

ЦИФРОВАЯ ТРАНСФОРМАЦИЯ

От случайной находки до нейросети: как искусственный интеллект меняет фармацевтику

До середины XX века новые лекарственные средства находили во многом случайно. Знаменитый пенициллин — классический пример: Александр Флеминг перед отпуском забыл помыть чашки Петри с бактериями, а вернувшись, увидел, что в одной из чашек выросла плесень, вокруг которой не было бактерий. Только с середины XX века ученые начали действовать системно, использовать для поиска кандидатов в лекарственные средства животные и клеточные модели. Сейчас начинается новый этап — этап цифровой трансформации фармацевтики.

Эксперименты на животных: первый шаг к порядку

Первые животные модели для поиска и тестирования лекарственных средств (ЛС) появились в середине XX века — фармакологи начали проверять гипотезы до того, как применять препараты на человеке.

Бум пришелся на 1970-е годы. Тогда создавали животные модели чуть ли не на конвейере: диабет, гипертония, рак — для каждой нозологии искали «зверьку», на котором можно было изучать пользу соединений. И практически сразу случился всплеск новых лекарств. Казалось, что вот он — правильный путь. Но у метода оказались минусы. Животное нужно вырастить, за ним нужно ухаживать, эксперименты длятся месяцами. Это сложно, медленно и очень дорого. И самое неприятное: то, что работает на животных, на человеке часто оказывается бесполезным.

В 1980–1990-х годах подросли новые технологии. Клеточные культуры, молекулярная биология, биохимические методы — теперь можно было тестировать вещества на клеточных культурах и просто в пробирке проверять их взаимодействие с целевыми белками. Скрининг потенциальных лекарственных средств стал гораздо быстрее и эффективнее. Вместо того чтобы месяцами ждать, умер ли животное, можно было измерить параметры взаимодействия за несколько часов. И снова случился всплеск — выросло количество новых лекарств, используемых в клинике.

А затем на сцену вышли компьютеры. Рост вычислительных мощностей, понимание структурных формул, накопление баз данных — всё это привело к рождению целого направления: компьютерного конструирования лекарств. Именно об этом и пойдет речь дальше.

Горькая правда: 15 лет и миллиарды долларов

Как выглядит реальный процесс разработки? Прежде чем мы окунемся в высокие технологии, давайте поймем масштаб проблемы. Разработка одного лекарства занимает от 10 до 15 лет. Это не опечатка. Целых полтора десятилетия от первой идеи до появления на прилавке.

А сколько это стоит? Огромные деньги — миллиарды долларов. Не каждая фармацевтическая компания может себе такое позволить.

Что входит в этот процесс? Вот только основные этапы:

- 1) поиск идеи и лекарственной мишени — надо определить, какой белок или процесс блокировать;
- 2) поиск веществ — нужно найти соединения, которые будут действовать на мишень;
- 3) оценка эффективности — необходимо убедиться, что вещество работает;
- 4) оценка побочных эффектов — а не убьет ли оно пациента заодно с болезнью;
- 5) доклинические исследования — всё еще на животных, но уже по строгим протоколам;
- 6) клинические исследования — три фазы на добровольцах и пациентах;
- 7) регистрация — подача документов в регуляторы (это еще год–полтора чистой бюрократии).

И это только верхушка айсберга. Считается, что на пути от идеи до аптеки нужно провести около 800 экспериментальных исследований.

Сколько новых лекарств выходит в мире каждый год? Всего около 100 новых соединений. Для целой планеты — это капля в море. А сколько соединений нужно перебрать, чтобы создать один препа-

рат? Примерно 10 тысяч. Десять тысяч кандидатов, из которых до финала дойдет только один.

И даже когда всё сделано правильно, шансы на успех — невысоки. Всего 10–15 % проектов, которые были начаты, доходят до финального хорошего результата. Максимум — 20 %. То есть из десяти перспективных идей в аптеку попадет максимум две.

Почему проекты проваливаются?

Даже если всё идет по плану, есть три главных «убийцы» лекарственных проектов.

Первое и самое частое — проблемы с эффективностью. Вы провели доклинические исследования на животных, всё блестяще. Начинаете клинические исследования на людях, но эффект — ноль. Почему? Человек отличается от животного. Метаболизм другой, пути передачи сигналов в клетках другие. Вещество может не дойти до нужной мишени или дойти, но у человека есть «обходные пути», которые сводят блокировку на нет.

Второе — токсичность. Лекарство может быть очень эффективным, но если оно заодно разрушает печень или почки — такой препарат никому не нужен. Никто не хочет лечить одну болезнь, получая в нагрузку повреждения внутренних органов.

Третье — биодоступность. Вещество показало себя отлично в пробирке и даже на животных. Но в организме человека оно не всасывается, слишком быстро выводится или не может добраться до мишени. Опять же — проблема экстраполяции от животного к человеку.

И четвертое, о чем редко говорят, — коммерческие причины. Да, даже работающее и безопасное лекарство могут закрыть. Почему? Представьте, что пять компаний одновременно разрабатывают препарат против одного и того же заболевания. Первая, выходящая на рынок, забирает все «сливки» — самые высокие цены, самые большие продажи. Остальные, зачастую понимая, что догнать лидера у них нет возможности, а денег потрачено уже очень много, проект закрывают.

У больших транснациональных компаний на полках лежат десятки таких замороженных исследований. Их не выбрасывают окончательно — вдруг через 20 лет откроются новые свойства молекулы или появится другое заболевание, при котором она будет полезна. И такое действительно бывает, но нечасто.

Мишень-центричная парадигма: главная идея современной фармакологии

Сейчас в мире доминирует подход, который называется «мишень-центричный». Звучит сложно, но на деле всё просто. Есть мишень — обычно это белок, который отвечает за развитие болезни. Есть лекарственное вещество (в науке его называют «лиганд») — химическое соединение, которое должно на эту мишень подействовать. Заблокировали мишень — вылечили болезнь.

Этот подход сформировался еще в XX веке и до сих пор остается основным. Хотя, конечно, есть нюансы: одно вещество может действовать на несколько мишеней сразу, и иногда это даже полезно — например, при аллергии или некоторых неврологических заболеваниях.

Когда в начале 2000-х годов закончился проект «Геном человека», ученые задали себе важный вопрос: «А на сколько мишеней вообще действуют современные лекарства?» Ответ оказался скромным: около 500 белков. С учетом бактерий, вирусов и грибов — но всё равно не так много.

Тогда авторы одного известного исследования сделали прогноз: через 10–20 лет количество мишеней вырастет до 5–10 тысяч белков. Прошло 20 лет. Сбылось? Сейчас в исследованиях разных компаний фигурирует чуть меньше 10 тысяч мишеней. Одна из

аналитических компаний ведет базу данных о том, на какие мишени действуют исследуемые соединения. Там указано, что более шести тысяч мишеней уже либо используются в клинике, либо находятся на разных стадиях клинических исследований. Конечно, это не значит, что на каждую из этих мишеней получится создать лекарство. Но сама цифра внушает оптимизм: медицина будущего будет иметь гораздо больше инструментов, чем сейчас.

Почему же фармацевты всё чаще смотрят в сторону ИИ и машинного обучения?

Ответ прост: мы столкнулись с объемами данных, которые не может переварить ни один человек. Посмотрим на биологическую сторону вопроса. У человека всего 23 хромосомы, которые кодируют около 20 тысяч белков. Но это только верхушка айсберга. Есть изоформы (разные версии одного белка), есть мутации, есть посттрансляционные модификации. С учетом всего этого разнообразия белковых форм у человека можно оценить в два миллиона. А еще белки взаимодействуют друг с другом — это создает сотни тысяч дополнительных объектов, которые нужно анализировать.

А теперь посмотрим на химию. За всё время исследований в мире использовано около 10 тысяч уникальных лекарственных соединений. Но если посмотреть, сколько из них реально применяется в России или Америке, то получится от полутора до двух тысяч уникальных соединений на страну (списки немного различаются).

А сколько соединений вообще было синтезировано и хоть как-то испытано? Около полутора-двух миллионов.

А сколько соединений синтезировано в принципе, даже если их не тестировали? Больше 150 миллионов.

А сколько можно теоретически сгенерировать, если соединять атомы разными способами? Здесь наука заходит в область генеративной химии — создания новых структур, которых еще никто не синтезировал. Оценки говорят о 10^{60} возможных соединениях. Это больше, чем атомов в Солнечной системе. Хранить такие данные невозможно, не то что тестировать. Именно поэтому нужны модели предсказания. Они позволяют сфокусироваться на маленькой области этого гигантского химического пространства и не перебирать всё подряд наугад.

Компьютерное конструирование лекарств

Существует два основных направления в компьютерном конструировании лекарств.

Первый подход — на основе мишени (target-based drug design). Для него нужно знать трехмерную структуру белка. Она может быть получена экспериментально (например, методом рентгеноструктурного анализа) или смоделирована на компьютере. И здесь в дело вступает искусственный интеллект, который помогает предсказывать 3D-структуру белков по их аминокислотной последовательности.

Что происходит дальше? Компьютер моделирует взаимодействие между молекулой лекарства и белком. Учитываются физико-химические свойства, эффекты взаимодействия на уровне атомов, рассчитывается энергия связывания. Это называется «докинг» — от английского docking, что значит «причаливание корабля к пирсу». Молекула «причаливает» к белку, и программа считает, насколько выгодно такое объединение. Это сложная вычислительная задача, но технология уже отработана и широко используется. Единственный минус — много ложных положительных результатов. Программа может сказать, что вещество подходит, а в реальном эксперименте — нет. Но комбинация разных подходов всё равно эффективнее, чем случайный перебор.

ЦИФРОВАЯ ТРАНСФОРМАЦИЯ

Второй подход — на основе структур уже известных веществ (ligand-based drug design). Он хорош тем, что для многих белков уже накоплены большие базы данных о том, какие соединения на них действуют, а какие — нет. Имея эту информацию, можно попытаться создать математическую модель, которая связывает особенности химической структуры с биологической активностью. Здесь как раз очень широко используются методы машинного обучения. Вы берете тысячи молекул с известной активностью, «скармливаете» их компьютеру — и он учится предсказывать, будет ли новая, еще не изученная молекула работать или нет.

Насколько это эффективно? Давайте сравним. Если вы тестируете соединения случайным образом, ваш успех (вероятность найти активное вещество) составит около 0,1 %. То есть на тысячу проверенных соединений — одна находка. Компьютерные методы повышают выход примерно в 200 раз — до 20 %. Десять проверенных соединений — и два из них активны. Это колоссальный прогресс.

Революция AlphaFold: 3D-структуру белка можно теперь получить за минуты

Как раньше определяли структуру белков? Для подхода, основанного на данных о мишени, самое важное — трехмерная структура белка. Экспериментальные методы определения структуры были разработаны еще в 70-х годах прошлого века. За 50 лет работы ученым по всему миру удалось определить структуру около 200 тысяч белковых комплексов (самих белков меньше, потому что многие изучали многократно). Но белков в природе — сотни миллионов. И для большинства из них структура неизвестна.

И тут на сцену выходит компания из Google (точнее, ее подразделение — DeepMind) с программой AlphaFold. Это метод на основе искусственного интеллекта, который по одной только аминокислотной последовательности белка предсказывает, как этот белок сворачивается в пространстве. Вы просто подаете на вход строку букв (последовательность аминокислот) — и получаете на выходе трехмерную модель: где какой атом находится, какие участки белка с чем взаимодействуют. И это занимает минуты, а иногда и секунды.

Каждые несколько лет проводятся международные соревнования по предсказанию структуры белков. AlphaFold выигрывает последние несколько раз подряд. Никто пока не смог обойти эту программу, хотя есть альтернативные разработки — например, ESMFold от компании Meta (деятельность в России признана экстремистской и запрещена).

Метод докинга — не просто теория. У него есть успешные практические примеры. Самый яркий — создание препаратов против протеазы ВИЧ. Именно с помощью компьютерного скрининга и докинга были найдены первые активные соединения, которые потом доработали до реальных лекарств. Так что это не фантастика, а работающий инструмент.

QSAR и машинное обучение: как научить компьютер видеть связь «структура — активность»

Подход на основе структур известных соединений — называется QSAR. Аббревиатура расшифровывается как «количественный анализ связи «структура — активность»» (от английского Quantitative Structure-Activity Relationship).

Алгоритм выглядит так:

- 1) сбор данных: получение коллекции структур соединений и их экспериментальной активности;
- 2) оцифровка структурных формул с помощью специальных математических описаний — дескрипторов (например, молекулярный вес, количество атомов определенного типа, форма молекулы, заряды и т. д.);
- 3) передача таблицы с дескрипторами на вход алгоритмам машинного обучения. Их существует больше десятка — разные методы регрессии, нейронные сети, «деревья решений»;
- 4) построение алгоритмом зависимости между числовым описанием структуры и экспериментальным эффектом;

5) проверка модели на независимых выборках — соединениях, которые не участвовали в обучении.

Если точность хорошая, модель готова к использованию. Ей можно «скормить» новую молекулу и предсказать, будет ли она активной.

Стоит сделать важное уточнение. Методы машинного обучения, которые используются в QSAR, появились довольно давно — еще в XX веке. Это и статистические методы, и нейронные сети первых поколений. Сегодня всё это входит в большой зонтичный термин «искусственный интеллект». Сюда же относятся и большие языковые модели (вроде ChatGPT), и генеративные сети. Мы все — свидетели огромного прогресса в этой области. И всё это потихоньку заходит в фармацевтику.

Как искали лекарства от COVID-19 компьютерными методами

Давайте пройдем по алгоритму, который запускается, когда появляется новый вирус.

Шаг первый: секвенирование генома. Как только вирус найден, ученые расшифровывают его генетический код. С помощью методов биоинформатики можно сразу определить, какие белки кодируются в этом геноме, даже не выделяя их физически, а просто анализируя последовательность.

Шаг второй: поиск аналогов. Посмотрели на последовательность генома — на какие известные вирусы это похоже? А вдруг для похожего вируса уже есть лекарство? Если да — можно сразу попробовать его использовать.

Шаг третий: поиск похожих белков. Даже если вирус новый, может оказаться, что какой-то его белок похож на белок другого вируса или даже на человеческий белок. А на тот белок уже есть ингибитор — лекарство или экспериментальное соединение. Тоже хороший вариант.

Шаг четвертый: докинг. Если ничего не подошло, нужно браться за моделирование. Для COVID-19 не стали экономить деньги — белки вируса кристаллизовали экспериментально за несколько недель, получили трехмерные структуры и запустили докинг.

Почему это важно? Белок — это не гладкий шар. У него есть полости, карманы, выпуклости. Активный центр, который нужно заблокировать, находится в определенном месте. Зная структуру, можно целенаправленно искать вещества, которые подойдут именно в этот карман.

Когда стало понятно, что ни одно из существующих лекарств на COVID-19 не действует, возникла новая проблема: а где вообще брать соединения для поиска? Синтезированных молекул — больше 100 миллионов, но и этого показалось мало. Исследователи обратились к генеративному искусственному интеллекту. Генерация — это конструирование новых объектов на основе обученной модели. В данном случае — новых химических структур, причем с заданными свойствами.

Главное требование — должна быть возможность синтезировать соединение. Бесплезно генерировать молекулы, которые нельзя получить в реальности. Также важны растворимость в воде (чтобы можно было тестировать) и возможность принимать в виде таблетки (пероральная доступность).

Во время эпидемии COVID-19 была запущена глобальная инициатива. Сгенерировали набор данных из миллиарда химических структур с помощью ИИ.

Задача была поставлена перед научным сообществом всего мира: проанализировать этот миллиард молекул тремя разными компьютерными методами (докинг, QSAR-модели и еще один подход на основе сходства между молекулами). Нужно было отобрать 10 тысяч самых перспективных структур, которые потенциально могли бы действовать на разные мишени вируса.

В проекте участвовали 130 команд из разных стран. Они прислали в сумме более 600 тысяч структур-кандидатов, каждый внес свой вклад порцию потенциально активных соединений.

Организаторы (эксперты-химики) отобрали из этого потока 820 соединений, которые разными методами предсказывались как активные. Все 820 были синтезированы и протестированы в реальных экс-

периментах. Результат: 28 соединений оказались действительно активными (порог активности — 20 микромоляр). Это 3 % успеха.

На первый взгляд — немного. Но вспомните: без компьютерных методов проанализировать миллиард структур было бы просто невозможно. А 3 % — это уже рабочий результат, который можно использовать для дальнейшей оптимизации.

Российский след: одна из первых в мире — программа PASS и ее возможности

Автор статьи участвовал в создании программы PASS (Prediction of Activity Spectra for Substances). Это одна из первых программ в мире (и первая в нашей стране), которая по структурной формуле соединения предсказывает спектр его биологической активности — то есть не только, будет ли вещество эффективным при каком-то заболевании, но и на какие белки оно подействует, какие побочные эффекты даст, не окажется ли токсичным. Позже появились и другие разработки, но программа PASS остается актуальной и в настоящее время. С развитием интернета программа стала доступной онлайн — ее ежегодно используют несколько тысяч ученых со всего мира. На данный момент есть несколько тысяч научных публикаций, в которых продемонстрировано, что прогнозы программы PASS подтвердились экспериментально.

PASS анализирует структуру соединения и формирует список предсказанных активностей — и положительных (от каких болезней поможет), и отрицательных (токсичность, побочные эффекты). Для каждой активности указывается вероятность подтверждения в эксперименте. На основе технологий, заложенных в PASS, создан целый набор веб-приложений для разных задач: предсказание метаболизма, предсказание цитотоксичности, действие на экспрессию генов и многое другое.

Есть такое понятие — «репозиционирование лекарств». Это когда препарат, разработанный для одной болезни, неожиданно оказывается эффективным при другом заболевании. И программа PASS помогла найти несколько таких случаев.

Вот некоторые конкретные примеры:

- одно вещество разрабатывалось для лечения гипертонии. А оказалось, что оно еще и положительно действует на мозг (ноотропный эффект);
- один антибиотик неожиданно показал эффективность при воспалительных заболеваниях кишечника.
- коллеги из Индии с помощью PASS подтвердили, что лекарственное растение, которое традиционно использовали при кожных заболеваниях, обладает еще и антидепрессантным действием.

От случайности к инженерии

Что мы имеем в итоге? Раньше создание лекарств было похоже на лотерею. Случайные находки, «метод тыка», миллионы потраченных долларов и десятилетия работы — без гарантии успеха. Сегодня эта область становится точной инженерной дисциплиной. Мы знаем структуры белков, умеем моделировать взаимодействия, предсказывать активность по химической формуле. Искусственный интеллект позволяет перебирать миллиарды вариантов там, где раньше мы едва справлялись с тысячами.

Конечно, проблема полностью не решена. Клинические исследования на людях — самая дорогая и долгая часть — пока не поддаются автоматизации. Но поиск активных соединений стал в 200 раз эффективнее. А от качества первых стадий во многом зависит успех всего проекта.

Поиск ЛС зависит от слаженной работы разных специалистов — фармакологов, медицинских химиков, биохимиков, биоинформатиков, фармацевтов, программистов. Никто не справится в одиночку. Но когда они работают вместе и к ним на помощь приходит искусственный интеллект, рождаются новые лекарства, которые через 10–15 лет (или быстрее, если ИИ продолжит ускоряться) окажутся в аптеках и спасут чьи-то жизни.

Автор: Алексей Александрович Лагунин, д.б.н., профессор РАН, заведующий кафедрой биоинформатики Института биомедицины (ИБФ)