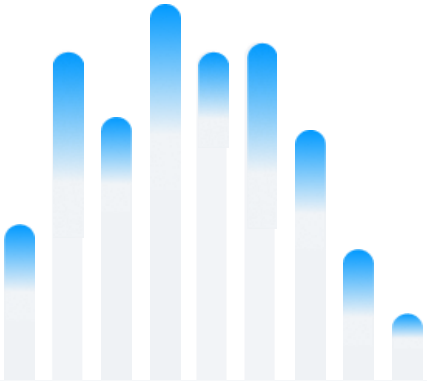
#наука и образование

Обработка и структурирование большого количества химической информации из разнородных источников с помощью ИИ

Федоров Максим Валериевич

Доктор химических наук,
член-корреспондент, и. о. директора
Института проблем передачи информации
им. А.А. Харкевича Российской академии наук

admin@syntelly.com



https://syntelly.ru

#проблема

Химическое пространство огромно

10^{60}
Расчетное количество синтетически доступных потенциальных кандидатов в биоактивные соединения

ЭКСПЕРИМЕНТАЛЬНЫЕ МЕТОДЫ НЕ СПРАВЛЯЮТСЯ С БОЛЬШИМ КОЛИЧЕСТВОМ НОВЫХ МОЛЕКУЛ

Человечеству известно уже более 250 миллионов молекул и это число экспоненциально растет с каждым днем

Поиск молекул вручную - неэффективен

Даже при наличии базы данных, на один запрос уходит до 90 минут. У исследователя таких запросов может быть до 200 в день. Почти все рабочее время тратится на поиск информации.

Экспертам нужны современные инструменты

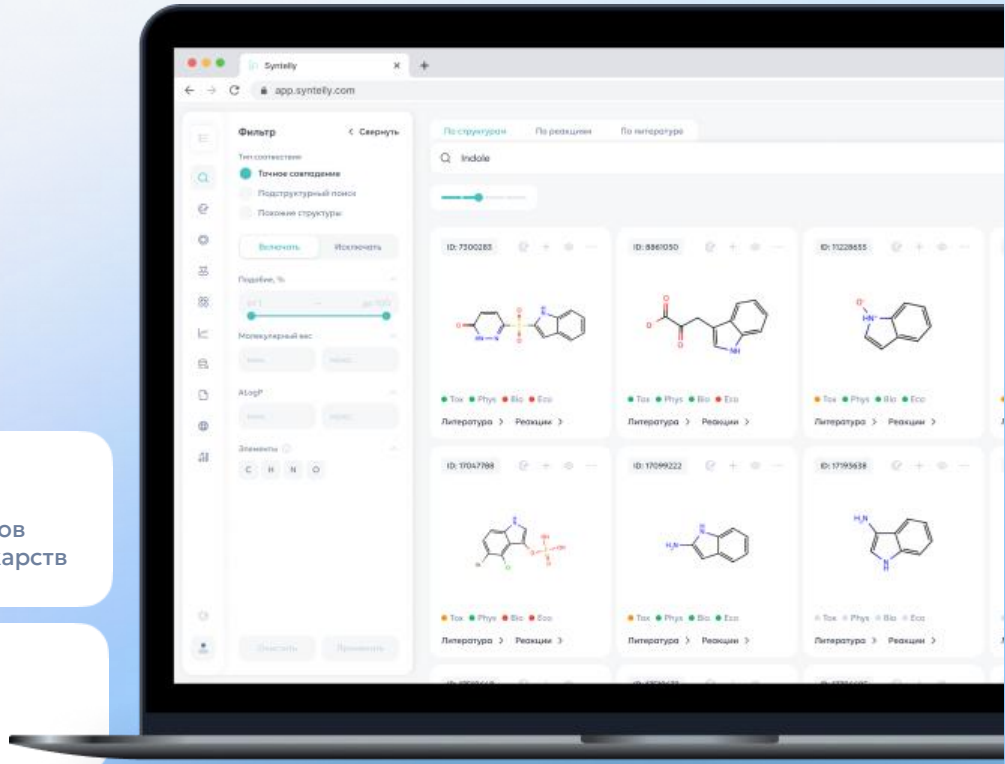
Необходимы новые инструменты на базе ИИ в качестве инструментов для исследований.
Искусственный интеллект ускоряет поиск в базах данных минимум в 20 раз, подбирая кластеры соединений с нужными свойствами, оценивая стоимость реакций и т.д.

Решение — провайдер химической информации

Специализированная SaaS платформа на основе искусственного интеллекта, позволяющая увеличить скорость и эффективность исследований в области органической химии.

Разработанный нами подход автоматического предсказания свойств химических соединений позволяет сократить на несколько порядков временные и денежные затраты.

Преимущества Синтелли — широкий функционал и большое количество модулей



В 20 раз быстрее

Поиск необходимой информации

На 3-4 года

Сокращение сроков на разработку лекарств

На 20%

Снижение затрат на исследования

Синтелли – ИИ-платформа для органической химии

Поиск информации

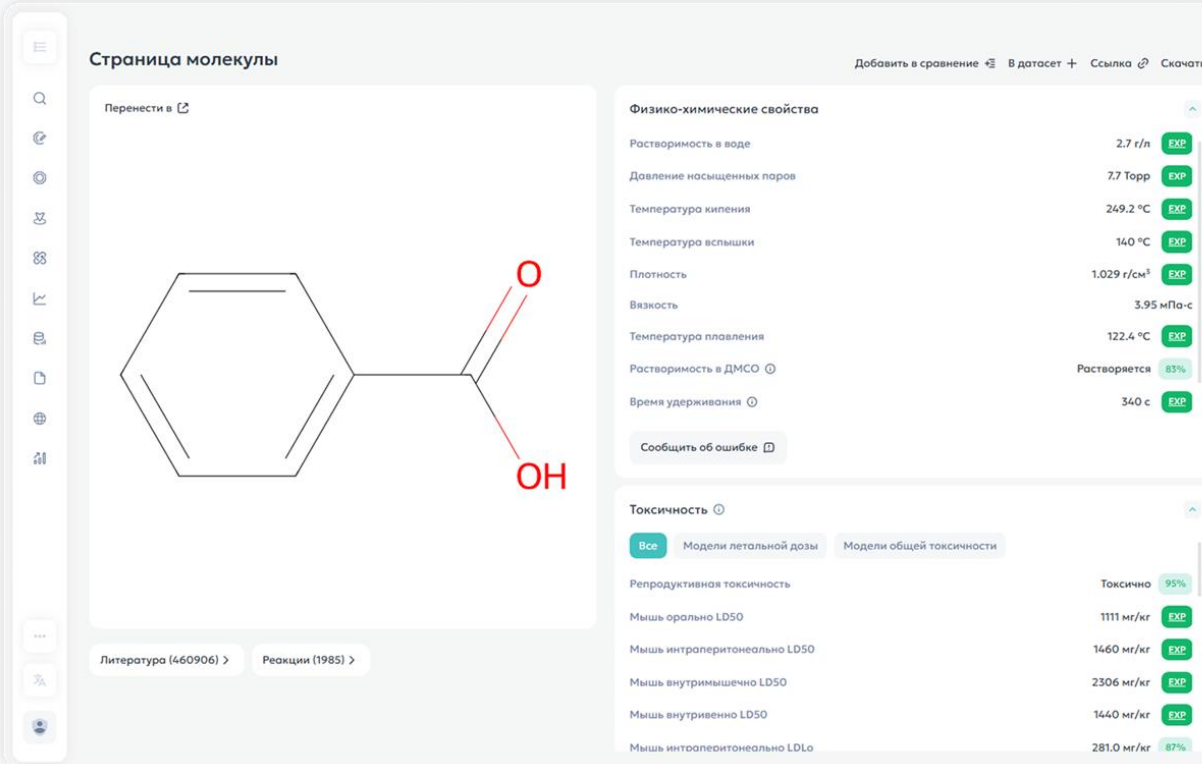
Структуры, экспериментальные свойства, реакции, публикации, патенты

Инструменты ИИ

Анализ данных, генерация новых молекул, прогнозирование свойств, эффективные методы синтеза

Удобство использования

Инструменты для совместной работы, лабораторный журнал, хранение и обработка данных



База данных Синтелли

Обновления каждый месяц

База данных Синтелли собрана из большого количества источников, таких как USPTO, WIPO, FIPS, PubMed, PubChem, Crossref и др. Дополняется с помощью собственного пайплайна нейросетевых модулей для извлечения информации из научных документов. Состоит преимущественно из малых органических молекул.

- Структурный
- Подструктурный
- Поиск по подобию
- Поиск по структурам Маркуша
- Комбинированный
- Полнотекстовый
- С учетом ограничивающих условий

7 млн
Реакций

20 млн
Патентов

150 млн
Научных публикаций

160 млн
Молекул

12,8 млрд
Данных о свойствах

30

ВУЗов и НИИ

10

клиентов

3500+

пользователей

Клиенты

Роспатент

SPLAT

Казанский федеральный университет

ИНФАМЕД К

INFAMED K

artgen biotech science for life

ГИЛС и НП

Государственный институт лекарственных средств и надлежащих практик

ТАТНЕФТЕХИМИНВЕСТ ХОЛДИНГ

Технологии Синтелли

С помощью ИИ происходит:

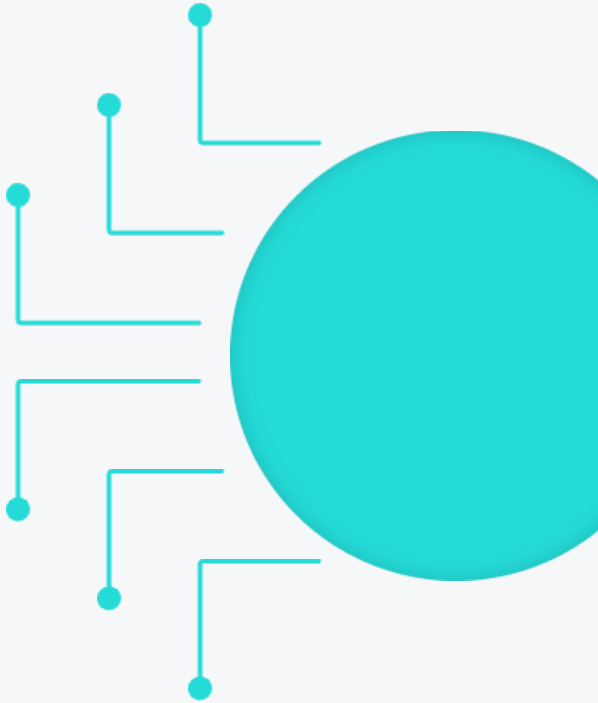
- Предсказание более 80 свойств молекул
- Генерация новых соединений
- Предсказание продуктов реакций и поиск ретросинтетических путей
- Оценка стоимости синтеза
- Визуализация химического пространства
- Предсказание спектров
- Извлечение химической информации из документов

Методы машинного обучения в Синтелли:

Классические методы, например модели градиентного бустинга, и различные виды нейронных сетей: полносвязные, графовые, трансформеры, а также специализированные модели компьютерного зрения и NLP.

Ядро Синтелли

Реализовано в масштабируемой микросервисной архитектуре с единой шиной обмена данными между модулями и базой данных.



Значимость NER для извлечения химических реакций

Автоматизированное извлечение структурированной информации о химических реакциях из неформализованных текстов (научные публикации, патенты, отчеты) критически важно для ускорения исследований в химии, фармацевтике и материаловедении.

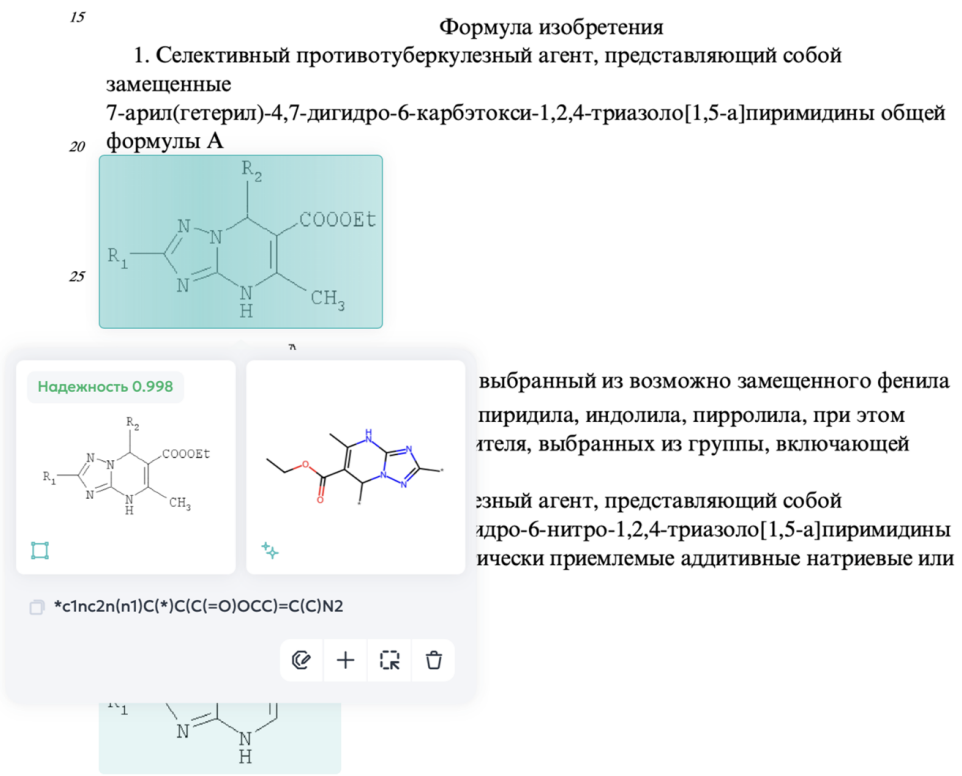
Синтелли реализует специализированную модель без интерактивного промптинга, реализующую парсинг текста в структурированные данные. Это обеспечивает высокую скорость, полную автоматизацию и простую интеграцию с ETL-системами.

Главное назначение системы — ускорение и автоматизация процессов анализа химических публикаций и формирование машинно-читаемых баз данных реакций для образовательных, исследовательских и научных организаций, а также компаний из фармацевтической, косметической и химической промышленности.

Научно-техническая база для NER в Синтелли

Синтелли обладает запатентованным методом автоматического распознавания химической информации непосредственно с изображений документов “Syntelly Img2Smiles” https://patents.google.com/patent/RU2774665C1/ru, что является ключевым технологическим элементом для извлечения структурных данных из научных публикаций, патентов и других химических источников.

Разработка Img2Smiles интегрирована в платформу для предоставления пользователям функционала автоматического и высококачественного преобразования изображений в машинно-читаемый формат



15 Формула изобретения

1. Селективный противотуберкулезный агент, представляющий собой замещенные 7-арил(гетерил)-4,7-дигидро-6-карбэтокси-1,2,4-триазоло[1,5-a]пиримидины общей формулы А

20

25

выбранный из возможно замещенного фенила
пиридила, индолила, пирролила, при этом
ителя, выбранных из группы, включающей

езный агент, представляющий собой
идро-6-нитро-1,2,4-триазоло[1,5-a]пиримидины
ически приемлемые аддитивные натриевые или

Существующие подходы

Модель / Год	Датасет (объём)	Описание решения	Качество
ChemicalTagger (2011)	~50 абзацев + ~10 000 патентов	Правила и грамматики (ANTLR + OSCAR4): выделение action-фраз (Add, Stir, Heat…), реагентов, продуктов, растворителей и условий; триггеры — глаголы-действия; связи через синтаксический парсер.	—
VinAI (ChEMU 2020)	1500 фрагментов из 170 патентов	BiLSTM–CNN-CRF для NER с BIO-тегами ролей (Starting Material, Reagent/Catalyst, Product, Solvent, Other, Time, Temperature, Yield, Example); EE не реализовано.	NER F1 = 94.3%
Mahendran et al. (ChEMU 2020)	1500 фрагментов из 170 патентов	BiLSTM–CRF vs BioBERT для NER; распознавание триггеров-глаголов; EE через правило «ближайший триггер» и CNN-классификатор пар (триггер→сущность).	NER ≈ 90%, EE ≈ 78%
IBM Seq2Seq (Vaucher et al., 2020)	~300 k «silver» + ~800 «gold» шагов	Transformer Seq2Seq: генерация полной последовательности лабораторных действий (Add, Stir, Heat, Filter…) с параметрами (вещество, кол-во, temp, time).	Полное совпадение = 60.8%; ≥ 75% совпадение = 82.4%
ChemRxnExtractor (Guo et al., 2021)	~800 реакций из JACS и др.	ChemBERT + CRF в два этапа: 1) BIO-теггер для продуктов; 2) multi-column BIO для ролей вокруг каждого продукта (Reactant/Solvent, Catalyst/Reagent, Temperature, Time, Yield).	Product F1 = 76.2%; Role linking F1 = 78.7%
ReactIE (Zhong et al., 2023)	599/96/111 (train/dev/test)	Weakly-supervised QA Seq2Seq (Flan-T5): задаются вопросы («What are the products?», «What is the catalyst?»), модель генерирует ответы-роли (Product, Reactant, Catalyst, Solvent, Reaction type, Temperature, Time, Yield).	Products F1 ≈ 91.1%; Roles F1 ≈ 81.6%
OpenChemIE (Qian et al., 2023)	~100 000 реакций из PDF-статей	Мультимодальный конвейер: BIO-NER хим. упоминаний + парсер таблиц и изображений → SMILES для IUPAC/тривиальных названий и R-групп; связи через идентификаторы на схемах.	exact match with Reaxys: F1 = 69.5%; P = 79%; R = 62%
MIT Fine-tuned LLM (2024)	100 000 пар текст→JSON (USPTO→ORD)	LLaMA-2 (7B) Fine-tune: генерация готового JSON по схеме ORD с ролями (Product, Reagent/Catalyst, Solvent, Temperature, Time, Yield, Reaction type); связи и разметка в структуре JSON.	JSON exact match = 91%; по отдельным полям = 92%

Наш пайплайн

1. PDF документ
2. Text Extraction/Preprocessing
3. Reaction protocol recognition
4. Chemical entity recognition
5. IUPAC2SMILES
6. Фильтрация
7. SMILES реакции + условия реакции

Преимущества использования NER

Потребительские преимущества:

1

Существенное сокращение времени на обработку научных текстов и формирование реакционной базы. Синтелли может обрабатывать до 3000 документов в сутки

2

Повышение полноты и точности извлечения благодаря комбинации NER-моделей.

3

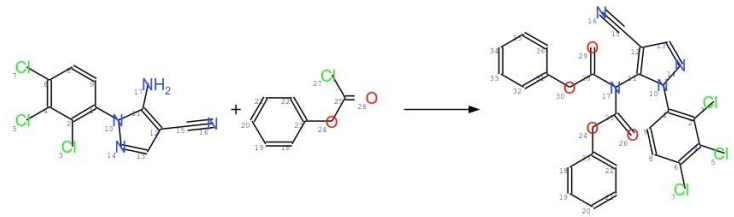
Возможность масштабирования на большие объемы данных и интеграции в корпоративные системы.

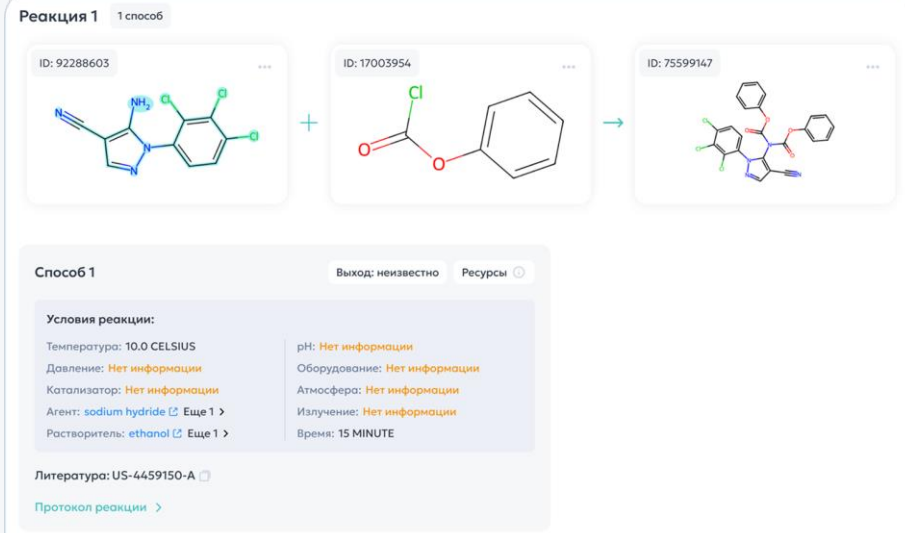
4

Улучшение воспроизводимости исследований за счёт стандартизированных описаний реакций в цифровом формате.

Конечный результат

Compound 4 reactant_compound_number | Sodium hydride chemical_general | (0.5 g mass) was added to a solution of 5-amino-4-cyano-1-(2,3,4-trichlorophenyl)pyrazole reactant | (5.8 g mass) in dimethylformamide solvent | (20 ml volume). After the initial exothermic reaction had subsided, the solution was heated on a steam-bath for 5 action_heat minutes and then cooled to 10° C temperature . Phenyl chloroformate reactant (2.8 ml volume) was then added slowly in small portions with occasional cooling with water chemical_general to keep the temperature below 50° C temperature . The mixture was then heated on a steam-bath for 10 action_heat minutes, cooled and evaporated to dryness. The solid residue was treated with diethyl ether chemical_general (300 ml). The ethereal suspension was filtered and the filtrate was evaporated to dryness. The solid residue thus obtained was boiled in ethanol solvent (200 ml volume) and separated by filtration, to give 4-cyano-5-di(phenoxycarbonyl)amino-1-(2,3,4-trichlorophenyl)pyrazole product (1.7 g mass), m.p. 206°-208° C physico_chemical_constant ., in the form of colourless crystals. EXAMPLE





Кейс Инфамед К

Цель: Разработать уникальный препарат с антисептическими и противовоспалительными свойствами.

Решение

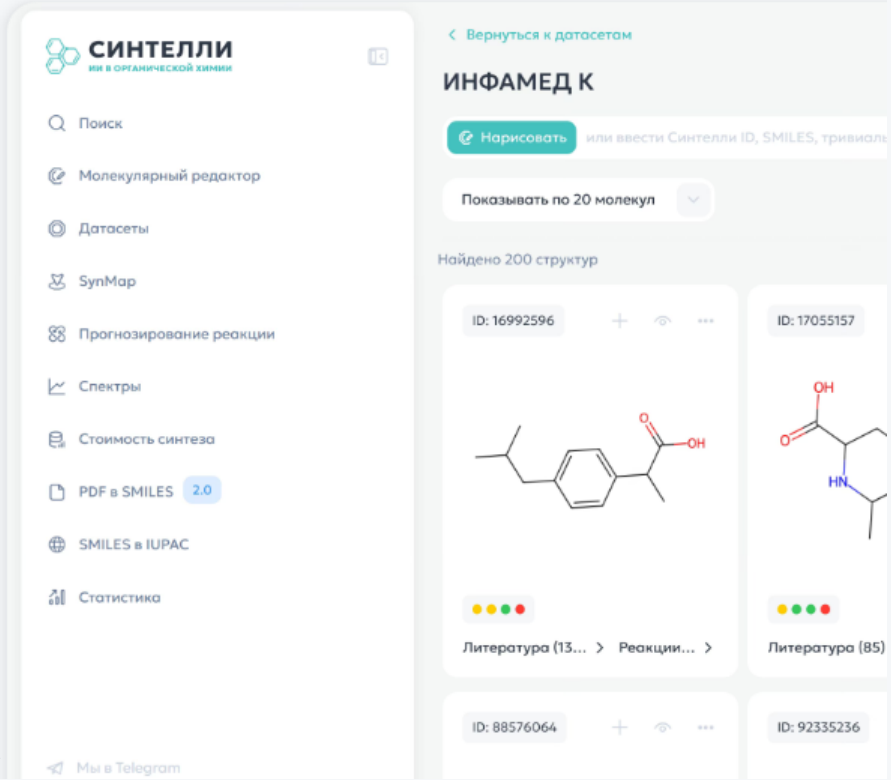
Использование модульной ИИ-платформы «Синтелли» для:

- Первичного скрининга более 200 молекул из миллионов возможных
- Отбора 38 перспективных соединений с учётом токсичности и патентной чистоты
- Предсказания спектров для подтверждения структуры (масс-спектры, ЯМР)
- Анализа научных публикаций и определения патентного пространства
- Определения путей синтеза для 6 приоритетных соединений

Результат

- Сокращение времени разработки нового препарата на 3–5 лет
- Снижение затрат на исследования за счет эффективного отсеивания неподходящих молекул
- Фокус на тестировании ограниченного количества молекул с повышенной вероятностью успеха

> «Ключевая ценность «Синтелли» для нас — значительное ускорение процесса разработки. Это позволяет сосредоточиться на глубокой оптимизации перспективных соединений и быстрее перейти к этапу клинических испытаний».
>
> — Константин Назаров, заместитель директора по производству ООО «ИНФАМЕД К»



Кейс Инфамед К

Цель: Разработать уникальный препарата с антисептическими и противовоспалительными свойствами.

Решение

Использование модульной ИИ-платформы «Синтелли» для:

- Первичного скрининга более 200 молекул из миллионов возможных
- Отбора 38 перспективных соединений с учётом токсичности и патентной чистоты
- Предсказания спектров для подтверждения структуры (масс-спектры, ЯМР)
- Анализа научных публикаций и определения патентного пространства
- Определения путей синтеза для 6 приоритетных соединений

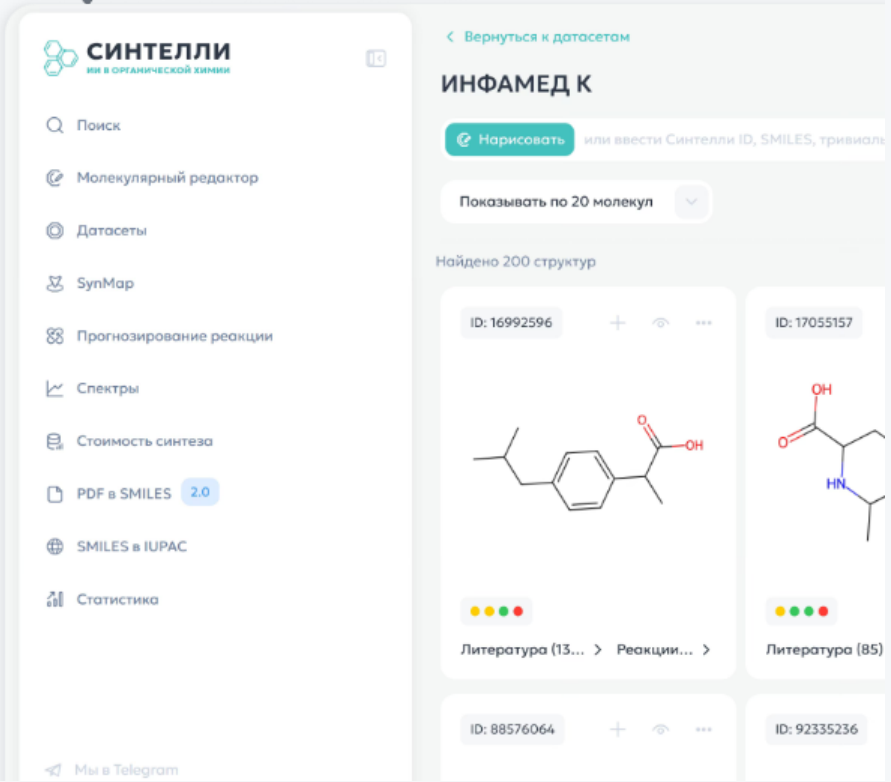
Результат

- Сокращение времени разработки нового препарата на 3–5 лет
- Снижение затрат на исследования за счет эффективного отсеивания неподходящих молекул
- Фокус на тестировании ограниченного количества молекул с повышенной вероятностью успеха

«Ключевая ценность «Синтелли» для нас — значительное ускорение процесса разработки. Это позволяет сосредоточиться на глубокой оптимизации перспективных соединений и быстрее перейти к этапу клинических испытаний».

— Константин Назаров, заместитель директора по производству ООО «ИНФАМЕД К»

Подробнее на сайте Синтелли по ссылке



СИНТЕЛЛИ
ИИ В ОРГАНИЧЕСКОЙ ХИМИИ

Синтелли — от гипотезы до открытия

Вместе мы создаём химию будущего