De Novo Drug Design with a Docking Score Proxy

Tomasz Danel
Jagiellonian University

Maciej Szymczak
Jagiellonian University

Łukasz Maziarka
Jagiellonian University

Igor Podolak
Jagiellonian University

Jacek Tabor
Jagiellonian University

Stanisław Jastrzębski
Jagiellonian University, Molecule.one

Abstract

De novo drug design has shown promise for generating molecules achieving high binding affinity. In the generation process, binding affinity of a molecule is typically approximated using a machine learning model trained on experimental data. We investigate how well the generated molecules are scored by docking, a popular computational method for assessing binding activity. We observe that optimizing for these proxies can produce poor binders, molecules that achieve poor docking scores or do not dock altogether. We make a simple recommendation for practitioners to include in the search process a machine learning based proxy that approximates the docking score. We demonstrate in-silico on the task of finding binders for the BACE receptors that it leads to more realistic molecules that achieve realistic docking scores within the range observed in the training set.

1 Introduction

De novo drug design is a computational method for generating novel compound structures from *scratch* [12]. These methods hold promise for tapping into the vast unexplored space of drug-like molecules. Typically, the generation process involves maximizing a predefined scoring function. This scoring function can include a number of factors such as toxicity or druglikeness.

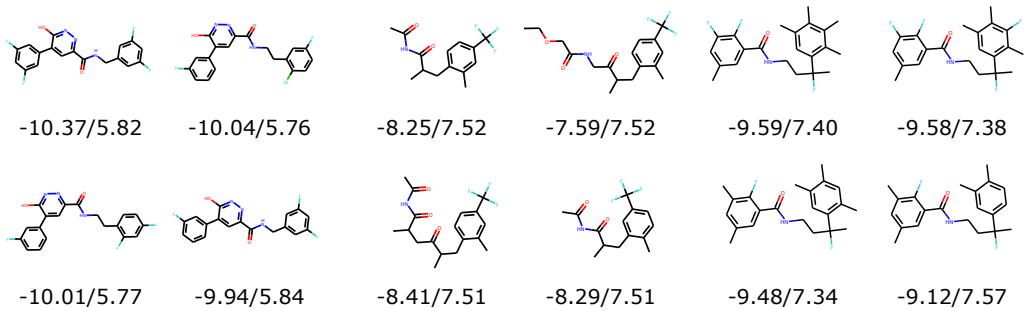
We turn our attention to the goal-oriented de novo design in which the generation process is aimed at finding binding molecules. In this case, the binding affinity is usually approximated by a machine learning model [8, 13, 9].

Guiding the search process using the output of a machine learning model has been shown to fail to generate meaningful samples in de novo models [11]. Specifically, the authors of [11] show that highly binding molecules according to a given model, are not highly binding according to the same model but trained on a different subset from the same dataset. In a similar spirit, [2] show that generally common de novo models fail to generate compounds with high docking scores, when optimized using a proxy for docking score.

We further investigate docking, a popular computational method for assessing ligand binding [7, 5]. Concretely, we use a popular open-source docking software SMINA [5]. Adding to recent investigations [11, 2], we show that de novo models that optimize for a proxy of ligand activity produce compounds that exhibit substantially lower docking scores than observed in the training dataset.

Our main contribution is to make a simple recommendation to practitioners to include docking score in the de novo optimization function. We leverage recent work that shows that outputs of docking programs can be reasonably well approximated using a deep neural network [4, 3]. This enables predicting docking scores rapidly, which makes it practical for use in de novo search. In a preeliminary in-silico study on the task of finding binders to the BACE receptors, we find that

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io



(a) DS (b) pIC50 (c) Both

Figure 1: Best compounds sampled by the models trained to optimise (a) docking score, (b) pIC50, or (c) both. As the reward, RF predictions with QED correction were used. Below each compound, its docking score and predicted pIC50 is shown.

adding such a scoring proxy to de novo models substantially improves the docking scores without compromising predicted pIC50 scores.

2 Experiments

Recent work has shown that de novo models, when applied to generating active binders, tend to produce unrealistic molecules [11]. Adding to these results, we investigate here, and propose ways to improve, the docking score of molecules generated by a state-of-the-art de novo model.

2.1 Experimental setup.

We search in-silico for molecules inhibiting the BACE-1 enzyme. The experimental data consists of 1513 molecules with activity towards β-secretase 1. We use pIC50 values to train a regression model $p_\theta(x)$ to predict inhibition and next apply REINVENT, a state-of-the-art model for de novo drug design, to generate compounds *de novo* that have low IC50. As the reward we use:

$$R_{DS} = \exp(\hat{p}_\theta(x)). \tag{1}$$

In experiments involving docking, we use SMINA docking software [5], due to its popularity and being freely available. We further choose Vinardo docking score over the default docking scoring function in SMINA [10]. We train another regression model $q_\xi(x)$ to predict Vinardo docking score and use this model to as an additional objective in de novo experiments. Finally, we also consider using both predictive models and optimise two targets with REINVENT. The joint reward is defined as follows:

$$R_{Both} = \exp(\hat{p}_\theta(x) - \hat{q}_\xi(x)). \tag{2}$$

We use two models to predict IC50 or docking score. The first model is Random Forest (RF) that uses ECFP features, and the second model is Molecule Attention Transformer (MAT) [6] which is a deep neural network working directly on molecular graphs. A random search of 50 hyperparameter sets was run to choose the best performing model, and then models were retrained on the full data using the chosen hyperparameters. We used a scaffold split to construct the validation subset.

QED correction. We observed that optimising either docking score or pIC50 may lead to generating big unrealistic compounds with repeated substructures, see Figure 3 in the Appendix. These structures may correlate well with the predicted target in the dataset, but are not likely to be good drug candidates even if they achieve high docking scores. To assess the druglikeness of a compound, a numeric score such as QED [1] can be used. QED is also shown in the aforementioned figure where it indicates low druglikeness of the proposals. To address the issue of producing non-druglike compounds, we multiply the reward by QED, which is a value between 0 and 1. The use of this measure imposes additional conditions on the generated structures, making them more compact and less repetitive.

2.2 Results

Generated molecules with high predicted binding affinity and/or high docking score to the BACE-1 enzyme are presented in Table 1. We first observe that optimizing just for affinity based on experimental data (pIC50) results in molecules that have low docking scores relatively to the training set. The average docking score in the training set is -9.13. In comparison, the average docking score for MAT/RF is -6.94/-7.77 without QED correction and -7.63/-8.78 with QED correction. They also often generate compounds that do not dock at all.

Adding docking score as an additional optimization target substantially improves the mean docking score, while still achieving high pIC50 scores. We observe the average docking score increases for MAT/RF to -7.47/-8.82 without QED correction, and -8.26/-9.16 with QED correction. Importantly, the predicted pIC50 score is comparable between using or not docking score proxy.

We also investigate the diversity of the proposed compound libraries. We compute diversity as $D = 1 - \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{i-1} s(c_i, c_j)$, where N is the number of compounds in the library, c_i is the i-th compound, and s is the Tanimoto similarity between compounds. We observe that RF creates less diverse drug proposals, collapsing into a narrow spectrum of compounds. With regard to the MAT results, its compounds are only slightly less diverse than originally in the dataset or in a random sample from the ChEMBL database.

Table 1: Results of the search for drug-like molecules binding to BACE-1. Each row corresponds to 256 generated molecules optimised towards the given target. The DS column shows the mean docking score and the average score of top 5% molecules. Analogically, the pIC50 column shows the mean pIC50 value and averaged top 5% values predicted by the corresponding base model.

(a) Results for search with QED correction.

Base model	Target	DS		pIC50		Undocked	Diversity
MAT	DS	-7.91	-9.23	5.05	5.15	1	0.74
	pIC50	-6.94	-7.21	6.60	7.64	6	0.82
	Both	-7.47	-8.49	5.49	6.92	1	0.70
RF	DS	-8.58	-10.06	5.52	5.75	3	0.57
	pIC50	-7.77	-8.01	7.39	7.51	1	0.41
	Both	-8.82	-9.28	7.32	7.50	56	0.42

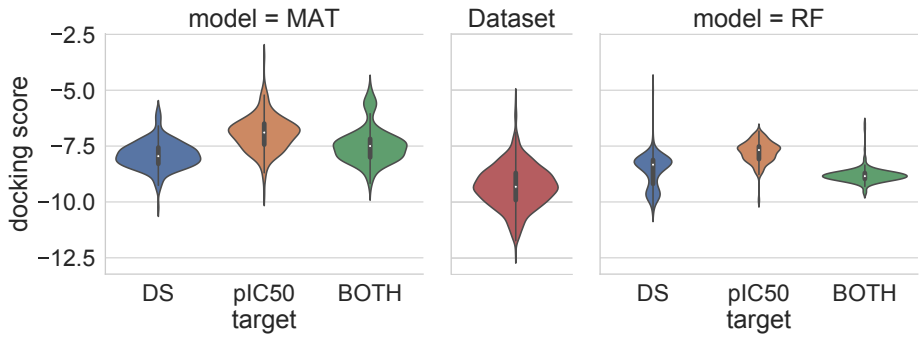
(b) Results for search without QED correction.

Base model	Target	DS		pIC50		undocked	diversity
MAT	DS	-9.21	-11.00	5.05	5.68	23	0.73
	pIC50	-7.63	-8.22	6.33	7.77	14	0.85
	Both	-8.26	-10.33	6.45	7.32	10	0.79
RF	DS	-9.32	-10.82	6.55	6.41	3	0.37
	pIC50	-8.78	-8.81	7.05	7.36	113	0.35
	Both	-9.16	-10.15	7.43	7.46	9	0.34

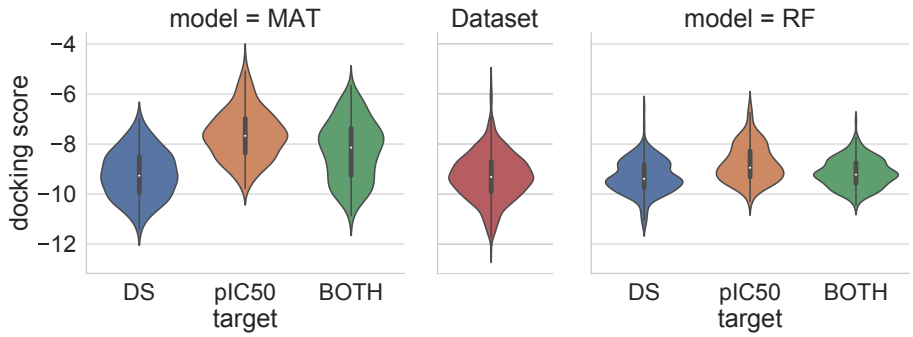
(c) Statistics of the molecules sampled from the prior model and experimental values in the dataset.

Base model	Target	DS		pIC50		undocked	diversity
ChEMBL prior	-	-8.16	-10.16	-	-	15	0.87
Dataset	-	-9.31	-11.30	6.52	7.25	-	0.82

The docking score distributions of the generated libraries, comparing to the dataset docking scores, are shown in Figure 2. Libraries optimised for the pIC50 clearly lose their ability to dock well since the distribution for this target is shifted towards higher values comparing to the dataset distribution. On the other hand, libraries optimised for both docking score and pIC50 maintain relatively low



(a) QED



(b) w/o QED

Figure 2: Distribution of the docking scores of generated compound libraries. The Vinardo docking score values for each model and each optimisation target are compared with the distribution of docking scores in the dataset. Plots show libraries (a) with QED correction and (b) without QED correction.

docking scores. This indicates that including docking score as a proxy task for predicting binding affinity preserves desirable ligand-protein interactions without impairing the pIC50 optimisation.

Finally, Figure 1 shows four best compounds for each target with the base model being RF and QED correction in the reward function. The problem with big repetitive structures was overcome while maintaining the novelty of the generated molecules – all compounds in Figure 1 have the Tanimoto distance to the training set greater than 0.6.

On the whole, we observe that adding docking proxy ('Both' in Table 1) improves docking score of the generated molecules with small or none negative impact on predicted pIC50 and the diversity of the sample. Hence, we using proxy for docking score as a practical and simple addition for de novo drug discovery.

3 Conclusions

Recent studies have shown limitations of current incarnation of de novo drug design for generating active binders [11, 2]. Adding to these studies, we investigated the docking score of molecules generated by REINVENT, a popular method for de novo drug design.

We found that optimizing for a proxy of experimental affinity (as in [8, 13, 9]) often results in unrealistic molecules attaining a poor docking score (compared to molecules in ChEMBL) or even failing to dock. As the main contribution, we proposed to add as an objective a computationally cheap proxy predicting the docking score (based on [4, 3]) which, as we show, constrains the search to more realistic molecules in terms of attained docking scores.

References

[1] 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, 2012.

[2] Tobiasz Cieplinski, Tomasz Danel, Sabina Podlewska, and Stanislaw Jastrzebski. We should at least be able to design molecules that dock well, 2020.

[3] Francesco Gentile, Vibudh Agrawal, Michael Hsing, Anh-Tien Ton, Fuqiang Ban, Ulf Norinder, Martin E. Gleave, and Artem Cherkasov. Deep docking: A deep learning platform for augmentation of structure based drug discovery. *ACS Central Science*, 6(6):939–949, 06 2020.

[4] Stanisław Jastrzębski, Maciej Szymczak, Agnieszka Pocha, Stefan Mordalski, Jacek Tabor, Andrzej J. Bojarski, and Sabina Podlewska. Emulating docking results using a deep neural network: A new perspective for virtual screening. *Journal of Chemical Information and Modeling*, 60(9):4246–4262, 09 2020.

[5] David Ryan Koes, Matthew P. Baumgartner, and Carlos J. Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 2013. PMID: 23379370.

[6] Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. Molecule attention transformer. *arXiv preprint arXiv:2002.08264*, 2020.

[7] Garrett M Morris, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell, and Arthur J Olson. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, 30(16):2785–2791, 12 2009.

[8] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, 2017.

[9] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de-novo drug design. *CoRR*, abs/1711.10907, 2017.

[10] Rodrigo Quiroga and Marcos A. Villarreal. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLOS ONE*, 11(5):1–18, 05 2016.

[11] Philipp Renz, Dries Van Rompaey, Jörg Kurt Wegner, Sepp Hochreiter, and Günter Klambauer. On failure modes of molecule generators and optimizers, Apr 2020.

[12] Benjamin Sanchez-Lengeling and volume = 361 number = 6400 year = 2018 issn = 0036-8075 eprint = https://science.sciencemag.org/content/361/6400/360.full.pdf journal = Science Aspuru-Guzik Alan, title = Inverse molecular design using machine learning: Generative models for matter engineering.

[13] Marwin H. S. Segler, Thierry Kogej, Christian Tyrchan, and Mark P. Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4(1):120–131, 2018.

A Additional experimental results

In Figure 3 we show QED for molecules generated without including QED as a n optimization target.

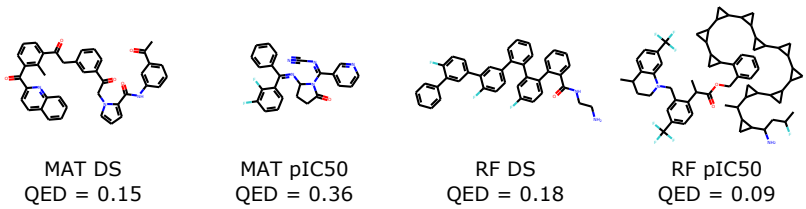
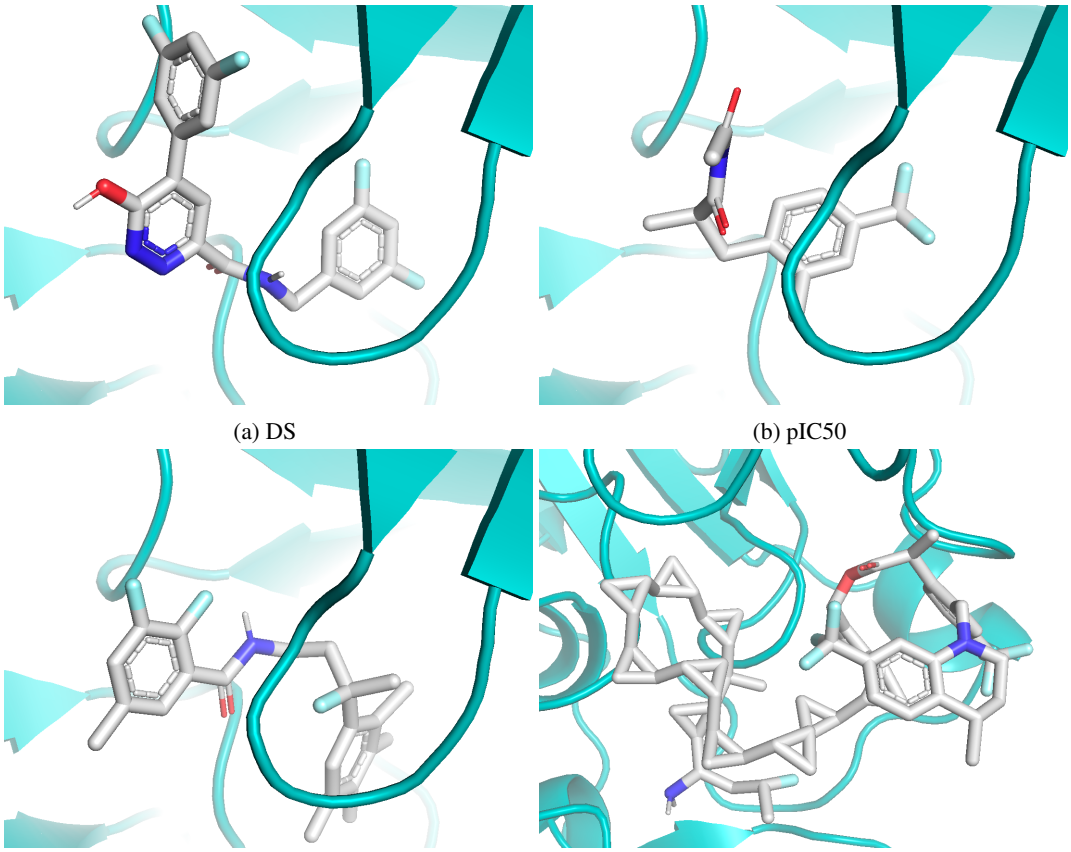


Figure 3: Exemplary compounds sampled with two predictive models (MAT and RF) and two optimisation targets (DS and pIC50) without QED correction. The druglikeness score is shown below each molecule.

B Docking poses

Figure 4 depicts docking poses of the generated compounds inside the binding pocket of BACE-1. It should be noted that the compound without QED correction failed to dock inside the binding pocket, but it rather binds to the protein surface.



(a) DS (b) pIC50

(c) Both (d) w/o QED

Figure 4: Docking poses of the proposed compounds; (a-c) show the best compounds optimised with models predicting docking score, pIC50, and both targets; (d) shows one compound without QED correction.