МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра биомедицинской информатики

Генерация описаний химических соединений методами глубокого обучения

Курсовая работа

Малыщика Акима Андреевича
студента 3 курса 3 группы
специальность "информатика"

Научный руководитель:
профессор кафедры БМИ
Тузиков Александр Васильевич

Минск, 2022

РЕФЕРАТ

Курсовая работа, 20 стр., 5 иллюстр., 9 источников.

Ключевые слова: ДЕСКРИПТОРЫ, МОЛЕКУЛЯРНЫЕ ФИНГЕРПРИНТЫ, SMILES.

Объекты исследования – дескрипторы химических соединений, алгоритмы машинного обучения на основе дескрипторов.

Цель исследования – изучение существующих молекулярных дескрипторов и алгоритмов на их основе.

Методы исследования – системный подход, изучение соответствующей литературы и электронных источников.

В результате исследования изучены графовые структуры, молекулярные фингерпринты и SMILES, исследованы алгоритмы обработки дескрипторов, рассмотрены архитектуры нейросетей на основе SMILES-описаний.

Области применения – хемоинформатика, медицина, фармацевтика.

Содержание

Введение

Компьютерное моделирование лекарств — относительно быстро развивающаяся и достаточно перспективная отрасль IT-индустрии, в которой активно используются алгоритмы машинного обучения. Генеративные нейронные сети применяются для получения новых соединений, которые потенциально могут стать основой для лекарственных средств. Использование алгоритмов глубокого обучения позволяет ускорить и удешевить существующий процесс создания лекарственных препаратов. Также, засчёт использования искусственного интеллекта появляется возможность рассмотрения соединений, которых нет в существующих на данный момент химических базах данных. Соответственно, методы машинного обучения могут позволить получить новые соединения, которые по сравнению с известными будут более эффективны к заданной молекулярной мишени.

Основная проблема такого подхода в том, что на данный момент не существует универсального метода кодирования химических соединений, который был бы лучше всех остальных в любой ситуации. Из анализа литературы можно сделать вывод, что наиболее перспективным форматом представления молекул в памяти компьютера сейчас являются SMILES-описания. Данный способ описания химических соединений позволяет применять алгоритмы машинного обучения и архитектуры нейронных сетей, используемые также при работе с естественными языками.

Таким образом, подходы к созданию генеративных моделей для химических веществ будут похожи на подходы, используемые в задачах NLP. Одним из наиболее перспективных методов на данный момент является обучение архитектуры с энкодером и декодером, в которой энкодер преобразует подаваемые на вход описания молекул, а обученный декодер используется как генератор новых химических соединений с необходимыми свойствами.

Цель работы: Получить описания химических веществ с заданными свойствами, которые будут эффективны к молекулярной мишени.

Задача работы: Разработать генеративную нейронную сеть для генерации описаний химических соединений.

Список использованных источников

1. Введение в хемоинформатику: Компьютерное представление химических структур: учеб. пособие / Т.И. Маджидов, И.И. Баскин, И.С. Антипин, А.А. Варнек. – Казань, Москва, Страсбург, 2020 – 176 с.

2. Чумаков А.А., Слижов Ю.Г. Система SMILES-кодирования молекулярных структур и её применение для решения научно-исследовательских задач. Национальный исследовательский Томский государственный университет, Электронное методическое пособие, 2017.–18 с.

3. M.A. Shuldau et al. Development of molecular autoencoders as generators of protein inhibitors: Application for prediction of potential drugs against coronavirus SARS-CoV-2 //Proceedings of the 15th Ibterbational Conference on Pattern Recognition and Information Processing (PRIP'2021), Sep. 21-24, 2021, Minsk, Belarus, 2021.

4. Fingerprints – Screening and Similarity. – 2019. – 1 с. – URL: https://www.daylight.com/dayhtml/doc/theory/theory.finger.html (дата обращения: 24.10.2021, 18:31)

5. A Practical Introduction to the Use of Molecular Fingerprints in Drug Discovery. – 2019. – 1 с. – URL: https://towardsdatascience.com/a-practical-introduction-to-the-use-of-molecular-fingerprints-in-drug-discovery-7f15021be2b1 (дата обращения: 02.11.2021, 15:28)

6. Mayr A. et al. DeepTox: toxicity prediction using deep learning //Frontiers in Environmental Science. – 2016. – 80 с.

7. Convolutional Networks on Graphs for Learning Molecular Fingerprints / David Duvenaud [и др.] // Harvard University, 2015 – 9 с.

8. SMILES – A Simplified Chemical Language. – 2019. – 1 с. – URL: https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (дата обращения: 15.11.2021, 22:03)

9. Bjerrum, E.J.; Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. – 2018. 131 с. https://doi.org/10.3390/biom8040131