

===== БИОИНФОРМАТИКА =====

Разработка генеративной нейронной сети глубокого обучения для компьютерного дизайна потенциальных ингибиторов коронавируса SARS-CoV-2

Шульдов Н.А.^{1,2}, Юшкевич А.М.¹, Фурс К.В.¹, Тузиков А.В.¹,
Андрианов А.М.^{*3}

¹Объединенный институт проблем информатики, Национальная академия наук
Беларуси, Минск, Республика Беларусь

²EPAM Systems, Минск, Республика Беларусь

³Институт биоорганической химии, Национальная академия наук Беларуси, Минск,
Республика Беларусь

Аннотация. Разработаны две генеративные модели глубокого обучения для компьютерного дизайна потенциальных ингибиторов основной протеазы (M^{Pro}) коронавируса SARS-CoV-2 – фермента, критически важного для репликации и транскрипции вируса и, поэтому, представляющего перспективную мишень для конструирования эффективных противовирусных препаратов. Для решения этой задачи сформирована обучающая библиотека малых молекул, содержащих элементы структуры, способные обеспечить специфические и эффективные взаимодействия потенциальных лигандов с каталитическим сайтом M^{Pro} SARS-CoV-2; разработана и реализована архитектура генеративных моделей, позволяющих генерировать новые высокоаффинные лиганды этого функционально важного белка SARS-CoV-2; проведено обучение и тестирование нейронной сети на соединениях из обучающей библиотеки и выполнена оценка результатов ее обучения и работы в двух разных режимах генерации. Использование генеративных моделей совместно с методом молекулярного докинга продемонстрировало их большой потенциал для заполнения неисследованных областей химического пространства новыми молекулами с заданными свойствами, что подтверждается полученными результатами, согласно которым из 4805 сгенерированных нейронной сетью соединений только одно присутствовало в исходном наборе данных.

Ключевые слова: методы машинного обучения, глубокое обучение, генеративные нейронные сети, коронавирус SARS-CoV-2, основная протеаза, виртуальный скрининг, молекулярный докинг, противовирусные препараты.

ВВЕДЕНИЕ

В настоящее время компьютерный дизайн потенциальных лекарств является важным инструментом, позволяющим значительно сократить время и затраты, необходимые для разработки новых терапевтических средств [1, 2]. В последние годы для разработки лекарств все чаще используются методы машинного обучения и, в частности, глубокого обучения, которые применяются на каждом этапе этого сложного многостадийного процесса не только для ускорения исследований, но и для оценки рисков и затрат в клинических испытаниях [3, 4]. Эти методы позволяют устанавливать математические взаимосвязи между эмпирическими данными о свойствах малых

*alexande.andriano@yandex.ru

молекул и экстраполировать их для прогнозирования физико-химических и биологических характеристик новых соединений [4–6]. С помощью машинного обучения можно исследовать количественные зависимости структура-активность (QSAR, Quantitative Structure-Activity Relationship) или структура-свойство (QSPR, Quantitative Structure-Property Relationship) и разрабатывать методы, которые могут предсказывать с высокой точностью влияние химических модификаций соединения на его биологическую активность, фармакокинетические и токсикологические характеристики [4–6]. Кроме того, методы машинного обучения могут быть успешно применены для решения задач перепрофилирования лекарств [7, 8], предсказания структуры белков [9], виртуального скрининга потенциальных лекарств и прогнозирования аффинности связывания лигандов с молекулярной мишенью [10–12]. В недавних работах методы глубокого обучения были использованы для скрининга баз данных химических соединений с целью идентификации антибактериальных и противовирусных ингибиторов, в том числе лекарственных веществ, потенциально активных в отношении ВИЧ-1 и SARS-CoV-2 [12–15]. Применение этих методов в комбинации с виртуальным скринингом молекулярных библиотек клинически одобренных лекарств привело к обнаружению ряда потенциальных препаратов против SARS-CoV-2, таких как Атазанавир, Ремдесивир, Калетра, Эналаприлат, Венетоклакс, Посаконазол, Даклатасвир, Омбитасвир, Торемифен, Никлозамид, Дексаметазон, Индометацин, Пралатрексат, Азитромицин, Палматин, и Саухиноне [15]. Кроме прогнозирования физико-химических свойств молекул и идентификации потенциальных лекарств в химических базах данных, машинное обучение также используется для создания новых соединений с заданными фармакологическими характеристиками [16–21]. Несмотря на то, что традиционный виртуальный скрининг молекулярных библиотек, основанный на подобию физико-химических характеристик соединений [22], предоставляет широкие возможности для идентификации новых потенциальных лекарств, он имеет определенные недостатки по сравнению с генеративными моделями. Одним из основных стимулов для использования генеративных моделей является более широкое исследование пространства молекулярных признаков. Поиск по подобию обеспечивает исследование химического пространства, ограниченного разнообразием имеющихся в распоряжении соединений, в то время как генеративные модели позволяют охватить молекулярное пространство признаков гораздо более широкого химического разнообразия. Вторая причина преимущества генеративных моделей – возможность наложения дополнительных условий в процессе генерации. Это позволяет проводить направленную генерацию новых молекул из химического пространства вместо их генерации «вслепую». Задаваемая величина свободной энергии связывания молекулы с молекулярной мишенью может быть использована в качестве критерия, позволяющего генерировать высокоаффинные лиганды из интересующего подмножества исследуемого химического пространства.

В последние годы генеративные модели глубокого обучения нашли широкое применение в исследованиях по разработке лекарств *de novo* [16–21]. Благодаря огромному прогрессу методов глубокого обучения в настоящее время разработаны генеративные модели с различной архитектурой и разными методами обучения, использующие разные типы и структуры данных [16–21]. Применение генеративных моделей глубокого обучения уже показало их способность генерировать молекулы, которые могут быть синтезированы, активны *in vitro*, стабильны и проявляют активность *in vivo* в моделях, связанных с различными заболеваниями [15]. В частности, в работе [19] с помощью условно-состязательного автоэнкодера был разработан ингибитор янус-киназы 3 (JAK3), представляющий новый класс иммуномодулирующих агентов. Кроме того, с использованием генеративного тензорного обучения с подкреплением были разработаны активные *in vivo* ингибиторы

рецепторов доменов дискоидина 1 и 2 (DDR1 и DDR2), а фармакокинетический профиль DDR1 был подтвержден экспериментами на мышах *in vivo* [21].

Однако, несмотря на то что генеративные модели глубокого обучения становятся все более распространенными в био- и хемоинформатике, их потенциал в этой области еще не раскрыт в полном объеме. В связи с этим разработка и применение генеративных методов глубокого обучения для компьютерного конструирования потенциальных лекарственных препаратов имеют большое научное и практическое значение.

Настоящее исследование посвящено разработке генеративного автоэнкодера на основе представления химических структур молекул в формате упрощенной системы линейного описания SMILES (Simplified Molecular Input Line Entry System) [23]. Одной из основных идей обучения модели в формате SMILES, а точнее, представления, получаемого векторизацией SMILES, была возможность преобразования данных этого формата в химические структуры молекул. Как было показано в исследовании [14], генеративные модели обеспечивают хорошую основу для обогащения результатов скрининга, однако использование молекулярных дескрипторов (фингерпринтов) может затруднить восстановление самих химических структур. Но описания молекул на основе SMILES могут стать необходимой альтернативой фингерпринтам в подходах хемоинформатики, когда важна возможность восстановления структуры химических соединений из их сжатого представления, используемого в генеративных моделях. Именно поэтому описания молекул в формате SMILES были выбраны в качестве архитектурной основы для разработки генеративных моделей глубокого обучения, предназначенных для генерации потенциальных ингибиторов выбранной белковой мишени.

Цель исследования: разработать генеративные модели глубокого обучения для рационального дизайна новых потенциальных ингибиторов основной протеазы (M^{Pro} ; Main Protease) SARS-CoV-2 – фермента, критически важного для репликации и транскрипции вируса и, поэтому, представляющего перспективную мишень для конструирования эффективных противовирусных препаратов [24].

В ряде исследований основная протеаза M^{Pro} SARS-CoV-2 использовалась в качестве мишени для скрининга клинически одобренных препаратов в качестве потенциальных ингибиторов вируса в надежде найти препараты, эффективные против COVID-19 [25]. Так как профили безопасности этих препаратов хорошо документированы, такой подход, сочетающий в себе структурный дизайн лекарств с методами виртуального скрининга и молекулярного моделирования, может существенно облегчить и ускорить процесс обнаружения противовирусных соединений с клиническим потенциалом с целью их перепрофилирования для лечения пациентов, инфицированных коронавирусом нового типа. Анализ литературы также показал, что в настоящее время проводятся исследования природных ингибиторов M^{Pro} , происходящих в основном из растений, морских организмов и микроорганизмов [25]. Кроме того, изучаются ковалентные пептидомиметики и малые молекулы, и при этом различные соединения проявляют противовирусную активность в инфицированных клетках человека [26]. Однако, принимая во внимание мутации SARS-CoV-2 [27], чрезвычайно актуальными также являются исследования по разработке новых противовирусных соединений, способных блокировать функционально важные сайты белков вируса COVID-19.

Для решения поставленной задачи были выполнены исследования, включающие:

1) формирование обучающей библиотеки малых молекул, содержащих элементы структуры, способные обеспечить специфические и эффективные взаимодействия потенциальных лигандов с каталитическим сайтом M^{Pro} SARS-CoV-2;

2) разработку и реализацию архитектуры генеративных моделей глубокого обучения, позволяющих генерировать новые высокоаффинные ингибиторы M^{Pro} SARS-CoV-2;

3) обучение и тестирование нейронной сети на соединениях из сформированной обучающей библиотеки;

4) оценку результатов обучения и работы автоэнкодера в двух разных режимах генерации.

МЕТОДЫ ИССЛЕДОВАНИЯ

Разработка генеративной нейронной сети на основе технологий глубокого обучения

Подготовка обучающего набора данных

Разработанный генеративный автоэнкодер построен таким образом, чтобы быть специфичным для целевого белка, и, следовательно, набор обучающих данных должен включать соединения, потенциально активные против выбранной молекулярной мишени. Как отмечалось выше, разработанную нейронную сеть адаптировали для генерации потенциальных ингибиторов M^{Pro} SARS-CoV-2, и, поэтому, для подготовки обучающего набора данных предварительно была сформирована виртуальная библиотека потенциальных анти-SARS-CoV-2 агентов, способных эффективно связываться и блокировать каталитический сайт фермента. Процедура подготовки этой молекулярной библиотеки была следующей:

а) Построение фармакофорных моделей и виртуальный скрининг химических баз данных

На первом этапе были построены фармакофорные модели, описывающие совокупность структурно-функциональных признаков, которые обеспечивают высокую аффинность связывания с M^{Pro} ряда известных ингибиторов генетически родственного SARS-CoV-2 вируса атипичной пневмонии SARS-CoV [24]. Эти ингибиторы, химические структуры которых заимствовали из работы [24], делили на группы в соответствии с их принадлежностью к определенному классу, а затем фармакофорные признаки репрезентативных представителей каждой группы усредняли с помощью веб-сервера PharmaGist [28]. В результате данного подхода были построены 16 фармакофорных моделей, соответствующих 16 группам ингибиторов M^{Pro} SARS-CoV, которые включали 6 классов пептидомиметиков и 10 классов низкомолекулярных химических соединений [24]. Полученный таким образом набор входных данных для проведения виртуального скрининга был дополнен фармакофорной моделью химического соединения X77, представляющего собой мощный нековалентный ингибитор как SARS-CoV, MERS-CoV, так и SARS-CoV-2 (<https://www.medchemexpress.com/x77.html>). Фармакофорную модель ингибитора X77 генерировали с помощью программного обеспечения веб-сервера Pharmit [29] с привлечением в качестве входных данных структуры комплекса X77–M^{Pro} в кристалле (Банк данных белков; PDB ID: 6w63; <https://www.rcsb.org>).

Для идентификации потенциальных ингибиторов M^{Pro} SARS-CoV-2 проводили фармакофорный поиск в 9 химических базах данных веб-сервера Pharmit (<http://pharmit.csb.pitt.edu/>) [29], содержащих более 213.5 миллионов химических структур. В результате проведенных исследований был получен набор из 711 102 соединений, удовлетворявших одной из 17 построенных фармакофорных моделей.

б) Молекулярный докинг

Соединения, идентифицированные в результате фармакофорного поиска, подвергали предварительному молекулярному докингу с M^{Pro} SARS-CoV-2. Затем эти соединения “фильтровали” на основе значений оценочной функции докинга с пороговой величиной, равной -7 ккал/моль, что соответствует стандартному порогу полуэффективной концентрации 10 мкМ, часто используемому при *in vitro* скрининге потенциальных лекарств [30]. Для 353 467 соединений, удовлетворявших заданному пороговому значению свободной энергии связывания, проводили уточняющий докинг со структурой M^{Pro} SARS-CoV-2, в результате которого были отобраны 347 732 соединения со свободной энергией связывания < -6 ккал/моль. Более высокий энергетический порог для уточняющего докинга по сравнению с предварительным докингом был выбран с учетом двух основных факторов. Во-первых, программа AutoDock Vina, использованная для уточняющего докинга, обладает более высокой точностью предсказания аффинности связывания лигандов с целевым белком, чем QuickVina 2 [31], которая была применена в случае предварительного докинга. Во-вторых, для уточняющего докинга рассматривали гораздо большее число ориентаций лиганда в активном центре белка (см. ниже подраздел «Вычислительный протокол молекулярного докинга»). Достигнутое с помощью двухэтапного докинга существенное сокращение размера обучающего набора данных позволило значительно уменьшить количество параметров разработанной нейронной сети и тем самым ускорить процесс ее обучения.

в) Пересмотр пространства SMILES и векторизации

Сформированную библиотеку соединений в формате линейного описания SMILES очищали от молекул, имевших молекулярную массу больше 1000 Да, и атомов, редко встречающихся в структурах лекарственных веществ химических элементов и изотопов, а также от некорректных записей SMILES. Структурные представления соединений, подготовленных в формате SMILES, получали с помощью Python 3 с привлечением модуля RDKit (<http://www.rdkit.org/>), который использовали ранее для очистки набора обучающих данных. На основе распределения частот символов SMILES из подготовленного набора данных удаляли соединения, имевшие хотя бы один элемент SMILES с частотой менее 0.001, а затем анализировали распределение длин записей SMILES. Соединения с представлением SMILES длиннее 120 символов исключали из дальнейшего рассмотрения (рис. 1,А). После применения всех фильтров набор данных составил 342 102 молекулы и соответствующих им представлений SMILES, каждое из которых было векторизовано в матрицу в соответствии с максимальной длиной и размером словаря SMILES-символов с добавленными символами начала и конца SMILES-строки, представленными "!" и "E", соответственно (рис. 1,Б).

Подготовленная на предыдущем этапе обучающая молекулярная библиотека объемом в 342 102 соединения и соответствующие им значения свободной энергии связывания с M^{Pro} SARS-CoV-2 сформировали набор данных для обучения и тестирования нейронной сети. Молекулярная библиотека была разделена на тренировочный, валидационный и тестовый наборы, включавшие соответственно 70 %, 15 % и 15 % от общего числа содержащихся в ней соединений. При формировании подмножеств использовали стратифицированное разбиение для сохранения равного распределения энергетической характеристики во всех трех подмножествах. Валидационный набор использовали для оценки способности модели восстанавливать входные SMILES-представления во время обучения, в то время как тестовые SMILES-описания применяли для генерации новых соединений путем добавления шума к их представлению на латентном слое автоэнкодера. Полученные таким образом наборы

данных использовали для обучения моделей, их валидации и тестирования новых, сгенерированных нейронной сетью молекул.

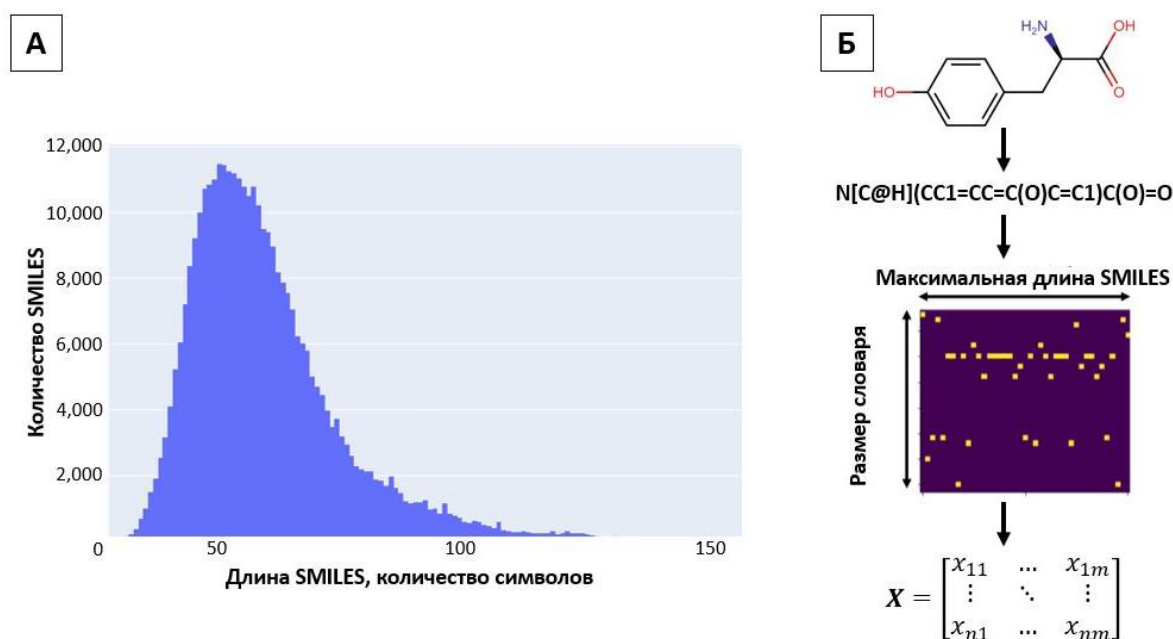


Рис. 1. Пересмотр пространства SMILES и векторизация: (А) Распределение длины SMILES для тренировочного набора; (Б) Процедура векторизации SMILES представлений.

Получение трехмерных структур сгенерированных соединений

Чтобы оценить способность модели глубокого обучения генерировать новые соединения, активные против целевого белка, необходимо было провести молекулярный докинг этих молекул с белком-мишенью. Очевидно, что для этого было необходимо наличие трехмерных (3D) структур сгенерированных молекул. Для получения 3D структур из линейного описания SMILES был разработан скрипт на языке Python 3 с использованием модуля RDKit (<http://www.rdkit.org/>). Процедура генерации включала следующие этапы: ввод SMILES-описаний соединений, проверку их достоверности, генерацию двумерных (2D) проекций координат атомов молекул, генерацию 3D координат их атомов, оптимизацию структур в силовом поле MMFF94, добавление атомов водорода и повторную оптимизацию в силовом поле MMFF94. Генерацию 3D координат атомов молекул осуществляли с помощью алгоритма ETKDGV3 [32].

а) Подготовка структуры белка

Структуру свободной формы M^{Pro} SARS-CoV-2 в кристалле заимствовали из Банка данных белков (PDB ID: 6Y84; <https://www.rcsb.org/pdb/>). Перед проведением докинга к структуре белка добавляли атомы водорода и задавали частичные атомные заряды с использованием модели Гастайгера [33]. Затем проводили оптимизацию структуры белка в силовом поле UFF [34] с помощью программного обеспечения OpenBabel [35]. Подготовленную таким образом структуру M^{Pro} SARS-CoV-2 использовали для предварительного и уточняющего молекулярного докинга как при подготовке обучающего набора данных, так и для молекулярного докинга с соединениями, сгенерированными нейронной сетью.

б) Подготовка структур лигандов

Подготовка структур лигандов для проведения предварительного докинга была такой же, как и для структуры белка. Эта процедура была реализована с использованием программы OpenBabel [35], однако включала дополнительный этап идентификации тепловых вращательных связей, который автоматически выполняется этим программным обеспечением. Подготовка структур лигандов перед уточняющим молекулярным докингом состояла из следующих этапов: 1) оптимизация в силовом поле MMFF94 [36, 37] для удаления стерически напряженных контактов и добавление атомов водорода, отсутствующих в исходной структуре; оба действия проводили с использованием модуля RDKit (<http://www.rdkit.org/>) в Python3, 2) добавление частичных атомных зарядов по схеме Гастайгера [33] и идентификация тепловых вращательных связей с помощью программного средства MGLTools (<http://mgltools.scripps.edu/>). Необходимо отметить, что перед проведением молекулярного докинга 3D структуры новых молекул получали из сгенерированных линейных представлений SMILES согласно процедуре, рассмотренной выше. Дальнейшие этапы подготовки сгенерированных нейронной сетью соединений к молекулярному докингу включали добавление частичных атомных зарядов и идентификацию тепловых вращательных связей, выполненную с помощью процедуры MGLTools.

Вычислительный протокол молекулярного докинга

Предварительный молекулярный докинг выполняли с помощью программы QuickVina 2 [31], а уточняющий – в программе AutoDock Vina [38] в приближении «жесткого» рецептора и «гибких» лигандов. В обоих случаях ячейка для докинга включала каталитический сайт M^{Pro} SARS-CoV-2 со следующими параметрами: $\Delta X = 19 \text{ \AA}$, $\Delta Y = 21 \text{ \AA}$, $\Delta Z = 23 \text{ \AA}$ с центром в точках $X = -20 \text{ \AA}$, $Y = 19 \text{ \AA}$, $Z = -26 \text{ \AA}$. Таким образом, объем ячейки для докинга составлял $19 \times 21 \times 23 = 9177 \text{ \AA}^3$. Значение параметра ширины охвата конформационного пространства лигандов, определяющего количество независимых запусков процедуры конформационного поиска, задавали равным 10 (предварительный докинг) и 50 (уточняющий докинг).

Глубокое обучение

а) Архитектуры моделей

В результате проведенных исследований были разработаны две генеративные модели глубокого обучения: 1) автоэнкодер, основанный на SMILES-представлениях молекул, который обучали без учителя [39] с использованием ячеек долгой краткосрочной памяти (LSTM; Long Shot-Term Memory) [40]; 2) автоэнкодер, основанный на описаниях SMILES, который обучали с частичным привлечением учителя [39] с использованием технологии LSTM. Во второй модели значение свободной энергии связывания химических соединений с молекулярной мишенью использовали как дополнительный параметр в латентном слое для обучения на основе результатов докинга соединений из тренировочного набора и как желаемое значение аффинности их связывания с M^{Pro} в режиме генерации новых соединений.

Высокоуровневая архитектура обеих моделей представлена на рисунке 2.

Первая (эмбединговая) модель (эмбедингом называют вектор чисел применительно к данным, которые он описывает) состоит из энкодера, латентного слоя, также являющегося входным слоем для гауссовского шума, и декодера (рис. 2). Эта модель получает на входной слой векторизованную матрицу описаний SMILES, которая проходит через слой LSTM, имеющий выходную размерность 64 нейрона. Особенность этой модели определяется тем, что сами выходные данные из ячеек LSTM

энкодера не используются, а вместо этого берутся скрытые векторы и векторы состояний ячеек. Эти векторы конкатенируются вместе, образуя 128-ми элементные векторы, которые пропускаются через полносвязный слой с выходной размерностью 32 нейрона. Данные на выходе из полносвязного слоя попадают на латентный слой, состоящий из 32 нейронов, и являются эмбедингами SMILES в контексте автоэнкодера. Далее эмбединги подаются параллельно на два полносвязных слоя с выходной размерностью 64 нейрона каждый, создавая начальные скрытые векторы и векторы состояния ячеек для слоя LSTM в декодере, имеющего такую же размерность, как и LSTM слой в энкодере. Существует также входной слой для декодера, используемый в качестве входных данных для LSTM, который в режиме обучения получает ту же векторизованную матрицу SMILES, что и вход энкодера, и, как обычная генеративная модель LSTM, предсказывает следующий символ. В режиме генерации на вход декодера подается символ начала строки “!”, запуская процесс посимвольной генерации новой строки SMILES. В этом режиме эмбединги используются для прогнозирования начальных состояний декодера LSTM, и они определяют, какая строка SMILES будет сгенерирована. Во всех полносвязных слоях (кроме последнего слоя) используется функция активации ReLu [41], а для последнего слоя – функция активации softmax [42].

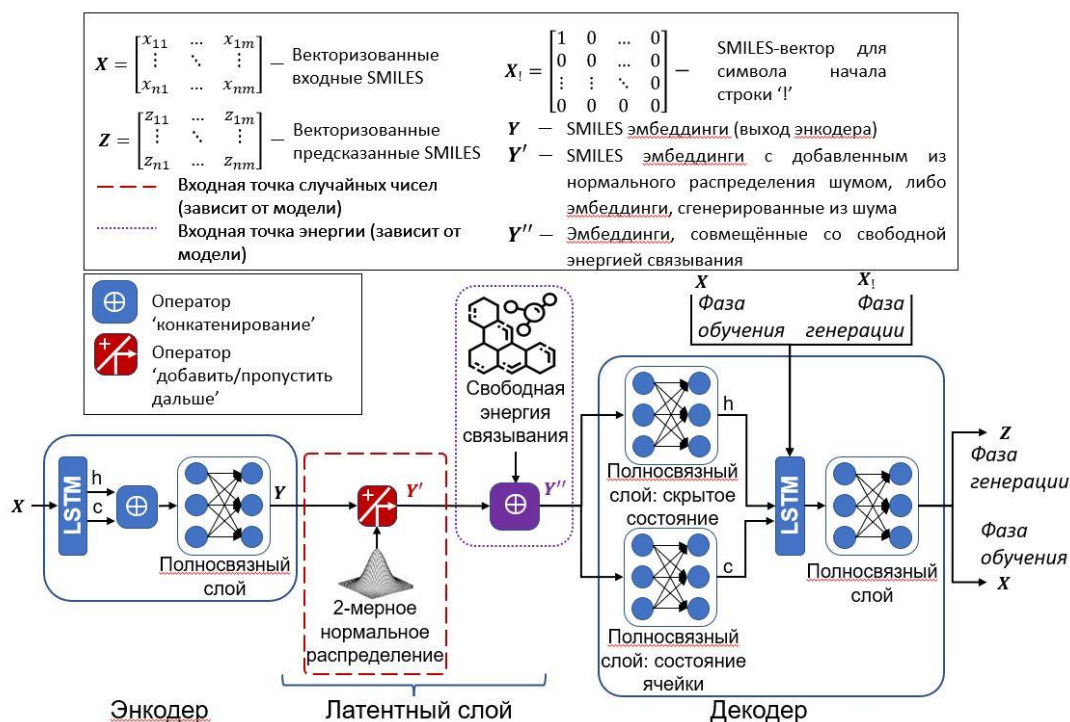


Рис. 2. Высокоуровневая архитектура разработанных моделей автоэнкодера.

[Открыть в полном размере](#)

Вторая (энергетическая) модель (рис. 2) отличается от эмбединговой модели дополнительным нейроном, расположенным на ее латентном слое и отвечающим за значение свободной энергии связывания молекулы с целевым белком. В отличие от эмбединговой модели, которая позволяет генерировать молекулы из случайных описаний SMILES, энергетическая модель, в дополнение к возможности вносить шум в эмбединги SMILES с целью улучшения их структур после декодирования, позволяет генерировать новые лиганды с заданной величиной энергии, увеличивая таким образом аффинность связывания.

б) Обучение моделей

Обе разработанные модели были составлены слой за слоем с использованием высокоуровневого интерфейса пакета TensorFlow 2.1 (<https://www.tensorflow.org/>). Модели прошли 150 эпох обучения с дополнительным использованием функций обратного вызова "Уменьшить скорость обучения на плато" и "Ранняя остановка", чтобы помочь им сойтись к лучшему локальному минимуму, а также избежать переобучения. В качестве оптимизатора использовали метод стохастической оптимизации градиентного спуска – метод Адама [43] – с начальным значением скорости обучения 0.005; при этом была выбрана категориальная функция потерь кросс-энтропии [44]. Прогресс оценки потерь для обеих моделей показан на рисунке 3.

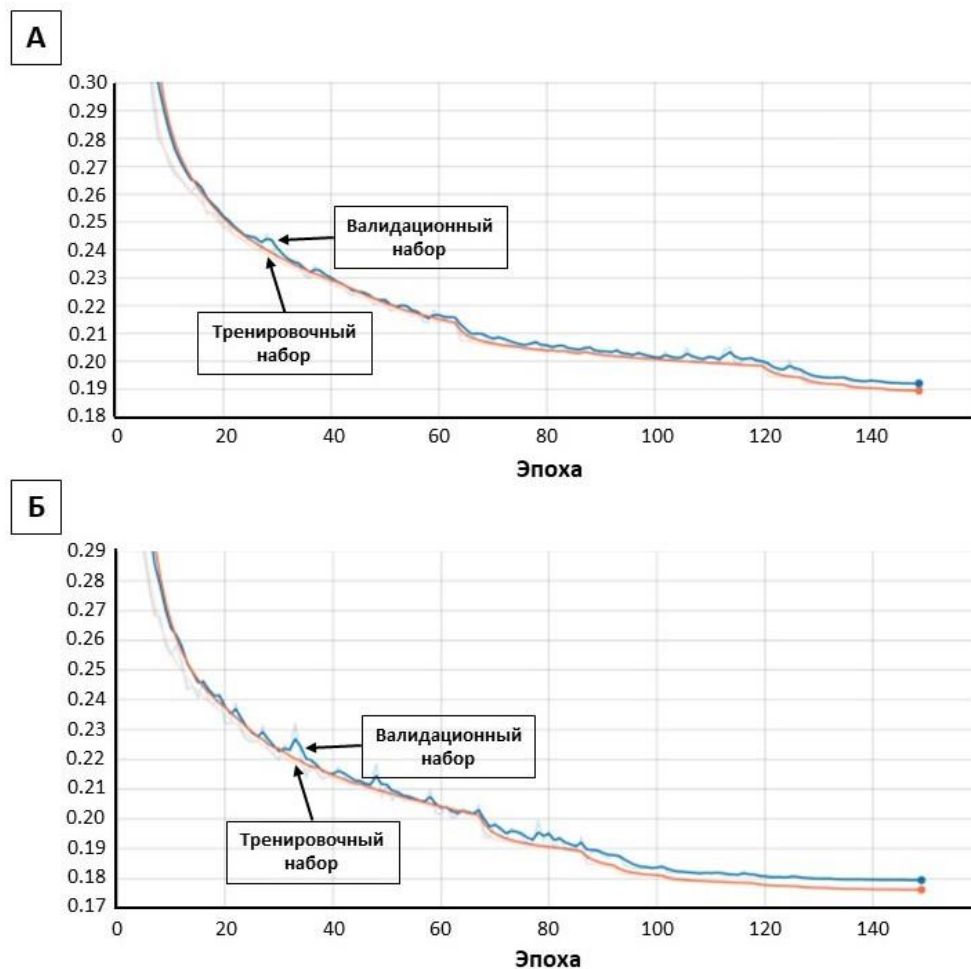


Рис. 3. Потери при обучении и валидации для моделей автоэнкодера: (А) Модель без учителя (эмбединги); (Б) Модель с частичным привлечением учителя (эмбединги и пороговое значение энергии связывания).

Генерация соединений и их анализ

В проведенном исследовании были рассмотрены два режима генерации соединений. В первом режиме генерацию выполняли из случайных чисел, взятых из нормального распределения с параметрами, полученными с помощью распределения тестовых данных на латентном слое для каждой компоненты вектора (режим «пропустить» в операторе «добавить/пропустить дальше», рис. 2). Для этого режима процесс генерации в случае энергетической модели предполагал задание априорного значения свободной энергии связывания для аппроксимации генерируемых молекул. При этом были

проведены вычислительные эксперименты для различных пороговых значений энергии. Основная идея второго режима генерации заключалась в использовании выборки лучших лигандов из тестового набора с целью добавления шума к их эмбединговым представлениям (режим «добавить» в операторе «добавить/пропустить дальше», рис. 2). Этот подход должен был изменить модифицированный лиганд и, в случае энергетической модели, также увеличить аффинность связывания молекулы к целевому белку, вынуждая автоэнкодер генерировать более перспективные соединения. Исследованные комбинации двух моделей автоэнкодеров и двух режимов генерации обобщены в таблице 1.

Таблица 1. Описание исследованных комбинаций генеративных LSTM-моделей и режимов генерации

Модель	Описание входных данных для генерации	Описание процесса генерации
Эмбединговая модель	Случайные векторы, полученные из многомерного нормального распределения, аппроксимирующего тестовый набор данных	Случайные числа используются в качестве эмбедингов и подаются на декодер
Эмбединговая модель	Эмбединги соединений со значением свободной энергии связывания ≤ -9 ккал/моль, выбранные из тестового набора данных	К эмбедингам добавляется искажение в виде гауссовского шума, и полученные представления подаются на декодер
Энергетическая модель	Случайные векторы, полученные из многомерного нормального распределения, аппроксимирующего тестовый набор данных, а также исходно задаваемые значения свободной энергии связывания	Случайные векторы используются как эмбединги и подаются в качестве латентного слоя на декодер с исходно заданными значениями свободной энергии связывания
Энергетическая модель	Соединения со значениями свободной энергии связывания ≤ -8 ккал/моль, выбранные из тестового набора данных, а также улучшенные значения свободной энергии связывания	Проводится расчет эмбедингов эталонных соединений, затем добавляется искажение в виде гауссовского шума, и полученные представления с заданными пороговыми значениями свободной энергии связывания подаются на декодер

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Как отмечалось выше, обе модели были протестированы с использованием каждого из двух режимов генерации. Полученные результаты оценивали на основе значений свободной энергии связывания, предсказанных методом молекулярного докинга, а также путем их сравнения с величинами, рассчитанными для 34 152 эталонных соединений, взятых из тестового набора данных. Эти соединения использовали в вычислительных экспериментах в качестве положительного контроля при оценке способности нейронной сети генерировать из этих молекул более перспективные соединения. Эталонные соединения выбирали случайным образом из тестового набора данных в различных режимах генерации. Результаты генерации соединений для всех проведенных вычислительных экспериментов обобщены в таблице 2.

Таблица 2. Количество сгенерированных нейронной сетью соединений и лучшие значения свободной энергии связывания, полученные с использованием двух генеративных моделей автоэнкодера и двух режимов генерации

Название модели; режим генерации	Количество соединений, прошедших процедуру молекулярного докинга	Наименьшее значение свободной энергии связывания, ккал/моль
Эмбединговая; случайные векторы	986	–10.6
Эмбединговая; эталонные соединения	2518	–10.3
Энергетическая; случайные векторы и значения энергии	594	–9.3
Энергетическая; эталонные соединения и улучшенные значения энергии	658	–10.2

Полученные результаты оценки качества работы разработанных моделей в двух режимах генерации приведены в таблице 3.

Таблица 3. Результаты генерации соединений с использованием двух генеративных моделей нейронной сети и двух режимов генерации

Название модели; метод генерации	Процент сгенерированных соединений с предсказанной величиной свободной энергией связывания			
	< –8 ккал/моль	< –9 ккал/моль	ниже по сравнению со значениями свободной энергией связывания эталонных соединений	ниже по сравнению с заданным пороговым значением свободной энергии связывания
Эмбединговая; случайные векторы	28.1 %	6.0 %	–	–
Эмбединговая; эталонные соединения	38.4 %	8.6 %	7.1 %	–
Энергетическая; случайные векторы и значения энергии	25.4 %	1.4 %	–	8.3 %
Энергетическая; эталонные соединения и улучшенные значения энергии	40.4 %	16.1 %	10.0 %	2.0 %

а) Эмбединговая модель; для генерации использованы случайные векторы

Несмотря на то, что эта модель не использует эталонные соединения и значения энергии связывания, она способна генерировать новые потенциальные ингибиторы целевого белка только путем генерации соединений на основе распределений эмбедингов, полученных из обучающего набора данных. Доля таких

модифицированных соединений среди молекул с низкой предсказанной энергией значительно выше: для 38.4 % лигандов значения свободной энергии связывания < -8 ккал/моль, а для 8.6 % – < -9 ккал/моль.

б) Эмбединговая модель; для генерации использованы эталонные соединения

Этот метод генерации использует известные лиганды целевого белка и пытается создать их модификации с более высоким химическим сродством к данной молекулярной мишени. Доля таких модифицированных соединений значительно выше: для 38.4 % лигандов значения свободной энергии связывания меньше -8 ккал/моль, а для 8.6 % – меньше -9 ккал/моль.

в) Энергетическая модель; для генерации использованы случайные векторы и пороговые значения энергии связывания.

Модель, обученная с частичным привлечением учителя, использует для входных молекулярных данных механизм “распутывания стиля и содержания” [45]. Согласно полученным результатам, для 31 % сгенерированных нейронной сетью соединений значения свободной энергии связывания находились в интервале 1 ккал/моль от заданной пороговой величины, а для 64 % – в интервале 2 ккал/моль.

г) Энергетическая модель; для генерации использованы эталонные соединения и пороговые значения энергии.

В данном вычислительном эксперименте пороговые значения энергии связывания для эталонных соединений смещены на 0.5 ккал/моль и 1.0 ккал/моль к более низким величинам. Эта комбинация модели и режима генерации показала лучшие результаты среди всех четырех комбинаций: 52 % соединений были распределены в диапазоне 1 ккал/моль относительно заданных пороговых значений; при этом 16.1 % соединений имели энергию связывания меньше -9 ккал/моль.

Вычислительный эксперимент с заданными пороговыми значениями свободной энергии связывания

Энергетическая модель в комбинации с гауссовским шумом была протестирована на способность к генерации соединений, демонстрирующих более низкие значения свободной энергии связывания при более низких величинах энергии, передаваемых «энергетическому» нейрону. При этом было проведено несколько вычислительных экспериментов по получению соединений с заданной энергией от -7 ккал/моль до -11 ккал/моль с шагом 1 ккал/моль. Результаты этих вычислительных экспериментов представлены на рисунке 4.

Для проверки, статистической значимости наблюдаемых различий между средними значениями энергии связывания для распределений, полученных в проведенных вычислительных экспериментах, был использован двусторонний t -критерий Стьюдента с нулевой гипотезой H_0 в предположении, что средние значения распределений энергии одинаковы. Критерий t применяли для всех пар распределений при различных пороговых значениях энергии (табл. 4). При этом ни одно из рассчитанных p -значений не было ниже уровня значимости 0.05, выбранного для отклонения нулевой гипотезы H_0 . В частности, самое низкое значение p , равное 0.22 (> 0.05), наблюдали между распределениями пороговых величин энергии -11 ккал/моль и -9 ккал/моль (табл. 4). Эти результаты позволяют предполагать, что распределения энергий в выборке действительно имеют одинаковые средние значения.

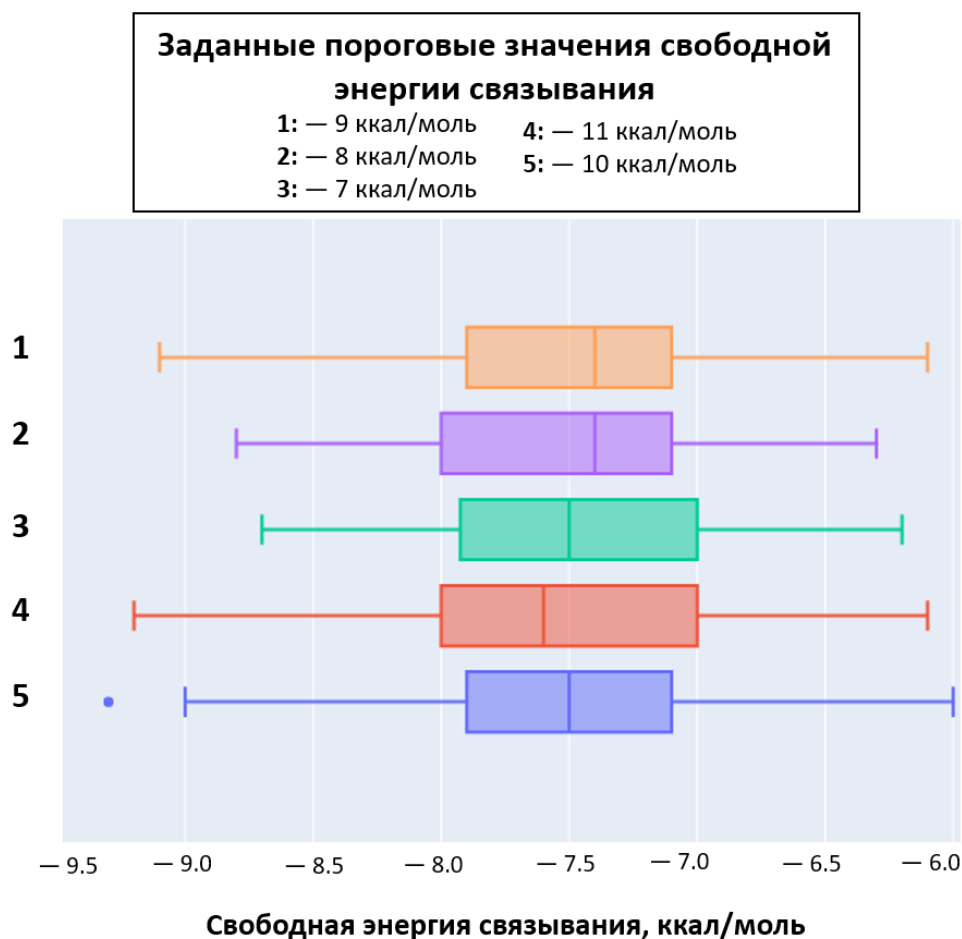


Рис. 4. График «диаграмма размаха», описывающий распределение результатов экспериментов с заданием различных пороговых значений свободной энергии связывания от -7 ккал/моль до -11 ккал/моль с шагом 1 ккал/моль.

Таблица 4. Результаты двустороннего t-теста для проверки нулевой гипотезы H_0

Пороговые значения энергии связывания и их парные комбинации, использованные для расчета p -значений на основе соответствующих нормальных распределений		p -значение
-11.0	-10.0	0.76
-11.0	-9.0	0.22
-11.0	-8.0	0.50
-11.0	-7.0	0.34
-10.0	-9.0	0.32
-10.0	-8.0	0.62
-10.0	-7.0	0.43
-9.0	-8.0	0.81
-8.0	-7.0	0.99
-8.0	-7.0	0.83

Эксперимент с использованием гауссовского шума

Этот эксперимент проводили для проверки целесообразности добавления гауссовского шума при генерации молекул из эталонных соединений с использованием «энергетического» нейрона. С этой целью были сгенерированы новые соединения, эмбединги которых изменяли добавлением гауссовского шума, а значения свободной энергии связывания, передаваемые «энергетическому» нейрону, устанавливали на более низком уровне по сравнению с эталонными соединениями. Затем эталонные соединения использовали повторно для генерации новых молекул, однако в этом случае задавали только более низкие значения энергии, чем у исходных лигандов. Данный эксперимент проводили для того, чтобы ответить на вопрос, способна ли модель модифицировать и улучшить свойства соединения при учете только «обновленного» значения свободной энергии связывания или же для этого требуются дополнительные пертурбации эмбедингов. Результаты этих экспериментов показаны на рисунке 5.

Как и в случае, рассмотренном выше, для сравнения средних значений двух распределений (рис. 5) использовали двусторонний t-критерий с нулевой гипотезой H_0 , предполагающей идентичность средних значений распределения энергии. Полученное значение p , равное 0.00068, оказалось ниже выбранного уровня значимости 0.05, что дает достаточно оснований для того, чтобы отвергнуть нулевую гипотезу H_0 и фактически предположить, что наблюдаемые различия между средними распределениями являются статистически значимыми. Следовательно, распределение энергий для соединений, полученных исключительно на основе улучшения аффинности связывания с целевым белком, имеет лучшее среднее значение. Этот результат предполагает, что добавление дополнительного шума только «запутывает» энергетическую модель при попытке извлечь выгоду из эмбедингов эталонных соединений.

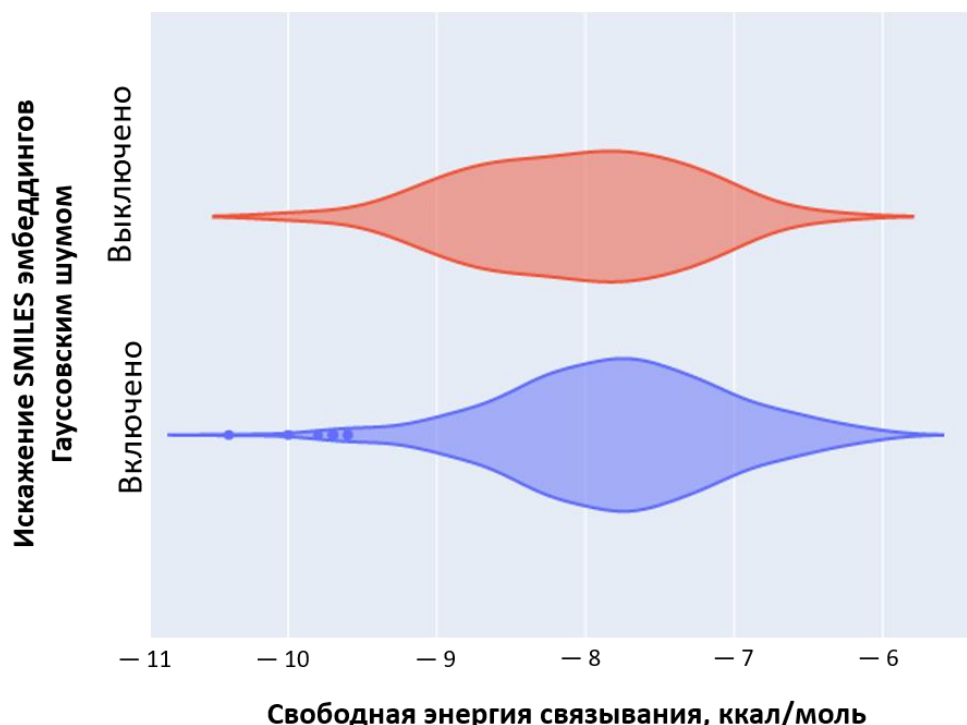


Рис. 5. Скрипичные графики, описывающие распределения энергии в экспериментах с добавлением гауссовского шума и без него для эмбедингов SMILES при генерации из эталонных соединений с использованием энергетической модели.

Сравнение моделей и режимов генерации

Для визуализации результатов эффективности работы двух разработанных моделей в двух режимах генерации, для каждой их комбинации было построено кумулятивное распределение свободной энергии связывания (рис. 6). Этот метод сравнения позволяет наблюдать общее качество сгенерированных экспериментальных выборок и проводить анализ распределения в целом без учета их количества. Из-за небольшого числа генерируемых соединений, демонстрирующих значения свободной энергии связывания меньше -6 ккал/моль (менее 2.8 % от всех генерируемых соединений), результаты генерации были отфильтрованы согласно этой пороговой величине, использованной ранее для получения обучающего набора данных. Для наглядности представления результатов сравнительного анализа для каждого эксперимента выбирали определенные подмножества. В случае эмбединговой модели были отобраны все соединения, сгенерированные из гауссовского шума (эксперимент I), и все соединения, сгенерированные с использованием эталонных соединений, которые были изменены с его помощью (эксперимент II). Последний режим генерации включал лишь эталонные соединения со значениями свободной энергии связывания меньше -9 ккал/моль. Для энергетической модели были выбраны три вычислительных эксперимента. Первый эксперимент (эксперимент III) генерировал соединения из гауссовского шума с использованием передаваемых в «энергетический» нейрон пороговых значений свободной энергии связывания, заданных в интервале от -9 ккал/моль до -11 ккал/моль с шагом 1 ккал/моль. Во втором эксперименте (эксперимент IV) использовали эталонные соединения, у которых соответствующие эмбединги изменяли путем добавления гауссовского шума, а пороговые значения энергии связывания понижали с шагом 0.5 ккал/моль и 1.0 ккал/моль относительно величин, полученных в результате докинга соответствующих эталонных соединений. Наконец, последний эксперимент (эксперимент V) также включал эталонные соединения, однако в этом случае изменение эмбедингов путем добавления шума не применяли, при этом в «энергетический» нейрон передавали улучшенное значение энергии.

Как видно из графика кумулятивного распределения свободной энергии связывания (рис. 6), модели из экспериментов II, IV и V генерировали больше соединений с более низкими значениями энергии, поскольку соответствующие им линии на графике расположены ниже, чем в других экспериментах. Сдвиг кумулятивной кривой вправо относительно других линий для каждой доли фиксированного набора данных f означает, что оставшая часть набора данных $1 - f$ для кривой, сдвинутой вправо, содержит больше соединений с более низкими значениями энергии. Например, при $f = 0.8$, можно наблюдать, что для экспериментов I и III лучшие 20 % сгенерированных соединений демонстрируют значения свободной энергии связывания выше -8.0 ккал/моль, а для экспериментов II и IV соответствующие значения немного ниже -8.0 ккал/моль. В то же время результаты эксперимента V значительно ближе к величине -8.5 ккал/моль.

Таким образом, сравнение моделей показало, что попытка генерировать новые соединения на основе уже существующих лигандов дает лучшие результаты по сравнению с их генерацией из гауссовского шума (табл. 3 и рис. 6). Кроме того, энергетической модели автоэнкодера удалось использовать введенный в латентный слой «энергетический» нейрон и получить больше соединений с более низкими значениями свободной энергии связывания с целевым белком, чем эмбединговой модели (рис. 6). При этом добавление гауссовского шума уменьшает преимущества использования «энергетических» нейронов (рис. 5). К сожалению, установка более низкого порогового значения свободной энергии связывания, передаваемого «энергетическому» нейрону, не имеет статистической значимости и не приводит к генерации более перспективных соединений (рис. 4). Обобщая все полученные данные,

можно сделать вывод, что архитектура модели автоэнкодера, использованная в эксперименте V, является наиболее перспективной из всех рассмотренных вариантов.

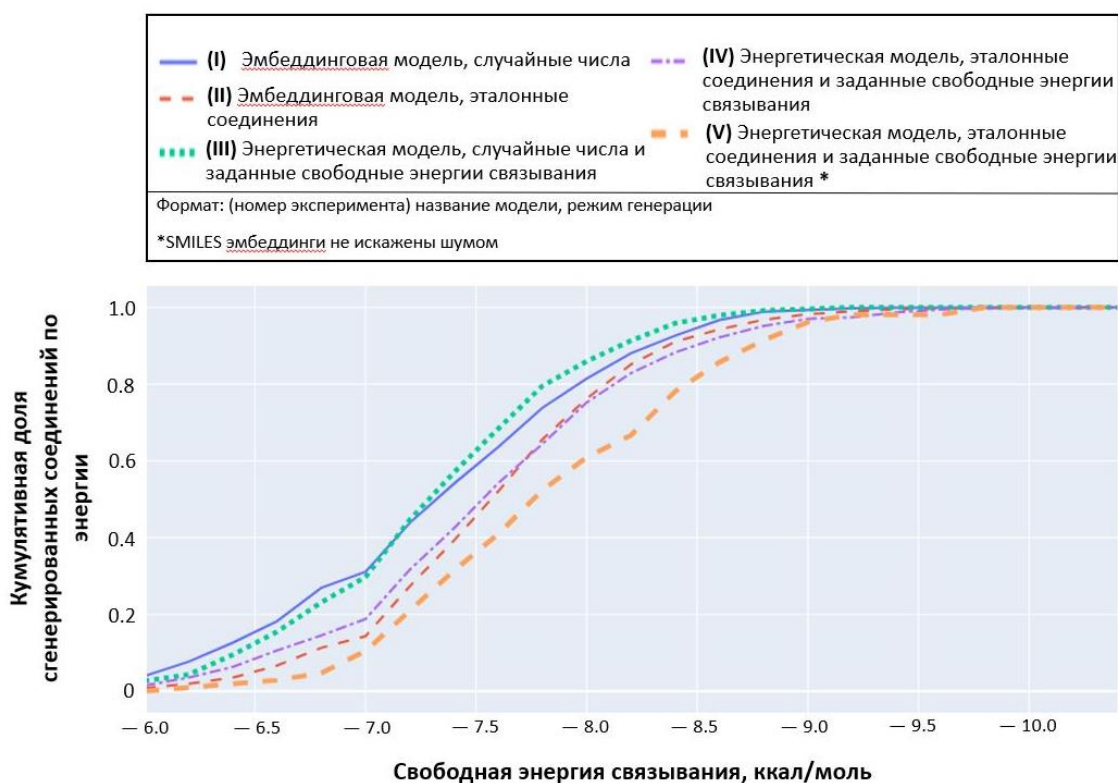


Рис. 6. Кумулятивные распределения свободной энергии связывания.

[Открыть в полном размере](#)

ЗАКЛЮЧЕНИЕ

С помощью технологий глубокого обучения разработана генеративная нейронная сеть для компьютерного дизайна новых потенциальных ингибиторов, способных блокировать каталитический сайт основной протеазы коронавируса SARS-CoV-2. Использование разработанной нейронной сети совместно с методом молекулярного докинга продемонстрировало ее значительный потенциал для заполнения неисследованных областей химического пространства новыми молекулами с заданными свойствами, что подтверждается полученными результатами, согласно которым из 4805 сгенерированных соединений только одно присутствовало в исходном наборе данных. Разработанные модели автоэнкодера в сочетании с традиционными методами молекулярного моделирования могут стать продуктивной основой для идентификации новых перспективных соединений против COVID-19, терапевтическое действие которых основано на блокаде основной протеазы SARS-CoV-2, критически важной для репликации и транскрипции вируса [24]. В связи с этим дальнейшее развитие данного исследования предполагает создание с помощью разработанной генеративной нейронной сети широкого набора потенциальных лигандов M^{Pro} SARS-CoV-2 с последующей компьютерной оценкой их противовирусной активности, идентификацией соединений-лидеров и биомедицинскими испытаниями.

Работа поддержана Белорусским республиканским фондом фундаментальных исследований (проекты Ф21КОВИД-002, X21COVID-003, Ф21АРМГ-001) и Союзом международных научных организаций ANSO (ANSO-CR-PP-2021-04).

СПИСОК ЛИТЕРАТУРЫ

1. Katsila T., Spyroulias G.A., Patrinos G.P., Matsoukas M.-T. Computational approaches in target identification and drug discovery. *Comput. Struct. Biotechnol. J.* 2016. V. 14. P. 177–184. doi: [10.1016/j.csbj.2016.04.004](https://doi.org/10.1016/j.csbj.2016.04.004)
2. Knight-Schrijver V.R., Chelliah V., Cucurull-Sanchez L., Le Novère N. The promises of quantitative systems pharmacology modelling for drug development. *Comput. Struct. Biotechnol. J.* 2016. V. 4. P. 363–370. doi: [10.1016/j.csbj.2016.09.002](https://doi.org/10.1016/j.csbj.2016.09.002)
3. Vamathevan J., Clark D., Czodrowski P., Dunham I., Ferran E., Lee G., Li B., Madabhushi A., Shah P., Spitzer M., Zhao S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 2019. V. 18. 6. P. 463–477. doi: [10.1038/s41573-019-0024-5](https://doi.org/10.1038/s41573-019-0024-5)
4. 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. *Front. Robotics and AI.* 2019. V. 6. P. 108. doi: [10.3389/frobt.2019.00108](https://doi.org/10.3389/frobt.2019.00108)
5. Dobchev D.A., Pillai G.G., Karelson M. In silico machine learning methods in drug development. *Curr. Top. Med. Chem.* 2014. V. 14. 16. P. 1913–1922. doi: [10.2174/1568026614666140929124203](https://doi.org/10.2174/1568026614666140929124203)
6. Cherkasov A., Muratov E.N., Fourches D., Varnek A., Baskin II, Cronin M, Dearden J., Gramatica P., Martin Y.C., Todeschini R., Consonni V., Kuz'min V.E., Cramer R., Benigni R., Yang C., Rathman J., Terfloth L., Gasteiger J., Richard A., Tropsha A. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* 2014. V. 57. No. 12. P. 4977–5010. doi: [10.1021/jm4004285](https://doi.org/10.1021/jm4004285)
7. Kinnings S.L., Liu N., Tonge P.J., Jackson R.M., Xie L., Bourne P.E. A Machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J. Chem. Inf. Model.* 2011. V. 51. No. 2. P. 408–419. doi: [10.1021/ci100369f](https://doi.org/10.1021/ci100369f)
8. Agastheeswaramoorthy K., Sevilimedu A. Drug REpurposing using AI/ML tools - for Rare Diseases (DREAM-RD): A case study with Fragile X Syndrome (FXS). 2020. *bioRxiv*. doi: [10.1101/2020.09.25.311142](https://doi.org/10.1101/2020.09.25.311142)
9. Senior A.W., Evans R., Jumper J., Kirkpatrick J., Sifre L., Green T., Qin C., Zidek A., Nelson A.W.R., Bridgland A., Penedones H., Petersen S., Simonyan K., Crossan S., Kohli P., Jones D.T., Silver D., Kavukcuoglu K., Hassabis D. Improved protein structure prediction using potentials from deep learning. *Nature.* 2020. V. 577. P. 706–710. doi: [10.1038/s41586-019-1923-7](https://doi.org/10.1038/s41586-019-1923-7)
10. Li H., Sze K.-H., Lu G., Ballester P.J. Machine-learning scoring functions for structure-based virtual screening. *WIREs Comput. Mol. Sci.* 2020. V. 11. Article No. e1478. doi: [10.1002/wcms.1478](https://doi.org/10.1002/wcms.1478)
11. Xiong G.-L., Ye W.-L., Shen C., Lu A.-P., Hou T.-J., Cao D.-S. Improving structure-based virtual screening performance via learning from scoring function components. *Brief. Bioinform.* 2020. V. 22. doi: [10.1093/bib/bbaa094](https://doi.org/10.1093/bib/bbaa094)
12. Stokes J.M., Yang K., Swanson K., Jin W., Cubillos-Ruiz A., Donghia N.M., MacNair C.R., French S., Carfrae L.A., Bloom-Ackermann Z., Tran V.M., Chiappino-Pepe A., Badran A.H., Andrews I.W., Chory E.J., Church J.M., Brown E.D., Jaakkola T.S., Barzilay R., Collins J.J. A deep learning approach to antibiotic discovery. *Cell.* 2020. V. 180. No. 4. P. 688–702. doi: [10.1016/j.cell.2020.01.021](https://doi.org/10.1016/j.cell.2020.01.021)
13. 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. *Brief. Bioinform.* 2021. V. 22. doi: [10.1093/bib/bbab258](https://doi.org/10.1093/bib/bbab258)
14. 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. *J. Biomol. Struct. Dyn.* 2021. V. 15. P. 1–19. doi: [10.1080/07391102.2021.1905559](https://doi.org/10.1080/07391102.2021.1905559)
15. Zhang Y., Ye T., Xi H., Juhas M., Li J. Deep learning driven drug discovery: Tackling Severe Acute Respiratory Syndrome Coronavirus 2. *Front. Microbiol.* 2021. doi: [10.3389/fmicb.2021.739684](https://doi.org/10.3389/fmicb.2021.739684)
 16. Mercado R., Rastemo T., Lindelöf E., Klambauer G., Engkvist O., Chen H., Bjerrum E.J. Practical notes on building molecular graph generative models. *ChemRxiv. Preprint.* 2020. doi: [10.26434/chemrxiv.12888383](https://doi.org/10.26434/chemrxiv.12888383)
 17. Arús-Pous J., Blaschke T., Ulander S., Reymond J.L., Chen H., Engkvist O. Exploring the GDB-13 chemical space using deep generative models. *J. Cheminform.* 2019. V. 11. Article No. 20. doi: [10.1186/s13321-019-0341-z](https://doi.org/10.1186/s13321-019-0341-z)
 18. Prykhodko O., Johansson S.V., Kotsias P.C., Arús-Pous J., Bjerrum E.J., Engkvist O., Chen H. A *de novo* molecular generation method using latent vector based generative adversarial network. *J. Cheminform.* 2019. V. 11. No. 1. Article No. 74. doi: [10.1186/s13321-019-0397-9](https://doi.org/10.1186/s13321-019-0397-9)
 19. Polykovskiy D., Zhebrak A., Vetrov D., Ivanenkov Y., Aladinskiy V., Mamoshina P., Bozdaganyan M., Aliper A., Zhavoronkov A., Kadurin A. Entangled conditional adversarial autoencoder for *de novo* drug discovery. *Mol. Pharmaceutics.* 2018. V. 15. No. 10. P. 4398–4405. doi: [10.1021/acs.molpharmaceut.8b00839](https://doi.org/10.1021/acs.molpharmaceut.8b00839)
 20. Zhang J., Mercado R., Engkvist O., Chen H. Comparative study of deep generative models on chemical space coverage. *ChemRxiv. Preprint.* 2020. doi: [10.26434/chemrxiv.13234289.v1](https://doi.org/10.26434/chemrxiv.13234289.v1)
 21. Zhavoronkov A., Ivanenkov Y.A., Aliper A., Veselov M.S., Aladinskiy V.A., Aladinskaya A.V., Terentiev V.A., Polykovskiy D.A., Kuznetsov M.D., Asadulaev A., Volkov Yu., Zholus A., Shayakhmetov R.R., Zhebrak A., Minaeva L.I., Zagribelnyy B.A., Lee L.H., Soll R., Madge D., Xing L., Guo T., Aspuru-Guzik A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 2019. V. 37. No. 9. P. 1038–1040. doi: [10.1038/s41587-019-0224-x](https://doi.org/10.1038/s41587-019-0224-x)
 22. Kim S., Chen J., Cheng T., Gindulyte A., He J., He S., Li Q., Shoemaker B.A., Thiessen P.A., Yu B., Zaslavsky L., Zhang J., Bolton E.E. PubChem in 2021: new data content and improved web interfaces. *Nucl. Acids Res.* 2021. V. 49. No. D1. P. D1388–D1395. doi: [10.1093/nar/gkaa971](https://doi.org/10.1093/nar/gkaa971)
 23. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 1998. V. 28. No. 1. P. 31–36. doi: [10.1021/ci00057a005](https://doi.org/10.1021/ci00057a005)
 24. Pillaiyar T., Manickam M., Namasivayam V., Hayashi Y., Jung S.H. An overview of Severe Acute Respiratory Syndrome-Coronavirus (SARS-CoV) 3CL protease inhibitors: peptidomimetics and small molecule chemotherapy. *J. Med.Chem.* 2016. V. 59. No. 14. P. 6595–6628. doi: [10.1021/acs.jmedchem.5b01461](https://doi.org/10.1021/acs.jmedchem.5b01461)
 25. Yan F., Gao F. An overview of potential inhibitors targeting non-structural proteins 3 (PL^{pro} and Mac1) and 5 (3CL^{pro}/M^{pro}) of SARS-CoV-2. *Comp. Struct. Biotechnol. J.* 2021. V. 19. P. 4868–4883. doi: [10.1016/j.csbj.2021.08.036](https://doi.org/10.1016/j.csbj.2021.08.036)
 26. Ullrich S., Nitsche C. The SARS-CoV-2 main protease as drug target. *Bioorganic Med. Chem. Lett.* 2020. V. 30. No. 17. doi: [10.1016/j.bmcl.2020.127377](https://doi.org/10.1016/j.bmcl.2020.127377)
 27. Forster P., Forster L., Renfrew C., Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Ssci. USA.* 2020. V. 117. No. 17. P. 9241–9243. doi: [10.1073/pnas.2004999117](https://doi.org/10.1073/pnas.2004999117)
 28. Schneidman-Duhovny D., Dror O., Inbar Y., Nussinov R., Wolfson H. J. Deterministic pharmacophore detection via multiple flexible alignment of drug-like molecules. *J. Comput. Biol.* 2008. V. 15. No. 7. P. 737–754. doi: [10.1089/cmb.2007.0130](https://doi.org/10.1089/cmb.2007.0130)
 29. Sunseri J., Koes D.R. Pharmit: interactive exploration of chemical space. *Nucl. Acids Res.* 2016. V. 44. P. W442–W448. doi: [10.1093/nar/gkw287](https://doi.org/10.1093/nar/gkw287)

30. Филимонов Д.А., Дружиловский Д.С., Лагунин А.А., Глориезова Т.А., Рудик А.В., Дмитриев А.В., Погодин П.В., Поройков В.В. Компьютерное прогнозирование спектров биологической активности химических соединений: возможности и ограничения. *Biomed. Chem.: Research and Methods*. 2018. V. 1. doi: [10.18097/BMCRM00004](https://doi.org/10.18097/BMCRM00004)
31. Alhossary A., Handoko S.D., Mu Y., Kwoh C.-K. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics*. 2015. V. 31. No. 13. P. 2214–2216. doi: [10.1093/bioinformatics/btv082](https://doi.org/10.1093/bioinformatics/btv082)
32. Wang S., Witek J., Landrum G.A., Riniker S. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *J. Chem. Inf. Model.* 2020. V. 60. P. 2044–2058. doi: [10.1021/acs.jcim.0c00025](https://doi.org/10.1021/acs.jcim.0c00025)
33. Gasteiger J., Marsili M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* 1978. V. 19. P. 3181–3184. doi: [10.1016/S0040-4039\(01\)94977-9](https://doi.org/10.1016/S0040-4039(01)94977-9)
34. Rappe A.K., Casewit C.J., Colwell K.S., Goddard W.A., Skiff W.M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* 1992. V. 114. P. 10024–10035. doi: [10.1021/ja00051a040](https://doi.org/10.1021/ja00051a040)
35. O'Boyle N.M., Banck M., James C.A., Morley C., Vandermeersch T., Hutchison G.R. Open Babel: An open chemical toolbox. *J. Cheminformatics*. 2011. V. 3. Article No. 33. doi: [10.1186/1758-2946-3-33](https://doi.org/10.1186/1758-2946-3-33)
36. Halgren T.A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* 1996. V. 17. P. 490–519. doi: [10.1002/\(SICI\)1096-987X\(199604\)17:5/6<490::AID-JCC1>3.0.CO;2-P](https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P)
37. Tosco P., Stiefl N., Landrum G. Bringing the MMFF force field to the RDKit: implementation and validation. *J. Cheminformatics*. 2014. V. 6. doi: [10.1186/s13321-014-0037-3](https://doi.org/10.1186/s13321-014-0037-3)
38. Trott O., Olson A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 2010. V. 31. P. 455–461. doi: [10.1002/jcc.21334](https://doi.org/10.1002/jcc.21334)
39. Epstein B., Meir R. Generalization bounds for unsupervised and semi-supervised learning with autoencoders. *arXiv: 1902.01449*. 2019. doi: [10.48550/arXiv.1902.01449](https://doi.org/10.48550/arXiv.1902.01449)
40. Hochreiter S., Schmidhuber J. Long Short-Term Memory. *Neural Comput.* 1997. V. 9. P. 1735–1780. doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)
41. Agarap A.F. Deep Learning using Rectified Linear Units (ReLU). *arXiv: 1803.08375v2*. 2019. doi: [10.48550/arXiv.1803.08375](https://doi.org/10.48550/arXiv.1803.08375)
42. Sharma Sid., Sharma Sim., Athaiya A. Activation functions in neural networks. *IJEAST*. 2020. V. 4. No. 12. P. 310–316. ISSN 2455-2143.
43. Kingma D.P., Ba J. Adam: A method for stochastic optimization. *arXiv: 1412.6980*. 2014. doi: [10.48550/arXiv.1412.6980](https://doi.org/10.48550/arXiv.1412.6980)
44. Janocha K., Czarnecki W.M. On loss functions for deep neural networks in classification. *Schedae Informaticae*. 2016. V. 25. P. 49–59. doi: [10.4467/20838476SI.16.004.6185](https://doi.org/10.4467/20838476SI.16.004.6185)
45. Shu R., Zhao S., Kochenderfer M.J. Rethinking style and content disentanglement in variational autoencoders. *6th International Conference on Learning Representations, {ICLR} 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. 2018. URL: <https://openreview.net/pdf?id=B1rQtWJDG> (дата обращения: 08.09.2022).

Рукопись поступила в редакцию 25.07.2022, переработанный вариант поступил 01.09.2022.
Дата опубликования 12.09.2022.

Development of a Deep Learning Generative Neural Network for Computer-Aided Design of Potential SARS-Cov-2 Inhibitors

Shuldau N.A.^{1,2}, Yushkevich A.M.¹, Furs K.V.¹, Tuzikov A.V.¹,
Andrianov A.M.³

¹United Institute of Informatics Problems, National Academy of Sciences of Belarus

²EPAM Systems, Minsk, Republic of Belarus

³Institute of Bioorganic Chemistry, National Academy of Sciences of Belarus

Abstract. Two generative deep learning models have been developed for the computer-aided design of potential inhibitors of the SARS-CoV-2 main protease (MPro), an enzyme critically important for the virus replication and transcription, and, therefore, presenting a promising target for the design of effective antiviral drugs. To solve this problem, we formed a training library of small molecules containing structural elements capable of providing specific and effective interactions of potential ligands with the SARS-CoV-2 M^{Pro} catalytic site. The architecture of generative models was developed and implemented to generate new high-affinity ligands of this functionally important SARS-CoV-2 protein. The neural network was trained and tested on the compounds from the training library, and the results of training and operation in two different generation modes were evaluated. The use of generative models in conjunction with the molecular docking demonstrated their great potential for filling the unexplored regions of the chemical space with novel molecules with pre-defined properties, which is confirmed by the obtained results according to which out of 4805 compounds generated by the neural network only one compound was present in the original data set.

Key words: machine learning methods, deep learning, generative neural networks, coronavirus SARS-CoV-2, main protease, antiviral drugs.