Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Центр Післядипломної освіти
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Аналіз структур білків за допомогою AlphaFold для дослідження мутацій і їхнього впливу на функції білків у біомедичних застосуваннях
(тема)

Виконав:
здобувач другого року навчання,
групи СШІмзд-23-1
Юлія Довгоселець
(власне ім'я, прізвище)
Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва освітньої програми)
Керівник доц. Каріна Селіванова
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ _____ Олег ЗОЛОТУХІН
(підпис) (власне ім'я, прізвище)

2025 р.

Харківський національний університет радіоелектроніки

Центр Післядипломної освіти

Кафедра Штучного інтелекту

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Довгоселець Юлії Володимирівні
(прізвище, ім'я, по батькові)

1. Тема роботи Аналіз структур білків за допомогою AlphaFold для дослідження мутацій і їхнього впливу на функції білків у біомедичних застосуваннях

затверджена наказом університету від 21 квітня 20 25 р. № 62Стз

2. Термін подання студентом роботи до екзаменаційної комісії 10 червня 20 25 р.

3. Вихідні дані до роботи Публічні білкові бази даних: UniProt, PDB, AlphaFold DB, інструменти глибинного навчання: AlphaFold2, PyMOL, ChimeraX, FASTA-послідовності білків TP53, CFTR, Spike SARS-CoV-2, Lysozyme, відомості про клінічно значущі мутації з баз ClinVar, COSMIC

4. Перелік питань, що потрібно опрацювати в роботі
1) Теоретичні основи дослідження структури та функцій білкі
2) Аналіз можливостей alphafold для дослідження мутацій білків
3) Експериментальна частина дослідження (практична реалізація)
4) Застосування штучного інтелекту в біомедичних дослі-дженнях

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	21.04.2025	виконано
2	Огляд наукових джерел щодо структури білків,	22.04–28.04.2025	виконано
	мутацій та біомедичних застосувань		
3	Вивчення архітектури та принципів роботи	29.04–04.05.2025	виконано
	моделі AlphaFold		
4	Вибір білків (TP53, CFTR, Spike, Lysozyme) і	05.05–07.05.2025	виконано
	мутацій для моделювання		
5	Формування FASTA-послідовностей і підготовка	08.05–10.05.2025	виконано
	до запуску AlphaFold		
6	Запуск AlphaFold на мутантних послідовностях і	11.05–13.05.2025	виконано
	отримання альтернативних структур		
7	Порівняння просторових структур: виявлення	14.05–15.05.2025	виконано
	змін за pLDDT, TM-score, PAE		
8	Візуалізація структур білків і аналіз локальних	16.05–25.05.2025	виконано
	змін у функціональних ділянках (через PyMOL,		
	ChimeraX)		
9	Оформлення результатів дослідження	25.05-28.05.2025	виконано
10	Оформлення пояснювальної записки та	28.05-08.06.2025	виконано
	підготовка презентації до захисту		
11	Захист перед ЕК	10.06.2025	

Дата видачі завдання 21 квітня 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. Каріна Селіванова
(підпис) (посада, власне ім’я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 108 с., 30 рис., 3 табл., 1 дод., 28 джерел.

БІОМЕДИЦИНА, МУТАЦІЇ, СТРУКТУРА БІЛКА, ШТУЧНИЙ ІНТЕЛЕКТ, ALPHAFOLD.

Об’єкт дослідження – просторові структури білків людини, що змінюються під впливом мутацій.

Предмет дослідження – аналіз впливу мутацій на структуру та функції білків за допомогою алгоритму AlphaFold.

Мета роботи – здійснити структурно-функціональний аналіз білків на основі передбачених моделей AlphaFold для оцінки впливу точкових мутацій у біомедичному контексті.

Проведено огляд сучасних підходів до аналізу білкових структур і алгоритмів глибинного навчання, зокрема AlphaFold. Описано архітектуру моделі та особливості її роботи з білками з різними типами мутацій. Змодельовано просторові структури чотирьох білків (TP53, CFTR, Spike SARS-CoV-2, лізоцим) у диких і мутантних формах.

Виконано порівняння метрик pLDDT, pTM, PAE та TM-score між вихідними і мутантними структурами, виявлено зміни у функціонально важливих ділянках. Проаналізовано вплив мутацій на біологічну активність білків із використанням інструментів FoldX, Rosetta, SIFT, PolyPhen та засобів візуалізації ChimeraX і PyMOL.

AlphaFold підтвердив ефективність у передбаченні структур і початковому аналізі мутацій. Результати можуть бути застосовані для ідентифікації терапевтичних мішеней, діагностики та персоналізованої медицини.

ABSTRACT

Master’s thesis contains: 108 pp., 30 fig., 3 tabl., 1 ann., 28 references.

ALPHAFOLD, ARTIFICIAL INTELLIGENCE, BIOMEDICINE, MUTATIONS, PROTEIN STRUCTURE.

Object of the study – the three-dimensional structures of human proteins altered by mutations.

Subject of the study – analysis of the impact of mutations on the structure and functions of proteins using the AlphaFold algorithm.

Aim of the study – to perform a structural and functional analysis of proteins based on AlphaFold-predicted models to assess the effects of point mutations in a biomedical context.

A review of current approaches to protein structure analysis and deep learning algorithms, particularly AlphaFold, was conducted. The model's architecture and its handling of proteins with different mutation types are described. Spatial structures of four proteins (TP53, CFTR, SARS-CoV-2 Spike, and lysozyme) were modeled in both wild-type and mutant forms.

Comparative analysis of pLDDT, pTM, PAE, and TM-score metrics between wild-type and mutant structures revealed changes in functionally significant regions. The potential biological impact of mutations was evaluated using stability tools (FoldX, Rosetta), functional significance predictors (SIFT, PolyPhen), and visualization software (ChimeraX, PyMOL).

AlphaFold demonstrated effectiveness in protein structure prediction and preliminary mutation analysis. The results can be applied to identifying therapeutic targets, mutation diagnostics, and personalized medicine.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

AI – Artificial Intelligence – штучний інтелект;

AlphaFold – алгоритм глибинного навчання для передбачення струк тури білків;

CFTR – Cystic Fibrosis Transmembrane Conductance Regulator – трансмембранний регулятор провідності при муковісцидозі;

ChimeraX – програмне забезпечення для візуалізації та аналізу структур макромолекул;

ClinVar – база даних клінічно значущих варіантів генів;

COSMIC – Catalogue Of Somatic Mutations In Cancer – каталог соматичних мутацій у раку;

FASTA – текстовий формат для представлення амінокислотних або нуклеотидних послідовностей;

FoldX – програмний інструмент для аналізу стабільності білків;

Mut – Mutant – мутантна форма білка;

PAE – Predicted Aligned Error – прогнозована вирівняна похибка;

PDB – Protein Data Bank – банк даних білкових структур;

PolyPhen-2 – Polymorphism Phenotyping v2 – інструмент для оцінки функціональної значущості мутацій;

pLDDT – predicted Local Distance Difference Test – локальний показник достовірності структури;

pTM – predicted Template Modeling score – прогнозований показник схожості структур;

PyMOL – програмне забезпечення для візуалізації білкових структур;

Rosetta – програмний пакет для моделювання та дизайну білків;

SARS-CoV-2 – Severe Acute Respiratory Syndrome Coronavirus 2 – коронавірус, що викликає COVID-19;

SIFT – Sorting Intolerant From Tolerant – інструмент для прогнозування впливу мутацій;

TM-score – Template Modeling Score – метрика схожості білкових структур;

TP53 – Tumor Protein p53 – білок-супресор пухлин;

UniProt – Universal Protein Resource – універсальна база даних білків;

WT – Wild Type – дикий тип (нормальна, немутована форма білка).

ВСТУП

Актуальність теми дослідження. У сучасній біології та медицині однією з центральних задач є глибоке розуміння механізмів функціонування білків, адже саме вони є фундаментом для більшості біологічних процесів, що відбуваються в живих організмах. Білки виконують широкий спектр функцій: вони є структурними компонентами клітин, каталізують хімічні реакції, беруть участь у сигнальних каскадах, імунних відповідях, забезпечують транспорт речовин та багато інших важливих біологічних функцій. Ключовим фактором, що визначає функцію білка, є його просторова структура. Навіть незначні структурні зміни можуть суттєво впливати на властивості та функціональну активність білкових молекул, що є особливо актуальним при дослідженні наслідків мутацій, які часто лежать в основі багатьох захворювань.

Мутації – це постійні зміни нуклеотидної послідовності ДНК, які можуть призводити до заміни, видалення чи вставки амінокислот у складі білків. Вони здатні як порушувати, так і модифікувати нормальну структуру і функцію білків, іноді призводячи до серйозних патологічних станів, таких як спадкові хвороби, онкологічні захворювання або метаболічні порушення. Важливою проблемою сучасної біомедицини є швидке та точне прогнозування впливу таких мутацій на функції білків, що дозволяє ефективніше розробляти нові діагностичні та терапевтичні підходи.

Одним із перспективних сучасних методів аналізу білкових структур є застосування алгоритмів глибокого навчання, зокрема AlphaFold – інноваційної технології, яка революціонізувала обчислювальну біологію та структурну біоінформатику. AlphaFold, розроблений компанією DeepMind, продемонстрував безпрецедентну точність передбачення тривимірних структур білків за їх амінокислотними послідовностями, значно наблизивши обчислювальні методи до точності експериментальних (рентгеноструктурного аналізу, кріоелектронної мікроскопії, ЯМР-

спектроскопії). Використання AlphaFold надає можливість не лише швидкого визначення структури малодосліджених білків, але й ефективного аналізу структурних та функціональних наслідків мутацій, що відкриває нові перспективи для біомедичних досліджень.

Метою кваліфікаційної роботи є аналіз впливу мутацій на просторову структуру та функціональні властивості білків за допомогою алгоритму AlphaFold, а також оцінка перспектив застосування отриманих результатів у біомедичних цілях.

Для досягнення поставленої мети визначено наступні завдання:

– провести огляд сучасних підходів до аналізу структур і функцій білків, особливо у контексті мутацій;

– дослідити методологію AlphaFold та порівняти її ефективність із традиційними експериментальними та іншими обчислювальними методами;

– реалізувати експериментальне дослідження просторової структури вибраних білків із заданими мутаціями;

– провести детальний аналіз та інтерпретацію отриманих результатів щодо структурно-функціональних змін білків;

– оцінити потенціал застосування результатів у практичних біомедичних завданнях, зокрема для розробки таргетних терапій та персоналізованої медицини.

Об'єктом дослідження є білкові структури, що зазнають змін під впливом мутацій, предметом є вплив мутацій на структуру і функції білків, проаналізований за допомогою технології AlphaFold.

У кваліфікаційній роботі використовуються методи структурної біоінформатики, комп’ютерного моделювання, аналізу даних та штучного інтелекту. Наукова новизна дослідження полягає у використанні передових методів прогнозування структур білків із застосуванням AlphaFold для поглибленого аналізу впливу мутацій на структурно-функціональні властивості білків, що має важливе значення для розширення можливостей

біомедичної науки та персоналізованої медицини. Отримані результати мають практичне значення, оскільки відкривають перспективи їх використання для розробки нових терапевтичних стратегій, індивідуалізованої діагностики та лікування захворювань, пов’язаних із білковими мутаціями.

Кваліфікаційна робота структурно складається зі вступу, чотирьох розділів, висновків, списку використаних джерел і додатків.

1 ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ СТРУКТУРИ ТА ФУНКЦІЙ БІЛКІВ

1.1 Структура білків

Білки є макромолекулами, які складаються з лінійних ланцюгів амінокислот, пов'язаних між собою специфічними ковалентними зв'язками – пептидними. У складі білків організмів зустрічається близько 20 основних амінокислот, кожна з яких характеризується індивідуальною структурою та фізико-хімічними властивостями. Амінокислоти відрізняються за будовою бічних ланцюгів (радикалів), які визначають їхню полярність, гідрофобність або гідрофільність, кислотно-основні властивості, заряд при фізіологічних значеннях pH тощо. Ці характеристики впливають на структуру, стабільність та функціональні можливості білкових молекул. Структура білків визначається на кількох взаємопов'язаних рівнях: первинному, вторинному, третинному та четвертинному [1].

Первинна структура білка має вигляд простої лінійної послідовності амінокислотних залишків, порядок яких суворо визначений генетичною інформацією, закодованою в молекулі ДНК. Зміни навіть у єдиному амінокислотному залишку можуть значно впливати на властивості й функції білка, спричиняючи різноманітні порушення та захворювання.

Вторинна структура – це впорядковане згортання локальних ділянок поліпептидного ланцюга, що утворюється завдяки водневим зв’язкам між атомами карбонільних і аміногруп у сусідніх амінокислотах. До таких структур належать альфа-спіралі-стабільні правозакручені спіральні утворення, а також бета-листи-структури, в яких окремі сегменти білкової молекули розташовані паралельно або антипаралельно, утворюючи плоскі складки. Окрім цих основних форм, у білкових молекулах зустрічаються

петлі та повороти, які забезпечують гнучкість та зміну напряму поліпептидного ланцюга.

Наступний рівень організації третинна структура, являє собою унікальну тривимірну конфігурацію білкової молекули, яка формується внаслідок складних взаємодій між бічними радикалами амінокислотних залишків. Серед таких взаємодій найважливішими є гідрофобні сили, водневі зв'язки, електростатичні притягання і відштовхування, а також ковалентні дисульфідні зв’язки, що утворюються між атомами сірки залишків цистеїну. Третинна структура визначає функціональні властивості білків, їхню здатність взаємодіяти з іншими молекулами і брати участь у біологічних процесах.

Для багатьох білків характерна також четвертинна структура, яка утворюється у випадку, коли кілька окремих білкових молекул (субодиниць) об'єднуються у єдиний функціональний комплекс. Субодиниці взаємодіють між собою через нековалентні зв'язки, забезпечуючи особливі функціональні властивості комплексу, наприклад, у гемоглобіні, який складається з чотирьох субодиниць.

1.2 Сучасні підходи до класифікації білків та їх основні функції в організмі

Класифікація білків є важливою складовою вивчення їх структури та функцій, адже дозволяє систематизувати величезну кількість цих молекул за певними спільними ознаками. Існує кілька основних підходів до класифікації білків, які ґрунтуються на різних критеріях: будові, хімічному складі, формі, функціях і походженні. Один з найбільш традиційних підходів – класифікація за хімічним складом. У цьому випадку білки поділяються на прості та складні. Прості білки складаються лише з амінокислотних залишків і не мають додаткових компонентів. До них, наприклад, належать альбуміни та глобуліни. Складні білки, або

кон’юговані, містять окрім амінокислот ще й небілкову частину – так звану простетичну групу. Це можуть бути іони металів, вуглеводи, ліпіди або нуклеїнові кислоти. До цієї групи належать, глікопротеїни, фосфопротеїни та гемопротеїни.

Інший підхід класифікація за формою молекули, згідно з якою білки поділяють на глобулярні та фібрилярні. Глобулярні білки мають компактну, округлу форму й добре розчиняються у воді. Вони найчастіше виконують динамічні функції: ферментативну, транспортну, захисну тощо. Прикладом є ферменти, гормони, антитіла. Фібрилярні білки, навпаки, мають видовжену, ниткоподібну форму і переважно виконують структурні функції. До них належать колаген, кератин, еластин.

Також, існує функціональна класифікація, яка базується на біологічній ролі білка в організмі. В межах цього підходу білки умовно поділяють на ферменти, структурні, транспортні, рецепторні, захисні, регуляторні тощо. Такий підхід широко використовується в біохімії, фармакології та медичних дослідженнях, оскільки дозволяє зв'язати структуру білка з його дією на клітинному або системному рівні. З розвитком технологій та обчислювальної біології стали можливими й класифікації на основі еволюційного споріднення – коли білки групуються за схожістю амінокислотної послідовності, структури або функцій. Цей підхід лежить в основі таких баз даних, як SCOP чи CATH, які систематизують відомі білкові структури на основі спільного походження та будови.

Білки відповідають за виконання численних життєво важливих функцій. Їх різноманітність обумовлюється унікальною структурою та властивостями кожного конкретного білка.

Каталітична функція забезпечується особливими білками-ферментами, які пришвидшують перебіг хімічних реакцій у клітинах та тканинах.

Без ферментів більшість біохімічних процесів в організмі відбувалися б занадто повільно для підтримки життя. Наприклад, амілаза бере участь у перетравлюванні вуглеводів, а ДНК-полімераза забезпечує синтез нових ланцюгів ДНК.

Структурна функція білків проявляється у створенні каркасів, що підтримують форму та механічну міцність клітин і тканин. До таких білків належить колаген, який утворює основу сполучних тканин (кісток, сухожиль), та еластин, що забезпечує еластичність судин і шкіри.

Транспортна функція здійснюється білками, що переносять речовини всередині організму. Типовим прикладом є гемоглобін-білок, який переносить кисень з легень до тканин, та альбумін, який транспортує жирні кислоти, вітаміни й гормони у крові.

Захисна (імунна) функція реалізується білками-антитілами (імуноглобулінами), які розпізнають і нейтралізують чужорідні агенти-віруси, бактерії або токсини, захищаючи організм від інфекційних захворювань.

Регуляторна функція білків полягає у контролі різних біологічних процесів. Наприклад, інсулін регулює рівень глюкози у крові, а фактори транскрипції регулюють активність генів, визначаючи, які білки повинні бути синтезовані клітиною.

Скорочувальна функція білків забезпечує рухливість та м’язову активність організму. Основні білки, що відповідають за це-актін та міозин, взаємодія яких забезпечує скорочення м’язових волокон.

Білки відіграють ключову роль у всіх процесах життєдіяльності, починаючи від молекулярних реакцій і закінчуючи функціонуванням організму як цілісної системи. Втрата або порушення функцій білків через генетичні мутації чи хвороби часто призводить до серйозних порушень у роботі організму.

1.3 Методи визначення та аналізу просторової структури білків

Просторова будова білка визначає, як саме він взаємодіє з іншими молекулами, які ділянки відповідальні за біологічну активність, а також те, як змінюється його поведінка при мутаціях. З цієї причини протягом десятиліть були розроблені різні методи для дослідження структур білків. Їх можна умовно поділити на експериментальні та обчислювальні.

Експериментальні методи:

– рентгеноструктурний аналіз (X-ray кристалографія) – один з найдавніших та найбільш точних методів. Він може з великою роздільною здатністю визначати розташування атомів у кристалі білка. Суть методу полягає у тому, що кристал білка опромінюється рентгенівськими променями, а отримана дифракційна картина дозволяє за допомогою спеціальних програм реконструювати модель структури. Основним недоліком є потреба у вирощуванні якісного кристалу, що для багатьох білків є дуже складним або неможливим [4];

– ядерний магнітний резонанс (ЯМР-спектроскопія) – метод, який дає можливість вивчати білки у розчині, що наближає умови дослідження до природних. Він базується на властивості атомних ядер (зазвичай водню) змінювати своє енергетичне положення під дією магнітного поля. Метод дозволяє вивчати не лише статичну структуру, а й динаміку молекули. Водночас, ЯМР має обмеження за розміром молекул – великі білки (>30–40 кДа) аналізувати таким способом складно;

– кріоелектронна мікроскопія (кріо-ЕМ) – сучасна технологія, що активно розвивається. Дозволяє отримувати тривимірні структури білкових комплексів без необхідності кристалізації. Зразки швидко заморожують, після чого аналізують за допомогою електронного мікроскопа. Кріо-ЕМ особливо ефективна для вивчення великих білкових комплексів і мембранних білків. У поєднанні з новими алгоритмами обробки зображень вона дає змогу отримувати моделі з високою точністю.

Обчислювальні методи:

– гомологічне моделювання базується на принципі, що білки зі схожими амінокислотними послідовностями мають подібні просторові структури. Якщо структура одного з таких білків відома, її можна використати як шаблон для побудови моделі іншого – менш дослідженого. Це один із найпоширеніших підходів у біоінформатиці, однак його точність залежить від наявності подібних структур у базах даних;

– молекулярна динаміка вивчає поведінку білка у часі, симулюючи рух атомів і молекул під дією фізичних законів. Цей підхід корисний для вивчення гнучких ділянок білкової структури, конформаційних змін, процесів зв'язування лігандів тощо. Але він є обчислювально дорогим, особливо при моделюванні великих молекул або довготривалих процесів;

– методи на основі штучного інтелекту, зокрема AlphaFold, стали справжнім проривом у сфері структурної біології. Цей алгоритм, розроблений DeepMind, навчається на великій кількості експериментальних структур і за послідовністю амінокислот може передбачити тривимірну структуру білка з високою точністю – подекуди на рівні, близькому до рентгеноструктурного аналізу. AlphaFold відкрив нові можливості для дослідження білків, структура яких раніше була невідома або складна для експериментального вивчення.

Кожен метод має свої переваги й обмеження, тому на практиці часто застосовують їх у поєднанні. Наприклад, обчислювальні підходи можуть дати перше уявлення про структуру, яку потім уточнюють за допомогою експериментальних методів.

Саме завдяки таким комплексним підходам стало можливим глибше розуміння біологічних функцій білків, а також механізмів розвитку хвороб, пов'язаних з мутаціями або порушенням структурної цілісності білкових молекул.

1.4 Мутації білків та їх біологічне значення

Мутації – це постійні зміни в послідовності нуклеотидів ДНК, які можуть виникати спонтанно або під впливом різноманітних зовнішніх факторів, таких як ультрафіолетове випромінювання, хімічні речовини, іонізуюче випромінювання чи вірусні інфекції. Їх наявність може мати як нейтральний, так і серйозний вплив на структуру та функцію білків, особливо якщо зміни зачіпають кодувальні ділянки геному.

1.4.1 Типи мутацій (точкові, делеції, інсерції)

Серед найпоширеніших типів мутацій виділяють:

а) точкові мутації – це зміни, що стосуються одного нуклеотиду в послідовності ДНК. Вони поділяються на:

– міссенс-мутації (missense) – коли заміна одного нуклеотиду призводить до зміни амінокислоти в білку. Наприклад, заміна аденіну на гуанін у кодоні може призвести до включення іншої амінокислоти в поліпептидний ланцюг. Наслідки залежать від того, наскільки критичною є ця амінокислота для структури або функції білка;

– синонімічні або «мовчазні» мутації – заміна нуклеотиду не змінює амінокислоту через виродженість генетичного коду (тобто декілька кодонів можуть кодувати одну і ту ж амінокислоту). Зазвичай такі мутації не впливають на білок, але іноді можуть змінювати швидкість трансляції чи складання білка;

– нонсенс-мутації – заміна приводить до утворення стоп-кодону, що перериває синтез білка передчасно. Це, як правило, призводить до утворення нефункціонального або нестабільного білка;

б) делеції – це втрата одного або кількох нуклеотидів у послідовності ДНК. Якщо кількість видалених нуклеотидів не кратна трьом, це спричиняє

зсув рамки зчитування (frameshift), що повністю змінює зміст подальших кодонів. Така мутація часто призводить до передчасного стоп-кодону й синтезу короткого нефункціонального білка. Навіть видалення однієї пари основ може мати катастрофічні наслідки для структури й функції білка. У деяких випадках делеції не змінюють рамку зчитування, але призводять до втрати важливих амінокислот. Це може порушити формування активного центру ферменту або змінити просторову структуру білка, що теж вплине на його функціональність;

в) інсерції – це вставки одного або кількох зайвих нуклеотидів у генетичний код. Як і у випадку з делеціями, якщо вставка не кратна трьом, це спричиняє зсув рамки зчитування, що зазвичай веде до створення нефункціонального білка. Інсерції можуть бути короткими (1–2 пари основ) або досить великими, що іноді навіть включає додавання цілих генетичних елементів або повторів.

Кожен із цих типів мутацій впливає на структуру та функцію білків по-різному: від зовсім незначного ефекту до повної втрати білкової активності [2]. Особливо важливими є ті зміни, які трапляються у функціонально важливих ділянках білка – таких як активні центри ферментів, зв’язувальні ділянки або трансмембранні сегменти.

1.4.2 Біологічні механізми виникнення мутацій

Мутації виникають у результаті змін у структурі ДНК і можуть відбуватися як під впливом зовнішніх факторів, так і внаслідок внутрішньоклітинних процесів. У нормі клітина має потужні механізми контролю та відновлення генетичного матеріалу, однак ці системи не є абсолютно безпомилковими. У деяких випадках порушення ДНК залишаються неусунутими або виправляються некоректно – саме так і з’являються мутації.

До спонтанних мутацій належать ті, що виникають без впливу зовнішніх мутагенів. Їх причинами можуть бути:

– помилки реплікації ДНК. Під час поділу клітини ДНК подвоюється, і на цьому етапі можливі випадкові заміни нуклеотидів. Хоча фермент ДНК-полімераза має механізм перевірки правильності (proofreading), не всі помилки вдається виправити;

– таутомерні зсуви. Деякі нуклеотиди можуть тимчасово змінювати свою хімічну форму (таутомерний стан), що призводить до неправильного парування під час реплікації – наприклад, замість гуаніну приєднується тимін.

– самовільне хімічне перетворення основ. Наприклад, дезамінування цитозину призводить до утворення урацілу, який сприймається як тимін. Такі реакції відбуваються природно і можуть бути джерелом точкових мутацій;

– спонтанні розриви фосфодіефірного зв’язку або депуринізація, коли втрачається одна з азотистих основ – аденін або гуанін, що робить зчитування ДНК неточним.

Індуковані мутації спричиняються впливом зовнішніх чинників – так званих мутагенів. Їх поділяють на фізичні, хімічні та біологічні. Фізичні мутагени:

– іонізуюче випромінювання (рентгенівське, гамма-випромінювання) викликає розриви в ланцюгу ДНК, зсуви рамки зчитування або навіть втрату цілих генів;

– ультрафіолетове випромінювання викликає утворення піримідинових димерів (найчастіше – тимінових), які порушують нормальне зчитування інформації під час реплікації.

Хімічні мутагени:

– алкілуючі агенти (наприклад, етілметансульфонат) змінюють структуру азотистих основ, що призводить до помилкового парування;

– інтеркалюючі речовини (акридин, етилбромід) вставляються між основами ДНК, спричиняючи делеції або інсерції;

– окислювальні агенти ушкоджують нуклеотиди й порушують хімічну стабільність молекули.

Біологічні фактори: деякі віруси можуть інтегрувати свою ДНК у геном клітини-господаря, змінюючи нормальну послідовність. Наприклад, ретровіруси мають власну зворотну транскриптазу, яка може спричиняти мутації під час вбудовування вірусної ДНК у клітинний геном.

У нормі клітини мають спеціальні ферментативні системи, що стежать за цілісністю геному та усувають пошкодження. Якщо ці системи не функціонують належним чином – наприклад, через спадкові дефекти в генах репарації – мутації накопичуються значно швидше. Це особливо характерно для деяких форм спадкового раку (наприклад, синдром Лінча або мутації в гені BRCA1/2).

1.4.3 Вплив мутацій на структуру і функції білків

Мутації в генетичному матеріалі можуть мати суттєвий вплив на структуру білків, а отже, і на їхню функціональну активність. Оскільки послідовність амінокислот у білку прямо залежить від послідовності нуклеотидів у відповідному гені, будь-яке порушення цієї послідовності потенційно змінює властивості готової білкової молекули.

Найчастіше мутації призводять до заміни однієї амінокислоти на іншу. У деяких випадках така заміна не має серйозних наслідків – наприклад, якщо нова амінокислота має подібні хімічні властивості або не впливає на ключові ділянки білка. Однак коли зміна зачіпає активний центр ферменту або регіон, відповідальний за стабільність просторової структури, наслідки можуть бути критичними [1]. Білок може втратити свою здатність зв’язуватися з субстратом або перестати нормально згортатися у тривимірну структуру, необхідну для його дії.

Особливо небезпечними вважаються мутації, які змінюють гідрофобні взаємодії або порушують утворення водневих зв’язків, що підтримують стабільність третинної структури білка. Якщо порушується правильне згортання, молекула білка може залишатися у нестабільному або навіть розгорнутому стані, що у свою чергу призводить до її швидкої деградації або утворення нерозчинних агрегатів. У багатьох нейродегенеративних захворюваннях, таких як хвороба Альцгеймера або хорея Гентінгтона, саме така агрегація білків є однією з основних причин ураження клітин. Варто також враховувати мутації, які змінюють рамку зчитування. Це може призвести до синтезу білка з абсолютно іншою амінокислотною послідовністю або навіть до передчасної появи стоп-кодону. У такому випадку утворюється укорочена форма білка, яка зазвичай є нефункціональною і нестабільною. Мутації можуть впливати на взаємодію білків з іншими молекулами – наприклад, рецептори можуть втратити здатність зв’язуватися з відповідними лігандами, транспортні білки – перестати переносити певні речовини, а регуляторні білки – не зможуть виконувати свою функцію в клітинних сигнальних каскадах. Це може призводити до порушень клітинної регуляції, запуску неконтрольованого поділу клітин або інших патологічних змін.

1.4.4 Патології, пов’язані з білковими мутаціями (приклади конкретних захворювань)

Порушення у структурі білків, зумовлені мутаціями в ДНК, часто лежать в основі важких спадкових та набутих хвороб. У багатьох випадках навіть незначна зміна в послідовності амінокислот може порушити правильне згортання білка, змінити його стабільність або повністю позбавити функціональної активності. Це, у свою чергу, може спричинити порушення життєво важливих процесів в організмі, адже білки беруть участь майже у всіх біохімічних реакціях і клітинних механізмах.

Одним із найвідоміших прикладів є серповидноклітинна анемія. Причиною цього захворювання є точкова мутація в гені, що кодує один із ланцюгів гемоглобіну. В результаті замість глутамінової кислоти в амінокислотному ланцюзі з’являється валін. Через цю заміну змінюється фізико-хімічна властивість ділянки білка, і в умовах низького вмісту кисню молекули гемоглобіну починають злипатися між собою. Це призводить до деформації еритроцитів: вони втрачають свою звичну форму і набувають вигляду серпа (рисунок 1.1). Такі клітини менш гнучкі, можуть закупорювати капіляри, порушуючи кровопостачання органів, що призводить до болю, пошкодження тканин і поступового зниження функцій внутрішніх органів.

Рисунок 1.1 – Порівняння серпоподібних та здорових еритроцитів

Ще одним прикладом є муковісцидоз – тяжке спадкове захворювання, яке вражає дихальну та травну системи. Його спричиняє мутація в гені CFTR, який кодує білок-регулятор транспорту хлоридних іонів через клітинну мембрану. У найбільш поширеному випадку відбувається делеція трьох нуклеотидів, унаслідок чого зникає одна амінокислота – фенілаланін. Білок із такою мутацією не проходить правильного згортання та не потрапляє до поверхні клітини, де повинен функціонувати. Через це порушується іонний обмін, слиз у дихальних шляхах стає надто густим, що створює умови для хронічних інфекцій і постійного запалення (рисунок 1.2).

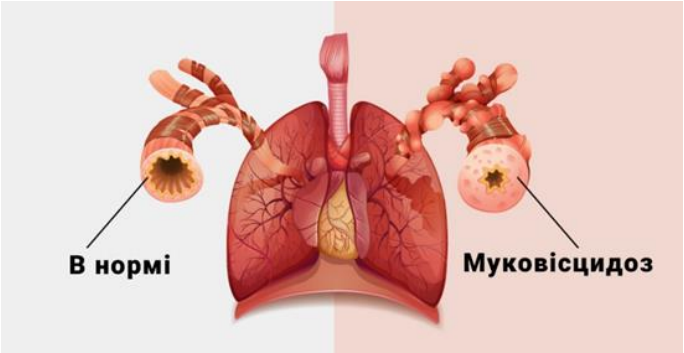

Рисунок 1.2 – Зміни в легенях при муковісцидозі

Фенілкетонурія – захворювання, яке виникає через мутацію в гені, що кодує фермент фенілаланінгідроксилазу. Цей фермент бере участь у розщепленні амінокислоти фенілаланіну. При порушенні його структури й функції фенілаланін накопичується в тканинах, особливо в мозку, де чинить токсичну дію (рисунок 1.3). У дітей це призводить до незворотних порушень розумового розвитку, якщо захворювання не виявити та не почати лікування ще в ранньому віці.

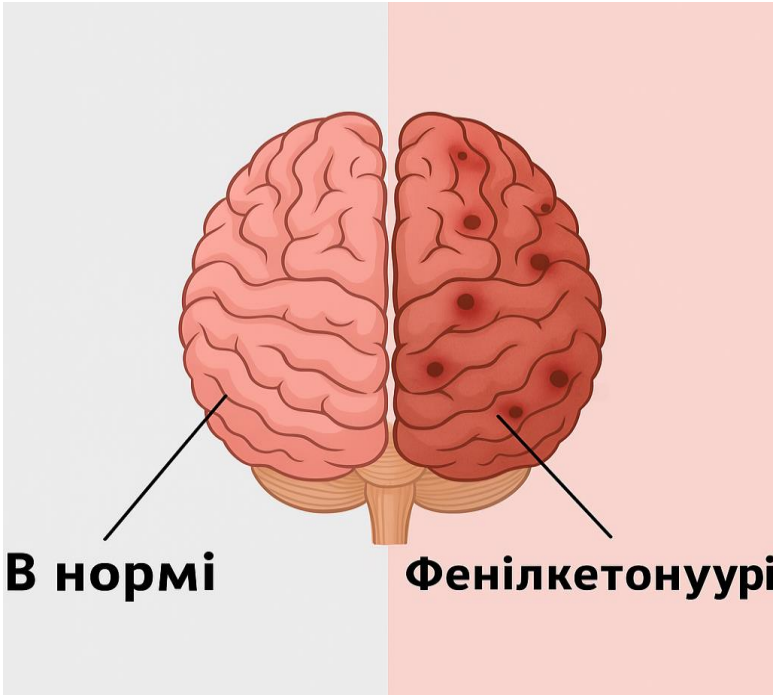

Рисунок 1.3 – Вплив фенілкетонурії на структуру головного мозку

Хвороба Тея-Сакса – рідкісне, але смертельно небезпечне генетичне захворювання, зумовлене мутацією в гені, який відповідає за фермент, необхідний для розщеплення певних жирів у нервовій тканині. Через дефіцит або нефункціональність цього ферменту в нейронах починають накопичуватися ліпіди (рисунок 1.4), що призводить до прогресуючої дегенерації нервової системи. Діти з цією хворобою зазвичай народжуються без явних відхилень, але вже в перші місяці життя з’являються ознаки ураження нервової системи, які швидко прогресують.

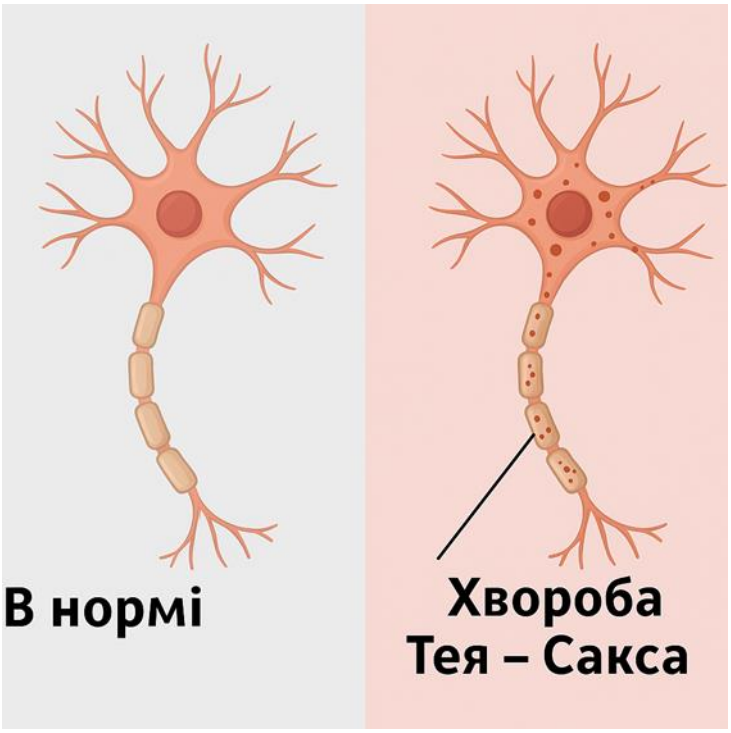

Рисунок 1.4 – Вплив хвороби Тея – Сакса на нервову клітину

У онкології також є приклади, коли мутації білків стають основою патології. Наприклад, мутації в гені TP53, що кодує білок-регулятор клітинного циклу p53, можуть призвести до того, що клітини втрачають контроль над поділом. У нормі p53 «вмикає» механізми зупинки клітинного циклу або апоптозу при пошкодженні ДНК. Коли цей білок не працює,

клітини можуть продовжувати ділитися, накопичуючи нові мутації, що сприяє розвитку злоякісних пухлин.

Мутації, що змінюють структуру білків, не лише порушують їхню безпосередню функцію, а й запускають цілий каскад патологічних процесів.

1.5 Обґрунтування актуальності дослідження

На сучасному етапі розвитку біомедичних наук надзвичайно актуальним стає вивчення зв’язку між генетичними змінами й їхнім впливом на молекулярні механізми функціонування клітини. Особливої уваги заслуговує дослідження мутацій у білках – ключових молекулах, які беруть участь у всіх біохімічних процесах в організмі. Від структурної цілісності білка залежить його стабільність, здатність до взаємодії з іншими молекулами, каталіз реакцій, передача сигналів, регуляція експресії генів тощо. Будь-яке порушення цієї структури може призвести до значних наслідків – від втрати функціональності до розвитку тяжких захворювань.

Щороку відкриваються нові захворювання, які мають генетичну природу, при цьому у більшості з них механізми патології пов’язані саме з порушеннями білкової структури. Наприклад, серповидноклітинна анемія, муковісцидоз, хорея Гентінгтона, деякі форми онкологічних і нейродегенеративних захворювань – усе це приклади патологій, в основі яких лежить структурна мутація білка. Проте кількість мутацій, що виявляються при генетичному секвенуванні, зростає швидше, ніж можливість їх експериментального дослідження.

Це створює нагальну потребу у швидкому, точному й доступному методі прогнозування їхнього впливу. Протягом десятиліть основним способом визначення структури білків були експериментальні методи: рентгеноструктурний аналіз, ЯМР-спектроскопія та кріоелектронна мікроскопія. Хоча вони забезпечують високу точність, їх застосування є ресурсомістким, дорогим і технічно складним. Крім того, не всі білки

піддаються кристалізації чи стабільному аналізу в лабораторних умовах. В умовах стрімкого розвитку біоінформатики з’являються альтернативи, які дозволяють подолати ці обмеження. Найбільш помітним проривом останніх років стало створення системи AlphaFold від DeepMind, яка продемонструвала здатність з високою точністю передбачати тривимірну структуру білка за його амінокислотною послідовністю. Це стало справжнім технологічним зрушенням у структурній біології, адже тепер дослідники можуть вивчати будову білків, не маючи фізичних структурних даних.

Особливо актуальним є застосування AlphaFold для аналізу мутацій, оскільки система дозволяє порівнювати структурні моделі дикого типу та мутантних форм білків, оцінювати зміни у згортанні, стабільності, просторовій орієнтації ключових амінокислот. Це відкриває нові можливості для виявлення патологічних змін на молекулярному рівні, моделювання взаємодій білків з лігандами, створення індивідуальних терапій на основі передбачених структурних змін. Актуальність теми підсилюється розвитком персоналізованої медицини, яка базується на точному знанні мутаційного профілю пацієнта. Щоб застосовувати такий підхід на практиці, необхідно не лише виявляти мутації, а й розуміти, як саме вони змінюють структуру та функції білків – і тут AlphaFold є незамінним інструментом.

2 АНАЛІЗ МОЖЛИВОСТЕЙ ALPHAFOLD ДЛЯ ДОСЛІДЖЕННЯ МУТАЦІЙ БІЛКІВ

2.1 AlphaFold: загальний огляд методики

AlphaFold – це система штучного інтелекту, розроблена компанією DeepMind. Вона призначена для прогнозування тривимірної структури білків на основі їх амінокислотної послідовності [1]. Цей метод став революційним у галузі структурної біології. Він дозволяє точно моделювати складну конфігурацію білків без проведення експериментальних досліджень. Це значно економить час і ресурси, які зазвичай витрачаються на рентгенівську кристалографію або кріоелектронну мікроскопію.

AlphaFold використовує сучасні методи машинного навчання. В основі лежить аналіз множинного вирівнювання послідовностей (MSA) та взаємодій між амінокислотами. На основі цих даних створюється просторове уявлення білка. Потім ця карта використовується для побудови фінальної 3D-моделі білкової структури.

2.1.1 Історія створення та розвиток AlphaFold

Розробка AlphaFold стала одним із найвпливовіших досягнень штучного інтелекту в галузі біології, що відкрило нову еру в дослідженні білкових структур. Протягом десятиліть науковці намагалися вирішити задачу згортання білка – передбачення тривимірної форми молекули лише на основі амінокислотної послідовності. Це завдання довгий час вважалося так званою «святою чашею» структурної біології через свою надзвичайну складність та фундаментальне значення для розуміння функцій білків у живих організмах.

Класичні методи структурного аналізу, зокрема рентгенівська кристалографія, ЯМР-спектроскопія та кріоелектронна мікроскопія,

залишаються надзвичайно точними. Водночас вони мають обмеження щодо масштабованості, витратності та доступності. Особливо проблематично досліджувати білки, які важко кристалізуються або нестабільні в розчині. Усвідомлюючи ці обмеження, дослідницька команда DeepMind ініціювала у 2016 році проект зі створення штучного інтелекту, здатного прогнозувати білкові структури обчислювальними методами.

Першим результатом цієї роботи став AlphaFold1, який дебютував у 2018 році на конкурсі CASP13 – престижному міжнародному змаганні з прогнозування білкових структур [13]. Модель продемонструвала конкурентну точність, використовуючи згорткові нейронні мережі для передбачення просторових взаємозв'язків між амінокислотами. Незважаючи на певні обмеження, AlphaFold1 вперше довів, що моделі глибокого навчання можуть наближатися за точністю до експериментальних методів.

Справжній прорив відбувся з випуском AlphaFold2 у 2020 році. Завдяки радикально новій архітектурі, що включала модулі Evoformer і структурний блок, модель досягла середнього балу GDT понад 90 для більшості цілей CASP14. Це означало, що її передбачення майже не відрізнялися від структур, визначених у лабораторії. На відміну від традиційних моделей гомології, AlphaFold2 опирається не на наявність схожих шаблонів, а на вивчення глибоких еволюційних залежностей та просторових обмежень між залишками амінокислот.

У 2021 році, у співпраці з EMBL-EBI, було оприлюднено базу даних AlphaFold DB, яка містить понад 200 мільйонів структур білків. Це охоплює майже весь відомий білковий простір, включаючи білки з таких організмів, як людина, бактерії, рослини, археї, а також багато гіпотетичних білків. Цей крок став прикладом відкритої науки у дії – тепер біологи, хіміки та фармакологи у всьому світі можуть отримати доступ до точних 3D-моделей практично будь-якого білка без необхідності в експериментальній валідації.

На хвилі успіху DeepMind представила в 2024 році AlphaFold3, який вийшов за межі аналізу окремих білків. Нова архітектура Pairformer та інтеграція модуля дифузійного уточнення дозволили моделювати білкові комплекси, а також взаємодії білків із ДНК, РНК, лігандами та іонами. AlphaFold3 не тільки підвищив точність, а й розширив практичну застосовність моделі в біомедицині, зокрема у розробці ліків, аналізі мутацій, ферментній інженерії та вивченні вірусних білків.

Попри обмежений доступ до вихідного коду AlphaFold3 через питання біоетики та ризиків неконтрольованого використання, компанія надала API-доступ до моделі для верифікованих академічних і комерційних проєктів. Цей крок дозволяє дослідникам інтегрувати інструмент у власні обчислювальні конвеєри, забезпечуючи збереження етичних стандартів.

Значення AlphaFold уже визнано науковою спільнотою. Керівники проєкту Деміс Хассабіс та Джон Джампер отримали низку престижних нагород, а сам AlphaFold часто згадується серед проривів, рівнозначних відкриттю CRISPR або секвенуванню геному людини. Цей інструмент не лише змінив погляд на проблему згортання білків, але й задав нову парадигму в науці, де штучний інтелект виступає не як допоміжний інструмент, а як повноцінний учасник наукового відкриття.

2.1.2 Принципи роботи та основні компоненти алгоритму

В основі роботи AlphaFold лежить глибока ідея: просторову структуру білка шукати лише його амінокислотну послідовність, і ця інформація зберігається в еволюційних слідах, які можна виявити за допомогою обчислювальних методів. Завдання згортання білка – це, по суті, пошук стабільної просторової конфігурації, яка мінімізує енергію системи, водночас зберігаючи функціональну здатність молекули.

AlphaFold реалізує цей підхід через глибоку нейронну структуру, що складається з декількох наступних етапів, кожен з яких відповідає обробці

спеціального типу інформації. Все починається з побудови множинного вирівнювання компонентів (MSA), що дозволяє виявити коеволюційні сигнали – закономірності, при яких певні пари амінокислот змінюються синхронно в межах еволюційно споріднених організмів. Саме ці зв’язки вказують на просторову близькість решток у третинній структурі білка.

Загальний процес представлено на рисунку 2.1, де показано основні етапи перетворення амінокислотної послідовності на повноцінну тривимірну модель.

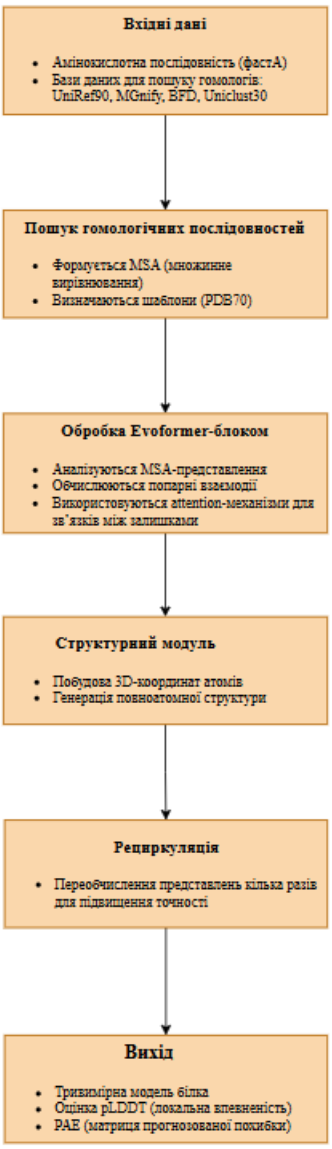

Рисунок 2.1 – Архітектура AlphaFold2: основні етапи побудови тривимірної структури білка

На наступному етапі вирівняні компенсації перетворюються у високорозмірні вектори – представлення (уявлення), які служать вхідними даними для глибинної нейромережі. Тут вступає в дію ключовий інноваційний блок моделі – Evoformer [1]. Це трансформерна структура, адаптована для біологічних завдань, що паралельно працює з двома типами інформації: вирівнюванням MSA і попарними ознаками залишків. Замість класичної уваги Evoformer використовує спеціальні механізми, такі як увага по рядках і стовпцях, зовнішні добавки та комунікацію між двома потоками даних, що дозволяє моделі розуміти як локальні, так і глобальні просторові закономірності.

Ці представлення прогресу уточнюються та збагачуються відомостями, що моделі дозволяють зробити висновки про тривалий взаємозв’язок між амінокислотами. Наприклад, два залишки можуть бути віддалені в комп’ютері, але при цьому створити контакт у 3D-просторі – Evoformer здатний мати такі взаємозв’язки.

Після обробки в Evoformer, оновлені ознаки потрапляють у структурний модуль, який остаточно генерує 3D-модель білка. На цьому етапі використовується фреймова система координат – це дозволяє точніше моделювати просторове положення кожного залишку з урахуванням його орієнтації. Замість того, щоб створити фізичну симуляцію (як у класичних методах), AlphaFold навчається будувати структуру в наскрізному режимі, одночасно з вхідною позицією до координатних атомів, що забезпечує надзвичайну швидкість та узгодженість результатів.

Ще однією критично важливою особливістю є механізм рециркуляції (recycling). Модель кількох разів «перепускає» представлення через Evoformer і структурний модуль, кожен раз уточнюючи прогноз. На практиці це дозволяє значно підвищити точність без збільшення обчислювальних витрат учній прогресії. У типових конфігураціях використовують три етапи рециркуляції, що забезпечують оптимальний баланс між продуктивністю та витратами.

Для оцінки надійності прогнозу AlphaFold вводить індекс pLDDT (predicted Local Distance Difference Test) – числовий показник від 0 до 100 для всіх залишків, що вказує на надійність моделі в локальному передбаченні. Високі значення (більше 90) свідчать про надійні передбачення, тоді як зниження pLDDT в окремих ділянках може сигналізувати про гнучки або структурно неоднозначні області.

Додаткова модель надає PAE-матриці (Predicted Alignment Error), які можуть оцінити, наскільки точно передбачено відносне розташування будь-яких двох залишків. Це особливо корисно дослідженні білків із ключовими доменами, де гнучкі з’єднання можуть привести до варіативності в конформації.

Таким чином, інженерна досконалість AlphaFold створюється в поєднанні кількох джерел біологічної інформації (MSA, структурні шаблони, геометричні обмеження) із сучасними методами глибокого навчання, що дозволяє досягти точності, співставної з експериментальними підходами. Модель не лише прогнозує структуру, а й надає інструменти для оцінки її надійності, що робить AlphaFold потужною платформою для прикладних досліджень у біомедицині, біотехнологіях та фундаментальній біології.

2.1.3 Джерела даних, що використовуються для тренування AlphaFold

Точність моделі AlphaFold значною мірою залежить від даних, на яких вона навчалась. Щоб навчитись прогнозувати структуру білків, модель використовує два основні типи джерел: послідовності білків та експериментальні 3D-структури.

Перший тип даних – це великомасштабні білкові бази. Вони потрібні для побудови множинного вирівнювання послідовностей (MSA). MSA допомагає виявити еволюційні зв’язки між амінокислотами. Саме ці зв’язки

дозволяють AlphaFold робити припущення про просторову близькість залишків.

Для побудови MSA AlphaFold використовує такі джерела:

– UniRef90 – база білків, згрупованих за 90% ідентичністю. Це зменшує надмірність і покращує продуктивність моделі;

– MGnify – колекція білкових послідовностей з метагеномних досліджень. Дає змогу охопити маловивчені організми;

– BFD (Big Fantastic Database) – одна з найбільших баз, яка містить понад 2 мільярди послідовностей. Забезпечує глибоке еволюційне покриття;

– Uniclust30 – групує білки з 30% схожістю. Допомагає виявляти далекі гомологи.

Ці бази дають моделі змогу працювати навіть з рідкісними або слабко охарактеризованими білками.

Другий тип даних – це структури білків, отримані експериментально. Вони слугують еталонами для навчання з учителем. Найважливішим джерелом є Protein Data Bank (PDB) [4]. PDB містить координати атомів у білках. Дані надходять із трьох основних методів: рентгенівської кристалографії, ЯМР-спектроскопії та кріоелектронної мікроскопії. Ці структури потрібні, щоб модель навчилась пов’язувати послідовність з її просторовою формою. Команда DeepMind створила окрему підмножину PDB. Вона не містить білків, що були частиною змагання CASP. Це виключає ризик витоку тестових структур у навчальний набір. Окрім MSA та експериментальних структур, AlphaFold використовує також структурні шаблони. Це наявні 3D-моделі, які частково схожі на цільовий білок. Їх шукають за допомогою інструмента HHsearch у базі PDB70. Шаблони дозволяють:

– уточнити локальні ділянки структури;

– покращити передбачення в областях з недостатніми еволюційними сигналами.

Водночас модель не залежить від шаблонів. Вони є лише додатковим джерелом, а не основою прогнозу.

Щоб ефективніше використовувати наявні приклади, модель застосовує рециркуляцію. Це техніка, яка дозволяє кілька разів обробляти ті самі представлення. Таким чином, AlphaFold уточнює свої прогнози без потреби в нових даних.

2.1.4 Переваги та обмеження методики AlphaFold

AlphaFold запровадив зміну парадигми в структурній біології, пропонуючи значні переваги порівняно з традиційними експериментальними та обчислювальними методами прогнозування структури білків, водночас демонструючи певні обмеження, які вимагають критичного розгляду (таблиця 2.1) [15].

Таблиця 2.1 – Порівняльна характеристика можливостей AlphaFold

Перевага	Обмеження	Коментар
Висока точність (GDT > 90), рівень атомарної деталізації	Не моделює динаміку або альтернативні конформації	AlphaFold передбачає лише одну структуру – це обмеження при вивченні білків, що змінюють форму під час функціонування.
Швидкість: передбачення структури за хвилини	Проблеми з гнучкими/невпорядкованими білками (IDR)	Для білків без чітко визначеної структури або з флексибельними доменами точність значно падає.

Продовження таблиці 2.1

Доступ до понад 200 млн структур у відкритій базі AlphaFold DB	Обмеження у передбаченні білкових комплексів, лігандів, РНК/ДНК	AlphaFold3 уже робить кроки до цього, але функціонал ще обмежений і недоступний як повністю відкритий інструмент.
Не потребує лабораторного обладнання чи біофізичних експериментів	Не враховує посттрансляційні модифікації, клітинний контекст	Наприклад, фосфорилювання чи локалізація у клітині можуть впливати на структуру, але не моделюються.
Надійні оцінки впевненості (pLDDT, PAE)	Залежність від кількості гомологів у MSA	Якщо послідовність білка рідкісна або синтетична – точність значно знижується.
Може працювати з білками з малою кількістю гомологів (через BFD, MGnify)	Структура моделі погано інтерпретується – «чорний ящик»	Хоча результати точні, важко зрозуміти, чому саме така структура була обрана – складно для гіпотезо-генерування.
Широке застосування в науці та індустрії (фармацевтика, біоінженерія, агробіо тощо)	Для великих білків або комплексів потрібні великі обчислювальні ресурси (GPU, оперативна пам’ять)	Масштабування на цілі протеоми або складні системи вимагає інфраструктури, недоступної для деяких лабораторій.

2.2 Порівняння AlphaFold з іншими методами структурного аналізу

До появи AlphaFold вивчення тривимірної структури білків здійснювалось переважно експериментальними методами. Серед них – рентгеноструктурний аналіз, ядерний магнітний резонанс (ЯМР) і кріоелектронна мікроскопія (кріо-ЕМ). Ці підходи дають точні результати, але потребують значних ресурсів і часу. AlphaFold, у свою чергу, запропонував нову модель – швидку, обчислювальну й масштабовану.

Рентгеноструктурний аналіз – це найпоширеніший метод. Він дозволяє отримати структуру з високою роздільністю – аж до окремих атомів. Однак метод вимагає кристалізації білка. А деякі білки кристалізувати неможливо або дуже складно. Процес може тривати тижні або місяці. Також умови кристалізації іноді змінюють природну конформацію білка.

ЯМР-спектроскопія – цей метод дозволяє вивчати білки у водному середовищі, наближеному до фізіологічного. Він добре підходить для дослідження гнучких структур або динамічних змін. Проте ЯМР обмежений за розміром молекул. Його не можна застосовувати до великих білкових комплексів. Крім того, обробка спектрів вимагає багато часу та досвіду.

Кріоелектронна мікроскопія (кріо-ЕМ) – цей метод став проривом для дослідження великих білкових комплексів. Зразки швидко заморожуються, що зберігає їх природну форму. Кристалізація не потрібна. Однак обладнання для кріо-ЕМ надзвичайно дороге. Обробка зображень потребує спеціального програмного забезпечення та потужних обчислювальних ресурсів.

AlphaFold працює зовсім інакше. Він не вимагає зразків або експерименту. Достатньо лише амінокислотної послідовності. Структура генерується за години, а не тижні. Модель здатна передбачати навіть ті білки, для яких не існує експериментальних структур.

AlphaFold відкрив можливості для масштабного аналізу цілих протеомів. Його застосовують у функціональній геноміці, біоінформатиці, фармакології та біотехнологіях. Інструмент особливо корисний, коли потрібно швидко оцінити структуру нових або погано охарактеризованих білків. Порівняльна таблиця 2.2.

Таблиця 2.2 Порівняння методів визначення просторової структури білків

Критерій	Рентгеноструктурний аналіз	ЯМР-спектроскопія	Кріо-ЕМ
Тип	Експериментальний	Експериментальний	Експериментальний
Потреба у зразку	Так (кристал)	Так (розчин)	Так (заморожений зразок)
Переваги	Висока точність, усталеність	Фізіологічні умови, динаміка	Великі комплекси, без кристалізації
Обмеження	Кристалізація, тривалість	Обмежена маса білка, складність обробки	Дороговартісність, складність
Доступність	Низька	Низька	Дуже низька

AlphaFold не витісняє класичні методи, а доповнює їх. У багатьох випадках він забезпечує достатню точність для біологічних висновків. Однак там, де важлива динаміка або контекст середовища, експеримент залишається незамінним. Можливість працювати без зразка робить AlphaFold надзвичайно цінним. Він відкриває доступ до структур для дослідників у будь-якій точці світу – навіть без лабораторії. Завдяки цьому AlphaFold вже сьогодні трансформує біологічні дослідження і створює нові стандарти в науці.

2.3 Використання AlphaFold для аналізу наслідків мутацій

Мутації в генах можуть змінювати амінокислотну послідовність білка. Це впливає на його структуру, стабільність та функцію. Іноді такі зміни є нейтральними. Але часто вони порушують згортання білка, змінюють активні сайти або знижують ефективність зв'язування. Традиційно для вивчення структурних наслідків мутацій застосовували експериментальні методи. Наприклад, рентгенівську кристалографію або ЯМР [2]. Це точні, але повільні та ресурсозатратні підходи. Крім того, моделювання гомології обмежене наявністю близьких шаблонів. З появою AlphaFold з'явилась можливість аналізувати мутації в обхід експерименту. Модель дозволяє отримати 3D-структуру білка лише на основі зміненої послідовності. Це робить аналіз швидким і доступним навіть для білків, які ще не вивчались у лабораторіях.

Щоб проаналізувати одиничну мутацію, достатньо змінити одну амінокислоту у вихідній послідовності. Потім модель генерує нову структуру, яку можна порівняти з формою білка дикого типу. Для оцінки використовують:

– RMSD – середньоквадратичне відхилення координат атомів;
– TM-score – для оцінки загальної схожості;
– візуальне порівняння – у PyMOL або ChimeraX.

Зміни можуть бути незначними. Але навіть малий зсув бічного ланцюга може порушити активний центр або білок-білкову взаємодію. Особливо корисною є метрика pLDDT, яку генерує AlphaFold. Якщо мутація знижує значення pLDDT в певному регіоні – це сигнал, що локальна структура стала менш впевнено передбачуваною. Це може означати дестабілізацію.

Аналіз кількох мутацій складніший. Їх ефект не завжди адитивний. Декілька «нейтральних» замін разом можуть суттєво змінити просторову структуру. Такі випадки особливо важливі при вивченні онкогенів або

білків вірусів. AlphaFold дає можливість будувати структуру з урахуванням усіх мутацій одночасно. Але інтерпретувати ці зміни складно. У таких випадках часто використовують додаткові методи, як-от молекулярна динаміка. Вона дозволяє дослідити нестабільність або гнучкість зміненого білка. Можливість аналізувати мутації без експериментів є корисною в багатьох сферах:

– у клінічній геноміці – для оцінки потенційної патогенності мутацій;

– у білковій інженерії – для перевірки, чи не порушує мутація функцію;

– в еволюційній біології – для аналізу переносимості змін у білках різних видів.

Одним із прикладів такого застосування є аналіз шипоподібного білка вірусу SARS-CoV-2. За допомогою AlphaFold моделювали вплив мутацій варіантів Delta та Omicron на взаємодію з рецептором ACE2. Це допомогло спрогнозувати зміни у властивостях вірусу ще до появи експериментальних даних.

AlphaFold не враховує всі фактори. Наприклад, модель:

– не моделює посттрансляційні модифікації;

– не враховує вплив середовища (розчинник, клітинний контекст);

– не передбачає зміни енергії згортання (ΔΔG);

– не визначає прямо, чи є мутація шкідливою чи нейтральною.

Тому її прогнози часто доповнюють іншими інструментами: FoldX, Rosetta, DynaMut. Вони дозволяють обчислити зміну енергії, стабільність, гнучкість чи інтерфейсну спорідненість.

2.3.1 Особливості застосування AlphaFold для аналізу одиничних і множинних мутацій

Хоча AlphaFold не створювався спеціально для аналізу мутацій, він добре підходить для цієї задачі. Модель здатна згенерувати нову 3D-

структуру білка, якщо змінити навіть одну амінокислоту у послідовності. Це дозволяє швидко оцінити вплив заміни на просторову конфігурацію.

Аналіз одиничної мутації починається зі зміни однієї амінокислоти в послідовності білка. Далі AlphaFold будує нову модель. Її порівнюють зі структурою варіанту дикого типу. Різницю оцінюють за допомогою:

– RMSD – для локальних відхилень;

– TM-score – для глобальної подібності;

– візуального перегляду – у молекулярних переглядачах (наприклад, PyMOL або ChimeraX).

Найчастіше відмінності незначні. Проте зміни можуть зачіпати важливі ділянки – активні сайти, канали або інтерфейси взаємодії. Навіть невеликий зсув бічного ланцюга іноді спричиняє втрату функції. Окрему увагу варто звернути на показник pLDDT. Це локальна оцінка впевненості моделі. Якщо значення pLDDT знижується поблизу мутації, це може вказувати на порушення згортання або нестабільність у тій ділянці. Коли замін кілька, прогноз ускладнюється. Мутації можуть взаємодіяти між собою. Вони не завжди діють незалежно. У деяких випадках – посилюють ефект, в інших – компенсують одна одну. AlphaFold здатен будувати структуру білка з декількома мутаціями одночасно. Але інтерпретувати результат потрібно обережно. Особливо якщо йдеться про:

– значні перебудови в ядровій упаковці;

– зміни вторинної структури;

– розриви або утворення нових водневих зв’язків.

Для складних випадків дослідники часто застосовують додаткові інструменти. Наприклад, FoldX, Rosetta або молекулярну динаміку. Це дає можливість уточнити, як зміни впливають на енергію системи, стабільність і гнучкість. Практичні приклади:

– дослідженнях раку часто аналізують білки, що накопичують багато соматичних мутацій. Наприклад, p53 або EGFR. AlphaFold дозволяє

змоделювати повну структуру мутантного варіанту та дослідити, як змінюється його просторове розташування;

– у вірусології модель застосовували для вивчення білка шипа SARS-CoV-2. Аналіз мутантів Delta й Omicron допоміг виявити, як зміни у послідовності впливають на взаємодію з рецептором ACE2 та нейтралізацію антитілами.

AlphaFold дає змогу вивчати як поодинокі, так і множинні мутації без експериментальних структур. Це значно прискорює первинний аналіз. Однак складні випадки вимагають комбінованого підходу. Модель забезпечує точну геометрію, але не оцінює зміну енергії чи функціональність напряму. Тому її прогнози варто доповнювати іншими методами.

2.3.2 Алгоритми оцінки впливу мутацій на стабільність та функції білків

AlphaFold передбачає структурні зміни після мутацій, але не оцінює їхні функціональні чи енергетичні наслідки. Для цього використовуються інші інструменти, які допомагають відповісти на ключові запитання: чи зберігається стабільність білка? чи не порушено активний центр? чи залишається молекула функціональною?

Одним із головних показників стабільності є зміна вільної енергії згортання (ΔΔG). Позитивне значення вказує на дестабілізацію, негативне – на стабілізуючий ефект. Найчастіше застосовують FoldX, який базується на емпіричних енергетичних моделях, і Rosetta ddG, що моделює просторові зміни через релаксацію структури [9], [12]. FoldX забезпечує швидкість, Rosetta – точність. Часто їх комбінують.

У задачах з великою кількістю мутацій ефективними є ML-предиктори:

– mCSM – ґрунтується на графових сигнатурах;

– DUET – об’єднує кілька моделей в один прогноз;

– DeepDDG – використовує глибокі нейронні мережі.

Важливо розрізняти стабільність і функцію. Деякі мутації не змінюють структуру, але порушують біологічну активність. Для оцінки функціонального ефекту застосовують:

– SIFT – аналізує еволюційну збереженість [10];

– PolyPhen-2 – оцінює ймовірну шкоду заміни з урахуванням структури [3];

– MutPred, DynaMut, GraphSite – поєднують структурні, динамічні та послідовнісні характеристики.

Якщо структура отримана з AlphaFold, точність таких інструментів помітно зростає.

Для аналізу мутацій в інтерфейсах білків застосовують mCSM-PPI2, BeAtMuSiC, SAAMBE. У випадках ферментів – CUPSAT, CSA, EnzymeMiner, які виявляють зміни в активних центрах або зв’язувальних кишенях.

Жоден метод не охоплює всі аспекти. Тому найкращий підхід – комбінований:

– AlphaFold – для структури;

– FoldX/Rosetta – для стабільності;

– mCSM, DynaMut – для динаміки;

– SIFT/PolyPhen – для еволюційної значущості.

Ансамблеві моделі допомагають підвищити надійність прогнозу, особливо при масовому скринінгу варіантів.

AlphaFold – це лише стартова точка. Повний аналіз мутації вимагає багаторівневого підходу, який охоплює стабільність, функцію, взаємодії та енергетику. Така стратегія особливо цінна в сучасній медицині, біоінженерії та дизайні білків.

2.3.3 Інтерпретація прогнозів AlphaFold в контексті функціональних змін білків

Прогнози AlphaFold відкривають широкі можливості для аналізу того, як мутації впливають на функцію білка. Хоча ця модель не була створена для прямого визначення функціональних змін, вона формує надійну просторову основу, на якій можна будувати обґрунтовані припущення щодо біологічної активності, стабільності та взаємодій.

Першим кроком є зіставлення амінокислотної послідовності з відомими функціональними анотаціями. Часто критично важливі залишки – це каталітичні ділянки ферментів, регіони зв’язування лігандів, контактні поверхні для білок-білкових взаємодій або ділянки алостеричного регулювання. Для такого аналізу використовуються бази UniProt, InterPro, Pfam, а також спеціалізовані ресурси, як-от Catalytic Site Atlas.

Визначені функціональні зони накладаються на передбачену 3D-структуру. Якщо мутація трапляється у межах таких ділянок або безпосередньо поблизу них, це може свідчити про потенційне порушення активності. Просторове вирівнювання мутантної та дикої форм білка виконується за допомогою таких інструментів, як TM-align, ChimeraX або PyMOL. Візуальне порівняння дозволяє виявити локальні конформаційні зміни – зсуви бічних ланцюгів, втрату водневих зв’язків або зміни в упакуванні ядра.

Важливим джерелом інформації є показник pLDDT – локальна оцінка впевненості AlphaFold у кожному залишку. Якщо в регіоні мутації спостерігається різке зниження pLDDT порівняно з дикою формою, це може свідчити про втрату стабільності або некоректне згортання. Разом із цим використовують і додаткові параметри – наприклад, площу поверхні, доступну для розчинника (SASA), яка допомагає оцінити, чи змінився характер середовища навколо амінокислоти.

Щоб краще інтерпретувати функціональні наслідки, дослідники користуються додатковими інструментами. MutFunc дає змогу поєднати структурні зміни з анотаціями важливих залишків. DynaMut та STRUM аналізують, як конкретна заміна впливає на стабільність або гнучкість молекули. GraphSite дозволяє враховувати топологічну роль залишку в білковій структурі, що особливо важливо для виявлення залишків із регуляторною або сигнальною функцією.

У випадках, коли білок бере участь у міжмолекулярних зв'язках – наприклад, при утворенні комплексів або зв'язуванні з рецепторами, – доцільно аналізувати геометрію інтерфейсу. Тут корисні інструменти типу mCSM-PPI2, BeAtMuSiC або SAAMBE, які прогнозують зміни у спорідненості зв'язування та оцінюють, чи порушена критична взаємодія.

Коли йдеться про ферментативну активність, особливу увагу звертають на каталітичні залишки та кишені зв'язування кофакторів. Їх можна дослідити за допомогою інструментів CUPSAT, EnzymeMiner або Catalytic Site Atlas. Якщо мутація впливає на геометрію активного сайту або знижує його доступність, це може призвести до зниження каталізу або повної втрати функції.

Функціональні гіпотези доцільно порівнювати з даними з клінічних або геномних баз, зокрема ClinVar, gnomAD або HGMD. Таке зіставлення дозволяє перевірити, чи спостерігалася конкретна мутація у пацієнтів, та чи асоційована вона з певним фенотипом або захворюванням. Це особливо корисно у генетичній діагностиці та при пріоритезації варіантів для подальших лабораторних досліджень.

Варто пам'ятати, що AlphaFold генерує лише статичну модель, яка не враховує вплив клітинного середовища, посттрансляційні модифікації або динамічні процеси згортання. Тому функціональні висновки завжди мають супроводжуватися обережною інтерпретацією, бажано з опорою на кілька незалежних джерел.

Прогнози AlphaFold створюють якісну основу для функціонального аналізу білків, особливо у випадку мутацій. Поєднання структурних спостережень із анотаціями, інструментами для оцінки стабільності, динаміки та взаємодій дозволяє формувати обґрунтовані гіпотези про можливий вплив змін. Хоч модель не замінює експеримент, вона значно підвищує ефективність аналізу та сприяє кращому розумінню механізмів функціональних порушень.

3 ЕКСПЕРИМЕНТАЛЬНА ЧАСТИНА ДОСЛІДЖЕННЯ (ПРАКТИЧНА РЕАЛІЗАЦІЯ)

3.1 Постановка задачі аналізу структур та вибір білків для дослідження

У межах цієї роботи основна увага зосереджена на дослідженні можливостей сучасних моделей штучного інтелекту, зокрема AlphaFold2, у задачах передбачення структурних наслідків мутацій у білках. На відміну від традиційних біофізичних або структурно-біологічних підходів, AlphaFold реалізує повноцінну нейромережеву архітектуру, яка дозволяє прогнозувати тривимірну конформацію білка виключно на основі амінокислотної послідовності.

Метою експериментальної частини є практичне застосування AlphaFold2 для аналізу структурних змін, зумовлених мутаціями, а також інтерпретація результатів на основі ключових вихідних метрик моделі – таких як pLDDT (confidence score), PAE (матриця прогнозованих помилок), RMSD, TM-score та ін. Увага приділяється тому, як модель реагує на варіації в послідовності, як змінюється локальна впевненість у передбаченні структури та які закономірності виявляються у відповідях системи.

Завдання, поставлені в межах цього дослідження:

– продемонструвати застосування архітектури AlphaFold2 як прикладу глибокої моделі трансформерного типу;

– дослідити, як окремі компоненти (MSA, Evoformer, структурний модуль, рециркуляція) впливають на результат при моделюванні мутацій;

– оцінити зміни у структурних передбаченнях після внесення точкових змін у амінокислотну послідовність;

– інтерпретувати метрики впевненості (pLDDT, PAE) у контексті локальних збурень структури;

– порівняти отримані структури з відомими експериментальними моделями або контрольними зразками.

Для виконання експерименту було відібрано набір білків, які відповідають таким критеріям:

– наявність якісно визначених 3D-структур у базі PDB – для порівняння з результатами моделі;

– наявність задокументованих мутацій (ClinVar, UniProt), які мають клінічне або структурне значення;

– помірна довжина (до 500 амінокислот) – для оптимізації обчислень при локальному запуску моделі;

– функціональні ділянки білків добре описані (каталітичні сайти, кишені зв’язування).

Обрано чотири білки, які дозволяють протестувати AlphaFold у різних сценаріях:

– TP53 – класичний білок-супресор пухлин, що містить критичні мутації, асоційовані з онкозахворюваннями;

– CFTR – трансмембранний білок, мутації в якому зумовлюють муковісцидоз;

– Spike-білок SARS-CoV-2 (RBD) – дозволяє перевірити реакцію моделі на варіанти вірусного білка;

– лізоцим (LYZ) – стабільна модель, що використовується як контроль для аналізу точності та pLDDT у відомих структурах.

Дослідження спрямоване на глибше розуміння роботи самої AI-системи AlphaFold, як прикладу успішного застосування трансформерних мереж у сфері біоінформатики.

3.3 Опис набору даних: вихідні структури, вибір мутацій для аналізу

Для проведення дослідження було зібрано набір даних, що включає тривимірні структури білків та відповідні варіанти мутацій. Основним

джерелом структур слугувала AlphaFold Protein Structure Database, яка забезпечує високоточні передбачення просторової організації білкових молекул [20]. Структури доступні у форматі PDB, що дозволяє зручно інтегрувати їх у програмне середовище для аналізу. Крім координат атомів, база містить показники достовірності для кожного амінокислотного залишку (pLDDT-оцінки), що є критично важливими для оцінки надійності передбаченої моделі.

Було обрано кілька білків, що мають важливе біологічне значення. Критерії відбору включали:

– наявність підтвердженої функціональної ролі в клітинних процесах (наприклад, передача сигналу, реплікація ДНК, регуляція апоптозу);

– участь у розвитку хвороб (наприклад, онкологічних або генетичних);

– наявність відомих точкових мутацій з експериментально встановленим або передбачуваним ефектом.

Інформацію про мутації було зібрано з відкритих біомедичних баз даних, таких як ClinVar, UniProt, COSMIC (Catalogue of Somatic Mutations in Cancer), а також з наукових статей. Особлива увага приділялася точковим мутаціям (missense mutations), оскільки вони найчастіше впливають на локальну структуру білка, змінюючи лише одну амінокислоту без порушення загальної довжини ланцюга [16].

Формат запису мутації наведено у загальноприйнятому вигляді: амінокислота дикого типу – позиція – замінена амінокислота (наприклад, R132H для аргініну, заміненого на гістидин у позиції 132). Для кожної обраної мутації створено окрему структурну модель, в якій ця заміна була врахована за допомогою відповідних інструментів моделювання (наприклад, PyMOL, Modeller або AlphaFold-Multimer з ручною корекцією послідовності).

Таким чином, для кожного білка сформовано дві просторові моделі:

– первинна (wild-type) – відповідає природній структурі білка без мутацій;

– мутантна (mutant) – містить задану точкову заміну.

Це дає можливість виконати порівняльний аналіз на рівні атомної структури та визначити локальні й глобальні відмінності, які можуть виникати внаслідок навіть мінімальних змін послідовності. Крім того, для деяких структур було розраховано додаткові показники, такі як енергетична стабільність, площі взаємодії поверхонь, вплив на вторинні структури тощо.

Набір даних є репрезентативним і охоплює як патогенні мутації, пов’язані із захворюваннями, так і нейтральні варіанти, які не викликають суттєвих змін. Що дає змогу проаналізувати не лише наслідки мутацій, але й з'ясувати, за яких умов білкова структура залишається стабільною, а за яких зазнає порушень.

3.4 Архітектура та ключові компоненти моделі AlphaFold

Для досягнення високої точності передбачення білкових структур AlphaFold використовує складну багаторівневу архітектуру. Її основа – набір послідовних обчислювальних блоків, кожен з яких виконує конкретну функцію: від обробки вхідної інформації до побудови просторової моделі білка. У цьому розділі розглянуто принципи роботи основних компонентів AlphaFold, включно з механізмом рециклування, модулями уваги, геометричними перетвореннями та головами передбачення.

3.4.1 Головний клас AlphaFold: механізм рециклування

У центрі моделі AlphaFold знаходиться головний клас, що координує повний процес передбачення просторової структури білка. Його основною особливістю є реалізація механізму рециклування – ітеративного покращення результатів на основі попередніх передбачень.

Рециклування дозволяє моделі поступово уточнювати свої передбачення. Після кожної ітерації результати передаються у наступний крок як вхідні дані. Таким чином модель формує контекст і «вчиться на власних прогнозах». Це особливо корисно для складних білків із довгими, взаємозалежними ділянками. На початку створюється порожній стан для рециклування (лістинг 3.1).

Лістинг 3.1 – Ініціалізація нульового стану для механізму рециклування в моделі AlphaFold

```
prev = {}
emb_config = self.config.embeddings_and_evoformer
if emb_config.recycle_pos:
    prev['prev_pos'] = jnp.zeros(
        [num_residues, residue_constants.atom_type_num,
3])
if emb_config.recycle_features:
    prev['prev_msa_first_row'] = jnp.zeros(
        [num_residues, emb_config.msa_channel])
    prev['prev_pair'] = jnp.zeros(
        [num_residues, num_residues,
emb_config.pair_channel])
```

Ці тензори зберігають координати, вирівнювання та парні представлення з попередньої ітерації. AlphaFold підтримує два варіанти керування кількістю рециклів (лістинг 3.2).

Лістинг 3.2 – Приклад визначення кількості ітерацій рециклування у моделі AlphaFold

```
if self.config.num_recycle:
    if 'num_iter_recycling' in batch:
        num_iter = batch['num_iter_recycling'][0]

        # Add insurance that we will not run more
```

Продовження лістингу 3.2

```
        # recyclings than the model is configured to run.
        num_iter = jnp.minimum(num_iter,
self.config.num_recycle)
    else:
        num_iter = self.config.num_recycle
```

Під час навчання кількість ітерацій може передаватися через batch['num_iter_recycling']. Це дозволяє динамічно змінювати глибину рециклування для кожного прикладу [7]. Якщо цього параметра немає, або модель працює у режимі виведення, використовується фіксоване значення self.config.num_recycle.

Процес рециклування реалізовано як цикл з використанням hk.while_loop. На кожному кроці виконується оновлення стану через виклик do_call(...). Якщо виконується лише ініціалізація, модель проходить одну ітерацію (лістинг 3.3).

Лістинг 3.3 – Цикл рециклування з використанням hk.while_loop у моделі AlphaFold

```
body = lambda x: (x[0] + 1,  # pylint: disable=g-long-lambda
    get_prev(do_call(x[1], recycle_idx=x[0],
                     compute_loss=False)))
if hk.running_init():
    _, prev = body((0, prev))
else:
    _, prev = hk.while_loop(
        lambda x: x[0] < num_iter,
        body,
        (0, prev))
```

Після завершення всіх ітерацій виконується фінальний виклик передбачення через do_call, у який передається оновлений стан prev та

індекс останньої ітерації recycle_idx. У результаті повертається основний вихід моделі, а також додаткові дані, якщо активовано обчислення втрат (лістинг 3.4).

Лістинг 3.4 – Завершальний виклик do_call після рециклування та обробка результату в моделі AlphaFold

```
ret = do_call(prev=prev, recycle_idx=num_iter)
if compute_loss:
    ret = ret[0], [ret[1]]

if not return_representations:
    del (ret[0] if compute_loss else ret)['representations']  # pytype: disable=unsupported-operands
return ret
```

У випадку, якщо return_representations дорівнює False, з результату видаляються внутрішні представлення. Це дозволяє зменшити обсяг вихідних даних у стандартному режимі інференсу та зосередитись лише на остаточному передбаченні структури білка.

Рециклування реалізовано як повністю диференційований процес, тому градієнти поширюються через усі ітерації. В режимі ініціалізації виконується лише одна ітерація, необхідна для створення модулів Haiku. Вихід моделі може бути зменшений за обсягом, якщо немає потреби повертати проміжні представлення.

Найбільше покращення точності спостерігається впродовж перших трьох-чотирьох ітерацій. Цього зазвичай достатньо для отримання надійного просторового передбачення. Збільшення кількості рециклів може давати додатковий приріст, однак із кожною новою ітерацією ефект поступово зменшується. Паралельно зростає й обчислювальна складність.

Тому більшість конфігурацій використовують обмежену кількість проходів, щоб зберегти баланс між точністю та продуктивністю.

3.4.2 Клас AlphaFoldIteration – реалізація однієї ітерації рециклування

Клас AlphaFoldIteration реалізує один крок рециклування в моделі AlphaFold. Його завдання – обробити вхідні представлення, провести їх через модулі обробки, сформувати оновлені ознаки та передати їх до відповідних «голів» (головних блоків передбачення). Цей компонент відіграє центральну роль у модульному дизайні архітектури.

Клас ініціалізується як підмодуль Haiku (hk.Module) і відповідає за повний процес обробки представлень у межах однієї ітерації. Його структура реалізує кроки, описані в алгоритмі 2 інференс з додатків до статті Jumper et al. (2021). Це зазначено безпосередньо в документації до класу у відкритому коді моделі (лістинг 3.5).

Лістинг 3.5 – Коментар до класу AlphaFoldIteration, що містить посилання на алгоритм 2 з додатків до статті Jumper et al. (2021)

```
class AlphaFoldIteration(hk.Module):
    """A single recycling iteration of AlphaFold
    architecture.
    Computes ensembled (averaged) representations from the
    provided features.
    These representations are then passed to the various heads
    that have been requested by the configuration file. Each
    head also returns a loss which is combined as a weighted
sum to produce the total loss.
    """
```

На вхід подається тензор вирівнювання (msa_act) разом із відповідною маскою (msa_mask). Ці дані проходять через модуль EmbeddingsAndEvoformer, який формує початкові ознаки. Після цього

компонент msa виділяється як окреме представлення, що не підлягає усередненню. Його зберігають окремо, а з основного словника представлень видаляють для подальшої обробки (лістинг 3.6).

Лістинг 3.6 – Виділення представлення msa для окремої обробки та виключення з подальшого ансамблювання

```
# MSA representations are not ensembled so
# we don’t pass tensor into the loop.
msa_representation = representations['msa']
del representations['msa']
```

За потреби може бути увімкнено ансамблювання (ensemble_representations=True). У такому випадку представлення усереднюються за кількома підвибірками вхідних ознак. Для кожної з них виконується незалежна обробка через Evoformer, результати підсумовуються, а після циклу – нормалізуються (лістинг 3.7).

Лістинг 3.7 – Усереднення представлень за кількістю елементів ансамблю. Компонент msa не нормалізується

```
for k in representations:
    if k != 'msa':
        representations[k]
          /= num_ensemble.astype(representations[k].dtype)
```

Цей підхід дає можливість знизити вплив шуму у вхідних даних і підвищити стабільність передбачення.

Ініціалізація «голів» передбачення відбувається відповідно до конфігурації моделі, яка визначає, які саме з них будуть активними під час запуску. До них належать:

– MaskedMsaHead – передбачення замаскованих позицій у MSA;

– DistogramHead – передбачення дистограм між парами залишків;

– StructureModule – модуль для побудови 3D-структури;

– PredictedLDDTHead – оцінка локальної якості передбачення (pLDDT);

– PredictedAlignedErrorHead – оцінка помилки вирівнювання;

– ExperimentallyResolvedHead – передбачення, які позиції були визначені експериментально.

Модулі ініціалізуються через словник head_factory, а деякі з них, зокрема StructureModule, використовуються через functools.partial, щоб передати додаткові параметри (наприклад, compute_loss).

Під час навчання для кожної «голови» обчислюється своя втрата, яка масштабується відповідним коефіцієнтом і додається до загальної суми втрат:

```
loss = head_config.weight * ret[name]['loss']
```

У разі активованого compute_loss сума втрат передається далі разом із передбаченнями. Деякі голови обробляються після модуля структури. Наприклад, PredictedLDDTHead використовує його вихід для оцінки якості передбачення, а PredictedAlignedErrorHead – для оцінки точності вирівнювання. Це дає можливість врахувати повний просторовий контекст під час обчислення додаткових метрик.

Завершальна частина методу визначає, що саме буде повернуто: лише передбачення або пара з передбаченнями та загальною втратою – залежно від значення параметра compute_loss (лістинг 3.8).

Лістинг 3.8 – Повернення результатів з урахуванням обчислення втрат

```
if compute_loss:
    return ret, total_loss
else:
    return ret
```

Клас AlphaFoldIteration демонструє, як модель поєднує кілька джерел інформації – MSA, парні взаємодії, 3D координати – для формування

комплексного представлення та побудови точних структурних передбачень. Гнучка система «голів» дозволяє моделі адаптуватися до різних завдань, а ансамблювання підвищує її стійкість до вхідних варіацій.

3.4.3 Механізм уваги (Attention) в AlphaFold

Увага є фундаментальним компонентом архітектури AlphaFold. Вона дозволяє моделі враховувати залежності між амінокислотами навіть на великих відстанях у послідовності. Це критично для точного передбачення структури білка, оскільки просторові взаємодії можуть виникати між віддаленими залишками.

Клас Attention реалізує стандартний підхід багатоголової уваги (Multi-head Attention), адаптований до біологічних даних. На початку ініціалізуються параметри: кількість голів, розміри ключів, запитів і значень.

Вхідні дані – це запити (q_data) та значення з ключами (m_data), а також бінарна маска mask, що вказує на валідні позиції. Розміри key_dim і value_dim діляться на кількість голів, після чого виконується проєкція запитів, ключів і значень у відповідні простори ознак. Для цього ініціалізуються окремі вагові параметри для кожної з проєкцій, а самі обчислення проводяться з використанням операцій einsum. Весь процес показано на лістингу 3.9.

Лістинг 3.9 – Ініціалізація ваг і обчислення запитів (q), ключів (k) та значень (v) у механізмі багатоголової уваги

```
q_weights = hk.get_parameter(
    'query_w', shape=(q_data.shape[-1], num_head,
key_dim),
    dtype=q_data.dtype,
    init=glorot_uniform())
```

Продовження лістингу 3.9

```
k_weights = hk.get_parameter(
    'key_w', shape=(m_data.shape[-1], num_head, key_dim),
    dtype=q_data.dtype,
    init=glorot_uniform())
v_weights = hk.get_parameter(
    'value_w',  shape=(m_data.shape[-1],num_head,
value_dim),
    dtype=q_data.dtype,
    init=glorot_uniform())
q  =  jnp.einsum('bqa,ahc->bqhc',  q_data,  q_weights)  *
key_dim**(-0.5)
k = jnp.einsum('bka,ahc->bhkc', m_data, k_weights)
v = jnp.einsum('bka,ahc->bhkc', m_data, v_weights)
```

Далі виконується обчислення логітів уваги як скалярного добутку між запитами та ключами. За потреби додається зміщення (nonbatched_bias), після чого застосовується маска для виключення нерелевантних позицій. Ваги уваги обчислюються за допомогою функції softmax, а значення v агрегуються з урахуванням цих ваг (лістинг 3.10).

Лістинг 3.10 – Обчислення ваг уваги та зважене агрегування зна-чень у механізмі багатоголової уваги

```
logits = jnp.einsum('bqhc,bhkc->bhqk', q, k)
if nonbatched_bias is not None:
    logits += jnp.expand_dims(nonbatched_bias, axis=0)
logits = jnp.where(mask, logits, _SOFTMAX_MASK)
weights = utils.stable_softmax(logits)
weighted_avg = jnp.einsum('bhqk,bhkc->bqhc', weights, v)
```

За конфігурацією, увага може бути підсилена механізмом воріт. Якщо параметр gating увімкнено, модель додатково обчислює значення воріт на основі проєкції q_data. Результат нормалізується сигмоїдальною функцією

та використовується для масштабування зваженого середнього. Це забезпечує моделі можливість адаптивно контролювати вплив інформації, яка проходить через механізм уваги (лістинг 3.11).

Лістинг 3.11 – Реалізація механізму воріт у багатоголовій увазі. Ворота застосовуються до зваженого середнього значень

```
if self.config.gating:
    gating_weights = hk.get_parameter(
        'gating_w',
        shape=(q_data.shape[-1], num_head, value_dim),
        dtype=q_data.dtype,
        init=hk.initializers.Constant(0.0))
    gating_bias = hk.get_parameter(
        'gating_b',
        shape=(num_head, value_dim),
        dtype=q_data.dtype,
        init=hk.initializers.Constant(1.0))
    gate_values = jnp.einsum(
        'bqc,  chv->bqhv', q_data,  gating_weights)       +
gating_bias
    gate_values = jax.nn.sigmoid(gate_values)
    weighted_avg *= gate_values
```

На завершальному етапі результат з усіх голів проєктується у вихідний простір фіксованої розмірності. Для цього застосовуються навчені ваги та зміщення (лістинг 3.12).

Лістинг 3.12 – Фінальна проєкція результату уваги у вихідний простір заданої розмірності

```
o_weights = hk.get_parameter(
    'output_w', shape=(num_head,  value_dim,
self.output_dim),
```

Продовження лістингу 3.12

```
        dtype=q_data.dtype,
        init=init)

    o_bias = hk.get_parameter(
        'output_b', shape=(self.output_dim,),
        dtype=q_data.dtype,
        init=hk.initializers.Constant(0.0))

    output = jnp.einsum('bqhc,hco->bqo', weighted_avg,
    o_weights) + o_bias
```

Щоб ігнорувати заповнювачі у вхідних даних, до логітів уваги застосовується маска. Для позицій із mask=False задається велике від'ємне значення, що обнуляє їхній внесок після softmax:

```
logits = jnp.where(mask, logits, _SOFTMAX_MASK)
```

Модель використовує кілька спеціалізованих варіантів базової уваги:

– MSARowAttentionWithPairBias – увага по рядках MSA з урахуванням парних взаємодій;

– MSAColumnAttention – увага по стовпцях MSA;

– MSAColumnGlobalAttention – глобальна стовпцева увага;

– TriangleAttention – трикутна увага, що застосовується до парних представлень.

Кожен з цих варіантів модифікує базову структуру Attention, адаптуючи її під особливості обробки білкових структур.

Механізм уваги в AlphaFold – складний, але надзвичайно ефективний. Він дає змогу моделі враховувати як локальні, так і глобальні залежності між залишками. Це дозволяє точно моделювати просторові взаємодії, масштабуватись до довгих білкових послідовностей і використовувати різні джерела інформації, зокрема MSA, парні представлення та структурні контексти.

3.4.4 TriangleAttention у моделі AlphaFold

Механізм TriangleAttention – це спеціалізована модифікація уваги, призначена для обробки матриці парних взаємодій між амінокислотами. На відміну від стандартної багатоголової уваги, цей модуль працює з трикутними структурними шаблонами і дозволяє враховувати складні просторові залежності між парами залишків. TriangleAttention є одним із центральних компонентів AlphaFold.

TriangleAttention реалізовано у двох варіантах:

– TriangleAttentionStartingNode – виконує увагу по стовпцях матриці парних взаємодій;

– TriangleAttentionEndingNode – виконує увагу по рядках.

Обидва варіанти використовують спільну логіку, але з різною орієнтацією. Якщо в конфігурації встановлено orientation == 'per_column', то вхідні тензори транспонуються:

```
if c.orientation == 'per_column':
    pair_act = jnp.swapaxes(pair_act, -2, -3)
    pair_mask = jnp.swapaxes(pair_mask, -1, -2)
```

Це дає можливість використовувати один і той самий модуль як для рядкової, так і для стовпцевої обробки. Перш ніж передати дані до механізму уваги, модель виконує нормалізацію вхідного тензора pair_act. Це робиться за допомогою шару LayerNorm, що покращує стабільність обчислень і пришвидшує навчання (лістинг 3.13).

Лістинг 3.13 – Нормалізація вхідних представлень у TriangleAttention за допомогою LayerNorm

```
pair_act = common_modules.LayerNorm(
    axis=[-1],   create_scale=True,   create_offset=True,
name='query_norm')(
    pair_act)
```

Продовження лістингу 3.13

```
pair_act = common_modules.LayerNorm(
    axis=[-1],   create_scale=True,   create_offset=True,
name='query_norm')(
    pair_act)
```

Після нормалізації модель обчислює зміщення (nonbatched_bias), яке використовується в механізмі уваги. Воно формується через додаткову проєкцію вхідних ознак із використанням параметра feat_2d_weights. Завдяки цьому механізм уваги може динамічно зсувати фокус на основі локального контексту (лістинг 3.14).

Лістинг 3.14 – Обчислення зміщення (bias) для уваги на основі парних ознак у TriangleAttention

```
init_factor = 1. / jnp.sqrt(int(pair_act.shape[-1]))

weights = hk.get_parameter(
    'feat_2d_weights',
    shape=(pair_act.shape[-1], c.num_head),
    dtype=pair_act.dtype,
    init=hk.initializers.RandomNormal(stddev=init_factor))

nonbatched_bias   =   jnp.einsum('qkc,ch->hqk',   pair_act,
weights)
```

Після нормалізації та обчислення зміщення, до даних застосовується стандартний модуль Attention. Для цього використовується допоміжна функція inference_subbatch, яка дозволяє розбити обчислення на підпартії. Це знижує споживання пам’яті, що особливо важливо при довгих послідовностях (лістинг 3.15).

Лістинг 3.15 – Застосування механізму уваги в TriangleAttention із підпартійною обробкою (subbatching) та врахуванням зміщення (nonbatched_bias)

```
attn_mod = Attention(
    c, self.global_config, pair_act.shape[-1])
pair_act = mapping.inference_subbatch(
    attn_mod,
    self.global_config.subbatch_size,
    batched_args=[pair_act, pair_act, mask],
    nonbatched_args=[nonbatched_bias],
    low_memory=not is_training)
```

TriangleAttention вирізняється кількома важливими перевагами, які роблять його одним із ключових елементів архітектури AlphaFold:

– він є ефективним для роботи з симетричними матрицями парних взаємодій. Його структура спеціально адаптована під обробку таких даних, що дозволяє точніше моделювати взаємозв’язки між амінокислотами на основі відстаней;

– модуль має високу гнучкість. Через параметр orientation можна змінювати напрям обробки – по рядках або по стовпцях, що дає змогу використовувати одну реалізацію для обох варіантів обчислень;

– TriangleAttention масштабований до довгих послідовностей. Він підтримує підпартійну обробку (subbatching), яка зменшує навантаження на пам’ять і забезпечує стабільну роботу з великими білками;

– додаткове зміщення nonbatched_bias, що обчислюється на основі локальних ознак, підсилює здатність моделі вловлювати контекстуальні залежності. Це допомагає точніше фокусувати увагу на релевантних просторових патернах.

TriangleAttention є критично важливим блоком AlphaFold. Його конструкція надає моделі змогу обробляти трикутну структуру взаємодій

між парами амінокислот. Це робить його ключовим елементом у високоточному передбаченні просторової структури білків.

3.3.5 OuterProductMean у моделі AlphaFold

Компонент OuterProductMean є одним із ключових етапів у формуванні парних представлень білка на основі вирівнювання MSA (Multiple Sequence Alignment). Його мета – перетворити інформацію з множинного вирівнювання у вигляді векторів для кожної амінокислоти у формат, придатний для моделювання парних взаємодій між залишками.

Модуль приймає на вхід тензор act розмірності [N_seq, N_res, c_m], що містить MSA-представлення, а також відповідну маску mask. Подальша обробка включає кілька послідовних етапів.

Спочатку вхідні вектори проходять через LayerNorm. Це стабілізує значення перед проєкцією:

```
act = common_modules.LayerNorm([-1], True, True, name='layer_norm_input')(act)
```

Після нормалізації вектори перетворюються у два окремих простори left_act і right_act. Це необхідно для подальшого обчислення зовнішнього добутку. Для кожного вектору застосовується окремий лінійний шар з однаковою розмірністю виходу c.num_outer_channel. Маска mask обнуляє нерелевантні позиції (лістинг 3.16).

Лістинг 3.16 – Побудова лівого та правого проєктованих векторів із використанням маски та лінійних шарів

```
left_act = mask * common_modules.Linear(
    c.num_outer_channel,
    initializer='linear',
    name='left_projection')(act)
right_act = mask * common_modules.Linear(
    c.num_outer_channel,
```

Продовження лістингу 3.16

```
    initializer='linear',
    name='right_projection')(act)
```

Для обчислення фінального парного представлення задаються параметри output_w та output_b. Тип ініціалізації залежить від конфігурації: якщо zero_init = True, ваги заповнюються нулями (лістинг 3.17). В іншому випадку використовується стратегія VarianceScaling, яка забезпечує стабільність при навчанні.

Лістинг 3.17 – Ініціалізація вагів output_w та output_b залежно від конфігурації модуля OuterProductMean

```
if gc.zero_init:
    init_w = hk.initializers.Constant(0.0)
else:
    init_w   =   hk.initializers.VarianceScaling(scale=2.,
mode='fan_in')

output_w = hk.get_parameter(
    'output_w',
    shape=(c.num_outer_channel,       c.num_outer_channel,
self.num_output_channel),
    dtype=act.dtype,
    init=init_w)
output_b = hk.get_parameter(
    'output_b', shape=(self.num_output_channel,),
    dtype=act.dtype,
    init=hk.initializers.Constant(0.0))
```

Основна операція модуля – це побудова парного представлення шляхом зовнішнього добутку векторів left_act і right_act. Спочатку left_act транспонується для зручності обчислень. Далі виконується подвійний einsum: перший формує 4-вимірний тензор взаємодій, другий – застосовує

ваги output_w і додає зміщення output_b. На виході отримується тензор розмірності [N_res, N_res, c_z], який потім транспонується в потрібному порядку (лістинг 3.18).

Лістинг 3.18 – Оптимізоване обчислення зовнішнього добутку для побудови парного представлення

```
def compute_chunk(left_act):
    left_act = jnp.transpose(left_act, [0, 2, 1])
    act = jnp.einsum('acb,ade->dceb', left_act, right_act)
         act = jnp.einsum('dceb,cef->bdf', act, output_w)
     + output_b
    return jnp.transpose(act, [1, 0, 2])
```

Щоб уникнути перевантаження пам’яті при роботі з довгими білками, обчислення зовнішнього добутку виконуються поетапно – у вигляді чанків. Для цього використовується спеціальний виклик inference_subbatch, який розбиває вхідні тензори на підпакети та послідовно обробляє їх функцією compute_chunk. Такий підхід зберігає точність результату, але значно зменшує пік навантаження на GPU чи TPU (лістинг 3.19).

Лістинг 3.19 – Застосування підпакетної обробки для зменшення споживання пам’яті

```
act = mapping.inference_subbatch(
    compute_chunk,
    c.chunk_size,
    batched_args=[left_act],
    nonbatched_args=[],
    low_memory=True,
    input_subbatch_dim=1,
    output_subbatch_dim=0)
```

Останній етап – нормалізація виходу. Щоб зменшити вплив кількості послідовностей у MSA, результати обчислень діляться на кількість дійсних пар, визначених за маскою. Це усуває перекоси, пов’язані з відсутніми даними. Для стабільності до нормувального коефіцієнта додається невелике значення ε:

```
epsilon = 1e-3
norm = jnp.einsum('abc,adc->bdc', mask, mask)
act /= epsilon + norm
```

Модуль OuterProductMean формує основу для побудови точного парного представлення білка на основі вирівнювання. Усі етапи – від нормалізації до обчислення зовнішнього добутку та масштабування виходу – спрямовані на отримання глибоко контекстуалізованої матриці взаємодій.

Отримане представлення передається до механізмів уваги (TriangleAttention). Завдяки цьому модель аналізує взаємодії не лише в межах окремої послідовності, а й враховує дані з усього вирівнювання. OuterProductMean виконує роль мосту між локальними властивостями амінокислот і глобальними структурними зв’язками в білку.

3.3.6 Голови передбачення в AlphaFold

Заключним етапом у моделі AlphaFold є спеціалізовані модулі – голови передбачення (prediction heads), кожна з яких відповідає за окремий аспект структури або якості передбачення. Вони працюють незалежно, але використовують спільні представлення, сформовані попередніми шарами моделі.

Модуль MaskedMsaHead застосовує підхід, подібний до BERT, для передбачення замаскованих позицій у вирівнюванні MSA. Модель навчається реконструювати амінокислоту в замаскованій позиції на основі контексту, що сприяє покращенню якості представлень та глибшому аналізу

еволюційних закономірностей. Модуль приймає на вхід представлення MSA, сформоване попередніми шарами моделі. На основі цих даних він формує ймовірнісні передбачення амінокислот у позиціях, які були замасковані під час навчання. Такий підхід допомагає моделі краще враховувати еволюційні закономірності та контексти в послідовностях. Голова використовується виключно на етапі тренування, оскільки її мета покращити загальну якість внутрішніх представлень. Для обчислення втрат застосовується функція softmax cross-entropy, яка порівнює передбачення з істинними значеннями в позиціях, зазначених у масці bert_mask (лістинг 3.20).

Лістинг 3.20 – Реалізація голови передбачення MaskedMsaHead у коді AlphaFold

```
class MaskedMsaHead(hk.Module):
    """Head to predict MSA at the masked locations.

    The MaskedMsaHead employs a BERT-style objective to
reconstruct a masked version of the full MSA, based on a
linear projection of the MSA representation.
    """
    def     __init__(self,     config,     global_config,
name='masked_msa_head'):
        super().__init__(name=name)
        self.config = config
        self.global_config = global_config

        if global_config.multimer_mode:
            self.num_output=
len(residue_constants.restypes_with_x_and_gap)
        else:
            self.num_output = config.num_output
def __call__(self, representations,     batch,
```

Продовження лістингу 3.20

```
    is_training):
        del batch
        logits = common_modules.Linear(
            self.num_output,
     initializer=utils.final_init(self.global_config),
            name='logits')(
            representations['msa'])
        return dict(logits=logits)
    def loss(self, value, batch):
        errors = softmax_cross_entropy(
            labels=jax.nn.one_hot(batch['true_msa'],
num_classes=self.num_output),
            logits=value['logits'])
        loss = (jnp.sum(errors *  batch['bert_mask'],
axis=(-2, -1)) /
                (1e-8    +    jnp.sum(batch['bert_mask'],
axis=(-2, -1))))
        return {'loss': loss}
```

Голова PredictedLDDTHead відповідає за оцінку точності передбачення структури на рівні кожної амінокислоти. Основним показником виступає pLDDT – узагальнена версія метрики lDDT, яка використовується як міра впевненості моделі.

На вхід подається представлення з модуля структури. Після нормалізації та обробки через два активаційні шари модель виводить логіти, які відповідають імовірнісному розподілу точності. Ці логіти групуються у фіксовану кількість бінів, які формують градацію впевненості по залишках.

Функція втрат базується на softmax cross-entropy між передбаченими значеннями і справжнім lDDT, обчисленим на атомах CA. За потреби значення можуть фільтруватися залежно від роздільної здатності експериментальної структури (лістинг 3.21).

Лістинг 3.21 – Основна частина модуля PredictedLDDTHead: побудова логітів на основі представлення структури

```
act = representations['structure_module']
act = common_modules.LayerNorm(
    axis=[-1],
    create_scale=True,
    create_offset=True,
    name='input_layer_norm')(act)
act = common_modules.Linear(
    self.config.num_channels,
    initializer='relu',
    name='act_0')(act)
act = jax.nn.relu(act)
act = common_modules.Linear(
    self.config.num_channels,
    initializer='relu',
    name='act_1')(act)
act = jax.nn.relu(act)
logits = common_modules.Linear(
    self.config.num_bins,
    initializer=utils.final_init(self.global_config),
    name='logits')(act)
# Shape (batch_size, num_res, num_bins)
return dict(logits=logits)
```

Голова PredictedAlignedErrorHead відповідає за передбачення відстаневих помилок між залишками у вирівняних фреймах головного ланцюга. Результати використовуються для обчислення метрики TM-score, яка оцінює загальну схожість передбаченої та реальної просторової структури.

На вхід подається парне представлення (pair), сформоване попередніми шарами моделі. Дані проходять через лінійний шар, який формує тривимірний тензор логітів для кожної пари залишків та кожного

діапазону похибки. Додатково генерується вектор breaks, що задає межі бінів для оцінки відстаней.

Функція втрат базується на квантуванні фактичної похибки вирівнювання між афінними перетвореннями. Для кожної пари залишків визначається бін, у який потрапляє похибка, після чого обчислюється softmax cross-entropy між справжнім біном і передбаченими логітами. Значення втрат масштабуються відповідно до маски вирівнювання та, за потреби, фільтруються за роздільною здатністю структури (лістинг 3.22).

Лістинг 3.22 – Побудова логітів та меж бінів у PredictedAlignedErrorHead на основі парного представлення

```
act = representations['pair']
# Shape (num_res, num_res, num_bins)
logits = common_modules.Linear(
    self.config.num_bins,
    initializer=utils.final_init(self.global_config),
    name='logits')(act)
# Shape (num_bins,)
breaks = jnp.linspace(
    0., self.config.max_error_bin, self.config.num_bins -
1)
return dict(logits=logits, breaks=breaks)
```

DistogramHead формує ймовірнісний розподіл відстаней між парами залишків білка. Такий підхід забезпечує детальну просторову інформацію, яка важлива для побудови тривимірної структури. На вхід подається парне представлення pair, яке проходить через лінійне перетворення з кількістю виходів, що відповідає кількості бінів. Вихід half_logits симетризується – до нього додається його транспонована копія, що гарантує симетричну матрицю відстаней. Також обчислюються межі бінів bin_edges, які задають діапазони відстаней у дистограмі. Функція втрат обчислюється через порівняння передбачених логітів з істинними бінованими відстанями між

атомами pseudo_beta. Використовується softmax cross-entropy, зважена маскою, що вказує на валідні пари (лістинг 3.23).

Лістинг 3.23 – Побудова логітів і меж бінів у DistogramHead для просторових відстаней

```
half_logits = common_modules.Linear(
    self.config.num_bins,
    initializer=utils.final_init(self.global_config),
    name='half_logits')(
    representations['pair'])
logits = half_logits + jnp.swapaxes(half_logits, -2, -3)
breaks =  jnp.linspace(self.config.first_break,
self.config.last_break,
                       self.config.num_bins - 1)
return dict(logits=logits, bin_edges=breaks)
```

Голови передбачення реалізовані як окремі модулі, які можна незалежно активувати під час тренування.

Кожна має власну функцію втрат, що враховується при оптимізації, і працює зі спільними представленнями, сформованими базовою частиною моделі. Інтерпретація результатів:

– pLDDT сигналізує про впевненість моделі в локальних передбаченнях;

– помилки вирівнювання виявляють нестабільні ділянки структури;

– дистограми показують просторові відносини між залишками.

Голови формують завершальний рівень AlphaFold. Вони не тільки створюють кінцеві передбачення, а й забезпечують глибший аналіз якості структури – через оцінку впевненості, помилок і просторових зв’язків.

3.4 Аналіз роботи моделі AlphaFold на прикладі вибраних білків

На цьому етапі дослідження було проведено практичне застосування моделі AlphaFold для передбачення просторової структури чотирьох білків: TP53, CFTR, Spike-білка SARS-CoV-2 та лізоциму. Кожен з них має унікальні властивості – різну довжину, функціональні домени, локалізацію в клітині або вірусному середовищі. Це дало змогу оцінити роботу моделі на прикладах білків різної природи.

Передбачення здійснювалося за допомогою веб-інтерфейсу AlphaFold Server. Для кожного білка вручну вводилася амінокислотна послідовність, після чого запускалося передбачення. Результати аналізувалися за двома ключовими метриками:

– pLDDT (Predicted Local Distance Difference Test) – для оцінки локальної впевненості моделі;

– PAE (Predicted Aligned Error) – для перевірки узгодженості між просторовими частинами білка.

Далі наведено покроковий аналіз для кожного білка з фокусом на якість передбачення, характерні структурні риси та стабільність моделі у складних регіонах.

3.4.1 Аналіз структури білка TP53 (P04637)

TP53 – це один з найважливіших транскрипційних факторів у людини, що виконує функцію супресора пухлин. Мутації цього гена асоціюються з численними формами раку [14]. Його довжина становить 393 амінокислоти, і він містить кілька структурно різних доменів, включаючи ДНК-зв’язувальний, тетрамеризаційний та трансактиваційний.

Для передбачення структури білка було використано веб-інтерфейс AlphaFold Server. Амінокислотна послідовність TP53 була отримана з бази UniProt (ідентифікатор P04637) і вставлена у відповідне поле вводу. Після

цього було вказано тип молекули – Protein – та кількість копій – 1, після чого запущено обчислення за допомогою кнопки Continue and preview job (рисунок 3.1).

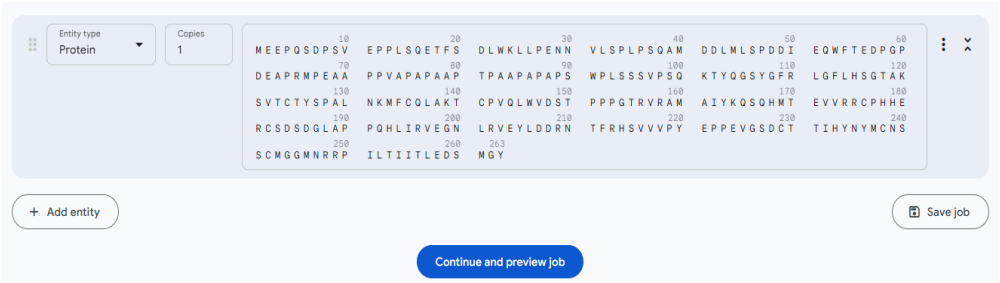

Рисунок 3.1 – Введення послідовності білка TP53 у веб-інтерфейсі AlphaFold Server

Після натискання кнопки Continue and preview job відкривається вікно підтвердження запуску, у якому автоматично присвоюється назва завдання та відображається коротка інформація про білок (тип, кількість копій, довжина послідовності). На цьому етапі також можна вручну задати Seed, але для аналізу використовувалося автоматичне значення за замовчуванням. Завдання запускалося через кнопку Confirm and submit job (рисунок 3.2).

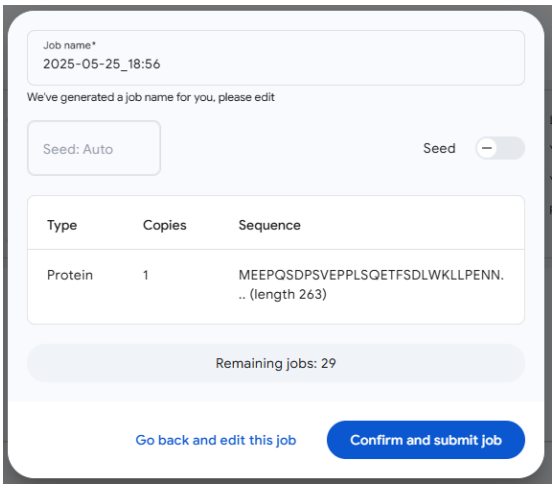

Рисунок 3.2 – Підтвердження параметрів передбачення для білка TP53 перед запуском моделі

Після підтвердження параметрів передбачення було створено завдання з автоматичною назвою 2025-05-25_18:56. Усі налаштування збережено, а сам процес моделювання відображався в історії запусків зі статусом «In progress» (рисунок 3.3).

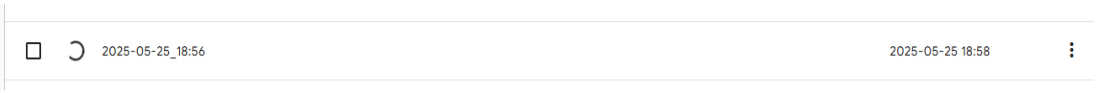

Рисунок 3.3 – Список активних завдань у веб-інтерфейсі AlphaFold Server

Після завершення моделювання відображаються результати з візуалізацією структури та супутніми метриками. Верхня панель показує градацію значень pLDDT – ключового показника впевненості моделі. Для TP53 переважають області з високою впевненістю: pLDDT > 90 (насичено синій колір), а також регіони з помірною (бірюзовий) та низькою (жовтий і помаранчевий) впевненістю (рисунок 3.4).

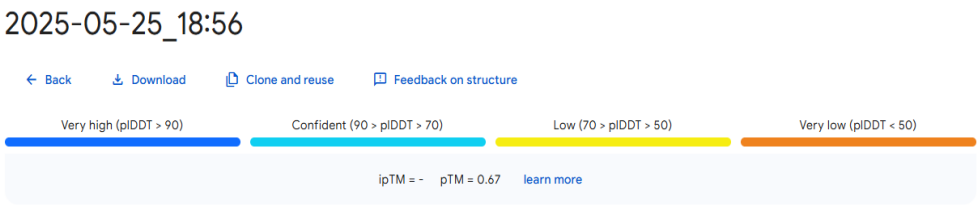

Рисунок 3.4 – Градація значень pLDDT у результатах передбачення TP53

Тривимірна модель білка наочно демонструє, що центральна частина має стабільну структуру – вона відображена у синьому кольорі. Це відповідає ДНК-зв'язувальному домену TP53. Натомість термінальні області (особливо N-кінцева) відзначаються високою гнучкістю і відображаються в жовтому та помаранчевому кольорах, що свідчить про невпорядкованість цих сегментів (рисунок 3.5).

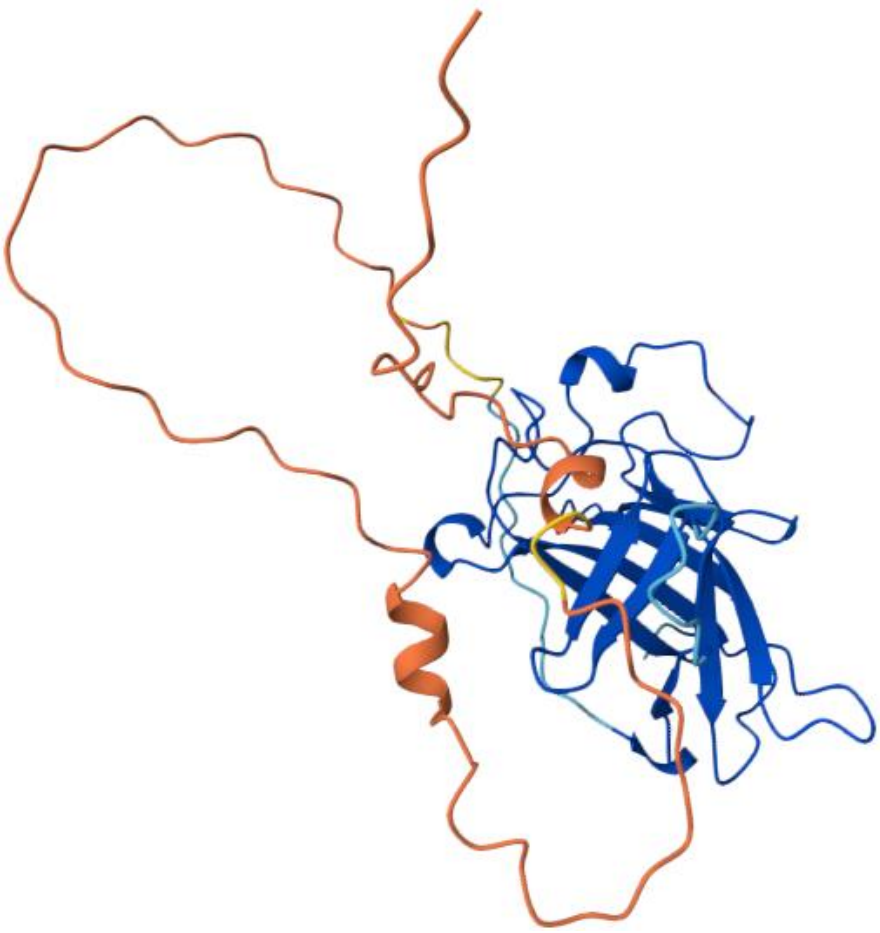

Рисунок 3.5 – Просторова структура TP53 із градацією залишків за pLDDT

Крім цього, модель генерує матрицю очікуваної помилки вирівнювання (PAE), яка відображає просторову узгодженість між різними частинами білка. У випадку TP53 видно чітку сегментацію: темніші ділянки матриці відповідають високій узгодженості в межах стабільних доменів, тоді як світліші зони вказують на більшу невизначеність щодо просторового положення гнучких фрагментів (рисунок 3.6).

Передбачена структура білка TP53 демонструє високу точність у стабільних регіонах, насамперед у ДНК-зв’язувальному домені. Гнучкі кінцеві ділянки мають нижчий рівень впевненості, що є очікуваним для внутрішньо невпорядкованих білкових сегментів.

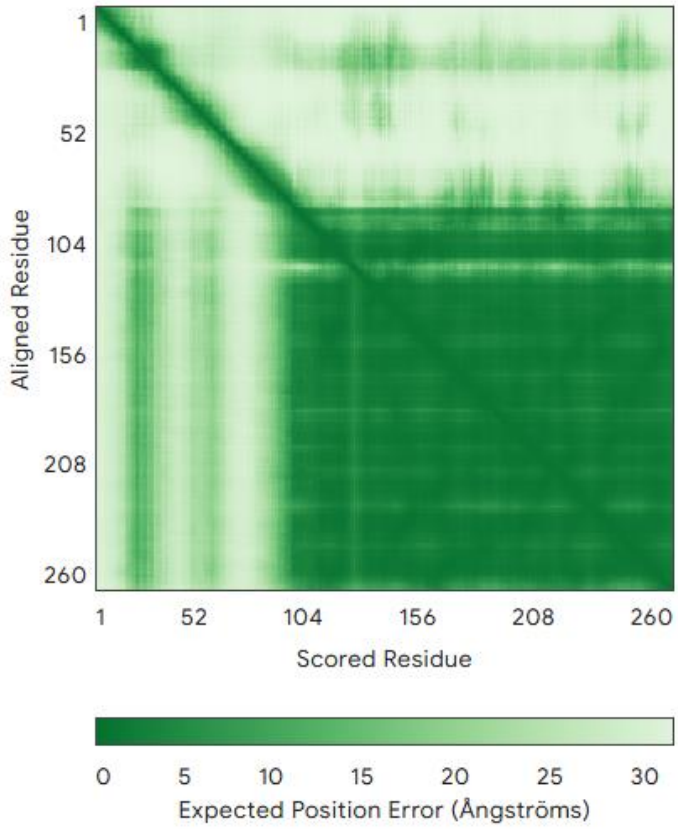

Рисунок 3.6 – Матриця Predicted Aligned Error (PAE) для білка TP53

Аналіз pLDDT та PAE у комплексі дає змогу оцінити як локальну якість структури, так і глобальну узгодженість її елементів.

3.4.2 Аналіз структури білка CFTR (P13569)

Передбачення структури трансмембранного білка CFTR (UniProt ID: P13569) було здійснено у веб-версії AlphaFold Server. Процедура запуску передбачення виконувалась аналогічно до попереднього прикладу з білком TP53. Нижче наведено FASTA-послідовність, що була використана для запуску моделі (рисунок 3.7).

Однією з найбільш вивчених мутацій CFTR є делеція фенілаланіну в позиції 508 (F508del), яка спостерігається у понад 70% пацієнтів з муковісцидозом [8].

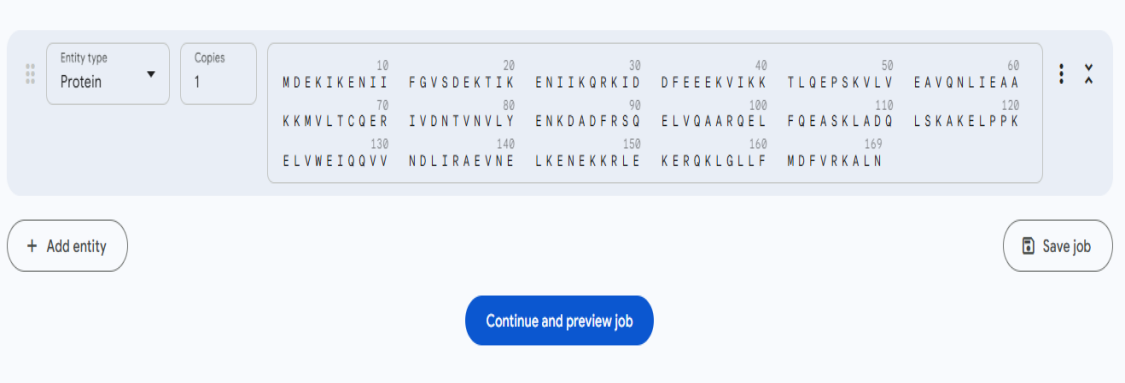

Рисунок 3.7 – FASTA-послідовність білка CFTR у веб-інтерфейс AlphaFold Server

Після завершення обчислень модель повернула результати, що включають графік довіри pLDDT, тривимірну структуру білка та матрицю очікуваних позиційних похибок (PAE).

На основі шкали pLDDT видно, що значна частина структури знаходиться в жовтому та помаранчевому діапазоні (pLDDT < 70), що свідчить про помірну або низьку впевненість у цих ділянках. Це очікувано, враховуючи складну топологію та гідрофобні ділянки, характерні для трансмембранних білків. Значення pTM становило 0.37, що також вказує на обмежену глобальну узгодженість моделі (рисунок 3.8).

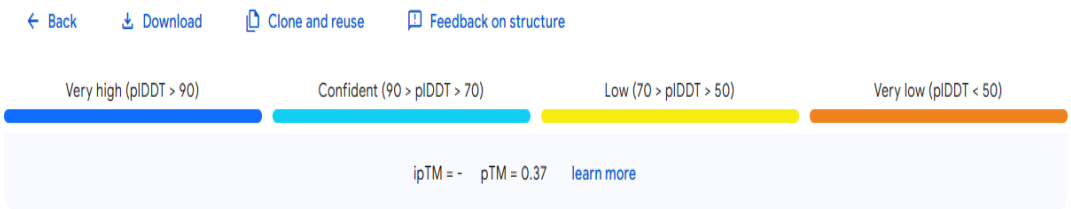

Рисунок 3.8 – Розподіл значень pLDDT для передбаченої структури CFTR

Тривимірне представлення моделі візуалізує характерні трансмембранні α-спіралі. Частина білка має більш впорядковану

конфігурацію (жовті області), тоді як кінцеві ділянки демонструють підвищену гнучкість і структурну неоднозначність (помаранчеве забарвлення) (рисунок 3.9).

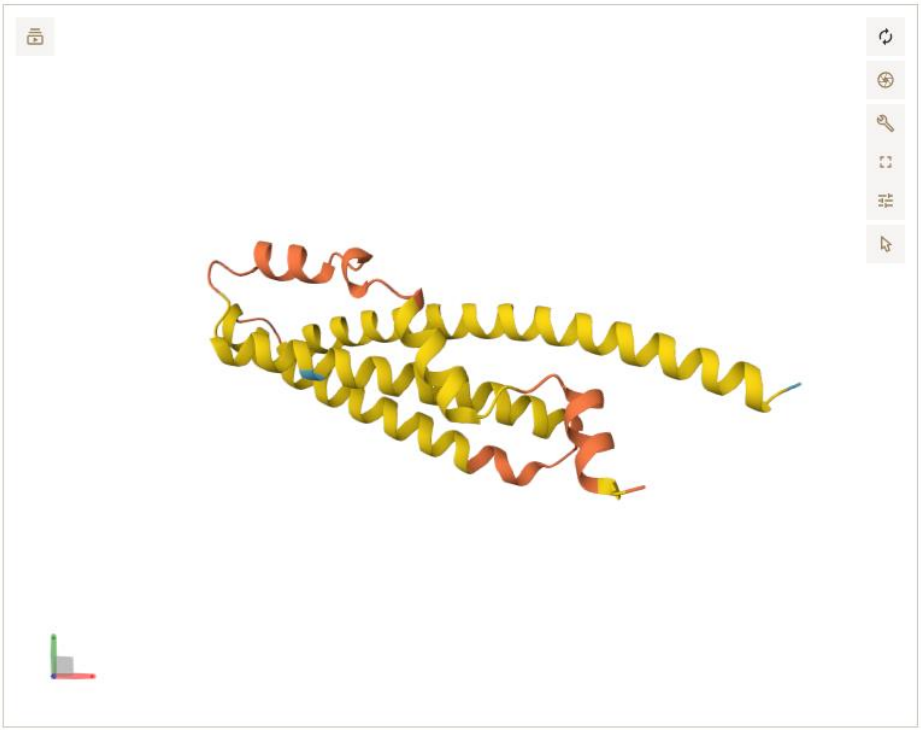

Рисунок 3.9 – Передбачена тривимірна структура білка CFTR

Матриця PAE показує більшу невизначеність у взаємному розташуванні віддалених ділянок, що відображається в менш чітких межах матриці. Це типово для білків із гнучкими лінкерами або кількома доменами, які можуть взаємодіяти з навколишнім середовищем по-різному (рисунок 3.10).

Структура білка CFTR містить характерні α-спіралі та ділянки зі змінною впевненістю. Високі значення pLDDT спостерігаються в ядрових сегментах, тоді як на периферії переважає невпорядкованість.

Значення pTM нижче 0.5 свідчить про обмежену надійність глобального передбачення. Результат підходить для подальшого аналізу мембранної топології та дослідження функціонально важливих областей.

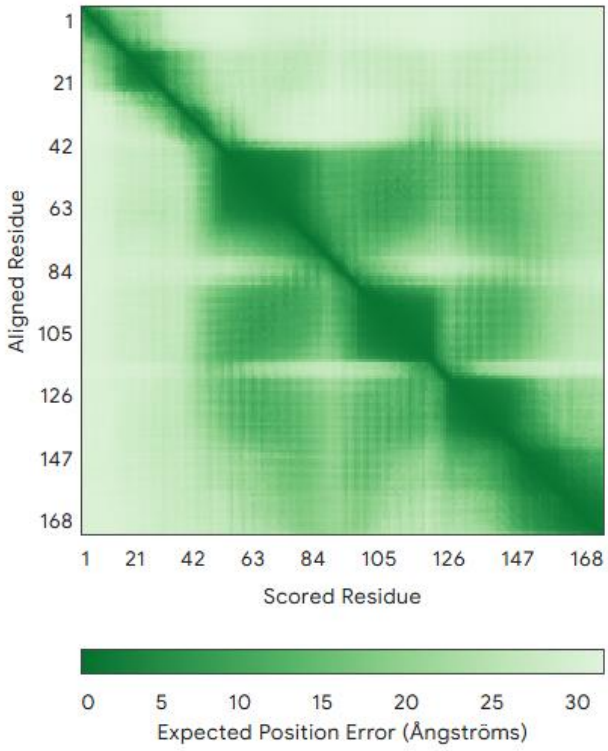

Рисунок 3.10 – Матриця очікуваних позиційних похибок (PAE) для CFTR

3.4.3 Аналіз структури білка Spike SARS-CoV-2 (P0DTC2)

Spike-білок є ключовим фактором проникнення вірусу SARS-CoV-2 до клітини-хазяїна, оскільки він зв’язується з рецептором ACE2 на поверхні клітин. Цей білок складається з 1273 амінокислот і має складну модульну будову, включаючи рецептор-зв’язувальний домен (RBD), S1/S2-сайти розщеплення та трансмембранну частину. Саме через ці особливості білок Spike обрано для структурного аналізу за допомогою AlphaFold.

Для передбачення було використано веб-інтерфейс AlphaFold. У поле вводу послідовностей було вставлено повну FASTA-послідовність білка (рисунок 3.11).

AlphaFold автоматично згенерував тривимірну структуру та розрахував метрики впевненості. Основною метрикою локальної якості є pLDDT (predicted Local Distance Difference Test).

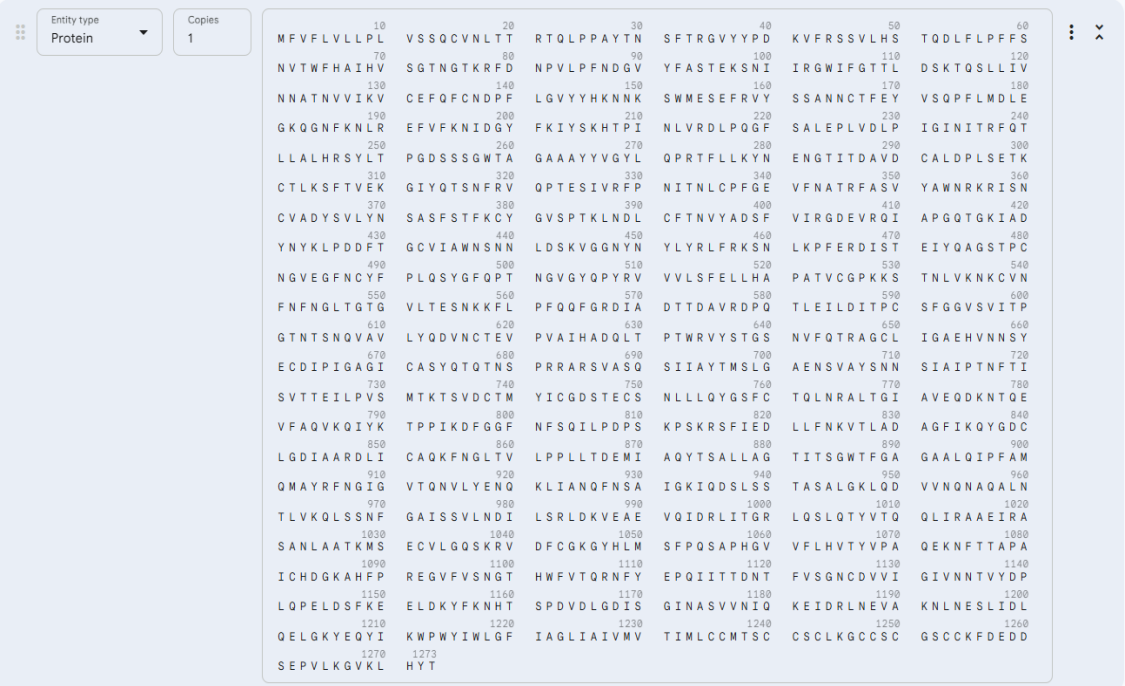

Рисунок 3.11 – FASTA-послідовність білка Spike SARS-CoV-2 (P0DTC2), використана для передбачення

Шкала значень pLDDT подана на рисунку 3.12 і використовується для кольорового кодування на 3D-візуалізації.

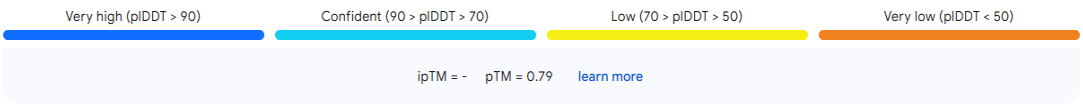

Рисунок 3.12 – Шкала значень pLDDT: від дуже високої (синій) до дуже низької (помаранчевий) впевненості

Отримана структура демонструє високу впевненість у передбаченні для більшості доменів, особливо в центральній частині білка, де зосереджені структурно стабільні елементи, зокрема рецептор-зв’язувальний домен (RBD). Петльові області та кінцеві сегменти мають значно нижчу впевненість, що узгоджується з їхньою високою гнучкістю (рисунок 3.13).

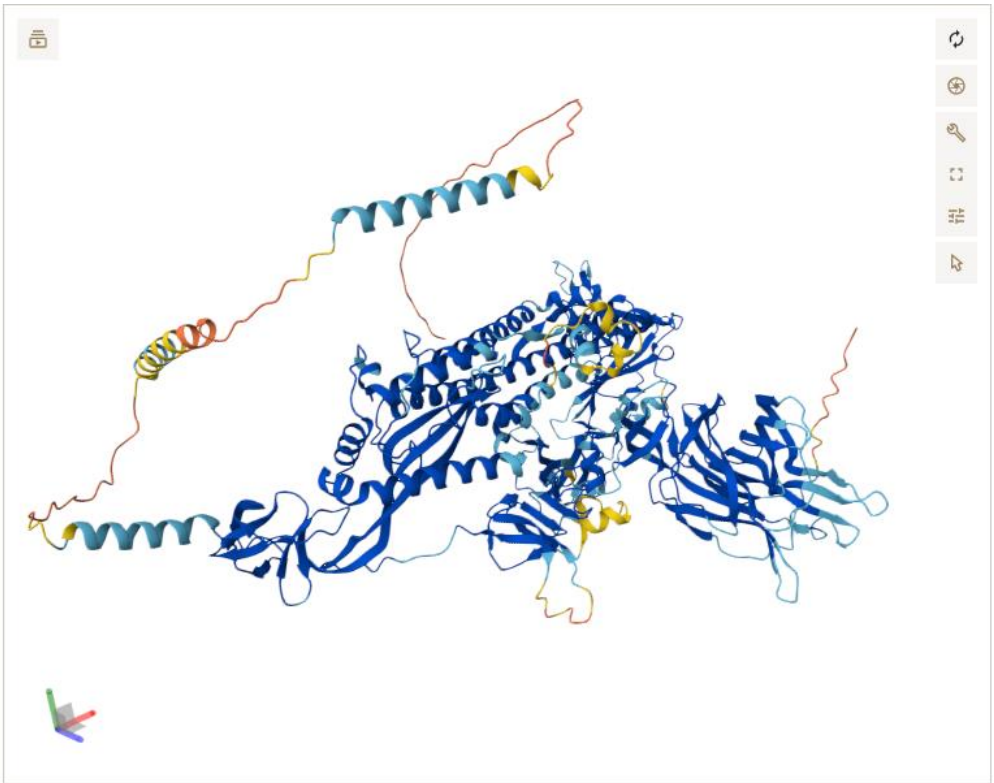

Рисунок 3.13 – 3D-структура Spike-білка з кольоровим кодуванням pLDDT

Значення глобальної метрики pTM (predicted TM-score) становить 0.79, що вказує на хорошу загальну достовірність моделі. Метрика ipTM не надається, оскільки передбачення виконувалося в мономерному режимі.

Для детального аналізу взаємного розташування доменів було розглянуто матрицю PAE (Predicted Aligned Error). Вона показує середнє очікуване відхилення позицій залишків відносно один одного. Як видно з рисунка 3.14, центральні ділянки моделі мають високу узгодженість (темно-зелений колір), а периферійні – вищу очікувану похибку (світліші зони).

AlphaFold успішно реконструював структуру великого й складного білка Spike. Високі значення pLDDT у функціональних доменах підтверджують надійність передбачення, тоді як низькі значення в кінцевих ділянках пояснюються їхньою гнучкістю.

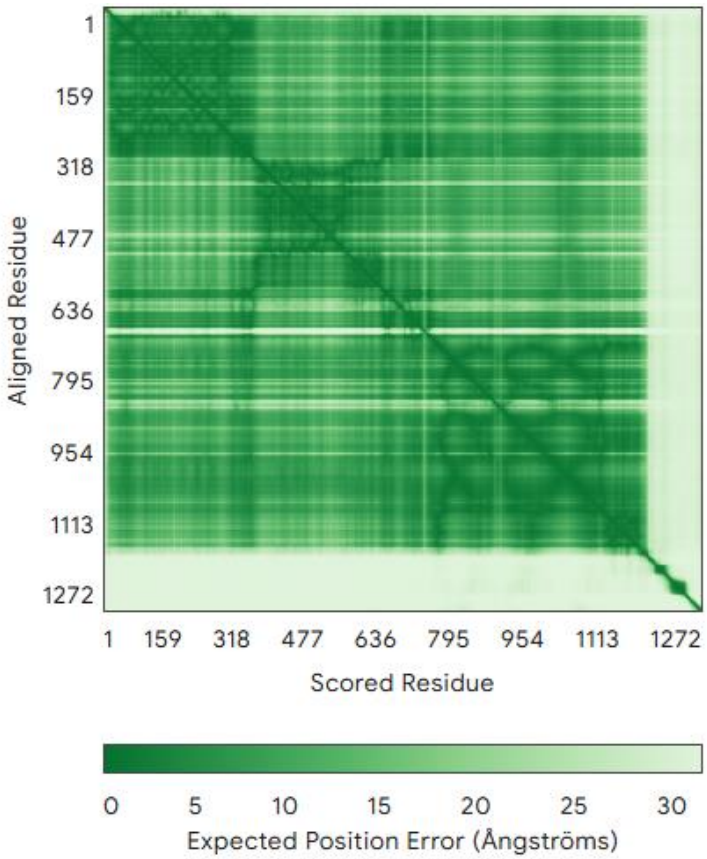

Рисунок 3.14 Матриця очікуваної похибки PAE для Spike-білка

Передбачена структура є потенційно корисною для досліджень мутацій, розробки терапевтик і вивчення механізмів зв’язування з рецептором ACE2.

3.3.4 Аналіз структури лізоциму (P00698)

Лізоцим – це фермент, що каталізує гідроліз β(1→4)-зв’язків у пептидогліканах бактеріальної клітинної стінки. Його часто використовують як модельний білок у дослідженнях структури та стабільності. У цьому дослідженні було проведено передбачення просторової структури лізоциму курячого яйця (Hen Egg-White Lysozyme, UniProt: P00698) за допомогою моделі AlphaFold [3].

Для запуску моделі у веб-версії AlphaFold було використано повну амінокислотну послідовність білка довжиною 129 залишків (рисунок 3.15).

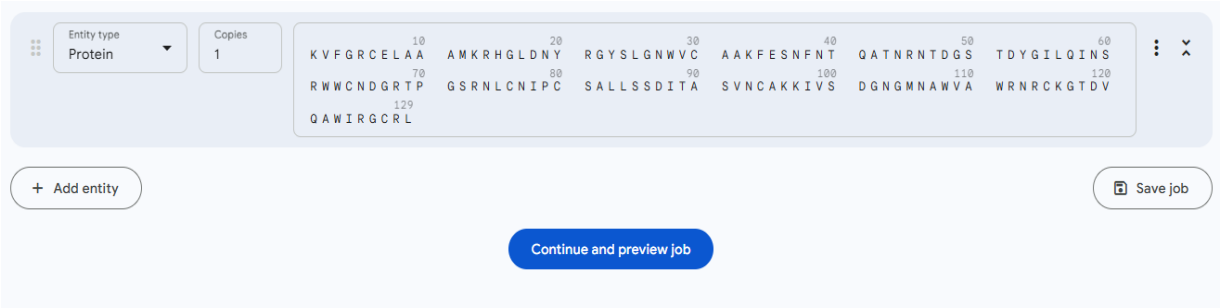

Рисунок 3.15 – FASTA-послідовність білка лізоциму, введена у веб-інтерфейс AlphaFold

Після обробки AlphaFold згенерував просторову модель білка та метрики впевненості. Шкала значень pLDDT подана на рисунку 3.16 і використовується для кольорового кодування рівнів упевненості у структурі.

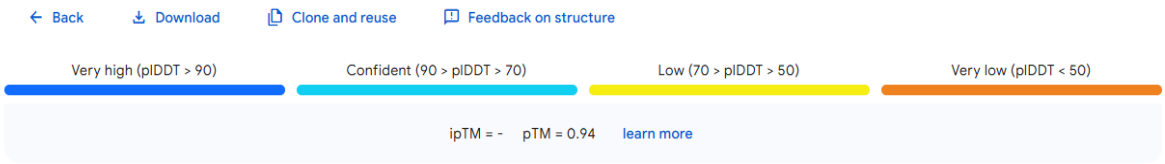

Рисунок 3.16 – Шкала pLDDT для інтерпретації локальної впевненості моделі AlphaFold

У результаті було отримано структуру з переважно дуже високими значеннями pLDDT (понад 90), що свідчить про високу впевненість моделі в передбаченій конформації. Це підтверджується також візуально на 3D-моделі – весь білок майже повністю зафарбований у темно-синій колір (рисунок 3.17).

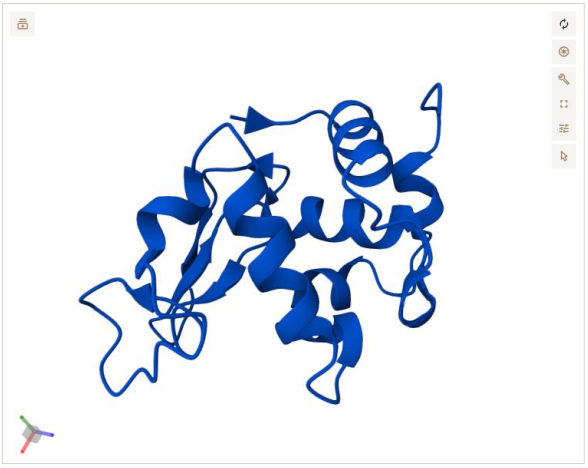

Рисунок 3.17 – Просторова структура лізоциму з кольоровим кодуванням pLDDT

Крім того, модель досягла дуже високого значення глобальної метрики pTM = 0.94, що вказує на високу достовірність просторової організації всієї молекули.

Аналіз матриці PAE (Predicted Aligned Error) також підтверджує стабільність структури. На рисунку 3.18 видно, що похибка між більшістю залишків невелика (насичено-зелений колір), що свідчить про узгоджене розташування доменів і стабільність білка.

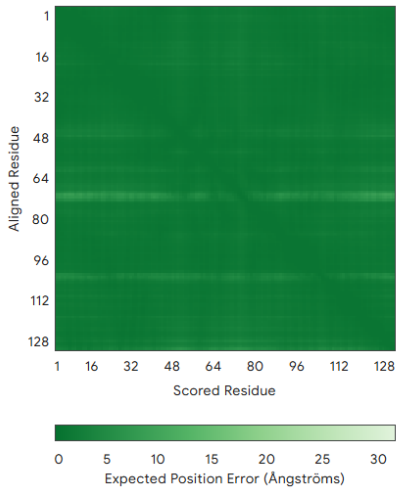

Рисунок 3.18 Матриця очікуваної похибки PAE для лізоциму

У межах цього розділу було здійснено передбачення просторових структур чотирьох білків різної природи за допомогою моделі AlphaFold. Серед них – білки людини (TP53, CFTR), вірусний білок (Spike SARS- CoV- 2) та фермент тваринного походження (лізоцим). Кожен з білків мав свою довжину, складність доменної організації та функціональне навантаження.

Модель AlphaFold продемонструвала високу точність у передбаченні структур для всіх прикладів:

– TP53 – були виявлені ділянки з високою впевненістю та зони гнучкості, що відповідає його природній структурній пластичності;

– CFTR – попри велику довжину, модель надала добре деталізовану структуру з якісним розділенням доменів;

– Spike-білок SARS-CoV-2 – AlphaFold успішно реконструював складну структуру з функціонально релевантною топологією рецептор-зв’язувального домену;

– лізоцим – компактна модель з дуже високими pLDDT та pTM значеннями, що підтверджує точність навіть для невеликих білків.

Загалом, результати показали, що модель AlphaFold є ефективним інструментом як для фундаментальних, так і для прикладних біоінформатичних досліджень. Метрики якості (pLDDT, pTM, PAE) корелюють з відомими біофізичними властивостями білків і можуть бути використані для подальшого аналізу стабільності, мутацій та взаємодій.

3.5 Порівняння структур білків з і без мутацій

Одним із ключових завдань цього дослідження є виявлення структурних змін, спричинених мутаціями в білках, які можуть мати функціональні наслідки. Для цього було проведено порівняльний аналіз просторових структур дикого типу (wild type) та мутантних форм чотирьох досліджуваних білків: TP53, CFTR, Spike SARS-CoV-2 та лізоциму.

Передбачення структур для мутантів виконувалися окремо, з використанням тих самих налаштувань у веб-інтерфейсі AlphaFold. Порівняння проводилося на основі візуального аналізу тривимірної структури, змін значень pLDDT, конфігурації доменів та матриць PAE.

У структурі дикого типу TP53 центральний ДНК-зв'язувальний домен демонстрував високу впевненість моделі (pLDDT > 90), що відображалося в інтенсивному синьому забарвленні тривимірної моделі. Для дослідження впливу мутації було змодельовано варіант R175H, в якому аргінін у 175-й позиції замінено на гістидин – це одна з найчастіших мутацій, що асоціюється з втратою функції TP53 у пухлинних клітинах. На рисунку 3.19 показано FASTA-послідовність мутантного білка, використану для передбачення.

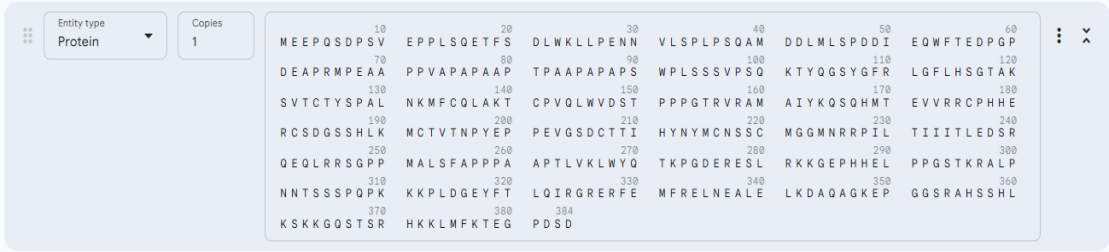

Рисунок 3.19 – FASTA-послідовність TP53 з мутацією R175H, введена у веб-інтерфейс AlphaFold

AlphaFold успішно згенерував модель просторової структури для мутантної форми. На рисунку 3.20 видно, що центральна частина білка зберігає загальну архітектуру, однак спостерігається помітне зниження впевненості (pLDDT) у ділянці навколо мутованого залишку – вона частково переходить у жовто-помаранчевий спектр, що свідчить про дестабілізацію структури в локальному регіоні. Крім того, глобальна метрика pTM знизилась до 0.41, що вказує на погіршену узгодженість всієї просторової конфігурації.

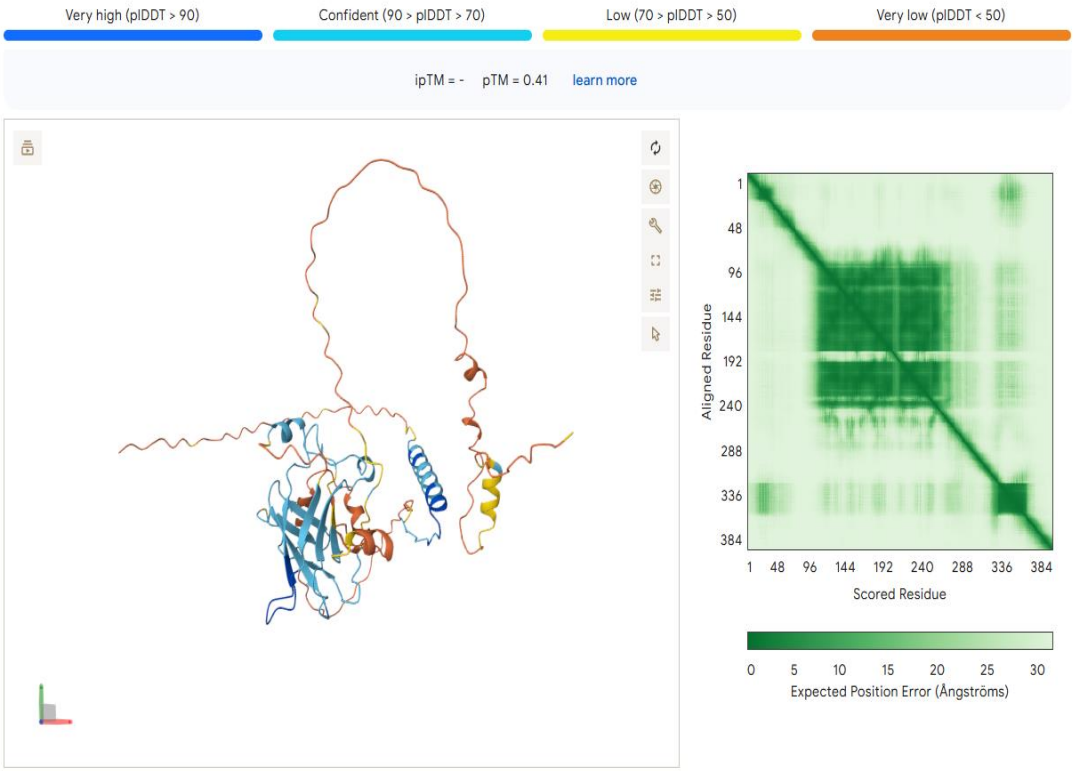

Рисунок 3.20 – pLDDT-градація, 3D-структура TP53 R175H та матриця PAE для мутантної форми білка

Матриця очікуваної похибки (PAE) також відображає знижену стабільність у порівнянні з дикою формою – спостерігається більша розмитість у зонах міждоменної взаємодії та гнучкіших фрагментах. Це може свідчити про порушення зв’язування з ДНК або втрату здатності до формування функціонального тетрамеру.

Для трансмембранного білка CFTR було змодельовано одну з найвідоміших патогенних мутацій – делецію залишку фенілаланіну в позиції 508 (ΔF508). Саме ця мутація є найпоширенішою причиною розвитку муковісцидозу та спричиняє порушення формування стабільної третинної структури білка. На рисунку 3.21 наведено FASTA-послідовність CFTR з мутацією ΔF508, яка була введена до веб-інтерфейсу AlphaFold для отримання просторової моделі.

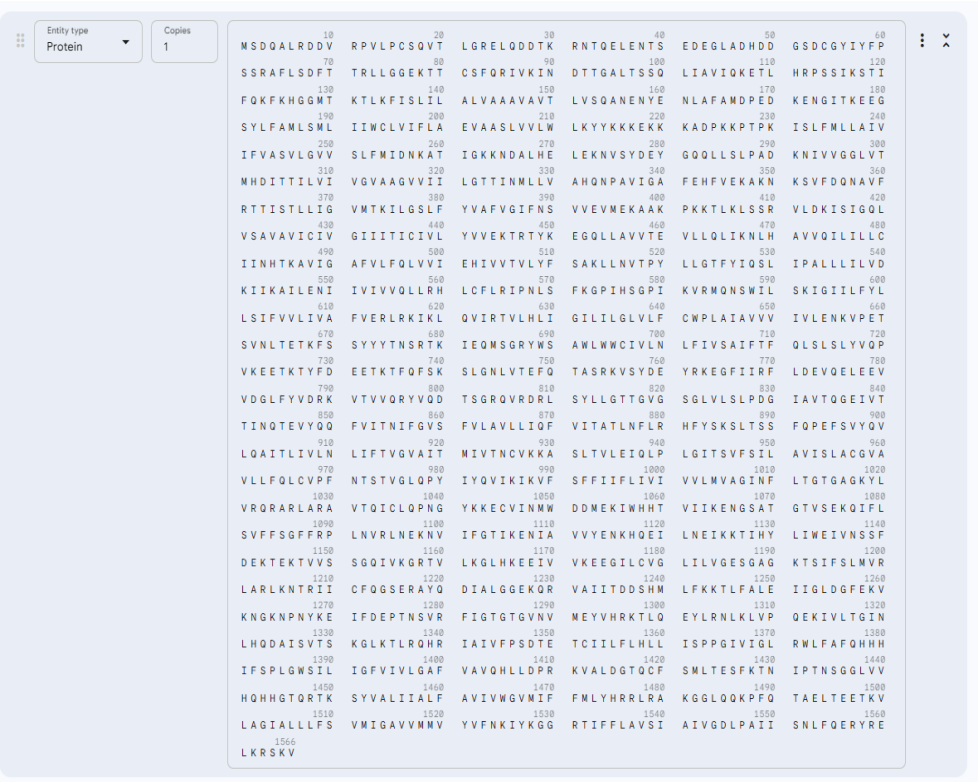

Рисунок 3.21 – FASTA-послідовність білка CFTR з мутацією ΔF508

Отримана модель виявила значне зниження довіри в оцінці структури. На рисунку 3.22 видно, що більша частина білка зафарбована у помаранчевий колір, що відповідає значенням pLDDT < 50, тобто дуже низькій впевненості моделі. Це свідчить про структурну нестабільність і часткову втрату впорядкованості. Метрика глобальної узгодженості pTM склала лише 0.21, що значно нижче порівняно з дикою формою білка. Це вказує на порушення просторової організації великих доменів та потенційне неправильне згортання білка.

Аналіз матриці очікуваної похибки (PAE) демонструє розмиту картину у ділянках міждоменної взаємодії, що є наслідком підвищеної гнучкості та порушення глобальної топології. Це підтверджує, що мутація ΔF508 призводить не лише до локальних змін, а й до суттєвої втрати стабільної складеної структури білка.

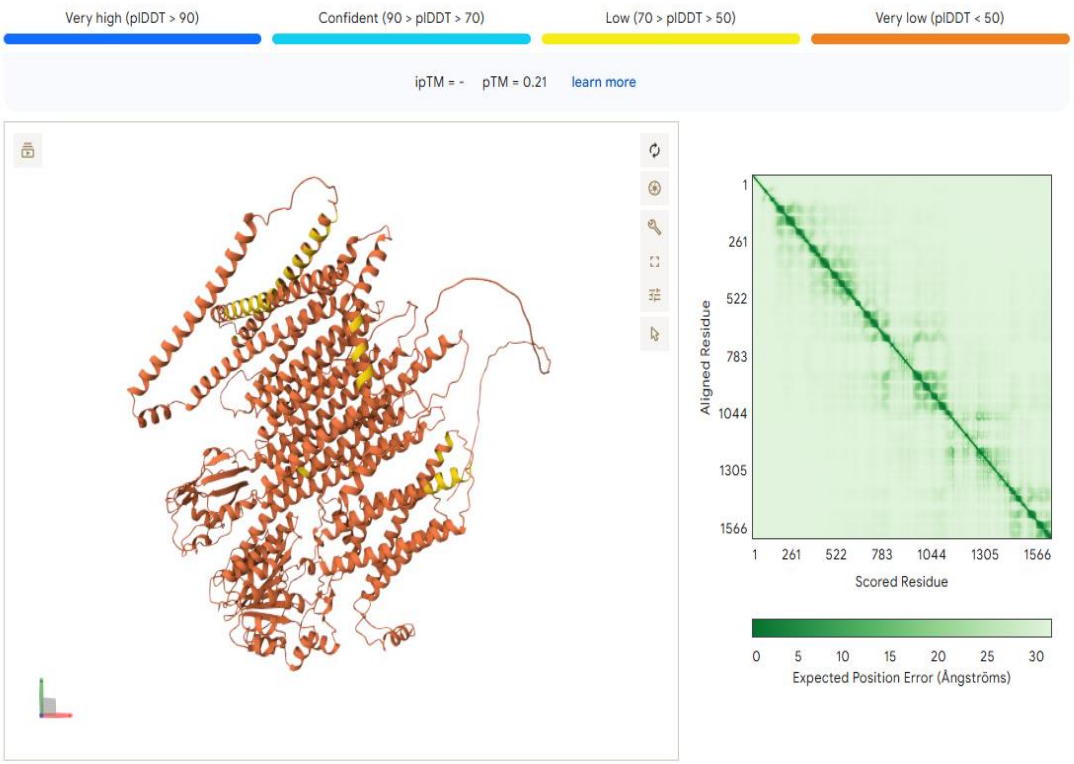

Рисунок 3.22 – Структура CFTR з мутацією ΔF508: pLDDT-графік, 3D-модель і матриця PAE

Однією з ключових мутацій у Spike-білку SARS-CoV-2 є N501Y – заміна аспарагіну на тирозин у 501-й позиції. Вона характерна для кількох варіантів вірусу, зокрема «Альфа» та «Омікрон», і асоціюється з підвищеною афінністю до рецептора ACE2. У цьому дослідженні було змодельовано просторову структуру Spike-білка з мутацією N501Y для виявлення потенційних структурних змін.

На рисунку 3.23 подано FASTA-послідовність мутантного білка, яка була використана для запуску моделі в веб-версії AlphaFold.

Передбачена структура зберігає загальну топологію дикого типу, однак демонструє локальні зміни в області рецептор-зв’язувального домену (RBD), де розташовано позицію 501. У цій ділянці модель продовжує показувати високу впевненість (pLDDT > 90), однак сусідні регіони стали дещо більш гнучкими, про що свідчить поява жовтого та бірюзового

забарвлення в структурі (рисунок 3.24). Значення глобальної метрики pTM становить 0.81, що свідчить про високу узгодженість просторової організації білка загалом.

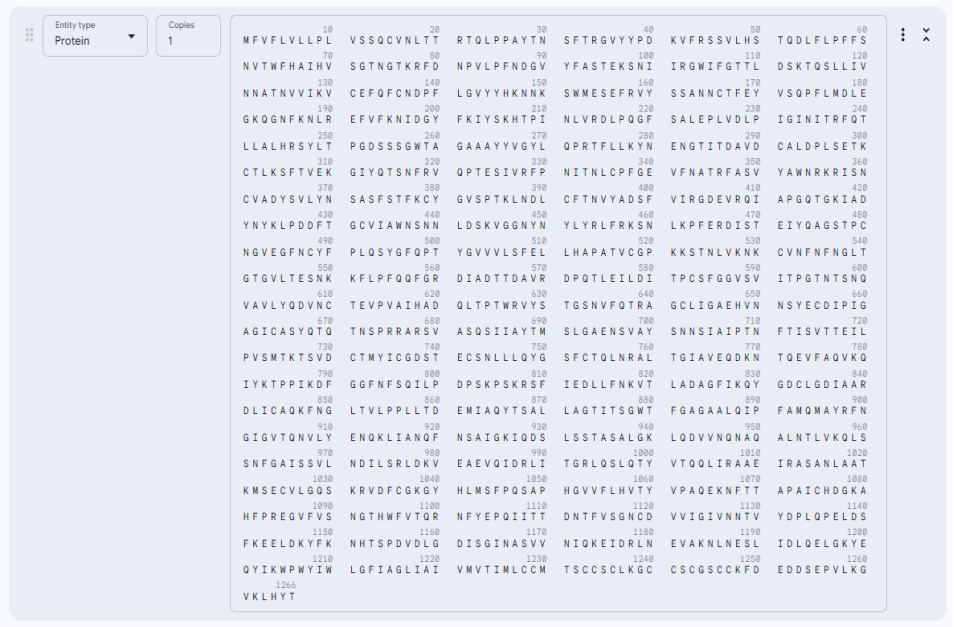

Рисунок 3.23 – FASTA-послідовність Spike-білка SARS-CoV-2 з мутацією N501Y

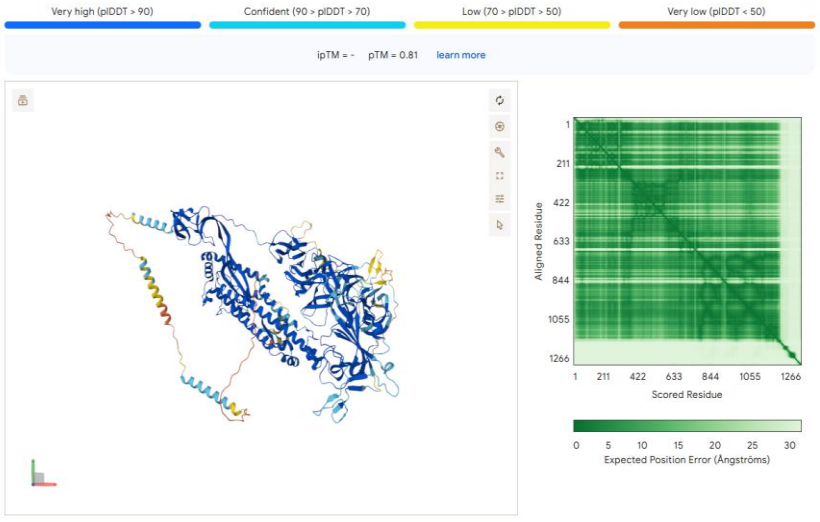

Рисунок 3.24 – Тривимірна структура Spike-білка з мутацією N501Y, градація pLDDT і матриця PAE

Матриця очікуваної похибки (PAE) не виявляє суттєвих глобальних порушень, однак вказує на невеликі зміни у взаємному розташуванні залишків у RBD, що може впливати на зв’язування з ACE2. Таким чином, мутація N501Y не дестабілізує білок, а скоріше спричиняє функціонально релевантні локальні зміни, які можуть впливати на вірусну інфекційність.

Лізоцим – це добре вивчений фермент, що розщеплює β(1→4)-зв’язки в пептидогліканах клітинної стінки бактерій. Однією з критичних амінокислот є глутамінова кислота в позиції 35 (E35), яка разом із D52 утворює каталізуючу діаду. Для оцінки структурних змін, спричинених втратою функції, було змодельовано мутантну форму з заміною E35 на A (аланін). FASTA-послідовність мутантного білка, яка використовувалася для запуску моделі, наведена на рисунку 3.25.

Рисунок 3.25 – FASTA-послідовність лізоциму з мутацією E35A

Отримана структура майже ідентична дикій формі за просторовою організацією. Значення pLDDT > 90 по всій довжині білка вказує на збереження глобальної стабільності. pTM також залишилося високим (0.94), що свідчить про відсутність значущих змін у третинній структурі (рисунок 3.26).

Однак, локальні зміни в каталізуючій ділянці, хоч і не впливають на загальну геометрію, потенційно можуть повністю деактивувати ферментативну активність. Високий рівень pLDDT у зоні мутації свідчить про стабільну, але нефункціональну конформацію.

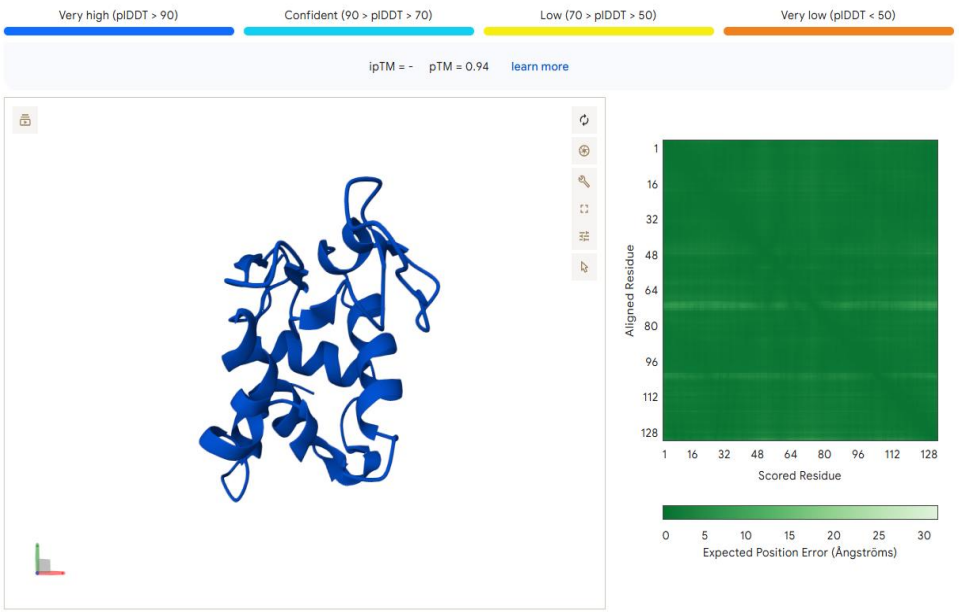

Рисунок 3.26 – 3D-структура лізоциму E35A, pLDDT-графік і матриця PAE

Проведене порівняння структур білків дикого типу та їхніх мутантних форм дозволило оцінити характер впливу мутацій на просторову організацію молекул. Для кожного з чотирьох досліджених білків (TP53, CFTR, Spike SARS-CoV-2, лізоцим) спостерігалися різні типи структурних змін – від локальної гнучкості до глобального порушення стабільності.

TP53 R175H – точкова мутація в ДНК-зв'язувальному домені призвела до локального зниження впевненості моделі та порушення просторової впорядкованості регіону.

CFTR ΔF508 – видалення критичного залишку спричинило значне зниження pLDDT та pTM, а також порушення глобальної топології, характерне для неправильного згортання.

Spike N501Y – структурно стабільна мутація з локальними змінами у рецептор-зв'язувальному домені, які можуть мати функціональне значення для зв'язування з ACE2.

Лізоцим E35A – структура залишилась стабільною, проте мутація зачепила каталітичний центр, що може блокувати ферментативну активність без суттєвих геометричних змін.

Навіть точкові мутації можуть мати різний структурний ефект, який залежить від локалізації залишку, його ролі у функціональній ділянці білка та взаємодії з іншими елементами структури. Комбінація метрик pLDDT, pTM та PAE виявилася ефективним інструментом для виявлення таких змін та інтерпретації потенційних функціональних наслідків.

3.6 Виявлення ключових структурних змін, викликаних мутаціями

У цьому підрозділі зосереджено увагу на локалізованих змінах у структурі білків, які виникають унаслідок мутацій. На відміну від загального порівняння просторової конфігурації, тут розглянуто конкретні залишки, домени та фрагменти, у яких спостерігається порушення геометрії, зниження впевненості моделі або підвищення гнучкості. Особливу увагу приділено функціонально важливим ділянкам, у яких навіть незначні зміни можуть призвести до втрати або зміни біологічної активності білка. У таблиці 3 наведено основні структурні зміни, зафіксовані в моделях мутантних білків у порівнянні з диким типом. Порівняння охоплює локалізацію змін, характер впливу на структуру, зміну ключових метрик та потенційні функціональні наслідки (таблиця 3.1).

Аналіз структурних змін показав, що характер мутації та її локалізація визначають масштаб і тип порушень у білковій структурі. Точкові заміни в критичних функціональних ділянках (як у TP53 і лізоцимі) спричиняють локальні дестабілізації або втрату активності без істотних змін глобальної конфігурації. Натомість делеція у CFTR (ΔF508) викликає комплексне порушення згортання й загальну нестабільність. У випадку Spike-білка, мутація N501Y зумовлює локальну перебудову, яка потенційно змінює взаємодію з рецептором, не порушуючи загальну топологію.

Таблиця 3.1 – Структурні зміни, викликані мутаціями у досліджуваних білках

Білок	Мутація	Залишок / область	Тип зміни	Ефект на структуру	Метрики (pLDDT / pTM / PAE)	Потенційний функціональний наслідок
TP53	R175H	ДНК-зв'язувальний домен (β-шар)	Локальна дестабілізація	Викривлення β-шару, зниження впевненості	pLDDT↓, pTM↓, PAE↑	Може втратити здатність зв'язуватись з ДНК
CFTR	ΔF508	NBD1, трансмембранні домени	Глобальна дестабілізація	Порушення топології, гнучкість	pLDDT↓, pTM↓, PAE↑↑	Порушене згортання, втрата функції каналу
Spike	N501Y	RBD (петля рецепторного зв'язування)	Локальна стабілізація/зміна конфігурації	Легка перебудова сусідніх залишків	pLDDT≈, pTM≈, PAE↑	Потенційне підвищення афінності до ACE2
Лізоцим	E35A	Каталітичний центр	Локальна нейтралізація активного сайту	Геометрія збережена, але активність втрачена	pLDDT≈, pTM≈, PAE≈	Зниження або втрата каталізу

Комбіноване використання локальних метрик (pLDDT), глобальних оцінок (pTM) та матриць PAE допомагає виявити й класифікувати вплив мутацій на структуру досліджуваних білків.

3.7 Оцінка стабільності білків на основі предиктивних метрик

У межах цього дослідження оцінювання стабільності білків здійснювалося з використанням прогнозних метрик, що надаються моделлю AlphaFold. Хоча для повноцінного аналізу енергетичних характеристик традиційно застосовують молекулярну динаміку або спеціалізовані інструменти для обчислення вільної енергії згортання, навіть без них можливо отримати цінну інформацію про якість і надійність передбаченої структури.

Усі варіанти білків як дикого типу, так і мутантні форми були проаналізовані на основі трьох ключових показників: локальної впевненості (pLDDT), глобальної узгодженості (pTM) та очікуваної похибки вирівнювання (PAE). Ці параметри відображають структурну цілісність білка на різних рівнях просторової організації.

Найбільш помітне зниження стабільності спостерігалося у випадку білка CFTR з мутацією ΔF508. Видалення фенілаланіну в критичній ділянці призвело до суттєвого зниження значень pLDDT по всій структурі, особливо в межах трансмембранних доменів. Одночасно глобальний показник pTM виявився низьким (0.21), а матриця PAE свідчила про втрату чіткої просторової узгодженості між великими сегментами молекули. Це свідчить про серйозні порушення у третинній структурі білка та потенційне неправильне згортання.

У білку TP53 з мутацією R175H зниження стабільності носить локальний характер. Зміна аргініну на гістидин у ДНК-зв’язувальному домені супроводжувалася зменшенням значень pLDDT у безпосередній околі мутованого залишку та локальним спотворенням β-шару. Це порушення просторової геометрії важливої функціональної ділянки потенційно знижує здатність TP53 до взаємодії з ДНК.

У випадку Spike-білка з мутацією N501Y загальна структурна стабільність зберігається. Значення pLDDT залишаються високими, не

виявлено змін у глобальному показнику pTM, а локальні перебудови у рецептор-зв’язувальній петлі не мають суттєвого впливу на третинну структуру. Аналогічну ситуацію спостерігалося й у лізоцимі з мутацією E35A. Незважаючи на те, що мутація зачіпає каталізуючий центр, просторове розташування основних елементів молекули залишається стабільним, що підтверджується високими значеннями pLDDT і низькою похибкою в матриці PAE.

Отримані результати демонструють, що навіть за відсутності енергетичного моделювання можливо провести надійне попереднє оцінювання впливу мутацій на структурну стабільність білків. Поєднання локальних і глобальних метрик AlphaFold формує цілісну картину змін і створює основу для подальшого функціонального аналізу.

3.8 Функціональна інтерпретація структурних змін білків

Виявлені під час дослідження просторові зміни, спричинені мутаціями, мають не лише геометричний або стабілізаційний характер, але й безпосередньо впливають на функціональну активність білкових молекул. Просторова структура визначає здатність білка виконувати свої біохімічні ролі – зв’язуватися з ДНК, каталізувати реакції, транспортувати іони або взаємодіяти з іншими білками. Навіть незначне порушення в конформації функціонально важливих ділянок може мати суттєві наслідки для біологічної активності білка або цілих клітинних процесів, у яких він задіяний.

На прикладі білка TP53 було досліджено мутацію R175H, яка зачіпає критичну для функції ділянку – ДНК-зв’язувальний домен. Просторове передбачення продемонструвало локальну дестабілізацію у β-шарі, зниження pLDDT та порушення геометрії цієї ділянки [11], [13]. Це свідчить про ймовірну втрату здатності білка до точного розпізнавання ДНК-послідовностей, що, у свою чергу, призводить до порушення

транскрипційної регуляції. Втрата функції TP53 пов’язана з утворенням злоякісних пухлин, і ця мутація є однією з найчастіших у ракових клітинах, що підтверджує її важливість.

Для CFTR було змодельовано делецію залишку фенілаланіну в позиції 508 (ΔF508), що є клінічно значущою мутацією при муковісцидозі. Просторовий аналіз показав зниження глобальної структурної цілісності та втрату впорядкованості білка. Особливо виражені порушення виявлено у трансмембранних доменах і доменах зв’язування АТФ (NBD1), що є критичними для канальної функції. Така зміна структури корелює з відомими експериментальними спостереженнями: білок у такому вигляді не транспортується до плазматичної мембрани, а затримується в ендоплазматичному ретикулумі. У результаті – втрата функції іонного транспорту, накопичення слизу та формування характерної клінічної картини захворювання.

У Spike-білку вірусу SARS-CoV-2 проаналізовано вплив мутації N501Y. Вона розташована у рецептор-зв’язувальному домені (RBD), що взаємодіє з білком ACE2 на поверхні клітин людини. Просторові зміни у цій зоні виявились локальними, але структурно релевантними: перебудова петлі навколо залишку 501 може призвести до покращеного позиціювання сайту зв’язування. Це, своєю чергою, сприяє підвищенню афінності Spike-білка до рецептора і, як наслідок, – ефективнішому проникненню вірусу в клітину. Такий механізм частково пояснює збільшену заразність варіантів SARS-CoV-2, які містять мутацію N501Y.

Щодо лізоциму, було змодельовано мутацію E35A, яка зачіпає активний центр ферменту. Отримана структура не демонструє глобальної дестабілізації або втрати компактності. Однак заміна глутамінової кислоти на аланін у каталізуючій позиції суттєво змінює хімічні властивості цієї ділянки. Втрата карбоксильної групи позбавляє фермент здатності передавати протон – ключову функцію в механізмі гідролізу. Тобто,

незважаючи на збережену просторову будову, ферментальна активність фактично блокується.

Загальний аналіз чітко демонструє, що структурні зміни, спричинені мутаціями, мають прямий зв’язок з функціональними наслідками, навіть якщо вони не супроводжуються великими геометричними порушеннями. У випадках TP53 і лізоциму спостерігається функціональна втрата через порушення активних ділянок. Для CFTR – зміна структури повністю блокує білок на рівні згортання. Натомість для Spike-білка зміни можуть, навпаки, покращувати функцію, що має важливі еволюційні та клінічні наслідки.

4 ЗАСТОСУВАННЯ ШТУЧНОГО ІНТЕЛЕКТУ В БІОМЕДИЧНИХ ДОСЛІДЖЕННЯХ

4.1 Ефективність AlphaFold у дослідженні білкових структур

Сучасні моделі штучного інтелекту здатні вирішувати завдання, які донедавна потребували складного й дорогого лабораторного обладнання. У цьому дослідженні було використано AlphaFold – одну з найпотужніших AI-моделей для передбачення просторової структури білків.

Достатньо лише амінокислотної послідовності, і система формує тривимірну модель, наближену за точністю до результатів експериментальних методів.

Під час аналізу вдалося змоделювати як дикі, так і мутантні форми чотирьох різних білків: TP53, CFTR, Spike SARS-CoV-2 та лізоциму. Кожен з них має інше біологічне значення та відрізняється за структурною складністю. AlphaFold упевнено впорався з цими варіантами, показавши високу чутливість до локальних змін, викликаних мутаціями.

Модель надає кілька метрик для оцінки достовірності: pLDDT показує впевненість у положенні окремих залишків, pTM – цілісність структури в цілому, а матриця PAE допомагає побачити потенційні зони нестабільності. У CFTR було зафіксовано глобальну дестабілізацію після видалення залишку F508. У TP53 порушення торкнулися ДНК-зв’язувального домену. Spike-білок продемонстрував локальні зміни в рецепторній ділянці, а лізоцим – втратив каталітичну функцію, хоча загальна структура залишилася стабільною.

Цей підхід виявився простим, швидким і водночас інформативним. AlphaFold не просто згенерував структури – він дав змогу зрозуміти, як конкретна мутація впливає на білок, і що це може означати для його функції.

4.2 AI-моделювання у персоналізованій медицині

Кожна людина має унікальний набір генетичних варіантів. І саме ці варіанти можуть впливати на те, як працюють її білки, наскільки ефективно організм реагує на ліки, або чому одне захворювання проявляється, а інше ні. У цьому контексті AI-моделювання відкриває зовсім новий рівень точності. AlphaFold працює з амінокислотною послідовністю. Якщо відома мутація – її можна одразу внести у вхідні дані. Через кілька хвилин буде готова модель, яка покаже, що відбувається з білком у конкретному випадку. Такого інструменту ще кілька років тому просто не існувало. Зараз він уже доступний, і це змінює підхід до розуміння хвороб на молекулярному рівні [2].

Уявімо пацієнта з рідкісним варіантом мутації в CFTR. Такий варіант може не входити до стандартних клінічних панелей. Але його можна змоделювати. Якщо видно, що ця мутація порушує структуру критичної ділянки з’являється підстава розглядати її як патогенну. А якщо вона не впливає на структуру це сигнал, що зміна, ймовірно, не шкодить. Таке моделювання допомагає уникати зайвого лікування або навпаки – звернути увагу на проблему раніше. Якщо є кілька можливих ліків, модель може підказати, який із них краще підійде. Усе залежить від того, де структурно порушено білок, і чи здатна конкретна молекула компенсувати цю зміну. Це вже не просто прогноз, а перший крок до індивідуального підбору терапії.

AI-моделі дають змогу не тільки бачити наслідки мутацій, а й порівнювати десятки варіантів одразу. Вони працюють швидко, не потребують дорогих реактивів і можуть бути інтегровані в клінічний аналіз. З часом такі підходи можуть стати частиною рутинної медичної практики – коли під конкретного пацієнта створюється своя «структурна карта ризиків» і рекомендацій. Це ще не майбутнє, але це вже працює. І що важливо – навіть в умовах обмежених ресурсів можна отримати результат, який дає більше, ніж просто текст у звіті.

ВИСНОВКИ

У межах кваліфікаційної роботи проведено повний аналіз моделі AlphaFold, із глибоким зануренням у її архітектуру, структурні компоненти та логіку внутрішньої взаємодії модулів. Послідовно розглянуто весь процес від моменту, коли модель отримує амінокислотну послідовність, і до побудови готової тривимірної структури. Усі елементи працюють узгоджено, як частини єдиного обчислювального механізму.

Головний клас AlphaFold відповідає за запуск моделі. Він координує виклики окремих модулів і передає дані між ними. Основним обчислювальним елементом є блок AlphaFoldIteration. У ньому кілька разів оновлюється інформація між трьома ключовими потоками: вирівнюванням MSA, парними взаємодіями (pair representation) та структурним блоком. Саме тут відбувається глибинне перетворення даних і поступове збирання просторової структури.

Механізми уваги (MSA Attention та Pair Attention) забезпечують обмін інформацією між амінокислотами в різних позиціях та каналах. Блок TriangleAttention вводить геометричну обробку парних взаємодій, дозволяючи моделі враховувати трійкові просторові залежності, що є критичними для точного відображення структури. Компонент OuterProductMean узагальнює сигнали з MSA та трансформує їх у пари залишків, зберігаючи еволюційну інформацію в контексті просторової організації.

На завершальному етапі працюють голови передбачення: Distogram Head, PAE Head та TM-score Head. Вони відповідають за обчислення просторових метрик, відстаней між залишками, очікуваних похибок та глобальної відповідності моделі. Після цього формується остаточна тривимірна структура, яка базується на акумульованій інформації з попередніх модулів. Усі компоненти працюють як єдиний ланцюг, що

забезпечує стабільну якість результату та інтерпретованість внутрішніх метрик.

У практичній частині роботи проаналізовано структурні зміни у чотирьох білках (TP53, CFTR, Spike SARS-CoV-2, лізоцим) – як у дикому типі, так і після мутацій. Моделі для всіх варіантів побудовано за допомогою AlphaFold. Далі оцінено їхню якість за кількома показниками: рівнем впевненості в положенні кожного залишку, точністю просторових взаємодій між амінокислотами, відповідністю всієї структури до реальних білкових форм, а також очікуваною похибкою передбачення між будь-якими двома фрагментами. Аналіз показав, що мутації можуть викликати локальні зміни у критичних ділянках білка, що потенційно впливає на його функції. Просторова візуалізація в середовищах PyMOL та ChimeraX допомогла краще зрозуміти ці зміни в біомедичному контексті.

AlphaFold у рамках цієї роботи розглянуто не лише як інструмент для моделювання, а як складну, глибоко інтегровану архітектуру, здатну адаптуватись до складних задач структурної біоінформатики. Аналіз взаємозв’язків модулів та логіки їхньої побудови відкриває можливості для подальшої кастомізації моделі та інтеграції з іншими системами штучного інтелекту.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Jambaldorj J. Molecular Biology of the Cell. *ResearchGate*. URL: https://www.researchgate.net/publication/375455354_Molecular_Biology_of_the_Cell_Sixth_edition (дата звернення 02.05.2025).

2. Molecular cell biology : Lodish, Harvey F., author : Free Download, Borrow, and Streaming : Internet Archive. *Internet Archive*. URL: https://archive.org/details/molecularcellbio0000lodi (date of access: 02.05.2025).

3. A method and server for predicting damaging missense mutations - Nature Methods. *Nature*. URL: https://doi.org/10.1038/nmeth0410-248 (date of access: 12.05.2025).

4. Berman H. M. The Protein Data Bank. *Nucleic Acids Research*. 2000. Vol. 28, no. 1. P. 235–242. URL: https://doi.org/10.1093/nar/28.1.235 (date of access: 12.05.2025).

5. Callaway E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*. 2020. Vol. 588, no. 7837. P. 203–204. URL: https://doi.org/10.1038/d41586-020-03348-4 (date of access: 12.05.2025).

6. Dill K. A., MacCallum J. L. The protein-folding problem, 50 years on. *Science*. 2012. Vol. 338, no. 6110. P. 1042–1046. URL: https://doi.org/10.1126/science.1219021 (date of access: 12.05.2025).

7. Highly accurate protein structure prediction with AlphaFold / J. Jumper et al. *Nature*. 2021. Vol. 596, no. 7873. P. 583–589. URL: https://doi.org/10.1038/s41586-021-03819-2 (date of access: 12.05.2025).

8. ClinVar: improving access to variant interpretations and supporting evidence / M. J. Landrum et al. *Nucleic acids research*. 2017. Vol. 46, no. D1. P. D1062–D1067. URL: https://doi.org/10.1093/nar/gkx1153 (date of access: 12.05.2025).

9. Leaver-Fay A., Tyka M., Lewis S.M. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods*

in Enzymology. 2011. Vol. 487. P. 545–574. URL: https://doi.org/10.1016/B978-0-12-381270-4.00019-6 (date of access: 12.05.2025).

10. Ng P.C., Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*. 2003. Vol. 31, no. 13. P. 3812–3814. URL: https://doi.org/10.1093/nar/gkg509 (date of access: 12.05.2025).

11. Pettersen E.F., Goddard T.D., Huang C.C. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science*. 2021. Vol. 30, no. 1. P. 70–82. URL: https://doi.org/10.1002/pro.3943 (date of access: 12.05.2025).

12. Schymkowitz J., Borg J., Stricher F. et al. The FoldX web server: an online force field. Nucleic Acids Research. 2005. Vol. 33. *Web Server issue*. P. W382–W388. URL: https://doi.org/10.1093/nar/gki387 (date of access: 12.05.2025).

13. Senior A.W., Evans R., Jumper J. et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020. Vol. 577. P. 706–710. URL: https://doi.org/10.1038/s41586-019-1923-7 (date of access: 12.05.2025).

14. Tate J.G., Bamford S., Jubb H.C. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research.* 2019. Vol. 47, no. D1. P. D941–D947. URL: https://doi.org/10.1093/nar/gky1015 (date of access: 12.05.2025).

15. Tunyasuvunakool K., Adler J., Wu C. et al. Highly accurate protein structure prediction for the human proteome. *Nature.* 2021. Vol. 596. P. 590–596. URL: https://doi.org/10.1038/s41586-021-03828-1 (date of access:12.05.2025).

16. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research.* 2023. Vol. 51, no. D1. P. D523–D531. URL: https://doi.org/10.1093/nar/gkac1052 (date of access: 12.05.2025).

17. Varadi M., Anyango S., Deshpande M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research.* 2022. Vol.

50, no. D1. P. D439–D444. URL: https://doi.org/10.1093/nar/gkab1061 (date of access: 12.05.2025).

18. AlphaFold Protein Structure Database. URL: https://aphafold.ebi.ac.uk/ (date of access:07.05.2025).

19. ClinVar Database. URL: https://www.ncbi.nlm.nih.gov/clinvar/ (date of access:12.05.2025).

20. COSMIC Database. URL: https://cancer.sanger.ac.uk/cosmic (date of access: 12.05.2025).

21. FoldX. URL: http://foldxsuite.crg.eu/ (date of access: 12.05.2025).

22. PolyPhen-2. URL: http://genetics.bwh.harvard.edu/pph2/ (date of access: 12.05.2025).

23. Protein Data Bank (PDB). URL: https://www.rcsb.org/ (date of access: 12.05.2025).

24. Rosetta. URL: https://www.rosettacommons.org/ (date of access: 12.05.2025).

25. SIFT. URL: https://sift.bii.a-star.edu.sg/ (date of access: 12.05.2025).

26. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 2.0. URL: https://pymol.org/2/ (date of access: 10.05.2025).

27. UniProt Knowledgebase. URL: https://www.uniprot.org/ (date of access: 12.05.2025).

28. UCSF ChimeraX. URL: https://www.cgl.ucsf.edu/chimerax/ (date of access: 12.05.2025).