

ХЕРСОНСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ДИЗАЙНУ
КАФЕДРА ІНФОРМАТИКИ І КОМП'ЮТЕРНИХ НАУК

Пояснювальна записка

до дипломної магістерської роботи

на тему:

**ВИКОРИСТАННЯ АВТОКОДУВАЛЬНИКІВ ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧ
КЛАСТЕРИЗАЦІЇ МАС-СПЕКТРІВ ХЛОРОРГАНІЧНИХ СПОЛУК З
КОНСОЛІДОВАНОГО РЕСУРСУ ХІМІЧНИХ БАЗ ДАНИХ**

Виконав: студент 6 курсу, групи 6КІ

спеціальності 122 «Консолідована інформація»

Сотніков Кирило Олександрович

(прізвище та ініціали)

Керівник: Літвиненко В.І.

(прізвище та ініціали)

Рецензент: _____

(прізвище та ініціали)

Хмельницький – 2024 р

Факультет	Інформаційних технологій та дизайну
Кафедра	Інформатики і комп'ютерних наук
Рівень вищої освіти	<u>другий (магістерський) рівень</u>
Галузь підготовки	<u>12 «Інформаційні технології»</u>
Освітньо-професійна програма	<u>Консолідована інформація</u> (назва)
Спеціальність	<u>122 «Компютерні науки»</u> (шифр і назва)

ЗАТВЕРДЖУЮ

Завідувача кафедри ІКН,
професор

_____ В.І. Литвиненко

«_____» _____ 2024 року

З А В Д А Н Н Я
НА ДИПЛОМНУ РОБОТУ СТУДЕНТУ

_____ Сотніков Кирило Олександрович

(прізвище, ім'я, по батькові)

1. Тема роботи Використання автокодувальників для розв'язання задач кластеризації мас-спектрів хлорорганічних сполук з консолідованого ресурсу хімічних баз даних
2. Керівник роботи Літвиненко Володимир Іванович, доктор технічних наук, професор кафедри інформатики і КН

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом ХНТУ від «25» вересня 2024 року №499-с.

3. Строк подання студентом роботи

4. Вихідні дані до роботи

5. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)

6. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)

7. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

8. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

[illegible]

Студент _____ Сотніков К.О.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Літвиненко В.І.
(підпис) (прізвище та ініціали)

Зміст

Вступ.....	12
РОЗДІЛ 1.....	13
ОСНОВНІ ПОНЯТТЯ.....	14
1.1.Автокодувальник.....	14
1.1.1.Базова архітектура.....	15
1.1.2.Різновиди	16
1.1.2.1.Розріджений автокодувальник (РАК).....	17
1.1.2.2.Знешумлювальний автокодувальник (ЗАК).....	18
1.1.2.3.Стягувальний автокодувальник (САК).....	18
1.1.2.4.Конкретний авто кодувальник.....	19
1.1.2.5.Варіаційний автокодувальник (ВАК).....	19
1.1.3.Тренування.....	19
1.2.Хлорорганічні сполуки (ХОС).....	20
1.2.1.Хлорорганічні сполуки (ХОС).....	20
1.2.2.Історія.....	20
1.2.3.Фізико-хімічні властивості.....	21
1.2.4. Дія на шкідливі організми.....	23
1.2.5.Токсикологічні властивості і характеристики	23
1.2.6.Безпечність.....	24
1.3. Основні поняття кластеризації та постановка задачі.....	24

1.4.Загальна характеристика кластеризації	27
1.4.1.Методи кластеризації	28
1.4.2.Підходи до кластеризації	30
1.4.3.Математичні характеристики кластера.....	30
1.4.4.Критерії зупинки кластеризації	31
ВИСНОВОК ДО РОЗДІЛУ 1.....	35
РОЗДІЛ 2.....	36
ІНСТРУМЕНТИ ТА БІБЛІОТЕКИ ЯКІ ВИКОРИСТОВУЮТЬСЯ В РОБОТІ , ТА ЇХ ПЛЮСИ І МІНУСИ.....	36
2.1. Автокодувальники (Autoencoders)	36
2.2. Алгоритм кластеризації K-Means	37
2.3. Бібліотеки для машинного навчання (TensorFlow/Keras),Scikit- learn.....	43
2.3.1.TensorFlow.....	43
2.3.2.Keras.....	45
2.3.3.Scikit-learn.....	47
2.4. Інструменти для візуалізації даних (Matplotlib, TSNE,K-Means).....	50
2.4.1.Matplotlib.....	50
2.4.2. t-SNE	55
2.5.Оптимізація алгоритмів кластеризації	59
2.6.Покращення моделей для зменшення розмірності	60
2.7.Покращення прогами.....	60

2.8. Бази даних хімічних сполук (PubChem, ChEMBL, ZINC).....	62
2.9.Опис середовища розробки PyCharm.....	73
ВИСНОВОК ДО РОЗДІЛУ 2.....	76
РОЗДІЛ 3.....	77
ЕТАПИ ВИКОНАННЯ РОБОТИ.....	77
3.1 Підготовка даних.....	77
3.2 Реалізація авто кодувальника.....	78
3.3 Виконання кластеризації K-Means.....	79
3.4 Візуалізація даних та результатів.....	80
3.5 Оцінка ефективності	80
ВИСНОВОК ДО РОЗДІЛУ 3	80
РОЗДІЛ 4.....	81
РЕЗУЛЬТАТИ РОБОТИ.....	81
4.1.1 Результати кластеризації хлорорганічних сполук	81
4.1.2 Аналіз візуалізації.....	81
4.1.3 Оцінка ефективності моделі.....	82
4.1.4 Обговорення.....	82
4.1.4.1. Аналіз отриманих результатів кластеризації.....	83
4.1.4.2. Порівняння використаних інструментів.....	83
4.1.4.3. Порівняння з альтернативними методами.....	84
4.1.4.4. Перспективи використання результатів.....	84

4.1.4.5. Обмеження та шляхи їх подолання.....	85
4.2.Скріншоти результат роботи.....	86
4.3.Пояснення результатів.....	88
ВИСНОВОК ДО РОЗДІЛУ 4.....	90
Висновок	92
Список використаних джерел.....	94
Анотації.....	96
Додатки	97
Код програми	97
Презентація	100

Реферат

Поязнювальна записка до випускної роботи : 104с.,9 рис., 27.джерел.1 табл

Об'єктом дослідження є хлорорганічні сполуки, представлені у вигляді числових дескрипторів, отриманих із відкритих хімічних баз даних (PubChem, ChEMBL, ZINC тощо).

Предметом дослідження є методи зменшення розмірності та кластеризації хімічних даних із використанням автокодувальників, а також їх застосування для аналізу структури та властивостей хлорорганічних сполук.

Мета роботи полягає у розробці та впровадженні методології використання автокодувальників для кластеризації хлорорганічних сполук з метою виявлення закономірностей у їхніх фізико-хімічних та структурних характеристиках.

Ключеві слова :ХОС, TensorFlow/Keras, Matplotlib, TSNE PubChem, ChEMBL, ZINC.

Abstract

Explanatory note to the final work: 104p., 9 fig., 27.sources.1 table

The object of the study is organochlorine compounds, presented in the form of numerical descriptors obtained from open chemical databases (PubChem, ChEMBL, ZINC, etc.).

The subject of the study is methods of dimensionality reduction and clustering of chemical data using autoencoders, as well as their application for analyzing the structure and properties of organochlorine compounds.

The purpose of the work is to develop and implement a methodology for using autoencoders for clustering organochlorine compounds in order to identify patterns in their physicochemical and structural characteristics.

Keywords: XOC, TensorFlow/Keras, Matplotlib, TSNE PubChem, ChEMBL, ZINC.

Вступ

Хлорорганічні сполуки є важливим класом хімічних речовин, які широко застосовуються у промисловості, сільському господарстві та медицині. Однак їх висока токсичність, біоаккумуляція та вплив на навколишнє середовище викликають значний науковий інтерес. Аналіз структурних та функціональних аспектів цих сполук є важливим для розуміння їх властивостей, зменшення екологічних ризиків та розробки нових матеріалів із бажаними характеристиками.

Одним із сучасних підходів до аналізу великомасштабних хімічних даних є використання методів машинного навчання, зокрема автокодувальників. Автокодувальники дозволяють ефективно зменшувати розмірність даних, виявляти приховані закономірності та кластеризувати хімічні сполуки за їх структурними чи фізико-хімічними характеристиками. Ця робота присвячена застосуванню автокодувальників для кластеризації хлорорганічних сполук із консолідованого ресурсу хімічних баз даних.

Об'єкт дослідження

Об'єктом дослідження є хлорорганічні сполуки, представлені у вигляді числових дескрипторів, отриманих із відкритих хімічних баз даних (PubChem, ChEMBL, ZINC тощо).

Предмет дослідження

Предметом дослідження є методи зменшення розмірності та кластеризації хімічних даних із використанням автокодувальників, а також їх застосування для аналізу структури та властивостей хлорорганічних сполук.

Мета роботи

Мета роботи полягає у розробці та впровадженні методології використання автокодувальників для кластеризації хлорорганічних сполук з метою виявлення закономірностей у їхніх фізико-хімічних та структурних характеристиках.

Наукова новизна одержаних результатів

1. Запропоновано підхід до використання автокодувальників для кластеризації хлорорганічних сполук, що дозволяє зменшити розмірність даних без втрати ключової інформації про структуру сполук.
2. Вперше створено латентний простір для хлорорганічних сполук, який відображає їхні ключові структурні та функціональні характеристики.
3. Отримані кластери сполук дали змогу ідентифікувати групи із схожими фізико-хімічними властивостями, що може бути використано для подальших досліджень у галузях екологічного аналізу та синтезу нових матеріалів.
4. Розроблено алгоритм інтерпретації латентного простору для прогнозування можливих властивостей хлорорганічних сполук.

Ця робота має теоретичне та практичне значення для хімічної інформатики, оскільки пропонує сучасний підхід до аналізу великих хімічних даних, що є актуальним у контексті розвитку технологій обробки та інтерпретації хімічної інформації.

Апробація результатів магістерської кваліфікаційної роботи.