exploitation rate is set to 30% and 70% of 200 samples. We use Adam optimizer with learning rate 2e-5. The linear layer applied to the latent space samples in the decoder is also trained with an Adam optimizer with a learning rate of 1e-6.

	Labeled Samples Pairs					
	A D	A As	A Ds	D As	D Ds	As Ds
MolWt	0.0468	0.3541	0.3400	0.2773	0.2664	0.0155
TPSA	0.0244	0.0668	0.0572	0.0593	0.0503	0.0277
MolLogP	0.1015	0.2643	0.2547	0.1591	0.1543	0.0645
FractionCSP3	0.1367	0.2421	0.2759	0.0927	0.1334	0.0810
NHOHCount	0.1169	0.2440	0.2861	0.0887	0.1231	0.0951
NOCount	0.0919	0.1968	0.2218	0.1111	0.1225	0.0754
NumHAcceptors	0.1149	0.2593	0.2884	0.1082	0.1235	0.0760
NumHDonors	0.1321	0.2534	0.2963	0.0857	0.1246	0.1077
NumRotatableBonds	0.0905	0.1933	0.2128	0.1074	0.1214	0.0935
RingCount	0.1751	0.3886	0.3911	0.1644	0.1858	0.0846
NumAromaticHeterocycles	0.1423	0.2490	0.2789	0.1359	0.1405	0.1182
NumAliphaticHeterocycles	0.1357	0.2556	0.2848	0.1117	0.1547	0.0924
NumAromaticCarbocycles	0.1440	0.2506	0.2860	0.1038	0.1478	0.1134
NumAliphaticCarbocycles	0.1509	0.2944	0.3193	0.0849	0.1130	0.0785
QED	0.1371	0.2460	0.2788	0.0997	0.1368	0.0799

Table 1: Wasserstein Distances between Actives and Decoys Samples Docked to ESR1 Receptor: to assess the quality of the sample provided by the agent, we provide Wasserstein distance of joint distribution of the docking score and various physico-chemical molecular descriptors for various pairs of labeled samples. The cost function in Wasserstein distance is choosen as square Euclidean norm. Considered descriptors are molecular weight (MolWt), partition coefficient of the molecule (MolLogP), fraction of sp^3-hybridized carbon atoms(FractionCSP3) and number of rotatable bonds, a measure of flexibility, count of NH and OH, total ring count, and specifically number of aromatic heterocycles, number of aromatic carbocycles, number of aliphatic heterocycles, number of aliphatic carbocycles, total polar surface area (TPSA), quantitative estimation of drug-likeness (QED-score). The ground truth ligands, decoys, respectively, for ESR1 receptor as available in DUD-E (label A, label D, respectively) and the samples sampled by RL agent which we label using the QSAR-RF model (labeled as As, Ds respectively).

Samples produced from the RL agent trained with reward DAFT(m; DRD3), DAFT(m; ESR1), DAFT(m; GRIA1) and DAFT(m; ABL1) were docked with the corresponding protein targets. For details on the docking see Appendix A.3. Ground truth samples $p_{data}(\bar{d}_i, \bar{\mathbf{c}}_i; \bar{m}, \rho)$ for benchmarking the DAFT agent were constructed from DUD-E ligands and decoys. We also computed basic physico-chemical molecular descriptors and we constructed the joint distribution with the docking scores. These descriptors include: molecular weight, partition coefficient of the molecule (log P), fraction of sp^3-hybridized carbon atoms and number of rotatable bonds, a measure of flexibility, count of NH and OH, total ring count, and specifically number of aromatic heterocycles, number of aromatic carbocycles, number of aliphatic heterocycles, number of aliphatic carbocycles, total polar surface area and quantitative estimation of drug-likeness (QED-score) [Bickerton et al., 2012]. We depict the dependencies of the chemical properties and docking score of each sample in the Figure 2.

Next we take DUD-E actives and DUD-E decoys. We also take the sample generated by RL agent with DAFT(m; DRD3), DAFT(m; ESR1), DAFT(m; GRIA1) and DAFT(m; ABL1) and label them with QSAR-RF binary classifier with active or decoy label. Then we compute the optimal transport matrix using Sinkhor algorithm [Flamary et al., 2021] with Square Euclidian norm as a cost function. We use this for computing the Wasserstein norm between the joint distributions of the docking scores and particular chemical property (as depicted in Figure 2) of different groups of molecules. In particular we consider pair of ground truth actives and ground truth decoys (A D), ground truth actives and sampled actives (A As), ground truth actives and sampled decoys (A Ds), ground truth decoys and sampled actives (D As), ground truth decoys and sampled decoys (D Ds), sampled actives and sampled decoys (As Ds). See Table 1 for the results. The values in column 'As Ds' of Table 1 indicates that the distributions of the generated decoys and actives are very close to each other and does not reflect the behaviour of the distances between 'A D' pair. Visual inspection of the distributions in Figure 2 indicates that $\kappa = 1.$ in $QED(m)$ term of the DAFT reward function pushed the optimisation into region of high QED values. Further CP term contains constrain on the number

	A D	A D-A Ds	A D - D As	A D - As Ds
init_reward	0.0244	-0.0328	-0.0349	-0.0033
10_SA	0.0244	-0.0436	-0.0504	-0.0464
0.1_SA	0.0244	-0.0422	-0.0367	-0.0162
10_QED	0.0244	-0.0401	-0.0386	0.0068
0.1_QED	0.0244	-0.0459	-0.0472	-0.0384
10_CP	0.0244	-0.0358	-0.0398	0.0073
0.1_CP	0.0244	-0.0449	-0.0461	-0.0304
10_act	0.0244	-0.0434	-0.0441	0.0061
0.1_act	0.0244	-0.0436	-0.0433	-0.0284
2_QED	0.0244	-0.0426	-0.0376	-0.0169
2_CP	0.0244	-0.0429	-0.0389	-0.0140
2_SA	0.0244	-0.0471	-0.0438	-0.0261
2_act	0.0244	-0.0436	-0.0377	-0.0172
0.5_act	0.0244	-0.0464	-0.0469	-0.0366
0.5_QED	0.0244	-0.0468	-0.0469	-0.0368
0.5_SA	0.0244	-0.0455	-0.0412	-0.0216
0.5_CP	0.0244	-0.0468	-0.0483	-0.0403
20_all	0.0244	-0.0352	-0.0415	-0.0272
50_all	0.0244	-0.0323	-0.0363	-0.0077
100_all	0.0244	-0.0342	-0.0355	-0.0083

Table 2: Scaling the hyperparameters $(\kappa_1, \kappa_2, \kappa_3, \kappa_4)$: The Wasserstein distances (with cost function as square Euclidean norm chosen as a cost function) of joint distributions of the docking score and TPSA. The ground truth ligands, decoys, respectively, for ESR1 receptor as available in DUD-E (label A, label D, respectively) and the samples sampled by RL agent which we label using the QSAR-RF model (labeled as As, Ds respectively). The values in columns A D-A Ds, A D-D As, and A D-As Ds refers to difference between the Wasserstein distance for the ground truth actives and decoys AD and the sampled samples from RL agent with scaled reward function. Values closer to 0 are more favourable.

of carbon rings which further restricts the region for the potential candidates. We can conclude that the parameter setting in our experiment leads to over-optimisation which restricts the selection of the potential candidates. Ideally, we would like to optimise our hyperparameters κ, such that they better reflect the region of the biologically active candidates.

In the second experiment we focus on ESR1 receptor and vary the hyper-parameters $\kappa_i, i = 1, \dots 4$ in the reward function. We consider two sets of experiments where we multiply only one of the κ_is by $0.1, 0.5, 2, 10$ while keeping other κs unchanged. This allows us to investigate the impact of each term in the reward function separately. We also study cases where we multiply all terms of the reward function by values $20, 50, 100$. We observe that the values of the Wasserstein distances changes conditioned on the particular chemical property. This experiment indicates that there is a risk of the inherent bias while hand-picking the hyper-parameters due to the human expert demand which can be in contradiction of the inner bias of the optimisation procedures in the ML model. The automated hyper-parameter optimisation can potentially help to tune the human expert requirements and the model design induced biases. See Table 2 for the results for total polar surface area.

5 Discussion

We provided a practical methodology which we believe other researchers interested in generative models for molecular drug designs find useful when constructing the experiments evaluating the performance of newly developed modelling approaches. In particular, we stated requirements on the reward function usable for the generative models of the novel chemical structures.

Our proposal is indeed incomplete and suffers from shortcomings such as dependence on other predictors which potentially accumulates the model error [Sculley et al., 2015], so monitoring the uncertainty of the predictors for biological activity used in the reward functions is inevitable

component of the model-pipeline. We discussed the biases introduced by the scores designed to quantify physico-chemical properties. Biases of this type will be shared with other compound scoring functions, either human expert designed or machine learnt.

The proposed function is general enough to target a wide spectrum of biological targets, but it is important to note that negative aspects, such as toxicity, environmental accumulation, side-effects, are not robustly protected against. Therefore, care must be taken to not rely blindly on compounds privileged by this workflow.

References

Ajay, W. Patrick Walters, and Mark A. Murcko. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *Journal of Medicinal Chemistry*, 41(18):3314–3324, aug 1998. doi: 10.1021/jm970666c.

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

Regine S. Bohacek, Colin McMartin, and Wayne C. Guida. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews*, 16(1):3–50, jan 1996. doi: 10.1002/(sici)1098-1128(199601)16:1<3::aid-med1>3.0.co;2-6.

Fiscato M. Segler M.- Vaucher A. C. Brown, N. Guacamol: Benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, page 1096–1108, 2019.

Bin Chen, Robert P. Sheridan, Viktor Hornak, and Johannes H. Voigt. Comparison of random forest and pipeline pilot naïve bayes in prospective QSAR predictions. *Journal of Chemical Information and Modeling*, 52(3): 792–803, mar 2012. doi: 10.1021/ci200615h.

Lénaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2257–2269. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/17f98ddf040204eda0af36a108cbdea4-Paper.pdf`.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. AutoDock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, jul 2021. doi: 10.1021/acs.jcim.1c00203.

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

Miklos Feher and Jonathan M. Schmidt. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *Journal of Chemical Information and Computer Sciences*, 43 (1):218–227, dec 2002. doi: 10.1021/ci0200467.

Rami Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurarlie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, LÃco Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

Sit CS. Grant LL. De novo molecular drug design benchmarking. *RSC Med Chem.*, pages 1273–1280, Jun 2021.

Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained bayesian optimization for automatic chemical design using variational autoencoders. *Chem. Sci.*, 11:577–586, 2020.

R Gómez-Bombarelli, JN Wei, D Duvenaud, Hernández-Lobato JM, and Aguilera-Iparraguirre J Hirzel TD Adams RP Aspuru-Guzik A. Sánchez-Lengeling B, Sheberla D. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci.*, 2:268–276, Feb 2018. doi: doi:10.1021/acscentsci.7b00572.

Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, David Duvenaud, Dougal Maclaurin, Martin Blood-Forsythe, and Hyun Sik et al. A, Chae. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials*, 15(10): 1120–1127, 2016.

Inbal Halperin, Buyong Ma, Haim Wolfson, and Ruth Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Genetics*, 47(4):409–443, may 2002. doi: 10.1002/prot.10115.

Corwin Hansch, Alka Kurup, Rajni Garg, and Hua Gao. Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms. *Chemical Reviews*, 101(3):619–672, feb 2001. doi: 10.1021/cr0000067.

John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7):1757–1768, 2012.

Yan A. Ivanenkov, Bogdan A. Zagribelnyy, and Vladimir A. Aladinskiy. Are we opening the door to a new era of medicinal chemistry or being collapsed to a chemical singularity? *Journal of Medicinal Chemistry*, 62 (22):10026–10043, 2019. doi: 10.1021/acs.jmedchem.9b00004. URL `https://doi.org/10.1021/acs.jmedchem.9b00004`.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

Andrea Karlova, Wim Dehaen, and Daniel Svozil. Molecular fingerprint vae. In *ICML Workshop on Computational Biology*, 2021.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL `http://arxiv.org/abs/1312.6114`.

T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990. doi: 10.1109/5.58325.

Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1945–1954. PMLR, 06–11 Aug 2017. URL `http://proceedings.mlr.press/v70/kusner17a.html`.

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

Georgia B. McGaughey, Robert P. Sheridan, Christopher I. Bayly, J. Chris Culberson, Constantine Kreatsoulas, Stacey Lindsley, Vladimir Maiorov, Jean-Francois Truchon, and Wendy D. Cornell. Comparison of topological, shape, and docking methods in virtual screening. *Journal of Chemical Information and Modeling*, 47(4): 1504–1519, jun 2007. doi: 10.1021/ci700052x.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, nov 2018. doi: 10.1093/nar/gky1075.

Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55 (14):6582–6594, 2012. doi: 10.1021/jm300687e. URL `https://doi.org/10.1021/jm300687e`.

I. V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

Pavel Polishchuk. Interpretation of quantitative structure–activity relationship models: Past, present, and future. *Journal of Chemical Information and Modeling*, 57(11):2618–2639, oct 2017. doi: 10.1021/acs.jcim.7b00274.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology*, 2020.

Tropsha A. Popova M, Isayev O. Deep reinforcement learning for de novo drug design. *Sci Adv.*, Jul 2018.

Van Rompaey D. Wegner-J. K. Hochreiter S. Klambauer G. Renz, P. On failure modes in molecule generation and optimization. *Drug discovery today*, 2019. URL `https://doi.org/10.1016/j.ddtec.2020.09.003`.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR. URL `http://proceedings.mlr.press/v32/rezende14.html`.

Aspuru-Guzik A Sanchez-Lengeling B. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, pages 360–365, Jul 2018.

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2503–2511. Curran Associates, Inc., 2015. URL `http://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf`.

Oleg Trott and Arthur J. Olson. AutoDock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, pages NA–NA, 2009. doi: 10.1002/jcc.21334.

David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, feb 1988. doi: 10.1021/ci00057a005.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/d60678e8f2ba9c540798ebbde31177e8-Paper.pdf`.

Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.

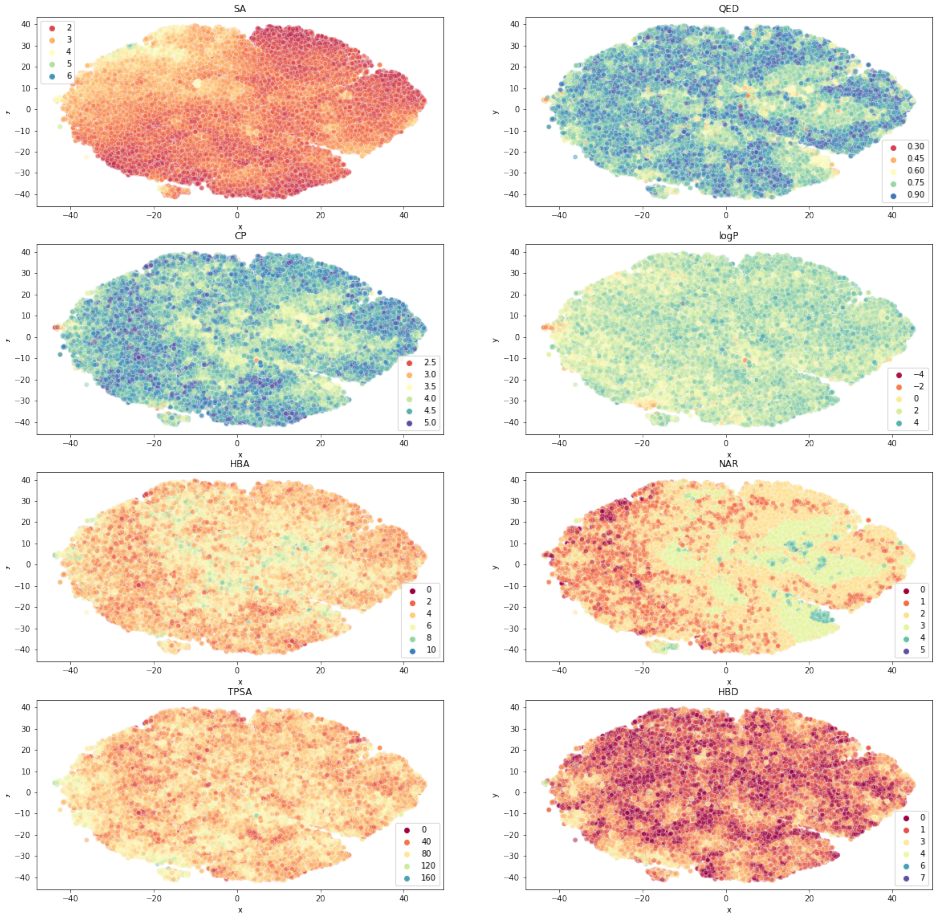

Figure 3: Visualisation of the Latent Space used for RL Training: the VAE is trained in 1.5mil SMILEs from benchmark MOSES dataset. Chemical properties used for training are: SA, QED and CP, the latter being composed of $logP$, number of hydrogen bonds donors, number of hydrogen bonds acceptors, number of aromatic rings and the topological polar surface area.

A Experiment Details

A.1 VAE

We use one-hot encoding for the SMILEs string as in Gómez-Bombarelli et al. [2018]. We examine the modification of the architecture inspired by Zhavoronkov et al. [2019]. To learn the distribution of the SMILE string, we use supervised VAE with 16-dimensional latents. The VAE encoder architecture is: a 2-layer bidirectional GRU with hidden dimension of 256 followed by 2-layer MLP with dimensions: 256, 32 and Leaky Relu activation with slope 0.1. The decoder architecture is: 1-linear layer: 128, dropout rate 0.2, followed by 3-layer GRU with 128 hidden dim followed by scaling linear layer matching the length of the input SMILE, which we set to 47. We train the decoder in teach-forcing manner, i.e. both the SMILE and its latent representation are pass to GRU-layers, while maximizing the logits corresponding to the correct tokens in the initial SMILE. We use structured prior distribution $\pi(z)s$ which is composed of mixture of Gaussians with mixture weights learned by tensor-train decomposition parameter as in [Oseledets, 2011, Zhavoronkov et al., 2019]. The chemical properties used in supervised training are: SA, QED and CP, see Figure 3 for visualisation of the trained latent space.

	Receptor			
Metrics	ABL1	DRD3	ESR1	GRIA2
valid molecules	0.798	0.888	0.757	0.855
unique molecules for sample size 1000	0.884	0.843	0.999	0.908
unique molecules for sample size 10000	0.706	0.645	0.984	0.752
unique molecules for sample size 20000	0.637	0.575	0.972	0.685
unique molecules for sample size 25000	0.619	0.552	0.969	0.664
Frechet ChemNet distance	20.7	25.1	8.71	16.2
simmilarity to nearest neighbour	0.515	0.552	0.496	0.53
fragment similarity	0.346	0.417	-0.00133	0.291
scaffold similarity	0.279	0.0299	0.527	0.233
internal diversity	0.759	0.758	0.82	0.775
LogP	0.352	0.711	0.217	0.231
Synthetic Accessibility Score	0.242	0.0664	0.0101	0.151
QED	0.0615	0.0489	0.0296	0.026
molecular weight	29.1	22.5	22.7	34.1
novelty	0.991	0.995	0.989	0.98

Table 3: MOSES Metrics: Samples from various reward functions evaluated using MOSES metrics. Metrics are computed based on comparison between the set generated from RL Agent and reference testset from the MOSES platform. Validity and uniqueness corresponds to unique and chemically valid generated SMILE strings. Novelty is the fraction of the molecules that are not present in the reference test set. Internal diversity is computed as arithmentic average of Tanimoto similarities between the sampled molecules subtracted from 1, i.e. closer to 1 corresponds to more diverse sample. The reported values for the chemical properties (molecular weight, LogP, SA score and QED) corresponds to 1-Wasserstein distance computed between the sample from RL agent and reference sample from test set.

VAE is trained for 10 epochs on MOSES training dataset using Adam optimizer with a learning rate of 1e-3 with batch-size 256.

A.2 Custom Property Setting

The custom property CP is designed to yield values close to 5 for molecules that resemble CNS penetrating drugs, similar to CNS_MPO, as defined by value ranges for the following properties: $\log P$, number of hydrogen bonds donors, number of hydrogen bonds acceptors, number of aromatic rings and the topological polar surface.

We split the values of the properties into custom ranges and assign them a custom weights as follows: for $\log P$ we have values -2, 0, 4, 5 and 7 with associated weight vectors for the corresponding intervals (0.05, 0.05 to 1., 1., 1. to 0.3., 0.3 to 0.05, 0.05). For the number of H-bond acceptors we have values 2 and 6, with weights (1., 1. to 0.05, 0.05). For the number of aromatic rings we have values 0, 1, 2 and 5 with weights (0.8, 0.8 to 1., 1., 1. to 0.4, 0.4). For the topological polar surface area we have values 60, 100 and 140 with weights (1., 1. to 0.9, 0.9 to 0.59, 0.59). For the number of H-bond donors we have values 1, 3 and 4 with weights (1., 1. to 0.9, 0.9 to 0.6, 0.6).

A.3 Docking

Molecular docking was performed using AutoDock Vina 1.2.0, an open-source tool for molecular docking [Eberhardt et al., 2021]. Protein and ligand preparation were standard: protein structures were extracted from the RCSB protein database and split into receptor and ligand. The receptor was prepared for docking using the ADFR Suite. All ligands were preprocessed using the Meeko package. Initial 3D conformation were generated using the MMFF forcefield. The coordinates of the ligand were extracted and used as the coordinates for the grid box. In all cases, the dimensions of the gridbox were set to 25x25x25 Å. The Vina scoring function, which was tuned using PDBBind information, was used. Vina uses a quasi-Newton optimization method, BGFS "Broyden-Fletcher-Goldfarb-Shanno" method to find the global optimum [Trott and Olson, 2009].