Функциональные характеристики программного обеспечения

Программный комплекс хемоинформатики на базе искусственного интеллекта для решения задач органической и медицинской химии

Москва
октябрь 2024

1. Аннотация

Настоящий документ описывает функциональные характеристики программного обеспечения (ПО) «Синтелли – программный комплекс хемоинформатики на базе искусственного интеллекта»

2. Функциональное назначение программы

ПО использует искусственный интеллект для оказания помощи пользователям в решении задач в области органической и медицинской химии. Функциональным назначением программы является:

1. поддержка пользователя в процессах экспериментального моделирования химических соединений с заданными свойствами;
2. представление информации о существующих химических соединениях, включенных в базу данных ИС для целей поиска информации о соединениях.

3. Эксплуатационное назначение программы

Программный комплекс Синтелли предназначен для использования в научно-исследовательских центрах и на предприятиях химической и фармакологической промышленности для поиска новых эффективных молекул-кандидатов в новые вещества и материалы (а том числе при разработке новых лекарственных препаратов). Также платформа может быть использована в организациях, занимающихся защитой интеллектуальной собственности, патентным поиском, в государственных, коммерческих и образовательных организациях.

4. Функциональные характеристики ПО

Программное обеспечение обеспечивает выполнение следующих функций:

1) ввод молекулярных структур в память ЭВМ с использованием графического интерфейса либо путем преобразования из химических нотаций;

2) поиск по химическим структурам, синонимам и различным идентификаторам (SMILES, InChI, название IUPAC, CAS-номер);
3) механизм структурного и подструктурного поиска молекул по критерию структурной близости;
4) генерация IUPAC имен (английский язык);
5) прогнозирование физико-химических, биологических, токсикологических свойств органических соединений (более 50 параметров);
6) расчет синтетической доступности молекулярных структур;
7) визуализация больших баз данных в виде точечных проекций на двумерную плоскость, а также навигация в химическом пространстве (анализ кластеров биоактивных соединений);
8) генерация новых структур органических соединений с заранее заданными пользователем свойствами (низкая токсичность, низкая синтетическая сложность и т.д);
9) Оптическое распознавание молекулярных структур из PDF (извлечение химических данных из различных неструктурированных печатных источников (PDF2SMILES 2.0))
10) работа с датасетами молекул и химических реакций;
11) экспорт молекулярных структур и рассчитанных значений в стандартные форматы данных, используемые в химии.
12) прогнозирование возможных продуктов реакции на основе реагентов, вступающих в нее, анализ синтеза искомой молекулы.

5. Компоненты программного обеспечения

№	Компонент продукта	Характеристики
1	Модуль «Поиск по структурам»	1. База данных свойств для уже изученных соединений (более 155 млн записей), включающая в себя: структуры органических соединений и известные свойства. Обеспечивает быстрый доступ к данным о соединениях , в том числе к экспериментальным данным; 2. Функционал сохранения истории поиска (поисковые запросы)

		3. Набор прогностических моделей на основе глубоких нейронных сетей для расчёта физико- химических, токсикологических, биологических свойств органических соединений: ● 10 физических свойств ● 4 экологических свойств ● 49 показателей токсичности ● ингибирование 5 цитохромов ● оценка мутагенности по тесту Эймса ● 2 различных способа оценки синтетической сложности ● вероятности того, что данная молекула может стать лекарственным средством и др Всего более 80 свойств в карточке структуры
2	Модуль поиска по литературным источникам	● База данных литературных источников, включающая публикации, патенты и заявки на патенты (160 млн записей). ● Механизм структурного и подструктурного поиска молекул по критерию структурной близости ● функционал полнотекстового поиска по патентной документации, комбинированный структурный и полнотекстовый поиск с учетом ограничивающих условий (дата публикации, автор, заявитель, владелец объекта интеллектуальной собственности) ● Возможность просмотра источника, где описано искомое соединение, ограничения поиска по дате публикации документа ● Функционал сохранения истории поиска (поисковые запросы)
3	Модуль «Молекулярный редактор»	Ввод и редактирование молекулярных структур с использованием графического интерфейса;

4	Модуль «Датасеты»	Поддержка пакетного режима обработки данных, функционал работы с дата-сетами молекул и химических реакций. Импорт в форматах: SDF, CSV, SMI. Экспорт в форматах: SDF, CSV.
5	Модуль визуализации химического пространства «SynMap»	Визуальный модуль анализа химического пространства (2D/3D), основанный на предобученной нейросетевой модели позволяет получить быстрое и наглядное представление об основных группах химических соединений которые есть в датасете. Модель производит проецирование структур химических соединений в координаты X и Y на двумерной плоскости. С помощью данного инструмента можно сравнивать и анализировать датасеты молекулы, накладывая их на карту различными слоями. Также данный модуль позволяет генерировать новые структуры с заранее заданными пользователем свойствами (низкая токсичность, низкая синтетическая сложность, высокая биологическая активность).
6	Модуль «Прогнозирование реакций»	Посвящен планированию синтеза органических соединений и включает в себя 2 функции: 1. Прогнозирование потенциальных продуктов реакции, на основе реагентов вступающих в нее (одностадийный органический синтез) 2. Анализа синтеза искомой молекулы (ретросинтетический анализ): деревья реакций
7	Модуль «Спектры»	Модуль позволяет прогнозировать спектральные данные ядерного магнитного резонанса (1H, 13C, 15N и 19F) для малых органических молекул. Результат представлен в виде набора "химический сдвиг - относительная интенсивность". Для спектров 1H также прогнозируется мультиплетность.
8	Модуль «Стоимость синтеза»	Аналитический инструмент, разработанный для оценки стоимости синтеза химических

		соединений. Необходимо ввести параметры желаемого синтеза: продукт, реагент, желаемый вес синтезируемого вещества и количество стадий реакции. Результатом является ТОП-5 схем реакций, упорядоченных по возрастанию стоимости. Это позволяет провести анализ по известным методикам и выбрать наиболее оптимальный путь синтеза с расчетом экономической эффективности. Модуль предоставляет возможности для детального анализа каждой схемы, редактирования таблицы стоимости и экспорта данных в форматах: Excel, PDF, CSV
9	Модуль «PDF2SMILES 2.0: Автоматическое извлечение структурных формул из документов в формате PDF»	Инструмент оптического распознавания химической документации: патентов, научных статей, протоколов испытаний, диссертаций и т.п. Данный модуль обеспечивает оптическое распознавание структур химических соединений и структур Маркуша, а также: ● функционал загрузки документов размером до 250 Мб ● функционал автоматического распознавания химической информации из патентных документов с помощью обновленного ансамбля нейронных сетей ● оценку надежности распознавания каждой отдельной структуры ● возможность редактирования распознанной структуры ● хранение распознанных документов пользователя (работа с коллекциями документов) ● возможность скачивания результатов распознавания в форматах: png, csv ● возможность сохранения распознанных структур в отдельный датасет
10	Модуль «SMILES в IUPAC»	Генерация систематических номенклатурных названий согласно правилам IUPAC на русском и английском языках

11	Модуль «Статистика»	Предоставляет пользователям данные о статистических параметрах (метриках) моделей машинного обучения, представленных в карточке структуры (RMSE, ROC AUC)

6. Внешнее окружение ПО

Продукт реализован по клиент-серверной модели. Серверная часть работает под руководством ОС Linux, клиентская – веб-порталом, поддерживающим работу с основными современными браузерами: Chrome, Яндекс.Браузер, Opera, FireFox, Safari.