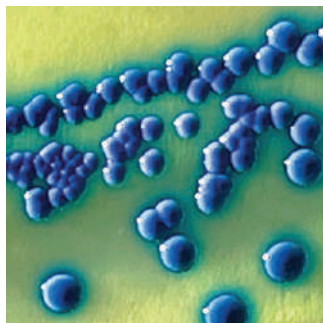




Межрегиональная ассоциация по клинической микробиологии и антимикробной химиотерапии Научно-исследовательский институт антимикробной химиотерапии ФГБОУ ВО СГМУ Минздрава России

Учредители:

Синопальников А.И.; Пискунов Г.Г.; Козлов Р.С.; Межрегиональная ассоциация по клинической микробиологии и антимикробной химиотерапии (МАКМАХ)

Главный редактор:

Синопальников А.И.

Адрес редакции:

214019, Смоленская обл., г. Смоленск, ул. Кирова, д. 46А

Эл. почта: info@cmac-journal.ru

Адрес для корреспонденции:

214019, г. Смоленск, а/я 5.

Тел./факс: +7(4812)52-06-02

Издатель МАКМАХ:

214019, г. Смоленск, ул. Кирова 46А. www.iacmac.ru

Адрес типографии:

214020, Россия, г. Смоленск, ул. Смольянинова, д. 1

Электронная версия журнала:

https://cmac-journal.ru

Подписка на сайте издателя:

https://service.iacmac.ru

Журнал зарегистрирован Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор).

Запись в реестре зарегистрированных СМИ: ПИ № ФС 77 – 86269 от 27.11.2023

Не распространяется через предпринятия связи

Тираж 3000 экз.

Свободная цена

Дата выхода – 04.08.2026

Журнал входит в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук

Присланные в редакцию статьи проходят рецензирование

Мнение редакции может не совпадать с точкой зрения авторов публикуемых материалов

Ответственность за достоверность рекламных публикаций несут рекламодатели

При перепечатке ссылка на журнал обязательна

Журнал является научным изданием для врачей, в связи с чем на него не распространяются требования Федерального закона от 29.12.2010 №436-ФЗ «О защите детей от информации, причиняющей вред их здоровью и развитию»

Иллюстрация для обложки предоставлена: Игорь Викторович Чеботарь (НИИ антимикробной химиотерапии ФГБОУ ВО СГМУ Минздрава России)

© Клиническая микробиология и антимикробная химиотерапия, 2026.

Содержание

Болезни и возбудители

- Шаповалова В.В., Шайдуллина Э.Р., Шипулин Г.А., Эйдельштейн М.В.
4 Анализ таксономической структуры рода *Stenotrophomonas* на основе геномных данных
- Андреев В.А., Привольнев В.В., Родин А.В., Новиков И.А., Голуб А.В.
13 Местное лечение раневой инфекции: что нового за последнее десятилетие?
- Григорян Д.А., Горбачева А.А., Мацвай А.Д.
25 Современные представления о роли вирусов как этиологического фактора сепсиса
- Домблдес Э.А., Чернуха М.Ю., Аветисян Л.Р., Киян Т.А., Жуховицкий В.Г., Гущин В.А., Кондратьева Е.И.
50 Этиологическая структура инфекций респираторного тракта у пациентов различных возрастных групп с первичной цилиарной дискинезией

Антимикробные препараты

- Резолюция по итогам круглого стола
64 Пневмонии. От пенициллина до наших дней: что изменилось за последние сто лет
- Зырянов С.К., Бондарева И.Б., Пуцман Г.А., Ченкуров М.С., Савоцкая А.М., Коваль В.Р.
72 Исследования фармакокинетики и фармакодинамики линезолида у недоношенных новорожденных

Антибиотикорезистентность

- Кузьменков М.Ю., Виноградова А.Г., Кузьменков А.Ю.
81 Методы машинного обучения в прогнозировании антибиотикорезистентности
- Белькова Ю.А., Козлов Р.С., Рачина С.А.
92 Роль мониторинга применения системных антимикробных препаратов в профилактике селекции лекарственной устойчивости

Опыт работы

- Фалалеева С.О., Ахмедова Э.И., Кирпичева А.В., Новикова Т.О., Курамшина В.Р., Желтова А.Р., Саушкина А.М.
103 Мониторинг этиологической структуры внебольничных инфекций мочевых путей и антибиотикорезистентности ведущего уропатогена – *Escherichia coli* у детей и подростков в Красноярском крае
- Полякова Е.А., Мательский Н.А., Горбич О.А., Горбич Ю.Л., Солнцева А.В.
108 Иммуногенетические предикторы тяжелого течения сепсиса: роль функциональных полиморфизмов
- Салмин А.В., Семенова О.В., Турсунова Н.В., Емельянова Е.П.
124 Микробиом нижних дыхательных путей у пациентов с туберкулезом легких и ВИЧ-инфекцией
- Симончик М.В., Тапальский Д.В., Карпова Е.В., Тапальский С.Д.
131 Антимутантная активность комбинации колистина и рифаксимины при селективной деконтаминации кишечника: валидация в динамической проточной системе

Методы машинного обучения в прогнозировании антибиотикорезистентности

Кузьменков М.Ю., Виноградова А.Г., Кузьменков А.Ю.

ФГБОУ ВО «Смоленский государственный медицинский университет» Минздрава России, Смоленск, Россия

Контактный адрес:

Михаил Юрьевич Кузьменков
Эл. почта: kuzmenkov.m.u@yandex.ru

Ключевые слова: антибиотикорезистентность, машинное обучение, нейронные сети, геном.

Конфликт интересов: авторы заявляют об отсутствии конфликтов интересов.

Антимикробная резистентность представляет собой серьезную глобальную проблему, угрожающую системам здравоохранения во всем мире. Высокая стоимость разработки новых препаратов и ограниченный срок их применения ввиду развития устойчивости осложняют ситуацию, что стимулирует активное использование различных подходов, которые могут внести вклад в решение проблемы. В данной статье анализируются современные методы прогнозирования антибиотикорезистентности отдельно взятого изолята. Обзор охватывает как традиционные алгоритмы машинного обучения, такие как логистическая регрессия и «случайный лес», так и передовые подходы, включающие сверточные и графовые нейронные сети, а также трансформерные языковые модели. Высокая изменчивость бактериальных геномов и неоднородность мутационных процессов затрудняют обобщающую способность моделей, а различия в клинических условиях формирования штаммов приводят к нестабильности предсказаний на новых данных. Это ставит под сомнение универсальность текущих решений и указывает на необходимость дальнейшего изучения методов сбора и обработки данных, а также учета ранее неизвестных и вновь возникающих мутаций.

Review

Machine learning methods in predicting antimicrobial resistance

Kuzmenkov M.Yu., Vinogradova A.G., Kuzmenkov A.Yu.

Smolensk State Medical University, Smolensk, Russia

Contacts:

Mikhail Yu. Kuzmenkov
E-mail: kuzmenkov.m.u@yandex.ru

Key words: antimicrobial resistance, machine learning, neural networks, genome.

Conflicts of interest: all authors report no conflicts of interest relevant to this article.

Antimicrobial resistance poses a significant global threat to healthcare systems worldwide. The high cost and short lifespan of new antibiotics due to the rapid evolution of resistance exacerbate this crisis, prompting exploration of diverse strategies to combat AMR. Within the context of AMR diagnostics, machine learning approaches for predicting bacterial resistance are under active investigation. This paper reviews state-of-the-art methods for predicting antibiotic resistance in individual bacterial isolates. The review encompasses both traditional ML algorithms, such as logistic regression and random forests, and more advanced techniques including convolutional and graph neural networks, as well as transformer language models. The inherent high variability of bacterial genomes and the heterogeneity of mutational processes hinder model generalizability, while variations in clinical conditions during strain development contribute to prediction instability on new datasets. This necessitates further research into improved data collection and preprocessing methods, as well as a more comprehensive consideration of both known and emerging resistance mechanisms.

Введение

Антимикробная резистентность (АМР) – это глобальная угроза для здравоохранения во всем мире, согласно данным 2019 г. в Европейском регионе ВОЗ с АМР было связано около 541 тыс. смертей [1]. В контексте диагностики АМР ведутся интенсивные исследования, направленные на разработку и валидацию прогностических моделей, базирующихся на методах машинного обучения, для предсказания фенотипической устойчивости бактериальных патогенов к антимикробным пре-

паратам. Существующие исследования охватывают широкий спектр алгоритмов машинного обучения (МО), а также различные источники данных, используемых для построения моделей, включая геномную информацию, данные о клинических характеристиках пациентов и результаты лабораторных анализов для прогнозирования как фенотипа и динамики АМР, так и поддержки врачебных решений [2–5]. Искусственный интеллект (ИИ) превосходит человеческие возможности в анализе данных,

что потенциально позволяет преодолеть ограничения, связанные с нехваткой информации и времени, обобщающей способностью, а также исключает субъективные факторы принятия решений. В контексте диагностики АМР, машинное обучение в перспективе сможет значительно повысить эффективность проводимых исследований [6]. Данный обзор литературы посвящен применению МО для прогнозирования АМР отдельного взятого изолята.

Классические методы машинного обучения для прогнозирования АМР на основе клинических данных

Машинное обучение представляет собой набор алгоритмов, способных вычислять (предсказывать) некоторую характеристику изучаемых данных по заданным признакам. Существует большое количество таких алгоритмов, самыми известными из которых являются нейронные сети, которые также отличаются большим разнообразием. Все эти методы с целью диагностики инфекций и выбора антимикробной терапии (АМТ) могут применяться для предсказания наличия того или иного свойства микроорганизмов: фенотип резистентности микроорганизма к определенному антимикробному препарату (АМП). Но для предсказания определенного свойства алгоритму машинного обучения необходимо обладать знаниями об изучаемом объекте. В исследовании Goodman K. и соавт. [7] древовидная модель на основе клинических данных (например, длительность приема антибиотиков и пребывания в стационаре) прогнозировала наличие бактериемии, вызванной продуцирующими бета-лактамазы расширенного спектра (БЛРС) *Escherichia coli*, *Klebsiella pneumoniae* или *Klebsiella oxytoca*. Таким образом, модели машинного обучения не предсказывали резистентность напрямую, но помогали сделать вывод о вероятности ее наличия. Чувствительность лучшей модели составила 90,8%. Кроме того, изучалась возможность прогнозирования чувствительности микроорганизмов к АМП в условиях отсутствия лабораторных данных [8]. В качестве данных для прогноза фенотипа резистентности, выступали некоторые характеристики пациента (возраст, пол и др.), тип материала (моча, кровь и др.), а также результаты окрашивания по Граму. Максимальные значения чувствительности и специфичности достигали 72% и 73% соответственно. Лучшей моделью стал алгоритм KNN (k-ближайших соседей). Конкретные виды микроорганизмов и АМП в публикации не приводятся.

Учитывая важность адекватной стартовой антимикробной терапии [9] особое значение приобретает быстрый выбор конкретного АМП для эмпирической терапии. Например, в публикации McGuire R. и соавт. [10] рассматривается возможность использования различных клинических данных пациентов (длительность АМТ в днях, наличие венозного катетера, локализация инфекции, наличие хирургического вмешательства и др.) для прогнозирования наличия или отсутствия резистент-

ности к карбапенемам у изолята, вызвавшего случай инфекции. Для построения модели был использован алгоритм градиентного бустинга над деревьями, главным принципом которого является создание множества древовидных моделей, которые последовательно дополняют и уточняют прогноз друг друга. Проверка качества модели продемонстрировала чувствительность 30% (наличие устойчивости к карбапенемам) и специфичность 99% (отсутствие устойчивости к карбапенемам). Похожее исследование проводилось Oonsivilai M. и соавт. [3]. Однако оно сосредоточено на прогнозировании фенотипа резистентности у бактерий при сепсисе у пациентов младше 16 лет в одной из детских больниц в Камбодже. В работе была предпринята попытка обучить модели, которые способны прогнозировать результаты окраски по Граму, а также резистентность к цефтриаксону, ампициллину и гентамицину. Особый интерес представляет набор из 35 собранных признаков, которые использовались при построении модели: возраст, масса тела к возрасту, количество посещений поликлиники, качество потребляемой воды и др. Лучшей моделью по результатам проведенного исследования был признан алгоритм «случайный лес», основным принципом его работы является построение множества простых древовидных моделей, а для формирования финального результата (итогового прогноза) рассчитывается усредненный результат прогноза всех деревьев. В данном исследовании максимальные результаты метрики (площадь под ROC-кривой – AUC), отражающей качество классификации, достигали значения 0,85 для прогнозирования устойчивости к ампициллину и гентамицину, 0,74 – для прогнозирования резистентности к цефтриаксону, и 0,71 для прогнозирования результатов окраски по Граму.

Многие исследования являются узконаправленными и рассматривают возможность применения машинного обучения для выявления резистентности конкретных микроорганизмов у определенных групп пациентов. Таким исследование является работа Henderson H. и соавт. [11], описывающая обучение моделей с предсказанием множественной лекарственной устойчивости (МЛУ) у представителей порядка Enterobacterales. Авторы собрали данные по 4534 пациентам с ВИЧ-инфекцией, у 343 (8%) из которых были выявлены Enterobacterales с МЛУ при исследовании клинического образца, взятого для диагностики бактериальной инфекции. Данные включали: пол, возраст пациента, количество госпитализаций за последний год, наличие хронических заболеваний, уровень CD4-клеток, вирусную нагрузку и др. В качестве модели машинного обучения была использована логистическая регрессия с регуляризацией. Максимальные результаты чувствительности и специфичности достигали 67% и 76% соответственно. Схожий набор признаков пациентов использовался в исследовании Hebert C. и соавт. [12], посвященном разработке моделей машинного обучения для прогнозирования резистентности микроорганизмов к АМП при инфекциях мочевыводящих путей. Модели обучались

на данных из электронной медицинской карты пациентов, включая демографические сведения и информацию о недавнем использовании АМП, пребывании в отделении реанимации и интенсивной терапии и др. В задачу моделей входило прогнозирование резистентности к цефазолину, цефтриаксону, ципрофлоксацину, цефепиму и пиперациллину-тазобактаму. В качестве метрики для оценки качества прогноза использовалась площадь под ROC-кривой – AUC, которая варьировалась от 0,65 для цефтриаксона до 0,69 для цефазолина.

Таким образом, в представленных исследованиях использовались клиничко-лабораторные данные пациентов для предсказания фенотипа резистентности бактерий с помощью так называемых «классических» методов машинного обучения: «случайный лес», метод опорных векторов, k-ближайших соседей и др. Однако их надежность признается ограниченной, так как резистентность определяется в первую очередь геномом бактерии, а клинические данные лишь опосредованно отражают эту картину и привязаны к конкретным условиям. Более перспективным представляется подход, при котором модели машинного обучения работают непосредственно с полными геномными последовательностями, предсказывая резистентность по наличию определенных генетических паттернов независимо от локальных клинических данных. Следует отметить, что в рамках данного обзора не рассматриваются «классические» биоинформационные подходы, основанные на технологии текстового поиска искомого гена, т.к. они не относятся к методам машинного обучения и не способны анализировать широкий геномный контекст, важный для построения прогностических моделей.

Глубокие нейронные сети для прогнозирования АМП на гетерогенных данных

В вопросах применения машинного обучения для прогнозирования резистентности бактерий ключевую роль играют те модели, которые обучаются на геномах бактерий. Наиболее часто к таким механизмам относят рекуррентные нейронные сети и трансформеры. Исторически первыми появились рекуррентные нейронные сети, однако после 2017 г. они уступили место трансформерам, которые были впервые описаны в статье «Attention Is All You Need» [13]. Архитектура данной нейронной сети особенно эффективна для обработки последовательностей данных, таких как текст, код, временные ряды и биологические последовательности (например, ДНК или белки). В отличие от рекуррентных нейронных сетей трансформеры используют механизм внимания, позволяющий им параллельно обрабатывать всю последовательность. Это значительно ускоряет обучение и позволяет обрабатывать более длинные последовательности [14]. Интересен подход к применению больших языковых моделей в диагностике антибиотикорезистентности, где в качестве основы использовали предобученную модель GPT-4 (OpenAI) [15]. Одна из версий модели оставалась неизменной, в то время как другая подвергалась

дополнительному обучению на пороговых значениях EUCAST для определения клинической категории чувствительности микроорганизмов к АМП [16]. Для дообучения использовались документы: EUCAST Breakpoint Tables (v13.1), EUCAST Expert Rules (v3.2) и EUCAST Guidelines for Resistance Mechanisms (v2.0). В дообученную модель отправлялась фотография чашки Петри с размеченными зонами подавления роста, а также таблицей, которая содержала числовые данные диаметров зон подавления роста. Также направлялся текстовый запрос (промт), в котором исследователи указывали, что модели необходимо сопоставить таблицу и изображение чашки Петри и сделать вывод о наличии определенного фенотипа (нет искомым маркеров резистентности, БЛРС, AmpC, продуцент карбапенемазы) на основе правил EUCAST. Данный вывод модель должна была оформить в виде таблицы и дать своим ответам вероятностную оценку, указав степень уверенности в собственных ответах, а также предложить дополнительные тесты для подтверждения результатов. В итоге исследователи добились чувствительности свыше 95%, но специфичность оставалась на относительно низком уровне, показав худший результат для обнаружения БЛРС – 69%.

Успешное прогнозирование фенотипа резистентности бактерий на основе генома демонстрируется в статье «Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning» [17]. В данной работе использовались данные полногеномного секвенирования 987 изолятов *E. coli* полученных из клинических образцов людей и животных (собственные данные исследователей), а также 1509 изолятов *E. coli*, полученных из открытых источников. Результаты определения чувствительности к ципрофлоксацину, цефотаксиму, цефтазидиму и гентамицину выступили прогнозируемым результатом. Следует отметить, что для дальнейшего обучения использовался не весь геном, а только матрица SNPs (Single Nucleotide Polymorphisms): исследователи выделяли гены, где были однонуклеотидные замены по сравнению с референсным геномом, получив сжатое представление генома. Кодирование происходило по одному из трех известных методов: Label Encoding (каждому нуклеотиду присваивается число: A → 1, C → 2, G → 3, T → 4, N → 0); One-Hot-Encoding (преобразование в бинарную матрицу, где каждый нуклеотид представлен вектором из 5 значений, 4 из которых – это число 0, и одно число это – 1, которая стоит на позиции соответствующей буквы – A, T, G, C или N); FCGR Encoding (Frequency Chaos Game Representation) – представление в виде фрактальных изображений (более сложный метод, который превращает ДНК-последовательность в 2D-изображение). Полный геном никогда не подавался в модели машинного обучения напрямую – сначала он сжимался в матрицу, а затем кодировался числовыми методами. В качестве моделей использовались логистическая регрессия, метод опорных векторов, «случайный лес» и свёрточная нейронная сеть. Оценка качества модели происходила при помощи метрики AUC. Конкретные значения получен-

Таблица 1. Метрики качества прогностических моделей при различных способах предобработки геномов

Антибиотик	Кодирование	Свёрточная нейронная сеть	Логистическая регрессия	«Случайный лес»	Метод опорных векторов
Ципрофлоксацин	Label Encoding	0,9	0,92	0,96	0,91
Ципрофлоксацин	One-Hot Encoding	0,93	0,93	0,96	0,92
Ципрофлоксацин	FCGR Encoding	0,92	0,92	0,95	0,91
Цефотаксим	Label Encoding	0,77	0,72	0,82	0,69
Цефотаксим	One-Hot Encoding	0,76	0,75	0,81	0,71
Цефотаксим	FCGR Encoding	0,8	0,8	0,81	0,79
Цефтазидим	Label Encoding	0,88	0,88	0,93	0,81
Цефтазидим	FCGR Encoding	0,89	0,84	0,93	0,86
Цефтазидим	One-Hot Encoding	0,86	0,88	0,92	0,85
Гентамицин	Label Encoding	0,85	0,88	0,95	0,83
Гентамицин	One-Hot Encoding	0,89	0,84	0,93	0,86
Гентамицин	FCGR Encoding	0,85	0,80	0,95	0,8

ных метрик качества обученных моделей представлены ниже в Таблице 1.

Исследователи указывают, что наилучший результат – «случайный лес», особенно для ципрофлоксацина (AUC = 0,96). Еще одним значимым результатом данного исследования стало выявление способности обученных моделей, использующих методы отбора признаков, идентифицировать гены, которые в наибольшей степени ассоциированы с устойчивостью к АМП. В частности отмечается, что было идентифицировано 19 геномных регионов, связанных с АМП, например, *nhaA* (Na⁺/H⁺-антипортер, регулирующий проницаемость мембраны); *marA* (участвует в развитии множественной лекарственной устойчивости); *murB* (участвует в синтезе пептидогликана).

Некоторые исследователи [18] подчеркивают значимость пространственной организации генов, что определяет их функциональную активность. В научной литературе всё чаще утверждается необходимость анализа геномов методами, аналогичными обработке текстовой информации. Основываясь на этой идее, многие исследователи применяли особый вид нейронных сетей – рекуррентные нейронные сети [19]. Однако этот метод в машинном обучении уходит в прошлое в виду своих технических ограничений, таких как продолжительное обучение, долгая работа и неспособность качественно обрабатывать длинные последовательности. Поэтому в рамках данного обзора он рассматриваться не будет.

Энкодерные трансформерные модели для прогнозирования АМП по геномным данным

На смену рекуррентным нейронным сетям в 2017 г. пришли трансформеры, которые учитывают взаимное расположение слов, отличаются высокой скоростью ра-

боты и обучения [14]. Трансформеры – это целый класс нейронных сетей, включающий множество подвидов, которые отличаются в зависимости от выполняемой задачи: классификация текста, перевод текста, генерация текста и т.д. Но в целом все нейронные сети, использующие трансформеры, делятся на два больших класса: энкодеры и декодеры. Задача энкодеров сводится к сворачиванию длинной текстовой последовательности в небольшой числовой вектор, который отражает смысл входного текста. Далее на этом векторе могут обучаться другие модели машинного обучения под конкретную задачу. Задача декодеров сводится к генерации новой последовательности текста. Трансформеры-энкодеры способны сжимать длинное текстовое представление в короткие вектора, а трансформеры-декодеры генерируют новую последовательность текста, исходя из поданных на вход данных, например, генерируя ответ на вопрос. Первой известной работой, в которой применялись трансформеры-энкодеры к геномам была публикация «DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome» 2020 г. [20]. Авторы исследования отмечают, что о геноме можно думать, как о длинном предложении, где последовательности букв А, Т, G, С составляют отдельные слова, то есть гены. В данном исследовании авторы взяли модель BERT (Bidirectional Encoder Representations from Transformers) [21], созданную, чтобы представлять слова или предложения в векторном виде – последовательность некоторых чисел, которые отражают суть изучаемого явления. Используя такие вектора, можно затем применять иные методы машинного обучения, такие как логистическая регрессия, например, для классификации предложения, поданного на вход BERT. Особенностью BERT является способ обучения: для модели не требуются истинные целевые прогнозируемые метки для ге-

номов, она пытается понять суть изучаемых данных исходя из самого контекста данных и не делает прогнозов. Процесс обучения основывался на задаче предсказания слов, которые были искусственно скрыты в предложении (маскированы). Например, модели предоставлялось предложение вида: «Бактерия имеет палочковидную форму и грамположительная, окрашиваясь по Граму в <пропуск> цвет». Задача BERT заключалась в определении наиболее вероятного слова для заполнения маскированного элемента <пропуск>, в данном случае – «фиолетовый». Таким образом, модель учитывала контекст, предшествующий и последующий за пропущенным словом, и формировала семантические связи между словами, часто встречающимися в едином контексте (например, «грамположительная», «фиолетовый», «бактерия»). Но поскольку геном – это своеобразный текст, то обучение модели BERT на нем должно дать модели представление о взаимных зависимостях генов между собой. То есть BERT может научиться восстанавливать пропущенные нуклеотидные последовательности из геномов. Именно такой идеей руководствовались авторы публикации «DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome» и обучили BERT на референсном геноме человека GRCh38.p13. Обученная модель получила название DNABERT. Для обучения модели вся геномная последовательность разбивалась на короткие k-меры длиной от 3 до 6. Наборы полученных данных, состоящие из 3, 4, 5 и 6 k-меров (всего 4 набора данных), использовались для обучения независимых моделей DNABERT. Каждый такой k-мер называется токеном. Например, токен (k-мер) из 3 элементов может выглядеть как «TGC» или «ATG» и т.д. Поскольку последовательность k-меров слишком длинная (она состоит из всего генома, разбитого по 3, 4, 5 или 6 нуклеотидов соответственно), в модель DNABERT k-меры подавались частями. Каждая часть состояла из 512 токенов (то есть 512 k-меров, которые идут по порядку в геноме). При этом 15% токенов были искусственно удалены. Именно эти 15% удаленных токенов и требовалось восстановить.

Конечная задача DNABERT сводится к тому, что каждый отдельный k-мер представляется моделью в виде вектора чисел, который несет в себе информацию о сущности последовательности из 3-6 нуклеотидов. Этот этап обучения называется предобучением. Можно сказать, что на этом этапе модель понимает «грамматику» и «синтаксис» ДНК, представляя результат в виде вектора чисел, но не осуществляет конкретных предсказаний. Далее, используя полученные векторные представления последовательностей нуклеотидов можно обучать модель под конкретные цели. К примеру, в оригинальной статье авторы дообучили модель для прогнозирования промоторов, сайтов связывания с транскрипционными факторами (TFBSs) и сайтов сплайсинга. Авторы отмечают высокую вычислительную сложность обучения модели DNABERT на различных геномных данных. В связи с этим, для оценки эффективности использования DNABERT в создании векторных представлений

геномов был проведен сравнительный анализ. В эксперименте были использованы различные модели машинного обучения (включая «случайный лес» и сверточную нейронную сеть), которым в качестве входных данных предоставлялись как последовательности нуклеотидов (в данном эксперименте геном мыши), так и их векторные представления, полученные с помощью предварительно обученной модели DNABERT (обученной на данных человека). Результаты показали, что модели, использующие векторные представления, сгенерированные DNABERT, продемонстрировали улучшенную производительность по сравнению с моделями, работающими непосредственно с последовательностью нуклеотидов. При этом важно отметить, что модель DNABERT не содержала информации о геноме мыши во время обучения.

В 2024 г. вышла новая версия описанной модели – DNABERT-2 [22]. Основное отличие кроется в том, что в DNABERT-2 для получения токенов вместо k-мер использовался особый метод токенизации – BPE (Byte Pair Encoding). Суть BPE заключается в том, что токеном становится не последовательность нуклеотидов фиксированной длины как при k-мерном представлении, а последовательность, которая встречается чаще остальных в данном геноме. Стоит отметить, что заслугой авторов DNABERT-2 является не только новая версия DNABERT, но и создание универсального публично доступного набора данных [23] GUE (Genome Understanding Evaluation), в котором собраны геномы человека и других живых организмов (всего 135 видов организмов) с общим размером в 32,48 млрд нуклеотидов. Указанные геномы имеют разметку для обучения моделей под 7 задач: предсказание промотора, предсказание основного промотора, предсказание места сплайсинга, классификация вариантов SARS-CoV-2, предсказание эпигенетических меток и предсказание мест связывания транскрипционных факторов у человека и мыши. Вместе с тем анализ существующей литературы показал, что на текущий день отсутствуют заметные исследования о предсказании резистентности бактерий к антибиотикам с использованием энкодеров DNABERT или DNABERT-2.

Декодерные трансформерные модели для прогнозирования АМП по геномным данным

Существует относительно небольшое число исследований, посвященных использованию декодеров для прогнозирования резистентности микроорганизмов к АМП [19]. Вместе с тем, есть отдельные попытки применения известных декодерных моделей, таких как GPT или Llama [24], для целей предсказания устойчивости бактерий к АМП. В статье «Can Large Language Models Predict Antimicrobial Resistance Gene?» 2025 г. [25] авторы использовали три вида декодерных моделей, основанных на трансформерах, для предсказания резистентности: ChatGPT4o-mini, Llama 3.1 и Claude 3.5. Для работы с такими декодерами на вход им передавался промпт с указанием, что необходимо сделать модели: «Укажите ан-

тибиотик (из списка: сульфаниламиды, аминогликозиды, бета-лактамы, гликопептиды, тетрациклины, фениколы, фторхинолоны, MLS, множественная лекарственная устойчивость), устойчивость к которому обеспечивает данная ДНК-последовательность: ATGAATCCCTATC...ACAAACTGCGAGGCAGTTCGCATGA».

В итоге модели без дообучения под данную задачу показывали точность около 0%. На втором этапе эксперимента авторы указанного исследования для каждой ДНК-последовательности искали 5 наиболее похожих генов в базах данных генов резистентности к АМП [26] при помощи алгоритма BLASTn [27]. Схожие гены передавались в промпт как дополнительный контекст вместе с описанием их функций и механизма резистентности. Генеративная модель (ChatGPT-4o-mini, LLaMA, Claude 3.5) использовала эти аннотации для предсказания класса антибиотика, к которому имеется резистентность. После такого дополнения доля случаев, когда модель не смогла определить класс резистентности для данной ДНК-последовательности уменьшилась до 73%, 39% и 14% для LLaMA, Claude 3.5 и ChatGPT-4o-mini соответственно. На третьем этапе исследователи дообучили указанные модели на базах данных CARD и MEGARes, которые содержат в себе названия генов и их детальное описание. Технические подробности формирования набора данных с использованием CARD и MEGARes для обучения в публикации не приводятся. Однако доля случаев, когда дообученные модели не смогли определить класс резистентности для входной ДНК-последовательности, уменьшилась до 0% для LLaMA и ChatGPT-4o-mini, а также до 11% для Claude 3.5. Очевидно, что работу таких моделей легко заменить поиском гена, ответственного за механизм резистентности при помощи классического алгоритма BLASTn. Поэтому применение декодерных моделей, по крайней мере, в том виде, в котором описано в статье, кажется излишним, а практическая ценность такого способа определения резистентности остается сомнительной.

Среди прочих заметно выделяется исследование «Sequence modeling and design from molecular to genome scale with Evo» [28]. Данная публикация посвящена новой глубокой нейросетевой архитектуре Evo – фундаментальной языковой модели типа декодер, предназначенной для анализа и генерации геномных последовательностей. Evo была предварительно обучена на огромном массиве данных «OpenGenome», состоящем исключительно из геномов бактерий, архей, вирусов прокариот и плазмид. Общий объем использованных данных составил более 300 млрд нуклеотидов. Обучение Evo происходило в два этапа: сначала модель обучалась на 8 тыс. нуклеотидов, затем контекст был расширен до 131 тыс. Такая схема обучения позволила модели научиться понимать и запоминать биологические закономерности на различных уровнях сложности, от отдельных генов до целых геномов. Несмотря на то, что Evo является моделью типа декодер, она успешно применяется в задачах предсказания свойств геномных последовательностей. Например, модель эф-

фективно прогнозирует «fitness» (приспособленность) белков и некодирующих РНК, уровень экспрессии генов, эссенциальность генов, а также структурные свойства и локализацию белков. Оценка приспособленности осуществляется путем сравнения логарифмической вероятности (log-likelihood) мутантной последовательности с логарифмической вероятностью исходной последовательности, т.е. дикого типа. Существенное снижение логарифмической вероятности после мутации указывает на ее делетеризующий («вредный») эффект или снижение приспособленности с точки зрения модели. Это обусловлено отклонением мутантной последовательности от биологических паттернов, усвоенных моделью Evo в процессе обучения. Расчет логарифмической вероятности последовательности осуществляется путем последовательного предсказания каждой буквы в последовательности. Например, для последовательности «ATGT» и ее мутантного варианта «ATGC» (замена «Т» на «С»), модель, получая каждый нуклеотид, предсказывает вероятность появления каждого из четырех возможных нуклеотидов на следующей позиции. Сравнение логарифмических вероятностей (log-likelihood) исходной и мутантной последовательностей, вычисляемое как разность $\log P(\text{исходная}) - \log P(\text{мутантная})$, позволяет оценить относительную вероятность каждой последовательности. Значительное уменьшение логарифмической вероятности мутантной последовательности по сравнению с исходной указывает на ее несоответствие распространенным биологическим паттернам и свидетельствует о делетеризующем эффекте мутации или снижении приспособленности (fitness) согласно модели. Предсказание уровня экспрессии генов моделью Evo также основано на вычислении вероятности последовательностей, таких как промоторы и сайты связывания рибосом (RBS). Последовательности с низкой вероятностью (низким log-likelihood), как правило, коррелируют с низкой экспериментальной экспрессией, так как они нетипичны или плохо узнаваемы клеточными механизмами.

В 2025 г. появилась улучшенная версия модели – Evo2 [29], которая также является моделью-декодером и обучена предсказывать следующий нуклеотид один за одним на основе предыдущих. В целом ее отличие заключается в более мощной обобщающей способности модели. Evo2 была обучена на 9,3 трлн нуклеотидов из тщательно подготовленного геномного атласа, охватывающего: бактерии, археи, эукариоты и бактериофаги, имеет чрезвычайно большое контекстное окно – 1 млн нуклеотидов (в отличие от 131 тыс. в Evo). Авторы исследования отмечают, что обобщающая способность новой модели настолько велика, что она способна выполнять задачи (предсказывать экспрессию, эссенциальность и т.д.) без дополнительного обучения, в то время как прошлая версия модели (Evo) нуждалась в таком дообучении под конкретную задачу. Evo2 осуществляет предсказания аналогично Evo: вычисляет логарифмическую вероятность (log-likelihood) исходной последовательности, затем вычисляет логарифмическую вероятность последовательности с мутацией и,

сравнивая эти два значения, модель оценивает влияние мутации. В процессе обучения нейронная сеть формирует внутренние представления входных данных в виде высокоразмерных векторов (эмбедингов), отражающих различные биологические характеристики геномных последовательностей (например, гены, промоторы, интроны). В отличие от моделей типа «черный ящик», в данном исследовании авторы применяют метод разреженного автокодировщика (Sparse Autoencoder, SAE) для анализа этих внутренних представлений и интерпретации признаков, извлеченных моделью Evo2 из геномных данных.

SAE (Sparse Autoencoder) – это специальный тип нейронной сети, построенный по типу энкодера. То есть эта сеть сначала сжимает высокоразмерные вектора в небольшие по размеру вектора, затем из этих небольших векторов пытается восстановить исходные данные. Таким образом, в сжатых внутренних векторах автоэнкодер сохраняет именно ту информацию, которая является наиболее важной и достаточной для восстановления исходных данных, отбрасывая при этом избыточные и шумовые признаки. Но особенностью SAE является то, что в ее скрытом представлении создается не просто небольшой вектор, а так называемый «разряженный» вектор, который имеет такую же размерность как входные данные или больше, но заполненный в основном нулями. Исследователи обучили модель SAE на восстановление векторов, взятых из глубины Evo2, а именно, из 26-ого слоя сети. Такой вектор, кодирует не просто один нуклеотид, а нуклеотид в контексте всей предыдущей последовательности. То есть по замыслу авторов такой вектор из глубины сети несет в себе информацию не только о конкретном виде нуклеотида, но и о том в каком гене он находится и какова его функция в геноме в целом. При помощи таких векторов SAE обучилась обнулять не нужные в векторе числа и увеличивать, наоборот, те признаки в векторе, которые несут наиболее важную для этого участка информацию. В статье описано, что исследователи проанализировали, какие именно нейроны (из этих 32768) активируются при подаче различных известных биологических последовательностей, позволяя понять биологический смысл отдельных нейронов SAE. Оказалось, что при подаче векторов, выражающих одинаковые участки геномов у разных организмов, SAE активирует одни и те же 64 признака (нейрона) в векторе из 32768 признаков (нейронов). Используя такие разреженные вектора, авторы смогли идентифицировать ряд биологически значимых участков геномов. В частности, были найдены нейроны, которые избирательно активируются на следующих типах геномных элементов: профаговые области (интегрированные вирусные последовательности в бактериальных геномах); CRISPR-спейсеры (участки ДНК, участвующие в адаптивном «иммунитете» бактерий); границы экзонов и интронов (участки, разделяющие кодирующие и не кодирующие фрагменты генов); сайты связывания транскрипционных факторов (регуляторные элементы ДНК, контролирующие экспрессию генов); элементы вторичной струк-

туры белков, такие как α -спирали и β -листы; гены тРНК (tRNA) и рРНК (rRNA). Эти результаты свидетельствуют о том, что модель Evo2 неявно освоила ключевые биологические признаки геномных последовательностей, которые затем были явно интерпретированы с помощью подхода на основе SAE. Авторы исследования не рассматривают возможности применения описанной модели для прогнозирования антимикробной резистентности. По субъективному мнению авторов настоящего обзора, векторизация геномных последовательностей или анализ вероятностных различий между изолятами могут лечь в основу новых моделей для предсказания АМП с учетом семантики генетического текста.

Гибридные подходы на основе нейронных сетей для прогнозирования АМП бактерий

Преимущество трансформеров при работе с геномом заключается в том, что они оценивают эффект от взаимодействия генов друг с другом. Однако трансформеры – это не единственный метод машинного обучения, который способен делать предсказания на основе взаимосвязей объектов друг с другом. Один из таких способов продемонстрирован в статье «A convolutional neural network highlights mutations relevant to antimicrobial resistance in *Mycobacterium tuberculosis*» [30], авторы которой использовали геномы *M. tuberculosis* для предсказания резистентности к 13 АМП. При этом исследователи обучали модель MD-CNN (Multi-Drug Convolutional Neural Network) предсказывать наличие или отсутствие резистентности сразу ко всем 13 АМП, а также строили модели SD-CNN (Single-Drug CNN), предсказывающие резистентность к каждому АМП отдельно.

Поскольку полная геномная последовательность слишком большая для существующей архитектуры нейронной сети и содержит много «шумовых» генов, перед подачей в модель данные полного геномного секвенирования прошли многоэтапную фильтрацию и обработку. На первом этапе были обработаны сырые риды с использованием PRINSEQ [31] и Kraken [32], затем последовательности были выравнены на референсный геном *M. tuberculosis* H37Rv с помощью инструмента BWA-MEM [33]. Этот алгоритм выполнял точное позиционирование ридов относительно стандартного генома, что критически важно для дальнейшего анализа мутаций. Для обеспечения надежности данных были исключены изоляты, у которых менее 95% генома было покрыто при глубине секвенирования 10x. Из всех оставшихся геномов были определены 18 генов, которые были отобраны на основе их известной или предполагаемой связи с устойчивостью к АМП. Далее последовательности кодировались в формате One-Hot-Encoding, где каждый нуклеотид (A, T, G, C) представляется вектором размерности 5. Финальная матрица для одной геномной последовательности имела размерность 5x10291x18, где 5 – это размерность вектора, 10291 это максимальное количество нуклеотидов гена наибольшего размера (длины), а 18 – количество отобранных генов. Важное

условие формирования такой матрицы – порядок каждого нуклеотида строго привязан к его позиции в референсном геноме. Полученная матрица передавалась в свёрточную нейронную сеть, которая выявляет локальные закономерности в многомерных данных, позволяет эффективно анализировать пространственные взаимосвязи мутаций внутри и между генами. Свёрточная нейросеть обрабатывает входную матрицу, последовательно применяя к ней свёрточные фильтры с фиксированными размерами, аналогично тому, как свёрточные сети анализируют изображения, выявляя характерные черты, такие как края и текстуры. В данном случае свёрточные фильтры изучают закономерности внутри последовательностей генов, выявляя значимые мутации, которые могут быть связаны с резистентностью к АМП. Таким образом, MD-CNN выполняет многозадачное предсказание, одновременно оценивая вероятность резистентности к разным АМП. Аналогичным образом работает SD-CNN (Single-Drug Convolutional Neural Network), однако данная архитектура обучена предсказывать устойчивость только к одному АМП. Вместо многоклассового выхода, как у MD-CNN, SD-CNN выдает одно значение, которое представляет вероятность наличия или отсутствия устойчивости конкретного изолята к выбранному АМП. Исследователи отмечают, что в среднем модель, обученная на предсказание резистентности сразу к нескольким АМП, показывает метрики лучше, чем отдельные модели, обученные предсказывать резистентность к конкретному антибиотику. Таким образом, локальные взаимодействия генов друг с другом могут эффективно улавливаться свёрточными нейронными сетями и использоваться для предсказания резистентности.

Продолжением исследования перспектив применения свёрточных нейронных сетей к геномам бактерий для предсказания их резистентности является публикация «Sequence-based modelling of bacterial genomes enables accurate antibiotic resistance prediction» [34], в которой исследователи описывают графовую нейронную сеть, учитывающую взаимодействие всех генов и нуклеотидов, а также в данном исследовании авторы перешли к задаче прогнозирования значения минимальной подавляющей концентрации (МПК), т.е. с точки зрения машинного обучения – к задаче регрессии. В данной публикации использовались геномы *M. tuberculosis* и *Pseudomonas aeruginosa*. Для каждого изолята было известно значение МПК конкретного АМП, которое использовалось в качестве целевой метки прогнозирования в процессе обучения модели. При отборе генов исследователи опирались на анализ литературы, где были собраны ранее известные гены, ассоциированные с АМП, а также на анализ вариабельности генов (рассчитывалось число мутаций на единицу длины гена и выбирались самые изменчивые гены). Последовательности нуклеотидов каждого гена были усечены до максимальной длины 2560 оснований. В дальнейшем применялось кодирование по методу One-Hot-Encoding, преобразующее каждый нуклеотид в четырехмерный вектор. В результате каждый ген был представлен матрицей разме-

ром 2560×4 , а матрицы генов короче 2560 оснований дополнялись нулями до указанного размера. Эти матрицы последовательно обрабатывались свёрточной нейронной сетью. При первой итерации учитывались нуклеотиды с 1 по 17, на второй итерации со 2 по 18 нуклеотид, затем с 3 по 19 и т.д. Далее свёрточная нейронная сеть преобразовывала все полученные последовательности нуклеотидов в один вектор размерности 64. Данный вектор отражает абстрактные признаки поданного на вход сети гена. Полученный вектор передавался во вторую нейронную сеть – графовую, нацеленную на работу с данными, которые помимо собственного значения также обладают связью друг с другом, которую можно выразить численно. В данной работе взаимосвязи между генами моделировались с использованием базы данных STRING, содержащей информацию о белок-белковых взаимодействиях. Каждый ген выступал в качестве узла в графе, ребра графа отражали силу взаимодействия между генами и взвешивались значениями из базы данных STRING (от 0 – отсутствие связи до 1 – максимальная достоверная связь). Обученная графовая нейронная сеть использовалась для предсказания значения МПК в отношении изучаемых видов бактерий. Оценка качества модели, выраженная коэффициентом детерминации (R^2), составила 0,66, т.е. модель объясняла 66% вариации в значениях МПК. Коэффициент корреляции Пирсона между предсказанными и истинными значениями МПК составил 0,87, что указывало на высокую степень линейной связи между ними.

В работе «Predicting antibiotic resistance genes and bacterial phenotypes based on protein language models» [35] предлагается двухуровневая система предсказания антибиотикорезистентности, которая работает не с нуклеотидными последовательностями, а с аминокислотными последовательностями и использует большие «белковые языковые модели» ProtBert-BFD и ESM-1b. В качестве обучающей выборки используются аминокислотные последовательности из DeepARG и HMD-ARG, дополненные 2000-ми белков, не связанных с механизмами резистентности, из базы данных HyperVR [36]. Модели ProtBert-BFD и ESM-1b являются энкодерами, которые превращают последовательность аминокислот, обозначенных буквами в последовательность векторов чисел. Далее эти векторы от каждой модели независимо передаются в разные рекуррентные сети, затем в линейные слои нейронной сети, которые и делают финальное предсказание. Каждое предсказание это вектор из 16 чисел, где наибольшее число будет означать отнесение рассматриваемого белка к одному из 15 типов резистентности (которые определены авторами подхода) либо отнесение белка к 16 классу (отсутствие резистентности). Следует отметить, что авторы для целей предсказания использовали не всю аминокислотную последовательность белка, а только первые 1024 аминокислоты. Сформированные каждой моделью предсказания затем умножаются на весовой коэффициент a и b соответственно, а затем суммируются. Весовые коэффициенты в начале обучения равны 0,5 и

изменяются в процессе обучения модели, чтобы определить, какой вклад в финальное предсказание должна вносить каждая модель. Авторы отмечают, что предсказание резистентности только на основе результата работы модели имеет высокую метрику ложноположительных срабатываний. Для того, чтобы снизить количество ложноположительных срабатываний, авторы статьи предложили использовать инструмент CARD (Comprehensive Antibiotic Resistance Database) – это специализированная база данных и набор инструментов по антибиотикорезистентности. Она содержит каталоги с генами и белками AMP [37]. Используя CARD авторы статьи проводят этап скрининга потенциально всех белков изолята, отбирая только те, которые алгоритм отнес к связанным с AMP, а затем оценивают оставшиеся белки своей моделью, предсказывая тип резистентности. В итоге метрика точности такого алгоритма превысила 90%. Авторы отмечают, что эта метрика выше, чем аналогичная метрика, полученная при использовании только CARD.

Заключение

В данном обзоре проанализированы современные методы машинного обучения для предсказания AMP. Рассмотренные подходы различаются по типу используемых данных (клинические или геномные), предсказываемым величинам (разделение на устойчивые и чувствительные изоляты, либо регрессионное предсказание МПК), и классам – от классических алгоритмов до сложных нейросетевых архитектур. Анализ показывает, что высокое качество прогноза часто достигается за счет предварительного отбора генетических признаков, на

пример, известных генов резистентности или участков с высокой мутационной нагрузкой. Однако существенным фактором, влияющим на качество предсказаний, остается этап отбора релевантных генов. Перспективным направлением представляется применение декодерных моделей для генерации векторных представлений нуклеотидных последовательностей, отдельных генов или целых геномов. Эти представления могут служить входными данными для последующего обучения моделей машинного обучения, решающих задачи бинарной классификации (чувствительность/резистентность) или регрессионного прогнозирования МПК. Поскольку многие изоляты бактерий формируются в различных клинических условиях, предобученные модели могут демонстрировать нестабильные результаты при анализе новых выборок. Это ставит под сомнение универсальность существующих решений и указывает на необходимость дальнейшего исследования способов учета ранее неизвестных мутаций. В перспективе дальнейшее развитие методов машинного обучения в области предсказания AMP представляется возможным не столько за счет усложнения архитектур нейросетей, сколько за счет разработки новых стратегий обработки геномных данных. В частности, актуальной задачей остается создание подходов, позволяющих учитывать не только уже известные мутации/гены, но и выявлять новые варианты изменений, оказывающих влияние на устойчивость к АМП, а также учитывать генетический контекст. Это требует усовершенствования методов сбора данных, оптимизации отбора признаков и разработки адаптивных, непрерывно обучающихся моделей, способных учитывать динамические изменения и контекст генома бактерий на уровне конкретной медицинской организации.

Литература

1. Mestrovic T. The burden of bacterial antimicrobial resistance in the WHO European Region in 2019: a cross-country systematic analysis. *Lancet Public Health*. 2022;7(11):e897-e913. DOI: 10.1016/S2468-2667(22)00225-0
2. Peiffer-Smadja N., Rawson T.M., Ahmad R., Buchard A., Georgiou P., Lescure F.-X., et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin Microbiol Infect*. 2020;26(5):584-595. DOI: 10.1016/j.cmi.2019.09.009
3. Oonsivilai M., Mo Y., Luangsanatip N., Lubell Y., Miliya T., Tan P., et al. Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia. *Wellcome Open Res*. 2018;3:131. DOI: 10.12688/wellcomeopenres.14847.1
4. Tang R., Luo R., Tang S., Song H., Chen X. Machine learning in predicting antimicrobial resistance: a systematic review and meta-analysis. *Int J Antimicrob Agents*. 2022;60(5-6):106684. DOI: 10.1016/j.ijantimicag.2022.106684
5. Luz C.F., Vollmer M., Decruyenaere J., Nijsten M.W., Glasner C., Sinha B. Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies. *Clin Microbiol Infect*. 2020;26(10):1291-1299. DOI: 10.1016/j.cmi.2020.02.003
6. Rajpurkar P., Chen E., Banerjee O., Topol E.J. AI in health and medicine. *Nat Med*. 2022;28(1):31-38. DOI: 10.1038/s41591-021-01614-0
7. Goodman K.E., Lessler J., Cosgrove S.E., Harris A.D., Lautenbach E., Han J.H., et al. A clinical decision tree to predict whether a bacteremic patient is infected with an

- extended-spectrum β -lactamase-producing organism. Clin Infect Dis. 2016;63(7):896-903. DOI: 10.1093/cid/ciw425
8. Feretzakis G., Loupelis E., Sakagianni A., Kalles D., Lada M., Christopoulos C., et al. Using machine learning algorithms to predict antimicrobial resistance and assist empirical treatment. Stud Health Technol Inform. 2020;272:75-78. DOI: 10.3233/SHTI200497
 9. Kollef M.H., Shorr A.F., Bassetti M., Timsit J.-F., Micek S.T., Michelson A.P., et al. Timing of antibiotic therapy in the ICU. Crit Care. 2021;25(1):360. DOI: 10.1186/s13054-021-03787-z
 10. McGuire R.J., Yu S.C., Payne P.R.O., Lai A.M., Vazquez-Guillamet M.C., Kollef M.H., et al. A pragmatic machine learning model to predict carbapenem resistance. Antimicrob Agents Chemother. 2021;65(7):e00063-21. DOI: 10.1128/AAC.00063-21
 11. Henderson H.I., Napravnik S., Kosorok M.R., Gower E.W., Kinlaw A.C., Aiello A.E., et al. Predicting risk of multidrug-resistant Enterobacterales infections among people with HIV. Open Forum Infect Dis. 2022;9(10):ofac487. DOI: 10.1093/ofid/ofac487
 12. Hebert C., Gao Y., Rahman P., Dewart C., Lustberg M., Pancholi P., et al. Prediction of antibiotic susceptibility for urinary tract infection in a hospital setting. Antimicrob Agents Chemother. 2020;64(7):e02236-19. DOI: 10.1128/AAC.02236-19
 13. Zhang Y., Liu C., Liu M., Liu T., Lin H., Huang C.-B., et al. Attention is all you need: utilizing attention in AI-enabled drug discovery. Brief Bioinform. 2023;25(1):bbad467. DOI: 10.1093/bib/bbad467
 14. Nerella S., Bandyopadhyay S., Zhang J., Contreras M., Siegel S., Bumin A., et al. Transformers and large language models in healthcare: a review. Artif Intell Med. 2024;154:102900. DOI: 10.1016/j.artmed.2024.102900
 15. Giske C.G., Bressan M., Fiechter F., Hinic V., Mancini S., Nolte O., et al. GPT-4-based AI agents—the new expert system for detection of antimicrobial resistance mechanisms? J Clin Microbiol. 2024;62(11):e00689-24. DOI: 10.1128/jcm.00689-24
 16. European Committee on Antimicrobial Susceptibility Testing. Available at: www.eucast.org. Accessed October 25, 2025.
 17. Ren Y., Chakraborty T., Doijad S., Falgenhauer L., Falgenhauer J., Goesmann A., et al. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. Bioinformatics. 2022;38(2):325-334. DOI: 10.1093/bioinformatics/btab681
 18. Quang D., Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44(11):e107-e107. DOI: 10.1093/nar/gkw226
 19. Li Y., Cui X., Yang X., Liu G., Zhang J. Artificial intelligence in predicting pathogenic microorganisms' antimicrobial resistance: challenges, progress, and prospects. Front Cell Infect Microbiol. 2024;14:1482186. DOI: 10.3389/fcimb.2024.1482186
 20. Ji Y., Zhou Z., Liu H., Davuluri R.V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021;37(15):2112-2120. DOI: 10.1093/bioinformatics/btab083
 21. Devlin J., Chang M., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, USA; 2019:4171-4186.
 22. Zhou Z., Ji Y., Li W., Dutta P., Davuluri R., Liu H. DNABERT-2: efficient foundation model and benchmark for multi-species genome. Available at: <https://arxiv.org/abs/2306.15006>. Accessed October 29, 2025. DOI: 10.48550/arXiv.2306.15006
 23. Genome Understanding Evaluation. Available at: <https://paperswithcode.com/dataset/gue>. Accessed October 28, 2025.
 24. Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.-A., Lacroix T., et al. LLaMA: open and efficient foundation language models. Available at: <https://arxiv.org/abs/2302.13971>. Accessed October 28, 2025. DOI: 10.48550/arXiv.2302.13971
 25. Yoo H. Can large language models predict antimicrobial resistance gene? Available at: <https://arxiv.org/abs/2503.04413>. Accessed October 28, 2025. DOI: 10.48550/arXiv.2503.04413
 26. Lakin S.M., Dean C., Noyes N.R., Dettenwanger A., Ross A.S., Doster E., et al. MEGARes: an antimicrobial resistance database for high throughput sequencing. Nucleic Acids Res. 2017;45(D1):D574-D580. DOI: 10.1093/nar/gkw1009
 27. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-410. DOI: 10.1016/S0022-2836(05)80360-2
 28. Nguyen E., Poli M., Durrant M.G., Kang B., Katrekar D., Li D., et al. Sequence modeling and design from molecular to genome scale with Evo. Science. 2024;386(6723):ead09336. DOI: 10.1126/science.ado9336
 29. Evo 2: genome modeling and design across all domains of life. Available at: <https://github.com/Arclnstitute/evo2>. Accessed October 28, 2025.
 30. Green A.G., Yoon C.H., Chen M.L., Ektefaie Y., Fina M., Freschi L., et al. A convolutional neural network highlights mutations relevant to antimicrobial resistance in *Mycobacterium tuberculosis*. Nat Commun. 2022;13(1):3817. DOI: 10.1038/s41467-022-31236-0
 31. Schmieder R., Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863-864. DOI: 10.1093/bioinformatics/btr026

32. Wood D.E., Salzberg S.L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 2014;15(3):R46. DOI: 10.1186/gb-2014-15-3-r46
33. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: <https://arxiv.org/abs/1303.3997>. Accessed October 28, 2025. DOI: 10.48550/arXiv.1303.3997
34. Wiatrak M., Weimann A., Dinan A., Brbić M., Floto R.A. Sequence-based modeling of bacterial genomes enables accurate antibiotic resistance prediction. Available at: www.biorxiv.org/content/10.1101/2024.01.03.574022v1. Accessed October 29, 2025. DOI: 10.1101/2024.01.03.574022
35. Wang B., Meng R., Li Z., Hu M., Wang X., Zhao Y., et al. Predicting antibiotic resistance genes and bacterial phenotypes based on protein language models. *Front Microbiol.* 2025;16:1628952. DOI: 10.3389/fmicb.2025.1628952
36. Ji B., Pi W., Zhang X., Peng S. HyperVR – a hybrid prediction framework for virulence factors and antibiotic resistance genes in microbial data. Available at: www.biorxiv.org/content/10.1101/2022.05.24.493218v1. Accessed October 29, 2025. DOI: 10.1101/2022.05.24.493218
37. Alcock B.P., Huynh W., Chalil R., Smith K.W., Raphe-nya A.R., Wlodarski M.A., et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* 2023;51(D1):D690-D699. DOI: 10.1093/nar/gkac920