ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Multiobjective Reinforcement Learning in Optimized Drug Design

Maryam Abbasi, Tiago Pereira, Beatriz P. Santos
Bernardete Ribeiro and Joel P. Arrais *

Department of Informatics Engineering (DEI)
Center for Informatics and Systems of the University of Coimbra (CISUC)
University of Coimbra, Coimbra, Portugal

Abstract. Machine learning has been increasingly applied with success in generating synthetically reasonable molecules. However, a complete system capable of both producing valid molecules and optimizing multiple traits has remained elusive. This paper employs multiobjective reinforcement learning to draw a framework to design compounds. Different multiobjective techniques have been evaluated, such as weighted sum and Chebyshev. The results show that the implemented model can be effectively optimized towards different and competing molecular properties. Nonetheless, the model implemented with the weighted sum scalarization technique with a weight of 0.55 for biological affinity is the one with the most appropriate trade-off for the different evaluated properties.

1 Introduction

The design of molecules with optimized key properties is fundamental in drug discovery. It is a multi-disciplinary and time-consuming process involving sophisticated methodologies and a high financial risk [1]. Nonetheless, as new diseases arise or new efficient ways of treatment for the exploration of existing conditions, it becomes more obvious that a reliable and efficient drug discovery pipeline is needed. The first step in this pipeline is to identify drug candidate molecules, lead compounds, which are the starting point whose design requires further structural optimization to improve the potency, selectivity, or pharmacokinetics. Lead compounds design is inherently a multiobjective problem where we have several objectives to satisfy. It is necessary to develop drugs that optimize the physicochemical properties such as absorption, distribution, metabolism, excretion, and toxicity. Besides, the candidate drugs should be valuable towards pharmacological properties, such as efficacy and target selectivity [2]. Therefore, by taking advantage of the potential of Machine Learning (ML) and multiobjective optimization algorithms, it is desirable to utilize the enormous datasets of chemical compounds and perform an efficient exploration toward the desirable properties for novel drug candidates [3].

Among the more successfully implemented methods that use the multiobjective technique without the use of machine learning, we highlight the work of Nicolau et al. [4]. The author combines evolutionary algorithms with local search techniques to generate molecules using a graph-based model. A widely

*This work has been supported by the Portuguese Research Agency FCT, through D4 - Deep Drug Discovery and Deployment (CENTRO-01-0145-FEDER-029266).

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

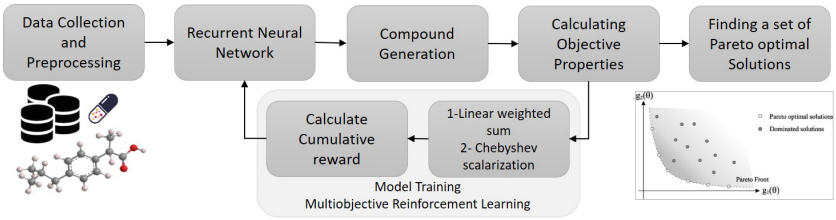


Fig. 1: The general framework for generation of multiple optimized molecular property.

used machine learning method is the variational autoencoder. It involves mapping molecules to a latent space in which the desired optimization is carried out, and, finally, the compounds are remapped again to the molecular space. In this type of approach, molecules are generally managed in string notation, SMILES format [5, 6]. Models based on the same principle but using graph representations of molecules have also been constructed, such as the work of [7]. Another strategy widely used for the de-novo generation of molecules is to define the problem through inverse design. The desired properties for the molecules are indicated a priori, and then, using virtual screening or Reinforcement Learning (RL), the compounds that satisfy these preferences are identified [8]. In this regard, Popova et al. [9], and Olivecrona et al. [10] built solutions using RL and artificial neural networks to guide a SMILES-based molecular generator through biologically promising chemical spaces. Also, Zhou et al. [11] combined RL with chemistry domain knowledge to build a multiobjective generator model from scratch that can produce molecules in compliance with chemical rules. Li et al. [7] implemented a graph-based generator to satisfy structural, physicochemical, and biological objectives simultaneously.

In this paper, we use a Recurrent Neural Network (RNN) architecture to construct a Generator that generates valid compounds from SMILES strings. The model was then retrained using Multiobjective RL (MORL) to find molecules with optimized drug-like properties. Figure 1 shows the general framework of this work. In MORL, the reward comes in the form of a vector where each element corresponds to an objective (molecular property). Therefore after generating the compounds, the vectors corresponding to distinct molecular properties are calculated. The MORL is based on the scalarization function that can be either a linear combination weighted sum [12], or nonlinear (Chebyshev scalarization) [13]. This method allows the accurate analysis of the Pareto front solutions, as it produces several non-dominated compounds simultaneously, which provides valuable information about trade-offs among the objectives at a low computational cost. The novelty introduced in this work is the strategies that guarantee that the molecules contain fundamental characteristics to be lead compounds, including the implementation of different multiobjective optimization methods.

2 Methods

2.1 Reinforcement Learning

Reinforcement learning is based on the formal framework of the Markov decision problems (MDP). RL is regarding how a number of actions should be performed by decision makers (or agents) in a specified context in order to

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

maximize the notion of cumulative reward. Here, the environment is the design of the molecule, and the goal is to find a policy π which selects an action for each state that can maximize the future rewards. Intuitively, we are trying to fit a function $Q(s, a)$ that predicts the future rewards of taking an action a on state s. A decision is made by choosing the action a that maximizes the Q function, which leads to larger future rewards. Mathematically, for a policy π, we can define the value of an action a on a state s to be

$$Q^{\pi}(s, a) = Q^{\pi}(m, t, a) = \mathbb{E}_{\pi}\left[\sum_{i=t}^{T} r_i\right]$$

where Q denotes taking an expectation with respect to π, and r_i denotes the reward at step i. This action-value function calculates the future rewards of taking action a on state s, and subsequent actions decided by policy π. Thus, we can define the optimal policy as $\pi^*(s) = \text{argmax}_a Q^{\pi^*}(s, a)$.

2.2 Multiobjective Reinforcement Learning

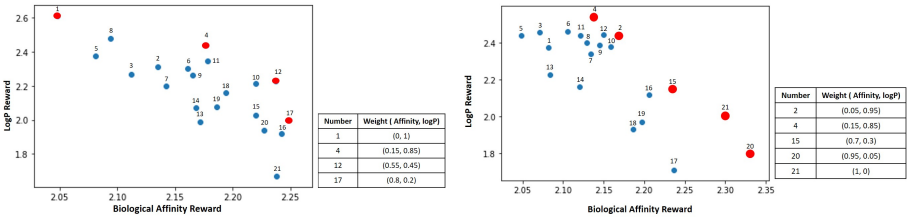
In the context of multi-objective reinforcement learning, the environment will return a vector of rewards at each time step t, with one reward for each objective, i.e. $\vec{r_t} = [r_{1,t}, \ldots, r_{k,t}]^T \in \mathbb{R}^k$ where k is the number of objectives. There exist various goals in multiobjective optimization. The aim may be to find a set of Pareto optimal solutions, or locate a single or multiple solutions that satisfy a decision maker's preference. In this article, we adapted the latter. Especially, to achieve multi-objective optimization, we used the scalarized reward framework, with the introduction of a user defined weight vector $w = [w_1, w_2, \ldots, w_k]^T \in \mathbb{R}^k$. After that, the scalarized reward can be computed by $r_{s,t} = w^T \vec{r_t} = \sum_{i=1}^{k} w_i r_{i,t}$.

Consequently, the objective of the MDP is to maximize the cumulative scalarized reward. To address this task, two types of scalarization strategies were explored to find a representative number of solutions that approximate the Pareto front. First, we applied a Linear Weighted Sum(LWS) with a uniform weights sampling. Second, a non-linear method was implemented to transform the vector of rewards into a scalar number called Chebyshev scalarization (Chev). Nevertheless, a common normalization approach was adopted in the strategies so that both objectives had the same relative importance a priori. This means that, for each generated molecule, after the reward assignment for each property, both values were normalized between 0 and 1. Therefore, it was guaranteed that both rewards were in the same range of values before scalarizing. Otherwise, it would not be easy to compare the two objectives having different ranges. Even so, to evaluate the models, the evolution of the scaled reward as well as the individual rewards were analyzed to prevent situations in which just one of the objectives is optimized, increasing the combined scaled reward in such a way that it gives the illusion that the model is contemplating the two goals when, in fact, only one of them is appropriately evolving.

3 Experimental Analysis and Results

This framework consists of two independently pre-trained models: the molecule generator and a biological affinity predictor for the target Adenosine A_{2A}. A

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.



(a) Linear Weighted Sum (b) Chebyshev

Fig. 2: Dominated and non-dominated solutions obtained by scalarization technique

dataset of 500,000 drug-like molecules extracted from ChEMBL was used to train the recurrent neural network generator [14]. The predictor model was trained following the procedure described in the work of Popova et al., employing 4,872 compounds and their respective biological affinities for the target (ChEMBL identifier: CHEMBL251) [9]. Moreover, the demonstration of the framework's ability to generate molecules considering several properties was performed by optimizing the biological affinity for the target Adenosine A_{2A} and lipophilicity through the partition coefficient. The biological affinity was measured in terms of pIC_{50}, and the higher this value, the greater the probability of the molecule to inhibit the desired target. Regarding the partition coefficient, the parameter employed was the ratio at the equilibrium of the concentration between octanol and water (logP). In this case, the RDKit tool was used to conduct its calculation [15]. This parameter influences the lipophilicity of the compounds, which is fundamental for the bioavailability of the candidate drugs to permeate the blood-brain barrier that protects the brain. According to Lipinski's rule of five, druglike molecules must have the logP between 1 and 4 [9]. The goal is to bias the RNN generator in order to maximize the pIC_{50} and the number of molecules generated with the logP in the range between 1 and 4. The importance allocated to each objective is determined by the weight assigned a priori to each one. The study of the best combination of weights is performed by testing several combinations, namely, through a uniform sampling of the weights between 0 and 1 with a step of 0.05. The results of the application of the two scalarization techniques are summarized in Figure 2. Each weight assignment is represented by a point whose coordinates are the rewards for each objective obtained at the end of the RL process.

It should be emphasized that the sum of the weights assigned to the two objectives must be 1, i.e., the higher the importance dedicated to one of them, the smaller will be the weight of the other objective. The results reflect this scenario since, for both scalarization techniques, there is competition between the two objectives. In most cases, the greater the weight assigned to the goal, the stronger its optimization will be and vice versa. Our goal is to identify the trade-off that best satisfies the different molecular purposes. In this sense, the solutions identified as red points are considered non-dominated since they have both objectives better optimized than the others, the dominated solutions. However, the non-dominated solutions are incomparable with each other as they all have one goal with a higher reward than the other. The identification of the best solution has

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

to be conducted with a more accurate assessment based on the direct analysis of the molecular properties. Therefore, non-dominated solutions were compared considering properties such as biological affinity, logP, validity, uniqueness, diversity, and synthesizability. Comparing the binding affinity property is done by calculating the distributions between molecules generated by the simple RNN generator and after retraining by MORL. The more significant the difference between both distributions - represented in the difference affinity column - the greater the biological affinity of molecules optimized towards the target. The results obtained for the different non-dominated solutions for each scalarization technique are shown in Table 1.

Solution	**Affinity biasing**	**% in range LogP**	**Valid (%)**	**Diversity**	**Unique (%)**	**SAS**
1 - Lws	0.12	90.19	77.78	0.79	63.49	1.98
4 - Lws	0.19	85.43	80.60	0.86	76.39	2.27
12 - Lws	**0.51**	**89.19**	**94.23**	**0.83**	**87.92**	**1.96**
17 - Lws	0.18	62.93	83.35	0.82	69.45	2.12
2 - Chev	0.39	91.67	85.55	0.85	43.42	2.38
4 - Chev	0.31	91.34	93.24	0.81	20.13	2.21
15 - Chev	0.21	69.18	77.63	0.76	75.4	2.01
20 - Chev	0.13	64.33	80.25	0.86	99.43	2.51
21 - Chev	0.48	76.92	86.53	0.81	85.61	2.51

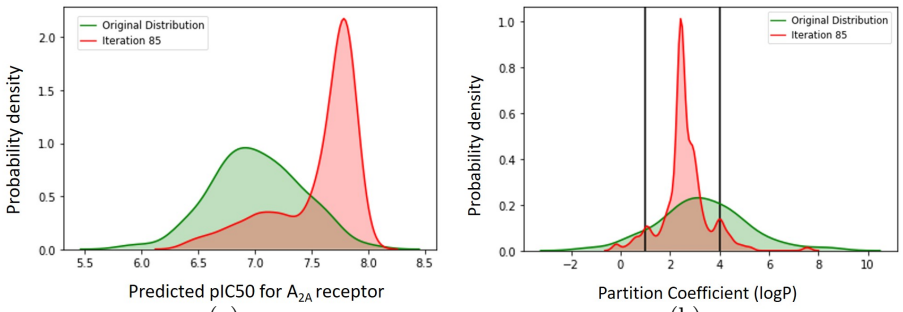
Table 1: Comparison of the non-dominated solutions obtained for each scalarization technique: Lws and Chev stands for linear weighted sum and Chebyshev, respectively.

From the analysis of the results, it is possible to identify a set of retrained generators that manage to generate molecules with interesting properties. In particular, the LWS scalarization technique with a weight of 0.55 for biological affinity is the one with the most appropriate trade-off for the different evaluated properties. In addition to a significant pIC_{50} difference, approximately 90% of the molecules fall within the optimal logP range. Furthermore, the Tanimoto diversity, as well as the rates of diversity and uniqueness, indicate that this model manages to generate synthesizable molecules with novelty and syntactically valid. Figure 3 shows the direct comparison between the molecules generated by the simple generator and the model highlighted in Table 1. The results demonstrate that despite the two objectives being competing, it was possible to identify a compromise solution that adequately satisfies both. In practice, the set of molecules obtained has a high probability of inhibiting the receptor and having a logP within the desired range. As a result, it is possible to affirm that most of the molecules are within the desired range of these properties, which are crucial for the lead compounds to be promising.

4 Conclusion

In real-world applications like lead optimization, it is often desired to optimize several different properties simultaneously. For example, we may want to optimize the selectivity of a drug while keeping the solubility in a specific range. This framework can tackle this problem by applying multiobjective rein-

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.



(a) (b)

Fig. 3: Comparison between simple and re-trained Generators regarding the biological affinity (a) and percentage of molecules inside the desired logP range (b).

forcement learning and considering different conflicting properties of a candidate molecule in order to be processed in the further drug discovery process.

References

[1] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung. Deep learning for molecular design a review of the state of the art. *Molecular Systems Design & Engineering*, 4(4):828–849, 2019.

[2] L. Di and E. H. Kerns. *Drug-like properties: concepts, structure design and methods from ADME to toxicity optimization*. Academic press, 2015.

[3] J. BO Mitchell. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(5):468–481, 2014.

[4] C. A. Nicolaou, J. Apostolakis, and C. S. Pattichis. De novo drug design using multiobjective evolutionary graphs. *Journal of Chemical Information*, 2009.

[5] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276, feb 2018.

[6] E. Putin, A. Asadulaev, Q. Vanhaelen, Y. Ivanenkov, A. V. Aladinskaya, A. Aliper, and A. Zhavoronkov. Adversarial Threshold Neural Computer for Molecular de Novo Design. *Molecular Pharmaceutics*, 15(10):4386–4397, oct 2018.

[7] Y. Li, L. Zhang, and Z. Liu. Multi-objective de novo drug design with conditional graph generative model. *Journal of Cheminformatics*, 10(1):33, dec 2018.

[8] N. Ståhl, G.. Falkman, A. Karlsson, G. Mathiason, and J. Boström. Deep Reinforcement Learning for Multiparameter Optimization in de novo Drug Design. *Journal of Chemical Information and Modeling*, 59(7):3166–3176, jul 2019.

[9] M. Popova, O. Isayev, and A. Tropsha. Deep reinforcement learning for de novo drug design. *Science Advances*, 4(7):eaap7885, jul 2018.

[10] M. Olivecrona, O. Blaschke, and H. Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, dec 2017.

[11] Z. Zhou, S. Kearnes, L. Li, R. N. Zare, and P. Riley. Optimization of Molecules via Deep Reinforcement Learning. *Scientific Reports*, 9(1):10752, dec 2019.

[12] K. Van Moffaert, M. M. Drugan, and A. Nowe. Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 191–199, 2013.

[13] T. T. Nguyen, N. D. Nguyen, P. Vamplew, S. Nahavandi, R. Dazeley, and C. P. Lim. A multi-objective deep reinforcement learning framework. *Engineering Applications of Artificial Intelligence*, 96:103915, 2020.

[14] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

[15] G. Landrum et al. Rdkit: Open-source cheminformatics. *https://www.rdkit.org/*, 2006.