Disease Centric Large Scale De Novo Design of Drug Candidate Molecules with Graph Generative Deep Adversarial Networks

Atabey Ünlü[1,2], Elif Çevrim[1,2], Ahmet Sarıgün[1,3], Heval Ataş[1,4], Altay Koyaş[4], Hayriye Çelikbilek[1], Deniz Cansen Kahraman[4], Abdurrahman Olğaç[5], Ahmet Rifaioğlu[6], Tunca Doğan[1,2,*]

[1]Biological Data Science Lab, Dept. of Computer Engineering, Hacettepe University, [2]Dept. of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, [3]Dept. of Chemistry, Middle East Technical University, [4]Cancer Systems Biology Lab, Graduate School of Informatics, Middle East Technical University, [5]Dept. of Pharmaceutical Chemistry, Faculty of Pharmacy, Gazi University, [6]Saez-Rodriguez Group, Institute for Computational Biomedicine, Heidelberg University, Germany; *To whom the correspondence should be addressed (tuncadogan@gmail.com)

Discovering novel drug candidate molecules is one of the most fundamental and critical steps in drug development. It is especially challenging to develop new drug-based treatments for complex diseases, such as various cancer subtypes, which have heterogeneous structure and affect multiple biological mechanisms. With the advancements in high-throughput screening technology, it is now possible to scan thousands of compounds simultaneously; but still, it is impossible to fully analyze the target and compound spaces due to the excessive number of protein-compound combinations. Furthermore, it is possible to design approximately 10^{60} small molecules that differ from each other by at least one atom or bond, indicating the nearly limitless potential of the theoretical space of drug-like molecules. Generative deep learning models, which create new data points according to a probability distribution at hand, have been developed with the purpose of picking completely new samples from a distribution space that is only partially known.

In this study, we propose a novel computational system, DrugGEN, for de novo generation of single and multi-target drug candidate molecules intended for specific drug resistant diseases, by constructing a new deep learning architecture that leverages the transformer architecture and graph neural networks in a generative adversarial setting (Figure 1a). The DrugGEN system optimizes two main processes: creation of a new molecule and transforming it to target a selected protein. To this end, we developed a two-fold end-to-end model that takes graph representations of small molecules and target proteins as input to stacked generative adversarial networks – sGAN (composed of 2 modules: GAN1 and GAN2), and outputs de novo drug candidate molecules specific to the given target proteins (Figure 1a). The main goal of GAN1 is learning molecular properties of drug-like small molecules, such as how the atoms and bonds should be arranged for a molecule to be chemically synthesizable and physically stable. GAN1 is composed of a transformer encoder-based generator and a graph convolutional network-based discriminator. Transformer encoder takes random gaussian noise as input and transforms it to a graph in the form of separate annotation and adjacency matrices (Figure 1b). These generated graphs (representing de novo molecules), are then fed to the discriminator along with real molecules, in which graph convolution and graph aggregation operations are applied to predict whether a data point is generated by the model or belongs to the real molecules set. The aim of the GAN2 module is modifying the previously generated de novo molecules to effectively bind to the selected target. GAN2 is composed of a transformer decoder (generator) (Figure 1c) and a graph convolutional (discriminator) network. Here, the transformer decoder architecture was re-designed to process both protein and small molecule feature graphs, creating a pseudo-interaction module. Final molecular products are sampled and sent to the GAN2 discriminator to be compared with the known inhibitors of the protein of interest.

The system was trained using all molecules in the ChEMBL database (~2M) and known inhibitors of the selected target protein (AKT1), in GAN1 and 2, respectively, to produce novel and effective inhibitory molecules against the hepatocellular carcinoma (HCC) disease, which is a deadly sub-type of liver cancer. The hyper-parameter values of the system were optimized via multiple rounds of training/validation experiments. Generated molecules were monitored based on their synthetic accessibility and quantitative estimation of drug-likeness. The overall model evaluation was done based on the percentage of valid molecules generated by the model, together with their uniqueness and novelty. All the metrics/scores were calculated using the RDkit library to keep results reproducible. The finalized system was run to design thousands of novel AKT1 inhibitors (Figure 1d, right hand-side). The resulting de novo molecule records are being evaluated by medicinal chemists, which will be followed by chemical synthesis of selected molecules and their utilization in wet-lab (in vitro) experiments for validating their inhibitory effects on drug resistant HCC cell lines. If the expected results are obtained, new drug candidate compounds of critical importance will be discovered for the treatment of HCC, and pre-clinical and clinical studies will be planned for future. DrugGEN has been developed as a generic system that can easily be used to design new molecules for other targets and diseases. All of the datasets, source code, results and pre-trained models of DrugGEN are freely available at https://github.com/HUBioDataLab/DrugGEN.

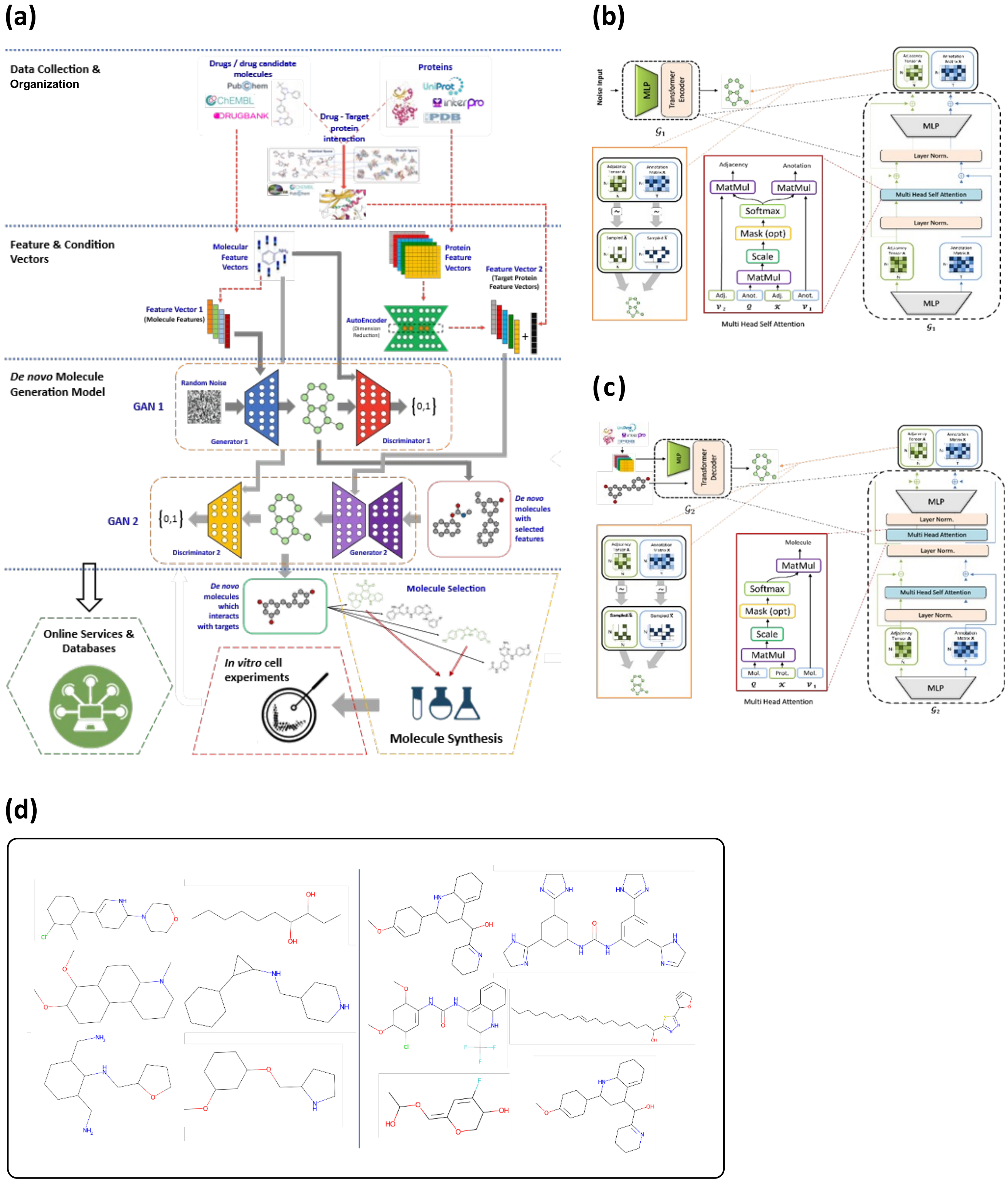

Figure 1. (a) Schematic representation of the DrugGEN system and the overall project, **(b)** detailed chart of the transformer encoder-based generator network in GAN1, **(c)** detailed chart of the transformer decoder-based generator network in GAN2, **(d)** valid and novel *de novo* molecule examples designed by the DrugGEN model (left: GAN1 output, right: GAN2 output for targeting AKT1).