

БИОИНФОРМАТИКА

BIOINFORMATICS



УДК 004.94+547.7+616-006+616-085
<https://doi.org/10.37661/1816-0301-2023-20-3-7-20>

Оригинальная статья
Original Paper

Генеративная нейронная сеть на основе модели гетероэнкодера для de novo дизайна потенциальных противоопухолевых препаратов: применение к Vcr-Abl тирозинкиназе

А. Д. Карпенко¹, Т. Д. Войтко², А. В. Тузиков¹, А. М. Андрианов^{3✉}

¹Объединенный институт проблем информатики
Национальной академии наук Беларуси,
ул. Сурганова, 6, Минск, 220012, Беларусь

²Белорусский государственный университет,
пр. Независимости, 4, Минск, 220030, Беларусь

³Институт биоорганической химии
Национальной академии наук Беларуси,
ул. Купревича, 5/2, Минск, 220084, Беларусь

✉E-mail: alexande.andriano@yandex.ru

Аннотация

Цели. Решается задача разработки генеративной модели гетероэнкодера для компьютерного дизайна потенциальных ингибиторов Vcr-Abl тирозинкиназы – фермента, активность которого является патофизиологической причиной хронического миелоидного лейкоза.

Методы. На основе рекуррентных и полносвязных нейронных сетей прямого распространения создана генеративная модель гетероэнкодера. Проведены обучение и тестирование этой модели на наборе химических соединений, которые содержат 2-ариламинопиримид, присутствующий в качестве основного фармакофора в структурах многих низкомолекулярных ингибиторов протеинкиназ.

Результаты. Разработанная нейронная сеть апробирована в процессе генерации широкого набора новых молекул и последующего анализа их химического сродства к Vcr-Abl тирозинкиназе методами молекулярного докинга.

Заключение. Показано, что разработанная нейронная сеть представляет собой перспективную математическую модель для de novo дизайна малых молекул, которые потенциально активны против Vcr-Abl тирозинкиназы и могут быть использованы для разработки эффективных противоопухолевых препаратов широкого спектра действия.

Ключевые слова: методы машинного обучения, глубокое обучение, генеративные нейронные сети, гетероэнкодеры, Vcr-Abl тирозинкиназа, молекулярный докинг, противоопухолевые препараты, хронический миелоидный лейкоз

Для цитирования. Генеративная нейронная сеть на основе модели гетероэнкодера для de novo дизайна потенциальных противоопухолевых препаратов: применение к Bcr-Abl тирозинкиназе / А. Д. Карпенко [и др.] // Информатика. – 2023. – Т. 20, № 3. – С. 7–20. <https://doi.org/10.37661/1816-0301-2023-20-3-7-20>

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 30.05.2023

Подписана в печать | Accepted 28.07.2023

Опубликована | Published 29.09.2023

A generative neural network based on a hetero-encoder model for de novo design of potential anticancer drugs: application to Bcr-Abl tyrosine kinase

Anna D. Karpenko¹, Timofey D. Vaitko², Alexander V. Tuzikov¹, Alexander M. Andrianov³✉

¹The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus,
st. Surganova, 6, Minsk, 220012, Belarus

²Belarusian State University,
av. Nezavisimosti, 4, Minsk, 220030, Belarus

³Institute of Bioorganic Chemistry
of the National Academy of Sciences of Belarus,
st. Kuprevicha, 5/2, Minsk, 220084, Belarus

✉E-mail: alexande.andriano@yandex.ru

Abstract

Objectives. The problem of developing a generative hetero-encoder model for computer-aided design of potential inhibitors of Bcr-Abl tyrosine kinase, an enzyme whose activity is the pathophysiological cause of chronic myeloid leukemia, is being solved.

Methods. A generative hetero-encoder model was designed based on the recurrent and fully connected neural networks of direct propagation. Training and testing of this model were carried out on a set of chemical compounds containing 2-arylamino-pyrimidine, which is present as the main pharmacophore in the structures of many small-molecule inhibitors of protein kinases.

Results. The developed neural network was tested in the process of generating a wide range of new molecules and subsequent analysis of their chemical affinity for Bcr-Abl tyrosine kinase using molecular docking methods.

Conclusion. It is shown that the developed neural network is a promising mathematical model for de novo design of small molecules which are potentially active against Bcr-Abl tyrosine kinase and can be used to develop effective broad-spectrum anticancer drugs.

Keywords: machine learning methods, deep learning, generative neural networks, hetero-encoders, Bcr-Abl tyrosine kinase, molecular docking, anticancer drugs, chronic myeloid leukemia

For citation. Karpenko A. D., Vaitko T. D., Tuzikov A. V., Andrianov A. M. *A generative neural network based on a hetero-encoder model for de novo design of potential anticancer drugs: application to Bcr-Abl tyrosine kinase*. Informatika [Informatics], 2023, vol. 20, no. 3, pp. 7–20 (In Russ.). <https://doi.org/10.37661/1816-0301-2023-20-3-7-20>

Conflict of interest. The authors declare of no conflict of interest.

Введение. В настоящее время методы машинного обучения получили существенное развитие и используются для решения многих задач, связанных с разными областями науки и техники. Применение этих методов в био- и хемоинформатике, а также в медицинской химии позво-

лило ускорить процесс создания новых лекарственных препаратов и повысить эффективность программ фармацевтических исследований [1, 2]. Современные алгоритмы машинного обучения используются для прогнозирования фармакологических свойств малых молекул, получения информации о молекулярных механизмах белок-белковых и белок-лигандных взаимодействий, исследования количественных зависимостей «структура – активность» и «структура – свойство», предсказания структуры белков и аффинности связывания лигандов с молекулярной мишенью и виртуального скрининга потенциальных лекарств [1, 2]. Среди самых ярких достижений технологий искусственного интеллекта необходимо выделить разработанную британской компанией Google DeepMind глубокую нейронную сеть AlphaFold 2 [3, 4], в основе которой лежит новый подход к машинному обучению, использующий физические и биологические данные о структуре белков и информацию о множественном выравнивании их аминокислотных последовательностей. С помощью этой программы оказалось возможным предсказывать на атомном уровне трехмерные структуры белков по их первичной структуре. Данные о структурах белков депонируются в базе данных белков AlphaFold, которая включает на сегодняшний день более 2 млн белковых структур (URL: <https://alphafold.ebi.ac.uk>) [5]. Использование предсказательных моделей нейронных сетей для скрининга баз данных химических соединений позволило идентифицировать ряд антибактериальных и противовирусных средств, в том числе ингибиторов ВИЧ-1 и SARS-CoV-2 [6–8]. Эти модели были также успешно применены для скрининга одобренных Управлением по санитарному надзору за качеством пищевых продуктов и медикаментов США лекарственных препаратов, направленного на их перепрофилирование для терапии COVID-19 [8] и лекарственно устойчивых форм туберкулеза [9]. В частности, авторы работы [9] использовали нейронную сеть глубокого обучения для виртуального скрининга ряда библиотек лекарственных соединений и обнаружили молекулу галицина, которая структурно отличается от обычных антибиотиков и проявляет бактерицидную активность против широкого филогенетического спектра патогенов, включая *Mycobacterium tuberculosis* и резистентные к карбапенемам энтеробактерии. Результаты работы [9] наглядно продемонстрировали эффективность применения методов глубокого обучения для прогнозирования потенциальных лекарств и, в частности, для расширения набора структурно различных антибактериальных средств.

Разработка эффективных алгоритмов глубокого обучения дала толчок к развитию нового направления исследований, ориентированного на *de novo* дизайн молекул с заданными фармакологическими свойствами и синтетической доступностью [10–15]. На сегодняшний день предложено большое число генеративных моделей глубокого обучения, которые продемонстрировали перспективность их использования для генерации новых молекул-кандидатов в лекарственные средства [10–15]. В качестве успешных применений генеративных нейронных сетей следует отметить разработку ингибитора янус-киназы 3 и активных *in vivo* ингибиторов рецепторов доменов дискоидина 1 и 2 [15]. Однако очевидно, что, несмотря на значительный прогресс в развитии алгоритмов глубокого обучения, их потенциал в области фармацевтических исследований в полной мере еще не раскрыт. Поэтому создание генеративных моделей глубокого обучения с различными видами архитектур и типами входных данных и разными методами обучения имеет большое научное и практическое значение.

Настоящее исследование посвящено разработке генеративной нейронной сети глубокого обучения для *de novo* дизайна потенциальных ингибиторов Bcr-Abl тирозинкиназы – фермента, играющего ключевую роль в патогенезе хронического миелоидного лейкоза (ХМЛ), характеризующегося неконтролируемым ростом миелоидных клеток в периферической крови и костном мозге [16].

В клинической практике для терапии ХМЛ используются несколько ингибиторов Bcr-Abl тирозинкиназы прямого взаимодействия с АТФ-связывающим карманом фермента, среди которых в первую очередь следует отметить такие препараты, как иматиниб, нилотиниб, понатиниб, дазатиниб и бозутиниб [17–20]. Однако все эти соединения проявляют высокую токсичность, вызывающую ряд гематологических и негематологических побочных эффектов [21]. Кроме того, у большинства пациентов после длительной химиотерапии возникает резистентность

к применяемым препаратам [21]. В связи с этим актуальным является поиск новых ингибиторов Vcr-Abl тирозинкиназы, обладающих меньшей токсичностью и снижающих риск возможного возникновения резистентности к используемым препаратам, связанной с их длительным применением.

Цель настоящего исследования заключалась в разработке генеративной нейронной сети глубокого обучения на основе модели гетероэнкодера для конструирования новых потенциальных ингибиторов Vcr-Abl тирозинкиназы и ее мутантной формы Vcr-Abl(T315I), резистентной к ряду противоопухолевых препаратов, используемых для лечения пациентов с ХМЛ [22–24]. Для достижения этой цели были проведены исследования, которые включали следующие этапы:

- разработку архитектуры гетероэнкодера – усовершенствованной версии автоэнкодера, способной одновременно обрабатывать входные данные о молекуле в нескольких разных форматах с целью получения более стабильных и экономичных в поддержке генеративных моделей, применимых для химических соединений различных классов и дающих улучшенные по сравнению с автоэнкодерами результаты;
- формирование обучающей библиотеки малых молекул, содержащих 2-ариламинопиридин – фрагмент, присутствующий в качестве основного фармакофора в структурах многих низкомолекулярных ингибиторов протеинкиназ [22];
- обучение и тестирование нейронной сети на наборе соединений из сформированной молекулярной библиотеки;
- оценку результатов работы гетероэнкодера;
- генерацию с помощью разработанной нейронной сети набора малых молекул с заданной энергией связывания с терапевтической мишенью;
- построение методами молекулярного докинга комплексов сгенерированных гетероэнкодером соединений с Vcr-Abl тирозинкиназой и ее мутантной формой Vcr-Abl(T315I) и предсказание их потенциальной ингибиторной активности с помощью оценочных функций AutoDock Vina [25], NNScore 2.0 [26] и RF-Score 4 [27];
- анализ результатов молекулярного докинга и отбор соединений-лидеров, перспективных для разработки новых ингибиторов Vcr-Abl тирозинкиназы.

Архитектура модели гетероэнкодера. Разработанная нейронная сеть основана на архитектуре гетероэнкодера, представляющего собой автоэнкодер, предназначенный для решения задач, в которых входные данные представлены в нескольких разных форматах [28–30]. Такая архитектура нейронной сети позволяет получить более информативное латентное пространство за счет большего числа начальных признаков, что расширяет возможности поиска зависимостей между ними в процессе обучения гетероэнкодера [28]. В настоящем исследовании реализована модель гетероэнкодера с тремя энкодерами и двумя декодерами, которая использует открытую библиотеку Keras (URL: <https://keras.io>), обеспечивающую работу с искусственными нейронными сетями (рис. 1). В этой модели входные данные задаются в строковых форматах SMILES (Simplified Molecular Input Line Entry System) и канонический SMILES [31–33], а также числовым вектором характеристики молекулы (URL: <https://www.rdkit.org/docs/source/rdkit.Chem.Descriptors.htm>) (рис. 1).

С учетом специфики входных данных были разработаны две подмодели: в качестве энкодеров для строковых форматов SMILES и канонический SMILES была выбрана архитектура с двумя слоями LSTM (Long Short-Term Memory). Входные данные обрабатываются двумя слоями LSTM, состоящими из 128 ячеек каждый, и полученные эмбединги для строкового формата передаются на полносвязный слой (dense encoder) нейронной сети (рис. 1).

Числовые характеристики молекул обрабатываются полносвязной нейронной сетью прямого распространения, которая представлена энкодером, состоящим из двух полносвязных слоев с размерностью 64 и 32, слоя батч-нормализации и дополнительного полносвязного слоя из 16 нейронов, результаты работы которого являются эмбедингами для числовых признаков. Эти эмбединги попадают на конкатенирующий слой, где образуют один вектор, который нормализуется на слое батч-нормализации и передается на полносвязный слой из 128 нейронов,

после чего задается желаемое значение энергии связывания молекулы с терапевтической мишенью. Результаты работы этого слоя, т. е. обработанные эмбединги и величина энергии связывания, образуют латентное пространство размерностью 129 (рис. 1).

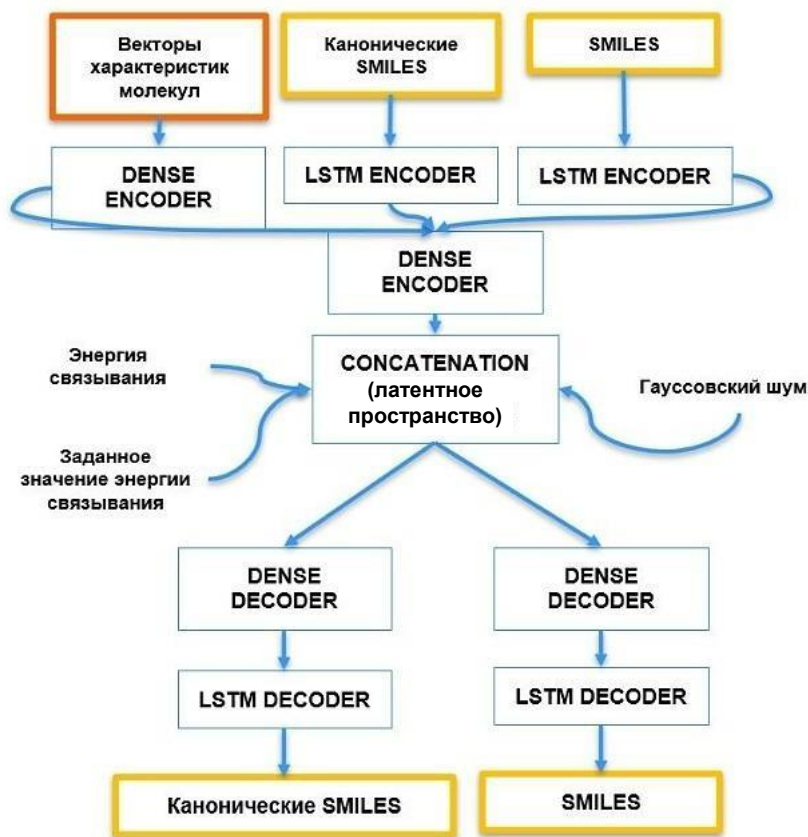


Рис. 1. Архитектура разработанной модели гетероэнкодера

Fig. 1. Architecture of the developed hetero-encoder model

В разработанную модель гетероэнкодера включены два одинаковых декодера (рис. 1), предназначенные для того, чтобы получить из векторов латентного пространства описание молекулы в двух строковых форматах. Декодеры функционируют следующим образом: вектор латентного пространства подается на два независимых полносвязных слоя размерностью 128 каждый и после их прохождения нормализуется на слоях батч-нормализации. На выходе генерируются два числовых вектора, которые передаются в качестве инициализирующих векторов на слой LSTM. На вход этого слоя дополнительно поступает строковый формат (для каждого слоя свой). Размерность слоя LSTM в декодерах также равна 128. После прохождения слоя LSTM данные передаются на полносвязный слой с функцией активации softmax, которая обрабатывает их таким образом, чтобы получить на выходе вероятности следующих символов. Для всех остальных полносвязных слоев используется функция активации ReLu, а для слоев LSTM – функция Tanh.

Разработанная модель гетероэнкодера имеет следующие особенности:

- во время подготовки входных данных в начало и конец строки добавляются символы для обучения слоев LSTM, поэтому на вход гетероэнкодера подается строка без последнего символа, а на выходе ожидается строка без первого символа;

- на латентный слой добавлен нейрон, позволяющий использовать в качестве дополнительного параметра свободную энергию связывания; этот нейрон не связан с энкодерами и используется только в декодерах для генерации молекул с желаемым химическим средством к терапевтической мишени;

- для более эффективного и стабильного обучения нейронной сети в ее кодирующей и декодирующей части используются слои батч-нормализации;
- на этапах кодирования и декодирования форматы данных не связаны друг с другом, что позволяет расширять архитектуру сети в случае необходимости ее перепрофилирования на другие терапевтические мишени;
- все энкодеры и декодеры обучаются вместе и одновременно в общей структуре гетероэнкодера.

Подготовка входных данных. Для формирования обучающей молекулярной библиотеки из базы данных PubChem (URL: <https://pubchem.ncbi.nlm.nih.gov/>) [34] были отобраны 120 000 соединений, содержащих ариламинопиримидин. Химические структуры этих соединений преобразовывали в форматы SMILES и канонический SMILES. Формат SMILES дает информацию о составе и химической структуре молекулы с использованием строки символов ASCII, а канонический SMILES представляет собой версию спецификации SMILES, включающую правила канонизации, которые позволяют записать формулу молекулы любого вещества однозначным образом. Эти правила касаются выбора первого атома в записи, направления обхода молекулярных циклов и выбора направления основной цепи молекулы при разветвлениях.

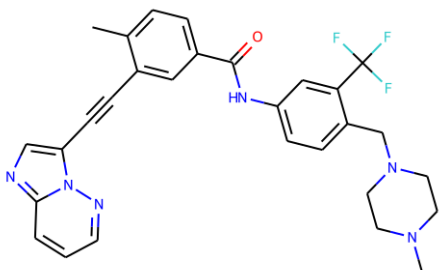
В табл. 1 дан пример описания химической структуры молекулы в форматах SMILES и канонический SMILES.

Таблица 1

Представление химической структуры молекулы в форматах SMILES и канонический SMILES

Table 1

Presentation of chemical molecular structure in SMILES and canonical SMILES formats

Химическая структура <i>Chemical structure</i>	
Химическая формула <i>Chemical formula</i>	C ₂₉ H ₂₇ F ₃ N ₆ O
Формат SMILES <i>SMILES format</i>	<chem>Cc1ccc(cc1C#Cc2cnc3n2cccc3)C(=O)Nc4ccc(c(c4)C(F)(F)F)CN5CCN(CC5)C</chem>
Канонический SMILES <i>Canonical SMILES</i>	<chem>Cc1ccc(C(=O)Nc2ccc(CN3CCN(C)CC3)c(C(F)(F)F)c2)cc1C#Cc1cnc2ccnnc12</chem>

Полученные молекулярные дескрипторы интегрировали в обучающую выборку, а затем преобразовали и отфильтровали с помощью процедуры, описанной ниже. Для каждой молекулы проверяли длины строковых форматов, и в тех случаях, когда они располагались вне диапазона 35–75 символов, молекулу удаляли из набора данных. Далее все атомы в строковой записи меняли на их односимвольные эквиваленты для предотвращения дополнительных трудностей при работе нейронной сети. Затем первые символы строк заменяли новым символом открытия строки, который до этого не встречался в выборке, и всем строкам дописывали символы завершения, причем таким образом, чтобы все строки после преобразования имели одинаковую длину. После этого строки преобразовывали в векторный формат. Сначала для каждого строкового формата извлекали уникальные символы и каждому из них присваивали уникальный индекс в рамках формата данных. После этого каждый символ строки заменяли числовым вектором с размерностью, равной числу уникальных символов в формате. Числовой вектор состоит из нулей и единственной единицы на месте индекса символа, т. е. каждая строка была представлена в виде матрицы, состоящей из нулей и единиц (этот метод также известен как One-Hot-

Encoding, URL: <https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>). В случае числовых эмбеддингов применяли процедуру стандартизации с целью уравновесить их влияние в процессе обучения. Таким образом, после фильтрации была получена выборка из 108 410 молекул в форматах, выбранных для обучения нейронной сети. Затем методом молекулярного докинга (программа AutoDock Vina, URL: <https://vina.scripps.edu>) [25] генерировали комплексы этих молекул со структурой Vcr-Abl тирозинкиназы в кристалле (URL: <https://www.rcsb.org>, PDB ID: 3OXZ) [35] и рассчитывали значения свободной энергии связывания. Молекулярный докинг проводили в приближении жесткого рецептора и гибких лигандов. Ячейка для докинга включала АТФ-связывающий сайт Vcr-Abl тирозинкиназы и имела следующие параметры: $\Delta X = 31 \text{ \AA}$, $\Delta Y = 23 \text{ \AA}$, $\Delta Z = 23 \text{ \AA}$ с центром в точках $X = 18 \text{ \AA}$, $Y = 8 \text{ \AA}$, $Z = 6 \text{ \AA}$. Значение параметра, характеризующего полноту поиска, задавали равным 100 [25]. Подготовленная обучающая библиотека объемом в 108 410 соединений и соответствующие им значения свободной энергии связывания сформировали набор данных для обучения и тестирования нейронной сети, который был разделен на тренировочный и тестовый поднаборы в пропорции 80 и 20 % соответственно от общего числа соединений.

Обучение гетерознкодера. Модель гетерознкодера включала 784 537 параметров (весов), из которых 781 369 параметров использовали для ее обучения. В процессе обучения применяли функцию потерь LF (Loss Function) следующего вида:

$$LF(s) = CCE(s) + 0,1 \cdot CCL(s),$$

где $CCE(s)$ – категориальная кросс-энтропия [35], s – молекула в формате SMILES, а $CCL(s)$ (CustomChemLoss) – функция, налагающая штрафы за нарушения стереохимии молекулы и отсутствие в ее структуре 2-ариламинопиримидина. Значение весового множителя штрафной функции выбирали путем перебора дискретного числа коэффициентов, направленного на определение такой величины этого параметра, при которой достигалась устойчивость обучения нейронной сети.

Категориальную кросс-энтропию $CCE(s)$ вычисляли по формуле

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

где $p(s_i)$ и $q(s_i)$ – соответственно истинная и предсказанная вероятности генерации символа s_i строки s .

Штрафную функцию $CCL(s)$ рассчитывали с помощью следующих критериев:

$$CCL(s) = \begin{cases} 0, & \text{если строка } s \text{ корректна и содержит 2-ариламинопирамидин;} \\ 1, & \text{если строка } s \text{ корректна, но не содержит 2-ариламинопирамидин;} \\ 5, & \text{если строка } s \text{ некорректна.} \end{cases}$$

В процессе обучения функция потерь для тренировочного набора изменялась в пределах от 1,867 до 1,0375, а для тестового набора – от 1,943 до 1,0445.

В качестве оптимизатора применяли метод стохастической оптимизации Адам [36]. Для обучения гетерознкодера использовали следующие параметры:

- коэффициент сохранения первого момента $\beta_1 = 0,9$;
- коэффициент сохранения второго момента $\beta_2 = 0,999$;
- сглаживающий параметр $\zeta = 10^{-7}$;
- объект, содержащий информацию о вычислительном узле $\eta = 0,005$;
- начальное значение скорости обучения 0,005;
- количество полных итераций обучения сети 25;
- размер подвыборки на одном шаге обучения 256.

Графики функции потерь для тренировочного и тестового наборов данных (рис. 2) свидетельствуют об их подобию и конечной сходимости, что позволяет сделать вывод об успешном обучении нейронной сети и отсутствии ее переобучения.

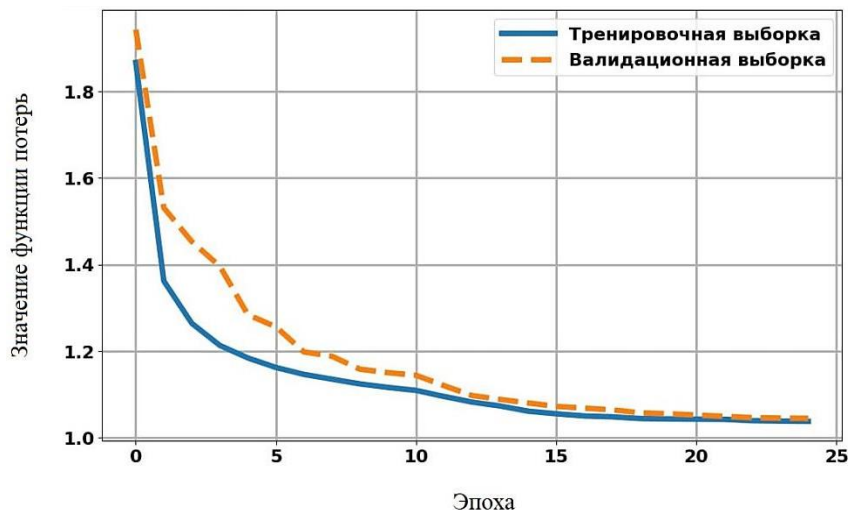


Рис. 2. Потери при обучении и валидации для разработанной модели гетероэнкодера

Fig. 2. Training and validation losses for the developed hetero-encoder model

Генерация соединений. Разработанную модель гетероэнкодера использовали для генерации широкого набора высокоаффинных лигандов Vcr-Abl тирозинкиназы с целью последующей идентификации потенциальных ингибиторов этого фермента методами молекулярного докинга. Для реализации процесса генерации с помощью кодирующей части модели получали представление латентного пространства из молекул обучающей библиотеки с энергией связывания ниже -9 ккал/моль. Далее в полученные векторы вносили некоторый стандартно распределенный шум для генерации новых латентных векторов, которые вместе с заданным значением энергии связывания подавались на декодирующую часть модели в качестве инициализирующих векторов, а стартовым символом для посимвольной генерации каждый раз являлся символ начала строки, добавленный ранее. Символы генерировались последовательно до получения символа окончания строки. В результате работы гетероэнкодера были получены линейные представления SMILES для 1117 молекул, которые очищали от дубликатов, проверяли на корректность, интерпретируемость и содержание 2-ариламинопиримидина с помощью модуля RDKit [37] и преобразовывали из формата SMILES в химические структуры. После процедуры фильтрации молекул были отобраны 1083 соединения, потенциальную ингибиторную активность которых против Vcr-Abl тирозинкиназы оценивали методом молекулярного докинга.

Оценка результатов работы гетероэнкодера. Для оценки эффективности работы гетероэнкодера с помощью программы AutoDock Vina (URL: <https://vina.scripps.edu>) были построены комплексы сгенерированных нейронной сетью соединений с рентгеновскими структурами Vcr-Abl тирозинкиназы (PDB ID: 3OXZ; URL: <https://www.rcsb.org>) и ее мутантной формы Vcr-Abl(T315I) (PDB ID: 3OY3; URL: <https://www.rcsb.org>) [35]. Молекулярный докинг выполняли по вычислительному протоколу, идентичному тому, который был использован при формировании обучающего набора данных. Согласно расчетным данным сгенерированные соединения имеют значения энергии связывания с нативной и мутантной Vcr-Abl тирозинкиназой, варьирующие в интервале от $-6,5$ до $-13,8$ ккал/моль (рис. 3).

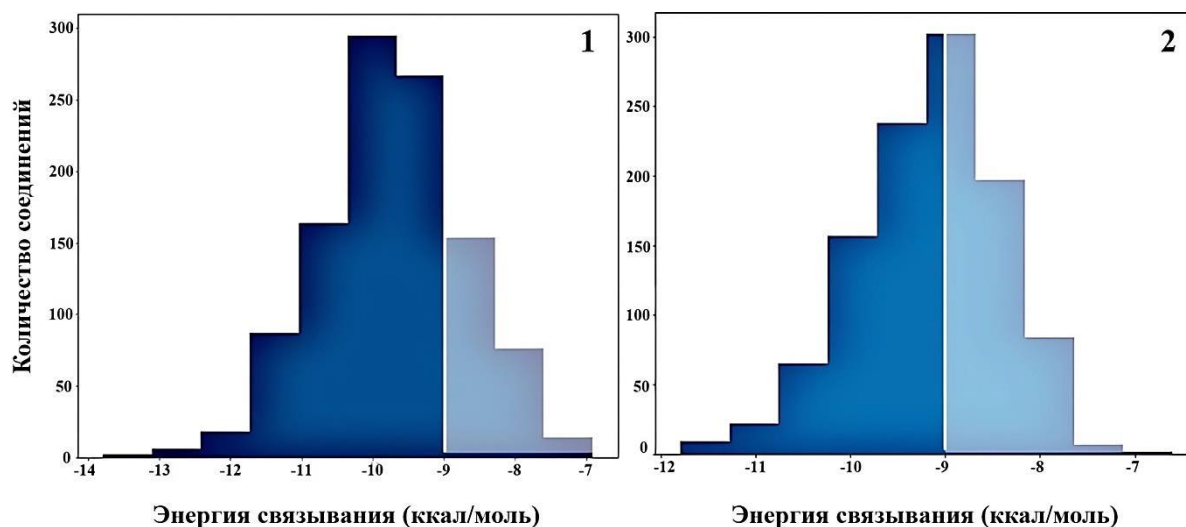


Рис. 3. Распределение энергии связывания сгенерированных соединений с нативной (1) и мутантной (Т315I) (2) тирозинкиназой

Fig. 3. Binding energy distribution of generated compounds with the native (1) and mutant (T315I) (2) tyrosine kinase

Для дальнейшего анализа были отобраны 569 молекул с энергией связывания от $-9,0$ до $-13,8$ ккал/моль, для которых проводили более точную оценку химического сродства к Bcr-Abl и Bcr-Abl(T315I) тирозинкиназе с помощью оценочных функций AutoDock Vina, NNScore 2.0 [26] и RF-Score-4 [27]. С этой целью для всех соединений определяли их ранги согласно каждой оценочной функции и на основе полученных данных вычисляли величину экспоненциального консенсусного ранга (exponential consensus rank, ECR) по формуле [38]

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

где $rank_{sf}$ – ранг соединения согласно оценочной функции sf ; σ_{sf} – параметр, контролирующий влияние оценочной функции sf на результаты консенсусного отбора (при расчетах ECR для всех рассматриваемых оценочных функций значение σ_{sf} задавали равным 10, предполагая, что их вклады в суммарную величину ECR одинаковы).

С целью идентификации соединений, потенциально активных против обеих терапевтических мишеней, для всех отобранных молекул рассчитывали перекрестный экспоненциальный консенсусный ранг ($crossECR$) по формуле

$$crossECR(i) = \frac{ECR_1(i)}{\max_i\{ECR_1(i)\}} + \frac{ECR_2(i)}{\max_i\{ECR_2(i)\}},$$

где $ECR_1(i)$ – значение ECR лиганда i для первой мишени (Bcr-Abl тирозинкиназы), а $ECR_2(i)$ – значение ECR лиганда i для второй мишени (Bcr-Abl(T315I) тирозинкиназы). Молекулы, имевшие наиболее низкие значения $crossECR$, относили к группе перспективных кандидатов на роль мультитаргетных противоопухолевых соединений, способных блокировать АТФ-связывающие сайты как Bcr-Abl тирозинкиназы, так и ее мутантной формы Bcr-Abl(T315I).

Анализ расчетных данных позволил идентифицировать четыре соединения-лидера, которые проявили высокое химическое сродство к нативной и мутантной (Т315I) тирозинкиназам. Химические структуры этих соединений показаны на рис. 4, а в табл. 2 приведены их физико-химические параметры, традиционно используемые в качестве основных фильтров для скрининга молекул на их способность быть эффективными при пероральном применении. Из данных табл. 2 следует, что соединения III и IV полностью удовлетворяют правилу пяти Липин-

ского, которое налагает ограничения на такие важные для потенциального лекарства характеристики, как всасывание, распределение, метаболизм и экскреция [39]. В то же время молекулы I и II обнаруживают лишь одно нарушение этого правила, связанное с небольшим превышением их молекулярной массы (табл. 2). Это позволяет предположить, что данные соединения также обладают лекарственными свойствами [39]. Идентифицированные соединения характеризуются низкими значениями свободной энергии связывания, предсказанными для комплексов лиганд/Bcr-Abl с помощью оценочных функций AutoDock Vina, NNScore 2.0 и RF-Score-4 (табл. 3). С учетом стандартной ошибки методов молекулярного докинга, составляющей ~ 2,9 ккал/моль [25], эти значения сопоставимы с величинами, рассчитанными для мощного противоопухолевого препарата понатиниба [18], использованного в качестве позитивного контроля (табл. 3).

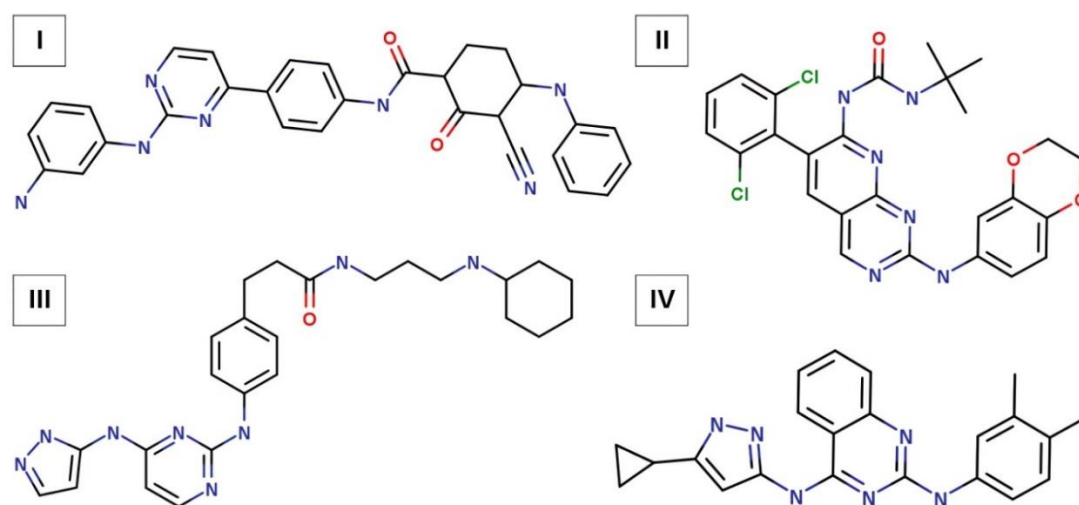


Рис. 4. Химические структуры идентифицированных соединений

Fig. 4. Chemical structures of identified compounds

Таблица 2

Физико-химические параметры идентифицированных соединений – потенциальных ингибиторов Bcr-Abl тирозинкиназы и ее мутантной формы Bcr-Abl(T315I)

Table 2

Physicochemical parameters of identified compounds, potential inhibitors of Bcr-Abl tyrosine kinase and its mutant form Bcr-Abl(T315I)

Соединение <i>Compound</i>	Химическая формула <i>Chemical formula</i>	Молекулярная масса (Да) <i>Molecular mass (Da)</i>	LogP	Число доноров водородной связи <i>Number of H-bond donors</i>	Число акцепторов водородной связи <i>Number of H-bond acceptors</i>
I	C ₃₀ H ₂₇ N ₇ O ₂	517,58	3,28	4	5
II	C ₂₆ H ₂₄ C ₁₂ N ₆ O ₃	539,41	4,79	3	6
III	C ₂₅ H ₃₄ N ₈ O	462,59	3,31	5	5
IV	C ₂₂ H ₂₂ N ₆	370,45	4,31	3	3

Примечание: приведенные данные получены с помощью веб-сервера SwissADME (URL: <http://www.swissadme.ch>), LogP – липофильность соединения.

Note: the given data were obtained using the SwissADME web server (URL: <http://www.swissadme.ch>), LogP – lipophilicity of the compound.

Таблица 3

Значения crossECR и энергии связывания (ккал/моль) для четырех сгенерированных нейронной сетью соединений I-IV и понатиниба (V) в комплексах с Bcr-Abl тирозинкиназой и ее мутантной формой Bcr-Abl(T315I)

Table 3

CrossECR values and binding energies (kcal/mol) for four neural network-generated compounds I–IV and ponatinib (V) in the complexes with Bcr-Abl tyrosine kinase and its mutant form Bcr-Abl(T315I)

Соединение <i>Compound</i>	Значение crossECR <i>CrossECR</i> value	Энергия связывания <i>Binding energy</i>					
		Bcr-Abl тирозинкиназы <i>Bcr-Abl tyrosine kinase</i>			Bcr-Abl(T315I) тирозинкиназы <i>Bcr-Abl(T315I) tyrosine kinase</i>		
		AutoDock Vina	RF-Score-4	NNScore 2.0	AutoDock Vina	RF-Score-4	NNScore 2.0
I	0,0674	-13,8	-11,5	-9,8	-11,3	-11,3	-8,9
II	0,0674	-13,0	-11,6	-10,1	-11,0	-11,5	-9,0
III	0,0835	-10,4	-11,6	-11,7	-10,0	-11,3	-11,3
IV	0,0931	-13,4	-11,3	-9,3	-11,4	-11,4	-5,8
V	0,0399	-12,0	-11,4	-12,2	-12,2	-11,2	-12,4

Полученные результаты показывают, что разработанная нейронная сеть представляет собой перспективную математическую модель для *de novo* дизайна малых молекул, которые потенциально активны против Bcr-Abl тирозинкиназы и ее мутантной формы Bcr-Abl(T315I) и могут быть использованы для разработки эффективных противоопухолевых препаратов широкого спектра действия.

Расчеты проводились с помощью вычислительной системы, имеющей следующие характеристики: Intel(R) Xeon(R) Platinum 8259CL CPU @ 2.50GHz x2, 12 GB RAM, GPU NVIDIA T4 16GB Memory, 2560 Cores, 160 TMUS.

Заключение. На основе рекуррентных и полносвязных нейронных сетей прямого распространения разработана модель гетероэнкодера для генерации новых потенциальных ингибиторов Bcr-Abl тирозинкиназы – фермента, играющего важную роль в развитии ХМЛ. Проведены обучение и тестирование нейронной сети, выполнен анализ результатов ее работы. В процессе апробации нейронной сети сгенерированы 1083 молекулы, химическое сродство которых к активным центрам нативной и мутантной (T315I) тирозинкиназ исследовано методом молекулярного докинга. В результате анализа полученных данных идентифицированы четыре соединения-лидера, представляющие значительный интерес для проведения дальнейших экспериментальных и теоретических исследований, включающих химический синтез молекул, биомедицинские испытания *in vitro* и оптимизацию их структур методами QSAR [40, 41], направленную на получение аналогов с улучшенной противоопухолевой активностью и приемлемыми фармакологическими свойствами.

Работа выполнена при поддержке Государственной программы научных исследований «Конвергенция 2025» (подпрограмма «Междисциплинарные исследования и новые технологии»), задание 3.04.1).

Вклад авторов. А. Д. Карпенко и Т. Д. Войтко разработали и реализовали архитектуру генеративной модели гетероэнкодера, обучили, протестировали и апробировали нейронную сеть. А. Д. Карпенко провела молекулярный докинг сгенерированных гетероэнкодером соединений с Bcr-Abl тирозинкиназой. А. М. Андрианов и А. В. Тузиков осуществляли руководство проектом и написали рукопись. Все авторы анализировали данные расчетов, обсуждали полученные результаты и внесли свой вклад в окончательную версию статьи.

References

1. 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, no. 6, pp. 463–477. <https://doi.org/10.1038/s41573-019-0024-5>
2. Lipinski C. F., Maltarollo V. G., Oliveira P. R., da Silva A. B. F., Honorio K. M. Advances and perspectives in applying deep learning for drug design and discovery. *Frontiers in Robotics and AI*, 2019, vol. 6, art. 108. Available at: <https://www.frontiersin.org/articles/10.3389/frobt.2019.00108/full> (accessed 07.08.2023). <https://doi.org/10.3389/frobt.2019.00108>
3. Cramer P. AlphaFold2 and the future of structural biology. *Nature Structural & Molecular Biology*, 2021, vol. 28, no. 9, pp. 704–705.
4. Bryant P., Pozzati G., Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, 2022, vol. 13, no. 1, art. 1265. Available at: <https://www.nature.com/articles/s41467-022-29480-5> (accessed 07.08.2023). <https://doi.org/10.1038/s41467-022-29480-5>
5. David A., Islam S., Tankhilevich E., Sternberg M. J. The AlphaFold database of protein structures: a biologist's guide. *Journal of Molecular Biology*, 2022, vol. 434, no. 2, p. 167336.
6. Timmons P. B., Hewage C. M. ENNAVIA is a novel method which employs neural networks for antiviral and anti-coronavirus activity prediction for therapeutic peptides. *Briefings in Bioinformatics*, 2021, vol. 22, iss. 6, art. bbab258. Available at: <https://academic.oup.com/bib/article/22/6/bbab258/6326528> (accessed 07.08.2023). <https://doi.org/10.1093/bib/bbab258>
7. Andrianov A. M., Nikolaev G. I., Shuldov N. A., Bosko I. P., Anischenko A. I., Tuzikov A. V. Application of deep learning and molecular modeling to identify small drug-like compounds as potential HIV-1 entry inhibitors. *Journal of Biomolecular Structure and Dynamics*, 2022, vol. 40, no. 16, pp. 7555–7573. <https://doi.org/10.1080/07391102.2021.1905559>
8. Zhang Y., Ye T., Xi H., Juhas M., Li J. Deep learning driven drug discovery: Tackling Severe Acute Respiratory Syndrome Coronavirus 2. *Frontiers in Microbiology*, 2021, vol. 12. Available at: <https://www.frontiersin.org/articles/10.3389/fmicb.2021.739684/full> (accessed 07.08.2023). <https://doi.org/10.3389/fmicb.2021.739684>
9. Stokes J. M., Yang K., Swanson K., Jin W., Cubillos-Ruiz A., ..., Collins J. J. A deep learning approach to antibiotic discovery. *Cell*, 2020, vol. 180, no. 4, art. e13, pp. 688–702. <https://doi.org/10.1016/j.cell.2020.01.021>
10. Mercado R., Rastemo T., Lindelöf E., Klambauer G., Engkvist O., ..., Bjerrum E. J. Practical notes on building molecular graph generative models. *ChemRxiv*, 2020. Available at: <https://chemrxiv.org/engage/chemrxiv/article-details/60c74f55567dfe705bec5672> (accessed 07.08.2023). <https://doi.org/10.26434/chemrxiv.12888383>
11. Arús-Pous J., Blaschke T., Ulander S., Reymond J. L., Chen H., Engkvist O. Exploring the GDB-13 chemical space using deep generative models. *Journal of Cheminformatics*, 2019, vol. 11, art. 20. Available at: <https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0341-z> (accessed 07.08.2023). <https://doi.org/10.1186/s13321-019-0341-z>
12. Prykhodko O., Johansson S. V., Kotsias P. C., Arús-Pous J., Bjerrum E. J., ..., Chen H. A de novo molecular generation method using latent vector based generative adversarial network. *Journal of Cheminformatics*, 2019, vol. 11, no 1, art. 74. Available at: <https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0397-9> (accessed 07.08.2023). <https://doi.org/10.1186/s13321-019-0397-9>
13. Polykovskiy D., Zhebrak A., Vetrov D., Ivanenkov Y., Aladinskiy V., ..., Kadurin A. Entangled conditional adversarial autoencoder for de novo drug discovery. *Molecular Pharmaceutics*, 2018, vol. 15, no. 10, pp. 4398–4405. <https://doi.org/10.1021/acs.molpharmaceut.8b00839s>
14. Zhang J., Mercado R., Engkvist O., Chen H. Comparative study of deep generative models on chemical space coverage. *Journal of Chemical Information and Modeling*, 2021, vol. 61, no. 6, pp. 2572–2581. <https://doi.org/10.26434/chemrxiv.13234289.v1>
15. Zhavoronkov A., Ivanenkov Y. A., Aliper A., Veselov M. S., Aladinskiy V. A., ..., Aspuru-Guzik A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, 2019, vol. 37, no. 9, pp. 1038–1040. <https://doi.org/10.1038/s41587-019-0224-x>
16. Köstler W. J., Zielinski C. C. Targeting receptor tyrosine kinases in cancer. *Receptor Tyrosine Kinases: Structure, Functions and Role in Human Disease*. New York, Springer, 2015, pp. 225–278.
17. Kantarjian H. M., Hochhaus A., Saglio G., De Souza C., Flinn I. W., ..., Hughes T. P. Nilotinib versus imatinib for the treatment of patients with newly diagnosed chronic phase, Philadelphia chromosome-positive, chronic myeloid leukaemia: 24-month minimum follow-up of the phase 3 randomised ENESTnd trial. *The Lancet Oncology*, 2011, vol. 12, no. 9, pp. 841–851. [https://doi.org/10.1016/S1470-2045\(11\)70201-7](https://doi.org/10.1016/S1470-2045(11)70201-7)

18. Tan F. H., Putoczki T. L., Stylli S. S., Luwor R. B. Ponatinib: a novel multi-tyrosine kinase inhibitor against human malignancies. *OncoTargets and Therapy*, 2019, vol. 12, pp. 635–645. <https://doi.org/10.2147/OTT.S189391>
19. O'Hare T. A decade of nilotinib and dasatinib: From in vitro studies to first-line tyrosine kinase inhibitors. *Cancer Research*, 2016, vol. 76, no. 20, pp. 5911–5913. <https://doi.org/10.1158/0008-5472.CAN-16-2483>
20. Brümmendorf T. H., Cortes J. E., de Souza C. A., Guilhot F., Duvillié L., ..., Gambacorti-Passerini C. Bosutinib versus imatinib in newly diagnosed chronic-phase chronic myeloid leukaemia: Results from the 24-month follow-up of the BELA trial. *British Journal of Haematology*, 2015, vol. 168, no. 1, pp. 69–81. <https://doi.org/10.1111/bjh.13108>
21. Bhullar K. S., Lagarón N. O., McGowan E. M., Parmar I., Jha A., ..., Rupasinghe H. P. V. Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular Cancer*, 2018, vol. 17, art. 48. Available at: <https://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-018-0804-2> (accessed 07.08.2023). <https://doi.org/10.1186/s12943-018-0804-2>
22. Koroleva E. V., Ignatovich Zh. I., Sinyutich Yu. V., Gusak K. N. Aminopyrimidine derivatives as protein kinases inhibitors. Molecular design, synthesis, and biologic activity. *Russian Journal of Organic Chemistry*, 2016, vol. 52, no. 2, pp. 139–177. <https://doi.org/10.1134/S1070428016020019>
23. Patel A. B., O'Hare T., Deininger M. W. Mechanisms of resistance to ABL kinase inhibition in CML and the development of next generation ABL kinase inhibitors. *Hematology/Oncology Clinics of North America*, 2017, vol. 31, no. 4, pp. 589–612. <https://doi.org/10.1016/j.hoc.2017.04.007>
24. Liu J., Zhang Y., Huang H., Lei X., Tang G., ..., Peng J. Recent advances in Bcr-Abl tyrosine kinase inhibitors for overriding T315I mutation. *Chemical Biology and Drug Design*, 2021, vol. 97, no. 3, pp. 649–664. <https://doi.org/10.1111/cbdd.13801>
25. Trott O., Olson A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 2010, vol. 31, no. 2, pp. 455–461. <https://doi.org/10.1002/jcc.21334>
26. Durrant J. D., McCammon J. A. NNScore 2.0: A neural-network receptor-ligand scoring function. *Journal of Chemical Information and Modeling*, 2011, vol. 51, no. 11, pp. 2897–2903. <https://doi.org/10.1021/ci2003889>
27. Wójcikowski M., Ballester P. J., Siedlecki P. Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific Reports*, 2017, vol. 7, no. 1, pp. 1–10.
28. Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks. *Science*, 2006, vol. 313, no. 5786, pp. 504–507.
29. Hwang M., Qian Y., Wu C., Jiang W. C., Wang D., ..., Hwang K. S. A local region proposals approach to instance segmentation for intestinal polyp detection. *International Journal of Machine Learning and Cybernetics*, 2023, vol. 14, no. 5, pp. 1591–1603.
30. Huang A., Ju X., Lyons J., Murnane D., Pettee M., Reed L. *Heterogeneous Graph Neural Network for Identifying Hadronically Decayed Tau Leptons at the High Luminosity LHC*. Available at: <https://arxiv.org/pdf/2301.00501.pdf> (accessed 07.08.2023).
31. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 1988, vol. 28, no. 1, pp. 31–36. <https://doi.org/10.1021/ci00057a005>
32. Weininger D., Weininger A., Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, 1989, vol. 29, no. 2, pp. 97–101.
33. O'Boyle N. M. Towards a Universal SMILES representation-A standard method to generate canonical SMILES based on the InChI. *Journal of Cheminformatics*, 2012, vol. 4, art. 22, pp. 1–14.
34. Kim S., Chen J., Cheng T., Gindulyte A., He J., ..., Bolton E. E. PubChem 2019 update: improved access to chemical data. *Nuclear Acids Research*, 2019, vol. 47(D1), pp. D1102–D1109.
35. Ho Y., Wookey S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access*, 2019, vol. 8, pp. 4806–4813.
36. Kingma D. P., Ba J. *Adam: A Method for Stochastic Optimization*, 2014. Available at: <https://arxiv.org/pdf/1412.6980.pdf> (accessed 07.08.2023).
37. Landrum G. *RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling*, 2013. Available at: https://www.rdkit.org/RDKit_Overview.pdf (accessed 07.08.2023).

38. Palacio-Rodríguez K., Lans I., Cavasotto C. N., Cossio P. Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Scientific Reports*, 2019, vol. 9, no. 1, art. 5142. Available at: <https://www.nature.com/articles/s41598-019-41594-3> (accessed 07.08.2023). <https://doi.org/10.1038/s41598-019-41594-3>

39. Lipinski C. A. Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 2004, vol. 1, no. 4, pp. 337–341.

40. Verma J., Khedkar V. M., Coutinho E. C. 3D-QSAR in drug design-a review. *Current Topics in Medicinal Chemistry*, 2010, vol. 10, no. 1, pp. 95–115. <https://doi.org/10.2174/156802610790232260>

41. Kuseva C., Schultz T. W., Yordanova D., Tankova K., Kutsarova S., ..., Mekenyan O. G. The implementation of RAAF in the OECD QSAR Toolbox. *Regulatory Toxicology and Pharmacology*, 2019, vol. 105, pp. 51–61. <https://doi.org/10.1016/j.yrtph.2019.03.018>

Информация об авторах

Карпенко Анна Дмитриевна, научный сотрудник, Объединенный институт проблем информатики Национальной академии наук Беларуси.
E-mail: rfe.karpenko@gmail.com

Войтко Тимофей Дмитриевич, студент, Белорусский государственный университет.
E-mail: timvaitko@gmail.com

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

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

Information about the authors

Anna D. Karpenko, Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: rfe.karpenko@gmail.com

Timofey D. Vaitko, Student, Belarusian State University.
E-mail: timvaitko@gmail.com

Alexander V. Tuzikov, Corresponding Member, D. Sc. (Phys.-Math.), Prof., Head of the Laboratory of Mathematical Cybernetics, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: tuzikov@newman.bas-net.by

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