Application of the LSTM-based deep generative model for *de novo* design of potential HIV-1 entry inhibitors

Danila A. Varabyeu
United Institute of Informatics Problems,
National Academy of Sciences of Belarus
Minsk, Republic of Belarus
daniel.vorobiov.2002@yandex.ru

Anna D. Karpenko
United Institute of Informatics Problems,
National Academy of Sciences of Belarus
Minsk, Republic of Belarus
rfe.karpenko@gmail.com

Keda Yang
Shulan International Medical College,
Zhejiang Shuren University,
Hangzhou 310015, China
kdyang@zjsru.edu.cn

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

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

***Abstract*—A Long Short-Term Memory (LSTM) autoencoder model for the design of novel inhibitors of gp120, the HIV-1 envelope glycoprotein critically important for the virus pathogenesis, was repurposed and used to generate a series of compounds potentially active against this therapeutic target. Training and validation of this neural network was carried out using a set of small-molecule compounds collected by a public web-oriented virtual screening platform Pharmit allowing one to search for small molecules based on their structural and chemical similarity to another small molecule. The trained neural network was then evaluated for validity, and the values of binding free energy to the target protein were estimated. As a result, it was shown that the LSTM-based autoencoder model is an effective tool for the design of potent inhibitors against gp120 and may be used for the development of new drugs able to combat other dangerous diseases.**

Keywords—machine learning, deep learning, generative neural networks, LSTM, autoencoders, molecular docking, antiviral drugs, HIV-1, GP120

I. Introduction

To date, the problem of discovering new drugs is a highly demanded and important for the world community. Many diseases that in the recent past were considered incurable, including those with unknown etiology, have been studied using *in silico* research. Computer-aided drug design is currently an important tool that can significantly reduce the time and costs required to develop new therapeutic agents. Using *in silico* studies, the development of drugs for therapy of many diseases, such as human immunodeficiency virus, chronic myeloid leukemia, COVID-19, diabetes mellitus, drug-resistant tuberculosis, and many others, was significantly accelerated [1-3].

Thus, one of the primary tasks in the discovery of new drugs consists in the finding new suitable drug candidates. Modern chemical databases include millions of different substances, making it impossible to find a suitable compound without using the automated search methods. At the same time, it is necessary that the automated search could not only select compounds with known properties, but also predict new molecules with the desired physicochemical parameters. To implement this approach, procedure of virtual screening of chemical databases was developed. Simultaneously, the use of deep learning algorithms can speed up and make the process of drug development cheaper. Additionally, the use of machine learning methods allows one to obtain compounds missing in the current chemical databases. Thus, machine learning methods can help to develop new compounds that will be more effective inhibitors of a given molecular target than the currently known drugs.

The goal of this work was to find new potent inhibitors of the CD4-binding site of the HIV-1 envelope protein gp120 using an LSTM generative model combined with pharmacophore-based virtual screening and molecular modeling techniques,

To reach the object of view, the following problems were solved: (i) selection of a suitable generative neural network model based on the convenience of the input data and speed of learning and generation, (ii) formation of a training dataset including compounds able to specifically and effectively interact with the CD4-binding site of gp120, preparation of therapeutic target, and molecular docking, (iii) training of the autoencoder model, (iv) generation of a number of new small-molecule compounds potentially active towards gp120, (v) estimation of the values of binding energy of the generated compounds to gp120 by molecular docking, (vi) analysis of the data from molecular docking and selection of the lead compounds promising for the development of novel effective inhibitors of the HIV-1 gp120 envelope protein.

II. Materials and Methods

LSTM Autoencoder Architecture

In the selected autoencoder model [4], the LSTM layers are used to process input data, after which the states from these layers are combined on the concatenating layer, forming embeddings. These embeddings are preliminary processed by a fully connected layer and combined on the concatenating layer with the data on the neuron responsible for the binding energy of a compound, forming a latent space.

The obtained elements of the latent space are fed to the decoder, where they are used to create the states of the LSTM layer. To do this, data from the latent space are fed to two independent fully connected layers, and then their outputs are transmitted as short-term and long-term states to the LSTM layer. The input data of the LSTM layer are the same as those expected at the output of the neural network. This allows one to improve the performance of the decoder and speed up the training of the model. The outputs from the LSTM layer are further go to the fully connected layer with the softmax activation function, which is used to process the data for obtaining the probabilities of the next character in the output. For all other fully connected layers, the ReLU activation function was used, and for the LSTM layers, the tanh activation function was applied. The general scheme of this LSTM-based autoencoder model is shown in Fig. 1.

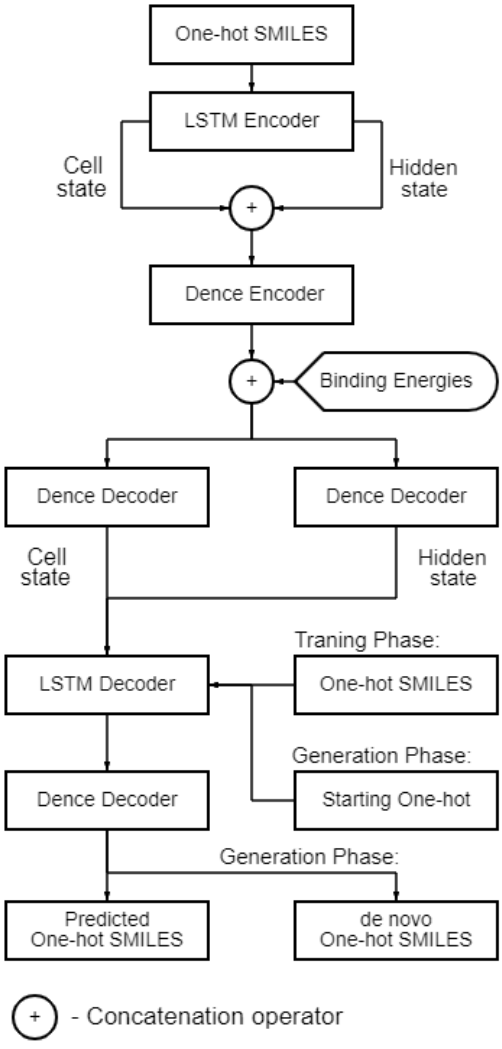

Fig. 1. Architecture of the LSTM-based autoencoder. model

Input Data Preparation

Before training the neural network, a training dataset of 77,184 small-molecule compounds was formed. For this purpose, the crystal structure of the potent HIV-1 entry inhibitor NBD-14204 bound to the viral envelope protein gp120 (PDB ID: 8F9Z) [5] was used to generate the pharmacophore model of this antiviral agent, perform the pharmacophore-based virtual screening of chemical databases (Pharmit; https://pharmit.csb.pitt.edu) [6] and molecular docking (AutoDock Vina; https://vina.scripps.edu) [7].

A pharmacophore model presents a set of steric and electronic properties necessary to ensure optimal molecular interaction with a specific biological target and to trigger (or block) its biological response [8, 9].

To run the virtual screening, the file describing the NBD-14204/gp120 complex in crystal (PDB ID: 8F9Z) [5] was downloaded from the PDB database [10] and processed using the UCSF Chimera software package [11] and web server SWISS-MODEL, an open access tool for automated comparative modeling of three-dimensional protein structures [12]. The virtual screening was performed using the chemical databases of the Pharmit web service [6], specifically CHEMBL30, ChemDiv, ChemSpace, MCULE-ULTIMATE, MolPort, NCI, LabNetwork, and ZINC-15. At the same time, the Lipinski "rule of five" [13] was used followed by filtering compounds by the root mean square deviations between the query features and the hit compound features [14] and restriction of binding energy value [6].

Before molecular docking, hydrogen atoms were added to the restored HIV-1 gp120 structure using the USCF Chimera software package [11]. Non-polar hydrogen atoms were removed from the protein structure using AutoDockTools-1.5.7 [15] and the source file was translated from the format .pdb to the .pdbqt one required for AutoDock Vina [7].

Molecular docking, an automated computer-aided algorithm that allows one to determine the ligand poses in the active site of a protein and calculate the values of bonding free energy, was carried out in the approximation of a rigid receptor and flexible ligands. The grid cell for docking included CD4-binding site of gp120 and had the following parameters: $\Delta X = 21.5$ Å, $\Delta Y = 21$ Å, $\Delta Z = 25$ Å centered at $X = -2.4$ Å, $Y = -16.57$ Å, $Z = 11.95$ Å. The value of the exhaustiveness parameter, which specifies the number of runs of individual samples, was equal to 100.

After molecular docking, the data were transferred to the SMILES canon format [16] and five lines were generated in the SMILES format for each line, where each SMILES line started with an atom different from others in the original structure, allowing one to expand the training dataset to 385,920 compounds. Further, compounds in which the characters were less common than in 0.07% of cases were removed from the training dataset.

To provide the input data, the SMILES strings were translated by one-hot encoding into matrices in which the first character is added "!" and the characters "E" are added after the actual sequence of SMILES to the end of the string. In these matrices, indexes in a line of 120 characters are located horizontally, and a dictionary of characters compiled based on the training dataset is settled vertically.

Autoencoder Training

Categorical cross-entropy (CCE) was used as a loss function and calculated using the formula

$$CCE(s) = -\sum_{s_i \in s} p(s_i) \log q(s_i),$$

where $p(s_i)$ and $q(s_i)$ are the true and predicted probabilities of generating the character si of the string s, respectively.

This function is a classic approach used in machine learning for solving multiclass classification problems [17]. For its optimal operation, it is required to select weights in such a way that this function reaches the minimum value on the training input data.

As an optimizer, the ADAM method [18], a classic adaptive gradient descent algorithm using moments, was used with a learning rate value of $\varepsilon = 0.008$.

After completing the training process, evaluation of the model was performed. Fig. 2 shows graphs of the loss function for one hundred training epochs.

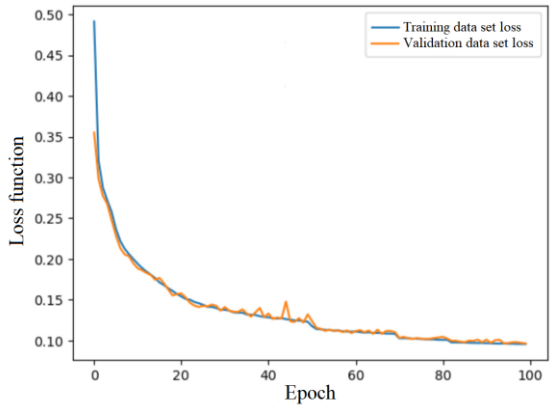

Fig. 2. Training and validation losses for the LSTM-based autoencoder model.

Compound Generation

To generate new data after designing the full-fledged autoencoder, the process was divided into three models. The first model performs the function of converting initial data into a vector of latent space, which is the coding part of the autoencoder. Using this model, one can get a representation of the latent space from the validation sample set by passing it to the input of the model. The second model converts the data from the latent space into the state vectors for the LSTM layer of the third model, by passing through a fully connected layer. The third model consists of an input and LSTM layers as well as of a fully connected layer having the weights with the same dimension as the autoencoder which are initialized by transferring the weights from the trained neural network. The model takes as input one one-hot encoded symbol and a vector of initial states for the LSTM layer from the previous model, and outputs one encoded symbol, which is presumably the next in the generated string. The generation starts with the beginning of the string character "!" and ends as soon as the end of the string character "E" is encountered.

III. RESULTS AND DISCUSSION

Using the above LSTM-based autoencoder, 46,846 novel compounds were generated in the linear SMILES format. These SMILES were cleaned from duplicates, checked for validity and interpretability using the RDKit module (http://www.rdkit.org/) [19] and converted to 2D and 3D chemical structures.

The values of binding free energy to the HIV-1 gp120 protein were evaluated for 5,000 molecules selected randomly from the sample of generated compounds. The results of the binding energy estimation for the original and generated molecules are presented in Fig. 3 in the form of two histograms.

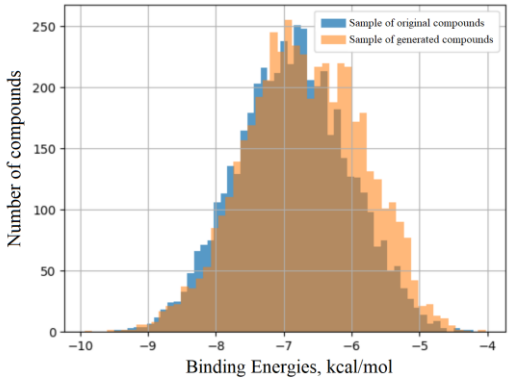

Fig. 3. Histograms of the distribution of the number of compounds by binding energy for the sample of original and generated compounds.

According to the molecular docking data, for, 159 of 5,000 generated compounds showed the values of binding free energy to gp120 close to or lower than the reference HIV-1 inhibitor NBD-14204 (for NBD-14204, this value was equal to −8.3 kcal/mol).

In addition, a more accurate assessment of the protein-ligand binding affinity was performed for the 5,000 selected compounds using scoring functions NNScore 2.0 [20] and RF-Score-4 [21]. For this purpose, the ranks of all compounds were calculated according to each scoring function and the values of the exponential consensus rank (ECR) were obtained for each compound by the formula [22]

$$ECR = \sum_{sf} \frac{1}{\sigma_{sf}} \cdot \exp\left(-\frac{rank_{sf}}{\sigma_{sf}}\right),$$

where $rank_{sf}$ is the rank of the compound according to the scoring function *sf*, σ_{sf} is the parameter that controls the influence of the scoring function sf on the results of consensus selection (ECR was calculated using $\sigma_{sf} = 10$ for all considered *sf*, since the contributions of the individual scoring functions were taken equal).

The results obtained for the five top-ranked compounds and NBD-14204 are shown in Table I.

TABLE I SCORING FUNCTIONS VALUES FOR THE FIVE TOP-RANKED COMPOUNDS AND NBD-14204

Ligand	ΔG_{VINA}, kcal/mol		$\Delta G_{RFScore4}$, kcal/mol	$\Delta G_{NNScore2}$, kcal/mol	*ECR*
	Orig.	Gen.			
I	-8.4	-9.1	-10.95	-10.32	0.289
II	-6.8	-8.9	-10.95	-10.38	0.288
III	-8.3	-9.1	-10.59	-10.15	0.281
IV	-8.5	-8.7	-10.74	-9.95	0.277
V	-6.3	-9.0	-10.21	-12.14	0.276
NBD-14204	-8.3	-8.3	-9.87	-7.83	–

The aqueous solubility, synthetic accessibility, and toxicity were evaluated for the 100 generated compounds with the best values of ECR using the SwissADME platform, a free web tool to evaluate drug-like properties of chemical compounds (http://swissadme.ch) [23]. Under the SwissADME predictions, the 84 compounds satisfy the physicochemical parameters commonly used as the basic filters to screen molecules for their ability to be effective drugs.

Chemical structures of the 3 top-ranked compounds and NBD-14204 are shown in Fig. 4.

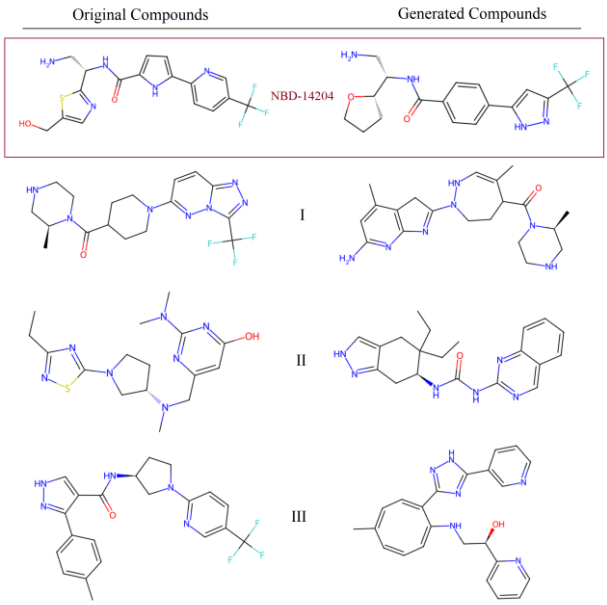

Fig. 4. An example of the generated compounds.

The further advancement of this study suggests the improving of the used deep generative model by switching to reinforcement learning to achieve the lower values of binding energies and their evaluation by molecular dynamics methods.

IV. Conclusion

The LSTM-based generative neural network [4] was repurposed for de novo design of potential inhibitors of the HIV-1 envelope glycoprotein gp120 playing a pivotal role in the virus attachment to the cellular receptor CD4. This generative autoencoder model was trained and tested, and the results of its operation were analyzed. During the validation of the neural network, 46,846 molecules were generated, and their inhibitory potential was evaluated using molecular docking tools.

As a result, eighty-four compounds that are of great interest for further studies were identified to be used as basic structures for the development of novel potent antiviral agents able to stop the HIV/AIDS spread.

References

[1] D. Barh, V. Chaitankar, E.C. Yiannakopoulou et al., "In Silico Models: From SimpleNetworks to Complex Diseases". Animal Biotechnology; Academic Press: Cambridge, MA, USA, pp. 385–404, 2014.

[2] J. Vamathevan et al., "Applications of machine learning in drug discovery and development," Nature Reviewers. Drug Discovery, vol. 18, no. 6, pp. 463–477, 2019.

[3] J. Meyers, B. Fabian and N. Brown "De novo molecular design and generative models," Drug Discovery Today, Vol. 26 (11), pp. 2707–2715, 2021. doi: 10.1016/j.drudis.2021.05.019

[4] Mikita A. Shuldau, Artsemi M. Yushkevich, Ivan P. Bosko, Alexander V. Tuzikov and Alexander M. Andrianov. Generative Autoencoders for Designing Novel Small-Molecule Compounds as Potential SARS-CoV-2 Main Protease Inhibitors. Communications in Computer and Information Science, vol. 1562, 2022, 120-136.

[5] F. Curreli, Y.D. Kwon, I. Nicolau et al., "Antiviral Activity and Crystal Structures of HIV-1 gp120 Antagonists," Int J Mol Sci. 2022 Dec 15;23(24):15999.

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

[7] O. Trott and A.J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," Journal of. Computational Chemistry, vol. 31, no. 2, pp. 455–461, 2010. doi: 10.1002/jcc.21334.

[8] C.G. Wermuth, C.R. Ganellin, P. Lindberg and L.A. Mitscher, "Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998)". Pure and Applied Chemistry. 70 (5): 1129–1143.

[9] Yang, S.-Y. "Pharmacophore modeling and applications in drug discovery: challenges and recent advances," Drug Discovery Today, 15(11-12), pp. 444– 450, 2010.

[10] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland,L. Iype and S. Jain, et al. "The protein data bank". Acta Crystallographica Section D: Biological Crystallography,58(6):899–907, 2002.

[11] E.F. Pettersen, T.D. Goddard, C.C. Huang and G.S. Couch et al., "UCSF Chimera — A visualization system for exploratory research and analysis," Journal of Computational Chemistry, 25(13), pp. 1605–1612, 2004.

[12] F. Kiefer, K. Arnold, M. Kunzli, L. Bordoli and T. Schwede, "The SWISS-MODEL Repository and associated resources," Nucleic Acids Research, 37(Database), D387–D392, 2009.

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

[14] Y. Maruyama, R. Igarashi, Y. Ushiku and A. Mitsutake, "Analysis of Protein Folding Simulation with MovingRoot Mean Square Deviation," J Chem Inf Model. 2023;63(5):1529-41. doi:10.1021/acs.jcim.2c01444.

[15] L. Ravi and K. Kannabiran, "A handbook on protein-ligand docking tool: AutoDock 4," Innovare Journal of Medical Sciences, pp. 28-33, 2016.

[16] 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. doi.org/10.1021/ci00057a005.

[17] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," IEEE Access, vol. 8, pp. 4806 4813, 2019.

[18] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[19] G. Landrum, "RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling," Greg Landrum, 8, 2013.

[20] J.D. Durrant and J.A. McCammon, "NNScore 2.0: A neural-network receptor–ligand scoring function," Journal of Chemical Information and Modeling, vol. 51, no. 11, pp. 2897–2903, 2011. doi: 10.1021/ci2003889.

[21] M. Wójcikowski, P.J. Ballester, and P. Siedlecki, "Performance of machine-learning scoring functions in structure-based virtual screening," Scientific Reports, vol. 7, no. 1, pp. 1–10, 2017.

[22] 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: 5142, 2019. doi: 10.1038/s41598-019-41594-3.

[23] A. Daina, O. Michielin and Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-like-ness and medicinal chemistry friendliness of small molecules. Sci. Rep. 2017; 7:1–13.