Application of semi-supervised GAN in combination with JT-VAE for generation of small molecules with high binding affinity to the KasA enzyme of *Mycobacterium tuberculosis*

Anna V. Gonchar
United Institute of Informatics Problems
National Academy of Sciences of Belarus
Minsk, Republic of Belarus
raphaelkyzy@gmail.com

Alexander V. Tuzikov
United Institute of Informatics Problems
National Academy of Sciences of Belarus
Minsk, Republic of Belarus
tuzikov@newman.bas-net.by

Konstantin V. Furs
United Institute of Informatics Problems
National Academy of Sciences of Belarus
Minsk, Republic of Belarus
ky6ujlo@gmail.com

Alexander M. Andrianov
Institute of Bioorganic Chemistry
National Academy of Sciences of Belarus
Minsk, Republic of Belarus
alexande.andriano@yandex.ru

***Abstract*—Semi-supervised generative adversarial neural network trained on molecular graph embeddings produced by Junction Tree Variational Autoencoder was implemented and applied for *de novo* design of new potential inhibitors of *Mycobacterium tuberculosis* protein KasA.**

Keywords—Mycobacterium tuberculosis (Mtb), KasA, generative adversarial neural network (GAN), semi-supervised learning, graph embeddings, virtual screening, molecular docking

I. Introduction

Over the years, the number of different types of data is growing and computer science has started to play an important role almost in every branch of science. Creation of the Protein Data Bank (PDB) (https://www.rcsb.org/) containing experimentally-determined 3D structures of proteins and nucleic acids, and different databases of small molecules like ZINC, DrugBank, PubChem, etc. has played an important role in the field of drug discovery, especially in Computer-Aided Drug Design (CADD). The term CADD describes a multi-disciplinary approach for a rational design of new chemical compounds and includes a vast variety of methods which use computational technologies to facilitate and accelerate the process of developing novel drug candidates [1]. Along with conventional molecular modeling methods, such as virtual screening, molecular docking and molecular dynamics, machine learning (ML) and its subfield deep learning (DL) have been getting more and more popular in CADD. Due to increasing availability of different biological data, quality and completeness, *de novo* design of molecules using deep generative models in combination with structure-based molecular modelling techniques has a great potential in the area of drug design ad discovery [1].

The main goal of this study is *de novo* design of small molecules potentially active against the KasA enzyme, β-ketoacyl-acyl carrier protein synthase I of *Mycobacterium tuberculosis* (*Mtb*) which plays an important role in mycobacterium cell wall biosynthesis. The loss of the KasA activity results in the *Mtb* cell lysis, testifying that this enzyme is a valuable target for the design of novel potent antitubercular inhibitors [2].

To solve the problem, we proposed a combined application of two generative networks, namely Junction Tree Variational Autoencoder (JT-VAE) [3] and semi-supervised generative adversarial neural network (SGAN). The JT-VAE model was used to produce graph embeddings which were used for the SGAN training. Molecules in the training dataset were separated into two groups according to their binding energy values to the target protein, low and high, calculated using molecular docking tools. This allowed SGAN to generate new molecules similar to those from the preferred group with the low values of binding energy.

II. Materials and Methods

A. JT- VAE

JT-VAE [3] is a VAE architecture that operates with molecular graph structures of compounds using specific encoder and decoder, while most other methods employ the SMILES representation [4]. The encoder is a neural network used to calculate a latent representation of a molecule in the continuous, high-dimensional latent space, and the decoder is a neural network used to decode a compound from coordinates in this space. In VAEs, the entire encoding-decoding process is stochastic.

In JT-VAE, each molecule is considered as a set of valid chemical substructures that are chosen from the component vocabulary which is formed from the training dataset [3]. These components are used as building blocks for a molecule during both encoding and decoding processes. Based on the components for each molecule, a junction tree scaffold is built with the specific decomposition algorithm. The original molecular graph and its associated junction tree are two complementary representations of a molecule [3]. The resulting latent vector of a compound is the latent vector of the molecular graph concatenated with the latent vector of the junction tree.

It is well known that JT-VAE generates only valid molecules owing to the component-by-component encoding-decoding approach, whereas the SMILES-based generative neural networks also produce invalid compounds.

B. S GAN

GAN consists of two neural networks, generator and discriminator that are trained simultaneously. Generator tries to generate data similar to those from a training dataset, while discriminator attempts to distinguish between real and fake data. GAN is a class of unsupervised algorithms, since explicit labels of any class of data are not used except for implicit “fake” and “real” labels.

SGAN is a GAN with additional class for discriminator allowing one to distinguish “low” or “high” values of binding free energy calculated by molecular docking methods. In this work, molecules exhibiting values of binding free energy lower than −8.2 kcal/mol were assigned to the “low” energy class and the others to “high”. The architecture of SGAN is shown in Fig. 1 where BN means batch normalization.

Our approach extends SGAN to molecular embeddings of the latent space of JT-VAE.

C. Target protein structure and control inhibitors

Since the current study was aimed at *de novo* designing potential inhibitors of KasA, structures of this protein and several known inhibitors which would serve as a positive control were needed. Two inhibitors of the catalytic activity of KasA, thiolactomycin-based analog TLM5 [5] and platensimycin [6], were used in the calculations as a positive control. TLM5 is a slow-onset inhibitor that interacts preferentially with the KasA acyl-enzyme form [5] which has not been deposited in PDB yet. However, it was shown [7] that the acylated KasA intermediate can be imitated by the C171Q KasA mutant, as the mutation Cys-171-Gln leads to the structural changes in the enzyme active site mimicking the acylation of Cys171. The TLM5/C171Q KasA structure in the crystal (PDB ID: 4C72, https://www.rcsb.org/structure/4c72) was therefore used in this work.

D. Training set preparation

a) Pharmacophore-based Virtual Screening

To form the training dataset, pharmacophore-based virtual screening of three molecular libraries from a web-oriented platform Pharmit [8], namely Zinc15, ChemSpace, and ChemDiv, was carried out. Using this web tool, two pharmacophore models were built based on the complex of C171Q KasA with TLM5. Additionally, the crystal structure of C171Q KasA bound to TLM was also used (PDB ID: 4C6X, https://www.rcsb.org/structure/4c6x), resulting in one more pharmacophore model. A number of filters imposing restrictions on the physicochemical parameters of molecules which are commonly taken as the basic criteria of their ability to be effective when taken orally were used during the screening process (Table I). Using Python 3 and mostly its package for cheminformatics RDKit (https://www.rdkit.org), duplicates were removed from the dataset and canonical kekulized SMILES representations were obtained for each compound. After the pharmacophore-based screening, the total number of compounds in the training dataset was 58,000.

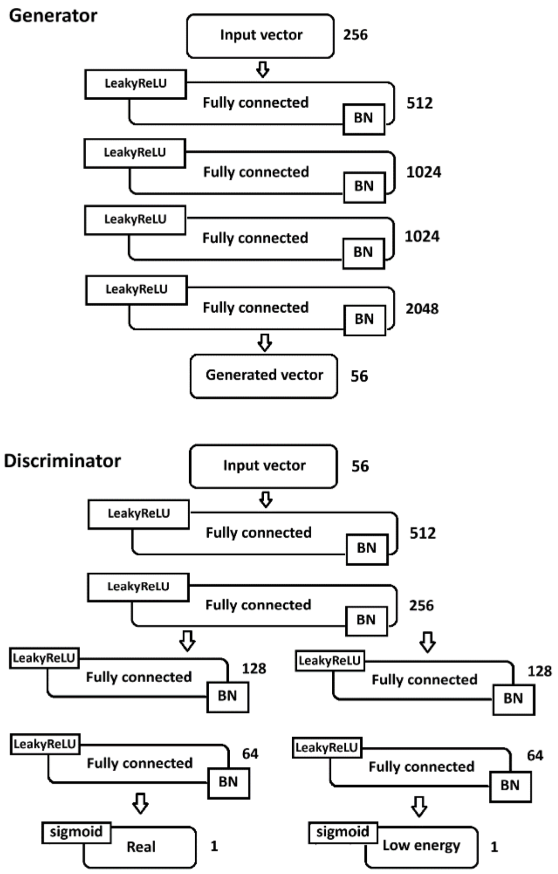

Fig. 1. The SGAN architecture

TABLE I. PHARMACOPHORE SEARCH FILTERS

M, Da	LogP	HBD	HBA	ΔG, kcal/mol	RMSD, Å
< 500	< 5	< 5	< 10	< 0	< 2

Footnote. The following notations are used: M – molecular weight, LogP – compound lipophilicity, HBD – number of H-bond donors, HBA – number of H-bond acceptors, **ΔG –** binding free energy, RMSD – the root mean squared deviation between the query features and the hit compound features [8].

b) Molecular Docking

To obtain the values of binding free energy for molecules from the training dataset, semi-flexible molecular docking of the unliganded C171Q KasA with the ligands was performed by the QuickVina2 program [9]. Structures of the enzyme and compounds from the training dataset were prepared for docking using the MGLTools program (https://ccsb.scripps.edu/mgltools/). The grid box for docking included the catalytic site of KasA with the following parameters: ΔX = 20.67 Å, ΔY = 24.8 Å, ΔZ = 16.46 Å centered at X = –7.24 Å, Y = –19.9 Å, Z = 6.75 Å. The value of the exhaustiveness parameter (i.e., the parameter defining the number of sampling performed by Vina) was set to 100. Distribution of the docking scores for the molecules from the training dataset is shown in Fig. 2.

E. SGAN training

The training dataset included 58,000 molecular graph embeddings from the latent space of JT-VAE. Number of epochs for training was 50. The probabilities of generator discriminator training were 0.8 and 0.2, respectively. Generator and discriminator loss functions G_{loss} and D_{loss} are given in formulas (1) and (2):

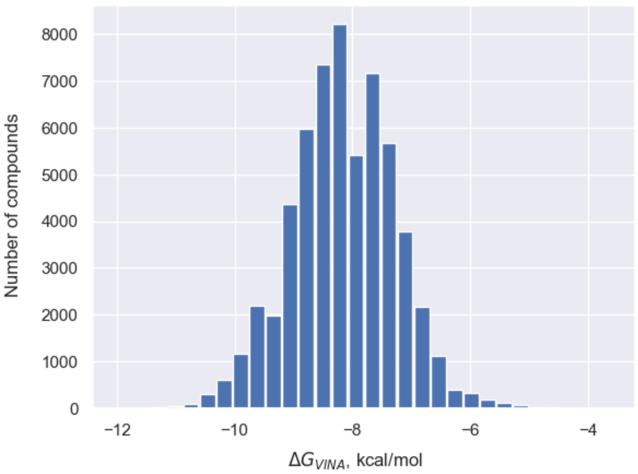

Fig. 2. Histogram of the distribution of binding free energy values calculated by molecular docking for molecules from the training dataset

$$G_{loss} = BCE\left(D_{out1}\left(G(noise)\right),1\right) + \\ + BCE\left(D_{out2}\left(G(noise)\right),1\right), \qquad (1)$$

$$D_{loss} = BCE\left(D_{out1}\left(G(noise)\right),0\right) + \\ + BCE\left(D_{out2}(real\_data), energy\_class\right), \qquad (2)$$

where D is the discriminator, D_{out1} is the output predicting the reality of the molecule, D_{out2} is the output predicting the energy class of the molecule, G is the generator, noise is the vector from the standard 256-dimensional Gaussian noise, *real_data* is the vector of the molecule from the training data, *energy_class* is 1 or 0 depending on whether the molecule belongs to the class of compounds with low binding energy or not.

The graphs of the loss functions of the generator and discriminator are shown in Fig. 3 with a solid blue and dashed red lines, respectively. Fig. 4 shows the discriminator predictions for 58,000 vectors from 56-dimensional standard Gaussian noise. In this Figure, the region of vectors that have a high probability of low binding energy values and are better than half the data in terms of the probability of being real is highlighted. We considered this region as the most promising for generation.

F. *Generation of new molecules*

50,000 vectors of length 56 were generated based on Gaussian noise and discriminator's predictions on them were obtained. To choose a promising region, the following conditions were set: the probability of data reality was above 15%, and the probability of the low energy class was above 80%. The number of molecular vectors that met the conditions was 1,395 and they were fed to JT-VAE decoder to get their SMILES representation. SMILES duplicates were then removed and the validity of generated molecules was checked. After this step, only 419 molecules left. In addition to sampling from promising region, the sampling from unpromising area was also conducted: in this case, the probability of real data was below 15%, and the probability of low energy class was below 30%. The number of molecules sampled was 9,413, but only 452 molecules were selected after decoding step, testifying that the vectors which are close in "Reality – Low energy" space are easily decoded into the same molecular graph representation.

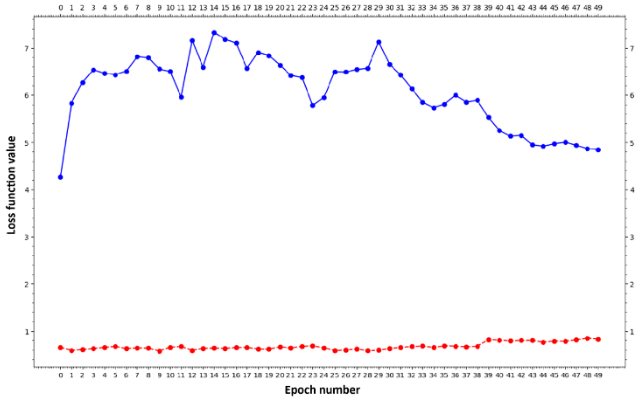

Fig. 3. Loss functions of the generator and discriminator for the developed neural network

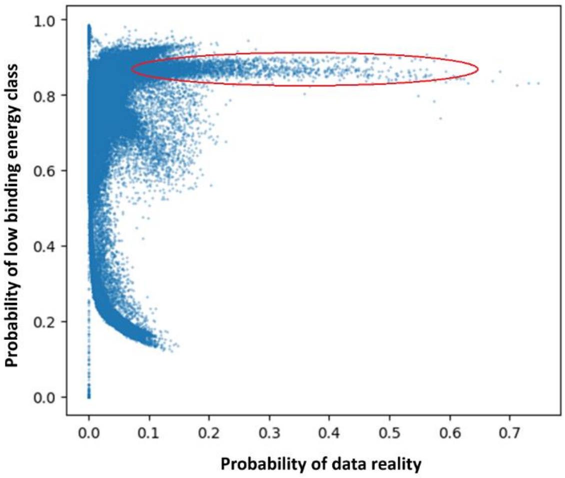

Fig. 4. SGAN discriminator predictions on vectors from Gaussian noise

G. *3D Structure Generation for new molecules*

3D structures of new generated molecules were derived from their 2D chemical sketches presented in the SMILES format. To do this, a stochastic algorithm for generating conformers ETKDG (Experimental-Torsion "basic Knowledge" Distance Geometry) [10, 11] from the RDKit package (https://www.rdkit.org) was used and the obtained conformations were optimized using the Merck molecular force field (MMFF) or the universal force field (UFF) in the cases when the MMFF optimization was unsuccessful.

III. Results and Discussion

To assess the SGAN model operation, molecular docking of the new generated molecules to the KasA enzyme was run with the computational protocol identical to the one used for compounds from the training dataset. As a result, 87% of molecules from the promising region showed the values of binding energy lower than –8.2 kcal/mol. For all these compounds, the average of binding energy was –9.24 kcal/mol, and the average molecular weight was 386 Da. At the same time, among the compounds from the unpromising region, only 10% of molecules had the energy values corresponding to the low class. The average energy for these molecules was –7.4 kcal/mol, and the average molecular weight was 269 Da.

These data showed that SGAN is able to generate new chemical compounds with the high affinity binding to the KasA protein. The fact that compounds from the unpromising region had a relatively small molecular weight correlates with the data according to which the binding free energy values predicted by the Vina scoring function often depend on the compound size (the greater molecular weight, the lower energy value). Therefore, the values of binding free energy were re-estimated in terms of scoring functions RF-Score-4 [12] and NNScore 2.0 [13]. Under the docking scores of three scoring functions, the ranks of the generated compounds were then calculated and the value of the exponential consensus ranking (ECR) function [14] for each molecule was obtained. The results obtained for the six top-ranked ligands and two reference compounds are shown in Table II. Analysis of the data of Table II shows that these ligands exhibit a strong attachment to the binding site of the KasA enzyme, as evidenced by the values of binding free energy which are lower than those predicted for the control inhibitors.

TABLE II. SCORING FUNCTIONS VALUES FOR THE SIX TOP-RANKED COMPOUNDS

Ligand	**ΔG_{VINA}, kcal/mol**	**$\Delta G_{RFScore4}$, kcal/mol**	**$\Delta G_{NNScore2}$, kcal/mol**	***ECR***
I	-11.2	-11.4	-12.9	0.262
II	-11.6	-11.3	-12.0	0.254
III	-10.8	-11.5	-10.9	0.243
IV	-11.1	-11.4	-10.4	0.234
V	-11.2	-11.1	-13.8	0.233
VI	-10.4	-11.6	-10.4	0.232
Control inhibitors				
Platensimycin	-9.6	-8.13	-9.48	
TLM5	-8.0	-8.27	-6.97	

Physicochemical properties of the identified compounds associated with the Lipinski's "rule of five" [15] were obtained by the SwissADME web tool (http://www.swissadme.ch). These properties met the "rule of five", and predicted synthesizability values indicate a high probability of synthetic availability of the selected compounds.

IV. Conclusion

Semi-supervised generative adversarial neural network trained on molecular graph embeddings produced by JT-VAE was developed to generate novel potential inhibitors of *Mycobacterium tuberculosis* protein KasA, one of the key enzymes responsible for mycobacterium cell wall biosynthesis. The results obtained showed that all 419 generated molecules are valid and 87% of these compounds show the values of binding free energy lower than –8.2 kcal/mol. Analysis of the six top-ranked compounds showed that these molecules form good scaffolds for the development of new antitubercular molecules with strong activity against *Mtb* and acceptable pharmacological properties.

The developed generative neural network model can also be repurposed for the designing new potential inhibitors of other therapeutic targets.

Acknowledgments

This work was supported by grants from the Belarusian Republican Foundation for Fundamental Research (F23-007) and the Consortium and the Drug Resistant Tuberculosis Portal Program (https://tbportals.niaid.nih.gov).

References

[1] J. Meyers, B. Fabian, and N. Brown, "De novo molecular design and generative models," Drug Discovery Today, vol. 26, no. 11, pp. 2707–2715, 2021. doi: 10.1016/j.drudis.2021.05.019.

[2] A. Bhatt, L. Kremer, A.Z. Dai, J.C. Sacchettini, and W.R. Jacobs, "Conditional depletion of KasA, a key enzyme of mycolic acid biosynthesis, leads to mycobacterial cell lysis," Journal of Bacteriology, vol. 187, no. 22, pp. 7596-606, 2005. doi: 10.1128/JB.187.22.7596-7606.2005.

[3] W Jin, R Barzilay and T Jaakkola, "Junction tree Variational autoencoder for molecular graph generation," International Conference on Machine Learning, vol. 80, pp. 2323– 2332, 2018.

[4] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," Journal of Chemical Information and Computer Sciences, vol. 28, no. 1, pp. 31–36, 1988.

[5] C. A. Machutta, G. R. Bommineni, S. R. Luckner, K. Kapilashrami, B. Ruzsicska, C. Simmerling, C. Kisker, and P. J. Tonge, "Slow onset inhibition of bacterial beta-ketoacyl-acyl carrier protein synthases by thiolactomycin," The Journal of biological chemistry, vol. 285, no. 9, pp. 6161–6169, 2010.

[6] J.D. Rudolf, L.-B. Dong, and B. Shen,."Platensimycin and platencin: Inspirations for chemistry, biology, enzymology, and medicine," Biochemical Pharmacology, vol. 133, pp. 139–151, 2017.doi:10.1016/j.bcp.2016.11.013.

[7] A. Witkowski, A. K. Joshi, Y. Lindqvist, and S. Smith, "Conversion of a beta-ketoacyl synthase to a malonyl decarboxylase by replacement of the active-site cysteine with glutamine," Biochemistry, vol. 38, no. 36, pp. 11643–11650, 1999.

[8] J. Sunseri and D. R. Koes, "Pharmit: interactive exploration of chemical space," Nucleic Acids Research, vol. 44, no. W1, pp. W442–W448, 2016.

[9] A. Alhossary, S.D. Handoko, Y. Mu, and C.-K. Kwoh, "Fast, accurate, and reliable molecular docking with QuickVina 2," Bioinformatics, vol. 31, pp. 2214–2216, 2015.

[10] S. Riniker and G.A. Landrum, "Better informed distance geometry: Using what we know to improve conformation generation," Journal of Chemical Information and Modeling, vol. 55, no. 12, pp. 2562–2574, 2015.. doi:10.1021/acs.jcim.5b00654.

[11] S. Wang, J. Witek, G. A. Landrum, and S. Riniker, "Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences," Journal of Chemical Information and Modeling, vol. 60, no. 4, pp. 2044–2058, 2020, doi: 10.1021/acs.jcim.0c00025.

[12] H. Li, K.S. Leung, M.H. Wong, P.J. Ballester, "Correcting the impact of docking pose generation error on binding affinity prediction," BMC Bioinformatics. 2016 Sep 22;17(Suppl 11):308. doi: 10.1186/s12859-016-1169-4.

[13] J.D. Durrant, J.A. McCammon, "NNScore 2.0: a neural-network receptor-ligand scoring function," J Chem Inf Model. 2011 Nov 28;51(11):2897-903. doi: 10.1021/ci2003889

[14] K. Palacio-Rodríguez, I. Lans, C.N. Cavasotto, and P. Cossio, "Exponential consensus ranking improves the outcome in docking and receptor ensemble docking," Scientific Reports, vol. 9, no.1,: Article 1, 2019. doi: 10.1038/s41598-019-41594-3.

[15] C.A. Lipinski, F. Lombardo, B.W. Dominy, and P.J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and evelopment settings," Advanced Drug Delivery Reviews, vol. 46, pp. 3–26, 2001. PMID: 11259830.