ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

БИОИНФОРМАТИКА

BIOINFORMATICS

УДК 51-76:577.322:539.19:004.94
https://doi.org/10.37661/1816-0301-2020-17-1-7-17

Поступила в редакцию 15.01.2020
Received 15.01.2020

Принята к публикации 10.02.2020
Accepted 10.02.2020

Разработка генеративной состязательной нейронной сети для идентификации потенциальных ингибиторов ВИЧ-1 методами глубокого обучения

Г. И. Николаев[1], Н. А. Шульдов[2], А. И. Анищенко[2], А. В. Тузиков[1], А. М. Андрианов[3]✉

[1]Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь
✉E-mail: alexande.andriano@yandex.ru

[2]Белорусский государственный университет, Минск, Беларусь

[3]Институт биоорганической химии Национальной академии наук Беларуси, Минск, Беларусь

Аннотация. Методами глубокого обучения разработан генеративный состязательный автоэнкодер для рационального дизайна потенциальных ингибиторов проникновения ВИЧ-1, способных блокировать участок белка gp120 оболочки вируса, критический для его связывания с клеточным рецептором CD4. Были выполнены исследования, включающие создание архитектуры автоэнкодера, формирование молекулярной библиотеки потенциальных лигандов белка gp120 ВИЧ-1 для обучения нейронной сети, молекулярный докинг лигандов с белком gp120 и расчет свободной энергии связывания, генерацию молекулярных дескрипторов химических соединений обучающего набора данных, обучение нейронной сети, оценку результатов обучения и работы автоэнкодера.

Рассмотрены результаты тестирования автоэнкодера на широком наборе соединений из молекулярной библиотеки ZINC. Показано, что совместное использование нейронной сети с виртуальным скринингом баз данных химических соединений формирует продуктивную платформу для идентификации базовых структур, перспективных для создания новых противовирусных препаратов, ингибирующих ранние стадии развития ВИЧ-инфекции.

Ключевые слова: методы глубокого обучения, генеративно-состязательный автоэнкодер, белок gp120, ингибиторы проникновения ВИЧ-1, методы молекулярного моделирования

Для цитирования. Разработка генеративной состязательной нейронной сети для идентификации потенциальных ингибиторов ВИЧ-1 методами глубокого обучения / Г. И. Николаев [и др.] // Информатика. – 2020. − Т. 17, № 1. – С. 7–17. https://doi.org/10.37661/1816-0301-2020-17-1-7-17

Development of a generative adversarial neural network for identification of potential HIV-1 inhibitors by deep learning methods

Grigory I. Nikolaev[1], Nikita A. Shuldov[2], Arseny I. Anischenko[2], Alexander V. Tuzikov[1], Alexander M. Andrianov[3]✉

[1]The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus
✉E-mail: alexande.andriano@yandex.ru

[2]Belarusian State University, Minsk, Belarus

[3]Institute of Bioorganic Chemistry of the National Academy of Sciences of Belarus, Minsk, Belarus

Abstract. A generative adversarial autoencoder for the rational design of potential HIV-1 entry inhibitors able to block the region of the viral envelope protein gp120 critical for the virus binding to cellular receptor CD4 was developed using deep learning methods. The research were carried out to create the architecture of the neural

network, to form virtual compound library of potential anti-HIV-1 agents for training the neural network, to make molecular docking of all compounds from this library with gp120, to calculate the values of binding free energy, to generate molecular fingerprints for chemical compounds from the training dataset. The training the neural network was implemented followed by estimation of the learning outcomes and work of the autoencoder. The validation of the neural network on a wide range of compounds from the ZINC database was carried out. The use of the neural network in combination with virtual screening of chemical databases was shown to form a productive platform for identifying the basic structures promising for the design of novel antiviral drugs that inhibit the early stages of HIV infection.

Key words: deep learning methods, a generative adversarial neural network, gp120 protein, HIV-1 entry inhibitors, molecular modeling

For citation. Nikolaev G. I., Shuldov N. A., Anischenko A. I., Tuzikov A. V., Andrianov A. M. Development of a generative adversarial neural network for identification of potential HIV-1 inhibitors by deep learning methods. *Informatics*, 2020, vol. 17, no. 1, pp. 7−17 (in Russian). https://doi.org/10.37661/1816-0301-2020-17-1-7-17

Введение. Современные методы компьютерного конструирования потенциальных лекарств значительно расширяют возможности фармацевтической индустрии, позволяя существенно сократить время и затраты, необходимые для создания новых терапевтических средств. Несмотря на то что эффективность компьютерных методов в создании лекарственных препаратов в настоящее время является общепризнанной, разработка новых математических подходов в сочетании с доступностью мощных и дешевых вычислительных ресурсов способствует их постоянному совершенствованию. Среди этих подходов важное место занимают методы машинного обучения (Machine Learning) и, в частности, методы глубокого обучения (Deep Learning), которые имеют большой потенциал для дальнейшего прогресса в данной области исследований. На сегодняшний день компьютерное конструирование потенциальных лекарств с помощью методов машинного обучения – одна из наиболее важных и быстро развивающихся областей хемоинформатики [1]. В отличие от физических моделей, основанных на физических закономерностях, таких же, как в квантовой химии или моделировании молекулярной динамики, в подходах машинного обучения реализуются алгоритмы распознавания образов для определения математических взаимосвязей между эмпирическими наблюдениями за малыми молекулами и их экстраполяциями для прогнозирования химических, биологических и физических свойств новых соединений. Кроме того, по сравнению с физическими моделями методы машинного обучения более эффективны и могут легко масштабироваться до больших наборов данных. Одним из преимуществ применения машинного обучения для конструирования лекарств является помощь исследователям в понимании и использовании взаимосвязи между химической структурой и ее биологической активностью (Quantitative Structure-Activity Relationship, QSAR) [2]. Современные методы машинного обучения могут использоваться для моделирования такой взаимосвязи или количественных отношений между структурой и свойством (Quantitative Structure-Property Relationship, QSPR), а также разработки интеллектуальных инструментов, способных достаточно точно предсказывать влияние химических модификаций соединения на его биологическую активность, фармакокинетические и токсикологические характеристики [1]. В связи с этим применение методов машинного обучения для компьютерного дизайна потенциальных лекарств имеет важное научное и практическое значение [3].

За последние несколько лет идея использования технологий искусственного интеллекта для ускорения процесса создания новых лекарственных препаратов и повышения эффективности программ фармацевтических исследований стала особенно востребованной в области хемоинформатики.

2018 г. ознаменовался впечатляющим числом проектов по сбору средств среди стартапов по поиску лекарств, полученных посредством использования искусственного интеллекта. Это свидетельствует о том, что работы по созданию нейронных сетей для идентификации потенциальных лекарств обладают серьезной привлекательностью для венчурных инвесторов. В настоящее время лондонская фармацевтическая компания BenevolentAI является лидером по сбору средств, достигнув в 2018 г. ошеломляющей отметки в 2 млрд долл. США (URL: https://www.linkedin.com/pulse/aimldl-drug-discovery-2018-year-review-andrii-buvailo). Компания

Atomwise, основанная в 2012 г. и ставшая пионером в использовании нейронных сетей для структурного проектирования лекарств, привлекла инвестиции в размере 45 млн долл. США для развития своей технологии открытия лекарств на основе алгоритмов глубокого обучения (URL: https://www.linkedin.com/pulse/aimldl-drug-discovery-2018-year-review-andrii-buvailo). Эта компания разработала нейронную сеть AtomNet и ежедневно тестирует с ее помощью 10 млн малых молекул для анализа их эффективности, прогнозирования токсичности и побочных эффектов с целью проверки на возможность использования в качестве лекарственных препаратов. Американская компания Insilico Medicine разрабатывает интеллектуальную систему, основанную на генеративных состязательных сетях, что позволит осуществлять процесс прогнозирования потенциальных лекарств от базового биологического моделирования и разработки биомаркеров до генерации молекул-лидеров, их оптимизации и доклинической проверки структур – кандидатов в лекарства (URL: https://www.linkedin.com/pulse/aimldl-drug-discovery-2018-year-review-andrii-buvailo).

В последние годы появилось большое число работ по применению методов машинного обучения для предсказания потенциальных ингибиторов ВИЧ-1 и резистентности вируса к анти-ВИЧ-препаратам (см., например, обзор [4]). Однако все эти исследования сконцентрированы на вирусных ферментах – обратной транскриптазе и протеазе. Соединения, блокирующие эти ферменты, не могут предотвращать проникновение вируса в клетку-мишень, что повышает внимание к ингибиторам ВИЧ-1, способным вмешиваться в ранние стадии жизненного цикла вируса путем блокирования процессов адсорбции и слияния мембран. Проникновение вирусного генома в клетку-хозяина – первый этап репликационного цикла ВИЧ-1 – представляет собой перспективную мишень для нескольких типов противовирусных препаратов, таких как ингибиторы связывания белка gp120 с первичным рецептором CD4, антагонисты корецепторов CCR5 и CXCR4 и ингибиторы слияния оболочки вируса с мембраной чувствительной клетки [5].

Цель настоящей работы – методами глубокого обучения создать генеративно-состязательную автоэнкодерную нейронную сеть для дизайна потенциальных ингибиторов ВИЧ-1, способных блокировать участок оболочки вируса, критический для его связывания с клеточным рецептором CD4. Для этого были проведены исследования, включающие:

– создание архитектуры состязательного автоэнкодера;

– формирование молекулярной библиотеки потенциальных лигандов белка gp120 ВИЧ-1 для обучения нейронной сети;

– молекулярный докинг лигандов с белком gp120 и расчет свободной энергии связывания;

– генерацию молекулярных дескрипторов (fingerprints) химических соединений обучающего набора данных;

– обучение нейронной сети;

– оценку результатов обучения и работы состязательного автоэнкодера.

Молекулярные дескрипторы предназначены для формального представления химических соединений в виде бинарных векторов фиксированной длины, что позволяет использовать такое представление для решения различных задач анализа и синтеза соединений методами машинного обучения.

Генеративно-состязательная нейронная сеть для идентификации потенциальных ингибиторов ВИЧ-1. Архитектура разработанного состязательного автоэнкодера состоит из двух нейросетей – автоэнкодера и дискриминатора, работающих во время обучения в соревновательном режиме. Такой режим позволяет настроить параметры автоэнкодера в режиме обучения и обеспечить получение качественных выходных данных на последующем этапе их генерирования. Задача дискриминатора состоит в том, чтобы отличать реальные данные от тех, которые генерирует автоэнкодер. Автоэнкодер представляет собой семислойную нейронную сеть, имеющую входной и выходной слои, латентный слой, а также четыре полносвязных слоя (рис. 1). На входной слой подаются молекулярные дескрипторы химических соединений, данные о которых проходят два полносвязных слоя (энкодер) и попадают на латентный слой, где к полученному результату добавляется численная оценка энергии связывания с молекулярной мишенью. Далее молекулярные дескрипторы проходят два полносвязных слоя (декодер) и попадают на выход, который, как и вход, представляет собой вектор молекулярного дескриптора. Работающая в таком режиме сеть уменьшает количество нейронов, поступающих на латентный

слой, который содержит сжатую информацию о векторе, поданном на вход сети, с последующим ее расширением на выходе. Латентный слой состоит их трех нейронов, два из которых получают значения от энкодера, а третий – значение энергии связывания с молекулярной мишенью. В рабочем режиме автоэнкодера на латентный слой, содержащий наиболее важную информацию об объекте, подаются случайные числа, которые затем проходят через декодер, генерирующий молекулярные дескрипторы молекул с требуемыми свойствами. Для генерации таких молекул важно, чтобы данные, поступающие на латентный слой после прохождения энкодера, имели нормальное распределение, которому обучены генератор случайных чисел и дискриминатор. Для обеспечения этого условия в процессе состязательного обучения энкодера и дискриминатора добивались того, чтобы энкодер был способен кодировать на латентный слой данные с нормальным распределением, а дискриминатор – отличать стандартное нормальное распределение (сгенерированные данные) от распределения, поступающего на латентный слой.

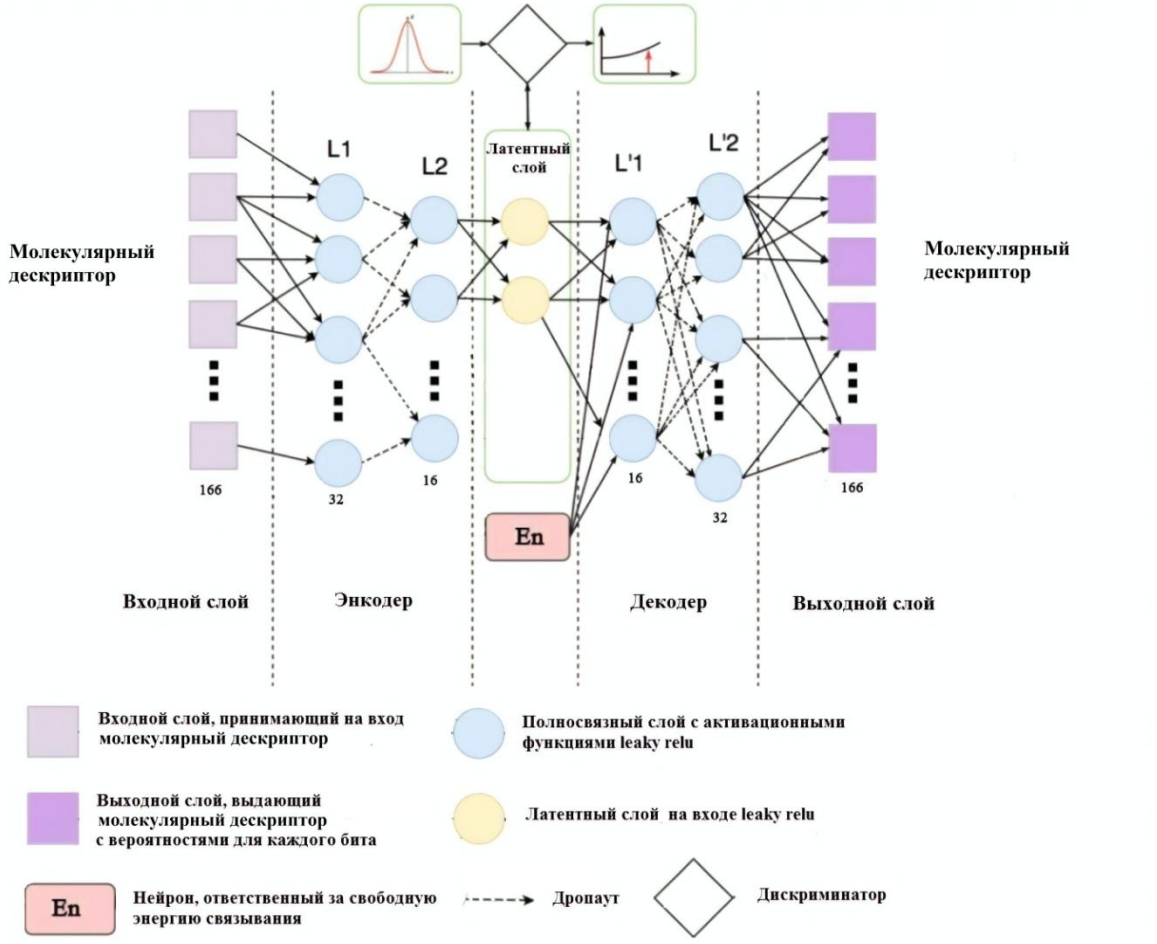


Рис. 1. Архитектура нейронной сети для генерации потенциальных ингибиторов ВИЧ-1, блокирующих CD4-связывающий сайт белка gp120 оболочки вируса

Разработанный состязательный автоэнкодер основан на модели нейронной сети, предназначенной для генерации химических соединений с противоопухолевой активностью [6], и имеет следующие особенности (рис. 1):

– на латентном слое используется нейрон, отвечающий за свободную энергию связывания. Он не взаимодействует с энкодером и подается только на вход декодера совместно с данными, полученными с помощью энкодера. Латентный слой состоит из трех нейронов;

– энкодер состоит из двух последовательных слоев L1 и L2 с 32 и 16 нейронами соответственно. Декодер включает два слоя – L'1 и L'2, содержащие 16 и 32 нейрона соответственно;

– дискриминатор состоит из четырех слоев, включающих 2, 16, 3 и 1 нейрон соответственно;

– на промежуточных слоях автоэнкодера используется активационная функция leaky relu [7]

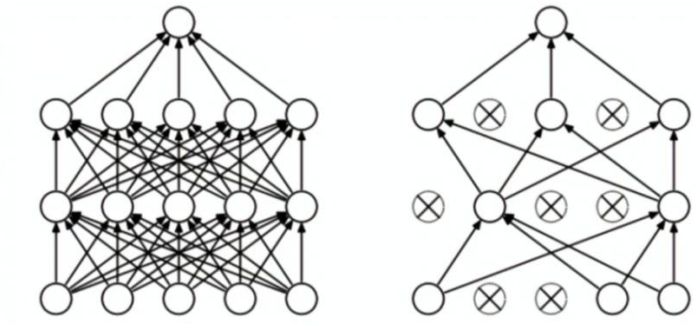
$$f(x) = \begin{cases} 0{,}01x \text{ при } x < 0, \\ x \text{ при } x \geq 0; \end{cases}$$

– на всех слоях дискриминатора используются сигмоидные активационные функции [8]

$$\sigma(x) = \frac{1}{1+e^{-x}};$$

– для дополнительного уровня защиты от переобучения между двумя полносвязными слоями энкодера и декодера добавлен слой дропаута, наличие которого позволяет нейронной сети полностью использовать свои параметры (веса) и выборочное случайное отключение во время обучения.

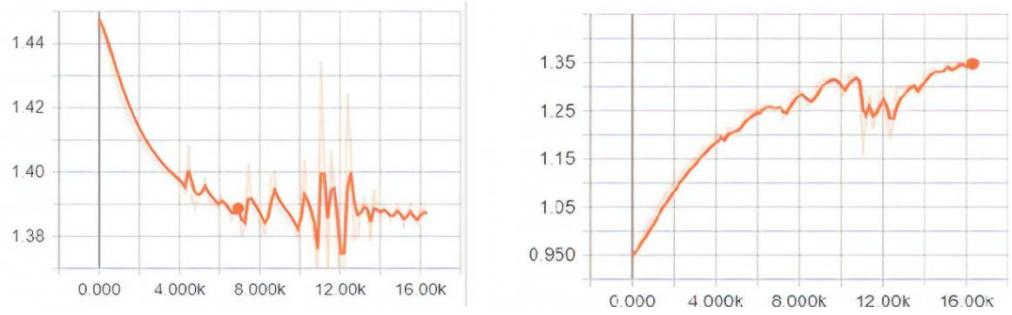
Основное назначение слоя дропаута заключается в том, чтобы вместо обучения одной сети обучить ансамбль из нескольких сетей, а затем усреднить полученные результаты. Схема работы слоя дропаута показана на рис. 2. Для обучения разработанного состязательного автоэнкодера использовался трехступенчатый итерационный процесс, который включал: 1) обучение дискриминатора различать заданное нормальное распределение от закодированного, полученного энкодером на латентном слое; 2) совместное обучение энкодера и декодера как автоэнкодера; 3) обучение энкодера сжимать данные таким образом, чтобы они представляли нормальное распределение.



а) *б)*

Рис. 2. Применение технологии дропаута: *а)* модель стандартной нейронной сети; *б)* модель нейронной сети с технологией дропаута, заключающейся в «выключении» из сети случайного набора нейронов, отмеченных на рисунке крестиком

Для дискриминационных моделей характерен более простой процесс обучения, чем для генеративных нейронных сетей. Поэтому при первых попытках обучить модель возникала ситуация, при которой функция потерь дискриминатора (ФПД) снижалась, а функция потерь энкодера росла (рис. 3).



а) *б)*

Рис. 3. Функции потерь дискриминатора (*а*) и энкодера (*б*) при более быстром обучении дискриминатора. На графиках представлены информативные части функций

Следует отметить, что при обучении нейронной сети сложно определить, действительно ли энкодер плохо обучается или он не способен «обмануть» дискриминатор. В связи с этим для улучшения процесса обучения были предприняты следующие меры:

– осуществлялось предварительное обучение дискриминатора до обучения всего автоэнкодера; предполагалось, что в этом случае дискриминатор будет отличать случайные числа, получаемые из нетренированного энкодера, от значений, соответствующих используемому нормальному распределению;

– в процессе обучения всей группы моделей дискриминатор тренировали не каждую эпоху*, а один раз в две эпохи, т. е. он обучался только в одну из двух эпох для модели автоэнкодера;

– для дискриминатора задавалась меньшая скорость обучения (learning rate), равная 0,001, в то время как для всего автоэнкодера она составляла величину, равную 0,005;

– в данные, сгенерированные из нормального распределения, добавлялся «небольшой шум», что затрудняло работу дискриминатора.

Данными изменениями была дополнена каждая эпоха (итерация) обучения, которая представляла собой трехступенчатый итерационный процесс:

1) энкодер и декодер обучались совместно как автоэнкодер;

2) дискриминатор обучался различать заданное нормальное распределение и закодированное «представление», полученное энкодером на латентном слое;

3) энкодер учился сжимать данные таким образом, чтобы они представляли собой нормальное распределение.

Дискриминатор и автоэнкодер обучались совместно в два этапа: реконструкции и регуляризации, выполняемые в каждом подмножестве из оригинальных данных. На этапе реконструкции (первой ступени итерации) автоэнкодер обновлял энкодер и декодер, чтобы минимизировать ошибку восстановления входных и выходных данных. На этапе регуляризации (второй ступени итерации) сначала обновлялась сеть дискриминатора, чтобы отличить истинные выборки (полученные с помощью генератора нормального распределения) от сжатых входных данных (данных на латентном слое, вычисленных автоэнкодером), а затем на третьей ступени итерации автоэнкодер обновлял свой энкодер, чтобы запутать сеть дискриминатора.

На рис. 4 изображены графики ФПД с учетом описанных выше изменений. Видно, что дискриминатор хорошо обучается классифицировать данные из нормального распределения, однако его способность правильно идентифицировать данные, полученные энкодером, падает.

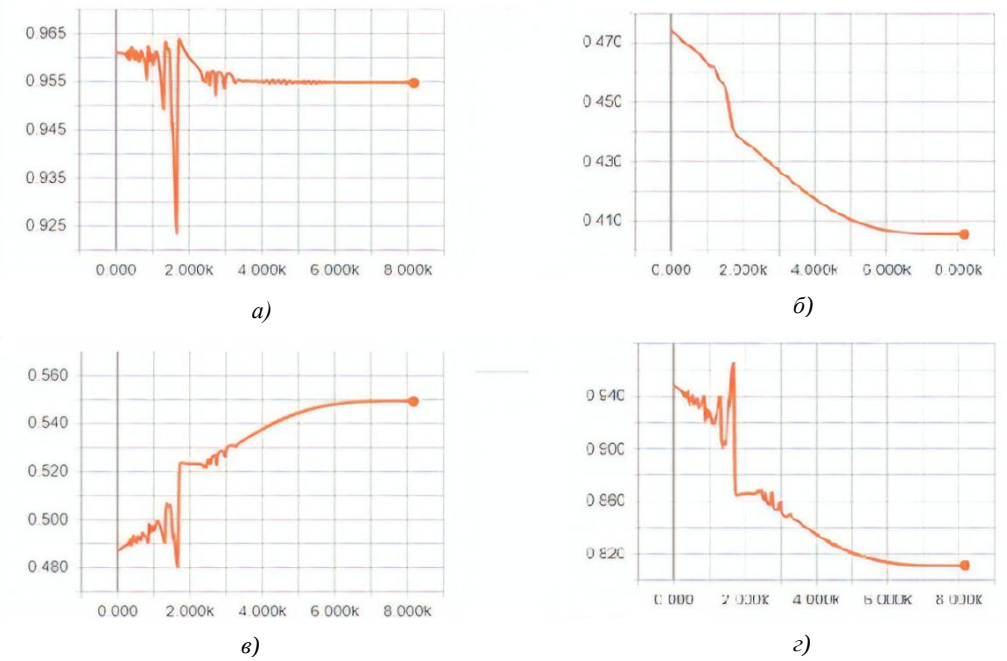
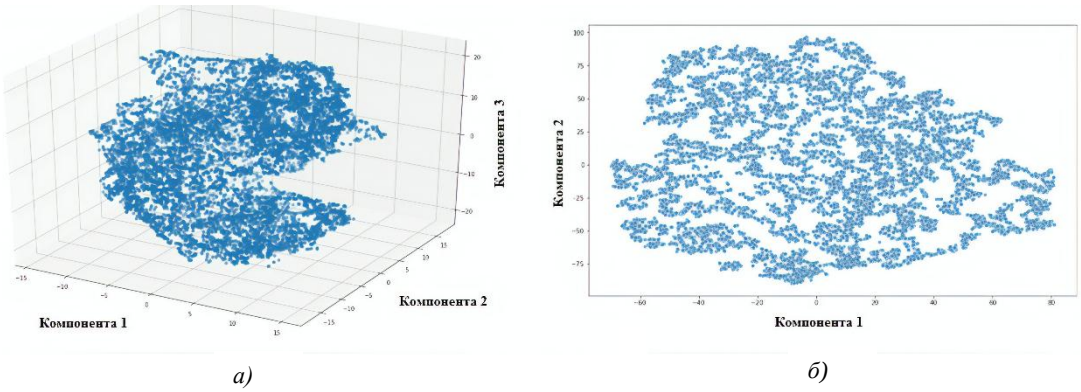


Рис. 4. Графики информативных частей ФПД (*а*), состоящей из ФПД для нормального распределения (*б*) и ФПД для закодированных данных (*в*), а также функция потерь энкодера (*г*)

*Эпоха – одна итерация в процессе обучения, включающая предъявление всех примеров из обучающего множества.

Для обучения автоэнкодера использовались следующие параметры:
количество эпох для главной версии модели, используемой для генерации, – 400;
скорость обучения всего автоэнкодера на первой ступени итерации – 0,005;
скорость обучения дискриминатора на второй ступени итерации – 0,001;
скорость обучения энкодера на третьей ступени итерации – 0,005;
параметр Batch size – 128;
оптимизатор – метод Adam [9].

В процессе обучения нейронной сети важно было убедиться, что в режиме генерации модель способна выдавать разные результаты для различных входных данных и производить множество молекулярных дескрипторов, а не генерировать их из небольшого числа возможных вариантов, поскольку существовала вероятность того, что такая ситуация отражает некий минимум исходных функций потерь. С целью снижения вероятности реализации такого сценария был использован алгоритм tSNE [10] для подвыборки из сгенерированных молекул (рис. 5), который представляет каждый сгенерированный молекулярный дескриптор двух- или трехмерной точкой таким образом, что похожие молекулярные дескрипторы отображаются близко расположенными точками, а непохожие дескрипторы – с большой вероятностью далеко отстоящими дуг от друга точками.



а) *б)*

Рис. 5. Результаты работы алгоритма tSNE для трех (*а*) и двух (*б*) компонент с подвыборкой из сгенерированных молекулярных дескрипторов

Создание молекулярной библиотеки для обучения автоэнкодера. Формирование обучающего набора данных выполнено в рамках подхода, использующего методологию клик-химии [11] для генерации наиболее вероятных структур-кандидатов биологически активных соединений. Для конструирования потенциальных лигандов с помощью программы DataWarrior (URL: http://www.openmolecules.org/help/basics.html) были созданы две молекулярные библиотеки. Первая библиотека включала отобранные из кластера Drug-Like базы данных ZINC [12, 13] небольшие молекулы (с молекулярной массой менее 250 Да) с азидной или алкиновой группой, содержащие ароматические фрагменты – элементы структуры, которые согласно данным об известных ингибиторах проникновения ВИЧ-1 [14, 15] играют ключевую роль для специфического взаимодействия с Phe43-полостью CD4-связывающего сайта белка gp120. Во вторую библиотеку были отобраны все низкомолекулярные соединения с молекулярной массой менее 250 Да, имеющие азидную или алкиновую группу. В результате работы программы DataWarrior первая библиотека включала 1388 соединений, а вторая библиотека – 3769. На следующем этапе эти соединения были использованы в качестве исходных реагентов для имитации реакции азид-алкинового циклоприсоединения с помощью программы AutoClickChem [16], которая рассматривала все возможные комбинации молекул из обеих библиотек. Это позволило получить набор из 1 655 301 гибридной молекулы, из которого 120 000 соединений, удовлетворяющие «правилу пяти» Липинского [17], были использованы для формирования обучающего набора данных. Оценку энергии связывания этих соединений с белком gp120 проводили методом молекулярного докинга – процедуры виртуального скрининга, позволяющей предсказать наиболее вероятные ориента-

ции лиганда в активном центре белка и рассчитать свободную энергию образования комплексов «лиганд-белок».

Генерация молекулярных дескрипторов MACCS (URL: http://www.dalkescientific.com/writings/NBN/fingerprints.html) в обучающем наборе данных осуществлялась с помощью программного пакета RDKit (URL: https://www.rdkit.org/) с открытым исходным кодом.

Молекулярный докинг лигандов из обучающего набора данных с белком gp120 выполнялся с помощью программы QuickVina 2 [18] с учетом конформационной подвижности лиганда. Трехмерная структура белка gp120 была выделена из комплекса этого гликопротеина с рецептором CD4 и антителом 17b (код 1GC1 из банка данных белков [19]). Атомы водорода добавлены к структуре белка gp120 с помощью программного пакета AutoDockTools. Ячейка для докинга представляла собой фрагмент белка gp120 с координатами $x \in$ (24 Å; 34 Å), $y \in$ (–15 Å; –5 Å), $z \in$ (78 Å; 88 Å), включающий Phe43-полость гликопротеина, т. е. объем ячейки составлял $10\times10\times10 = 1000$ Å^3. Для каждого лиганда генерировали девять моделей комплекса, лучших по значению оценочной функции. При этом параметр, характеризующий полноту поиска (охват конформационного пространства), был задан равным 50.

Оценка результатов обучения и работы автоэнкодера. Для тестирования работы автоэнкодера с помощью программного пакета RDKit была создана библиотека молекулярных дескрипторов MACCS (URL: http://www.dalkescientific.com/writings/ NBN/fingerprints.html) для 21 325 567 соединений из библиотеки Drug-Like базы данных ZINC [12, 13] и рассчитаны пять молекулярных дескрипторов для сгенерированных автоэнкодером молекул при пороговом значении энергии связывания с белком gp120, равном 5 ккал/моль. В результате виртуального скрининга созданной библиотеки для каждой из этих молекул с подобными молекулярными дескрипторами были найдены лиганды, которые показаны в таблице. При этом в качестве меры подобия молекулярных дескрипторов использовалось расстояние Хэмминга, определяемое в теории кодирования как число пар несовпадающих компонент сравниваемых векторов [20], и коэффициент Танимото, который вычислялся по формуле [21]

$$T(a, b) = \frac{N_c}{N_a+N_b-N_c},$$

где T – коэффициент Танимото, принимающий значения от 0 до 1; N_a – количество элементов в первом векторе; N_b – количество элементов во втором векторе; N_c – количество одинаковых элементов в двух векторах. В процессе скрининга библиотеки молекулярных дескрипторов из базы данных ZINC отбирались соединения, для которых коэффициент Танимото удовлетворял условию $T > 0{,}85$ [21]. В таблице молекулярные дескрипторы представлены строками из 166 бит, в которых каждый бит соответствует присутствию либо отсутствию определенного свойства или структурного фрагмента (URL: http://www.dalkescientific.com/writings/NBN/fingerprints.html). В векторах молекулярных дескрипторов, полученных после декодирования, 1 или 0 обозначает наличие либо отсутствие соответствующего структурного признака. Для каждого сгенерированного нейронной сетью лиганда приведены пять лучших по критериям R и T соединений из базы данных ZINC.

Идентифицированные в базе данных ZINC соединения подвергались процедуре докинга с белком gp120 и рассчитывалась энергия связывания с Phe43-полостью CD4-связывающего сайта оболочки ВИЧ. Молекулярный докинг проводился с помощью программы QuickVina 2 [18] с использованием вычислительного протокола, идентичного тому, что был применен при создании обучающего набора данных.

Анализ результатов молекулярного докинга найденных соединений с белком gp120 показал (таблица), что совместное использование нейронной сети с виртуальным скринингом библиотеки молекулярных дескрипторов позволяет идентифицировать лиганды с более низкой по сравнению с заданным пороговым значением энергией связывания. При этом идентифицированное в базе данных ZINC соединение с кодом ZINC000026430653 – аналог трех сгенерированных нейронной сетью лигандов – характеризуется величиной энергии связывания с белком gp120, сопоставимой со значением –9,5±0,1 ккал/моль, измеренным для комплекса CD4-gp120

методом изотермической титрационной калориметрии [22]. Эта величина, равная –8,8 ккал/моль, близка к значениям оценочной функции QuickVina 2 [18], полученным ранее [23–26] для высокоаффинных лигандов белка gp120, сконструированных методами молекулярного моделирования, а также для ингибиторов ВИЧ-1 NBD-11021 и NBD-14010, представляющих новое поколение полных функциональных антагонистов клеточного рецептора CD4 [27]. В частности, предсказанные QuickVina 2 значения свободной энергии Гиббса для ингибиторов NBD-11021 и NBD-14010 равны –8,4 и –8,6 ккал/моль соответственно.

Результаты тестирования нейронной сети

Молекулярные дескрипторы сгенерированных нейронной сетью лигандов	Коды соединений в базе данных ZINC	Расстояние Хэмминга R, критерий Танимото T	Энергия связывания, ккал/моль
00000000000000000000000000000000 00000000000000000000000100000000 01001000010000101011001000000 00010100010000000000010000000001 11011000010010101000010110010101 1010110101111110	ZINC000026430653	R = 3; T = 0,96	–8,8
	ZINC000037104033	R = 3; T = 0,96	–7,1
	ZINC000002786698	R = 4; T = 0,95	–7,0
	ZINC000055836809	R = 3; T = 0,96	–6,9
	ZINC000055843838	R = 4; T = 0,95	–6,5
00000000000000000000000000000000 00000000000000000000000100000000 01001000010000101011001000000 00011110010000000000010000000001 11011000010010101000010110010101 0010111101111110	ZINC000026430653	R = 5; T = 0,94	–8,8
	ZINC000037104033	R = 6; T = 0,93	–7,1
	ZINC000002786698	R = 6; T = 0,93	–7,0
	ZINC000163393594	R = 4; T = 0,95	–6,0
	ZINC000128895014	R = 4; T = 0,95	–5,9
00000000000000000000000000000000 00000000000000000000000100000000 01001000010000101011001000000 00011110010000000000010000000001 11011001010000001100010110010101 0010111101111110	ZINC000035245594	R = 6; T = 0,93	–6,6
	ZINC000163489237	R = 7; T = 0,92	–6,6
	ZINC000052221501	R = 7; T = 0,92	–6,6
	ZINC000600676089	R = 6; T = 0,93	–6,4
	ZINC000004006242	R = 7; T = 0,92	–5,9
00000000000000000000000000000000 00000000000000000000000100000000 01001000010000101011001000000 00010100010000000000010000000001 11011000010010001000010110010101 1010110101111110	ZINC000026430653	R = 4; T = 0,95	–8,8
	ZINC000002786698	R = 5; T = 0,94	–7,0
	ZINC000055836809	R = 4; T = 0,95	–6,8
	ZINC000037104033	R = 4; T = 0,95	–6,7
	ZINC000685198234	R = 4; T = 0,95	–6,0
00000000000000000000000000000000 00000000000000000000000100000000 01001000010000101011001000000 00011110010000000000010000000001 11011001010010101100010110010101 0010111111111110	ZINC000182934853	R = 7; T = 0,92	–7,4
	ZINC000771860139	R = 5; T = 0,94	–6,6
	ZINC000012991344	R = 7; T = 0,92	–6,0
	ZINC000128895014	R = 7; T = 0,92	–6,0
	ZINC000163393499	R = 7; T = 0,92	–6,0

Заключение. Результаты исследования свидетельствуют о том, что разработанная нейронная сеть представляет собой эффективную математическую модель для виртуального скрининга баз данных химических соединений, направленного на поиск малых молекул с высоким сродством к белку gp120 и разработку на их основе новых антиВИЧ-препаратов широкого спектра действия.

Разработанный автоэнкодер может быть использован для генерации молекулярных дескрипторов химических соединений, способных блокировать участок оболочки вируса, критический для его связывания с клеточным рецептором CD4. Анализ полученных результатов показывает, что совместное использование нейронной сети с виртуальным скринингом молекулярных библиотек формирует продуктивную платформу для идентификации базовых структур, перспективных для создания новых противовирусных препаратов, терапевтическое действие которых основано на ингибировании ранних стадий развития ВИЧ-инфекции.

References

1. Cherkasov A., Muratov E. N., Fourches D., Varnek A., Baskin I. I., …, Tropsha A. QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry*, 2014, vol. 201457, pp. 4977–5010.

2. Ali S. M., Hoemann M. Z., Aubé J., Georg G. I., Mitscher L. A., Jayasinghe L. R. Butitaxel analogues: Synthesis and structure-activity relationships. *Journal of Medicinal Chemistry*, 1997, vol. 40, pp. 236–241.

3. Vamathevan J., Clark D., Czodrowski P., Dunham I., Ferran E., …, Zhao S. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 2019, vol. 18(6), pp. 463–477.

4. Dubey A. Machine learning approaches in drug development of HIV/AIDS. *International Journal of Molecular Biology: Open Access*, 2018, vol. 3(1), pp. 23–25.

5. Li W., Lu L., Li W., Jiang S. Small-molecule HIV-1 entry inhibitors targeting gp120 and gp41: a patent review (2010-2015). *Expert Opinion on Therapeutic Patents*, 2017, vol. 27, pp. 707–719.

6. Kadurin A., Aliper A., Kazennov A., Mamoshina P., Vanhaelen Q., Khrabrov K., Zhavoronkov A. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*, 2017, vol. 8, pp. 10883–10890.

7. Xu B., Wang N., Chen T., Li M. *Empirical Evaluation of Rectified Activations in Convolutional Network*, 2015. Available at: https://arxiv.org/abs/1505.00853 (accessed 12.11.2019).

8. Rudoy G. I. The Choice of the Activation Function in the Prediction of Neural Networks. *Machine Learning and Data Analysis*, 2011, no. 1, pp. 16–39. Available at: https://arxiv.org/abs/1412.6980 (accessed 12.11.2019).

9. Kingma D., Ba J. *Adam: A Method for Stochastic Optimization*, 2014.

10. Van der Maaten L. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, vol. 9, pp. 2579–2605.

11. Kolb H. C., Finn M. G., Sharpless K. B. Click chemistry: Diverse chemical function from a few good reactions. *Angewandte Chemie International Edition*, 2001, vol. 40, no. 11, pp. 2004–2021.

12. Irwin J. J., Shoichet B. K. ZINC – a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 2005, vol. 45, no. 1, pp. 177–182.

13. Irwin J. J., Sterling T., Mysinger M. M., Bolstad E. S., Coleman R. G. ZINC: a free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 2012, vol. 52, no. 7, pp. 1757–1768.

14. Courter J. R., Madani N., Sodroski J., Schön A., Freire E., …, Smith A. B. 3rd. Structure-based design, synthesis and validation of CD4-mimetic small molecule inhibitors of HIV-1 entry: Conversion of a viral entry agonist to an antagonist. *Accounts of Chemical Research*, 2014, vol. 47, pp. 1228–1237.

15. Curreli F., Kwon Y. D., Zhang H., Scacalossi D., Belov D. S., …, Debnath A. K. Structure-based design of a small molecule CD4-antagonist with broad spectrum anti-HIV-1 activity. *Journal of Medicinal Chemistry*, 2015, vol. 58, pp. 6909–6927.

16. Durrant J. D., McCammon J. A. AutoClickChem: click chemistry in silico. *PLOS Computational Biology*, 2012, vol. 8, no. 3, e1002397. https://doi.org/10.1371/journal.pcbi.1002397

17. Lipinski C. A., Lombardo F., Dominy B. W., Feeney P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 2001, vol. 46, no. 1–3, pp. 3–26.

18. Alhossary A., Handoko S. D., Mu Y., Kwoh C. K. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics*, 2015, vol. 31, no. 13, pp. 2214–2216.

19. Kwong P. D., Wyatt R., Robinson J., Sweet R. W., Sodroski J., Hendrickson W. A. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, 1998, vol. 393, pp. 648–659.

20. Blahut R. E. *Theory and Practice of Error Control Codes*. Addison-Wesley, 1983, 500 p.

21. Tanimoto T. T. *IBM Internal Report 17th*. IBM Corp., Armonk, New York, 1957.

22. Myszka D. G., Sweet R. W., Hensley P., Brigham-Burke M., Kwong P. D., …, Doyle M. L. Energetics of the HIV gp120-CD4 binding reaction. *Proceedings of the National Academy of Sciences*, 2000, vol. 97, pp. 9026–9031.

23. Andrianov A. M., Nikolaev G. I., Kornoushenko Y. V., Xu W., Jiang S., Tuzikov A. V. In silico identification of novel aromatic compounds as potential HIV-1 entry inhibitors mimicking cellular receptor CD4. *Viruses*, 2019, vol. 11, E746. https://doi.org/10.3390/v11080746

24. Andrianov A. M., Nikolaev G. I., Kornoushenko Y. V., Huang J., Jiang S., Tuzikov A. V. Virtual screening and identification of potential HIV-1 inhibitors based on cross-reactive neutralizing antibody N6. *Doklady of the National Academy of Sciences of Belarus*, 2019, vol. 63, no. 4, pp. 445–456.

25. Andrianov A. M., Nikolaev G. I., Kornoushenko Y. V., Karpenko A. D., Huang J., Jiang S., Tuzikov A. V. Identification of functional mimetics of the neutralizing anti-HIV antibody N6 by virtual screening and

molecular modeling N6. *Doklady of the National Academy of Sciences of Belarus*, 2019, vol. 63, no. 5, pp. 561–571.

26. Andrianov A. M., Nikolaev G. I., Kornoushenko Y. V., Huang J., Jiang S., Tuzikov A. V. In silico identification of high-affinity ligands of the HIV-1 gp120 protein, potential peptidomimetics of neutralizing antibody N6. *Mathematical Biology and Bioinformatics*, 2019, vol. 14, no. 2, pp. 430–449.

27. Curreli F., Kwon Y. D., Belov D .S., Ramesh R. R., Kurkin A. V., …, Debnath A. K. Synthesis, antiviral potency, in vitro ADMET, and X-ray structure of potent CD4 mimics as entry inhibitors that target the Phe43 cavity of HIV-1 gp120. *Journal of Medicinal Chemistry*, 2017, vol. 60, pp. 3124–3153.

Информация об авторах

Николаев Григорий Игоревич, научный сотрудник, Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь.
E-mail: reshaemvsem@gmail.com

Шульдов Никита Андреевич, студент, Белорусский государственный университет, факультет прикладной математики и информатики, Минск, Беларусь.
E-mail: nickshuldov29@gmail.com

Анищенко Арсений Игоревич, студент, Белорусский государственный университет, факультет прикладной математики и информатики, Минск, Беларусь.
E-mail: BatsilaBox3@gmail.com

Тузиков Александр Васильевич, член-корреспондент, доктор физико-математических наук, профессор, директор, Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь.
E-mail: tuzikov@newman.bas-net.by

Андрианов Александр Михайлович, доктор химических наук, главный научный сотрудник, Институт биоорганической химии Национальной академии наук Беларуси, Минск, Беларусь.
E-mail: alexande.andriano@yandex.ru

Information about the authors

Grigory I. Nikolaev, Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus.
E-mail: reshaemvsem@gmail.com

Nikita A. Shuldov, Student, Belorussian State University, Faculty of Applied Mathematics and Computer Science, Minsk, Belarus.
E-mail: nickshuldov29@gmail.com

Arseny I. Anishenko, Student, Belorussian State University, Faculty of Applied Mathematics and Computer Science, Minsk, Belarus.
E-mail: BatsilaBox3@gmail.com

Alexander V. Tuzikov, Corresponding Member, Dr. Sci. (Phys.-Math.), Professor, Director, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus.
E-mail: tuzikov@newman.bas-net.by

Alexander M. Andrianov, Dr. Sci. (Chem.), Chief Researcher, Institute of Bioorganic Chemistry of the National Academy of Sciences of Belarus, Minsk, Belarus.
E-mail: alexande.andriano@yandex.ru