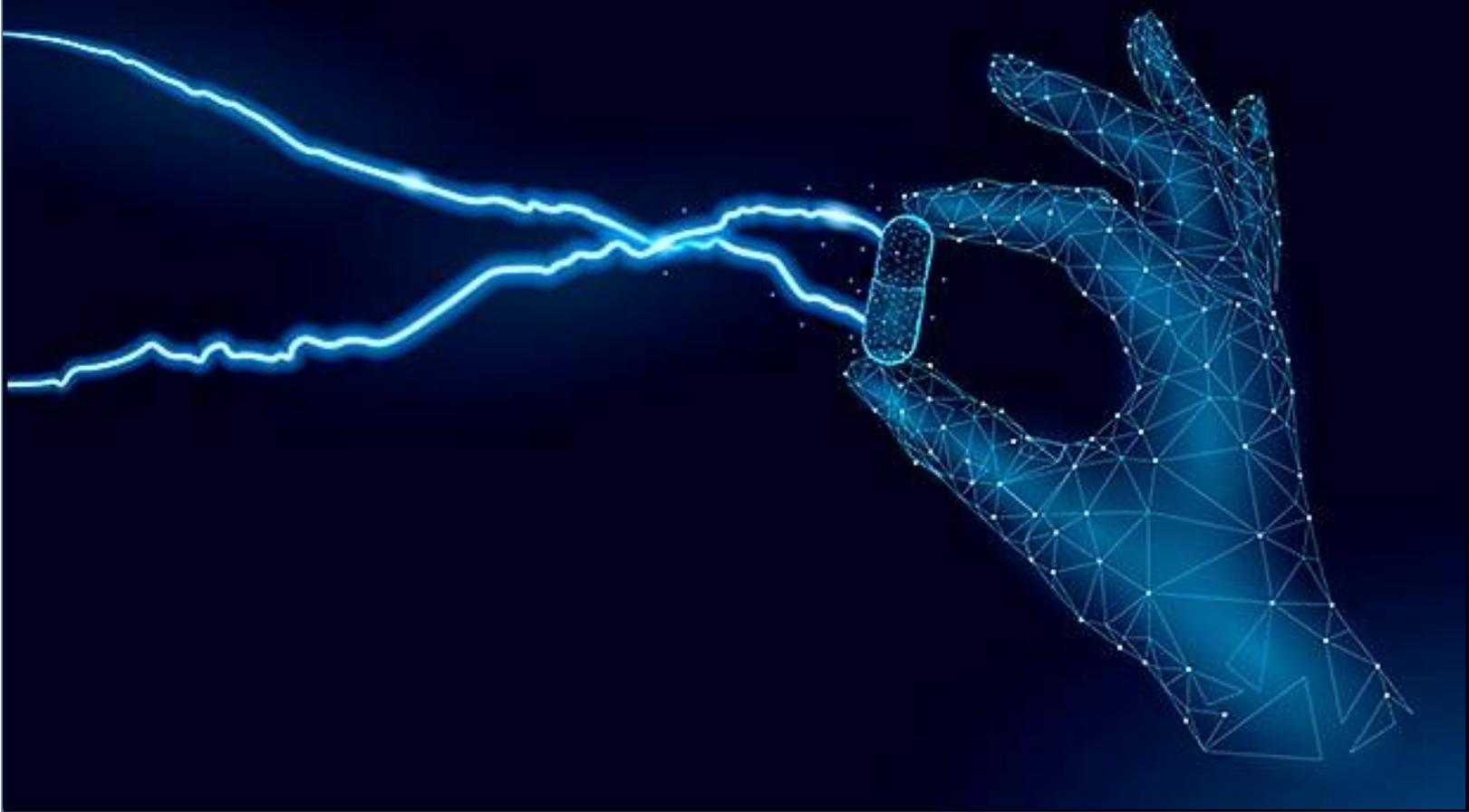
# To explore drug space smarter: artificial intelligence in drug design for G protein-coupled receptors

Liu, X.

# Chapter 7

Conclusions and future perspectives



Having provided a review about computational approaches for *de novo* drug design and four research projects in the previous chapters, I am well versed in cutting-edge AI technologies, especially deep learning, applied in different scenarios of *de novo* drug design. In the following paragraphs, I will draw conclusions of this thesis and give a future outlook to illustrate its appropriateness in drug discovery and to bring forward other promising scopes for its application.

## 7.1. Conclusions

Drug discovery is a time- and resource-consuming process. To this end, computational approaches that are applied in *de novo* drug design play an important role to improve the efficiency and decrease the costs to develop novel drugs. Over several decades, a variety of methods have been proposed and applied in practice [1]. Traditionally, drug design problems are always taken as the combinational optimization in discrete chemical space, such as evolutionary algorithms [2,3], heuristic search algorithms [4], simulated annealing algorithms [5], *etc.*. Hence optimization methods were exploited to search for new drug molecules that meet multiple objectives. With the accumulation of data and the development of machine learning methods, computational drug design methods have gradually shifted to a new paradigm. There has been particular interest in the potential application of deep learning methods to drug design [6]. In **Chapter 2**, we gave a brief description of these two different *de novo* methods, compared their application scopes and discussed their possible development in the future.

Over the last ten years deep learning has progressed tremendously in both image recognition, natural language processing and other data rich fields [7]. In drug discovery, recurrent neural networks (RNNs) have been shown to be an effective method to generate novel chemical structures in the form of SMILES [8]. However, ligands generated by current methods have so far provided relatively low diversity and do not fully cover the whole chemical space occupied by known ligands. In **Chapter 3**, we therefore propose a new method (*DrugEx*) to discover *de novo* drug-like molecules. *DrugEx* is an RNN model (generator) trained through a special exploration strategy integrated into reinforcement

learning. As a case study we applied our method to design ligands for the adenosine $A_{2A}$ receptor. From ChEMBL data, a machine learning model (predictor) was created to predict whether generated molecules are active or not. Based on this predictor as the reward function, the generator was trained by reinforcement learning without any further data. We then compared the performance of our method with two previously published methods, *REINVENT* [9] and *ORGANIC* [10]. We found that the candidate molecules our model designed and predicted to be active, had a larger chemical diversity and better covered the chemical space of known ligands compared to the state-of-the-art (SOTA).

Although deep learning has led to breakthroughs in drug discovery, most of its applications only focus on a single drug target to generate drug-like active molecules. This is in spite of the reality that drug molecules often interact with more than one target which can have desired (polypharmacology) or undesired (toxicity) effects. In polypharmacology ideal drugs are required to bind to multiple specific targets to enhance efficacy or to reduce the development of resistance [11]. In **Chapter 4**, we extended our *DrugEx* algorithm with multi-objective optimization to generate drug molecules towards multiple targets or one specific target while avoiding off-targets (the two adenosine receptors, $A_1$AR and $A_{2A}$AR, and the potassium ion channel hERG). In our model, we applied an RNN as the *agent* and machine learning predictors as the *environment*, both of which were pre-trained in advance and then interplayed under the reinforcement learning framework. The concept of evolutionary algorithms was merged into our method such that *crossover* and *mutation* operations were implemented by the same deep learning model as the *agent*. During the training loop, the agent generates a batch of SMILES-based molecules. Subsequently scores for all objectives provided by the *environment* are used for constructing Pareto ranks of the generated molecules with non-dominated sorting and Tanimoto-based crowding distance algorithms. Here, we adopted GPU acceleration to speed up the process of Pareto optimization. The final reward of each molecule is calculated based on the Pareto ranking with the ranking selection algorithm [12]. The agent is trained under the guidance of the reward to make sure it can generate more desired molecules after convergence of the training process. All in all we demonstrated the generation of compounds with a diverse

predicted selectivity profile toward multiple targets, offering the potential of high efficacy and lower toxicity.

Due to the huge chemical space in which feasible drug-like molecules are searched for, rational drug design always starts from specific molecular scaffolds as the core to which side chains are added or modified. With the rapid growth of deep learning methods and their application in drug discovery, a variety of approaches has been developed for *de novo* drug design. However, earlier versions of *DrugEx* are trained under fixed objectives and do not allow users to input any prior information, like most goal-directed methods. In order to improve its generality, *DrugEx* was updated to design drug molecules based on multiple scaffolds given by users. In **Chapter 5** we extended the transformer model [13], which is a multi-head self-attention deep learning model containing an encoder and a decoder, to deal with each molecule as a graph. The encoder of the graph transformer receives the input graph of the scaffolds containing multiple fragments and its decoder outputs the graph-based molecule containing given scaffolds. Each molecule was generated by growing and connecting procedures for the fragments in given scaffolds that were unified into one model. Moreover, we trained this generator under the reinforcement learning framework to increase the number of active ligands. As proof our proposed method was applied to design adenosine $A_{2A}$ receptor ligands which were compared with SMILES-based methods. The results demonstrated its effectiveness as most of the generated molecules contained the given scaffolds and had a high virtual affinity towards the adenosine $A_{2A}$ receptor.

Despite the rapid growth of AI techniques in drug discovery, widespread adoption of new *de novo* drug design approaches in the fields of medicinal chemistry and chemical biology is still lagging behind the most recent developments. It is urgently needed to establish a close collaboration between diverse teams of experimental and theoretical scientists. To accelerate the adoption of both modern and traditional *de novo* molecular generators, we developed *GenUI* (Generator User Interface), a software platform that makes it possible to integrate molecular generators within a feature-rich graphical user interface that is easy to use by experts of varying backgrounds. *GenUI* is implemented as a web service and its

interfaces offer access to cheminformatics tools for data preprocessing, model building, molecule generation, and interactive chemical space visualization. Moreover, the platform is easy to extend with customizable frontend React.js components and backend Python extensions. *GenUI* is open source which has integrated *DrugEx* as a proof of principle. In **Chapter 6**, we presented the architecture and implementation details of GenUI and discuss how it can facilitate collaboration in the disparate communities interested in *de novo* molecular generation and computer-aided drug discovery.

## 7.2. Further perspectives

With the four projects mentioned above we catch a glimpse of the overwhelming power of AI in drug *de novo* design. However, it is impossible to make a thorough investigation of its capability in every scope of drug discovery with only four years study. In my view, there are still a plethora of promising issues about the development and application of AI to design chemical compounds that attract researchers' interest and are worth addressing.

### 7.2.1. New AI technologies

Deep learning is the most attractive branch in AI and it is still growing rapidly. First convolutional neural networks and recurrent neural networks achieved a breakthrough in image recognition and natural language processing [7]. Consequently the transformer model was proposed based on a self-attention mechanism in 2017 and achieved SOTA performance in language processing [13]. Subsequently, a large number of variants have been developed. For example, BERT, which is the encoder part of the transformer and is pre-trained with large amounts of data, improved the performance of sequence data prediction dramatically [14]. This led to more and more researchers employing it to construct predictive models for biological and chemical data [15,16]. In addition, GPT-3, which is also derived from the transformer model, achieved SOTA performance in many sequence generation tasks [17]. Moreover, transformer-based methods can also deal with graph data [18], allowing it to be applied to graph-based molecular design. Therefore they are promising algorithms to be used in drug design.

With respect to the huge number of parameters in the complicated architectures of deep learning, there are also many new methods to effectively train these models. For example, when dealing with image generation, generative adversarial networks [10] and variational autoencoders [19] are commonly used to train the model to generate the most similar samples. When introducing different computational methods in drug design in **Chapter 2**, we discussed the possibility of the combination of deep learning methods and optimization methods. Afterwards, we proposed a new kind of training method through simulating the idea of evolutionary algorithms **in Chapter 4**. Moreover, there are many studies about the application of evolutionary algorithms [20] or Bayesian optimization [21] to update the parameters in deep learning models. With the architecture of deep learning becoming more and more complex, it is worth discussing about how to effectively train models to avoid the issues of the local minimum and overfitting.

### 7.2.2. Different constraint conditions

Besides the objectives mentioned in the previous chapters, such as affinity for adenosine receptors and the drug-likeness score, the ideal drug molecule also needs to meet more objectives in reality. In addition to the affinity for one or more given targets, it also needs to have qualified ADME (absorption, distribution, metabolism, and excretion) properties [22] and low toxicity. More specifically, some of these requirements can be conflicting and cannot be satisfied simultaneously. Therefore, an important issue is to orchestrate the many objectives for effective drug design. However, most of current studies just simply transform the multi-objectives into a single objective with the weighted sum of these scores in order to guide the training of deep learning models. Actually, there are plenty of multi-objective optimization methods [12] being developed as mentioned in **Chapter 2**. These methods are worth exploring their integration with deep learning models.

Another important property of generated drug molecules is their synthesizability. However, the most current SMILES-based and Graph-based models cannot directly guarantee that the generated molecules can be synthesized [23]. Therefore, it is critical to predict the synthesizability of these generated molecules, which determines if they could be

experimentally tested in practice. For example, some researchers combined deep reinforcement learning and Monto Carlo tree search to put forward to methods to predict retrosynthesis score [24,25] for given molecules and provide the feasible synthetic schemes [26]. Moreover, some other groups directly generate molecule base on reaction, in which each molecule in the training set are decomposed as a reaction tree [27]. And the aim of the model is choosing the reaction from the library step by step. In the end, the molecule construct with the whole reaction tree is generated.

In **Chapter 3 & 4**, all of the model conditions were fixed. This allowed the model to be trained well, but it cannot interact with users by receiving continued and updated information. If the conditions are changed, the model has to be trained again, which is an inconvenient and time-consuming process. In order to improve the generality of the model we proposed a new method in **Chapter 5** in which an end-to-end model received scaffold information from users. General speaking, it can also take other information as input to design bioactive molecules conditionally. For example, it can be used for lead optimization, *i.e.* the input can be an inactive or toxic ligand, and the output should be a similar ligand but active or safe, respectively. Moreover, now that proteochemometric modelling (PCM) has been proposed for many years to take the information of both drug and target information as input and predict their affinity [28], it can also be used to construct inverse PCM models, which take protein information as input to design its active ligands [29]. Considering that the full sequence length of some proteins is too large to be dealt with by current deep learning models, protein descriptors can also be used as input information.

### 7.2.3. Designing various kind of molecules

In this thesis, we only focus on the generation of small organic molecules, but there are other biological/chemical molecules to be designed. For example, natural products have always been the effective components of traditional Chinese medicine, but their physico-chemical properties are distinct from classical drug molecules. For instance unlike small synthetic molecules most of the natural products do not adhere to the Rule of 5 [30]. Compared with classical drug molecules, natural products also have different advantages

as drug candidates. Natural products have been optimized by long-term natural evolution to have particular bioactivities, including the regulation of endogenous defense mechanisms through the interaction with other organisms, which is the possible reason for its key role in therapeutic areas especially for infectious diseases and cancer [31]. Moreover, their use in traditional medicine may provide insights regarding efficacy and safety, covering a wider area of chemical space compared with small organic molecules [32]. Now that there are several AI methods for the retrosynthesis of organic molecules [26], they also provide a valuable direction to exploit these methods in the synthesis pathway prediction of natural products.

Besides small organic molecules, peptides and proteins are important macromolecules for medicine. For example, some antimicrobial peptides can be used as drugs to inhibit the growth of a variety of microbes. The data representation of a peptide is a sequence of amino acid residues, which is feasible to be designed with deep learning models [33]. Moreover, there are variable domains in Fab regions of antibodies which determine specificity and efficacy to recognize the antigen. This part of the antibody also needs to be designed and can be generated by AI methods [34].

## 7.3. Final notes

The main thrust of this thesis is a comprehensive study about the application of AI technologies in *de novo* drug design. An integrative Python-based toolkit named *DrugEx* was developed to facilitate the accessibility of our methods to other researchers. In order to decrease the threshold for experimental researchers who are not familiar with computer coding, this tool was also used as the engine integrated into a web-based graphic toolkit named *GenUI* which has powerful capabilities of interactions with users and developers. These two software packages are my main contributions to the scientific community. Generally speaking, the highlight of this thesis is sufficiently embodied on the cover page. Faced with the huge chemical space of drug-like molecules (unveiling of the capsule at the bottom), AI is an effective approach to rapidly narrow down the search scope. AI itself is a mimic of the human brain running *in silico* (the logo in the center). The chip located in

the center of the brain consists of a variety of different electronic components. Seven tandem diodes resemble the protein structure of a GPCR which has seven transmembrane domains. Its intracellular domain with the G protein (a total of four subunits represented by four gears) forms a virtual document which recodes with digits if the GPCR is activated or not. The component in the lower right side is like a magnifying glass that is identifying the active ligands after exploring the huge chemical space with this virtual lab. I hope the readers could be beneficial from this thesis to have broad and deep understanding of the role that AI methods play in drug discovery.

# Reference


1. Schneider G, Fechner U (2005) Computer-based de novo design of drug-like molecules. Nat Rev Drug Discov 4 (8):649-663. doi:10.1038/nrd1799
2. van der Horst E, Marques-Gallego P, Mulder-Krieger T, van Veldhoven J, Kruisselbrink J, Aleman A, Emmerich MT, Brussee J, Bender A, Ijzerman AP (2012) Multi-objective evolutionary design of adenosine receptor ligands. Journal of chemical information and modeling 52 (7):1713-1721. doi:10.1021/ci2005115
3. Lameijer EW, Kok JN, Back T, Ijzerman AP (2006) The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules. Journal of chemical information and modeling 46 (2):545-552. doi:10.1021/ci050369d
4. Gillet VJ, Newell W, Mata P, Myatt G, Sike S, Zsoldos Z, Johnson AP (1994) SPROUT: recent developments in the de novo design of molecules. J Chem Inf Comput Sci 34 (1):207-217. doi:10.1021/ci00017a027
5. Sengupta S, Bandyopadhyay S (2012) De novo design of potential RecA inhibitors using multi objective optimization. IEEE/ACM Trans Comput Biol Bioinform 9 (4):1139-1154. doi:10.1109/TCBB.2012.35
6. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug discovery today. doi:10.1016/j.drudis.2018.01.039
7. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521 (7553):436-444. doi:10.1038/nature14539
8. Segler MHS, Kogej T, Tyrchan C, Waller MP (2018) Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Cent Sci 4 (1):120-131. doi:10.1021/acscentsci.7b00512
9. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. Journal of cheminformatics 9 (1):48. doi:10.1186/s13321-017-0235-x
10. Benjamin S-L, Carlos O, Gabriel L. G, Alan A-G (2017) Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). doi:10.26434/chemrxiv.5309668.v3
11. Chaudhari R, Tan Z, Huang B, Zhang S (2017) Computational polypharmacology: a new paradigm for drug discovery. Expert Opin Drug Discov 12 (3):279-291. doi:10.1080/17460441.2017.1280024
12. Emmerich MTM, Deutz AH (2018) A tutorial on multiobjective optimization: fundamentals and evolutionary methods. Nat Comput 17 (3):585-609. doi:10.1007/s11047-018-9685-y
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin IJae-p (2017) Attention Is All You Need.arXiv:1706.03762
14. Devlin J, Chang M-W, Lee K, Toutanova KJae-p (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.arXiv:1810.04805
15. Chithrananda S, Grand G, Ramsundar BJae-p (2020) ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction.arXiv:2010.09885
16. Wang S, Guo Y, Wang Y, Sun H, Huang J (2019) SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. Paper presented at the Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Niagara Falls, NY, USA,
17. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C,

McCandlish S, Radford A, Sutskever I, Amodei DJae-p (2020) Language Models are Few-Shot Learners.arXiv:2005.14165

18. Yun S, Jeong M, Kim R, Kang J, Kim HJJae-p (2019) Graph Transformer Networks.arXiv:1911.06455
19. Gomez-Bombarelli R, Wei JN, Duvenaud D, Hernandez-Lobato JM, Sanchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A (2018) Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Cent Sci 4 (2):268-276. doi:10.1021/acscentsci.7b00572
20. Young SR, Rose DC, Karnowski TP, Lim S-H, Patton RM (2015) Optimizing deep learning hyper-parameters through an evolutionary algorithm. Paper presented at the Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments, Austin, Texas,
21. Griffiths RR, Hernandez-Lobato JM (2020) Constrained Bayesian optimization for automatic chemical design using variational autoencoders. Chem Sci 11 (2):577-586. doi:10.1039/c9sc04026a
22. Kirchmair J, Goller AH, Lang D, Kunze J, Testa B, Wilson ID, Glen RC, Schneider G (2015) Predicting drug metabolism: experiment and/or computation? Nat Rev Drug Discov 14 (6):387-404. doi:10.1038/nrd4581
23. Liu X, IJzerman AP, van Westen GJP (2021) Computational Approaches for De Novo Drug Design: Past, Present, and Future. Methods Mol Biol 2190:139-165. doi:10.1007/978-1-0716-0826-5_6
24. Genheden S, Thakkar A, Chadimova V, Reymond JL, Engkvist O, Bjerrum E (2020) AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. Journal of cheminformatics 12 (1):70. doi:10.1186/s13321-020-00472-1
25. Thakkar A, Chadimova V, Bjerrum EJ, Engkvist O, Reymond JL (2021) Retrosynthetic accessibility score (RAscore) - rapid machine learned synthesizability classification from AI driven retrosynthetic planning. Chem Sci 12 (9):3339-3349. doi:10.1039/d0sc05401a
26. Segler MHS, Preuss M, Waller MP (2018) Planning chemical syntheses with deep neural networks and symbolic AI. Nature 555 (7698):604-610. doi:10.1038/nature25978
27. Ghiandoni GM, Bodkin MJ, Chen B, Hristozov D, Wallace JEA, Webster J, Gillet VJ (2020) Enhancing reaction-based de novo design using a multi-label reaction class recommender. J Comput Aided Mol Des 34 (7):783-803. doi:10.1007/s10822-020-00300-6
28. van Westen GJ, Wegner JK, Geluykens P, Kwanten L, Vereycken I, Peeters A, Ijzerman AP, van Vlijmen HW, Bender A (2011) Which compound to select in lead optimization? Prospectively validated proteochemometric models guide preclinical development. PLoS One 6 (11):e27518. doi:10.1371/journal.pone.0027518
29. Grechishnikova D (2021) Transformer neural network for protein-specific de novo drug generation as a machine translation problem. Sci Rep 11 (1):321. doi:10.1038/s41598-020-79682-4
30. Mullard A (2018) Re-assessing the rule of 5, two decades on. Nat Rev Drug Discov 17 (11):777. doi:10.1038/nrd.2018.197
31. Atanasov AG, Waltenberger B, Pferschy-Wenzig EM, Linder T, Wawrosch C, Uhrin P, Temml V, Wang L, Schwaiger S, Heiss EH, Rollinger JM, Schuster D, Breuss JM, Bochkov V, Mihovilovic MD, Kopp B, Bauer R, Dirsch VM, Stuppner H (2015) Discovery and resupply of pharmacologically active plant-derived natural products: A review. Biotechnol Adv 33 (8):1582-1614. doi:10.1016/j.biotechadv.2015.08.001
32. Atanasov AG, Zotchev SB, Dirsch VM, International Natural Product Sciences T, Supuran CT (2021) Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov 20 (3):200-216. doi:10.1038/s41573-020-00114-z
33. Wang C, Garlick S, Zloh M (2021) Deep Learning for Novel Antimicrobial Peptide Design.

Biomolecules 11 (3). doi:10.3390/biom11030471

34. Saka K, Kakuzaki T, Metsugi S, Kashiwagi D, Yoshida K, Wada M, Tsunoda H, Teramoto R (2021) Antibody design using LSTM based deep generative model from phage display library for affinity maturation. Sci Rep 11 (1):5852. doi:10.1038/s41598-021-85274-7