

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«Київський політехнічний інститут імені Ігоря Сікорського»

Фізико-технічний інститут
Кафедра фізики енергетичних систем

«На правах рукопису»

УДК 544.112, 004.896

«До захисту допущено»

Завідувач кафедри

_____ А. А.Халатов

(підпис)

(ініціали, прізвище)

Магістерська дисертація

зі спеціальності 105 Прикладна фізика та наноматеріали

на тему: **Моделювання молекулярної структури комплексу білка з лігандом**

Виконав: студент 6 курсу, групи ФФ-72м

_____ Докашенко Богдан Вікторович _____.

(прізвище, ім'я, по батькові)

Науковий керівник доцент, к.ф.-м.н. Пономаренко С. М.

(посада, науковий ступінь, вчене звання, прізвище та ініціали)

Рецензент Старший науковий співробітник Інституту фізики

НАН України, к.ф.-м.н. Бойко О.П. _____.

(посада, науковий ступінь, вчене звання, прізвище та ініціали)

Засвідчую, що у цій магістерській
дисертації немає запозичень з праць
інших авторів без відповідних посилань.

Студент _____

(підпис)

Національний Технічний Університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Фізико-технічний інститут

Кафедра фізики енергетичних систем

Рівень вищої освіти – другий (магістерський) за освітньо-професійною програмою

Спеціальність 105 Прикладна фізика та наноматеріали

«ЗАТВЕРДЖЕНО»

Завідувач кафедри

_____ **Халатов. А. А.** _____

(підпис)

(ініціали, прізвище)

_____ 2018 р.

ЗАВДАННЯ

на магістерську дисертацію студенту

Докашенко Богдана Вікторовича

1. Тема дисертації: «Математичне та програмне забезпечення моделювання молекулярного докінгу протеїнів»,
науковий керівник дисертації Пономаренко С. М., доцент, к.ф-м.н.
затверджені наказом по університету _____.
2. Термін подання студентом дисертації: «__» _____ 2018 р.
3. Вхідними даними для дисертації є база даних сполук RCSB, молекули ліганду, молекули білків, нейронні мережі. Вихідними є система здатня моделювати комплекс білка з лігандом, за короткий термін ніж це можливо зараз (декілька годин), за достатньою точністю (середнє квадратичне відхилення не більше 3).
4. Перелік завдань, які потрібно розробити: ознайомитися з існуючими системами, які займаються молекулярним докінгом, виділити їх переваги та недоліки, зробити аналіз характеристик цих систем; ознайомитися з існуючими методами молекулярного докінгу, виділити їх переваги та недоліки, зробити

аналіз характеристик цих методів; ознайомитися з методами машинного навчання, проаналізувати їх, обрати методи, які підійдуть для молекулярного докінгу; обрати найкращий метод та за його допомогою розробити математичну модель докінгу; порівняти результати запропонованої моделі із результатами існуючих моделей; проаналізувати отримані результати.

5. Перелік ілюстративного матеріалу: концептуальна модель мережі, ілюстраційна модель процесу молекулярного докінгу, візуальне представлення алгоритму розробленої системи молекулярного докінгу.

6. Дата видачі завдання: «__» _____ 2017 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Грунтовне ознайомлення з предметною областю		
2	Ознайомлення з існуючими алгоритмами та методами молекулярного докінгу, їх аналіз.		
3	Ознайомлення з існуючими рішеннями для моделювання молекулярного докінгу, їх аналіз.		
4	Дослідження області машинного навчання, аналіз існуючих методів, вибір необхідного методу для поставленої задачі.		

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
5	Розробка математичної моделі моделювання молекулярного докінгу за допомогою методів машинного навчання.		
6	Розробка архітектури програмного забезпечення для моделювання молекулярного докінгу		
7	Аналіз та порівняння отриманих результатів.		
8	Висновки до розробленої роботи. Публікація		

Студент

Докашенко Б. В.

Науковий керівник дисертації

Пономаренко С. М.

РЕФЕРАТ

Дисертацію виконано на 73 аркушах, вона містить перелік посилань на використані джерела з 29 найменувань. У роботі наведено 24 рисунки та 3 таблиці.

Актуальність теми. Задача молекулярного моделювання є невід'ємною частиною фундаментальних досліджень, спрямованих на вивчення механізмів функціонування білків, а також прикладних фармацевтичних задач, таких як створення нових лікарських сполук, отримання високопродуктивних ферментів, комп'ютерний аналіз молекулярних взаємодій тощо. Докінг також використовують в процесі віртуального високопродуктивного скринінгу (сканування) баз даних, який значно знижує витрати проектів, спрямованих на пошук нових ефективних і селективних лігандів. Взаємодія між молекулами є основою для біологічних процесів. Використовуючи ці взаємодії, живі організми підтримують складні регуляторні та метаболічні механізми. Експериментальна дослідницька робота є основою для кращого розуміння біологічних процесів. Комп'ютерне моделювання є важливим для того, щоб отримати інформацію від великої кількості багатоваріантних експериментальних даних. Внаслідок цього, розвиток і оптимізація алгоритмів докінгу є на даний момент областю активних наукових розробок.

Мета і задачі дослідження. Метою дисертаційної роботи є спрощення та пришвидшення процесу моделювання молекулярного докінгу, автоматизація процесу, підвищення точності, покращення методів створення молекулярних сполук задля фармацевтичних цілей.

Для досягнення вказаної мети було розв'язано такі задачі:

- ознайомитися з існуючими системами, які займаються молекулярним докінгом, виділити їх переваги та недоліки, зробити аналіз характеристик цих систем;

- ознайомитися з існуючими методами молекулярного докінгу, виділити їх переваги та недоліки, зробити аналіз характеристик цих методів;
- ознайомитися з методами машинного навчання, проаналізувати їх, обрати методи, які підійдуть для молекулярного докінгу;
- обрати найкращий метод та за його допомогою розробити математичну модель докінгу;
- порівняти результати запропонованої моделі із результатами існуючих моделей;
- проаналізувати отримані результати.

Об'єктом дослідження є алгоритми, математичні моделі, біохімічні моделі докінгу, молекули ліганд, молекули протеїнів, взаємозв'язки між молекулами, нейронні мережі, генеративно-зв'язкові нейронні мережі, Backpropagation нейронні мережі, програми для візуалізації та докінгу молекул (Open Babel, AutoDock), python бібліотеки для роботи з молекулами.

Предметом дослідження є математичне та програмне забезпечення моделювання докінгу молекул протеїнів за допомогою нейромережі.

Методи дослідження. Для розв'язання поставленої задачі використовувалися такі методи: методи машинного навчання (для розробки прогнозуючої моделі молекулярного докінгу), методи оптимізації (для знаходження оптимальних фільтрів та вагів при навчанні нейронних мереж), методи фільтрації (для розмиття координат розташування атомів), конформаційні методи (для зміни розташування атомів в молекулі).

Наукова новизна одержаних результатів: уперше поставлену задачу моделювання молекулярного докінгу було успішно виконано за допомогою нейромережі; розроблена система моделювання молекулярного докінгу в порівнянні із стандартними біохімічними системи працює в декілька разів швидше та при цьому дає майже таку саму, а іноді і ліпшу точність. Крім того, при використанні стандартних підходів дослідник повинен володіти глибокими

знаннями з молекулярного докінгу та обирати серед різних методів та настроювати багато параметрів. При застосуванні розробленої системи цього не потрібно.

Практичне значення одержаних результатів. Розроблені методи, математичне та програмне забезпечення для моделювання молекулярного докінгу дозволяє отримати спрощення та пришвидшення процесу моделювання молекулярного докінгу, автоматизацію процесу, підвищення точності, покращення методів створення молекулярних сполук.

Публікації. Результати дисертації викладено в публікації та тезах доповіді у Моделювання молекулярної структури комплексу білка з лігандом за допомогою нейромережі//Наука онлайн: Міжнародний електронний науковий журнал - 2018. - №11 5-6.

Ключові слова: *молекулярний докінг, протеїн, ліганд, конформації, машинне навчання, генеративно-змагальна мережа, метод зворотного поширення помилки, конволюційні нейронні мережі, розвернуті нейронні мережі.*

ABSTRACT

The thesis is presented in 73 pages. It contains bibliography of 29 references. Twenty-four figures and three tables are given in the thesis.

Topic relevance. The task of molecular modeling is an integral part of fundamental research aimed at studying the mechanisms of protein function, applied pharmaceutical tasks such as the creation of new drug compounds, the production of high-performance enzymes, computer analysis of molecular interactions, etc. Docking is also used in the process of virtual high-performance screening (scanning) databases, which significantly reduces the cost of projects aimed at finding new effective and selective ligands. Interaction between molecules is the basis for biological processes. Using these interactions, living organisms support complex regulatory and metabolic mechanisms. Experimental research work is the basis for a better understanding of biological processes. Computer simulations are important in order to obtain information from a large number of multivariate experimental data. As a result, the development and optimization of docking algorithms is at present the area of active scientific development.

Research goal and objectives. The aim of the dissertation is to simplify and accelerate the molecular docking simulation process, to automate the process, to improve accuracy and to improve the methods for the creation of molecular compounds for pharmaceutical purposes.

To accomplish this goal, the following tasks were solved:

- systematize existing systems which deal with molecular docking, to highlight their advantages and disadvantages and to analyze the characteristics of these systems;
- systematize existing methods of molecular docking, to highlight their advantages and disadvantages, to analyze the characteristics of these methods;
- to get familiar with the methods of machine learning, to analyze them, to choose methods which are suitable for molecular docking;

- choose the best method and use it to develop a mathematical and computer model of docking;
- compare the results of the proposed model with the results of existing models;
- analyze the results.

Object of research is algorithms, mathematical models, biochemical docking models, ligand molecules, protein molecules, interactions between molecules, neural networks, generic-connected neural networks, back propagation neural networks, programs for visualization and doping of molecules (Open Babel , AutoDock), python libraries for working with molecules.

Subject of research is the mathematical and software modelling of the doping of protein molecules using a generic-competitive network.

Methods of research. To solve the task, the following methods were used: the methods of machine learning (for the development of the prediction model of molecular docking), optimization methods (for finding optimal filters and weights for training the neural networks), filtration methods (for blurring atomic location coordinates), conformational methods (to change the location of atoms in the molecule).

Scientific contribution consists of the following: for the first time the problem of molecular docking simulation was performed using generative-adversarial network; the developed system of molecular docking modelling operates several times faster and gives almost the same and sometimes better accuracy in comparison with standard biochemical systems. In addition, when using standard approaches, the researcher must have deep knowledge of molecular docking and choose between different methods and adjust many parameters. When applying the developed system, it is not required.

Practical value of obtained results. The developed methods, mathematical and software for modelling the molecular docking allows to get simplification and acceleration of molecular docking modelling process, to automate the process, increase the accuracy, to improve methods of molecular compounds creation.

Publications The results of the dissertation are presented in the publication and abstracts of the report in Modeling the molecular structure of the protein complex with the ligand using the neural network. // Science Online: International Electronic Journal of Journalism - 2018. - №11 5-6.

Keywords: *molecular docking, protein, ligand, conformation, machine learning, generic-adversarial network, method of reverse error propagation, convolutional neural networks, deployed neural network.*

ЗМІСТ

ПЕРЕЛІК ПОЗНАЧОК, СКОРОЧЕНЬ, СИМВОЛІВ І СПЕЦІАЛЬНИХ ТЕРМІНІВ	13
ВСТУП.....	14
РОЗДІЛ 1.	17
ЗАГАЛЬНІ ВІДОМОСТІ ТА СУЧАСНИЙ СТАН ПРОБЛЕМИ	17
1.1.1 Оціночні функції	17
1.1.2 Алгоритм конформаційного пошуку.....	22
1.1.3 Водневі зв'язки.....	26
1.1.4 Гідрофобні взаємодії.....	28
1.1.5 Оцінки вільної енергії гідрофобних взаємодій.	28
1.1.6 Стекінг	30
1.1.7 Консенсусний підхід	31
1.1.8 Опис існуючих технічних рішень.....	33
1.2 Порівняння існуючих технічних рішень.....	35
1.3 Штучна нейронна мережа.....	37
1.3.1 Вхідний шар	42
1.3.2 Згортковий шар.....	42
1.3.3 Підвиборчий шар.....	46
1.3.4 Повнозв'язковий шар	47
1.3.5 Вибір функції активації	49
1.3.6 Перевернуті або розгорнуті нейромережі.....	51
1.3.7 Метод зворотного поширення помилок.....	53
Висновки до розділу 1.....	56
РОЗДІЛ 2.	58
РОЗРОБКА СИСТЕМИ ТА ЇЇ ДЕТАЛІ.....	58
2.1 Опис алгоритму роботи розробленої системи.....	58
2.2 Навчання системи.....	62

Висновки до розділу 2.....	63
РОЗДІЛ 3.	64
АНАЛІЗ ОТРИМАНИХ РЕЗУЛЬТАТІВ	64
3.1 Порівняння з існуючими системами	64
3.2 Оцінка точності.....	65
Висновки до розділу 3.....	69
ВИСНОВКИ.....	70
СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ	71

ПЕРЕЛІК ПОЗНАЧОК, СКОРОЧЕНЬ, СИМВОЛІВ І СПЕЦІАЛЬНИХ ТЕРМІНІВ

GAN – (Generative adversarial network) генеративно-змагальна мережа, що реалізована системою двох штучних нейронних мереж, які змагаються одна з одною в рамках гри з нульовою сумою та використовуються в навчанні без вчителя.

Ліганд – молекула, що оборотно з'єднана з білком нековалентним зв'язком. Ліганд може бути будь-якою молекулою, що формує комплекс з білком.

PDB – (Protein Data Bank) база даних трьохвимірних структур білків, а також формат файлів, які містять набір атомів будь-яких молекул в стандартному представленні.

Молекулярний докінг – це метод молекулярного моделювання, метою якого є пошук та передбачення найбільш вірогідного та достовірного розташування та конформації молекул, що формують комплекс з білком, по відношенню до молекул протеїну для утворення нових стійких сполук.

Backpropagation - ітеративний градієнтний алгоритм, який використовується з метою мінімізації помилки роботи багатошарового перцептрону та отримання бажаного виходу.

ПЗ – програмне забезпечення.

ВСТУП

В даний час методи комп'ютерного молекулярного моделювання стають невід'ємною частиною фундаментальних досліджень, спрямованих на вивчення молекулярних механізмів функціонування білків, а також і прикладних проектів, пов'язаних з раціональним дизайном нових лікарських сполук. Метод молекулярного моделювання, метою якого є пошук найбільш достовірної орієнтації і конформації молекули ліганда в центрі зв'язування білка-мішені, називається молекулярним докінгом. Молекулярний докінг дозволяє передбачати просторову структуру комплексу рецептор-ліганд і вільну енергію його утворень, виходячи з даних про просторову структуру рецептора, відомої з дозволом в кілька ангстрем (наприклад, отриманої за допомогою рентгеноструктурного аналізу), і хімічною структурою ліганду. Переваги цього методу очевидні: встановлення ключових амінокислотних залишків в активному центрі білка, що дозволяє вивчати структурно-динамічні основи ферментативних реакцій на атомному рівні; раціональний дизайн лігандів і/або рецепторів з наперед заданими селективністю, кінетичними властивостями і т.і.

Підхід молекулярного докінгу може бути використаний для моделювання взаємодії між невеликою молекулою та білком на атомному рівні, що дозволяє характеризувати поведінку малих молекул у місці зв'язування цільових білків, а також висвітлювати фундаментальні біохімічні процеси [1]. Процес стикування передбачає два основні етапи: передбачення конформації ліганда, що передбачає знаходження його положення та орієнтації в цих ділянках та оцінку спорідненості зв'язків. Задача молекулярного моделювання є невід'ємною частиною фундаментальних досліджень, спрямованих на вивчення механізмів функціонування білків, а також прикладних фармацевтичних задач, таких як створення нових лікарських сполук, отримання високопродуктивних ферментів, комп'ютерний аналіз молекулярних взаємодій тощо. Докінг також використовують в процесі віртуального високопродуктивного скринінгу

(сканування) баз даних, який значно знижує витрати проектів, спрямованих на пошук нових ефективних і селективних лігандів. Взаємодія між молекулами є основою для біологічних процесів. Використовуючи ці взаємодії, живі організми підтримують складні регуляторні та метаболічні механізми. Експериментальна дослідницька робота є основою для кращого розуміння біологічних процесів. Комп'ютерне моделювання є важливим для того, щоб отримати інформацію від великої кількості багатоваріантних експериментальних даних. Внаслідок цього, розвиток і оптимізація алгоритмів докінгу є на даний момент областю активних наукових розробок.

Через те, що експериментальні методи визначення структури молекулярних комплексів є дорогими, трудомісткими і не завжди досяжними, для прогнозування нових молекулярних сполук краще використовувати обчислювальні методи.

Протягом останніх років в літературі з'явилася велика кількість досліджень, орієнтованих на порівняння різних алгоритмів докінгу і використовуваних ними оціночних функцій між собою [1-5]. Такі дослідження покликані допомогти вибрати з численних програм докінгу найбільш підходящу для певного класу з'єднань або для біомолекулярних систем зі специфічними властивостями. Нажаль, єдиного універсального методу для вирішення даної задачі на сьогодні не існує. При моделюванні процесу докінгу кожен раз використовують різні набори методів, які не гарантують успіху. але й досі не визначено остаточного підходу або алгоритму, який вважався би найефективнішим та найточнішим. Крім того, результати теоретичних досліджень досить часто не співпадають з експериментальними результатами та майже всі програми для молекулярного докінгу працюють досить довго. Тож, знаходження оптимального, ефективного, швидкого та точного методу докінгу залишається відкритим питанням, а сам процес докінгу є областю активних наукових розробок.

Таким чином, через всі ці складнощі під час молекулярного докінгу за допомогою біохімічних методів було запропоновано проаналізувати методи машинного навчання та реалізувати математичне та програмне забезпечення молекулярного докінгу за допомогою методів машинного навчання яке б зуміло би подолати існуючі проблеми подібних систем такі як швидкість роботи, оцінка результатів та можливо навіть точність.

РОЗДІЛ 1.

ЗАГАЛЬНІ ВІДОМОСТІ ТА СУЧАСНИЙ СТАН ПРОБЛЕМИ

1.1.1 Оціночні функції

У загальному випадку тут можна виділити дві незалежні (хоча і взаємопов'язані) складові процедури докингу - алгоритм конформаційного пошуку та оціночну функцію (ОФ).

Існують два основні підходи до моделювання докингу: взаємозалежність форми та симуляція. При першому підході білок та ліганд описуються як додаткові поверхні та увага приділяється саме геометричній відповідності. При другому підході ліганд знаходить потрібне положення ітераційним шляхом. Кожна ітерація включає в себе переміщення, обертання та конформацію ліганд (конформаційний пошук), після чого заново обчислюється енергійна оцінка системи (оціночна функція). Зазвичай при молекулярному докингі використовують саме цей підхід, адже процес його роботи фізично ближчий до того, що відбувається насправді і зазвичай більш точний. Візуальне представлення процесу молекулярного докингі можна побачити на рисунку 1.1.

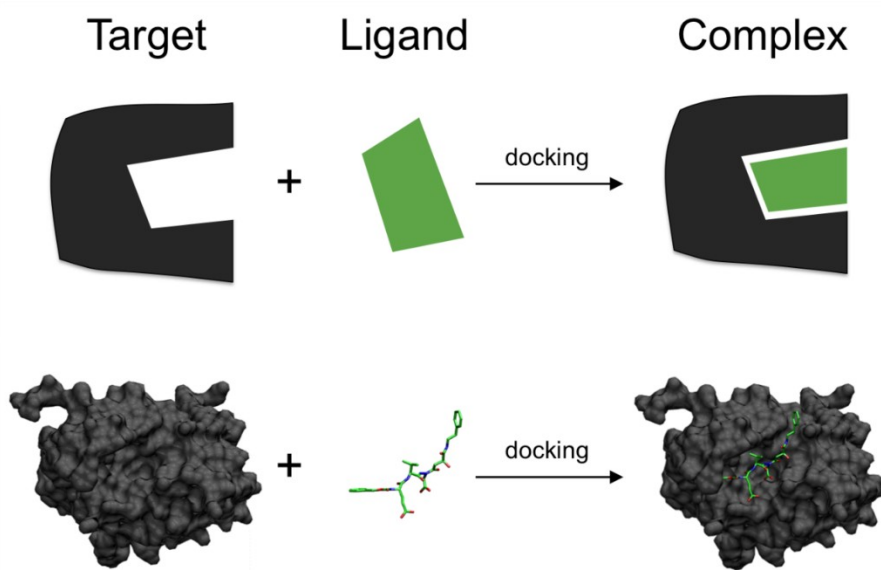


Рис. 1.1 – Процес молекулярного докингі

У класичному варіанті молекулярного докінгу завдання алгоритму конформаційного пошуку зводиться до перебору конформаційного простору комплексу за рахунок варіювання торсійних кутів ліганда і його переміщення в цілому відносно нерухомої структури білка-мішені. Сучасні алгоритми конформаційного пошуку в переважній більшості випадків знаходять конформації, близькі до експериментальних, за порівняно короткий час. Проте є фактори, що також впливають на успішність докінгу, які часто не враховуються в стандартних алгоритмах. Один з таких факторів - конформаційна рухливість білка-мішені, що в більшості випадків супроводжує зв'язування ліганда. Діапазон рухливості може бути різним - починаючи з невеликого «підстроювання» бічних ланцюгів і закінчуючи масштабними доменними рухами [5]. Ці рухи відіграють велику роль: якщо взяти структуру білка, оптимальну для зв'язування даного ліганду, то результат докінгу напевно буде більш точним, ніж якщо взяти будь-яку іншу (наприклад, апоформу). На перший погляд самим логічним вирішенням цієї проблеми є облік рухливості білка в програмі докінгу. На жаль, сучасні обчислювальні можливості не дозволяють проводити таке моделювання за прийнятний час, так як молекула білка дуже велика, і облік рухливості по всіх ступенях свободи може привести до так званого «комбінаторного вибуху» (астрономічному збільшенню числа можливих варіантів). Лише в деяких програмах передбачена обмежена рухливість сайтів зв'язування білка (як правило, на рівні невеликої адаптації конформацій бічних ланцюгів залишків активного центру). Інший підхід до цієї проблеми складається в докінгу в кілька різних конформацій одного і того самого білка з подальшим вибором кращих рішень з кожного запуску докінгу. Третій підхід - знайти деяку універсальну структуру білка-мішені, за участю якої докінг давав би досить хороші результати для різних класів лігандів. При цьому зменшується число «пропущених» (але правильних) рішень, однак також сильно зростає і число невірних варіантів [6,7].

Оціночні функції (ОФ), які використовуються в процесі докінгу, служать для обчислення приблизної енергії комплексів і ранжирування різних передбачуваних конформацій ліганда в сайті зв'язування на кожному кроці конформаційного пошуку (рис.1.2).

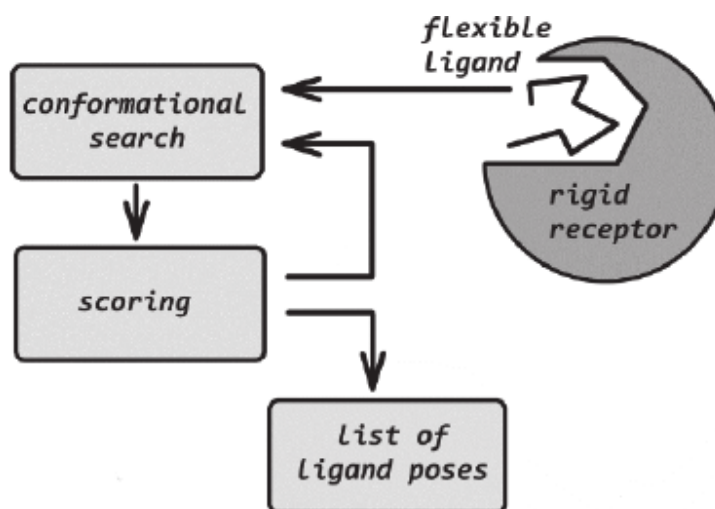


Рис. 1.2 – Схема процедури молекулярного докінгу

Але ОФ не завжди добре справляються з цим завданням, вносячи помилки при моделюванні структури комплексу рецептор-ліганд там, де вони повинні вказувати вірне рішення. Заздалегідь передбачаючи неточність в ОФ, зазвичай розглядають не єдину структуру комплексу, а цілий набір можливих варіантів, розраховуючи на те, що серед них виявиться правильний або хоча б близький до нього. З цього набору можна вибрати найбільш правдоподібний варіант, ґрунтуючись, наприклад, на відомі експериментальні дані про роль тих чи інших амінокислотних залишків активного центру розглянутого білка в зв'язуванні лігандів. Фактично, в даному випадку використовується додатковий фільтр, що оптимізує положення ліганда для даного конкретного білка-мішені [5-9].

ОФ показує, яка з орієнтацій ліганда в сайті зв'язування рецептора найбільш правдоподібна або, якщо порівнюються кілька різних лігандів, - який з них володіє найбільшою спорідненістю до білка-мішені. Зазвичай енергію

зв'язування ліганда з рецептором розкладають на окремі незалежні складові - терми, що відображають різні фізичні взаємодії. Лінійна комбінація цих термів і є ОФ. Всі ОФ, що застосовуються в сучасних алгоритмах докінгу, можна умовно розділити на три типи:

- засновані на силових полях,
- емпіричні,
- статистичні ОФ.

Детальний опис цих типів ОФ можна знайти далі.

Останнім часом емпіричні ОФ знайшли широке застосування в алгоритмах докінгу. Популярність використання емпіричних ОФ обумовлена тим, що з накопиченням все більшої кількості експериментальних даних про будову і біохімію молекулярних комплексів для них стає досить просто скласти навчальну вибірку і конструювати емпіричні ОФ. При створенні таких ОФ можливі два підходи. Перший полягає у використанні експериментальних даних про зв'язуванні різних лігандів з досліджуваним рецептором [12, 13]. Отримана білок-специфічна ОФ, що враховує структурні особливості сайту, дозволяє ефективно проводити пошук нових лігандів для цього білка по базах даних низькомолекулярних сполук. Другий підхід заснований на створенні лігандспецифічних критеріїв. Він менш поширений, так як вимагає великої кількості експериментальних даних про взаємодії окремого класу лігандів з різними білками. Проте на даний момент вже існують кілька успішних прикладів створення таких специфічних ОФ для пептидів, вуглеводів, АТР.

Результатом докінгу, як правило, є не одна структура комплексу білок-ліганд, а цілий набір найбільш ймовірних (з кращими значеннями ОФ) орієнтацій ліганда в сайті зв'язування. Тому навіть у разі, якщо недостатня точність використовуваної в процесі докінгу ОФ не дозволяє вибрати з отриманого набору «правильний» варіант, що відповідає нативній орієнтації ліганда, завжди є можливість переранжирувати цей набір по більш ефективному критерію. Такий метод отримав назву консенсусного докінгу. Треба відзначити, що багато

програмних пакетів для молекулярного докінгу використовують кілька різних ОФ, що істотно полегшує реалізацію такого підходу на практиці. Крім того, було відмічено, що серед різних ОФ одні більш ефективні в процесі пошуку можливих конформацій і орієнтацій ліганда в сайті, а інші - в фінальному ранжируванні цих варіантів по значенням енергії взаємодії білок-ліганд [14].

Використання недостатньо адекватної оціночної функції може призводити як до помилок у визначенні точного розташування ліганда в сайті зв'язування, так і до помилкового визначення сполук-лідерів при віртуальному скринінгу. Обов'язково потрібна перевірка достовірності оціночної функції для даної конкретної мішені на основі, наприклад, ретгенівських структур комплексів ліганда і рецептора (при їх наявності) або знаходження справжніх лігандів в спеціально складеній базі випадкових сполук.

Також розповсюдженою проблемою є наявність молекул води. Найчастіше зв'язування ліганда і рецептора опосередковується молекулою води. При використанні стандартних методів докінгу потенційна наявність цієї молекули не враховується ніяк. Вирішення цієї проблеми можливе при використанні молекулярної динаміки в явно заданому розчиннику або за допомогою попереднього розміщення необхідної молекули в сайті зв'язування (складний і неоднозначний варіант).

Крім того, додаткові навантаження на часові затрати додаються через рухомість рецепторів та велику кількість конформацій ліганд.

Вирішення цих проблем можливо за допомогою різних методів уточнення структури – мінімізації енергії, молекулярної динаміки, залучення додаткових даних. Проте, уточнення структури сайту зв'язування - найважливіше завдання, оскільки помилка на цій стадії може привести до великих і необґрунтованих фінансових затрат.

1.1.2 Алгоритм конформаційного пошуку

У класичному варіанті молекулярного докингу завдання алгоритму конформаційного пошуку зводиться до перебору конформоційного простору комплексу за рахунок варіювання торсіонних кутів ліганду і його переміщення в цілому відносно нерухомої структури білка-мішені. сучасні алгоритми конформаційного пошуку в переважній більшості випадків знаходять конформації, близькі до експериментальних, за порівняно короткий час. Проте є фактори, які також впливають на успішність докингу, які часто не враховуються в стандартних алгоритмах. Один з таких факторів - конформаційна рухливість білка-мішені, в більшості випадків супроводжує зв'язування ліганду. Діапазон рухливості може бути різним - починаючи з невеликого «підстроювання» бічних ланцюгів і закінчуючи масштабними доменними рухами. Ці рухи відіграють велику роль: якщо взяти структуру білка, оптимальну для зв'язування даного ліганду, то результат докингу напевно буде точніше, ніж якщо взяти будь-яку іншу (наприклад, апоформу).

На перший погляд самим логічним вирішенням цієї проблеми є огляд рухливості білка в програмі докингу. На жаль, сучасні обчислювальні кошти не дозволяють проводити таке моделювання за прийнятний час, так як молекула білка дуже велика, і огляд рухливості по всіх ступенях свободи може привести до так званого «комбінаторного вибуху» (астрономічному збільшенню числа можливих варіантів). Лише в деяких програмах передбачена обмежена рухливість сайтів зв'язування білка (як правило, на рівні невеликої адаптації конформацій бічних ланцюгів залишків активного центру). Інший підхід до цієї проблеми складається в докингу кількох різних конформацій одного і того ж білка з подальшим вибором кращих рішень з кожного запуску докингу. Третій підхід - знайти деяку універсальну структуру білка-мішені, за участю якої докинг давав би досить хороші результати для різних класів лігандів. При цьому

зменшується число «Пропущених» (але правильних) рішень, однак сильно також зростає і число невірних варіантів.

Оціночні функції (ОФ), які використовуються в процесі докину, служать для обчислення приблизної енергії комплексів і ранжирування різних передбачуваних конформацій ліганда в сайті зв'язування на кожному кроці конформаційного пошуку. Але ОФ не завжди добре справляються з цим завданням, додаючи помилки при моделюванні структури комплексу рецептор-ліганд там, де вони повинні вказувати вірне рішення. Заздалегідь передбачаючи неточність в ОФ, зазвичай розглядають не єдину структуру комплексу, а цілий набір можливих варіантів, розраховуючи на те, що серед них виявиться правильний або хоча б близький до нього. З цього набору можна вибрати найбільш правдоподібний варіант, ґрунтуючись, наприклад, на відомих експериментальних даних про роль тих чи інших амінокислотних залишків активного центру розглянутого білка в зв'язуванні лігандів. Фактично, в даному випадку використовується додатковий фільтр, що оптимізує положення ліганда для даного конкретного білка-мішені. До вирішення проблеми ранжирування результатів докину можна підійти і з іншого боку - використовувати лігандоспецифічне ОФ, що підвищують ймовірність знайти правдоподібне рішення в разі досить вузького класу хімічних речовин, наприклад, нуклеотидів, пептидів і ін. Такі ОФ враховують важливі для розпізнавання цих сполук взаємодії з рецептором, - наприклад, стекінг в разі нуклеотидів, водневі зв'язки пептидів з основним ланцюгом білка-мішені та інші типи контактів. ОФ показує, яка з орієнтацій ліганда в сайті зв'язування рецептора найбільш правдоподібна або, якщо порівнюються кілька різних лігандів, - який з них володіє найбільшим спорідненістю до білка-мішені. Зазвичай енергію зв'язування ліганда з рецептором розкладають на окремі незалежні складові - терми, що відображають різні фізичні взаємодії. Лінійна комбінація цих термів і є ОФ. Всі ОФ, що застосовуються в сучасних алгоритмах докину, можна умовно розділити на три типи: 1) засновані на силових полях, 2) емпіричні і 3) статистичні ОФ.

Оцінка енергії взаємодії рецептор-ліганд на основі силового поля здається найбільш природним підходом, оскільки терми силового поля безпосередньо відповідають тим чи іншим взаємодіям, що визначає молекулярне розпізнавання. Прикладом ОФ, заснованої на емпіричному силовому полі, може служити функція, яка включена в широко відомий пакет для докингу DOCK. Зважаючи на те що ліганди, як правило, ковалентно не пов'язані з рецептором, ОФ на основі молекулярномеханічних силових полів включають терми, що характеризують невалентних взаємодій: в основному, електростатичні і Ван-дер-ваальсово сили. Головним обмеженням ОФ на основі силових полів є те, що вони розраховані на оптимізацію структури молекули, з огляду на зміну лише ентальпійного складової енергії взаємодії. Тим часом, зв'язування ліганда з рецептором супроводжується ефектом десольватації, а також зміною ентропії, що не враховуються в розрахунках молекулярної механіки. Ця особливість суттєво обмежує застосування такого типу ОФ в молекулярному докинг. З цієї точки зору істотною перевагою володіють так звані «Емпіричні» ОФ. На відміну від ОФ, заснованих на силових полях, вони включають терми, що описують міжмолекулярні контакти найчастіше більш примітивним способом, - без проведення прямих аналогій з парними межмолекулярними фізичними взаємодіями. Передбачувальна здатність ОФ залежить не тільки від конкретного виду термів, що описують ті чи інші взаємодії, але і від вагових коефіцієнтів при термах, що визначаються виходячи з параметризації з використанням навчальних наборів експериментальних даних про структури комплексів, - як у функції Chemscore, розробленої в 1997 р.

В даному випадку міжмолекулярні взаємодії представлені у вигляді лінійної комбінації термів, що описують різні види контактів: водневі зв'язки (Chb), гідрофобні взаємодії (Clip), взаємодії з іонами металів (Cmet); Крім того, враховується число «заморожених» при зв'язуванні торсіонних кутів Hrotb, що відбивають зміна ентропії (Crot). Завдяки спрощеному вигляді термів можна істотно скоротити розрахунки і заощадити обчислювальні ресурси. Наприклад,

координаційні зв'язки з іонами металів $M(rlr)$ або гідрофобні контакти $L(rlr)$ можуть бути описані з допомогою відстаней rlr між відповідними атомами ліганда і рецептора, хоча таке наближення і не є фізично коректним. Водневі зв'язки описуються емпіричними геометричними параметрами (відстань між донором і акцептором rAD і кут між ними і атомом водню αADH), а не їх енергетичними характеристиками

Ефективність емпіричних ОФ сильно залежить від використовуваних при їх параметризації навчальних наборів. Останнє, звичайно, накладає певні обмеження на область застосування цих ОФ - надійність одержуваних результатів не може бути гарантована в тому випадку, якщо молекули білка-мішені і / або ліганда істотно відрізняються за характером міжмолекулярних взаємодій від навчальної вибірки. Наприклад, ОФ, яка найбільш ефективна при розгляді комплексів, де домінують полярні контакти (водневі зв'язки, сольові містки і т.д.), може видавати помилкові результати у випадках, коли гідрофобні контакти відіграють ключову роль в розпізнаванні ліганда рецептором, і навпаки.

Третій тип ОФ - статистичні, засновані на кривих радіального розподілу атомів в структурах комплексів ліганд-рецептор, отриманих експериментально. В Надалі ці криві можуть бути перетворені в статистичні потенціали, які, однак, більшою мірою призначені для передбачення орієнтації ліганда в активному сайті, ніж для оцінки вільної енергії зв'язування. Найбільш відомі з таких ОФ - PMF і ASP, де парні потенціали параметризовані для кожного типу атомів і визначають середню об'ємну щільність атомів одного типу і їх радіальну функцію розподілу щодо кожного атома іншого типу.

Статистичні ОФ неявним чином характеризують ті особливості взаємодії ліганда з рецептором, які складно описати в явному вигляді, - наприклад взаємодії ароматичних груп. У той же час даний підхід сильно залежить від поділу атомів на типи (як правило, число типів атомів в таких схемах істотно

перевищує число хімічних елементів) і, як і емпіричні ОФ, від складу навчальної вибірки. Крім того, інтерпретація статистичних ОФ в загальноприйнятих термах різних взаємодій (електростатичних, гідрофобних і т.д.) є скрутною.

Останнім часом емпіричні ОФ знайшли широке застосування в алгоритмах докину. Популярність використання емпіричних ОФ обумовлена тим, що з накопиченням все більшої кількості експериментальних даних про будову і біохімії молекулярних комплексів для них стає досить просто скласти навчальну вибірку і конструювати системоспецифічні або настроюються ОФ. При створенні системоспецифічних ОФ можливі два підходи. Перший полягає у використанні експериментальних даних про зв'язуванні різних лігандів з досліджуваним рецептором. Отримана білок-специфічна ОФ, що враховує структурно особливості сайту, дозволяє ефективно проводити пошук нових лігандів для цього білка по базах даних низькомолекулярних сполук. Другий підхід заснований на створенні лігандспецифічних критеріїв. Він менш поширений, так як вимагає великої кількості експериментальних даних про взаємодії окремого класу лігандів з різними білками. Проте на даний момент вже існують кілька успішних прикладів створення таких специфічних ОФ для пептидів, вуглеводів, АТР.

Однією з важливих проблем, що виникають при виборі найбільш підходящих критеріїв оцінки для дослідження конкретної системи білок-ліганд, є облік різних (найбільш значущих) міжмолекулярних взаємодій в розглянутому комплексі. Зупинимось докладніше на методах обліку деяких з цих взаємодій.

1.1.3 Водневі зв'язки

Водневі зв'язки часто відіграють визначальну роль у формуванні структури біологічних макромолекул - білків і нуклеїнових кислот, - а також в освіті їх комплексів з низькомолекулярними речовинами. В емпіричних ОФ наявність в

молекулі водневих зв'язків визначається по простих геометричних критеріям – віддалі між донором і акцептором і розі між ними і атомом водню.

Нерідко для певного типу лігандів характерне утворення певних мотивів (комбінацій) водневих зв'язків з білком. Ці комбінації водневих зв'язків вносять свій внесок в специфічне впізнавання лігандів в активному центрі білка-мішені. Пошук і облік таких мотивів в ОФ може істотно поліпшити предсказательную здатність докину на тому спектрі мішеней / лігандів, для яких характерний цей мотив. З цією метою представляють інтерес дослідження зв'язування пептидних і нуклеотидних лігандів з білками.

Зокрема, на основі аналізу експериментально встановлених просторових структур комплексів білок-пептид з баз даних Brookhaven Protein Data Bank (PDB) і PDBbind було показано, що атоми азоту і кисню основному ланцюзі пептидних лігандов статистично достовірно утворюють водневі зв'язки переважно з головним ланцюгом білка-рецептора і в меншій мірі - з атомами бічних груп амінокислотних залишків. Пояснюється це тим, що часто в таких білках ділянку β -тяжа експонований в активний центр надає пептидному ліганду або його аналогу можливість утворити велику кількість саме таких зв'язків, також прийнявши витягнуту форму β -тяжа, - як, наприклад, в матриксних мметалопротеїна. Для нуклеотидних лігандів, тобто таких, до складу яких входять азотисті підстави (ATP, ADP, GTP, NAD, FAD і ін.), також характерні певні мотиви водневих зв'язків з сайтом зв'язування в білку-рецепторі. Причому в даному випадку вони під чому нагадують сітку водневих зв'язків в парах нуклеотидів в подвійній спіралі ДНК. Мотиви водневих зв'язків можуть бути враховані в ОФ у вигляді суми вкладів окремих зв'язків, складових мотив.

1.1.4 Гідрофобні взаємодії

Полярні молекули мають властивість об'єднуватися один з одним в полярному розчиннику, прагнучи, таким чином, мінімізувати площа контакту з ним. На макроскопічному рівні це спостерігається у вигляді так званих гідрофобних (Ліпофільних) взаємодій, які відіграють найважливішу роль у формуванні просторової структури біологічних макромолекул і цілих систем, таких, як, наприклад, ліпідна біслойних мембрана. Також, в ряді випадків вони відповідальні і за специфічне впізнавання рецептор-ліганд.

Незважаючи на широку поширеність і фундаментальне значення для структури біомакромолекул і їх комплексів, природа фізичних сил, що лежать в основі гідрофобного ефекту, не визначена досить точно для їх кількісного опису.

Це пояснюється тим, що гідрофобний ефект складається з безлічі різних міжмолекулярних взаємодій: електростатичного, ван-дер-ваальсова, ентропійного ефекту та ін. В зв'язку з цим до цих пір не запропоновано універсального методу чисельної

1.1.5 Оцінки вільної енергії гідрофобних взаємодій.

Проте усвідомлення ключової ролі гідрофобних взаємодій в організації біомолекулярних систем дало поштовх розвитку емпіричних підходів до оцінки гідрофобности. Найпростіший спосіб, який використовується, зокрема, в функції CChemscor, - розділити всі атоми на гідрофобні (вуглець, сірка, галогени) і гідрофільні (Інші атоми). Число контактів між гідрофобними атомами може служити приблизною оцінкою величини гідрофобних взаємодій ліганду з рецептором.

Однак цей підхід має серйозні обмеження і з його допомогою не завжди можна домогтися коректного передбачення. Так, гетероциклічний фрагмент в складі ліганда (Наприклад, аденін в АТР) буде ідентифікований як

гідрофільний, хоча відомо, що аденін, що входить до складу АТР, переважно зв'язується в гідрофобних кишнях активних сайтів і, отже, має гідрофобні властивості.

Більш точні методи, що враховують при розрахунку властивостей окремих частин молекули не тільки тип хімічного елемента, але і вплив оточення, засновані на експериментальних даних по розподілу органічних сполук між полярною і неполярною фазами (зазвичай вода / октанол). Логарифм коефіцієнта розподілу речовини між двома середовищами - $\log P$ - традиційно служить кількісною мірою його гідрофобності. Для складання ОФ часто використовується лінійне наближення, згідно якому молекулярний коефіцієнт $\log P$ може бути представлений у вигляді лінійної комбінації атомних констант гідрофобності. В системі параметризації кожен атом молекули відноситься до одного з 120 топологічних типів, що включають явно задані атоми водню. Також відомі інші системи параметризації – без атомів водню, фрагментів та ін. Емпіричні атомні константи гідрофобності використовують при вивченні просторового розподілу гідрофобних / гідрофільних властивостей, наприклад, на поверхні молекули. Згідно з концепцією молекулярного гідрофобного потенціалу (МГП), гідрофобні властивості розраховують на поверхні молекули або в будь-якій іншій точці простору, використовуючи функцію, залежну від відстані. Необхідно відзначити, що, як уже було сказано вище, суворого опису гідрофобних сил не існує, тому що використовуються функції $g(r_{ij})$ носять, по суті, інтуїтивний і емпіричний характер. При цьому в деяких роботах було показано, що одні з них краще описують властивості низькомолекулярних сполук, а інші - макромолекул, таких, як білки.

Для характеристики вкладу гідрофобних взаємодій при утворенні комплексів ліганд-рецептор запропоновані різні критерії, що оцінюють відповідність їх гідрофобних / гідрофільних властивостей. Існують методи, засновані на порівнянні величини і знака констант гідрофобності контактують атомів або значень МГП двох молекул на поверхні інтерфейсу. Інші методи

оцінюють гідрофобні взаємодії просто по числу контактів між гідрофобними атомами або хімічними групами, або за часткою гідрофобної поверхні, екранованої від розчинника, або по комплементарності гідрофобних областей на поверхні двох молекул (утворення енергетично вигідного контакту між цими областями). Для кількісної оцінки комплементарності гідрофобних властивостей систем типу АТР-білок ми використовували параметр, що враховує частку заглибленою гідрофобної поверхні ліганда.

1.1.6 Стекінг

Серед різних типів міжмолекулярних контактів на особливу увагу заслуговує стекінг ароматичних кілець, також грає важливу роль в молекулярному розпізнаванні. Хоча багато лігандів, в тому числі і лікарські сполуки, містять ароматичні фрагменти і групи, стекінг часто не враховується явно при складанні ОФ.

Стекінг-взаємодії спостерігаються між двома ароматичними групами, внаслідок чого вони приймають певну орієнтацію один щодо одного в просторі. Найвідоміший приклад - подвійна спіраль ДНК, де азотисті основи внаслідок стекінг розташовуються паралельно один одному. Також можливо і перпендикулярний («Т-образне») взаємне розташування ароматичних кілець, що показано, наприклад, для бензолу, крім того, ароматичні сполуки мають тенденцію брати участь в π -катионному взаємодії, при якому утворюється контакт між позитивно зарядженими групами і електронним хмарою кільця.

Ароматичні контакти, так само як і водневі зв'язку, можуть бути описані з допомогою геометричних критеріїв. У наших дослідженнях взаємодій АТР з різними білками було показано, що аденіну властиво освіту стекінг з бічними ланцюгами амінокислотних залишків білка, переважно з фенілаланином, що узгоджується і з результатами інших робіт. Для того щоб виявити параметри контактів, утворених в результаті стекінг, ми пропонуємо аналізувати взаємну

орієнтацію двох ароматичних фрагментів з'єднань в термінах відстані між центрами циклів і кута між їх площинами.

Діапазон значень параметрів, що визначають наявність або відсутність стекінг в розрахункових алгоритмах, до сих пір залишається не з'ясованим і в оціночних критеріях вибирається досить довільно. Картину погіршує ще й той факт, що багато ароматичні сполуки прагнуть розташуватися не тільки паралельно, а й перпендикулярно один до одного, як це було показано для амінокислот в білках і в моделях, що складаються з простих вуглеводнів, - бензолу, нафталіну і т.д.

Уточнення взаємного розташування ароматичних кілець для конкретних систем підвищує ефективність оцінки якості та достовірності структур комплексів білок- ліганд, що передбачаються методами молекулярного моделювання. Наприклад, в результаті аналізу просторових структур комплексів різних білків з лігандами, що містять у своїй структурі гуанін, було встановлено, що для таких лігандів характерне утворення «паралельного» стекінг з відстанню від площині кільця гуаніну до центру ароматичного фрагмента амінокислотного залишку (h), рівним 3-4 Å, і іноді - «Т-образних» контактів з $h = 4.5-5.5$ Å, як у випадку з тирозином. Варто зауважити, що гуанідинових група аргініну, так само як і ароматичні амінокислоти, схильна утворювати паралельні контакти з аденін.

1.1.7 Консенсусний підхід

Результатом докинг, як правило, є не одна структура комплексу білок-ліганд, а цілий набір найбільш ймовірних (з кращими значеннями ОФ) орієнтацій ліганда в сайті зв'язування. Тому навіть у разі, якщо недостатня точність використовуваної в процесі докинг ОФ не дозволяє вибрати з отриманого набору «Правильний» варіант, відповідний нативної орієнтації ліганда, завжди є можливість переранжировать цей набір по більш ефективному критерієм. такий

метод отримав назву консенсусного докинг. Треба відзначити, що багато програмні пакети для молекулярного докинг використовують кілька різних ОФ, що істотно полегшує реалізацію такого підходу на практиці. Крім того, було відмічено, що серед різних ОФ одні більш ефективні в процесі пошуку можливих конформацій і орієнтацій ліганда в сайті, а інші - в фінальному ранжируванні цих варіантів по значенням енергії взаємодії білок-ліганд, як це, наприклад, реалізовано програмі LeadFinder.

Як приклад консенсусного докинг може служити дослідження, присвячене порівнянню двох ОФ, реалізованих в програмі докинг GOLD. Одна з них, Goldscore, використовує терми силових полів - ван-дер-ваальсові взаємодії і водневі зв'язку. Інша - емпірична ОФ Chemscore - враховує водневі зв'язку, гідрофобні взаємодії і координаційні зв'язки з іонами металів.

Різні орієнтації ліганда в сайті зв'язування, згенеровані за допомогою Goldscore, потім були отранжировані за значеннями Chemscore, що істотно підвищило частку вірно передбачених структур комплексів білок-ліганд. Зворотна операція - докинг з застосуванням Chemscore і подальше ранжування рішень докинг по Goldscore - також дозволила поліпшити результати в порівнянні з застосуванням кожного з ОФ у окремо. В літературі зустрічається ще й інший термін – консенсусне ранжування. Він означає дещо інший підхід: тут ОФ комбінують не за рахунок послідовного їх застосування, а виробляють ранжування результатів докинг по критерієм, який представляє собою зважену комбінацію різних ОФ. Системоспецифічні (настроюються) ОФ також можуть бути використані в «Консенсусному докинг». Вхідні в рівняння вагові коефіцієнти різних термів взаємодії можуть бути оптимізовані для окремих випадків, - наприклад, для конкретного білка-мішені або певного класу лігандів. Так наприклад, відомо, що необхідною умовою ефективного пригнічення тромбіну є воднева зв'язок інгібітора з бічним ланцюгом залишку Asp189 в активному сайті тромбіну, а для циклінзалежної кінази 2 інгібітори повинні утворювати водневий зв'язок з атомом азоту основному ланцюзі залишку Leu83.

Такого роду дані можуть бути використані і в стандартних ОФ в якості додаткових фільтрів або додаткових термів поряд з вже існуючими.

1.1.8 Опис існуючих технічних рішень

В даний час існує понад 50 програмних забезпечень для моделювання молекулярного докінгу. У цьому розділі буде наведено детальний опис та порівняння деяких з існуючих програмних забезпечень.

Молекулярний докінг описує процес, в якому молекула ліганда поміщається в активний сайт білкової мішені в тривимірному просторі. Важливими є два аспекти: передбачення аффіності (спорідненості) між лігандом і білком та прогноз правильної позиції ліганда в активному сайті білка. Головною метою усіх програм для моделювання молекулярного докінгу є передбачення точної конформації ліганд (його розташування та орієнтація в молекулі-мішені) з найвищою скоринговою оцінкою без помилок та за допустимий час. Зв'язування молекули ліганда до рецептора оцінюється його взаємодоповнюваністю з точки зору форми та фізико-хімічної взаємодії з цією білковою мішенню.

Програмне забезпечення Dock – одна з найбільш популярних програм докінгу, яка швидко розвивається. Висока швидкість роботи при використанні простих оціночних функцій і можливість паралельних обчислень дозволяють проводити ефективний віртуальний скринінг великих бібліотек сполук. Використання більш складних оціночних функцій, що враховують ефекти розчинника і зміну конформації рецептора, дозволяє проводити докладне дослідження зв'язування конкретного ліганда з рецептором. Особливістю алгоритму, який використовується в даному програмному забезпеченні, є те, що потенціали ліганд-рецепторної взаємодії обчислюються в вузлах просторової сітки, що покриває сайт зв'язування. Пошук сайту зв'язування можливий по геометричним параметрам; при цьому поверхня рецептора апроксимується набором сфер, що стосуються поверхні в двох точках і розташованих зовні від

поверхні. У першій версії програми поверхню ліганда представлялася набором аналогічних сфер, розташованих усередині поверхні. Ключове наближення: атоми ліганду можуть розташовуватися тільки в центрах сфер. У поточній версії програми автоматично відбувається поділ молекули ліганда на якірний фрагмент, для якого ведеться пошук оптимальної орієнтації, і зовнішні заступники, для яких ведеться дослідження всіх можливих конформацій.

Програмне забезпечення AutoDock – найпопулярніша з усіх програм докінгу. Найчастіше AutoDock використовують для дослідження поверхні рецептора з метою виявлення потенційного сайту зв'язування (якщо він невідомий). Швидкість роботи дозволяє за розумний час досліджувати всю поверхню рецептора, що неможливо за допомогою, наприклад, Dock. Дане програмне забезпечення використовує генетичний алгоритм пошуку оптимальної конформації. Відсутня підтримка віртуального скринінгу, різних оціночних функцій, паралельних обчислень. У версії AutoDock 4 враховується рухливість молекули рецептора.

Програмне забезпечення FlexX – програма для докінгу, яка може бути застосована як для уважного дослідження механізму взаємодії ліганда з рецептором, так і для віртуального скринінгу. На основі даних про структуру сайту зв'язування рецептора конструюється протомолекула, що складається з фрагментів CO, CH₄ та NH, на яку накладається структура ліганду. При конформаційному пошуку молекула фрагментується таким чином, щоб мінімізувати число конформацій, що перебираються. Оцінка енергії зв'язування проводиться за допомогою емпіричної оціночної функції. Алгоритм програми відрізняється високою швидкістю роботи.

FRED - точна і швидка програма молекулярного докінгу, за словами розробників, швидкість розрахунку-оцінки досягає 10 конформерів (ліганда) в секунду, а оцінка спорідненості за допомогою традиційних оціночних функцій проводиться після проходження фільтра на предмет стерическое

комплементарності ліганд-рецепторного комплексу і його фармакофорні особливостей .

Glide - потужна програма молекулярного докінгу, що дозволяє також проводити швидкий молекулярний докінг баз даних хімічних сполук (віртуальний скринінг). В якості алгоритму докінгу використовує метод Монте Карло.

GOLD - програма для ліганд-білкового докінгу на основі генетичного алгоритму. Програма повністю враховує конформаційну рухливість ліганд і часткову гнучкість бічних ланцюгів амінокислотних залишків білкової молекули. Використовуються оціночні функції (scoring functions): GoldScore, ChemScore, а також певні користувачем.

Окрім вищезгаданих програмних забезпечень існують також eHiTS, Surflex-DOck, FRED, Molegro та інші.

1.2 Порівняння існуючих технічних рішень

В результаті роботи програм докінгу ми отримуємо набір конформацій ліганда (лігандів), оптимально розташованих в сайті зв'язування рецептора. Ступінь достовірності цих результатів і складність вибору правильного рішення повністю залежать від використовуваної структури мішені, використаного методу та оціночних функцій. Отже, порівняємо основні характеристики розглянутих прикладних забезпечень для моделювання молекулярного докінгу. Порівняльна таблиця представлена в таблиці 1.1.

Таблиця 1.1 – Порівняльні характеристики існуючих систем

Назва ПЗ	AutoDock	FlexX	Glide	GOLD
Доступ до ПЗ	Безкоштовне для академічних цілей	Платне (6 тижнів пробний період)	Платне	Платне (2 місяці пробний період)
ОС, які підтримує	Unix, Mac OSX, Linux, SGI	Unix, Linux, SGI, Sun Windows	Unix, Linux, SGI, IBM AIX	Linux, SGI, Sun, IBM, Windows
Алгоритми/методи Докінгу	Генетичний алгоритм, генетичний алгоритм Ламарка, імітація отжигу	Інкрементальна конструкція	Монте Карло	Генетичний алгоритм
Оціночні функції	AutoDock	FlexXScore, PLP, ScreenScore, DrugScore	GlideScore, GlideCom	GoldScore, ChemScore, задана користувачем функція
Переваги	Велика точність, достатньо швидко працює	Велика швидкість, велика точність для малих структур	Велика швидкість для баз даних хімічних сполук	Добре підходить для лігандів із великою кількістю

				рухливих зв'язків
Недоліки	Відсутня підтримка віртуального скринінгу, різних оціночних функцій, паралельних обчислень	Мала точність для великих структур	Не завжди вдається отримати сполуку, інколи мала точність	Довго працює

Таким чином, як бачимо з порівняльної таблиці 2.1, усі розглянуті програмні забезпечення для молекулярного докінгу використовують зовсім різні алгоритми для пошуку конформацій ліганд та різні оціночні функції. Також бачимо, що всі ПЗ мають як переваги, так і недоліки.

1.3 Штучна нейронна мережа

Штучна нейронна мережа (ШНМ) - математична модель, а також її програмне або апаратне втілення, побудована за принципом організації та функціонування біологічних нейронних мереж - мереж нервових клітин живого організму. Це поняття виникло при вивченні процесів, що протікають в мозку, і при спробі змодельовати ці процеси. Проблема створення систем штучного інтелекту, здатних до інтелектуального аналізу інформації, розвивається починаючи з появою перших обчислювальних машин. Перша робота, яка тепер

за загальним визнанням вважається що відноситься до штучного інтелекту, була виконана Уорреном Мак—Каллоком і Уолтером Піттсом. Мак—Каллок і Піттс запропонували модель, що складається зі штучних нейронів, в якій кожен нейрон характеризувався тим, що знаходиться у «ввімкненому» або «вимкненому» стані, а перехід у « ввімкнений » стан відбувався у відповідь на стимуляцію достатньої кількості сусідніх нейронів. Схему пристрою, який моделює систему людського сприйняття запропонував нейрофізіолог Френк Розенблат, який він назвав «персептрон». Модель нейрона, запропонована Уорреном Мак—Каллоком і Уолтером Піттсом у 1943 році, широко використовується і наразі у теорії штучних нейронних мереж. З появою глибинного навчання, моделі штучних нейронних мереж стали найбільш потужними системами розпізнавання образів. Моделі глибоких нейронних мереж виникли через появу більш потужних обчислювальних пристроїв, велику кількість даних для їх навчання та деяких модернізацій архітектури мережі. У 2012 році, Алекс Крижевський зі своєю командою з великим відривом перемогли у змаганні з розпізнавання зображень – ImageNet, навчивши глибоку згорткову нейронну мережу AlexNet.

Revolution of Depth

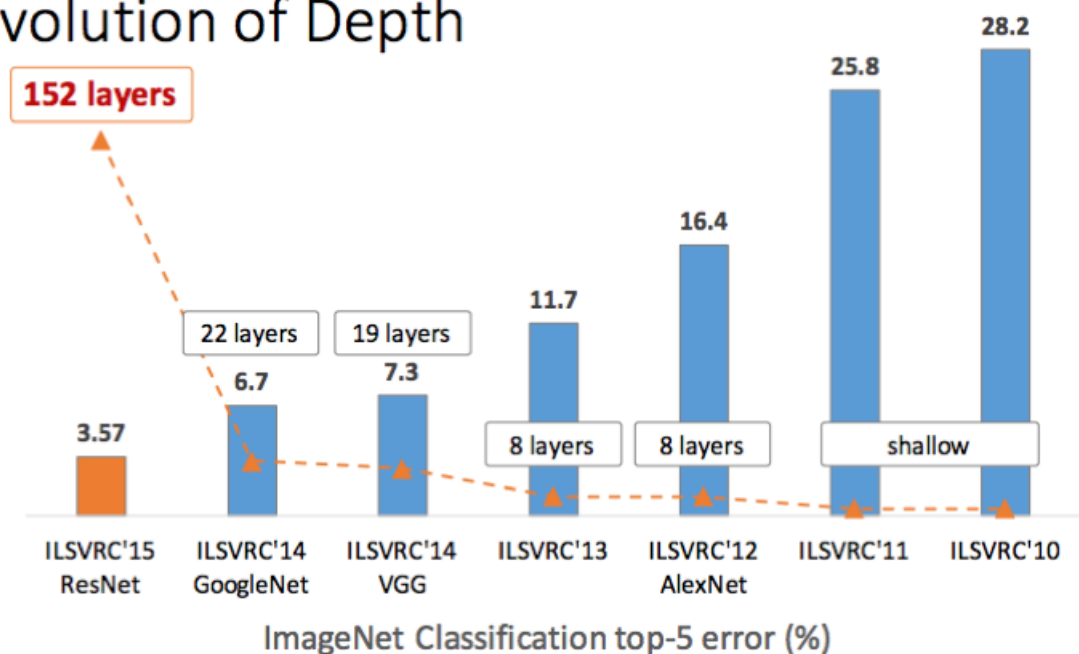


Рис. 1.3 Вплив глибини нейромережі на показник помилки

Ця подія стала визначною для розвитку глибинного навчання. AlexNet використовує модель штучного нейрона як основну структурну одиницю. Отже аналіз та моделювання біологічного нейрона у 1943 році виявилось визначним фактором для створення найпотужнішої системи розпізнавання образів. Подальший аналіз біологічних нервових систем показує, що мозок використовує велику кількість алгоритмів та структурних одиниць, та нейрон – це лише один з прикладів, який доводить здатність таких систем до інтелектуальної обробки інформації.

ШНМ є система з'єднаних і взаємодіючих між собою простих процесорів (штучних нейронів). Такі процесори зазвичай досить прості (особливо в порівнянні з процесорами, використовуваними в персональних комп'ютерах). Кожен процесор подібної мережі має справу тільки з сигналами, які він періодично отримує, і сигналами, які він періодично посиляє іншим процесорам. І, тим не менше, будучи з'єднаними в досить велику мережу з керованим взаємодією, такі окремо прості процесори разом здатні виконувати досить складні завдання.

З точки зору машинного навчання, нейронна мережа являє собою окремий випадок методів розпізнавання образів, дискримінантного аналізу, методів кластеризації і т.і.

З математичної точки зору, навчання нейронних мереж - це багатопараметрична завдання нелінійної оптимізації.

З точки зору кібернетики, нейронна мережа використовується в задачах адаптивного управління і як алгоритми для робототехніки.

З точки зору розвитку обчислювальної техніки та програмування, нейронна мережа - спосіб вирішення проблеми ефективного паралелізму [2].

А з точки зору штучного інтелекту, ШНМ є основою філософської течії коннективізма і основним напрямком в структурному підході з вивчення можливості побудови (моделювання) природного інтелекту за допомогою комп'ютерних алгоритмів.

Нейронні мережі не програмуються в звичному сенсі цього слова, вони навчаються. Можливість навчання - одне з головних переваг нейронних мереж перед традиційними алгоритмами. Технічно навчання полягає в знаходженні коефіцієнтів зв'язків між нейронами. В процесі навчання нейронна мережа здатна виявляти складні залежності між вхідними даними і вихідними, а також виконувати узагальнення. Це означає, що в разі успішного навчання мережа зможе повернути вірний результат на підставі даних, які були відсутні в навчальній вибірці, а також неповних і / або «зашумлених», частково спотворених даних.

Опишемо архітектуру та принцип роботи згорткової (конволюційної) нейронної мережі, з якої складається дискримінатор.

Згорткові мережі було натхнено біологічними процесами,[4] в яких схему з'єднання нейронів натхнено організацією зорової кори тварин. Окремі нейрони кори реагують на стимули лише в обмеженій області зорового поля, відомій як рецептивне поле. Рецептивні поля різних нейронів частково перекриваються таким чином, що вони покривають усе зорове поле.

Робота згорткової нейронної мережі зазвичай інтерпретується як перехід від конкретних особливостей зображення до більш і більш абстрактних деталей на кожному шарі до виділення понять високого рівня. При цьому мережа сама виробляє необхідну ієрархію абстрактних ознак (послідовності карт ознак), фільтруючи незначні деталі і виділяючи істотне.

У звичайному перцептроні, який представляє собою повнозв'язну нейронну мережу, кожен нейрон пов'язаний з усіма нейронами попереднього шару, причому кожен зв'язок має свій персональний ваговий коефіцієнт. У згортковій нейронній мережі під час операції згортки використовується лише обмежена матриця ваг невеликого розміру, яку «рухають» по всьому оброблюваному шару (на самому початку - безпосередньо по вхідному зображенню), формуючи після кожного зсуву сигнал активації для нейрона наступного шару з аналогічною позицією. Тобто для різних нейронів вихідного

шару використовуються одна і та ж матриця ваг, яку також називають ядром згортки. Її інтерпретують як графічне кодування якої-небудь ознаки, наприклад, наявність похилої лінії під певним кутом. Тоді наступний шар, що вийшов в результаті операції згортки такою матрицею ваг, показує наявність даної ознаки в оброблюваному шарі і її координати, формуючи так звану карту ознак (англ. Feature map). В згортковій нейронній мережі набір ваг не один, а ціла гама, що кодує елементи зображення (наприклад лінії і дуги під різними кутами). При цьому такі ядра згортки не закладаються дослідником заздалегідь, а формуються самостійно шляхом навчання мережі класичним методом зворотного поширення помилки. Прохід кожним набором ваг формує свій власний примірник карти ознак, роблячи нейронну мережу багатоканальною (багато незалежних карт ознак на одному шарі). Також слід зазначити, що при переборі шару матрицею ваг її пересувають зазвичай не на повний крок (розмір цієї матриці), а на невелику відстань.

Операція субдискретизації (англ. Subsampling, англ. Pooling, також перекладається як «операція підвибірки» або операція об'єднання), виконує зменшення розмірності сформованих карт ознак. У даній архітектурі мережі вважається, що інформація про факт наявності шуканого ознаки важливіше точного знання його координат, тому з кількох сусідніх нейронів карти ознак вибирається максимальний і приймається за один нейрон ущільненої карти ознак меншої розмірності. За рахунок цієї операції, крім прискорення подальших обчислень, мережа стає більш інваріантною до масштабу вхідного зображення. На рисунку 1.4 можна побачити типову архітектуру згорткової мережі.

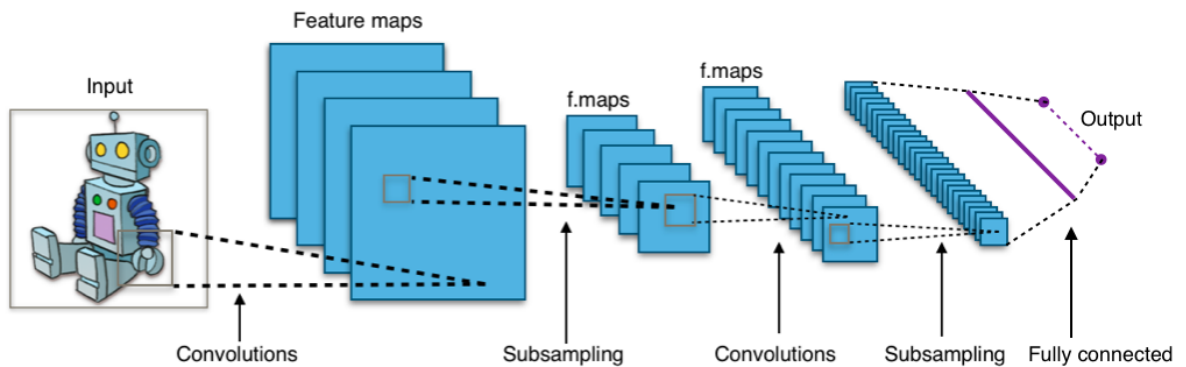


Рис. 1.4 – Типова архітектура згорткової мережі

1.3.1 Вхідний шар

Вхідний шар враховує двовимірну топологію зображень і складається з декількох карт (матриць), карта може бути одна, в тому випадку, якщо зображення представлено в відтінках сірого, інакше їх 3, де кожна карта відповідає зображенню з конкретним каналом (червоним, синім і зеленим).

Вхідні дані кожного конкретного значення пікселя нормалізуються в діапазон від 0 до 1, за формулою:

$$f(p, min, max) = \frac{p - min}{max - min}, \quad (1.1)$$

де f – функція нормалізації, p – значення кольору значення кольору від 0 до 255, min – мінімальне значення пікселя, max – максимальне значення пікселя.

1.3.2 Згортковий шар

Шар згортки включає в себе для кожного каналу свій фільтр, ядро згортки якого обробляє попередній шар за фрагментами (підсумовуючи результати матричного твору для кожного фрагмента). Вагові коефіцієнти ядра згортки (невеликий матриці) невідомі і встановлюються в процесі навчання. На рисунку

1.5 зображено нейрони згорткового шару, які перетворені за декількома вихідними каналами.

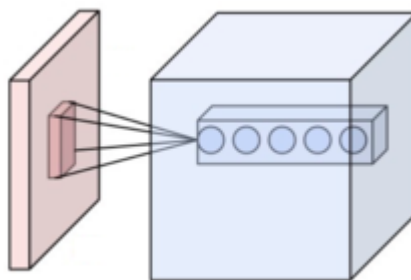


Рис. 1.5 – Візуальне зображення нейронів згорткового шару, які перетворені за декількома вихідними каналами

Кількість карт визначається вимогами до задачі, якщо взяти велику кількість карт, то підвищиться якість розпізнавання, але збільшиться обчислювальна складність. Виходячи з аналізу наукових статей, в більшості випадків пропонується брати співвідношення один до двох, тобто кожна карта попереднього шару (наприклад, у першого згорткового шару, попереднім є вхідний) пов'язана з двома картами згорткового шару.

Розмір усіх карт згорткового шару однаковий та розраховується за формулою:

$$(w, h) = (mW - kW + 1, mH - kH + 1), \quad (1.2)$$

де (w, h) - розмір згорткової карти, mW – ширина попередньої карти, mH - висота попередньої карти, kW – ширина ядра, kH – висота ядра.

Ядро являє собою фільтр або вікно, яке ковзає по всій області попередньої карти і знаходить певні ознаки об'єктів. Наприклад, якщо мережу навчали на зображеннях людей, то одне з ядер могло б в процесі навчання видавати

найбільший сигнал в області очей, рота, брів або носа, інше ядро могло б виявляти інші ознаки. Розмір ядра зазвичай беруть в межах від 3x3 до 7x7. Якщо розмір ядра маленький, то воно не зможе виділити будь-які ознаки, якщо занадто велике, то збільшується кількість зв'язків між нейронами. Також розмір ядра вибирається таким, щоб розмір карт згорткового шару був парних, це дозволяє не втрачати інформацію при зменшенні розмірності в підвиборчому шарі.

Ядро являє собою систему поділюваних ваг або синапсів, це одна з головних особливостей згорткової нейромережі. У звичайній багатошаровій мережі дуже багато зв'язків між нейронами, що вельми уповільнює процес детектування. У згортковій мережі - навпаки, загальні ваги дозволяють скоротити число зв'язків і дозволити знаходити одну і ту саму ознаку по всій області зображення.

На початку значення кожної карти згорткового шару рівні 0. Значення ваг ядер задаються випадковим чином в області від -0.5 до 0.5. Ядро ковзає по попередній карті і робить операцію згортки за формулою:

$$(f * g)[m, n] = \sum_{k, l} f[m - k, n - l] * g[k, l], \quad (1.3)$$

де f – вихідна матриця зображення, g – ядро згортки.

При цьому в залежності від методу обробки країв вихідної матриці результат може бути менше вихідного зображення, такого ж розміру або більшого розміру, відповідно до рисунка 1.6.

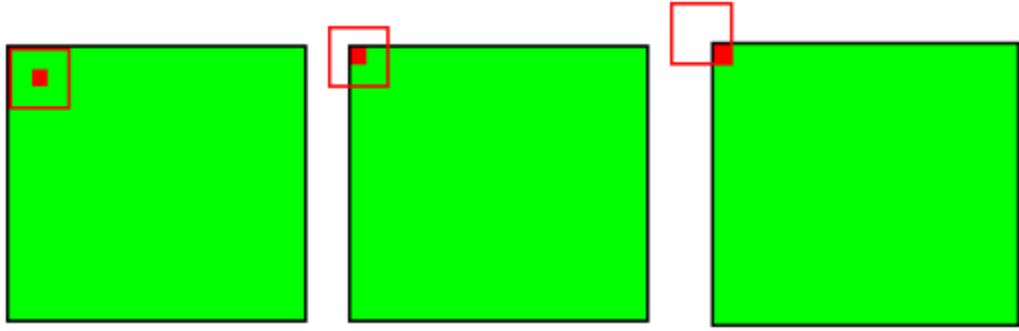


Рис. 1.6 – Види згортки вихідної матриці

Даний шар можна описати формулою:

$$x^l = f(x^{l-1} * k^l + b^l), \quad (1.4)$$

де x^l – вихід шару l , $f()$ – функція активації, b^l – коефіцієнт зрушення шару l , $*$ - операція згортки входу x з ядром k .

При цьому, за рахунок крайових ефектів, розмір вихідних матриць зменшується за формулою:

$$x_j^l = f(\sum_i x_i^{l-1} * k_j^l + b_j^l), \quad (1.5)$$

де x_j^l – карта признаков j (вихід шару l), $f()$ – функція активації, b^l – коефіцієнт зрушення шару l для карти признаков j , $*$ - операція згортки входу x з ядром k .

На рисунку 1.7 наведено приклад операції згортки та отримання значень згорткової карти.

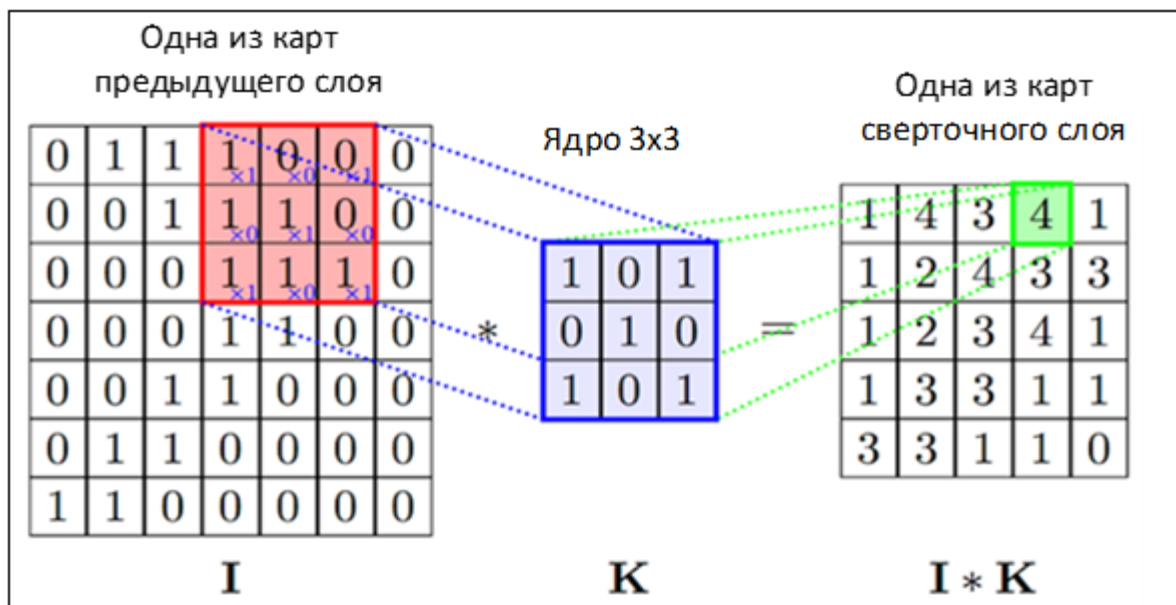


Рис. 1.7 – Операція згортки та отримання значень згорткової карти

1.3.3 Підвиборчий шар

Підвиборчий шар також, як і згортковий має карти, але їх кількість співпадає з попереднім шаром. Мета шару - зменшення розмірності карт попереднього шару. Якщо на попередній операції згортки вже були виявлені деякі ознаки, то для подальшої обробки настільки докладне зображення вже не потрібно, і воно зменшується до менш докладного. До того ж фільтрація вже непотрібних деталей допомагає не перенавчатися.

У процесі сканування карти попереднього шару ядром підвиборчого шару (фільтром), ядро не перетинається на відміну від згорткового шару. Зазвичай, кожна карта має ядро розміром 2x2, що дозволяє зменшити попередні карти згорткового шару в 2 рази. Вся карта ознак поділяється на ділянки 2x2 елемента, з яких вибираються максимальні за значенням. На рисунку 1.8 можна побачити принцип формування нової карти.

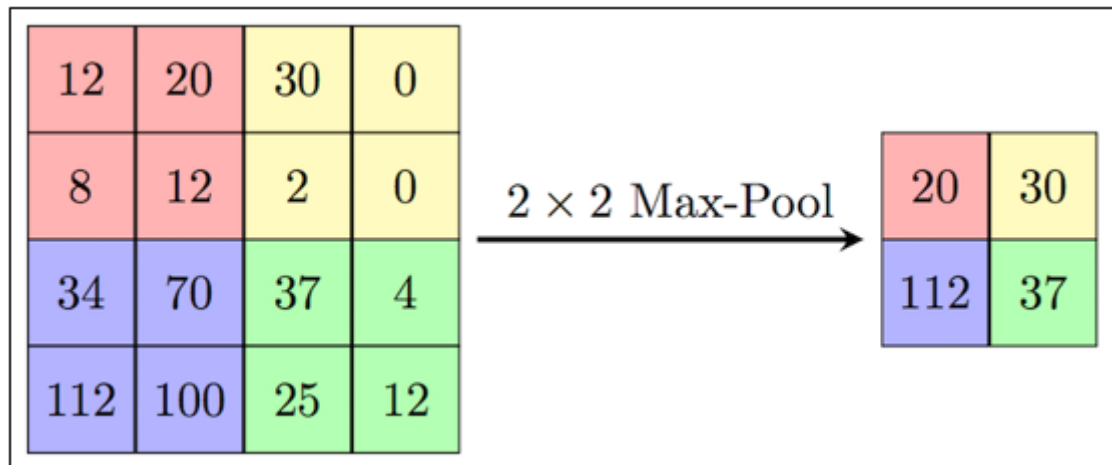


Рис. 1.8 – Приклад операції підвибірки

Також даний шар може бути описаний наступною формулою:

$$x^l = f(a^l * \text{subsample}(x^{l-1}) + b^l), \quad (1.6)$$

де x^l – вихід шару l ;

$f()$ – функція активації;

a^l, b^l – коефіцієнти зсуву шару l ;

$\text{subsample}()$ – операція вибірки локальних максимальних значень.

1.3.4 Повнозв'язковий шар

Останній з типів шарів – це шар звичайного багат шарового персептрона. Мета шару - класифікація, що моделює складну нелінійну функцію, оптимізуючи яку, поліпшується якість розпізнавання. Приклад повнозв'язних шарів зображений на рисунку 1.9.

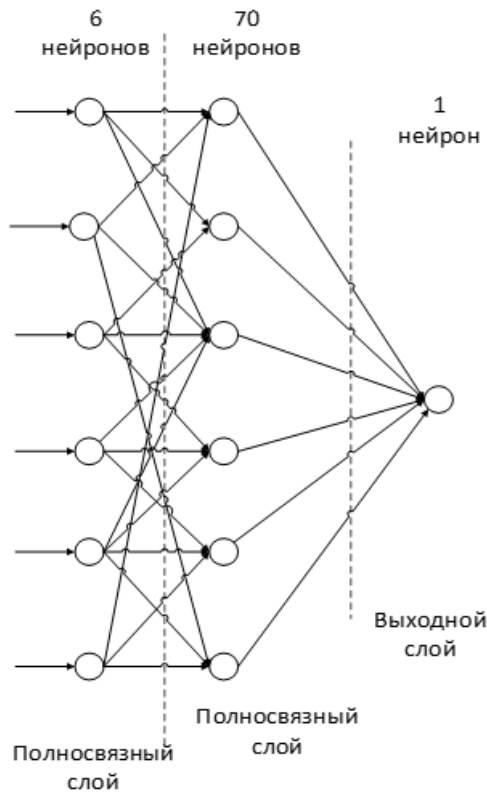


Рис. 1.9 – Приклад повнозв'язних шарів

Нейрони кожної карти попереднього подвигорочного шару пов'язані з одним нейроном прихованого шару. Таким чином число нейронів прихованого шару дорівнює числу карт подвигорочного шару, але зв'язку можуть бути не обов'язково такими, наприклад, тільки частина нейронів будь-якої з карт подвигорочного шару бути пов'язана з першим нейроном прихованого шару, а частина, що залишилася з другим, або все нейрони першої карти пов'язані з нейронами 1 і 2 прихованого шару. Обчислення значень нейрона можна описати формулою:

$$x_j^l = f(\sum_i x_i^{l-1} * w_{i,j}^{l-1} + b_j^{l-1}), \quad (1.7)$$

де x_j^l – карта признаков j (вихід шару l);

$f()$ – функція активації;

b^l – коефіцієнт здвику шару l ;

$w_{i,j}^{l-1}$ – матриця вагових категорій шару l .

Таким чином, приклад загальної архітектури згорткової нейронної мережі зображений на рисунку 1.10.

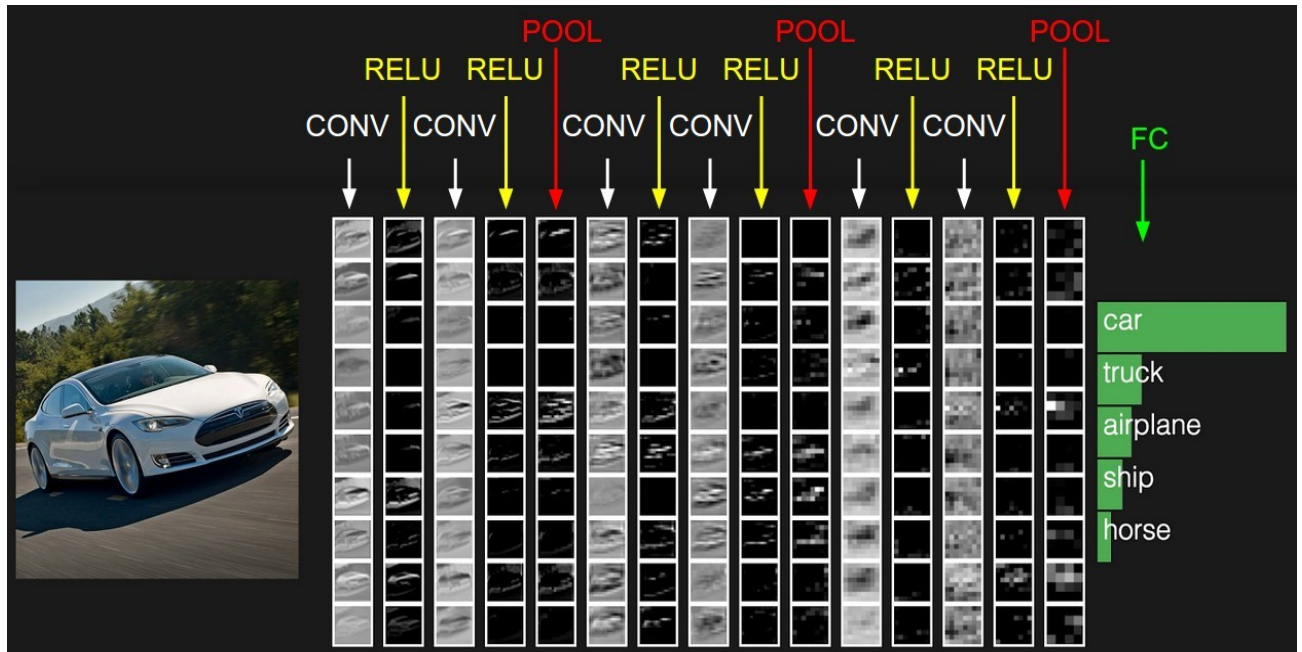


Рис. 1.10 – Приклад архітектури згорткової нейронної мережі

1.3.5 Вибір функції активації

Одним з етапів розробки нейронної мережі є вибір функції активації нейронів. Вид функції активації багато в чому визначає функціональні можливості нейронної мережі і метод навчання цієї мережі. Класичний алгоритм зворотного поширення помилки добре працює на двошарових і тришарових нейронних мережах, але при подальшому збільшенні глибини починає відчувати проблеми. Одна з причин - так зване загасання градієнтів. У міру поширення помилки від вихідного шару до вхідного на кожному шарі відбувається домноження поточного результату на похідну функції активації. Похідна у традиційній сігмоїдальній функції активації менша одиниці на всій області визначення, тому після декількох шарів помилка стане близькою до нуля. Якщо ж, навпаки, функція активації має необмежену похідну (як, наприклад,

гіперболічний тангенс), то може статися вибухове збільшення помилки у міру поширення, що призведе до нестійкості процедури навчання.

В даній роботі будемо використовувати функцію активації – виправлена лінійна одиниця (англ. ReLU – rectified linear unit), яка визначена наступною формулою:

$$f(s) = \max(0, s), \quad (1.8)$$

де s – вхідне значення нейрона. Також зображення даної функції представлено на рисунку 1.11.

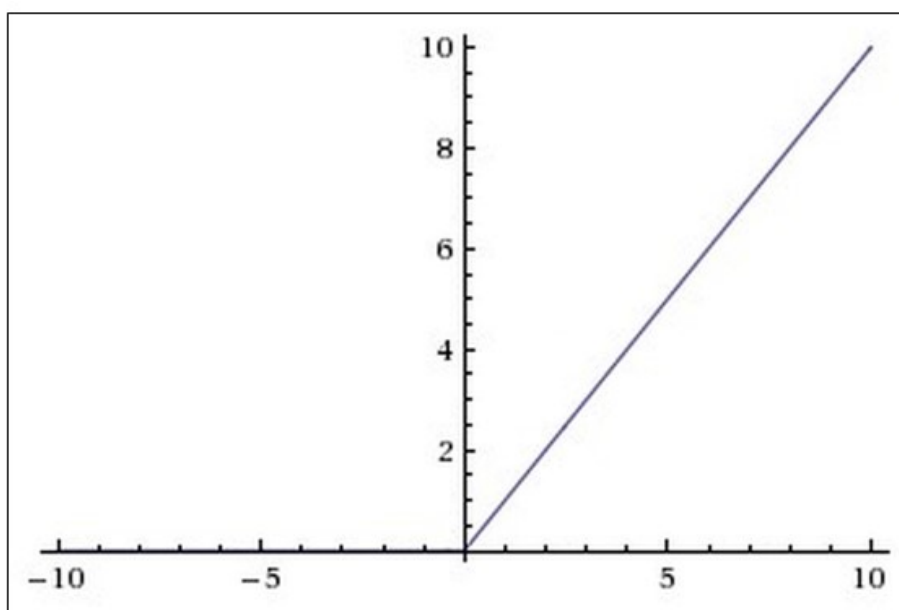


Рис. 1.11 – Графік функції активації ReLU

Було обрано дану функцію активації через наступні переваги:

- її похідна дорівнює або одиниці, або нулю, і тому не може статися розростання або загасання градієнтів, тому що помноживши одиницю на дельту помилки ми отримаємо дельту помилки, якщо ж ми б використовували іншу функцію, наприклад, гіперболічний тангенс, то дельта помилки могла, або зменшитися, або зрости, або залишитися такою

ж, тобто, похідна гіперболічного тангенса повертає число з різним знаком і величиною, що може сильно вплинути на загасання або розростання градієнта. Більш того, використання даної функції приводить до проріджування ваг;

- обчислення сигмоїд і гіперболічного тангенса вимагає виконання ресурсномістких операцій, таких як возведення в степінь, в той час як ReLU може бути реалізований за допомогою простого порогового перетворення матриці активацій в нулі;
- відсікає непотрібні деталі в каналі при негативному виході.

1.3.6 Перевернуті або розгорнуті нейромережі

Основною відмінністю між конволюційною та деконволюційною нейронними мережами є те, що в конволюційній нейронній мережі вхідний сигнал піддається декільком шарам згортки та субдискретизації. Деконволюційна нейронна мережа навпаки прагне згенерувати вхідний сигнал у вигляді суми згорток карт ознак з урахуванням застосовуваних фільтрів. Для вирішення даного завдання, використовується широкий спектр інструментів теорії розпізнавання образів, наприклад алгоритми усунення розмитості (deblurring).

На вхід цієї нейромережі подається вектор, який кодує клас, визначений тип стільця, і ще один вектор, який кодує геометричні параметри камери. На виході ця нейромережа синтезує зображення стільця і маску, яка відокремлює стілець від фону. Далі показано відмінність такої розгорнутої нейромережі від традиційної — в тому, що вона все робить в зворотному порядку (Рис. 1.12).

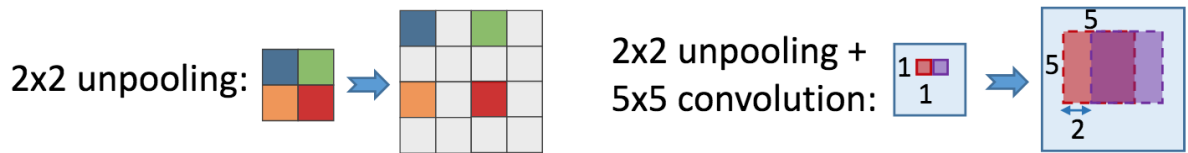


Рис. 1.12 – Ілюстрація «деконволюції» у вигляді обернених конволюційних операцій

Зображення у тепер не на вході, а на виході. Уявлення, які виникають у цій нейромережі, спочатку є просто векторами, а в деякий момент перетворюються в набори зображень. Поступово зображення комбінуються один з одним за допомогою узагальнених згортки, і на виході виходять картинки. Приклад архітектури даної мережі зображено на рисунку 1.10.

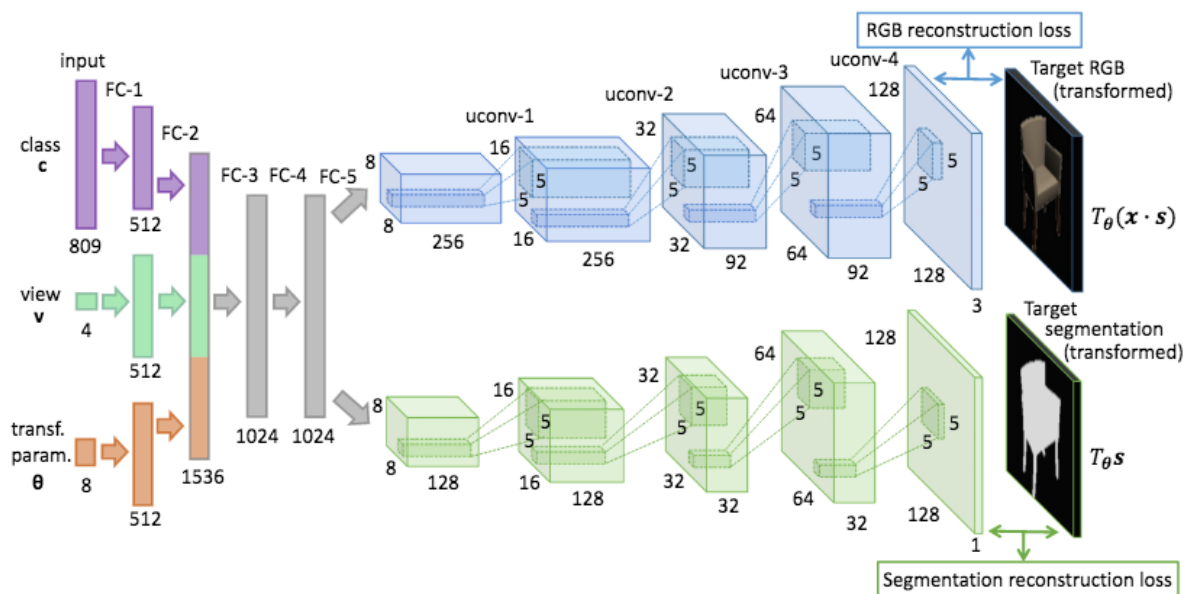


Рис. 1.13 – Приклад архітектури деконволюційних нейронних мереж

Таким чином, за допомогою деконволюційних нейронних мереж можемо отримати зображення з вектору або карти його ознак (Рис. 1.11).

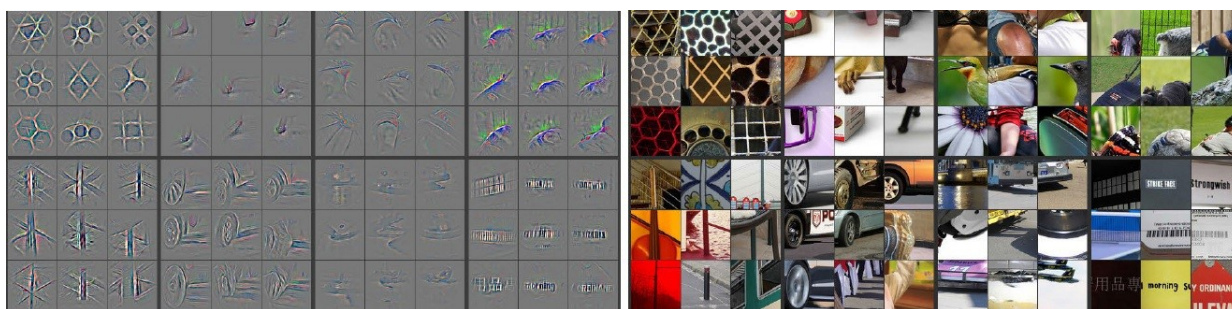


Рис. 1.14 – Приклад результатів роботи деконволюційної нейронної мережі
(зліва – карти ознак, справа – результуюче зображення)

1.3.7 Метод зворотного поширення помилок

Навчання алгоритмом зворотного поширення помилок передбачає два проходи по всім шарам мережі: прямого і зворотного. При прямому проході вхідний вектор подається на вхідний шар нейронної мережі, після чого поширюється по мережі від шару до шару. В результаті генерується набір вихідних сигналів, який і є фактичною реакцією мережі на даний вхідний образ. Під час прямого проходження всі синаптичні ваги мережі фіксовані. Під час зворотного проходження всі синаптичні ваги налаштовуються відповідно до правила корекції помилок, а саме: фактичний вихід мережі віднімається з бажаного, в результаті чого формується сигнал помилки. Цей сигнал з часом поширюється по мережі в напрямку, протилежному напрямку синаптичних зв'язків. Звідси і назва - алгоритм зворотного поширення помилки. Синаптичні ваги налаштовуються з метою максимального наближення вихідного сигналу мережі до бажаного.

Метою навчання мережі алгоритмом зворотного поширення помилок є таке підстроювання її ваг, щоб деяка множина входів приводила до необхідної множини виходів. Для стислості ці множини входів і виходів будуть називатися векторами. При навчанні передбачається, що для кожного вхідного вектора існує

парний йому цільовий вектор, що задає необхідний вихід. Разом вони називаються навчальною парою. Мережа навчається на багатьох парах.

Алгоритм зворотного поширення помилки наступний:

1. Ініціалізувати синаптичні ваги маленькими випадковими значеннями.
2. Вибрати чергову навчальну пару з навчальної множини; подати вхідний вектор на вхід мережі.
3. Обчислити вихід мережі.
4. Обчислити різницю між виходом мережі і необхідним виходом (цільовим вектором навчальної пари).
5. Відкоригувати ваги мережі для мінімізації помилки.
6. Повторювати кроки з 2 по 5 для кожного вектора навчальної множини до тих пір, поки помилка на всій множині не досягне прийняттого рівня.

В якості оціночної функції для МЗПП оберемо метод найменших квадратів:

$$E(w_k) = \frac{1}{2} \sum (t_k - o_k)^2, \quad (1.9)$$

де w_k -- ваги зав'язків, t_k - правильні значення, o_k – значення вузлів.

Для зменшення помилки використовується стохастичний градієнтний спуск, тобто потрібно «рухатися» в протилежну сторону від градієнта:

$$\Delta w_{i,j} = -\mu \frac{\partial E}{\partial w_{i,j}}, \quad (1.10)$$

де $0 < \mu < 1$ – множник, що задає швидкість «руху», яку також можна рахувати як:

$$\Delta w_{i,j} = -\mu \delta_j w_{i,j}, \quad (1.11)$$

$$\text{де } \delta_j = \begin{cases} -o_j(1 - o_j)(t_j - o_j), & \text{для вузлів останнього рівня} \\ -o_j(1 - o_j) \sum_{k \in \text{Outputs}} \delta_k w_{j,k}, & \text{для внутрішнього вузла мережі} \end{cases}$$

Існує два режими реалізації методу зворотного поширення помилки:

- Стохастичний градієнтний спуск.
- Пакетний (batch) градієнтний спуск.

Для пакетного градієнтного спуску функція втрат обчислюється для всіх зразків разом після закінчення епохи, і потім вводяться поправки вагових коефіцієнтів нейрона відповідно до методу зворотного поширення помилки.

Стохастичний метод одразу ж після обчислення виходу мережі на одному зразку вводить поправки в вагові коефіцієнти.

Пакетний метод більш швидкий і стабільний, але він має тенденцію зупинятися і застрягати в локальних мінімумах. Тому для виходу з локальних мінімумів потрібно використовувати особливі прийоми, наприклад, алгоритм імітації отжигу.

Стохастичний метод повільніший, але від того, що він не здійснює точного градієнтного спуску, а вносить «шуми», використовуючи градієнт, що розрахований не до кінця, він здатний виходити з локальних мінімумів і може привести до кращого результату.

У вигляді компромісу рекомендують також застосовувати міні-пакети, коли поправка шуканих вагів здійснюється після обробки декількох зразків (міні-пакета), тобто, рідше ніж при стохастичному спуску, але частіше ніж при пакетному.

Висновки до розділу 1

Недоліки в першу чергу полягають у тому, що немає чіткого представлення розподілу генератора за дійсними даними. Також під час навчання дискримінатора його треба добре синхронізувати з генератором (зокрема, генератор не слід навчати без оновлення дискримінатора, щоб уникнути "сценарію Helvetica", в якому генератор згортає забагато значень). Перевагами є те, що не потрібно використовувати Маркові ланцюги, для отримання градієнтів використовується тільки метод зворотного поширення помилок. Крім того, досить багато різноманітних функцій можуть бути включені в модель.

Вищезазначені переваги в основному зменшують обчислювальну складність. Але також в генеративно-змагальних мережах можуть бути статистичні переваги. Наприклад, самі тренувальні дані у процесі навчання не змінюються проходячи через дискримінатор, змінюються тільки градієнти. Іншим плюсом є те, що даний тип мережі може представляти розподіл даних вироджено, у той час як Марковим ланцюгам потрібно щоб розподіл був розмитим.

В якості генератору та дискримінатору можна обирати різні архітектури нейронних мереж. Далі будуть наведені переваги, через які було обрано саме згорткову та розвернуту нейронні мережі.

При використанні згорткових нейронних мереж (ЗНМ) алгоритм використовує порівняно мало попередньої обробки, в порівнянні з іншими алгоритмами класифікування зображень. Це означає, що мережа самостійно знаходить необхідні фільтри, що в традиційних алгоритмах розроблялися вручну. Ця незалежність конструювання ознак від апріорних знань та людських зусиль є великою перевагою. Також згорткові мережі можуть швидко працювати на послідовній машині і швидко навчатися за

рахунок чистого розпаралелювання процесу згортки по кожній карті, а також зворотної згортки при поширенні помилки по мережі.

Застосування розвернутих нейронних мереж (РНМ) дає такі переваги, що для вихідних зображень можна отримати великий набір фільтрів, які охоплюють всю структуру зображення, використовуючи примітивні уявлення. Таким чином, виходять фільтри, що застосовуються до всього зображення, а не до кожного маленького шматочка вихідного зображення. Це є великою перевагою, так як з'являється більш повне розуміння процесів, що відбуваються при навчанні згорнутих нейромереж. Підхід з використанням РНМ заснований на методі пошуку глобального мінімуму, а також використанні фільтрів отриманих при навчанні ЗНМ, і призначений для зведення до мінімуму погано обумовлених витрат, які виникають в згортковому підході.

Використання нейромереж для докингу є новою задачею і практично ніким не імплементувалась. Подібні алгоритми мають бути швидшими за звичайні та можливо точнішими оскільки нейромережа повинна сама визначити правила докинга. Але машинне навчання дуже чутливе до кількості прикладів для навчання, є багато вільних баз даних де можна дістати з'єднані протеїни з лігандами, наприклад на <https://www.rcsb.org/> є близько 80000 експериментально зроблених з'єднань, чого має бути достатньо для навчання.

РОЗДІЛ 2.

РОЗРОБКА СИСТЕМИ ТА ЇЇ ДЕТАЛІ

2.1 Опис алгоритму роботи розробленої системи

Існує багато можливостей електронного зберігання молекул. В нашому випадку маємо трьохвимірні структури молекул, що записані в форматі .pdb. Приклад структури такого файлу зображено на рисунку 2.1.

```

HETATM 32 C3 PSI B 100 -11.370 15.781 3.271 0.50 25.00 C
HETATM 33 O3 PSI B 100 -13.577 15.844 6.493 0.50 25.00 O
HETATM 34 C4 PSI B 100 -12.596 15.131 6.328 0.50 25.00 C
HETATM 35 N4 PSI B 100 -11.629 16.367 4.458 0.50 25.00 N
HETATM 36 O4 PSI B 100 -11.511 11.436 8.275 0.50 25.00 O
HETATM 37 C5 PSI B 100 -12.829 12.077 8.419 0.50 25.00 C
HETATM 38 N5 PSI B 100 -12.539 13.842 6.657 0.50 25.00 N
HETATM 39 O5 PSI B 100 -13.000 12.664 9.710 0.50 25.00 O
HETATM 40 C6 PSI B 100 -13.056 11.697 10.774 0.50 25.00 C
HETATM 41 OS PSI B 100 -10.864 19.015 1.093 0.50 25.00 O
CONNECT 1 2
CONNECT 2 1 8 10
CONNECT 3 11 26 27
CONNECT 4 12 29 30
CONNECT 5 25 32
CONNECT 6 13 34 35

```

Рис. 2.1 – Приклад структури файлу pdb

Для подальшої роботи з даними було вирішено представляти молекули у вигляді кубів. Для кожного типу атомів будується куб, який розбитий сіткою, розмір комірки якої дорівнює 1 Å (Рис. 2.2). В кожній комірці сітки одиницею визначається присутність атому та робиться розмиття за допомогою функції Гауса.

Данні було підготовлено таким чином щоб при тренуванні система отримувала оригінальний куб білка та передбачувала можливе положення в майбутньому ліганда.

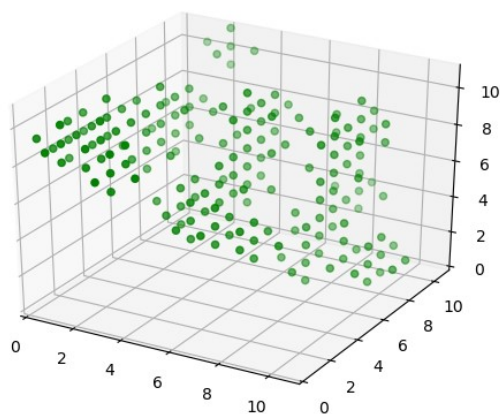


Рис. 2.2 - Візуалізація представлення розташування атомів

У Гауссового розмиття є одна важлива властивість - сепарабельність. Це дає можливість розділити алгоритм на дві частини - розмиття по координаті x і розмиття по y . Таким чином коефіцієнти не потрібно розраховувати для всіх сусідів, досить знайти для одного стовпчика або рядка. Коефіцієнти можна знайти за формулою Гаусса:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (2.1)$$

де μ — математичне сподівання, а σ — дисперсія.

Для молекулярного докінгу за допомогою розробленої моделі застосовано наступний алгоритм: на вхід системи подається окремо молекули ліганда та білка, які представляються у вигляді куба. До моделі подається окремо молекула білка, на виході отримуємо куби, що зображують можливе розташування різних типів атомів ліганда, як зображено на рисунку 2.3. Червоним позначено ймовірні розташування атомів, які були отримані в результаті роботи системи, зеленим — розташування атомів заданого ліганда.

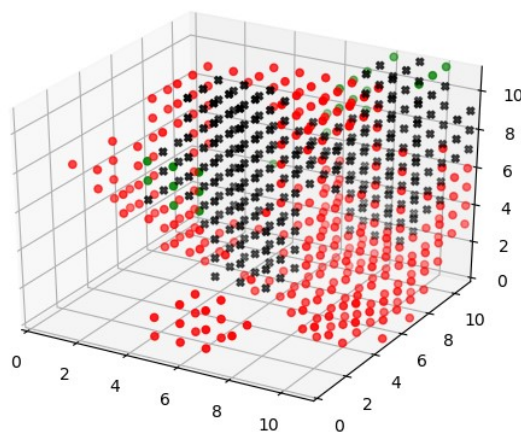


Рис. 2.3 – Візуалізація спрогнозованого та дійсного розташування атомів ліганда

Після чого за допомогою оціночної функції та порівняння розташування згенерованих кубів ліганд та реальних, визначається кінцеве розташування молекули ліганда відносно молекули білка. Причому порівнюється не тільки оригінал молекули ліганда, а й її конформації. Конформація - просторове розташування атомів в молекулі певної конфігурації, обумовлене поворотом навколо однієї або декількох одинарних сигма-зв'язків. Тобто деякі атоми в молекулі можуть змінювати своє розташування. В нашому випадку для знаходження конформацій було використано вже існуючу бібліотеку в python – ProDy.

Далі за допомогою методу induced fit підбирається конформація білка. Алгоритм induced fit робить припущення, що початкова взаємодія між лігандом та білком-мішенню є відносно слабкою, але ці слабкі взаємодії швидко викликають конформаційні зміни в протеїні, що зміцнює зв'язування. Переваги індукованого механізму виникають внаслідок стабілізуючого ефекту міцного зв'язування ферментів. Таким чином, даний алгоритм можна представити у вигляді ключа та замка, але принцип дії зворотній – замість того, що підбирати необхідний ключ до замку, підбирається необхідний замок до ключа. В

розробленій системі моделювання молекулярного докінгу метод *induced fit* реалізований за допомогою ПЗ OpenBabel. Графічне представлення роботи методу *induced fit* можна побачити на рисунку 2.4.

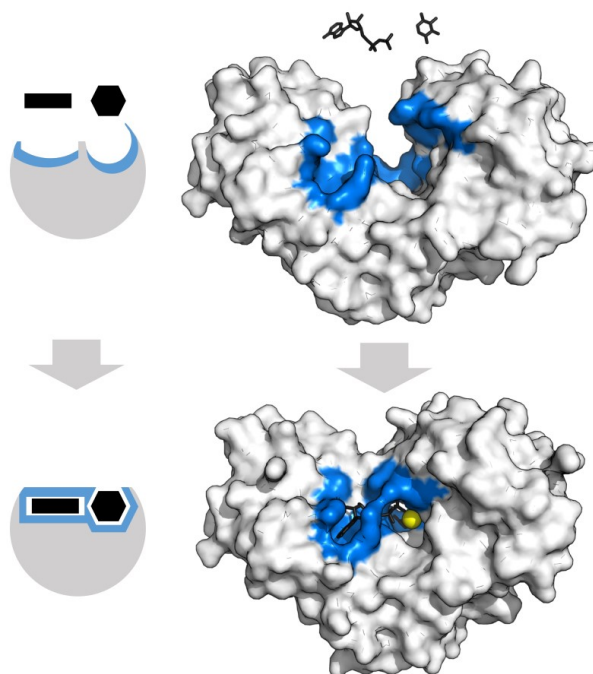


Рис. 2.4 – Графічне представлення роботи методу *induced fit*

На кожному кроці, коли отримуємо один з варіантів розташування ліганда перевіряємо основне правило, яке каже, що енергія сполуки має бути меншою, за суму енергій окремо молекул білка та ліганда:

$$\Delta G(binding) = G(complex) - [G(protein) + G(ligand)] < 0, \quad (2.2)$$

де $G(complex)$ – енергія сполуки, $G(protein)$ – енергія білка, $G(ligand)$ – енергія ліганда. G – енергія Гіббса. це величина, зміна якої в ході хімічної реакції дорівнює зміні внутрішньої енергії системи. Енергія Гіббса показує, яка частина від повної внутрішньої енергії системи може бути використана для хімічних перетворень або отримана в їх результаті в заданих умовах і дозволяє встановити

принципову можливість протікання хімічної реакції в заданих умовах. Математично це термодинамічний потенціал такого вигляду:

$$G = U + PV - TS, \quad (2.3)$$

де U – внутрішня енергія, P – тиск, V – об’єм, T – абсолютна температура, S – ентропія.

Таким чином, повне візуальне представлення розробленої системи молекулярного докінгу можна побачити на рисунку 2.5.

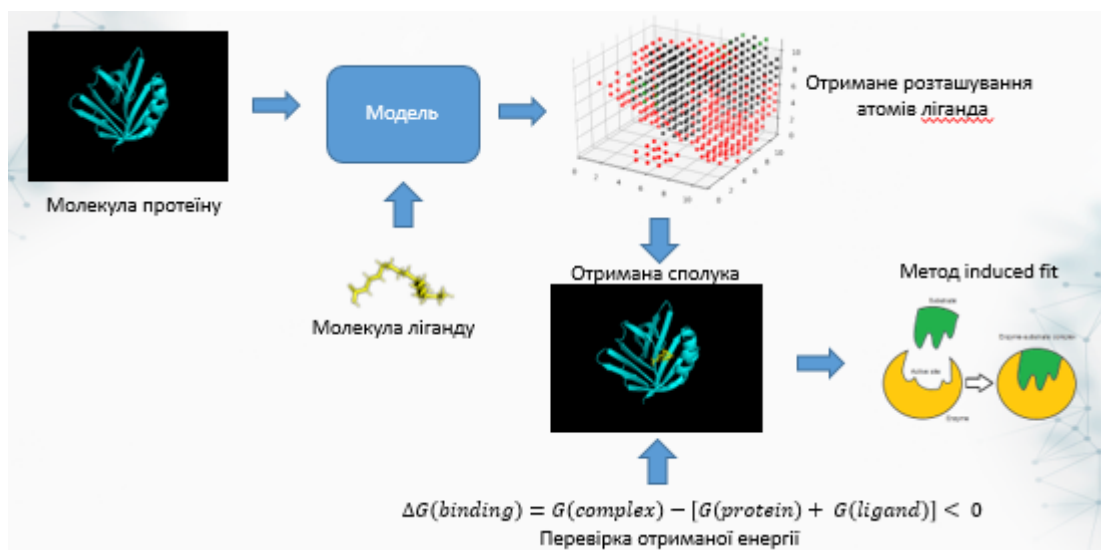


Рис. 2.5 – Схематичне представлення алгоритму роботи розробленої системи

2.2 Навчання системи

Система навчалась на базі даних RCSB яка містить відомі сполуки комплексу білків з лігандами. Використання Гаусового розмиття виявилось необхідним після великої кількості спроб тренування моделей на звичайних кубах. Оскільки куби були заповнені майже всюди нулями, модель швидко навчалась видавати лише нулі. Запровадження розмиття вирішило цю проблему,

але виявилось недостатнім для успішного подальшого навчання. Протягом часу було вирішено застосувати звичайну нейромережу замість зі звичайними нейронами, оскільки згорткові шари та MaxPooling швидко привчались видавати шароподібні відповіді. Це обумовлено тим, що найчастіше скупчення атомів відбувалось як раз в центрі куба.

Звичайна глибока нейромережа зуміла після довгого навчання видавати бажаний результат. Оскільки зрозуміло, що нейромережа не зможе видавати завжди тільки правильні координати, було вирішено що більш сприятливим варіантом має бути випадок коли вона видасть більше більше відповідей ніж є насправді, але цим захопить більшу кількість правильних координат.

Висновки до розділу 2

Була розроблена та навчена система здатня видавати сприятливи відповіді на тестових образцах. Було вдосконалено попередні ідеї та їх реалізація. Була відкинута ідея використання генеративно-змагальної нейромережі оскільки не вдалося отримати адекватних результатів. Майбутня перевірка за стандартним датасетом має показати результати яких вдалось отримати.

РОЗДІЛ 3. АНАЛІЗ ОТРИМАНИХ РЕЗУЛЬТАТІВ

3.1 Порівняння з існуючими системами

Таким чином, за допомогою розробленої системи моделювання молекулярного докінгу отримали досить високу точність, яка взмозі конкурувати із існуючими системними рішеннями. Порівняльний аналіз точності існуючих систем із розробленою представлений в таблиці 6.1.

Таблиця 3.1 Порівняння систем моделювання молекулярного докінгу

PDB code	AutoDock	Flex X	Glide	Gold	Наше моделювання
DC-SIGN 2IT6	3.5	0.61	0.79	1.68	0.97
DC-SIGN 2XR5	4.21	7.26	2.01	4.31	1.22
DC-SIGN 2XR6	5.48	7.44	6.49	6.82	4.34
1aaq	11.76	1.51	2.21	1.76	2.5
1abe	0.31	0.41	0.17	0.29	0.2
2cgr	1.47	1.03	0.39	0.79	0.3
3cra	2.78	1.75	0.85	1.37	1.1

Як бачимо з таблиці 3.1, не можна виділити один найкращий молекулярний докер, для різних молекул отримується зовсім різна точність. Отже, отримали систему, яка за точністю є ліпшою для деяких типів молекул та яка є в рази швидшою.

3.2 Оцінка точності

Тестування алгоритму на різних відомих білка, та оцінка казує що повторний запуск системи може давати кращі результати моделювання. Для приклада в таблиці 3.2 можна побачити різні сполуки та їх RMSD за кожную прогонку системи у порівнянні з відомим експериментальним результатом.

Таблиця 3.2 Порівняння структур комплексу білка з лігандом після чотирьох запусків

ahap	acjp	cbsp	ptbp
3.614	4.209	3.324	4.143
3.821	1.704	2.799	1.358
2.697	3.644	1.645	3.457
2.949	4.624	1.149	2.305

Є два типу оцінок RMSD для подібних систем: коли рахується перший отриманий результат (RMSD1) або коли робиться десять запусків на кожному прикладі і рахується середнє серед найкращих результатів (RMSD2). RMSD2 для добре працюючих систем має становити не більше 2, наша система мала оцінку в 2.5, що є добре, але недостатньо. Зміни в системі на етапі induced fit дозволили зменшити RMSD2 та зменшити кількість негативних відповідей. Подальше покращення можливо в цьому ж напрямку додаванням умов для переміщення ліганда.

Приклади отриманої сполуки та дійсної сполуки представлені на рисунках нижче. На них можна побачити оригінальну структуру комплексу білка з лігандом зверху та результат моделювання у порівнянні з оригіналом.

Як ми бачимо результат моделювання є дуже близьким до оригіналу, але іноді має трохи інші кути нахилу.

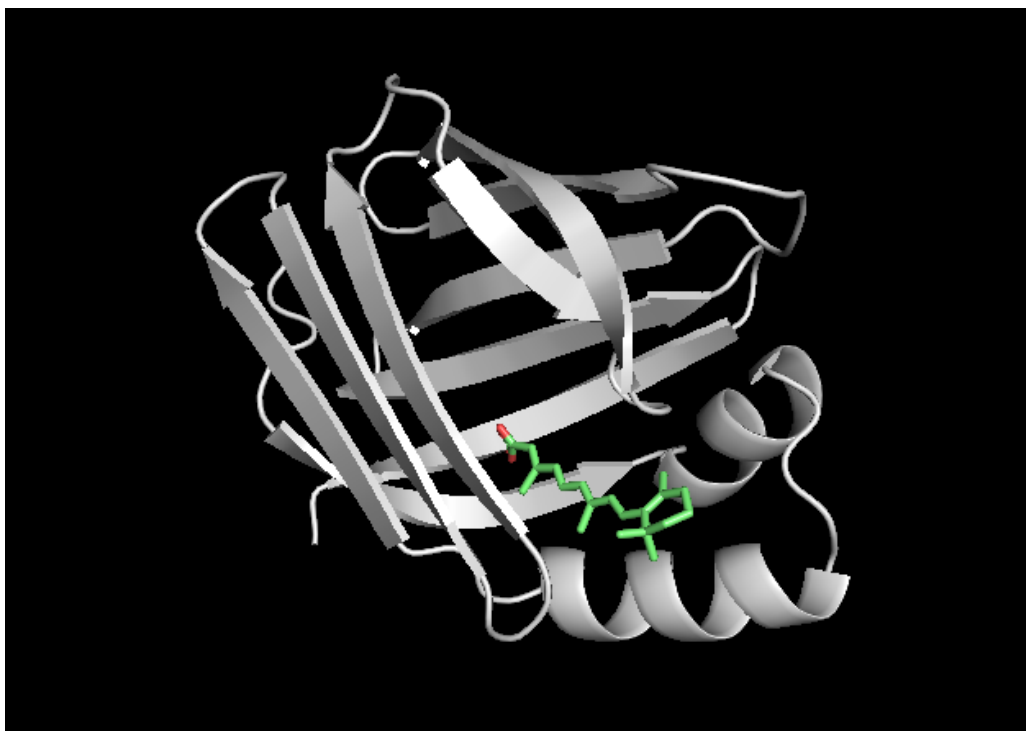


Рис. 3.1 – Дійсна сполука білка та ліганду

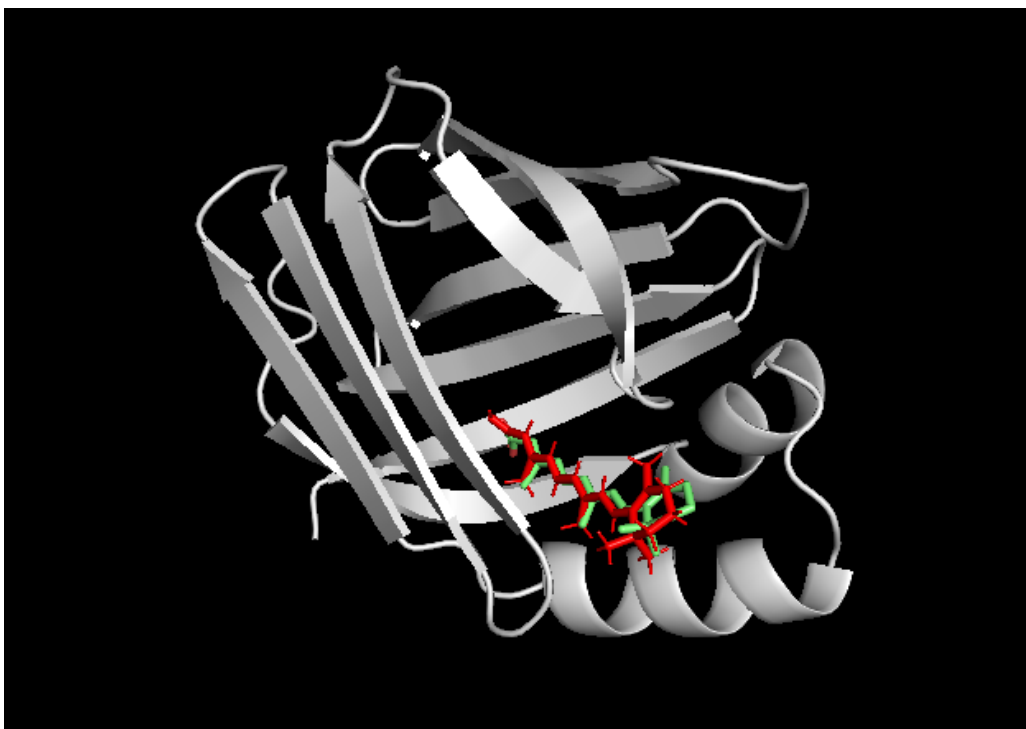


Рис. 3.2 – Дійсна сполука білка з лігандом разом із спрогнозованим розташуванням ліганду

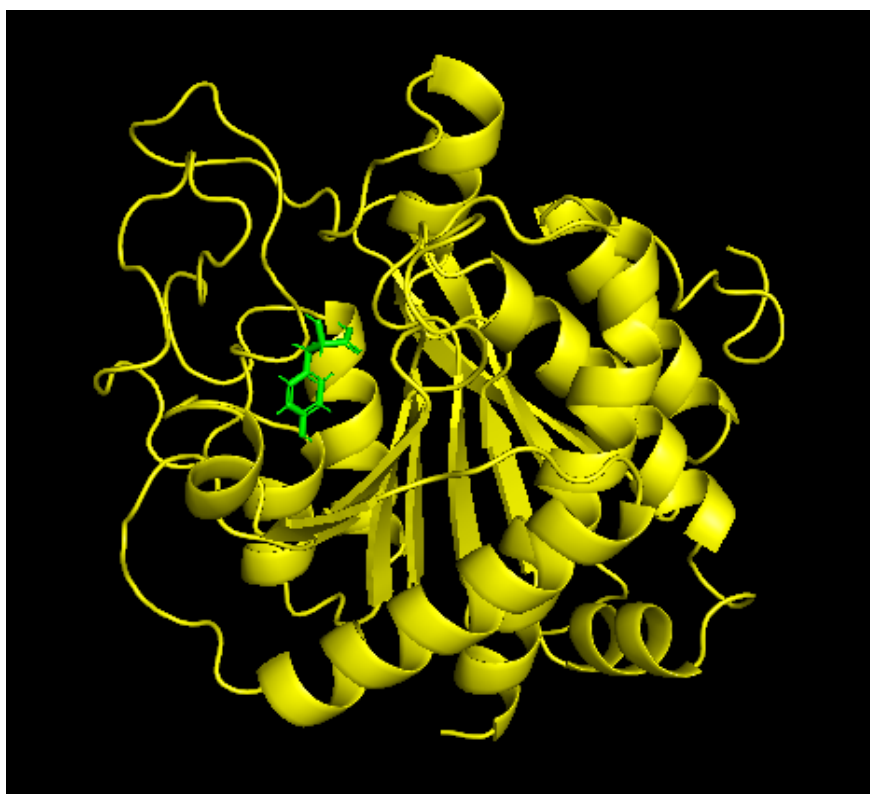


Рис. 3.3 – Дійсна сполука білка та ліганду

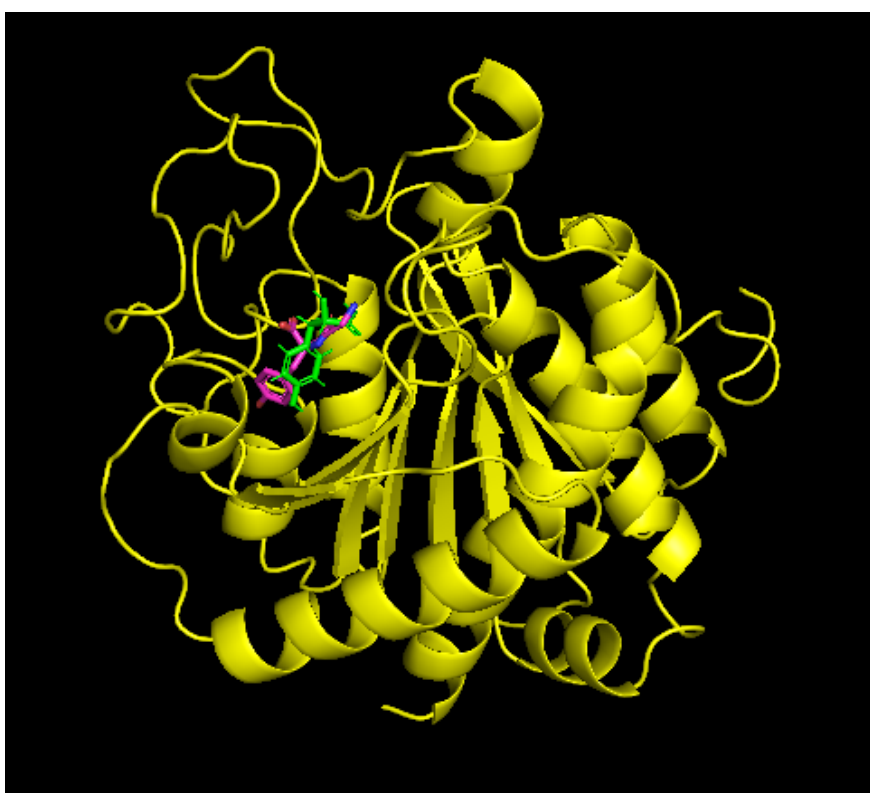


Рис. 3.4 – Дійсна сполука білка з лігандом разом із спрогнозованим розташуванням ліганду

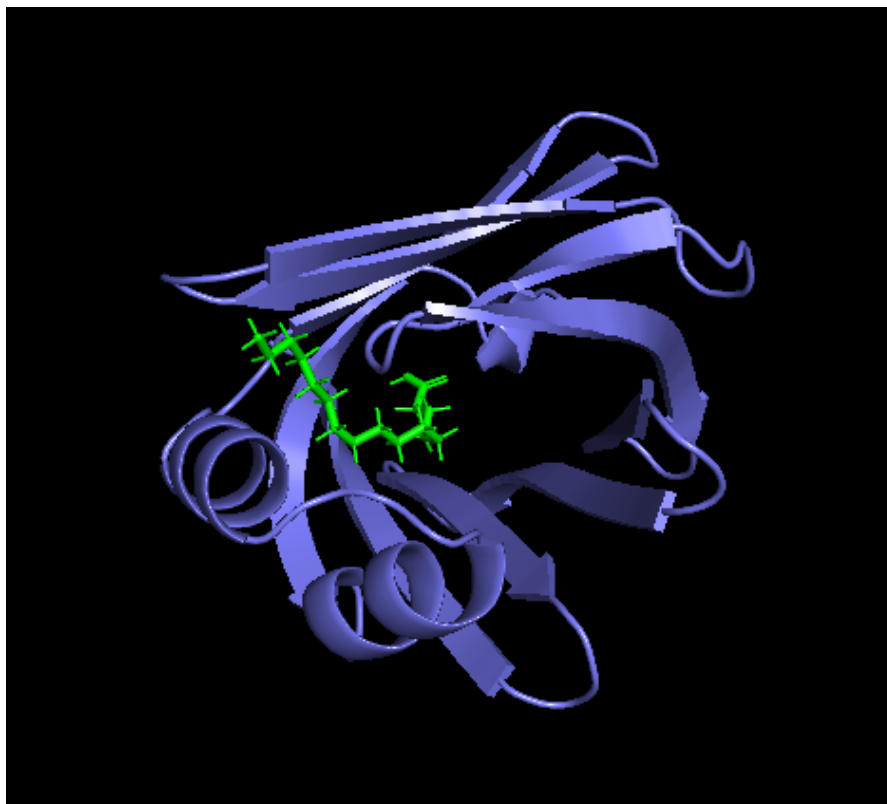


Рис. 3.5 – Дійсна сполука білка та ліганду

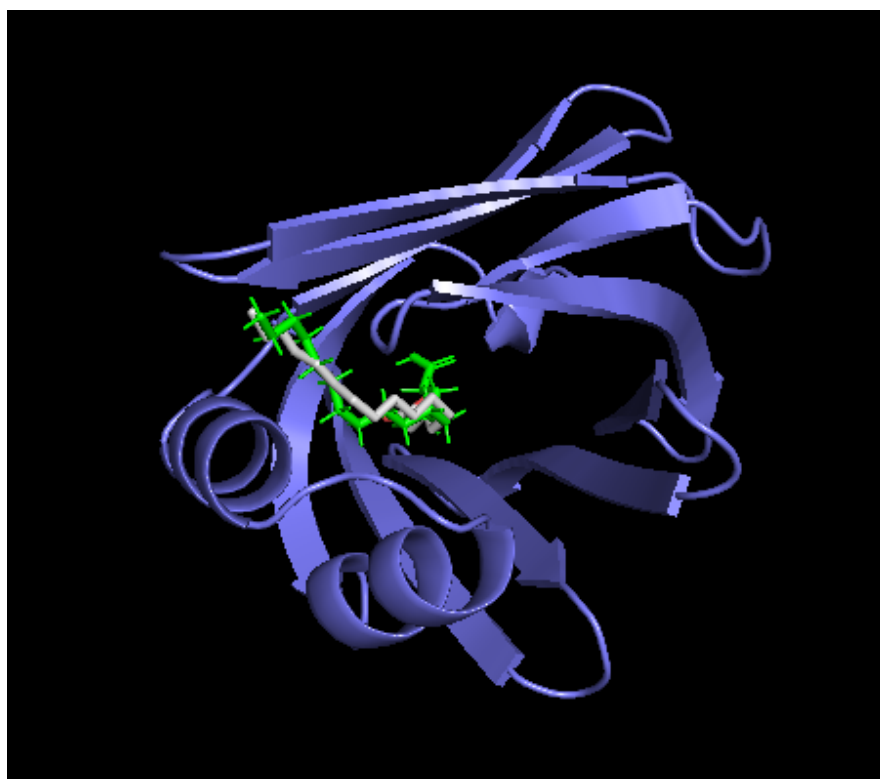


Рис. 3.6 – Дійсна сполука білка з лігандом разом із спрогнозованим розташуванням ліганду

Висновки до розділу 3

Одна із головних невирішених ще проблем є розташування ліганда на 180° відносно відомої експериментально позиції. Не відомо чи є це неправильним розташуванням оскільки іноді ліганди є практично симетричними. Також залишається відкритим питання про те чи дійсно існуюча методика оцінки є коректною оскільки вона оцінює лише по одному відомому результату і чи є це справді єдиним вірним рішенням до проведення реальних експериментів не відомо.

ВИСНОВКИ

Отже, при порівнянні результатів молекулярного докінгу за допомогою стандартних біохімічних алгоритмів із запропонованим алгоритмом використання нейронних мереж було виявлено, що при використанні запропонованого алгоритму процес докінгу проходить в декілька разів швидше. Таким чином, тривалість докінгу зменшується з кількох годин до 20-30 хв. Крім того, запропонована модель узагальнює різні підходи до моделювання, враховує різні взаємозв'язки між атомами для різних моделей, таким чином досліднику не потрібно робити це власноруч.

Але на сьогодні розроблена модель має меншу точність ніж існуючі. У той час як звичайні алгоритми досягають точності у 1-3 Å, запропонована модель дає точність в середньому 2.5 Å, що не є критичним та може бути цілковито поліпшено за допомогою навчання моделі на більшій кількості даних та за допомогою подальшого підбору параметрів.

Загалом, задача по створенню системи будуєчу комплекси структури білка з лігандом відповідно до усіх фізичних вимог за допомогою нейромережі – зроблено. Були проведені тести і проаналізовані результати, що показали що система має достатні можливості для реального використання в наукових цілях у майбутньому. Поставлені цілі можна вважати виконаними оскільки, тривалість процесу моделювання молекулярної структури комплексу білка з лігандом зменшено з кількох годин до 20 хвилин, достатня точність отримана і має всі можливості для подальшого покращення. Використані усі заплановані техніки та інструменти.

Майбутнє покращення має мати на меті зменшення RMSD2 до рівня 2 Å, а якщо можливо, то і RMSD1 такого ж рівня, що буде означати повну відповідність до усіх інших подібних систем, а також її покращення та вигідне використання у наукових цілях.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. McConkey BJ, Sobolev V, Edelman M. The performance of current methods in ligand-protein docking. *Current Science*. 2002;83:845–855.
2. *Joseph Gomes, Bharath Ramsundar, Evan N. Feinberg, and Vijay S. Pande* Atomic Convolutional Networks for predicting Protein-Ligand binding affinity. Режим доступу: <https://arxiv.org/pdf/1703.10603.pdf>
3. Пырков Т.В., Озеров И.В., Балицкая Е.Д., Ефремов Р.Г. (2010), Молекулярный докинг: роль невалентных взаимодействий в образовании комплексов белков с нуклеотидами и пептидами. Режим доступу: [http://www.rjbc.ru/2010/4/2010_36_4\(2\).pdf](http://www.rjbc.ru/2010/4/2010_36_4(2).pdf)
4. Goodfellow, Ian J.; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron & Bengio, Yoshua (2014), "Generative Adversarial Networks", [arXiv:1406.2661](https://arxiv.org/abs/1406.2661)
5. Salimans, Tim; Goodfellow, Ian; Zaremba, Wojciech; Cheung, Vicki; Radford, Alec & Chen, Xi (2016), "Improved Techniques for Training GANs", [arXiv:1606.03498](https://arxiv.org/abs/1606.03498)
6. Betts M.J., Sternberg M.J. // *Protein Eng.* 1999. V. 12. P. 271–283.
7. Zhong H., Tran L.M., Stang J.L. // *J. Mol. Graph. Model.* 2009. V. 28. P. 558–575.
8. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.
9. Kokh D.B., Wenzel W. // *J. Med. Chem.* 2008. V. 51. P. 5919–5931.
10. Ferrara P., Curioni A., Vangrevelinghe E., Meyer T., Mordasini T., Andreoni W., Acklin P., Jacoby E. // *J. Chem. Inf. Model.* 2006. V. 46. P. 254–263.
11. Thomas M.P., McInnes C., Fischer P.M. // *J. Med. Chem.* 2006. V. 49. P. 92–104.

12. Kitchen D.B., Decornez H., Furr J.R., Bajorath J. // *Nat. Rev. Drug. Discov.* 2004. V. 3. P. 935–949.
13. Durrant, J. D. & McCammon, J. A. Nnscore 2.0: a neural-network receptor–ligand scoring function. *Journal of chemical information and modeling* 51, 2897–2903.
14. Moitessier N., Englebienne P., Lee D., Lawandi J., Corbeil C.R. // *Br. J. Pharmacol.* 2008. V. 153. P. S7–S26.
15. Amini A., Shrimpton P.J., Muggleton S.H., Sternberg M.J. // *PROTEINS.* 2007. V. 69. P. 823–831.
16. Wang, R., Lai, L. & Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design* 16, 11–26.
17. Wang, R., Fang, X., Lu, Y. & Wang, S. The pdbbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry* 47, 2977–2980.
18. Catana C., Stouten P.F.W. // *J. Chem. Inf. Model.* 2007. V. 47. P. 85–91.
19. Stroganov O.V., Novikov F.N., Stroylov V.S., Kulkov V., Chilov G.G. // *J. Chem. Inf. Model.* 2008. V. 48. P. 2371–2385.
20. Deepchem: Deep-learning models for drug discovery and quantum chemistry. <https://github.com/deepchem/deepchem>.
21. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design* 30, 595–608.
22. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50, 742–754.
23. LeCun, Y. et al. Comparison of learning algorithms for handwritten digit recognition. In *International conference on artificial neural networks*, vol. 60, 53–60.
24. Wallach, I., Dzamba, M. & Heifets, A. Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery.

25. Durrant, J. D. & McCammon, J. A. Nnscore 2.0: a neural-network receptor–ligand scoring function. *Journal of chemical information and modeling* 51, 2897–2903.
26. Duvenaud, D. K. et al. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, 2224–2232.
27. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics* 134.
28. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* 98.
29. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15.