УДК 004.89:60 *DOI 10.55421/3034-4689_2026_29_4_118*

С. К. Долгих, Л. Ю. Кошкина

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ В БИОТЕХНОЛОГИИ: ДИЗАЙН БЕЛКОВ, ФЕРМЕНТОВ И ГЕНОМНОЕ РЕДАКТИРОВАНИЕ

Ключевые слова: искусственный интеллект, биотехнология, большие языковые модели, глубокое обучение, машинное обучение, аминокислотная последовательность, белковая инженерия, ферментная инженерия, геномное редактирование, target-off эффект, CRISPR Cas-9, белковая языковая модель ProGen, интерпретируемость моделей.

Представление биологических последовательностей – белков, ферментов и нуклеиновых кислот – в виде формализованных «языков» открыло новые возможности для их анализа, моделирования и генерации. На основе анализа публикаций российских и зарубежных авторов выявлены следующие перспективные направления использования больших языковых моделей в биотехнологии: дизайн белков, ферментная инженерия, геномное редактирование. Статья посвящена анализу современных подходов к использованию больших языковых моделей и глубокого обучения (Deep Learning, DL) в инженерии белков и ферментов, а также в задачах геномного редактирования, включая оценку непреднамеренных модификаций CRISPR Cas-9 систем. Особое внимание уделяется биотехнологическому потенциалу данных методов, их ограничениям, проблемам интерпретируемости и биобезопасности. На основе анализа актуальных исследований обсуждаются перспективы интеграции языковых моделей в молекулярные и промышленные биотехнологические процессы.

S. K. Dolgikh, L. Yu. Koshkina

LARGE LANGUAGE MODELS IN BIOTECHNOLOGY: PROTEIN DESIGN, ENZYMES AND GENOMIC EDITING

Keywords: artificial intelligence, biotechnology, large language models, deep learning, machine learning, amino acid sequence, protein engineering, enzyme engineering, genomic editing, target-off effect, CRISPR Cas-9, ProGen, interpretability of models

The description of biological sequences – proteins, enzymes and nucleic acids – in the form of formalized "languages" has opened up new possibilities for their analysis, modeling and generation. Based on the analysis of publications by Russian and foreign authors, the following promising areas of use of large language models in biotechnology were identified: protein design, enzyme engineering, genomic editing. The article is devoted to the analysis of modern approaches to the use of large language models and deep learning (Deep Learning, DL) in protein and enzyme engineering, as well as in genomic editing tasks, including the assessment of unintentional modifications Cas-9 CRISPR systems. Particular attention is paid to the biotechnological potential of these methods, their limitations, problems of interpretability and biosafety. Based on the analysis of current research, the prospects for integrating language models into molecular and industrial biotechnological processes are discussed.

Введение

Современная биотехнология характеризуется стремительным ростом объёмов данных, получаемых в результате экспериментов в геномике, протеомике и синтетической биологии [1]. Анализ и интерпретация таких данных требуют вычислительных подходов, способных выявлять нелинейные зависимости между структурой биологических последовательностей и их функциональными свойствами. В этой связи методы искусственного интеллекта (ИИ) становятся неотъемлемым элементом биотехнологических исследований и разработок, охватывая широкий спектр задач – от открытия лекарственных соединений до оптимизации биокаталитических процессов [2-3].

Адаптация больших языковых моделей (*large language models*, LLM) к биологическим данным основана на представлении аминокислотных и нуклеотидных последовательностей как символьных цепочек, обладающих собственной статистической структурой. Такой подход позволяет обучать модели на больших корпусах биологических последовательностей без явного задания биофизических правил, что существенно расширяет возможности вычислительного анализа и проектирования биомолекул [4].

Применение LLM в инженерии белков продемонстрировало возможность генерации новых аминокислотных последовательностей, обладающих заданными или ранее не наблюдавшимися свойствами. Ряд исследований показал, что модели, обученные на масштабных базах последовательностей, способны воспроизводить функционально значимые паттерны и создавать белки, сохраняющие биологическую активность в экспериментальных условиях [5, 6]. Результаты подтверждают гипотезу о том, что языковые модели способны захватывать скрытые закономерности, связывающие последовательность, структуру и функцию белков.

Ключевым аспектом белковой инженерии остаётся задача установления связи между первичной структурой белка и его трёхмерной конфигурацией, определяющей функциональные характеристики. Глубокие нейронные сети, включая архитектуры, используемые в языковых моделях, продемонстрировали высокую эффективность в задачах предсказания пространственной структуры белков, что стало важным этапом в развитии вычислительной биологии [7]. Интеграция структурного моделирования с генеративными

возможностями LLM открывает новые перспективы рационального дизайна белков.

Аналогичный прогресс наблюдается в инженерии ферментов, где ключевые цели – повышение каталитической эффективности, стабильности и субстратной селективности. Методы машинного обучения (*Machine Learning, ML*), опирающиеся на анализ последовательностей и экспериментальных данных, позволяют резко сократить объемы лабораторного скрининга и ускорить направленную эволюцию ферментов [8]. Специализированные модели глубокого обучения также эффективно справляются с предсказанием функций ферментов по их последовательностям, что является преимуществом в эпоху экспоненциального роста биологических баз данных [9].

Помимо белковой и ферментной инженерии, методы DL (*Deep Learning* – глубокое обучение) и языковые модели играют всё более значимую роль в области геномного редактирования. Технологии CRISPR (*Clustered Regularly Interspaced Short Palindromic Repeats*, т.е. кластерные регулярно расположенные короткие палиндромные повторы) обладают высоким потенциалом для биомедицины и синтетической биологии, однако их применение сопровождается рисками возникновения непреднамеренных модификаций (*off-target* эффектов). Разработка интерпретируемых моделей глубокого обучения для прогнозирования таких эффектов позволяет повысить безопасность и надёжность геномных вмешательств, что является критически важным для клинического и промышленного применения [10].

Несмотря на значительный прогресс, использование LLM в биотехнологии сопряжено с рядом ограничений, включая зависимость от качества обучающих данных, ограниченную интерпретируемость результатов и потенциальные риски двойного назначения. Вопросы ответственного применения ИИ, биобезопасности и контроля над генеративными моделями приобретают особую актуальность по мере их внедрения в биотехнологические процессы [4].

Целью настоящей статьи является анализ современных подходов к применению больших языковых и глубоких моделей в инженерии белков, ферментов и геномном редактировании, а также оценка их биотехнологического потенциала и существующих ограничений.

1. Большие языковые модели и дизайн белков

Применение LLM в белковой инженерии базируется на фундаментальной идее представления аминокислотных последовательностей в виде формализованного языка, где каждая аминокислота рассматривается как токен, а белковая последовательность – как предложение или текст. Такой подход позволяет использовать архитектуры, изначально разработанные для обработки естественного языка, для анализа и генерации биологических последовательностей.

Одним из подходов, продемонстрировавших эффективность языковых моделей в данной области, является модель ProGen, основанная на автрорегрессионном языковом моделировании белковых последовательностей [5]. ProGen представляет собой мощную нейронную сеть с 1,2 миллиардами параметров, обученную на обширном общедоступном датасете, включающем 280 миллионов белковых последовательностей [6]. Одной из ключевых особенностей ProGen является механизм условной генерации: процесс создания последовательностей управляется специальными тегами, которые передаются модели в качестве дополнительных входных данных. В задачах обработки естественного языка такими тегами служат, например, стиль текста, тематика, даты или другие семантические сущности. В случае белков роль управляющих тегов играют аннотации функциональных свойств — семейство белка, действующий биологический процесс или молекулярная функция.

В исследовании [6] показано сходство условных языковых моделей и ProGen (рис.1).

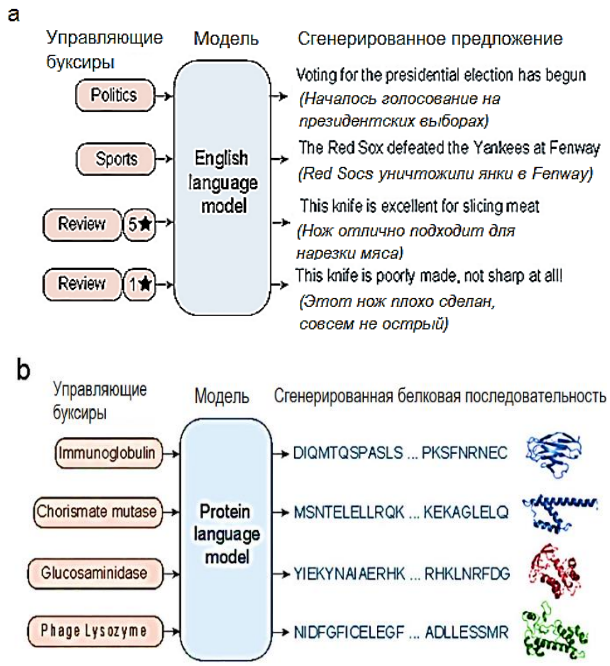


Рис. 1 – Условные языковые модели (a) и условная белковая языковая модель ProGen (b) [6]

Fig. 1 – Conditional language models (a) and the ProGen conditional protein language model (b) [6]

В рамках подхода автрорегрессионного языкового моделирования белковых последовательностей обучение модели осуществляется на больших корпусах белков, что позволяет ей выявлять статистические закономерности, связанные с эволюционными и функциональными ограничениями. Авторы подчёркивают, что методы отбора ProGen могут быть использованы для генерации набора образцов со статически высокой приспособленностью [5]. Результаты исследования показывают, что модель

умеет определять структурно и функционально значимые белки (рис.2).

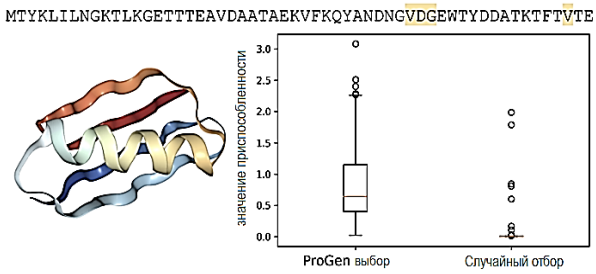


Рис. 2 – Кристаллическая структура белка GB1 (слева) и значения приспособленности образцов, выбранных с помощью ProGen, по сравнению со случайным выбором (справа)

Fig. 2 – Crystal structure of the GB1 protein (left) and fitness scores for samples selected using ProGen, compared to a random selection (right)

Дальнейшее развитие направления связано с переходом от статистического сходства к экспериментальному подтверждению функциональности сгенерированных белков. В работе [6] авторы демонстрируют, что большие языковые модели способны генерировать аминокислотные последовательности, которые не только структурно правдоподобны, но и сохраняют биологическую активность *in vitro* (в искусственных условиях). Экспериментальные результаты данной работы демонстрируют успешную экспрессию и функциональную активность сгенерированных белков в различных семействах, что делает эти данные особенно ценными для анализа. Искусственные белки остаются активными, даже если они отличаются (максимальная идентичность 40–50%, имеется в виду идентичность наиболее значимого соединения) от известных природных белков (здесь выбросы указывают на образцы с высокой активностью) (рис. 3).

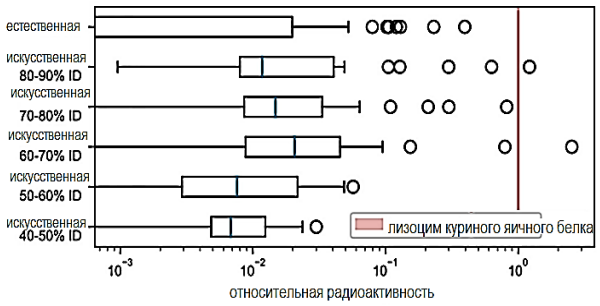


Рис. 3 – Активность искусственных белков [6

Fig. 3 – Activity of synthetic proteins [6]

Ключевым фактором, определяющим функциональность белка, является его трёхмерная структура. В обзоре [7] показано, что современные нейронные сети и методы Монте-Карло (в частности Rosetta Monte Carlo, RMC) способны эффективно моделировать взаимосвязь между первичной последовательностью и пространственной организацией белка. Метод RMC – стохастический алгоритм оптимизации, используемый в программном пакете Rosetta для предсказания пространственной структуры белков (фолдинга), дизайна новых белковых последовательностей и белок-белкового докинга.

В контексте белкового дизайна языковые модели могут рассматриваться как генераторы первичных последовательностей (метод RoseTTAFold), тогда как модели структурного предсказания (метод AlphaFold) выступают в роли фильтров и оценочных функций. Авторы обзорного исследования [7] указывают: «RoseTTAFold и AlphaFold обучены предсказывать структуру не на основе отдельных аминокислотных последовательностей, а на основе выравнивания множества гомологичных последовательностей, и они учатся извлекать богатую структурную информацию из этих эволюционных данных». Такой двухэтапный подход (RoseTTAFold + AlphaFold) позволяет существенно сократить пространство поиска и повысить вероятность получения функционально активных белков.

Таким образом, исследования [5–7] демонстрируют, что большие языковые модели способны захватывать сложные закономерности, лежащие в основе структуры и функции белков, и использовать их для генерации новых биомолекул.

2. Ферментная инженерия на основе DL

Ферменты являются ключевыми элементами большинства биотехнологических процессов, включая биокатализ, синтез лекарственных соединений и промышленное биопроизводство. Их каталитическая активность, стабильность и специфичность определяют эффективность и экономическую целесообразность технологических решений.

Традиционные подходы к инженерии ферментов включают классические методы белковой инженерии, направленные на модификацию структуры и свойств уже известных ферментов, а также использование ферментов как инструментов для генной инженерии (рестриктазы, лигазы), для разделения изомеров (ацилирование, гидролиз) и поиска новых ферментов в микроорганизмах с помощью зондов и скрининга, что позволяет улучшать их активность, стабильность и специфичность для биотехнологических нужд.

Приведем основные классические подходы.

- ✓ *Модификация с использованием ферментов-«ножниц».* Ферменты (рестриктазы и лигазы) используются для «вырезания» генов или их фрагментов из одной ДНК и «вставки» в другую (вектор), что лежит в основе клонирования и создания рекомбинантных белков [11].
- ✓ *Использование ферментов для разделения изомеров*, а именно, стереоспецифический гидролиз и избирательное ацилирование [12].
- ✓ *Поиск и скрининг новых ферментов*. Прямой скрининг экстрактов и использование ДНК-зондов [13].
- ✓ *Направленная эволюция.* Её основы заложены в классической инженерии – это систематическая мутация генов ферментов, отбор и скрининг

получившихся вариантов для получения улучшенных характеристик, как устойчивость или активность [14].

Следует отметить, что традиционные подходы к инженерии ферментов, такие как направленная эволюция [15] и рациональный дизайн [16, 17], требуют значительных ресурсов для экспериментов и часто сопровождаются ограниченной предсказуемостью результатов. В этой связи методы глубокого обучения становятся важным инструментом для ускорения и оптимизации процессов ферментной инженерии.

Авторы исследования [8] отмечают: «…стратегии, основанные на данных, включая статистическое моделирование, машинное обучение и глубокое обучение, в значительной степени способствовали пониманию взаимосвязи последовательности, структуры и функции ферментов». В работе представлен *data-driven* подход к инженерии ферментов, основанный на анализе экспериментальных данных и аминокислотных последовательностей с использованием методов ML. Исследователи показывают, что модели, обученные на комбинации последовательностной и функциональной информации, способны выявлять скрытые закономерности, определяющие каталитические свойства ферментов. Такой подход позволяет не только прогнозировать функциональные характеристики известных ферментов, но и предсказывать перспективные мутации, повышающие эффективность биокатализа.

Особое значение в статье [8] придаётся использованию ML и DL как инструмента навигации по пространству возможных последовательностей. Классические модели машинного обучения, как правило, опираются на тщательно подобранные дескрипторы — физически, химически или статистически обоснованные признаки, извлекаемые из данных. В отличие от них, модели DL используют каскад из множества слоёв искусственных нейронных сетей, которые автоматически и иерархически формируют сложные многомерные представления (тензоры) признаков, извлекая скрытые паттерны без необходимости ручного конструирования дескрипторов. Авторы подчёркивают, что даже для относительно коротких белков количество потенциальных вариантов мутаций чрезвычайно велико, что делает полный экспериментальный перебор практически невозможным. Применение глубокого и машинного обучения позволяет сузить область поиска, что сокращает число лабораторных экспериментов.

Дополнительным направлением развития вычислительных методов в ферментной инженерии является автоматическая аннотация и предсказание функций ферментов по аминокислотной последовательности – одно из самых востребованных направлений, так как количество известных последовательностей в базах данных (например, UniProt) растет в разы быстрее, чем возможности исследователей проверить их функции «мокрыми» методами. В условиях экспоненциального роста биологических баз данных значительная часть обнаруженных белков остаётся плохо охарактеризованной с функциональной точки зрения. В этой связи особый интерес представляет модель EZYDeep, предложенная в работе [9], которая использует методы DL для классификации и предсказания функций ферментов на основе их последовательностей. Здесь авторы демонстрируют, что модель EZYDeep превосходит традиционные методы аннотации ML по точности и устойчивости, особенно при работе с неполными или эволюционно удалёнными последовательностями. Сравнение EZYDeep с другими методами представлено в работе [9], где наилучшие значения выделены жирным шрифтом (таблица 1).

Таблица 1 - Превосходство EZYDeep над HECNet и DEEPre

Table 1 - Superiority of EZYDeep over HECNet and DEEPre

–	EZYDeep		HECNet		DEEPre	
–	Acc	F_1	Acc	F_1	Acc	F_1
Уровень 0	98.56	98.25	94.0	94.1	95.9	95.9
Уровень 1	98.76	98.80	93.06	82.8	91.8	87.3
Уровень 2	98.23	94.59	93.5	70.6	88.8	63.4
Уровень 3	98.11	97.04	93.3	74.1	86.9	53.3

Модель обучается на репрезентативных наборах данных и способна автоматически извлекать информативные признаки, не требуя применения ручной инженерии [9].

Процесс работы EZYDeep следующий. На входе последовательности производится проверка последовательности, в случае положительного результата осуществляется переход. Далее следует кодирование данных, задается ферментный и/или неферментный прогноз, и по условию получают: либо неферментный прогноз, либо прогнозирование Enzyme Commission number (уникальный классификационный шифр, присваиваемый ферментам Международным союзом биохимии и молекулярной биологии IUBMB).

Важно отметить, что методы, представленные в [8] и [9], хорошо сочетаются с генеративными подходами, основанными на языковых моделях. В таком интегрированном сценарии большие языковые модели могут использоваться для генерации новых ферментных последовательностей, тогда как специализированные модели DL – для оценки их функционального потенциала и классификации. Эти методы позволяют перейти от итеративного подхода к более системному и предсказуемому проектированию биокатализаторов, что имеет прямое значение для развития промышленной и медицинской биотехнологии.

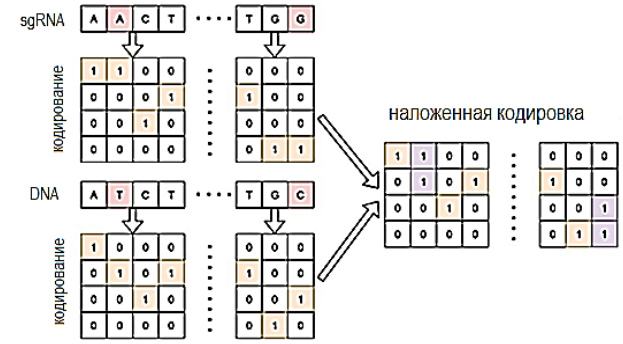
3. Геномное редактирование и интерпретируемый ИИ

Технологии геномного редактирования, в частности системы CRISPR Cas-9 (CRISPR-ассоциированный фермент Cas-9 – нуклеаза, «молекулярные ножницы», управляемые РНК,

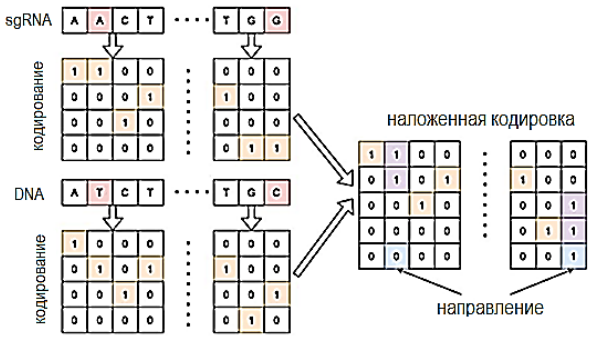
которые способны разрезать двуцепочечную ДНК в точно заданном месте), стали одним из наиболее значимых инструментов современной биотехнологии и молекулярной биологии. Несмотря на высокую точность и относительную простоту использования, CRISPR Cas-9 сопровождается рисками возникновения *off-target* эффектов – нежелательных разрезов ДНК вне целевого участка, которые могут приводить к мутациям и нарушению функций генома. В этой связи задача точного и надёжного прогнозирования *off-target* активности приобретает ключевое значение.

Методы DL широко применяются для оценки вероятности *off-target* эффектов, однако многие из них представляют собой «чёрные ящики», что ограничивает их использование в задачах, связанных с клинической безопасностью и регуляторными требованиями. В работе [10] предложен подход CRISPR-DIPOFF, направленный на объединение высокой точности прогнозирования с интерпретируемостью моделей машинного обучения. Авторы подчёркивают, что интерпретируемость является критическим фактором при использовании алгоритмов в контексте геномного редактирования, где необходимо понимать причины и механизмы получаемых предсказаний.

Ключевым этапом разработки модели CRISPR-DIPOFF является представление входных биологических данных. В работе [10] направляющие РНК (sgRNA) и соответствующие участки ДНК кодируются с использованием *one-hot encoding* (рис. 4).



(A) 4-канальное кодирование



(B) 5-канальное кодирование

Рис. 4 – Пример однократного кодирования

Fig. 4 – Example of single-pass encoding

Как показано на рис. 4 (A), последовательности sgRNA и ДНК независимо кодируются в четырёхканальном формате, после чего объединяются с помощью логической операции OR. Дополнительно вводится отдельный канал (рис. 4 (B) [10]), отражающий направление несовпадений между sgRNA и ДНК, что позволяет модели учитывать позиционные и структурные особенности связывания.

Такой способ кодирования сохраняет биологически значимую информацию о различиях между целевыми и потенциальными *off-target* участками и обеспечивает корректную подачу данных в нейронную сеть. Подход демонстрирует адаптацию методов представления данных, характерных для задач обработки последовательностей, к специфике геномного редактирования.

Архитектура моделей глубокого обучения, используемых в CRISPR-DIPOFF, основана на рекуррентных нейронных сетях. Закодированная входная матрица последовательно обрабатывается одним или двумя однонаправленными либо двунаправленными рекуррентными слоями (RNN, LSTM или GRU), за которыми следуют полносвязные скрытые слои с функцией активации ReLU и слоями исключения (dropout). Такая архитектура позволяет моделировать контекстные зависимости между позициями нуклеотидных последовательностей, что имеет принципиальное значение для корректного прогнозирования *off-target* активности.

Для достижения оптимального качества предсказаний в работе [10] проводится систематический подбор гиперпараметров моделей (внешних конфигурационных настроек алгоритма машинного обучения, задаваемых исследователем вручную до начала процесса обучения).

В таблице 2 [10] представлен перечень используемых гиперпараметров и диапазоны их значений, включая размеры скрытых слоёв, коэффициенты дропаута и параметры оптимизации.

Результаты выбора наилучших конфигураций для различных архитектур приведены в таблице 3 [10], где показаны значения метрик качества, включая AUPRC, на валидационной и тестовой выборках.

Сравнение эффективности различных архитектур глубокого обучения представлено в работе [10]. Несмотря на близкие значения точности и AUROC для разных моделей, метрики, чувствительные к дисбалансу классов, такие как F1-мера и AUPRC, демонстрируют существенные различия. Авторы показывают, что двунаправленная LSTM-модель обеспечивает наиболее сбалансированное соотношение показателей качества, что отражено в её архитектуре [10]. Модель имеет двунаправленную сеть LSTM, за которой следуют два скрытых слоя и выходной слой.

Особое внимание в работе о подходе CRISPR-DIPOFF уделяется вопросам интерпретируемости [10]. Авторами представлен анализ вкладов отдельных признаков и нейронов в итоговое предсказание модели, показаны наиболее значимые

признаки, влияющие на классификацию *off-target* событий, что позволяет выявить критические позиции в последовательностях sgRNA и ДНК. Тепловые карты активаций скрытых слоёв нейронов, приведённые в работе [10], позволяют визуально проследить процесс обработки входных данных моделью.

Таблица 2 - Список гиперпараметров и их возможных значений, использованных в исследовании [10]

Table 2 - List of hyperparameters and their possible values used in the study [10]

Наименование гиперпараметра	**Диапазон/ значения**	**Шаг/размер для диапазона**
Размер батча/пакета	32-256	удвоение количества шагов
Эпоха (полный проход по всем данным)	10-100	10
Количество рекуррентных слоев	1,2	–
Двунаправленные рекуррентные нейронные сети	Истина, Ложь	–
Вероятность дропаута	0.1-0.5	0.05
Скрытый слой	32-512	удвоение количества шагов
Количество скрытых слоев	0-6	1
Темп обучения	$1*10^{-5}$-0.5	изменение порядка величин 5, 10, 50, 100

Таблица 3 - Гиперпараметры лучших моделей RNN, LSTM и GRU, а также их оценки AUPRC на валидационном и тестовом выборках

Table 3 - Hyperparameters of the best RNN, LSTM, and GRU models, along with their AUPRC scores on the validation and test sets

Наименование гиперпараметра	**Тип модели**		
	RNN	**LSTM**	**GRU**
Размер	256	512	128
LSTM слои	2	1	2
Двунаправленная LSTM	Истина	Истина	Истина
Скрытые слои	0	2	0
Вероятность дропаута	0.4	0.4	0.1
Размер батча	256	64	64
Эпоха	60	50	30
Темп обучения	0.00100	0.00010	0.00050
AUPRC на валидационной выборке	0.7643	0.7403	0.7427
AUPRC на тестовом наборе	0.5711	0.7208	0.6859

Таким образом, подход CRISPR-DIPOFF демонстрирует, что интерпретируемые модели глубокого обучения могут эффективно применяться для прогнозирования *off-target* эффектов CRISPR Cas-9 без существенного снижения точности. В контексте представленного обзора данное исследование служит показательным примером применения интерпретируемого искусственного интеллекта в геномном редактировании, дополняя генеративные и предсказательные подходы, используемые в инженерии белков и ферментов.

Заключение

В работе представлен анализ современных подходов к применению больших языковых и глубоких моделей в инженерии белков, ферментов и геномном редактировании.

Следует отметить, что применение больших языковых моделей (LLM) в биотехнологии на сегодняшний день вышло за рамки обработки текстов и сосредоточено на анализе биологических последовательностей как «языков» жизни.

Рассмотренные исследования показывают, что методы искусственного интеллекта становятся важным инструментом вычислительного проектирования биомолекул, дополняя, а в ряде случаев превосходя традиционные экспериментальные подходы.

Показано, что LLM эффективно применяются для генерации и анализа аминокислотных последовательностей, захватывая эволюционные и функциональные закономерности, лежащие в основе структуры и функции белков [5–7].

В инженерии ферментов методы DL и ML позволяют повысить эффективность поиска и оптимизации биокатализаторов, а также автоматизировать предсказание их функций на основе последовательностной информации, то есть, на основе данных, обрабатываемых с учетом их временной последовательности и предыдущих состояний системы [8, 9].

В области геномного редактирования интерпретируемые модели глубокого обучения играют ключевую роль в прогнозировании *off-target* эффектов CRISPR Cas-9, что способствует повышению безопасности и надёжности данных технологий [10]. Это подчёркивает значимость разработки объяснимых алгоритмов при использовании искусственного интеллекта в биотехнологических прикладных проектах.

Вместе с тем применение больших языковых моделей в биотехнологии сопряжено с рядом ограничений, включая зависимость от качества обучающих данных, необходимости экспериментальной валидации и учета требований биобезопасности.

Дальнейшее развитие данной области связано с интеграцией генеративных и интерпретируемых моделей [18], а также с формированием ответственных подходов к использованию искусственного интеллекта в биотехнологии [19].

Литература

1. Э.Г. Милкова, *Colloquium-Journal*. № 8-5 (60), 4-5, (2020).

2. И.А. Филиппов, *Актуальные исследования.* № 32-1 (267), 6-8, (2025).
3. Liu Y.-T., Zhang L.-L., Jiang Z.-Y., Tian X.-S., Li P.-L., Wu P.-H., et al. Applications of Artificial Intelligence in Biotech Drug Discovery and Product Development, *MedComm*, **6**, 8, e70317 (2025). DOI: 10.1002/mco2.70317.
4. Wheeler N. E. Responsible AI in biotechnology: balancing discovery, innovation and biosecurity risks, *Frontiers in Bioengineering and Biotechnology*, **13**, 1537471 (2025). DOI: 10.3389/fbioe.2025.1537471.
5. Madani A., McCann B., Naik N., Keskar N. S., Anand N., Eguchi R. R., Huang P.-S., Socher R. ProGen: Language Modeling for Protein Generation, *BioRxiv*, 2020.03.07.982272 (2020). DOI: 10.1101/2020.03.07.982272.
6. Madani A., Krause B., Greene E. R., et al. Large language models generate functional protein sequences across diverse families, *Nature Biotechnology*, **41**, 8, 1099–1106 (2023). DOI: 10.1038/s41587-022-01618-2.
7. Baek M., Baker D. Deep learning and protein structure modeling, *Nature Methods*, **19**, 1, 13–14 (2022). DOI: 10.1038/s41592-021-01360-8.
8. Jiang Y., Ran X., Yang Z. J. Data-driven enzyme engineering to identify function-enhancing enzymes, *Protein Engineering, Design and Selection*, **36**, gzac009 (2023). DOI: 10.1093/protein/gzac009.
9. Boulahrouf K., Aliouane S., Chehili H., Skander Daas M., Belbekri A., Hamidechi M. EZYDeep: A Deep Learning Tool for Enzyme Function Prediction based on Sequence Information, *The Open Bioinformatics Journal*, **16**, e187503622306270 (2023). DOI: 10.2174/18750362-v16-230705-2023-7.
10. Toufikuzzaman M., Hassan Samee M. A., Rahman M. S., et al. CRISPR-DIPOFF: an interpretable deep learning approach for CRISPR Cas-9 off-target prediction, *Briefings in Bioinformatics*, **25**, 2, bbad530 (2024). DOI: 10.1093/bib/bbad530.
11. Р.Ф. Хайруллин, Р.Г. Киямова, А.А. Ризванов. *Экспрессия рекомбинантных белков в E.coli:* учеб. пособие. Казань: Изд-во Казан. ун-та, 2018. 142 с.
12. В.К. Османов. *Инженерная энзимология*: учебно-методическое пособие. Н. Новгород: Изд-во Нижегородской гос. медицинской академии, 2014. 68 с.
13. А. Н. Синякова, Е. В. Костина, *Молекулярная биология*. **57**, 3, 440-457 (2023).
14. А.С. Черкашина, О.О. Михеева, В.Г. Акимкин, *Вестник Московского университета.* Серия 2: Химия. **65**, 2, 113-120 (2024).
15. Т.Е. Тюгашев, О.С. Федорова, Н.А. Кузнецов, *Молекулярная биология*, **57**, 2, 209-219 (2023).
16. В.И. Тишков, А.А. Алексеева, И.В. Голубев, В.В. Федорчук, И.С. Каргов, С.А. Зарубина, И.А. Долина, Д.Л. Атрошенко, Г.С. Захарова, А.А. Полозников, Т.С. Виролайнен, Р.П. Ковалевский, А.В. Степашкина, Т.А. Чубарь, И.В. Упоров, А.В. Скляренко, С.В. Яроцкий, С.С. Савин, В сборнике: *Биотехнология: состояние и перспективы развития. материалы VIII Московского Международного Конгресса*. ЗАО «Экспо-биохим-технологии», РХТУ им. Д.И. Менделеева. 2015. С. 452-453.
17. В.И. Тишков, А.А. Пометун, А.В. Степашкина, В.В. Федорчук, С.А. Зарубина, И.С. Каргов, Д.Л. Атрошенко, П.Д. Паршин, М.Д. Шеломов, Р.П. Ковалевский, К.М. Бойко, М.А. Эльдаров, Э. Д'Оронцо, Ф. Секундо, С.С. Савин, *Вестник Московского университета*. Серия 2: Химия. **59**, 2, 70-77 (2018).
18. С.К. Долгих, Л.Ю. Кошкина, *Вестник Технологического университета*, **28**, 10, 86-90, (2025).
19. Е.П. Попечителев, *Основы биотехтоники. Технические системы – инструмент практической деятельности человека*. Старый Оскол : ТНТ, 2026. 376 с.

References

1. E.G. Milkova, *Colloquium-Journal*. No. 8-5 (60), 4–5, (2020).
2. I.A. Filippov, *Current Research.* No. 32-1 (267), 6–8, (2025).
3. Liu Y. -T., Zhang L.-L., Jiang Z.-Y., Tian X.-S., Li P.-L., Wu P.-H., et al. Applications of Artificial Intelligence in Biotech Drug Discovery and Product Development, *MedComm*, **6**, 8, e70317 (2025). DOI: 10.1002/mco2.70317.
4. Wheeler N. E. Responsible AI in biotechnology: balancing discovery, innovation, and biosecurity risks, *Frontiers in Bioengineering and Biotechnology*, **13**, 1537471 (2025). DOI: 10.3389/fbioe.2025.1537471.
5. Madani A., McCann B., Naik N., Keskar N. S., Anand N., Eguchi R. R., Huang P.-S., Socher R. ProGen: Language Modeling for Protein Generation, *BioRxiv*, 2020.03.07.982272 (2020). DOI: 10.1101/2020.03.07.982272.
6. Madani A., Krause B., Greene E. R., et al. Large language models generate functional protein sequences across diverse families, *Nature Biotechnology*, **41**, 8, 1099–1106 (2023). DOI: 10.1038/s41587-022-01618-2.
7. Baek M., Baker D. Deep learning and protein structure modeling, *Nature Methods*, **19**, 1, 13–14 (2022). DOI: 10.1038/s41592-021-01360-8.
8. Jiang Y., Ran X., Yang Z. J. Data-driven enzyme engineering to identify function-enhancing enzymes, *Protein Engineering, Design and Selection*, **36**, gzac009 (2023). DOI: 10.1093/protein/gzac009.
9. Boulahrouf K., Aliouane S., Chehili H., Skander Daas M., Belbekri A., Hamidechi M. EZYDeep: A Deep Learning Tool for Enzyme Function Prediction based on Sequence Information, *The Open Bioinformatics Journal*, **16**, e187503622306270 (2023). DOI: 10.2174/18750362-v16-230705-2023-7.
10. Toufikuzzaman M., Hassan Samee M. A., Rahman M. S., et al. CRISPR-DIPOFF: an interpretable deep learning approach for CRISPR Cas-9 off-target prediction, *Briefings in Bioinformatics*, **25**, 2, bbad530 (2024). DOI: 10.1093/bib/bbad530.
11. R.F. Khairullin, R.G. Kiyamova, A.A. Rizvanov. *Expression of Recombinant Proteins in E. coli:* textbook. Kazan: Kazan University Press, 2018. 142 pp.
12. V.K. Osmanov. *Engineering Enzymology*: teaching manual. Nizhny Novgorod: Nizhny Novgorod State Medical Academy Press, 2014. 68 pp.
13. A.N. Sinyakova, E.V. Kostina, *Molecular Biology*. **57**, 3, 440–457 (2023).
14. A.S. Cherkashina, O.O. Mikheeva, V.G. Akimkin, *Bulletin of Moscow University.* Series 2: Chemistry. **65**, 2, 113–120 (2024).
15. T.E. Tyugashev, O.S. Fedorova, N.A. Kuznetsov, *Molecular Biology*, **57**, 2, 209–219 (2023).
16. V.I. Tishkov, A.A. Alekseeva, I.V. Golubev, V.V. Fedorchuk, I.S. Kargov, S.A. Zarubina, I.A. Dolina, D.L. Atroshenko, G.S. Zakharova, A.A. Poloznikov, T.S. Virolainen, R.P. Kovalevsky, A.V. Stepaskina, T.A. Chubar, I.V. Uporov, A.V. Sklyarenko, S.V. Yarotsky, S.S. Savin, In: *Biotechnology: Current Status and Prospects for Development. Proceedings of the 8th Moscow International Congress*. Expo-Biokhim-Technologii CJSC, D.I. Mendeleev Russian Chemical Technology University. 2015. pp. 452–453.
17. V.I. Tishkov, A.A. Pometun, A.V. Stepashkina, V.V. Fedorchuk, S.A. Zarubina, I.S. Kargov, D.L. Atroshenko, P.D. Parshin, M.D. Shelomov, R.P. Kovalevsky, K.M. Boiko,

M.A. Eldarov, E. D'Oronzo, F. Secondo, S.S. Savin, *Bulletin of Moscow University*. Series 2: Chemistry. **59**, 2, 70–77 (2018).
18. S.K. Dolgikh, L.Yu. Koshkina, *Herald of Technological University*, **28**, 10, 86–90, (2025).
19. E.P. Popechitelev, *Fundamentals of Biotechtonics. Technical Systems – Tools of Human Practical Activity*. Stary Oskol: TNT, 2026. 376 pp.

© С. К. Долгих – магистрант по направлению 2.09.04.01 «Информатика и вычислительная техника» (Проектирование интеллектуальных компьютерных систем), Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия, godvelonsky@gmail.com; **Л. Ю. Кошкина** – кандидат технических наук, доцент кафедры Химической кибернетики, Казанский национальный исследовательский технологический университет, Казань, Россия, KoshkinaLYu@corp.knrtu.ru

© S. K. Dolgikh – Master-student in 2.09.04.01 "Informatics and computer engineering", Great Peter St.-Petersburg Polytechnic University, Sankt-Petersburg, Russia, godvelonsky@gmail.com; **L. Yu. Koshkina** – PhD (Technical Sci.), Associate Professor of the Kazan National Research Technological University, Kazan, Russia, KoshkinaLYu@corp.knrtu.ru.

Дата поступления рукописи в редакцию – 03.02.26.
Дата принятия рукописи в печать – 02.03.26.