МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХЕРСОНСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ
Факультет комп'ютерних наук, фізики та математики
Кафедра інформатики, програмної інженерії та економічної кібернетики

MACHINE LEARNING ДЛЯ РЕАЛІЗАЦІЇ ПРОГРАМНИХ ЗАСОБІВ ВИЯВЛЕННЯ ПОТЕНЦІЙНИХ ІНГІБІТОРІВ БІЛКА У МОЛЕКУЛЯРНОМУ ДОКІНГУ

Кваліфікаційна робот (проєкт)

на здобуття ступеня вищої освіти «бакалавр»

Виконав: студент 4 курсу 431 групи
Спеціальності 121 Комп'ютерні науки
Батуашвілі Ерік
Керівники: доктор фізико-математичних наук, професор Песчаненко Володимир Сергійович; доктор педагогічних наук, професор Співаковський Олександр Володимирович
Рецензент: кандидатка педагогічних наук, доцентка Єрмакова-Черченко Наталія Олександрівна

Херсон – 2021

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

PDB	-	Protein Data Bank
VMD	-	Visual Molecular Dynamics
MMTK	-	Molecular Modelling Toolkit
MERTK	-	Tyrosine-protein kinase Mer
PMV	-	Python Molecule Viewer
RBVI	-	Resource for Biocomputing, Visualization and Informatics

ВСТУП

Актуальність теми. Створення ліків та дослідження їх сприяння на людський організм є необхідністю у наш час. Процес винаходження нових ліків складається з пошуку сполук, які будуть відповідати необхідним характеристикам на важливих ділянках структури рецепторів (білків). Сучасні програми дозволяють здійснити такий пошук за допомогою обчислювального експерименту, котрий полягає у прикладанні просторових образів ліганд і білка. Ця процедура називається програмним докінгом. Найбільш адекватним підходом при цьому є використання даних про структуру білка, отриманих рентгеноструктурним методом; у випадку їх відсутності відомості про просторову будову ділянки зв'язування можуть бути отримані шляхом аналізу структури вже відомих ефективних лігандів. Комп'ютерні програми, які здійснюють програмний докінг, дозволяють виявляти комплементарні молекули в великих базах хімічних структур, в тому числі в каталогах комерційно доступних сполук. Докінг як засіб пошуку потенційних лігандів серед десятків і сотень тисяч органічних сполук повинен зв’язуватися з оцінкою енергії взаємодії ліганд з рецепторами, так само як і з процедурами, що дозволяють в певних межах максимізувати цю енергію. Ця робота полягає у використанні нейронної мережі для виявлення оптимального біологічно активного молекулярного комплексу. Найкращі рішення будуть обрані опорними для створення біологічно активних компонентів.

Об’єкт дослідження: структура живих організмів.

Предмет дослідження: комп’ютеризація біохімічних процесів.

Мета дослідження: розробка та використання штучного інтелекту у процесах поєднування живих клітин; моделювання структур біохімічних даних та їх моделей у просторі.

Досягнення мети дослідження передбачало розв’язання таких завдань:

- Дослідження клітин та процесів у структурах живих організмів;
- Пошук та підготовка молекулярних структур для реалізації процесу докінгу;
- Реалізація молекулярного докінгу та обробка результатів;
- Створення штучної нейронної мережі;
- Перевірка доцільності використання засобів машинного навчання.

РОЗДІЛ 1
ТЕОРЕТИЧНА ОСНОВА МОЛЕКУЛЯРНОГО ДОКІНГУ

1.1. Визначення молекулярного докінгу

Молекулярний докінг є ключовим інструментом у структурній молекулярній біології та комп'ютерному дизайні ліків. Мета ліганд-білкового докінгу полягає в тому, щоб передбачити оптимальний режим(и) зв'язування ліганд з білком відомої тривимірної структури. Вдалі методи стикування шукають багатовимірні простори і використовують скорінг функцію, яка правильно ранжує кандидатські стикування. Докінг може бути використаний для виконання віртуального скринінгу на великих бібліотеках сполук, ранжирування результатів і пропозиції структурних гіпотез про те, як ліганди інгібують мішень.

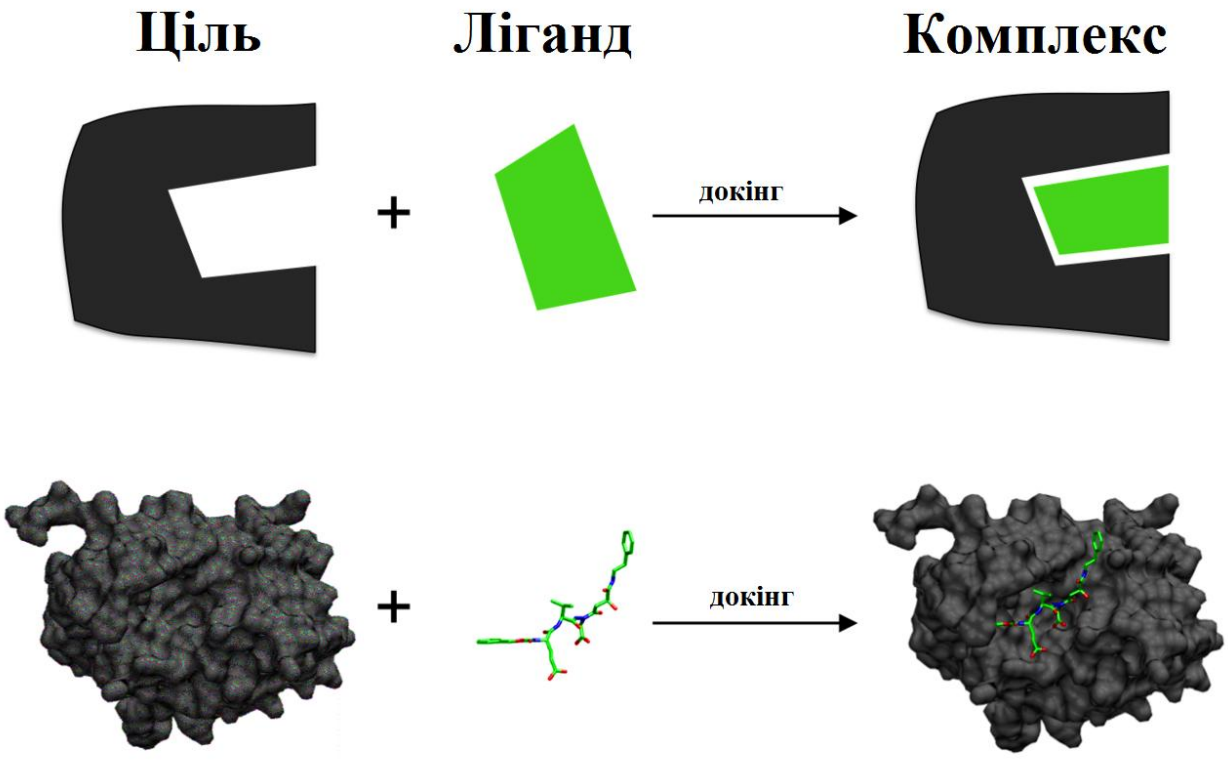


Рис.1.1. Діаграма докінгу малої молекули ліганди (зелена) з білковим рецептором (сіра)

Молекулярний докінг використовується для моделювання процесу молекулярного впізнавання. Зазвичай необхідно знайти оптимальну конформацію ліганду (рис.1.1). Дане положення досягається в разі, коли вільна енергія зв'язування мінімальна.

Існує багато програм для теоретичного докінгу білків. У більшості випадків білок-рецептор фіксується в просторі, а ліганд повертається навколо нього. При цьому для кожної конфігурації поворотів обчислюються оціночні розрахунки по скоринговій функції. Скорингова функція заснована на поверхневій компліментарності, електростатичних взаємодіях, Ван-дер-Ваальсовському відштовхуванні та утворення водневих зв’язків. Проблема при цьому пошуку в тому, що обчислення по всьому конфігураційному простору вимагають багато часу на обчислення та рідко приводять до єдиного рішення.

1.2. Підходи до моделювання докінгу

Існують різні підходи при моделюванні докінгу. Один з підходів використовує техніку відповідності, яка описує білок і ліганд як додаткові поверхні. Інший підхід моделює фактичний процес докінгу, в якому обчислюються попарні енергії взаємодії. У обох підходів є істотні переваги, а також деякі обмеження.[1]

Жорстким називається докінг, при якому довжини зв'язків, кути і торсіонні кути партнерів докінгу залишаються незмінними в процесі моделювання. Однак в результаті взаємодії з іншим білком або лігандом відбуваються конформаційні зміни як самого білка, так і бічних ланцюгів. Рухливість остова, в свою чергу, розділяється на два типи: рухливість великих ділянок білка - доменів, так званий рух зсуву, і рухливість окремих частин, таких як петлі. В даному випадку жорсткий докінг некоректно описує взаємодії. Тому існують деякі додаткові алгоритми гнучкого докінгу. Вони допускають конформаційні зміни, в результаті чого даний підхід дозволяє отримувати оцінки взаємодій найбільш наближені до природних. Однак підрахунок всіх можливих конформаційних змін з урахуванням руху на даному рівні розвитку комп'ютерів зайняв би величезний час. Більш того, велика кількість ступенів свободи також може призводити до збільшення кількості помилково позитивних результатів. У зв'язку з даними проблемами,

виникає необхідність раціонально вибирати невелику підмножину можливих конформаційних змін для проведення моделювання.

Гнучкий докінг також може бути використаний в рамках докінгу низькомолекулярного з'єднання. Однак в даному випадку дозволяється обертання навколо будь-яких зв'язків в молекулі самого ліганду, білок при цьому залишається жорсткою структурою.

Докінг також можна розділити на одноразовий і послідовний. Послідовний докінг застосовується, в основному, для докінгу декількох низькомолекулярних сполук (лігандів). Після докінгу одного з лігандів в окремий файл зберігається структура білка з даним лігандом. Далі алгоритм повторюється і реалізується докінг для другого ліганду в раніше збережену структуру. Даний підхід може бути корисний при пошуку аллостеричних центрів.

Геометрична відповідність (методи визначення взаємозалежності форми) описується для білка і ліганди як ряд особливостей, які визначають їх оптимальну взаємодію. Ці особливості можуть включати як саму молекулярну поверхню, так і опис додаткових особливостей поверхні. В цьому випадку молекулярна поверхня рецептора описується з точки зору її доступності для розчинника, а молекулярна поверхня ліганди описується з точки зору її відповідності опису поверхні рецептора. Взаємозалежність між двома поверхнями складає опис відповідності форми, яке може допомогти виявити різні положення ліганд. В іншому підході потрібно описати гідрофобні особливості білка, використовуючи повороти в атомах головного ланцюга.

При моделюванні докінгу білок і ліганд відокремлені деякою фізичною відстанню, і ліганд знаходить свої координати відносно активного центру білка після певного числа ітерацій. Кожна ітерація включає в себе перетворення твердого тіла, такі як переміщення і обертання, а також внутрішні зміни структури ліганду, включаючи кутові обертання. Кожен з цих кроків в просторі змінює повну енергетичну оцінку системи, і, отже, вона обчислюється після кожного руху. Очевидна перевага цього методу полягає в тому, що це дозволяє досліджувати гнучкість ліганду під час моделювання, тоді як методи взаємозалежності форми повинні використовувати деякі інші підходи, щоб дізнаватися про рухливості ліганду. Інша перевага полягає в тому, що процес фізично ближче до того, що

відбувається в дійсності, коли білок і ліганд наближаються один до одного після молекулярного розпізнавання. Незручність полягає у великих затратах часу для оцінки оптимального рішення докінгу, так як необхідно досліджувати досить великий енергетичний ландшафт.

1.3. Механізми докінгу

Для проведення докінгу першою вимогою є структура необхідного білка. Зазвичай структуру визначають за допомогою біофізичних методів, таких як рентгенівська кристалографія, ЯМР-спектроскопія або кріо-електронна мікроскопія (кріо-ЕМ), але також може бути отримана в результаті побудови моделювання гомології. Ця структура білка та база даних про потенційні ліганди служать вхідними матеріалами для програми докінгу. Успіх програми залежить від двох компонентів: алгоритму пошуку та функції оцінки.

Теоретично, алгоритм пошуку простору складається з усіх можливих орієнтацій та конформацій білка-рецептора в парі з лігандом. Однак на практиці з сучасними обчислювальними ресурсами неможливо повністю (вичерпно) дослідити простір пошуку - це передбачало б перерахування всіх можливих положень кожної молекули (молекули динамічні та існують у комплексі конформаційних станів) та всіх можливих обертальних та поступальних орієнтацій ліганд щодо білка при заданому рівні зернистості. Більшість використовуваних докінгових програм враховують весь конформаційний простір ліганду, а декілька намагаються змоделювати гнучкий білковий рецептор. Кожен знімок пари називається позою.[2][3]

Різні стратегії конформаційного пошуку були застосовані до ліганду та до рецептора. До них належать:

- систематичні або стохастичні крутильні пошуки обертових зв’язків;
- моделювання молекулярної динаміки;
- генетичні алгоритми для еволюції нових малоенергетичних конформацій і де оцінка кожної пози виконує функцію скорингу, що використовується для вибору осіб для наступної ітерації.

Конформації ліганду можуть генеруватися у відсутності рецептора, а згодом стикуватися або конформації можуть створюватися на льоту в присутності порожнини, що зв'язує рецептор, або з повною гнучкістю обертання кожного двогранного кута за допомогою стикування на основі фрагментів. Оцінка енергії поля сили найчастіше використовується для вибору енергетично обґрунтованих конформацій, а також використовуються методи, засновані на знаннях.

Пептиди - це одночасно дуже гнучкі та відносно великі молекули, що робить моделювання їх гнучкості складним завданням. Було розроблено ряд методів, що дозволяють ефективно моделювати гнучкість пептидів під час стикування білків-пептидів.

Обчислювальна здатність сучасних ЕОМ збільшилася, що зробило можливим використання більш складних обчислювальних методів у розробці лікарських препаратів. Однак боротьба з гнучкістю рецепторів у методологіях стикування залишається ще складним питанням. Основною причиною цієї складності є велика кількість ступенів свободи, які доводиться враховувати при розрахунках. Однак нехтування ними в деяких випадках може призвести до поганих результатів стикування з точки зору передбачення пози прив'язки.

Для імітації гнучкості рецепторів часто використовуються кілька статичних структур, експериментально визначених для одного і того ж білка в різних конформаціях. Альтернативно, бібліотеки ротамерів бічних ланцюгів амінокислот, які оточують порожнину зв'язування, можна шукати, щоб отримати альтернативні, але енергетично обґрунтовані білкові конформації.

Програми стикування генерують велику кількість потенційних поз ліганду, деякі з яких можна негайно відхилити через зіткнення з білком. Решта обчислюються за допомогою заданої скорингової функції, яка приймає позу як вхід і повертає число, що вказує на ймовірність того, що поза

представляє сприятливу взаємодію зв’язування та ранжує один ліганд відносно іншого.

Більшість скорингових функцій - це силові поля молекулярної механіки, засновані на фізиці, які оцінюють енергію пози в межах місця зв'язування. Різні внески у прив’язку можна записати як адитивне рівняння:

$$\triangle G_{bind} = \triangle G_{solvent} + \triangle G_{conf} + \triangle G_{int} + \triangle G_{rot} + \triangle G_{t/t} + \triangle G_{vib}$$

Компоненти складаються з ефектів розчинника, конформаційних змін білка та ліганду, вільної енергії внаслідок взаємодії білок-ліганд, внутрішніх обертань, енергії асоціації ліганду та рецептора з утворенням єдиного комплексу та вільної енергії внаслідок зміни режимів коливань. Низька (негативна) енергія вказує на стабільну систему і, отже, на ймовірну взаємодію, що зв'язує.

Альтернативний підхід полягає у отриманні на основі знань статистичного потенціалу взаємодій з великої бази даних білково-лігандних комплексів, таких як банк даних білків (Protein Data Bank), та оцінки відповідності пози відповідно до цього висновку.

Існує велика кількість структур рентгенівської кристалографії для комплексів між білками та лігандами з високою спорідненістю, але порівняно менше для лігандів з низькою спорідненістю, оскільки пізніші комплекси, як правило, менш стабільні і, отже, важче кристалізуються. Функції оцінки, навчені цими даними, можуть правильно стикувати ліганди з високою спорідненістю, але вони також даватимуть правдоподібні конформації для лігандів, які не зв'язуються. Це дає велику кількість хибнопозитивних випадків, тобто ліганди, яким передбачається зв'язуватися з білком, насправді не поєднуються.

Одним із способів зменшити кількість помилкових випадків є перерахунок енергії найвищих оцінювальних конформацій, використовуючи

більш точні, але обчислювально більш інтенсивні методи, такі як метод Пуассона-Больцмана.

1.4. Оцінка алгоритмів докінгу

Взаємозалежність між вибіркою та скоринговою функцією впливає на здатність докінгу при прогнозуванні правдоподібних позицій або споріднених властивостей для нових сполук. Таким чином, для визначення його передбачувальної здатності зазвичай потрібна оцінка протоколу стикування (коли доступні експериментальні дані). Оцінка стикування може проводитися за допомогою різних стратегій, таких як:

- розрахунок точності стикування;
- кореляція між шкалою стикування та експериментальною реакцією або визначенням коефіцієнта збагачення;
- відстань між іонзв'язуючою частиною та іоном в активному центрі;
- наявність індукційних моделей.

Точність докінгу являє собою один захід для кількісної оцінки придатності програми процесу докінгу шляхом раціоналізації здатності прогнозувати правильну позу ліганду щодо експериментально спостережуваного.

Докінг-екрани також можна оцінити шляхом збагачення анотованих лігандів відомих сполучних речовин із великої бази даних передбачуваних незв’язуючих молекул. Таким чином, успіх докінг-екрану оцінюється за його здатність збагачувати невелику кількість відомих активних сполук у верхніх рядах екрану серед набагато більшої кількості молекул у базі даних. Площа під кривою робочої характеристики приймача широко використовується для оцінки його продуктивності.

Результати потрапляння з док-екранів піддаються фармакологічному підтвердженню. Тільки перспективні дослідження є переконливим доказом придатності методики для конкретної цілі.

Потенціал стикувальних програм відтворювати режими зв'язування, як це визначається рентгенівською кристалографією, можна оцінити за допомогою ряду наборів тестів стикування.

Для малих молекул існує кілька наборів базових даних для стикування та віртуального скринінгу, наприклад набір Astex Diverse, що складається з високоякісних рентгенівських кристалічних структур білка-ліганду або довідника корисних рецепторів для оцінки ефективності віртуального скринінгу.

Оцінка результатів докінгових програм щодо їхнього потенціалу відтворювати режими пептидного зв’язування може бути оцінена за оцінкою ефективності стикування та оцінки (LEADS-PEP).

За результатами докінгу можливо виявити, чи існує вдалий зв’язок молекули до заданого білка, та сформувати файл з інформацією про взаємне положення тіл у просторі, їх енергетичні рівні тощо. Ці дані можуть бути оцінкою докінгу, його вдалістю. За величиною показників фізичних характеристик можливо визначити, чи будуть на практиці ці з’єднання вдалими та почати лабораторні тестування.

Для проведення молекулярного докінгу використовуюсь такі програмні засоби, як:

- PyMol;
- BioPython;
- UCSF Chimera;
- PMV;
- Coot;
- CCP4mg;
- mmLib;
- VMD;
- MMTK.

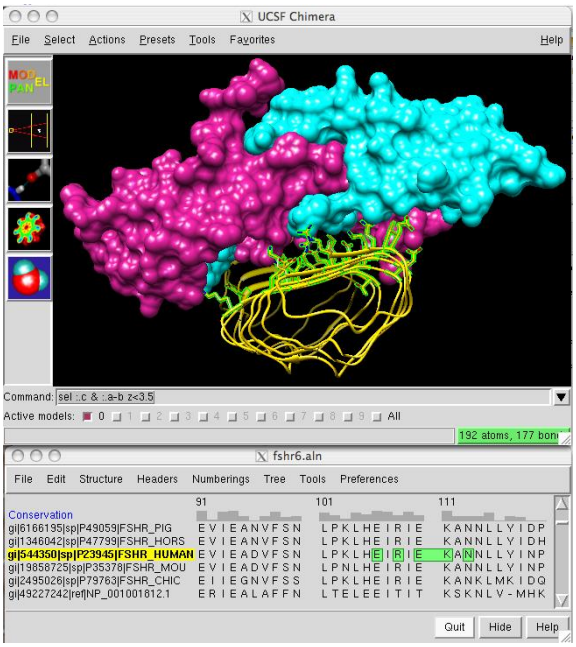


Рис. 1.2. Інтерфейс програми UCSF Chimera

Найбільш поширеною є програма Chimera, котра дає багато можливостей у моделюванні процесу програмного докінгу.[4][5]

UCSF Chimera (або просто Chimera) - це розширювана програма для інтерактивної візуалізації та аналізу молекулярних структур та супутніх даних, включаючи мапи щільності, надмолекулярні збірки, вирівнювання молекулярних ланцюгів, результати стикування, траєкторії та конформаційні положення.[6] Можна створювати високоякісні зображення та анімації. Chimera включає повну документацію та може бути завантажена безкоштовно для некомерційного використання.[7]

Chimera розробляється ресурсом з біообчислень, візуалізації та інформатики (RBVI) при Університеті Каліфорнії, Сан-Франциско.

Також важливим ресурсом є ZINC. ZINC - це колекція комерційно доступних хімічних сполук, приготованих спеціально для віртуального скринінгу. ZINC використовують дослідники (як правило, люди, які пройшли підготовку в якості біологів або хіміків) у фармацевтичних компаніях, біотехнологічних компаніях та дослідницьких університетах.[8][9]

ZINC відрізняється від інших хімічних баз даних, оскільки представляє біологічно значущу тривимірну форму молекули, котру можна ефективно використати для процесу молекулярного докінгу.[10][11]

РОЗДІЛ 2
ВИКОРИСТАННЯ MACHINE LEARNING ДЛЯ ПОШУКУ ОПТИМАЛЬНИХ ІНГІБІТОРІВ БІЛКА

2.1. Визначення machine learning

Machine Learning – це підрозділ штучного інтелекту, що вивчає методи будування алгоритмів, здібних до навчання. Алгоритми Machine Learning будують модель на основі зразкових (навчальних) даних для того, щоб робити прогнози або приймати рішення, не будучи явно запрограмованими для цього. Алгоритми Machine Learning використовуються в найрізноманітніших додатках, таких як фільтрація даних та комп'ютерне бачення, де важко або нездійсненно розробити звичайні алгоритми для виконання необхідних завдань.[11]

Підмножина машинного навчання тісно пов’язана з обчислювальною статистикою, яка фокусується на прогнозуванні за допомогою комп’ютерів. Вивчення математичної оптимізації доставляє методи, теорію та сфери застосування до галузі машинного навчання. Видобуток даних - це суміжна галузь дослідження, зосереджена на аналізі дослідницьких даних за допомогою безконтрольного навчання.

Під математичною оптимізацією мається на увазі вибір найкращого елемента (з урахуванням певного критерію) з деякого набору доступних альтернатив. Своєрідні оптимізаційні проблеми виникають у всіх кількісних дисциплінах - від обчислювальної техніки та техніки до досліджень операцій та економіки.

У найпростішому випадку задача оптимізації полягає в максимізації або мінімізації реальної функції шляхом систематичного вибору вхідних значень з дозволеного набору та обчислення значення функції. Узагальнення теорії та техніки оптимізації на інші формулювання становить велику область прикладної математики. Більш загально, оптимізація включає пошук

найкращих доступних значень деякої цільової функції з урахуванням певного простору, включаючи безліч різних типів цільових функцій та різних типів просторів.

Машинне навчання передбачає виявлення комп’ютерами, як вони можуть виконувати завдання, не будучи явно запрограмованими для цього. Це передбачає вивчення комп’ютерів на основі даних, що дозволяють виконувати певні завдання. Для простих завдань можна запрограмувати алгоритми, що повідомляють машині, як виконувати всі кроки, необхідні для вирішення проблеми. З боку комп'ютера навчання не потрібно. Для більш просунутих завдань для людини може бути складно вручну створювати необхідні алгоритми. На практиці більш ефективним є допомога машині у розробці власного алгоритму.

Дисципліна машинного навчання використовує різні підходи для навчання комп’ютерів виконанню завдань, коли не існує повністю задовільного алгоритму. У випадках, коли існує величезна кількість потенційних відповідей, одним із підходів є позначення деяких правильних відповідей як дійсних. Потім це використовують як навчальні дані для комп’ютера для вдосконалення алгоритму, який він використовує для визначення правильних відповідей.

Для молекулярного докінгу іноді використовують спрощений геометричний докінг. Процес виявлення оптимальних конформацій рецепторів виявляють за допомогою фіксування рецептору у просторі та накладання на нього ліганд. Такий підхід спрощує моделювання, але не враховує деякі фізичні властивості, котрі необхідно реалізувати окремо.

Машинне навчання також має тісні зв'язки з оптимізацією: багато проблем формулюються як мінімізація певної функції збитків. Функції збитків виражають невідповідність між передбаченнями моделі, що навчається, та фактичними екземплярами проблеми. Різниця між двома полями випливає з мети узагальнення: хоча алгоритми оптимізації можуть

мінімізувати втрати на навчальному наборі, машинне навчання займається мінімізацією втрат на невидимих зразках.[12]

2.2. Класифікація методів машинного навчання

Через велику кількість задач різних типів у машинному навчанні є декілька способів (методів) реалізації: навчання з вчителем, навчання без вчителя, навчання з частковим залученням та навчання з підкріпленням (рис.2.1).

Рис.2.1. Методи машинного навчання

Навчання з вчителем - це завдання машинного навчання у вивченні функції, яка відображає вхідні дані на виході на основі прикладів пар вхід-вихід. Воно виводить функцію із позначених навчальних даних, що складаються з набору навчальних прикладів. У навчанні з вчителем кожен приклад являє собою пару, що складається з вхідного об'єкта (як правило, вектора) та бажаного вихідного значення (контрольного сигналу). Цей метод навчання аналізує навчальні дані та виробляє виведену функцію, яка може бути використана для відображення нових прикладів. Оптимальний алгоритм дозволить правильно визначати мітки класів для невидимих екземплярів. Для

цього потрібно, щоб алгоритм навчання узагальнював дані до невидимих ситуацій коректним способом.[13]

Навчання з частковим залученням вчителя - це підхід до машинного навчання, який поєднує невелику кількість позначених даних із великою кількістю непозначених даних під час навчання. Навчання під наглядом поєднує в собі навчанням без нагляду (без позначених навчальних даних) та контрольованим навчанням (лише з позначеними даними про навчання.

Непозначені дані, коли використовуються разом із невеликою кількістю позначених даних, можуть значно покращити точність навчання. Для отримання позначених даних для навчальної проблеми часто потрібен кваліфікований людський агент або фізичний експеримент. Таким чином, витрати, пов'язані з процесом позначення, можуть зробити неможливими великі, повністю позначені навчальні набори, тоді як отримання непозначених даних є відносно недорогим. У таких ситуаціях навчання з частковим залученням вчителя може мати велике практичне значення.

Навчання без вчителя – це один із методів машинного навчання, який шукає раніше не виявлені зразки у наборі даних без раніше існуючих позначок та з мінімальним наглядом людини. На відміну від навчання з вчителем, яке зазвичай використовує заздалегідь позначені людиною дані, навчання без вчителя дозволяє моделювати щільність ймовірності над вхідними даними. У цьому методі нейронна мережа намагається самостійно знайти кореляції в даних, витягуючи корисні ознаки і аналізуючи їх.

Двома основними засобами, що використовуються в навчанні без вчителя, є основний компонентний і кластерний аналіз. Кластерний аналіз використовується при навчанні без вчителя для групування наборів даних із загальними атрибутами для екстраполяції алгоритмічних зв’язків. Кластерний аналіз - це розділ машинного навчання, який групує дані, які не були позначені чи класифіковані. Кластерний аналіз визначає спільність даних і реагує на основі наявності або відсутності таких спільних ознак у

кожному новому фрагменті даних. Цей підхід допомагає виявити аномальні точки даних, які не входять до жодної групи.

Навчання з підкріпленням – це метод машинного навчання, який стосується того, як програмні компоненти повинні вживати міри в середовищі, щоб максимізувати можливу винагороду. При прийнятті рішення вивчаються зворотний зв'язок, нові тактики і рішення, котрі здатні привести до більшого виграшу. Цей підхід використовує довгострокову стратегію - так само як в шахах, де наступний найкращий хід може не допомогти виграти в кінцевому рахунку.

Навчання з підкріпленням відрізняється від навчання з вчителем тим, що йому не потрібні подані позначені пари введення та виведення та відсутністю необхідності явно коригувати неоптимальні дії. Натомість основна увага приділяється пошуку балансу між дослідженням даних та їх експлуатацією.[14]

2.3. Моделі машинного навчання

Використання машинного навчання передбачає створення моделі, яка навчається на деяких навчальних даних і після цього може обробляти додаткові дані для прогнозування.

Штучні нейронні мережі (далі ШНМ) - це обчислювальні системи, нечітко натхненні біологічними нейронними мережами, які складають мозок тварин. Такі системи вчаться виконувати завдання, розглядаючи приклади, як правило, без програмування будь-яких правил, що стосуються конкретних завдань.

ШНМ - це модель, заснована на сукупності з'єднаних одиниць або вузлів (рис.2.2), що називаються штучними нейронами, які моделюють нейрони біологічного мозку. Кожне з'єднання подібно синапсам біологічному мозку та може передавати інформацію від одного штучного нейрона до іншого. Штучний нейрон, який отримує сигнал, може його обробити, а потім сигналізувати про додаткові штучні нейрони, підключені до нього.

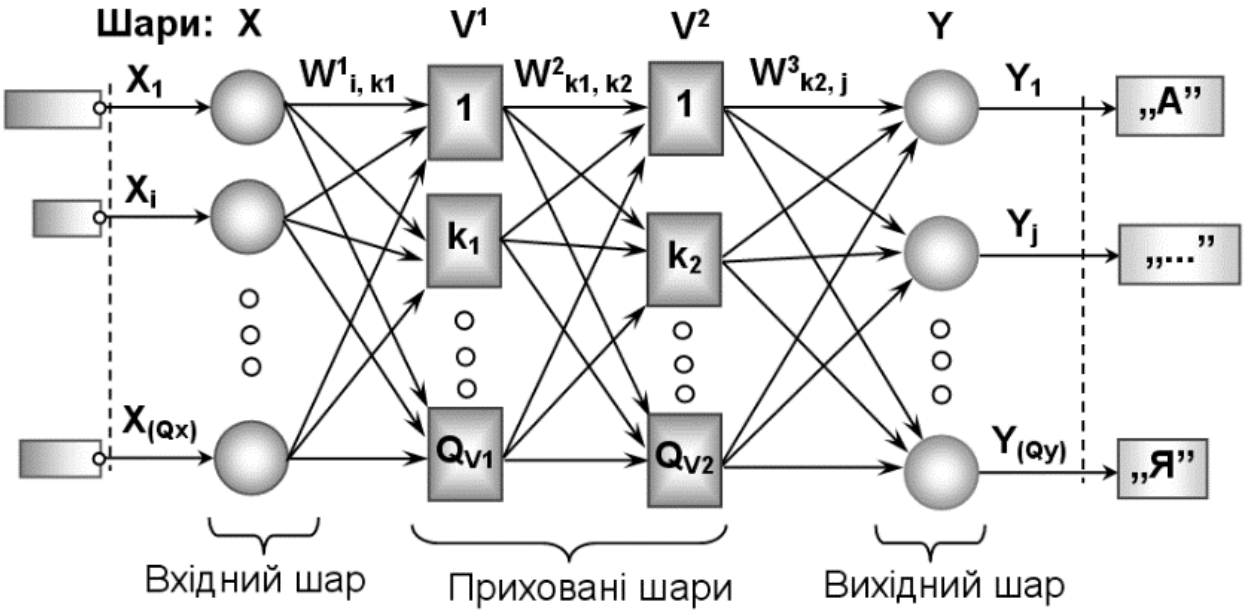


Рис.2.2. Схема устрою штучної нейронної мережі

У загальних реалізаціях ШНМ сигнал при з'єднанні між штучними нейронами є дійсним числом, і вихід кожного штучного нейрона обчислюється деякою нелінійною функцією суми його входів. Зв'язки між штучними нейронами називаються краями, границями. Штучні нейрони та краї зазвичай мають вагу, яка регулюється в процесі навчання. Вага збільшує або зменшує силу сигналу при з'єднанні. Штучні нейрони можуть мати такий поріг, що сигнал надсилається лише в тому випадку, якщо сукупний сигнал здатен перетнути поріг. Як правило, штучні нейрони агрегуються в шари. Різні шари можуть виконувати різні типи перетворень на своїх входах. Сигнали рухаються від першого, вхідного шару до останнього, вихідного шару.

Початковою метою підходу ШНМ було вирішення проблем таким самим чином, як це робив би мозок людини. Однак з часом увага переходила до виконання конкретних завдань, що призводило до відхилень від біології. Штучні нейронні мережі використовувались для різноманітних завдань, включаючи комп'ютерний зір, розпізнавання мови, машинний переклад, фільтрацію соціальних мереж, гру в настільні та відеоігри та медичну діагностику.

Дерева рішень (рис.2.3) використовують прогностичну модель для переходу від спостережень за предметом до висновків про цільову вартість елемента. Це один із підходів прогнозного моделювання, що використовується у статистиці, видобутку даних та машинному навчанні.

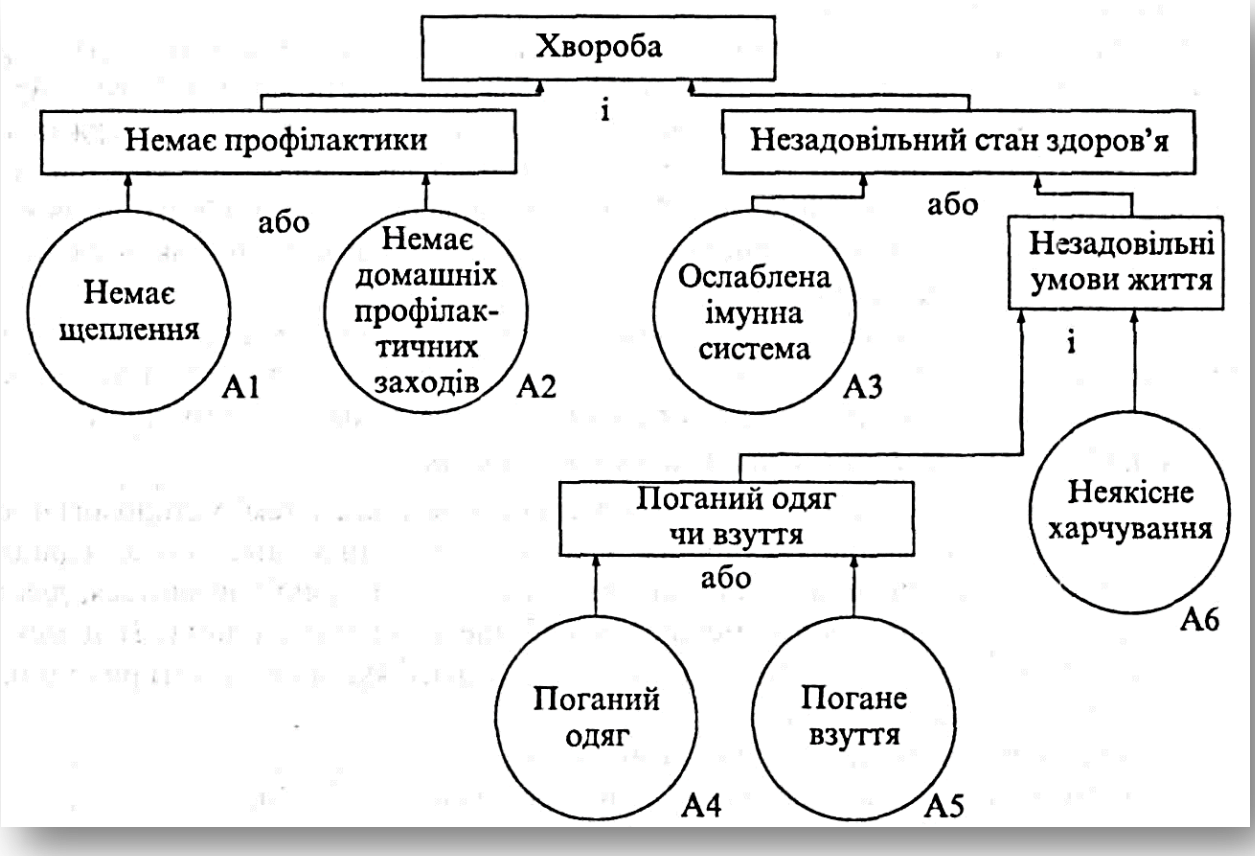


Рис.2.3. Дерево прийняття рішень для визначення ризику захворюваності

Моделі дерев, де цільова змінна може приймати дискретний набір значень, називаються деревами класифікації; у цих структурах листя представляють позначення класів, а гілки - сполучення ознак, які ведуть до цих позначень класів. При аналізі рішень дерево рішень може використовуватися для візуального та явного представлення рішень та прийняття рішень.

Регресійний аналіз охоплює велику різноманітність статистичних методів для оцінки взаємозв'язку між вхідними змінними та пов'язаними з

ними ознаками. Найбільш поширеною формою є лінійна регресія (рис.2.4), де проводиться одинарна лінія, яка найкраще відповідає даним згідно з математичними критеріями.

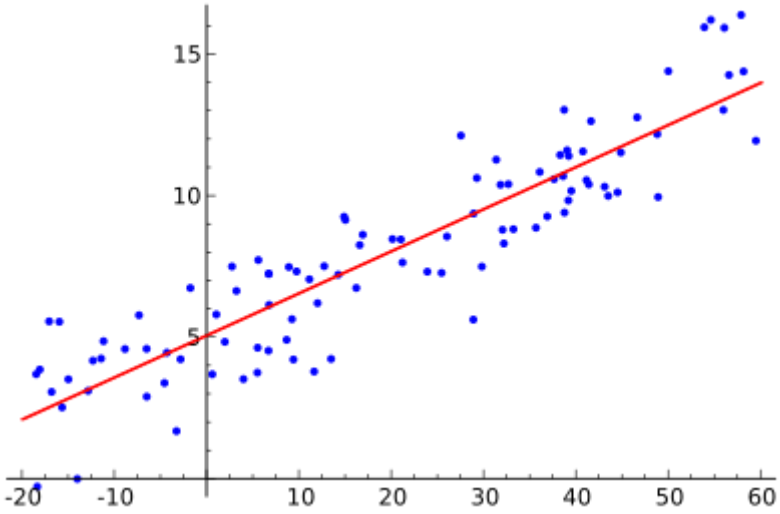


Рис.2.4.Лінійна регресія на визначеному наборі даних

Маючи справу з нелінійними проблемами, моделі переходу включають поліноміальну регресію, логістичну регресію або навіть регресію ядра, що вводить нелінійність, використовуючи переваги ядрових методів, щоб неявно зіставити вхідні змінні з багатовимірним простором.

Метод опорних векторів (далі МОВ) являє собою набір пов’язаних контрольованих методів навчання, що використовуються для класифікації та регресії. Алгоритм навчання МОВ будує модель, яка передбачає збіжності нових прикладів у відповідні категорії. МОВ є неімовірнісним, двійковим, лінійним класифікатором. МОВ можуть ефективно виконувати нелінійну класифікацію, неявно відображаючи свої входи у багатовимірні простори.

Залежно від початкового набору даних та кількості ресурсів для споживання висовують той чи інший метод навчання, найгнучкішими будуть ресурсовитратні методи, а найдешевшими – методи із заздалегідь заданою інформацією та набором даних.[15]

Для біохімічних процесів, де кількість інформації велика, не є оптимальним використання гнучких методів машинного навчання. На

офіціальному ресурсі ZINC15, де зібрано інформацію про винайдені ліки, буде обрано необхідну інформацію і за нею навчено нейронну мережу розпізнавати біологічно активні молекули.

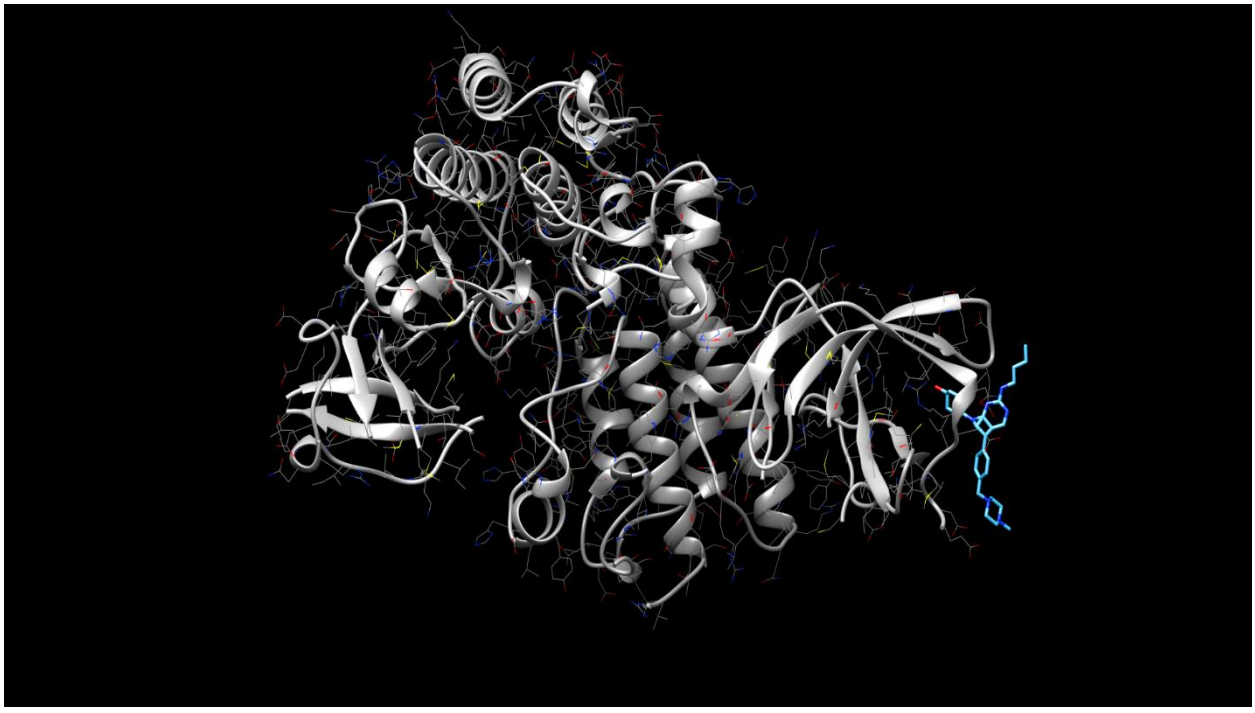
Через достатню кількість даних про вже існуючі молекули можливо також розробити просторове бачення для моделі нейронної мережі. На практиці для просторового розпізнавання потрібно для кожної потенційної точки стикування сканувати декілька положень, що буде проблемою з точки зору пам’яті.

Вирішено обирати потенційного кандидата-ліганду за допомогою характеристик, визначаючих біохімічні зв’язки. Після навчання нейронної мережі буде передано файл PDB зі структурою білка-рецептору, після чого занести масив кандидатів-ліганд.

РОЗДІЛ 3
СТВОРЕННЯ ПРОГРАМНИХ ЗАСОБІВ ПОШУКУ ОПТИМАЛЬНИХ БІОЛОГІЧНО АКТИВНИХ СПОЛУК

3.1. Створення програмних засобів для аналізу можливостей біохімічних процесів

За допомогою роботи програм SwissDock та UCSF Chimera було проведено молекулярний докінг заздалегідь відомої пари онкобілок-інгібітор (інгібуюча молекула) MER (MERTK) та UMC2025 (дані узяті з ресурсу ZINC15). За результатами була сформована тривимірна модель, на якій візуально представлені білок та інгібуюча молекула, їх фізичні властивості



(вільна енергія, взаємоположення у просторі ті інші) (див. рис. 3.1).[16]

Рис. 3.1. Докінг MER та UMC2025

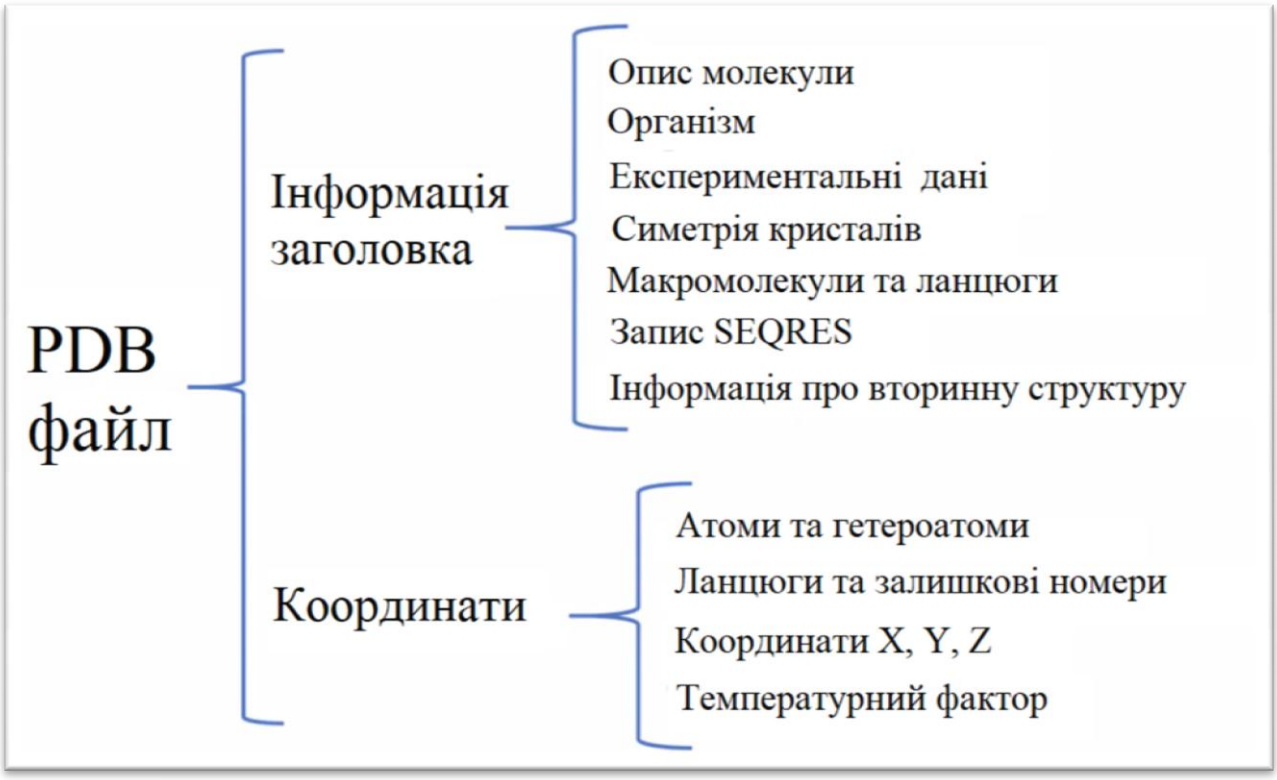


Рис. 3.2. Структура PDB файлу

На рисунку 3.2 зображено складові файлу формату PDB (Protein Data Bank). [17][18] Вони є основними керуючими елементами при заданні молекулярних сполук. За ними нейронна мережа буде отримувати необхідні дані та робити висновки.

У PDB файлах буде розглядатися наступне:

- Інформація залоговка:
  - Опис молекули (назва, опис, асиметричні одиниці);
  - Організм, у якому було досліджено дану сполуку;
  - Експериментальні дані (назва ти типологія походження, спосіб отримання, перелік авторів, ресурси з опублікованими результатами, якість, з якою зроблена модель та інше);

- Симетрія кристалів (має у собі опис операцій, котрі необхідні для формування на асиметричному органі для утворення кристалу);
- Макромолекули та ланцюги (опис наявності мутацій, наявність інших молекул);
- Записи SEQRES (ці записи містять у собі оригінальні послідовності білкових та нуклеїнових кислот; якщо деякі сегменти протеїну невпорядковані, координати атомів можуть не існувати для них. Але вони будуть перелічені серед записів SEQRES);
- Інформація про вторинну структуру;

- Координати:
  - Атоми та гетероатоми (HETATM містять у собі інформацію про розташування малих молекул, ATOM містять у собі записи протеїнів);
  - Ланцюги та залишкові номери (серійний номер атому);
  - Координати X, Y, Z;
  - Температурний фактор.

Для обробки PDB файлів було обрано бібліотеку BioPython, котра дає користувачу багато можливостей над маніпуляцією файлів такого типу.

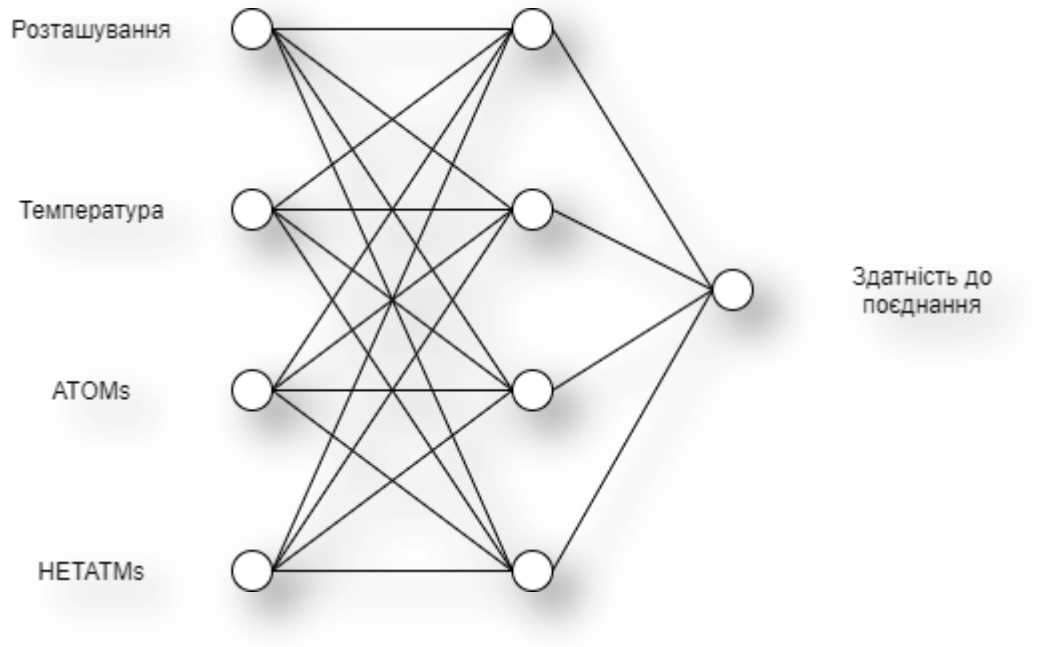
3.2. Створення штучної нейронної мережі

На основі розглянутих структур файлів використано мову Python та бібліотеки BioPython, numpy для обробки даних штучної нейронної мережі.

Нейронна має наступні складові:

1. Вхідний масив, отриманий із фізичних властивостей файлу PDB за допомогою засобів бібліотеки BioPython (розглядатися будуть лише наступні фізичні властивості: розташування атомів, температурний фактор, ATOMs та HETATMs);

2. Прихований шар з кількістю нейронів, котра дорівнює кількості розглянутих фізичних властивостей;
3. Шар виводу інформації про можливість виникнення інгібуючої



реакції.

Рис. 3.3. Граф нейронної мережі

На мові програмування Python з бібліотекою BioPython було виокремлено необхідні дані для передачі на вхідний шар нейронної мережі. На основі отриманих даних з ресурсу RCSB PDB були сформовані тестові дані сполук із заздалегідь відомими інгібіторами у формати .mol. У ресурсі PubChem було знайдено необхідні речовини, що утворюють біологічно активний комплекс. За допомогою сервісу SwissDock був проведений докінг декількох сполук з онкобілком MER: деякі з них були інгібуючими молекулами, інші – ні. На основі отриманих фізичних властивостей цілісних файлів PDB та часткових елементів кожного з них нейронна мережа повинна знайти закономірність, за якою відбувається інгібування заданої молекули.[19][20]

До нейронної мережі було передано в якості вхідного масиву даних файли mol та PDB, за допомогою парсера котрі мають необхідний вигляд. Далі, нейронна мережа проходить навчання за заданою вибіркою (рис.3.4-3.5).

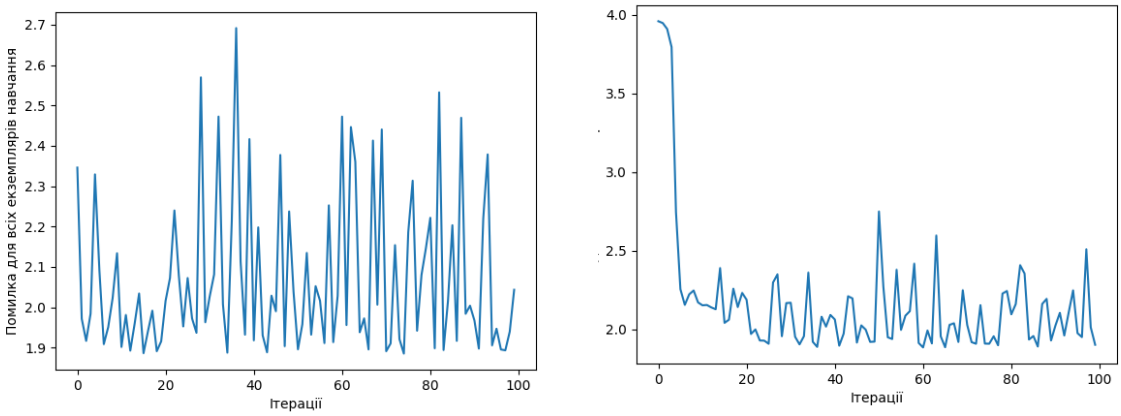


Рис.3.4-3.5. Відображення помилок для екземпляра нейронної мережі (до та після навчання)

Таким чином, нейронна мережа може з деякою вірогідністю передбачити, чи буде заданий молекулярний комплекс біологічно активною структурою (зокрема для білка MER) без необхідності проводити молекулярний докінг.

ВИСНОВОК

У процесі виконання кваліфікаційної роботи було проаналізовано дані молекулярного докінгу та нейронних мереж, можливостей їх взаємного використання для підвищення ефективності розробки нових ліків. На основі отриманої інформації було розроблено програмні засоби, необхідні для забезпечення виявлення потенційних молекул інгібіторів.

У першій частині кваліфікаційної роботи розглянуті основи теоретичного докінгу: процес з’єднання молекул у просторі, основні терміни для роботи з даними результатів докінгу, програми для реалізації стикування, просторово відображено можливі проблеми, котрі можуть виникати і є типовими для розглянутого типу білка рецептору.

У другій частині кваліфікаційної роботи розглянуто структуру нейронної мережі, можливості її навчання та способи взаємодії з молекулярним докінгом та його відомими результатами.

У третій частині кваліфікаційної роботи були проведені заходи щодо підготовки та проведення молекулярного докінгу деякої вибірки комерційно відомих сполук ліків. На основі цього було сформовано нейронну мережу, що може з деякою вірогідністю передбачити, чи є дана до нього молекула зв’язна.

Результати кваліфікаційної роботи можуть бути використані при реалізації програмних засобів виявлення потенційних молекул-інгібіторів, після чого передані медичним працівникам.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Гюнтер Х. Введення в курс спектроскопії ЯМР. – Пер. с англ. – М., 1984.
2. Габуда С. П., Плетнев Р. Н., Федотов М. А. Ядерний магнітний резонанс в неорганічній хімії. – М: Наука. – 1988.- 214 с.
3. Молекулярна біологія. Структура та біосинтез нуклеїнових кислот / Під ред. А.С. Спіріна. М.; Старша школа. 1990. – 352 с.
4. Goddard, T. D.; Huang, C. C.; Meng, E. C.; Pettersen, E. F.; Couch, G. S.; Morris, J. H.; Ferrin, T. E. (2017). "UCSF chimerax: meeting modern challenges in visualization and analysis". Protein Science. 27 (1): 14–25.
5. Hauser AS, Windshügel B (Dec 2015). "A Benchmark Data Set for Assessment of Peptide Docking Performance". Journal of Chemical Information and Modeling.: 361 с.
6. Chapman, Brad; Chang, Jeff (August 2000). "Biopython: Python tools for computational biology". ACM SIGBIO Newsletter.: 152 с.
7. UCSF Chimera Documentation [Електронний ресурс] / Режим доступу: https://www.cgl.ucsf.edu/chimera/docindex.html - Дата доступу 06.04.2021.
8. RCSB PDB-101 [Електронний ресурс] / Режим доступу: https://pdb101.rcsb.org/ - Дата доступу – 08.04.2021.
9. PubChem [Електронний ресурс] / Режим доступу: https://pubchem.ncbi.nlm.nih.gov/ - Дата доступу – 05.04.2021.
10. RCSB PDB: Data API Documentation [Електронний ресурс] / Режим доступу: https://data.rcsb.org/#rest-api – Дата доступу – 08.04.2021.
11. ZINC15 [Електронний ресурс] / Режим доступу: https://zinc15.docking.org/ – Дата доступу – 08.04.2021.
12. RealPython [Електронний ресурс] / Режим доступу: https://realpython.com/python-ai-neural-network/#vectors-and-weights. – Дата доступу 23.02.2021.
13. TensorFlow [Електронний ресурс] / Режим доступу: https://playground.tensorflow.org - Дата доступу – 08.04.2021.

14. SpringerLink [Електронний ресурс] / Режим доступу: https://link.springer.com/referenceworkentry - Дата доступу - 01.03.2021.

15. Towards Data Science [Електронний ресурс] / Режим доступу: https://towardsdatascience.com – Дата доступу 01.04.2021.

16. Protein Data Bank [Електронний ресурс] / Режим доступу: https://www.rcsb.org – Дата доступу – 05.04.2021.

17. BioPython Documentation [Електронний ресурс] / Режим доступу: https://biopython.org/wiki/Documentation – Дата доступу - 06.04.2021.

18. NumPy Documentation [Електронний ресурс] / Режим доступу: https://numpy.org/doc/ - Дата доступу – 06.04.2021.

19. Journal of biological chemistry [Електронний ресурс] / Режим доступу: https://www.jbc.org/article/S0021-9258(20)34445-8/fulltext - Дата доступу - 06.04.2021.

20. Python Documentation [Електронний ресурс] / режим доступу: https://docs.python.org/3/library/parser.html - Дата доступу - 07.04.2021.