

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ВИСОКИХ ТЕХНОЛОГІЙ

Завідувач кафедри супрамолекулярної хімії

Проф. Сергій Вікторович Рябухін

Протокол № _____ засідання кафедри

від “ _____ ” _____ 20 _____ р.

**РОЗРОБКА НОВИХ ІЗОФОРМОСЕЛЕКТИВНИХ ІНГІБІТОРІВ
ХЛОРИДНИХ КАНАЛІВ CLC-КА ТА CLC-КВ.**

Випускна кваліфікаційна робота магістра
студента спеціальності

102 Хімія

Уяздовського Мирослава Ігоровича

Науковий керівник від установи

Завідувач кафедри молекулярної біотехнології та біоінформатики

к.б.н Нипорко Олексій Юрійович

Науковий керівник

Технічний директор

ТОВ «Рецептор»

к.б.н, Старосила Сергій Анатолійович

Оцінка захисту роботи

Київ 2022 р.

АНОТАЦІЯ

Уяздовський М.І. Розробка нових ізоформоселективних інгібіторів хлоридних каналів СІС-Ка та СІС-Кв. – Випускна кваліфікаційна робота магістра за спеціальністю 121 Хімія ОП «Хемоінформатика».

Робота присвячена розробці низькомолекулярних інгібіторів СІС-Ка та СІС-Кв. Для цього було зроблено універсальну модель машинного навчання для передбачення активності малих молекул відносно білків-мішеней. Отриману модель використали для віртуального скринінгу щодо двох ізоформ іонного каналу СІС-К.

В результаті, було отримано біля $17 \cdot 10^3$ сполук для подальшого докінгу, з яких 38 сполук були відібрані як потенційні ізоформоселективні інгібітори іонних каналів СІС-Ка та СІС-Кв.

Отримані сполуки можуть бути використані для подальшої розробки антигіпертензивних лікарських засобів; розроблена модель машинного навчання може бути використана для скринінгу сполук відносно інших білків-мішеней.

Ключові слова: Машинне навчання; графові нейронні мережі;

докінг; ізоформоселективність; СІС-К.

ANNOTATION

Uiazdovskyi M.I. Development of novel isoform-selective inhibitors of chloride channels ClC-Ka and ClC-Kb. - Graduation thesis of the master's degree in the expertise 121 Chemistry, EP "Chemoinformatics".

The paper is devoted to developing low molecular weight inhibitors of ClC-Ka and ClC-Kb. A universal model of machine learning was made to predict the activity of small molecules against target proteins to do this. The obtained model was used for virtual screening against two isoforms of the ClC-K ion channel.

As a result, about $17 * 10^3$ compounds were obtained for further docking, of which 38 compounds were selected as potential isoform-selective inhibitors of ClC-Ka and ClC-Kb ion channels.

The obtained compounds can be used for further developing antihypertensive drugs; the developed machine learning model can screen compounds against other target proteins.

Keywords: Machine learning; graph neural networks; docking; isoform-selectivity; ClC-K.

2.1.9	Тюнінг	моделі	25
DTI.....			
2.2	Дані для обробки моделлю DTI.....		25
2.3	Гомологічне	моделювання	структури
	Ca.....	СІС-	26
2.4	Проведення молекулярної динаміки білків СІС-Ка та СІС-Кб для отримання конформацій білків-мішеней для молекулярного докінгу.....		27
2.5	Молекулярний докінг відібраних за допомогою DTI сполук в СІС-Ка та СІС-Кб для чисельного визначення їх афінності та передбачення селективності.....		27
3.	РЕЗУЛЬТАТИ	ТА	30
ОБГОВОРЕННЯ.....			
ВИСНОВКИ.....			44
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....			45

ВСТУП

Для деяких захворювань, пов'язаних з нирками, два гомологи CLC-K, CLC-Ka і CLC-Kb, є особливо переконливими мішенями через їх важливу роль у підтримці електролітного балансу. У людей мутації втрати функції CLC-Kb призводять до синдрому Барттера III типу, розладу супроводжуваного втратою солі, що характеризується низьким кров'яним тиском (1) Тому CLC-Kb є потенційною молекулярною мішенню для антигіпертензивних терапевтичних засобів.

Селективність низькомолекулярних інгібіторів до CLC-Ka та CLC-Kb є критичною оскільки два вищезгадані канали також експресуються у внутрішньому вусі, де вони відіграють важливу роль, хоча і дублюють одне одного: порушення роботи обох гомологів призводить до глухоти, а окреме порушення роботи одного з них не впливає на слух (2). Розробка селективних малих молекул, які специфічно націлені на CLC-Ka або CLC-Kb, становить серйозну проблему в дизайні ліганду через >90% ідентичності послідовності між двома гомологами.

Методи *in silico* показали свою ефективність як в скринінгу ізоформоселективних інгібіторів CLC-Ka та CLC-Kb (52), так і в гомологічному моделюванні структури білка-мішені (34). Але попередні методи віртуального скринінгу стикнулися з обмеженнями пропускну здатності, нівелювати які повинні моделі машинного навчання.

Метою роботи було передбачення потенційно ізоформоселективних інгібіторів іонних каналів CLC-Ka та CLC-Kb.

Відповідно до мети роботи були поставлені такі завдання:

1. Навчити графову нейронну мережу для проведення первинного віртуального скринінгу із передбаченням активності сполук відносно різних молекулярних мішеней.

2. Провести первинний віртуальний скринінг бібліотеки хімічних сполук за із застосуванням графової нейронної мережі для ідентифікації потенційно активних сполук щодо іонних каналів CLC-Ка та CLC-Кб. Зробити аналіз результатів віртуального скринінгу.
3. Провести гомологічне моделювання структури CLC-Ка.
4. Провести моделювання молекулярної динаміки ізоформ хлоридних каналів CLC-Ка та CLC-Кб з метою отримання серії конформацій білків для молекулярного докінгу.
5. Провести молекулярний докінг кращих сполук, отриманих в результаті скринінгу графовою моделлю, та вибрати серед них селективні та афінні щодо CLC-Ка та CLC-Кб.

РОЗДІЛ 1. ЗАГАЛЬНА ХАРАКТЕРИСТИКА ХЛОРИДНОГО КАНАЛУ ClC-K

1.1 Історія відкриття, функції та пов'язані патології ClC-K

Велике сімейство хлоридних каналів ClC і хлорид/протонних обмінників було відкрито Йенчем і його співробітниками, які клонували хлоридний канал електричного органу електричного скату (названого ClC-0) (3).

Деякі члени цієї родини відіграють важливу роль у стані здоров'я та хворобливих станах, таких як міопатія, остеопетроз, хвороба Дента та синдром Бартера (БС) (4,5,6). У нирках два канали ClC, пов'язані з регуляторною субодиницею Barttin, відіграють ключову роль у всмоктуванні NaCl, таким чином беручи участь у контролі артеріального тиску (7). Зокрема, хлоридний канал ClC-Kb, і меншою мірою канал ClC-Ka, керують базолатеріальним етапом всмоктування хлорид йонів в товстій частині висхідного коліна, дистальному звивистому каналці та вставлених клітинах з'єднувального каналця, що збирає протоки (7). Вони також необхідні для регулювання концентрації сечі (5,8). БС — це аутомно-рецесивне захворювання, що характеризується втратою NaCl нирками, гіпокаліємічним метаболічним алкалозом і вторинним гіперальдостеронізмом із нормальним або низьким кров'яним тиском (7).

Серед чотирьох типів генетичних варіантів БС три пов'язані з дефектами ниркових хлоридних каналів:

- БС типу 3 пов'язаний з втратою функції ClC-Kb (9, 10);
- тип 4a пов'язаний з мутаціями в гені BSND, що кодує білок Barttin (11);
- тип 4b - це дигенне захворювання через мутації втрати функції двох генів, що кодують ниркові хлоридні канали, CLCNKA і CLCNKB (12).

Мишачі моделі, позбавлені ClC-K2 (ортолог ClC-Kb) або ClC-K1 (ортолог ClC-Ka), демонструють контрастні фенотипи. Тварини ClC-K2 knock-out (КО) демонструють ознаки людського синдрому Бартера (13), тоді як ClC-K2-K1 КО тварини демонструють нецукровий діабет (8, 14). Тим не менш, поліморфізм ClC-Ka , а також ClC-Kb асоціюється зі схильністю до гіпертонії (15, 16).

Гіпертонія, одне з найпоширеніших захворювань людини, є основним фактором ризику серцево-судинних захворювань та ниркової недостатності (17). Генетичні дослідження (18) підтвердили новаторську гіпотезу про те, що гіпертонія обов'язково включає надмірне надходження солі (NaCl) до нирок (19).

Комбінована терапія у пацієнтів з артеріальною гіпертензією включає діуретики, спрямовані на товсту частину висхідного коліна та дистальний звивистий каналець. Проте існує загальна думка, що було б доцільно диверсифікувати доступні ліки як засіб проти резистентної гіпертензії. У цьому контексті ClC-Ka і, зокрема, ClC-Kb , який присутній уздовж усього дистального нефрона, де відбувається тонко налаштована абсорбція NaCl , є цінними мішенями для розробки нових ліків у галузі гіпертонії (20, 21).

Кілька ключових досліджень в цьому напрямку були проведені для того, щоб ідентифікувати інгібітори для цих двох каналів і окреслити можливу область зв'язування лікарського засобу на ClC-Ka . Проте в даний час фармакологічні відомості стосовно цих каналів залишається неповними (22- 27).

1.2 Структура ClC-K .

Як і інші білки ClC , хлоридний канал ClC-Kb є трансмембранним білком, активним як гомодимер (28). Кожен протомер довжиною 687 амінокислот складається з ТМ домену, який містить 17 α -спіралей, і двох внутрішньоклітинних доменів цистатіонін-бета-синтази (CBS).

Незважаючи на його високу біологічну важливість, структура і функції ClC-Kb досить довго залишалися не вивченими. Тим не менш, існує кілька гомологів ClC-Kb з відомими тривимірними структурами (або їх частиною), які можна використовувати, щоб отримати уявлення про структуру-функцію.

Повідомлялося про експериментальні структури еукаріотичного транспортера CLC SmCLC (29), бактеріального транспортера ClC *Escherichia coli* EcClC (30) та цитоплазматичного домену хлоридного каналу ClC-Ka (31). Gradogna та ін. (32) побудували структурну модель для ClC-Ka, використовуючи як шаблон бактеріальний транспортер EcClC, який має меншу ідентичність послідовності з ClC-Kb (28% для 17 ТМ α -спіралей), ніж канал SmCLC (33% для 17 ТМ α -спіралей). Однак, хоча транспортер EcClC можна вважати цікавим шаблоном для дослідження трансмембранних областей каналів ClC-Kb або ClC-Ka, йому бракує внутрішньоклітинного цитоплазматичного домену. Важливо, що структура ClC-K великої рогатої худоби була нещодавно визначена за допомогою крио-електронної мікроскопії (33), і вона остаточно допомогла отримати нові знання про цей рецептор.

В результаті одному із досліджень (34) була остаточно розроблена структурна модель каналу ClC-Kb людини, із використанням всіх доступних відповідних структурних 3D-шаблонів. Початкові структурні моделі потім уточнювали за допомогою мінімізації енергії та інших інструментів моделювання з білком, вбудованим у модель клітинної мембрани.

На рисунку 1.1 зображений розподіл електростатичних потенціалів по трансмембранній (ТМ) частині ClC-Kb людини.

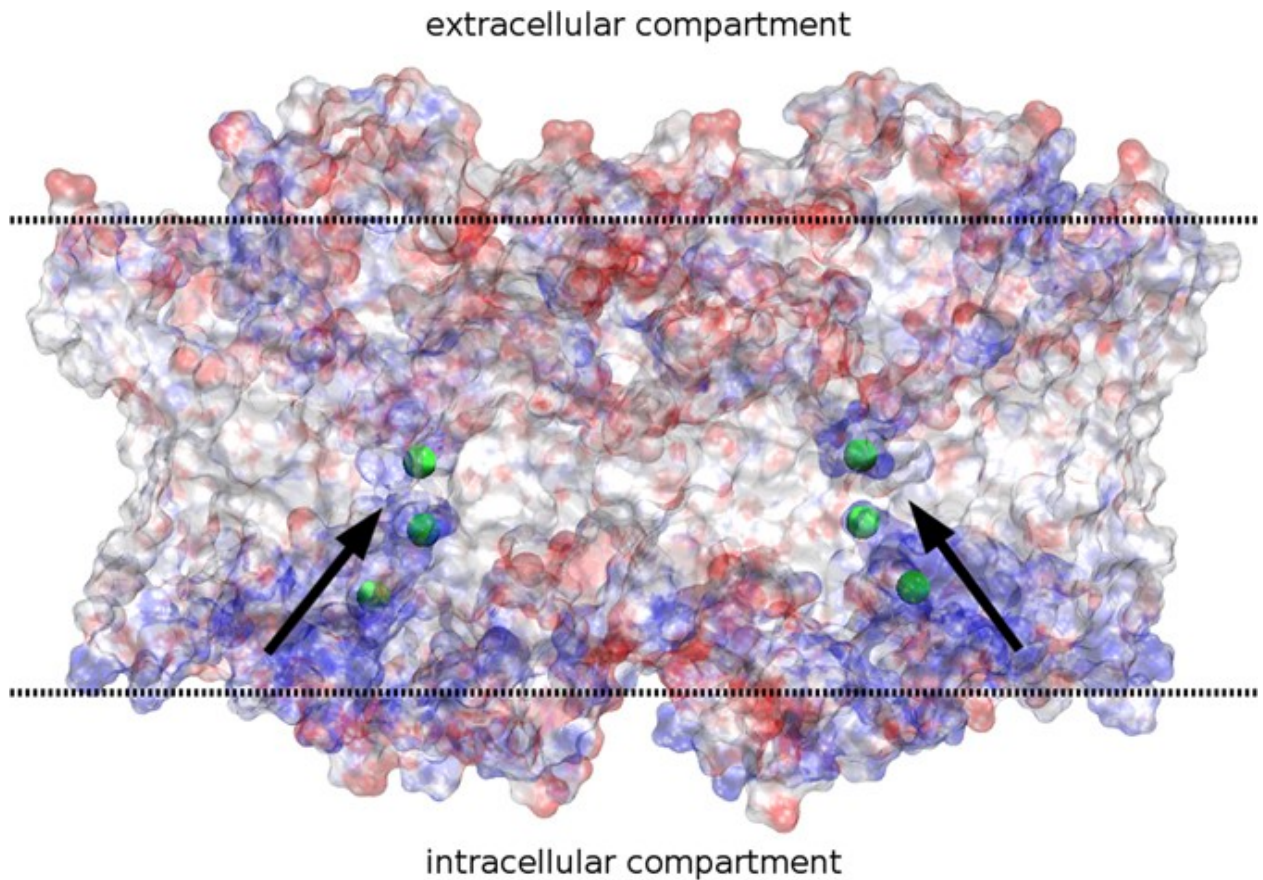


Рисунок 1.1 Електростатичні потенціали ТМ частини каналу СІС-К ν людини (34). Розріз каналу через поздовжню вісь мембрани зображений пунктирними лініями. Електростатичний потенціал, розрахований для всієї системи (білка та мембрани), відображається на поверхні білка від -5 (червоний) до +5 (синій) ккал*моль⁻¹. Стрілки показують шлях іонів від внутрішньоклітинного компартменту до іонних воріт. Для відображення внутрішньої частини каналу використовується прозоре зображення поверхні.

Перевірили передбачену структуру за допомогою кількох протоколів структурного аналізу, а потім зіставили та проаналізували в 3D кілька відомих і

нових амінокислотних замін. Такий аналіз, коли амінокислотна варіація ідентифікована у пацієнтів, називається біоструктурною патологією (35, 36).

Дана структура у подальшому була використана для пошуку ізоформоселективних інгібіторів білків ClC-Ка та ClC-Кб.

На рисунку 1.2 зображений шлях хлориду в каналі ClC-Кб.

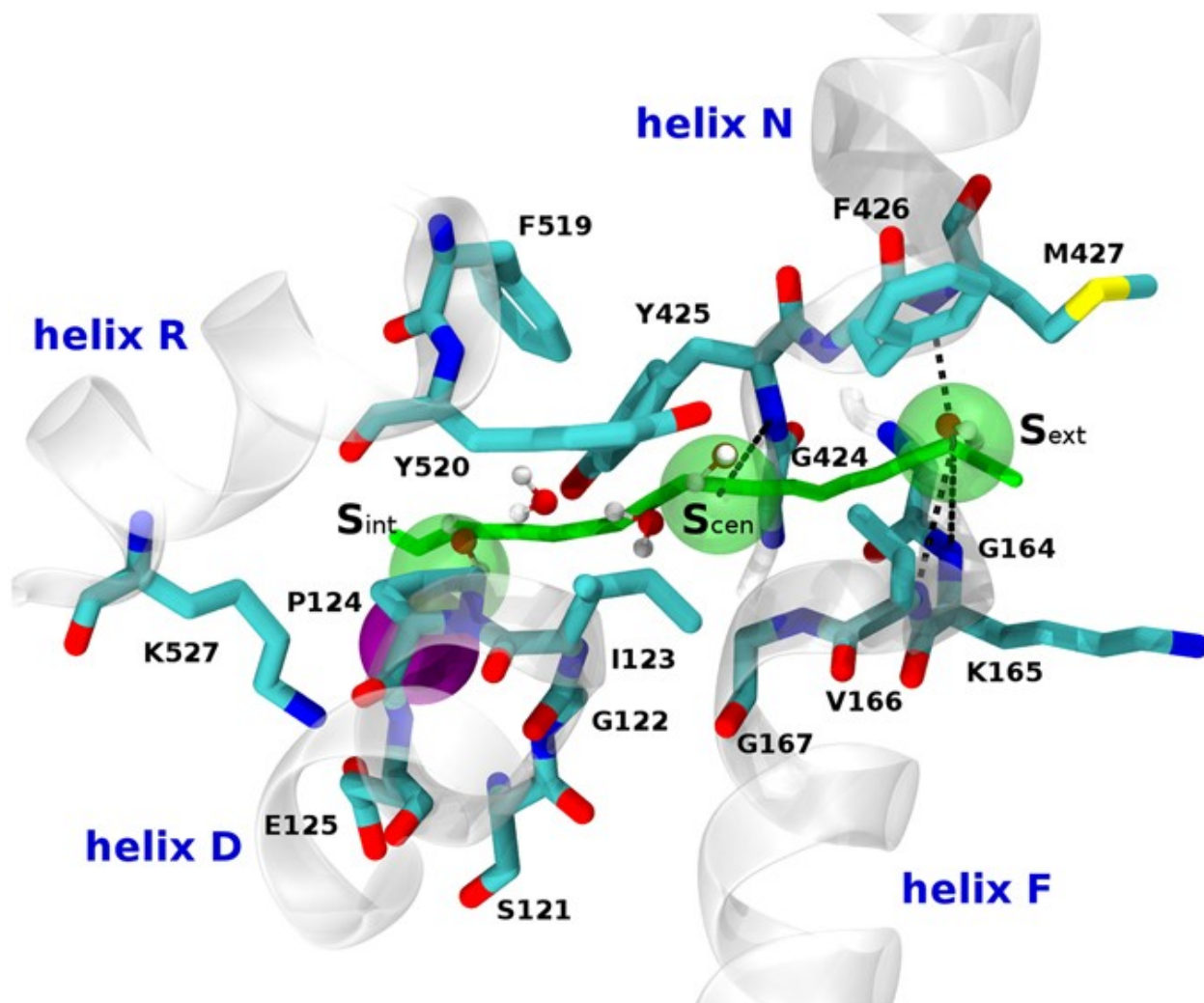


Рисунок 1.2 Шлях хлориду в каналі ClC-Kb (34). Показані три експериментально відомі положення хлорид-іона (Sext, Scen і Sint) у CmCLC та EcClC (зелені сфери). Прогнозований шлях хлорид-іона, отриманий в результаті моделювання MD, показаний зеленою паличкою. Важливі спіралі зображені у вигляді ескізу. Залишки, що утворюють пору каналу, представлені у вигляді паличок, а чорні пунктирні лінії представляють імовірні електростатичні взаємодії між трьома можливими положеннями хлорид-іонів і NH-групою білкового скаффолду. Хлорид-іон, який взаємодіє з порою під час моделювання MD, представлений у вигляді фіолетової сфери. Молекули води, присутні в моделюванні, представлені у вигляді кульки та палиці.

1.3 Блокатори CLC-K

Малі молекули, які зв'язують і модулюють специфічні гомологи CLC, можуть служити цінними інструментами для зв'язування складної поведінки воріт та різноманітних фізіологічних функцій у сімействі CLC. Низькомолекулярні модулятори CLC також є перспективними як терапевтичні агенти (37, 38).

Для деяких захворювань, пов'язаних з нирками, два гомологи CLC, CLC-Ka і CLC-Kb, є особливо переконливими мішенями через їх важливу роль у підтримці електролітного балансу. У людей мутації втрати функції CLC-Kb призводять до синдрому Барттера III типу, розладу супроводжуваного втратою солі, що характеризується низьким кров'яним тиском (1). Даний фенотип припускає CLC-Kb як потенційну мішень для антигіпертензивних терапевтичних засобів.

Аналогічно, моделі на КО мишах показують, що транспорт CLC-Ka Cl⁻ необхідний для концентрації сечі (39, 40) і таким чином припускають, що інгібітори можна використовувати для лікування гіпонатріємії (41), стану, при якому концентрація натрію в крові низька. Гіпонатріємія розвивається, коли нирки не можуть ефективно виводити воду, і може виникнути у пацієнтів із серцевою недостатністю, цирозом, нирковою недостатністю або іншими станами (41, 42). Лікування гіпонатріємії може бути складним, і без втручання може виникнути стійке неврологічне ураження (42, 43).

Для деяких захворювань, пов'язаних з нирками, два гомологи CLC-K, CLC-Ka і CLC-Kb, є особливо переконливими мішенями через їх важливу роль у підтримці електролітного балансу. У людей мутації втрати функції CLC-Kb призводять до синдрому Барттера III типу, розладу супроводжуваного втратою солі, що характеризується низьким кров'яним тиском (1). Даний фенотип припускає CLC-Kb як потенційну мішень для антигіпертензивних терапевтичних засобів.

Попередні повідомлення про інгібітори CLC в основному описують низькоафінні (мікромолярні та мілімолярні IC_{50}) сполуки, які не мають селективності. Найпотужнішим відомим інгібітором CLC є пептидний токсин GaTx2, який селективно інгібує CLC-2 з $IC_{50} \sim 20$ пМ. Однак, незважаючи на його дію, інгібування GaTx2 насичується на $\sim 50\%$ (44), що обмежує його використання в якості фармакологічного зонда. Повідомляється, що два класи низькомолекулярних інгібіторів демонструють помірну дію та селективність щодо CLC-Ка порівняно з CLC-Кб.

Перший, набір стилбенових дисульфонатів, є змішаним класом інгібіторів транспорту аніонів (45, 46, 47). Один з них, 4,4'-диізотіоціано-2,2'-стилбендисульфонова кислота (DIDS), був використаний як хімічний інструмент для підтвердження димерної архітектури каналу Torpedo CLC-0 (48). Згодом лабораторні випробування показали, що DIDS гідролізується в буферному розчині з утворенням олігомерів політіосечовини, які є більш потужними антагоністами CLC-0, CLC-Ка і CLC-ec1, ніж сам DIDS (49).

Пентамер DIDS є найпотужнішим і селективним відомим непептидним інгібітором CLC з IC_{50} проти CLC-Ка 0,5 мкМ і приблизно в 100 разів збільшеним IC_{50} проти CLCКб (49). Однак дослідження зв'язку структура-активність (SAR) та подальша розробка інгібіторів на основі DIDS серйозно ускладнені через низьку розчинність у воді даних олігомерних молекул, а також синтетичні проблеми систематичної модифікації скафолду DIDS.

Другий клас інгібіторів CLC базований на арилзаміщеній бензофуранкарбоновій кислоті (50, 51). Найкращі з цих сполук, MT-189 і SRA36, є приблизно в 10 разів менш активними, ніж пентамер DIDS, але краще піддаються синтетичним маніпуляціям. У той час як SRA36 інгібує як CLC-Ка, так і CLC-Кб з однаковою ефектністю, MT189 має майже втричі більшу афінність до CLC-Ка порівняно з CLC-Кб ($IC_{50} = 7$ проти 20 мкМ) (50).

Була описана розробка, синтез та оцінка бібліотеки N-арильованих похідних бензimidазолу (BIM) (52), одна з яких (BIM1) демонструвала >20-кратну селективність щодо CLKa у порівнянні CLC-Kb, тобто двох найбільш близьких споріднених гомологів CLC людини. Але не зважаючи на відносно високу селективність дана сполука має мікромольну активність, що робить її не дуже придатною для потенційного терапевтичного застосування.

РОЗДІЛ 2. МАТЕРІАЛИ ТА МЕТОДИ ДОСЛІДЖЕНЬ

Задача, яку необхідно було вирішити полягала у ресурсно-ефективному первинному віртуальному скринінгу бібліотеки хімічних сполук з метою виявити потенційно активні виключно до однієї з ізоформ білку CIC-K (тільки а або тільки б). Виключна селективність лише до однієї з ізоформ білку CIC-K є необхідною задля уникнення побічного ефекту (глухоти), що виникає в наслідок неселективного інгібування обох ізоформ білку (2). Дана задача була вирішена за рахунок моделі машинного навчання DTI, описаної нижче. На жаль, подібні моделі машинного навчання не можуть остаточно вирішити задачу селективності (53), тому потенційна селективність отриманих сполук була з'ясована за допомогою докінгу.



Загальний хід експерименту описаний нижче на рисунку 2.1.

Рисунок 2.1 Загальний хід експерименту.

2.1. Розробка моделі машинного навчання DTI для передбачення активності малих молекул щодо різних молекулярних мішеней

2.1.1 Підготовка набору даних для тренування моделі машинного навчання DTI

Загальновідомо, що моделі машинного навчання потребують вхідних даних (зразків) для тренування із подальшим набуттям можливості видачі релевантних результатів (54). Крім того, модель машинного навчання потребує як позитивних, так і негативних зразків для успішного вирішення задачі класифікації(55).

Для створення тренувального набору даних були використані дані з наступних джерел:

- Бази ChEMBL (<https://www.ebi.ac.uk/chembl/>);
- Бази STITCH (Search Tool for InTeractions of Chemicals, <http://stitch.embl.de/>);
- Набору штучно згенерованих негативних зразків.

2.1.2 Підготовка набору даних із використанням бази даних ChEMBL

Задля отримання набору даних, пари білок-ліганд були вивантажені із використанням SQL скрипту та вбудованого API (Application Programming Interface) бази даних ChEMBL.

```
SELECT  
  
    assays.assay_type  
  
    , assays.description  
  
    , activities.assay_id  
  
    , activities.standard_relation  
  
    , activities.standard_value  
  
    , activities.standard_units  
  
    , activities.standard_type
```

```
, target.chembl_id as target_chembl_id  
  
, compound.canonical_smiles  
  
FROM assays  
  
INNER JOIN activities  
  
    ON assays.assay_id = activities.assay_id  
  
    AND standard_units IN ('nM', 'ug.mL-1')  
  
    AND standard_value IS NOT NULL  
  
    AND standard_type IN ('IC50', 'Ki', 'Kd', 'EC50', 'AC50', 'Activity')  
  
INNER JOIN target_dictionary AS target  
  
    ON target.tid = assays.tid  
  
    AND tax_id = 9606  
  
    AND target_type = 'SINGLE PROTEIN'  
  
INNER JOIN compound_structures AS compound  
  
    ON compound.molregno = activities.molregno  
  
INNER JOIN molecule_dictionary as molecule  
  
    ON molecule.molregno = activities.molregno  
  
    AND molecule_type = 'Small molecule'
```

2.1.3 Підготовка набору даних із використанням бази даних STITCH

Дані зі STITCH вивантажувалися напряму, оскільки сервіс пропонує до завантаження вже готові пакети пар білок-ліганд.

Важливо зауважити, що характеристика активності комплексу білок-ліганд у даній базі даних не є настільки явною, як у випадку із ChEMBL, оскільки задається не напряму через параметри біологічної активності, а через так званий «score» (рахунок), котрий може знаходитись у межах від 0,000 до 1,000. В свою чергу, низькі значення рахунку в даній базі даних означають не відсутність активності, а відсутність даних про неї (56).

В результаті, критерієм відбору позитивних прикладів пари білок-ліганд слугував рахунок, що перевищував або дорівнював 0,400.

2.1.4 Підготовка набору негативних зразків із використанням генерації на основі відбитків за Морганом

Можна помітити, що сумарна кількість позитивних зразків становила 780141, в той час як негативних нараховувалося 449034. Такий набір даних не підходить для моделі машинного навчання, що займається класифікацією великого масиву даних, в якому переважають негативні зразки (55). Тому постала необхідність отримати набір негативних зразків, котрий доведе відношення позитивних зразків до негативних зразків до 1:9 відповідно.

Ця задача виявилася принципово вирішуваною, оскільки виходячи із структурних властивостей білків, стійка взаємодія білка із випадковою молекулою у переважній більшості випадків буде неможливою (57).

В результаті було реалізовано такий підхід до генерації негативних зразків. Для білка обирали всі пари білок-ліганд із позитивних зразків, знімали з них відбитки за Морганом (Morgan fingerprinting) (58) та проводили скринінг пулу лігандів отриманого шляхом вилучення структур лігандів з пар білок-ліганд отриманих із баз даних ChEMBL та STITCH.

Для кожної отриманої сполуки оцінювали подібність структури (58) у порівнянні з позитивними зразками, і якщо подібність становила менше 0,3, то сполука обиралася як негативний зразок.

2.1.5 Опис репрезентації ліганду у моделі

Для репрезентації ліганду у DTI було одночасно використано два підходи :

- Модифіковані молекулярні графи, що використовувалися у підході для передбачення властивостей біологічно активних молекул Chemi-net (59);
- Стандартні відбитки за Морганом (60).

Для генерації обох репрезентацій лігандів на основі ізоморфних SMILES використовувався Python API хемоіноформатичного пакету RDKit v. 2021.03, який має відкритий код.

2.1.6 Репрезентація білку у моделі

Білок репрезентували у вигляді графу білка на рівні залишків на основі тривимірних білкових структур, створених за допомогою AlphaFold (62) у форматі PDB (63).

Приблизно 50% білків у наборі даних мають відомі тривимірні структури, депоновані в Protein Data Bank. Але AlphaFold (<https://alphafold.ebi.ac.uk/>) використовували для прогнозування структури всіх білків, задля уніфікації підходу та уникання неавтоматизованої обробки експериментально визначених структур. Це пояснюється тим, що структури білків є часто не повними, дещо помилковими, також для них типовим є включення додаткових лігандів у кристал, включно із тими, які є не характерними даному білку.

Загалом, тривимірні структури були створені вручну з використанням прискореної реалізації алгоритму AlphaFold в Google Colaboratory: ColabFold (64).

Відфільтровані тривимірні структури перетворювали в граfi рівня залишків за допомогою Biopython v. 1.79 (65) та Pteros(66, 67). На базі Open Drug Discovery Toolkit (68) розробили підхід для кодування властивостей білка в характеристиках граней графа. Грань створювали, якщо дві амінокислоти утворювали нековалентний зв'язок у межах певної відстані.

Особливості грані визначали тип даного нековалентного зв'язку (табл. 3).

В таблиці 2.1 наведені типи нековалентних зв'язків і відповідні границі відстаней, які використовували для генерації графу та створення граней.

Таблиця 2.1 типи нековалентних зв'язків і границі відстаней для генерації графу та створення граней

Тип зв'язку	Границі відстаней (Å)
Гідрофобний контакт	4
Водневий зв'язок	3.5
Сольовий місток	4
Катіон-пі взаємодія	5
Перпендикулярний пі-стекінг	5

Паралельний пі-стекінг	5
------------------------	---

2.1.7 Архітектура моделі DTI

Використовували ГНМ (графові нейронні мережі) для вилучення ознак із графів ліганду та білка, а потім повністю з'єднаних (ПЗ) шарів нейронної мережі.

Робили тюнінг різних гіперпараметрів.

Налаштовували DTI на один тип ГНМ або комбінацію кількох типів ГНМ, що забезпечує найкращі результати перехресної перевірки набору даних контрольного тесту. Були розглянуті такі типи ГНМ : Graph Attention Network (GATv1) (74), Graph Attention Network, яка виправляє проблему статичної уваги (GATv2) (75), присутню в GATv1, Graph Convolutional Network (GCN) (76), Graph Isomorphism Network (GIN) (77), Graph Isomorphism Network with incorporated edge features (GINE) (78), ГНМ для передачі нейронних відомостей для квантової хімії (GQC) (79) і GCN для вивчення молекулярних відбитків (GMF) (80).

Використовували лише відбитки лігандів за Морганом, лише графи лігандів або обоє одночано, підключали ПЗ шарів після виходу об'єднання ГНМ та використовували різні типи комбінації графів.

Модель DTI створювали з допомогою PyTorch (81) з відкритим вихідним кодом для машинного навчання, а ГНМ , створювали за допомогою PyTorch Geometric (82).

2.1.8 Метрики оцінки моделі DTI

Обирали 3 метрики оцінки, які використовуються більшістю авторів базових підходів.

Коефіцієнт кореляції Метьюса (ϕ) (83):

Модель класифікації прогнозує клас кожного екземпляра даних, приписуючи кожному зразку його передбачувану мітку (позитивну чи негативну): таким чином, наприкінці процедури класифікації кожен зразок потрапляє в один із наступних чотирьох випадків:

- Фактичні позитиви, які є правильно передбаченими позитивами, називаються справжніми позитивами (TP);
- Фактичні позитиви, які є помилково прогнозованими негативами, називаються хибнегативними (FN);
- Фактичні негативи, які є правильно прогнозованими негативами, називаються справжніми негативами (TN);
- Фактичні негативи, які є помилково передбаченими позитивними результатами, називаються хибнопозитивними (FP).

Цей розподіл може бути представлений у вигляді таблиці 2×2 , яка називається *матрицею плутанності* $M = \begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$, яка повністю описує результат задачі класифікації.

Одним із статистичних показників для опису якості класифікації моделей є MCC. Він лежить у межах $[-1,1]$, де -1 – це повністю протилежний до бажаного результат, а 1 – ідеальне передбачення. Цей показник обраховується за формулою:

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (1)$$

Також була використана метрика площа під кривою ROC (receiver operating characteristic) – AUC(Area Under Curve) (84):

$$A = \int_{-\infty}^{\infty} y(T)x'(T)dT \quad (2)$$

де x' - похідна від частоти хибно-позитивних результатів, y - частота істинно-позитивних результатів, T – граничний параметр.

Також було розраховано Precision-Recall AUC (85):

Де precision дозволяє встановити частку істинно правильних результатів серед усіх, які були ідентифіковані як правильні у межах [0,1]:

$$precision = \frac{TP}{TP+FP} \quad (3)$$

Де accuracy є статистичним показником, що вказує на частку правильно спрогнозованих результатів серед усіх спрогнозованих результатів у межах [0,1]:

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

2.1.9 Тюнінг моделі DTI

Налаштування проводили з використанням платформи оптимізації гіперпараметрів Optuna (87). Для навчання моделі використовували пакетний розмір 32, оптимізатор Адама з швидкістю навчання 0,0001 і функцію середніх квадратичних втрат помилок. Кількість епох становила 30.

2.2 Дані для обробки моделлю DTI

Пошук потенційних інгібіторів проводили для ізоформ а та b хлоридного каналу ClC-K, що були представлені у вигляді графу за методом, що наведений вище. Не проводили декомпозицію структур на окремі домени.

Структури обох ізоформ ClC-K були отримані за допомогою AlphaFold у Google Colaboratory: ColabFold (64).

Сполуки із реальної синтезованої бази Chemspace (<https://chem-space.com/>), котра містить 3 174 252 сполук, переводили у графовий вигляд. Запускали віртуальний скринінг утвореного хімічного простору проти ізоформ ClC-K із використанням моделі DTI.

Модель DTI, що використовували, здатна працювати в трьох режимах:

1. Передбачення імовірності зв'язування;
2. Сортування бази даних білків за імовірністю зв'язування зі сполукою (база нараховує близько 9000 білків людини);
3. Поліфармакологія, тобто модель дає для кожної сполуки список білків, для яких вірогідність зв'язування зі сполукою перевищує 95%.

Для подальшого молекулярного докінгу обирали 0,5% сполук із результатів скринінгу, для яких різниці вірогідностей зв'язування із ізоформами ClC-K були максимальними.

Наприклад, якщо сполука зв'язувалася із ClC-Ka з імовірністю 98%, а з ClC-Kb із імовірністю 3%, то різниця імовірностей становила 95%. Якщо ця різниця потрапляла в топ 0,5%, то сполука вважалася потенційним селективним інгібітором хлоридного каналу ClC-Ka або ClC-Kb.

2.3 Гомологічне моделювання структури ClC-Ka

Гомологічне моделювання проводили за допомогою програми Modeller <https://bio.tools/modeller> з налаштуваннями за замовченням.

Первинну структуру білка брали із бази даних Uniprot за ідентифікатором A0A131XDW4 (<https://www.uniprot.org/uniprot/A0A131XDW4>).

У якості шаблону використовували структуру бичачого СІС-К отриману за допомогою кріоелектронної мікроскопії (33) та вивантажену із структурної бази PDB.

Структура мала ідентифікатор 5TQQ (<https://www.rcsb.org/structure/5TQQ>).

2.4 Проведення молекулярної динаміки білків СІС-Ка та СІС-Кв для отримання конформацій білків-мішеней для молекулярного докінгу

Для кожної ізоформи білку з метою подальшого отримання різних поз брали кожну 10-ту із 100 (бралися рівномірно по всій траєкторії) конформацій з молекулярної динаміки. Молекулярну динаміку проводили у пакеті GROMACS 2021.4 (<https://www.gromacs.org/>) із використанням силового поля CHARMM (<https://manual.gromacs.org/current/user-guide/force-fields.html>).

2.5 Молекулярний докінг відібраних за допомогою ДТІ сполук в СІС-Ка та СІС-Кв для чисельного визначення їх афінності та передбачення селективності

Проводили автоматизований докінг у програмі AutoDock 1.56 (<https://autodock.scripps.edu/>).

Налаштування докінгу виглядали так:

```
npts 98 122 108          # num.grid points in xyz
gridfld CLCKa_60_A.maps.fld    # grid_data_file
spacing 0.173             # spacing(A)
```

```

receptor_types A C HD N NA OA SA # receptor atom types
ligand_types A C HD N NA OA SA # ligand atom types
receptor CLCKa_60_A.pdbqt # macromolecule
gridcenter 121.66 78.938 82.751 # xyz-coordinates or auto
smooth 0.5 # store minimum energy w/in rad(A)
map CLCKa_60_A.A.map # atom-specific affinity map
map CLCKa_60_A.C.map # atom-specific affinity map
map CLCKa_60_A.HD.map # atom-specific affinity map
map CLCKa_60_A.N.map # atom-specific affinity map
map CLCKa_60_A.NA.map # atom-specific affinity map
map CLCKa_60_A.OA.map # atom-specific affinity map
map CLCKa_60_A.SA.map # atom-specific affinity map
elecmap CLCKa_60_A.e.map # electrostatic potential map
dsolvmap CLCKa_60_A.d.map # desolvation potential map
dielectric -0.1465 # <0, AD4 distance-dep.diel;>0, constant

```

На білках заряди розраховували за методом Kollman, а на лігандах Gasteiger.

Скоринг докінгу проводився за такими параметрами:

1. Усереднена вільна енергія, обчислена AutoDock, за всіма конформерами однієї сполуки.

2. Максимальна вільна енергія, обчислена AutoDock, за всіма конформерами однієї сполуки.

3. STD за вільною енергією, обчисленою AutoDock, за всіма конформерами однієї сполуки.

4. RMSD вільної енергії обчислене за всіма конформерами однієї сполуки.

Підход до пошуку потенційно селективних інгібіторів ізоформ СІС-К був подібний до підходу аналізу даних отриманих за допомогою ГНМ. Так само обиралися сполуки не з найбільшою афінністю, а із найбільшою різницею у передбачуваних афінностях до ізоформ.

Як критерій відбору біологічно активних сполук обрали різницю вільних енергій зв'язування між ізоформами СІС-К. Обирали сполуки різниця усереднених вільних енергій яких становила $>1,8$ кДж/моль.

В подальшому до обраних сполук застосовували набір власних алгоритмів Receptor.AI (<https://receptor.ai/>) для передбачення ADMET, що оцінювали 13 передбачених параметрів.

У якості останнього етапу фільтрації молекул використовувалася візуальна оцінка правильності положення ліганду у межах сайту зв'язування молекули-мішені.

3. РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ

Під час створення набору даних для навчання DTI із використанням бази даних ChEMBL задля класифікації пари як позитивного або негативного зразку використовувалися показники активності: IC50 / EC50 / Ka/ Kd / Ki. В якості узагальнюючого кількісного показника активності для класифікації використовувалося значення концентрації 10^3 нМ для IC50 / EC50/AC50. Тобто пари для яких показник IC50 / EC50/AC50 був менший за 103 нМ вважався позитивним зразком, а більше – негативним.

Сумарна кількість зразків отриманих за допомогою бази даних ChEMBL була такою:

Позитивних: 484979;

Негативних: 449034.

В той час як сумарна кількість зразків отриманих за допомогою бази даних STITCH була такою була такою:

Позитивних: 295162;

Негативних: 0.

Для створення набору негативних зразків генерували по 9 потенційно неактивних сполук на кожну активну пару білок-ліганд. Тобто, наприклад, якщо для білка існувало 2 активних ліганди, алгоритм генерував 18 негативних зразків.

В результаті, сумарна кількість зразків у наборі негативних зразків була такою:

Позитивних: 0;

Негативних: 6582139.

Сумарна кількість зразків по всім трьом наборам для навчання DTI була такою:

Позитивних: 780157;

Негативних: 7031173.

Графічне представлення набору даних можна побачити на рисунку 3.1. На ньому в наочному вигляді передане приблизне співвідношення у розмірів наборів зразків отриманих трьома вищезгаданими методами.



Рисунок 3.1 Графічне представлення набору даних для тренування моделі машинного навчання. П: - кількість позитивних зразків; Н: - кількість негативних зразків.

Для репрезентації ліганду у моделі DTI розраховували по 2 1024 бітних вектори (61):

- Вектор класичних відбитків за Морганом із радіусом 2;
- Вектори інваріантів відбитків на основі ознак з таким саме радіусом.

Далі обидва вектори конкатинувалися до єдиного 2048 бітного вектора.

Сам ліганд репрезентували на атомарному рівні у вигляді графу таким чином: положення важких атомів бралось у якості вузлів графу.

Нижче, у таблиці 3.1 можна побачити характеристики представлення лігандів у вигляді графу.

Таблиця 3.1 характеристики представлення лігандів у вигляді графу

Назва	Р озмір
Властивості вузлів	
Тип атома з унітарним кодом (C, N, O, F, S, Cl, Br, P, I)	9
Маса атома (у масштабі від мінімального до максимального)	1

Ступінь атома (масштабується за min-max)	1
Загальна кількість зв'язаних атомів водню (у масштабі min-max)	1
Гібридизація атомів з унітарним кодом (sp2, sp3)	2
Властивості граней	
Тип зв'язку з унітарним кодом (одинарний, подвійний, потрійний, ароматичний)	4
Зв'язок спряжений (1 - так, 0 - ні)	1
Зв'язок у циклі (1 - так, 0 - ні)	1

Граф для представлення білку-мішені орієнтували через водневі зв'язки, сольові містки та взаємодії катіон-пі, які є некомутативними і вимагають уточнення ролей для кожного із залучених залишків. Наприклад, на грані, створеній водневим зв'язком між вузлами a і b, важливо визначити, який вузол є донором, а який акцептором. Таким чином, грані a-b присвоюються ознаки грані, які відрізняються від ознак грані b-a. Так само сольові містки вимагають розрізнення позитивних (катіонний залишок) і негативних (аніонний залишок) вузлів, а взаємодії катіон-пі – між катіонними та ароматичними залишками. Навпаки, пі-стекінг та гідрофобні контакти є симетричними.

Загальною проблемою з представленням графів у машинному навчанні (ML) є їх великий розмір, що означає величезний обсяг пам'яті та повільну швидкість навчання. Щоб зменшити розмір представлення графу, використовували наступний підхід.

Для кожного з 23 стандартних типів амінокислот призначали сім характеристик AARHY7 (69) і матрицю blosum62 23×23 відповідно до вирівнювань гомологічних білкових послідовностей (70), отриманих за допомогою GraphSol (70, 71).

Щоб створити основні компоненти:

1. для кожної з 20 амінокислот об'єднали AARHY7 та вектор із відповідних замінів blosum62 для відповідної амінокислоти;

2. реплікували об'єднані вектори відповідно до відносної частоти амінокислот у протеомі людини (72), завдяки цьому враховували частоту під час створення основних компонентів;

3. проводили аналіз головних компонентів (principal component analysis, PCA) на отриманих даних (73).

Брали 8 основних компонентів, які покривають 92% мінливості даних.

Крім того, були використані структурно-залежні та залежні від послідовності ознаки вузла та грані (Таблиця 2.4.).

Властивості вузла на рівні залишків для представлення графа білка наведені у Таблиці 3.2.

Таблиця 3.2 Властивості вузла на рівні залишків для представлення графа білка

Назва	P
-------	---

	озмір
Властивості вузла	
В-фактор (масштабований за середнім і стандартним відхиленням)	1
Площа поверхні, доступної для розчинника (масштабована за середнім значенням і стандартним відхиленням)	1
Кут фі (у градусах, поділений на 180)	1
Кут пси (у градусах, поділений на 180)	1
Унітарним кодом, що належить до вторинної структури (альфа-спіраль, ізольований залишок бета-містка, ланцюг, 3-10 спіраль, поворот, згин)	6
Основні компоненти з AAPHY7 і blosum62	8
Властивості грані	
Гідрофобний контакт (1 - так, 0 - ні)	1
Водневий зв'язок (від донора до акцептора; 1 - так, 0 - ні)	1
Водневий зв'язок (від акцептора до донора; 1 - так, 0 - ні)	1
Сольовий місток (від катіона до аніону; 1 - так, 0 - ні)	1
Сольовий місток (від аніона до катіону; 1 - так, 0 - ні)	1

Пі-катіонна взаємодія (від ароматичного кільця до катіону; 1 - так, 0 - ні)	1
Пі-катіонна взаємодія (від катіону до ароматичного кільця; 1 - так, 0 - ні)	1
Паралельний пі-стекинг (1 - так, 0 - ні)	1
Перпендикулярний пі-стекинг (1 - так, 0 - ні)	1

Хоча використання білкових структур із бази даних AlphaFold забезпечує однорідність даних, воно має деякі недоліки. Зокрема, цей підхід вносить невизначеність, якщо декілька експериментальних структур розв'язуються та використовуються як шаблони. У таких випадках AlphaFold потенційно може створити суміш між кількома альтернативними конформаціями білка, що не має функціонального значення. Очевидним покращенням було б використання всієї доступної експериментально визначеної структури білка разом із прогнозами AlphaFold.

Іншим простим удосконаленням є використання графів білка на атомному рівні замість графів на рівні залишків, а також використання більш повних функцій вузлів і граней. Зокрема, на додаток до В-факторів можна використовувати інші показники гнучкості білка, такі як взаємні кореляційні матриці рухів.

Були використані такі типи ГНМ при розробки архітектури DTI: Graph Attention Network (GATv1) (74), Graph Attention Network, яка виправляє проблему статичної уваги (GATv2) (75), присутню в GATv1, Graph Convolutional Network (GCN) (76), Graph Isomorphism Network (GIN) (77), Graph Isomorphism Network with incorporated edge features (GINE) (78), ГНМ для передачі нейронних

відомостей для квантової хімії (GQC) (79) і GCN для вивчення молекулярних відбитків (GMF) (80).

Використовували лише відбитки лігандів за Морганом, лише графи лігандів або обоє одночано.

Архітектура моделі DTI представлена на рисунку 3.2.

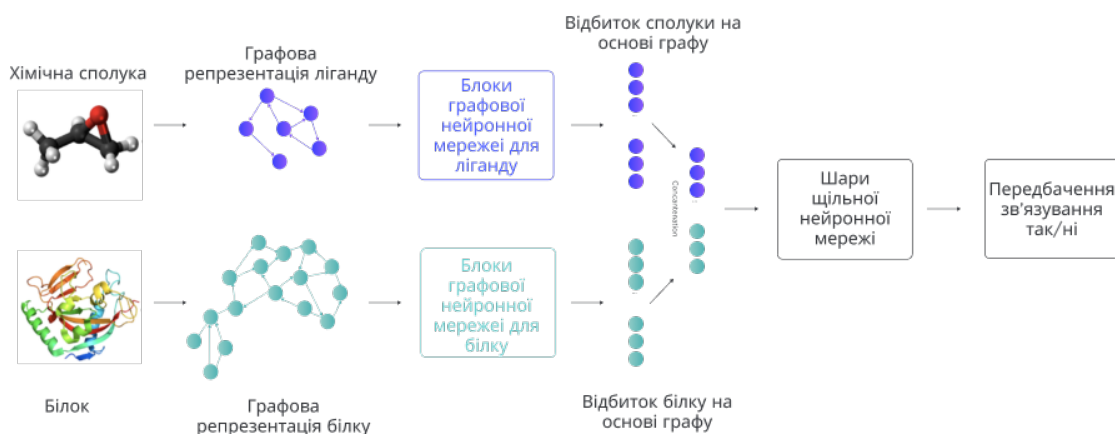


Рисунок 3.2 Архітектура моделі DTI.

Показники метрик статистичної оцінки DTI вказують на високу якість даної моделі.

В результаті отримали показники метрик наведені у Таблиці 3.3

Таблиця 3.3 показники метрик оцінки моделі МН

Метрика		Значення
Коефіцієнт кореляції Метью		0,839510
ROC	AUC	0,982

	Precision-Recall AUC	0,920
--	----------------------	-------

Для тюнінгу моделі DTI набір даних було розбито на 10 рівних частин.

Для підбору гіперпараметрів для моделі було застосовано 10% наявних даних. 10% наявних даних використовували для тестування фінальної моделі.

Згодом було застосовано 8 навчених моделей для прогнозування афінності тестового набору.

Нарешті, середнє значення для кожного показника було розраховано та порівняно з базовими підходами.

Налаштовували DTI на тип ГНМ або комбінацію типів ГНМ (максимум 2 різні послідовні типи ГНМ для ліганду або білка); кількість шарів ГНМ ; кількість багатопроходних механізмів уваги і регуляризацій для GATv1 і GATv2; розмір вихідної вибірки (кількість ознак вихідного вузла); використання функції активації між шарами ГНМ та типом функції (ReLU, Leaky ReLU, Sigmoid); використання пакетної нормалізації між шарами ГНМ; використання відбитків за Морганом, ліганд графу або обох типів ознак ліганду; включення шарів ПЗ після виходу об'єднання ГНМ ; тип об'єднання графів (середнє, додавання, максимальне) або комбінація типів; кількість шарів ФК; використання пакетної нормалізації в шарах ПЗ; випадання між шарами ПЗ; розмір виходу для кожного рівня ПЗ.

В результаті скринінгу за допомогою моделі DTI отримали 16473 сполук з передбаченою високою селективністю. Тобто обирали сполуки, для яких імовірність зв'язування з однією ізоформою була набагато більшою, ніж із іншою.

Для гомологічного моделювання структури ClC-Ka використовувалася структура бичачого ClC-K отримана за допомогою криоелектронної мікроскопії. Дана структура знаходилася у відкритому доступі в базі даних PDB.

Структура візуалізована за допомогою Mol* Viewer (88) зображена на рисунку 3.3.

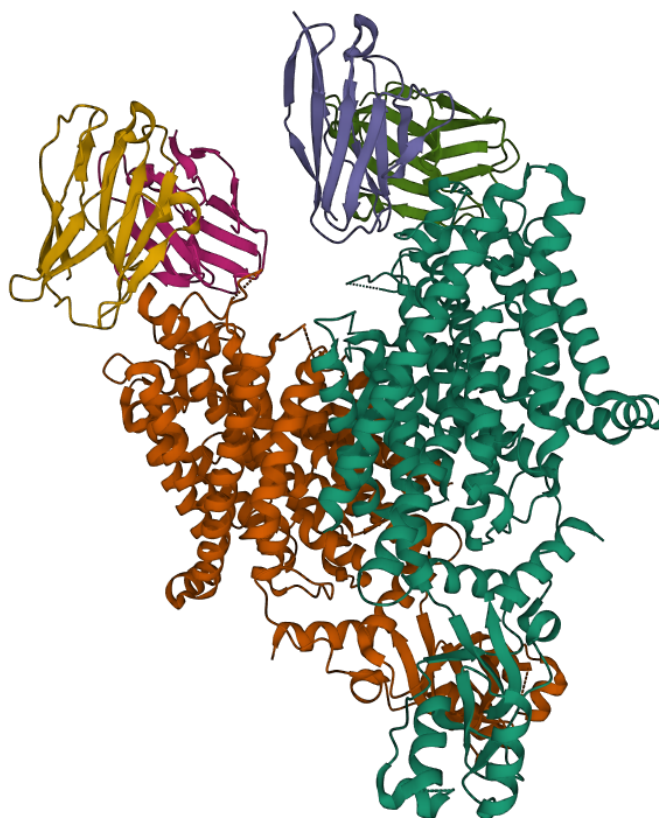


Рисунок 3.3 Структура бичачого хлоридного каналу ClC-K. Червоним та зеленим виділені домени CBS 1 та CBS 2 відповідно.

Конформації із МД ізоформ ClC-Ka та b декомпозували на окремі домени CBS 1 та CBS 2. Також була зроблена спроба вилучити петлю, що перекривала вхід до пори іонного каналу утворюваного ClC-K. Але результати докінгу в подальшому показали недоцільність даного підходу.

На рисунку 3.4 приведена для прикладу 50-та конформаціямолекулярної

динаміки домену CBS 2 СІС-Кб.

Рисунок 3.4 Структура домену CBS 2 СІС-Кб людини. МД отриманий за допомогою силового поля CHARMM у GROMACS 2021.4. Спосіб візуалізації: Biovia DS Visualizer.

На докінг подавали 16473 попередньо відібраних сполук. Ліганди автоматично готували із використанням вбудованої функції підготовки лігандів.

В результаті було проведено докінг попередньо обраних сполук проти окремих доменів білку СІС-К (СBS 1 та СBS 2) для всіх обраних кадрів молекулярної дигаміки. Докінг був гнучким відносно лігандів.

В результаті подальшої оцінки селективності сполук за допомогою молекулярного докінгу були обрані 473 сполуки, для яких було передбачено високу афінність по відношенню виключно до СІС-Ка або виключно СІС-Кб.

В результаті оцінки перспективних сполук за допомогою алгоритмів передбачення ADMET та візуального контролю було обрано 38 перспективних сполук.

Нижче, на рисунках 3.5 та 3.6 приведені структури 2 сполук селективних відносно СІС-Кб. Кількість приведених сполук є обмеженою у зв'язку із нерозголошуваністю отриманих даних.

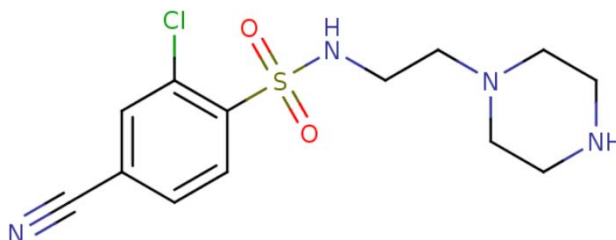


Рисунок 3.5 Структура сполуки (1) селективної до ізоформи СІС-Кб людини.

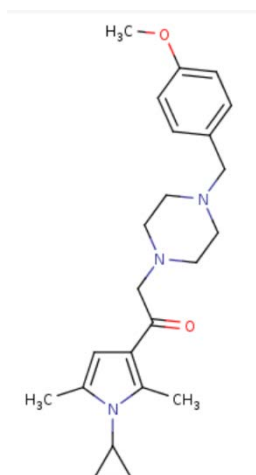


Рисунок 3.6 Структура сполуки (2) селективної до ізоформи СІС-Кв людини.

Нижче у таблиці 3.4 наведені характеристики перспективних сполук.

Таблиця 3.4 відомості про деякі перспективні сполуки

Сполука	SMILES	Енергія зв'язування для ізоформи а	Енергія зв'язування для ізоформи б
сполука (1)	<chem>C1.C1.C1C1=C(C=CC(=C1)C#N)S(=O)(=O)NCCN1CCNCC1</chem>	-5,143	-7,2035
сполука (2)	<chem>OC(=O)C(O)=O.COC1=CC=C(CN2CCN(CC(=O)C3=C(C)N(C4CC4)C(C)=C3)CC2)C=C1</chem>	-4,382	-6,434

Було помічено, що обрані інгібітори в переважній більшості були селективні щодо ізоформи СІС-Кв.

На рисунку 3.7 приведено приклад комплексів білок (СІС-Кв)-лігнад (сполука (2)), для якої передбачили високу селективність, перспективні фармакокінетичні характеристики, та для яких положення у сайті зв'язування оцінили як вірне.

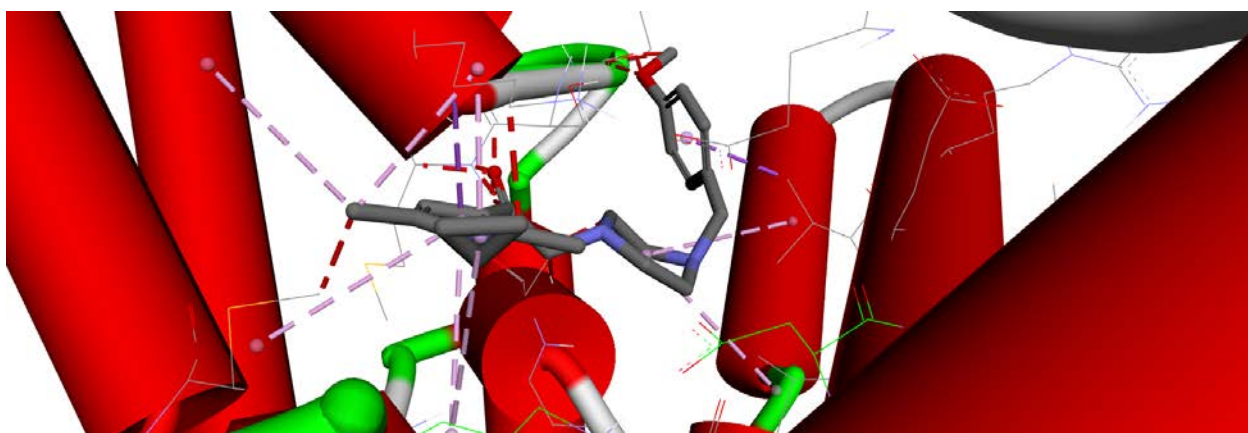


Рисунок 3.7 Сполука 2 у сайті зв'язування CBS 2 СІС-Кв. Спосіб візуалізації: Biovia DS Visualizer. Пурпуровий пі-сігма взаємодія, сірий – алкільна взаємодія, рожевий – пі-алкільна взаємодія.

На рисунку 3.8 приведено приклад комплексів білок (СІС-Кв)-лігнад (сполука (1)), для якої передбачили високу селективність, перспективні фармакокінетичні характеристики, та для яких положення у сайті зв'язування оцінили як вірне.

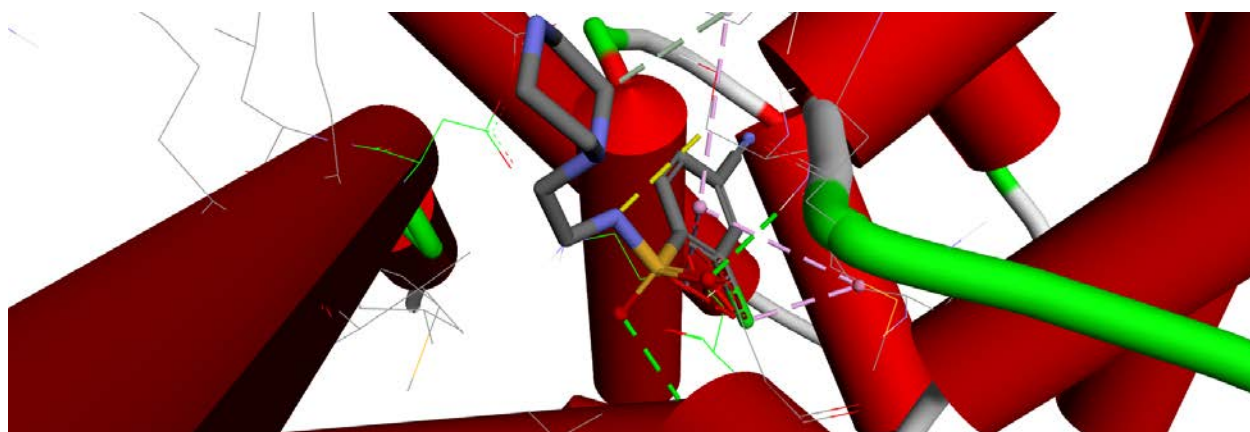


Рисунок 3.8 Сполука 1 у сайті зв'язування CBS 2 ClC-Kb. Спосіб візуалізації: Biovia DS Visualizer. Пурпуровий пі-сігма взаємодія, сірий – алкільна взаємодія, рожевий – пі-алкільна взаємодія, зелений колір – водневий зв'язок, жовтий – взаємодія водень-вуглець.

ВИСНОВКИ

1. Розробили нову модель глибокого навчання для оцінки афінності сполук з білком (DTI). Відмінною особливістю цієї моделі є графова репрезентація як білка, так і лігандів, які зберігають значну кількість інформації про їх зв'язування і просторове розташування, не завдаючи надмірного обчислювального навантаження. Метрики оцінки моделі показали такі значення: Коефіцієнт кореляції Метьюса - 0,839510, ROC- AUC - 0,982, Precision-Recall AUC - 0,920.

2. Було проведено віртуальний скринінг бази $\sim 3 \cdot 10^6$ сполук і відібрано $\sim 17 \cdot 10^3$ потенційно селективних сполук щодо ізоформ хлоридного каналу ClC-Ka або ClC-Kb для подальшого молекулярного докінгу.

3. Ідентифіковано 38 потенційно ізоформоселективних інгібіторів хлоридних каналів ClC-K при аналізі комплексів білок-ліганд (різниця усереднених вільних енергій зв'язування з різними лізоформами яких становила $>1,8$ кДж/моль), отриманих за допомогою молекулярного докінгу та даних з

алгоритмів передбачення ADMET. Селективність сполук переважно направлена на ізоформу ClC-Kb.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

- (1) Matsumura, Y.; Uchida, S.; Kondo, Y.; Miyazaki, H.; Ko, S. B. H.; Hayama, A.; Morimoto, T.; Liu, W.; Arisawa, M.; Sasaki, S.; Marumo, F. Overt Nephrogenic Diabetes Insipidus in Mice Lacking the CLC-K1 Chloride Channel. *Nature Genetics*, 1999, 21, 95–98. <https://doi.org/10.1038/5036>.
- (2) Estévez, R.; Boettger, T.; Stein, V.; Birkenhäger, R.; Otto, E.; Hildebrandt, F.; Jentsch, T. J. Barttin Is a Cl⁻ Channel β -Subunit Crucial for Renal Cl⁻ Reabsorption and Inner Ear K⁺ Secretion. *Nature*, 2001, 414, 558–561. <https://doi.org/10.1038/35107099>.
- (3) Jentsch, T. J.; Steinmeyer, K.; Schwarz, G. Primary Structure of Torpedo Marmorata Chloride Channel Isolated by Expression Cloning in *Xenopus* Oocytes. *Nature*, 1990, 348, 510–514. <https://doi.org/10.1038/348510a0>.

- (4) Jentsch, T. J. CLC Chloride Channels and Transporters: From Genes to Protein Structure, Pathology and Physiology. *Critical Reviews in Biochemistry and Molecular Biology*, 2008, 43, 3–36. <https://doi.org/10.1080/10409230701829110>.
- (5) Uchida, S.; Sasaki, S. FUNCTION OF CHLORIDE CHANNELS IN THE KIDNEY. *Annual Review of Physiology*, 2005, 67, 759–778. <https://doi.org/10.1146/annurev.physiol.67.032003.153547>.
- (6) StÄ¶lting, G.; Fischer, M.; Fahlke, C. CLC Channel Function and Dysfunction in Health and Disease. *Frontiers in Physiology*, 2014, 5. <https://doi.org/10.3389/fphys.2014.00378>.
- (7) Andrini, O.; Keck, M.; Briones, R.; Lourdel, S.; Vargas-Poussou, R.; Teulon, J. CIC-K Chloride Channels: Emerging Pathophysiology of Bartter Syndrome Type 3. *American Journal of Physiology-Renal Physiology*, 2015, 308, F1324–F1334. <https://doi.org/10.1152/ajprenal.00004.2015>.
- (8) Matsumura, Y.; Uchida, S.; Kondo, Y.; Miyazaki, H.; Ko, S. B. H.; Hayama, A.; Morimoto, T.; Liu, W.; Arisawa, M.; Sasaki, S.; Marumo, F. Overt Nephrogenic Diabetes Insipidus in Mice Lacking the CLC-K1 Chloride Channel. *Nature Genetics*, 1999, 21, 95–98. <https://doi.org/10.1038/5036>.
- (9) KONRAD, M.; VOLLMER, M.; LEMMINK, H. H.; VAN DEN HEUVEL, L. P. W. J.; JECK, N.; VARGAS-POUSSOU, R.; LAKINGS, A.; RUF, R.; DESCHÊNES, G.; ANTIGNAC, C.; GUAY-WOODFORD, L.; KNOERS, N. V. A. M.; SEYBERTH, H. W.; FELDMANN, D.; HILDEBRANDT, F. Mutations in the Chloride Channel Gene *CLCNKB* as a Cause of Classic Bartter Syndrome. *Journal of the American Society of Nephrology*, 2000, 11, 1449–1459. <https://doi.org/10.1681/asn.v11i81449>.
- (10) Simon, D. B.; Bindra, R. S.; Mansfield, T. A.; Nelson-Williams, C.; Mendonca, E.; Stone, R.; Schurman, S.; Nayir, A.; Alpay, H.; Bakkaloglu, A.; Rodriguez-Soriano, J.; Morales, J. M.; Sanjad, S. A.; Taylor, C. M.; Pilz, D.; Brem, A.; Trachtman, H.; Griswold, W.; Richard, G. A.; John, E.; Lifton, R. P. Mutations in the Chloride Channel Gene, *CLCNKB*, Cause Bartter's Syndrome

- Type III. *Nature Genetics*, 1997, 17, 171–178. <https://doi.org/10.1038/ng1097-171>.
- (11) Birkenhäger, R.; Otto, E.; Schürmann, M. J.; Vollmer, M.; Ruf, E.-M.; Maier-Lutz, I.; Beekmann, F.; Fekete, A.; Omran, H.; Feldmann, D.; Milford, D. V.; Jeck, N.; Konrad, M.; Landau, D.; Knoers, N. V. A. M.; Antignac, C.; Sudbrak, R.; Kispert, A.; Hildebrandt, F. Mutation of BSND Causes Bartter Syndrome with Sensorineural Deafness and Kidney Failure. *Nature Genetics*, 2001, 29, 310–314. <https://doi.org/10.1038/ng752>.
- (12) Schlingmann, K. P.; Konrad, M.; Jeck, N.; Waldegger, P.; Reinalter, S. C.; Holder, M.; Seyberth, H. W.; Waldegger, S. Salt Wasting and Deafness Resulting from Mutations in Two Chloride Channels. *New England Journal of Medicine*, 2004, 350, 1314–1319. <https://doi.org/10.1056/nejmoa032843>.
- (13) Hennings, J. C.; Andrini, O.; Picard, N.; Paulais, M.; Huebner, A. K.; Cayuqueo, I. K. L.; Bignon, Y.; Keck, M.; Cornière, N.; Böhm, D.; Jentsch, T. J.; Chambrey, R.; Teulon, J.; Hübner, C. A.; Eladari, D. The ClC-K2 Chloride Channel Is Critical for Salt Handling in the Distal Nephron. *Journal of the American Society of Nephrology*, 2016, 28, 209–217. <https://doi.org/10.1681/asn.2016010085>.
- (14) Grill, A.; Schießl, I. M.; Gess, B.; Fremter, K.; Hammer, A.; Castrop, H. Salt-Losing Nephropathy in Mice with a Null Mutation of the *Clcnk2* Gene. *Acta Physiologica*, 2016, 218, 198–211. <https://doi.org/10.1111/apha.12755>.
- (15) Barlassina, C.; Dal Fiume, C.; Lanzani, C.; Manunta, P.; Guffanti, G.; Ruello, A.; Bianchi, G.; Del Vecchio, L.; Macciardi, F.; Cusi, D. Common Genetic Variants and Haplotypes in Renal *CLCNKA* Gene Are Associated to Salt-Sensitive Hypertension. *Human Molecular Genetics*, 2007, 16, 1630–1638. <https://doi.org/10.1093/hmg/ddm112>.
- (16) Jeck, N.; Waldegger, S.; Lampert, A.; Boehmer, C.; Waldegger, P.; Lang, P. A.; Wissinger, B.; Friedrich, B.; Risler, T.; Moehle, R.; Lang, U. E.; Zill, P.; Bondy, B.; Schaeffeler, E.; Asante-Poku, S.; Seyberth, H.; Schwab, M.; Lang, F.

Activating Mutation of the Renal Epithelial Chloride Channel CLC-Kb Predisposing to Hypertension. *Hypertension*, 2004, 43, 1175–1181.

<https://doi.org/10.1161/01.hyp.0000129824.12959.f0>.

- (17) Lawes, C. M.; Hoorn, S. V.; Rodgers, A. Global Burden of Blood-Pressure-Related Disease, 2001. *The Lancet*, 2008, 371, 1513–1518.
[https://doi.org/10.1016/s0140-6736\(08\)60655-8](https://doi.org/10.1016/s0140-6736(08)60655-8).
- (18) Lifton, R. P.; Gharavi, A. G.; Geller, D. S. Molecular Mechanisms of Human Hypertension. *Cell*, 2001, 104, 545–556. [https://doi.org/10.1016/s0092-8674\(01\)00241-0](https://doi.org/10.1016/s0092-8674(01)00241-0).
- (19) Guyton, A. C. Blood Pressure Control—Special Role of the Kidneys and Body Fluids. *Science*, 1991, 252, 1813–1816.
<https://doi.org/10.1126/science.2063193>.
- (20) Denton, J. S.; Pao, A. C.; Maduke, M. Novel Diuretic Targets. *American Journal of Physiology-Renal Physiology*, 2013, 305, F931–F942.
<https://doi.org/10.1152/ajprenal.00230.2013>.
- (21) McCallum, L.; Lip, S.; Padmanabhan, S. The Hidden Hand of Chloride in Hypertension. *Pflügers Archiv - European Journal of Physiology*, 2015, 467, 595–603. <https://doi.org/10.1007/s00424-015-1690-8>.
- (22) Liantonio, A.; Accardi, A.; Carbonara, G.; Fracchiolla, G.; Loiodice, F.; Tortorella, P.; Traverso, S.; Guida, P.; Pierno, S.; De Luca, A.; Camerino, D. C.; Pusch, M. Molecular Requisites for Drug Binding to Muscle CLC-1 and Renal CLC-K Channel Revealed by the Use of Phenoxy-Alkyl Derivatives of 2-(p-Chlorophenoxy)Propionic Acid. *Molecular Pharmacology*, 2002, 62, 265–271.
<https://doi.org/10.1124/mol.62.2.265>.
- (23) Liantonio, A. Investigations of Pharmacologic Properties of the Renal CLC-K1 Chloride Channel Co-Expressed with Barttin by the Use of 2-(p-Chlorophenoxy)Propionic Acid Derivatives and Other Structurally Unrelated Chloride Channels Blockers. *Journal of the American Society of Nephrology*, 2004, 15, 13–20. <https://doi.org/10.1097/01.asn.0000103226.28798.ea>.

- (24) Liantonio, A.; Picollo, A.; Babini, E.; Carbonara, G.; Fracchiolla, G.; Loiodice, F.; Tortorella, V.; Pusch, M.; Camerino, D. C. Activation and Inhibition of Kidney CLC-K Chloride Channels by Fenamates. *Molecular Pharmacology*, 2005, 69, 165–173. <https://doi.org/10.1124/mol.105.017384>.
- (25) Liantonio, A.; Picollo, A.; Carbonara, G.; Fracchiolla, G.; Tortorella, P.; Loiodice, F.; Laghezza, A.; Babini, E.; Zifarelli, G.; Pusch, M.; Camerino, D. C. Molecular Switch for CLC-K Cl⁻ Channel Block/Activation: Optimal Pharmacophoric Requirements towards High-Affinity Ligands. *Proceedings of the National Academy of Sciences*, 2008, 105, 1369–1373. <https://doi.org/10.1073/pnas.0708977105>.
- (26) Liantonio, A.; Imbrici, P.; Camerino, G. M.; Fracchiolla, G.; Carbonara, G.; Giannico, D.; Gradogna, A.; Mangiatordi, G. F.; Nicolotti, O.; Tricarico, D.; Pusch, M.; Camerino, D. C. Kidney CLC-K Chloride Channels Inhibitors. *Journal of Hypertension*, 2016, 34, 981–992. <https://doi.org/10.1097/hjh.0000000000000876>.
- (27) Zifarelli, G.; Liantonio, A.; Gradogna, A.; Picollo, A.; Gramegna, G.; De Bellis, M.; Murgia, A.; Babini, E.; Conte Camerino, D.; Pusch, M. Identification of Sites Responsible for the Potentiating Effect of Niflumic Acid on CLC-K α Kidney Chloride Channels. *British Journal of Pharmacology*, 2010, 160, 1652–1661. <https://doi.org/10.1111/j.1476-5381.2010.00822.x>.
- (28) Fahlke, C. Physiology and Pathophysiology of CLC-K/Barttin Channels. *Frontiers in Physiology*, 2010, 1. <https://doi.org/10.3389/fphys.2010.00155>.
- (29) Feng, L.; Campbell, E. B.; Hsiung, Y.; MacKinnon, R. Structure of a Eukaryotic CLC Transporter Defines an Intermediate State in the Transport Cycle. *Science*, 2010, 330, 635–641. <https://doi.org/10.1126/science.1195230>.
- (30) Dutzler, R.; Campbell, E. B.; MacKinnon, R. Gating the Selectivity Filter in CLC Chloride Channels. *Science*, 2003, 300, 108–112. <https://doi.org/10.1126/science.1082708>.

- (31) Markovic, S.; Dutzler, R. The Structure of the Cytoplasmic Domain of the Chloride Channel ClC-Ka Reveals a Conserved Interaction Interface. *Structure*, 2007, 15, 715–725. <https://doi.org/10.1016/j.str.2007.04.013>.
- (32) Gradogna, A.; Fenollar-Ferrer, C.; Forrest, L. R.; Pusch, M. Dissecting a Regulatory Calcium-Binding Site of CLC-K Kidney Chloride Channels. *Journal of General Physiology*, 2012, 140, 681–696. <https://doi.org/10.1085/jgp.201210878>.
- (33) Park, E.; Campbell, E. B.; MacKinnon, R. Structure of a CLC Chloride Ion Channel by Cryo-Electron Microscopy. *Nature*, 2016, 541, 500–505. <https://doi.org/10.1038/nature20812>.
- (34) Louet, M.; Bitam, S.; Bakouh, N.; Bignon, Y.; Planelles, G.; Lagorce, D.; Miteva, M. A.; Eladari, D.; Teulon, J.; Villoutreix, B. O. In Silico Model of the Human ClC-Kb Chloride Channel: Pore Mapping, Biostructural Pathology and Drug Screening. *Scientific Reports*, 2017, 7. <https://doi.org/10.1038/s41598-017-07794-5>.
- (35) Review. *Biological Chemistry Hoppe-Seyler*, 1996, 377, 1–18. <https://doi.org/10.1515/bchm3.1996.377.1.1>.
- (36) Villoutreix, B. O.; Teleman, O.; Dahlbäck, B. *Journal of Computer-Aided Molecular Design*, 1997, 11, 293–304. <https://doi.org/10.1023/a:1007912929828>.
- (37) Verkman, A. S.; Galiotta, L. J. V. Chloride Channels as Drug Targets. *Nature Reviews Drug Discovery*, 2008, 8, 153–171. <https://doi.org/10.1038/nrd2780>.
- (38) Suh, K.; Yuspa, S. Intracellular Chloride Channels: Critical Mediators of Cell Viability and Potential Targets for Cancer Therapy. *Current Pharmaceutical Design*, 2005, 11, 2753–2764. <https://doi.org/10.2174/1381612054546806>.
- (39) Uchida, S. Physiological Role of CLC-K1 Chloride Channel in the Kidney. *Nephrology Dialysis Transplantation*, 2000, 15, 14–15. https://doi.org/10.1093/ndt/15.suppl_6.14.

- (40) Matsumura, Y.; Uchida, S.; Kondo, Y.; Miyazaki, H.; Ko, S. B. H.; Hayama, A.; Morimoto, T.; Liu, W.; Arisawa, M.; Sasaki, S.; Marumo, F. Overt Nephrogenic Diabetes Insipidus in Mice Lacking the CLC-K1 Chloride Channel. *Nature Genetics*, 1999, 21, 95–98. <https://doi.org/10.1038/5036>.
- (41) Denton, J. S.; Pao, A. C.; Maduke, M. Novel Diuretic Targets. *American Journal of Physiology-Renal Physiology*, 2013, 305, F931–F942. <https://doi.org/10.1152/ajprenal.00230.2013>.
- (42) Buffington, M. A.; Abreo, K. Hyponatremia. *Journal of Intensive Care Medicine*, 2015, 31, 223–236. <https://doi.org/10.1177/0885066614566794>.
- (43) McGreal, K.; Budhiraja, P.; Jain, N.; Yu, A. S. L. Current Challenges in the Evaluation and Management of Hyponatremia. *Kidney Diseases*, 2016, 2, 56–63. <https://doi.org/10.1159/000446267>.
- (44) Thompson, C. H.; Olivetti, P. R.; Fuller, M. D.; Freeman, C. S.; McMaster, D.; French, R. J.; Pohl, J.; Kubanek, J.; McCarty, N. A. Isolation and Characterization of a High Affinity Peptide Inhibitor of ClC-2 Chloride Channels. *Journal of Biological Chemistry*, 2009, 284, 26051–26062. <https://doi.org/10.1074/jbc.m109.031724>.
- (45) Salhany, J. M. (1996). Allosteric effects in stilbenedisulfonate binding to band 3 protein (AE1). *Cellular and Molecular Biology (Noisy-le-Grand, France)*, 42(7), 1065-1096.
- (46) Romero, M. F.; Chen, A.-P.; Parker, M. D.; Boron, W. F. The SLC4 Family of Bicarbonate Transporters. *Molecular Aspects of Medicine*, 2013, 34, 159–182. <https://doi.org/10.1016/j.mam.2012.10.008>.
- (47) Cabantchik, Z. I.; Greger, R. Chemical Probes for Anion Transporters of Mammalian Cell Membranes. *American Journal of Physiology-Cell Physiology*, 1992, 262, C803–C827. <https://doi.org/10.1152/ajpcell.1992.262.4.c803>.
- (48) Miller, C.; White, M. M. Dimeric Structure of Single Chloride Channels from Torpedo Electropex. *Proceedings of the National Academy of Sciences*, 1984, 81, 2772–2775. <https://doi.org/10.1073/pnas.81.9.2772>.

- (49) Matulef, K.; Howery, A. E.; Tan, L.; Kobertz, W. R.; Du Bois, J.; Maduke, M. Discovery of Potent CLC Chloride Channel Inhibitors. *ACS Chemical Biology*, 2008, 3, 419–428. <https://doi.org/10.1021/cb800083a>.
- (50) Liantonio, A.; Picollo, A.; Carbonara, G.; Fracchiolla, G.; Tortorella, P.; Liodice, F.; Laghezza, A.; Babini, E.; Zifarelli, G.; Pusch, M.; Camerino, D. C. Molecular Switch for CLC-K Cl⁻ Channel Block/Activation: Optimal Pharmacophoric Requirements towards High-Affinity Ligands. *Proceedings of the National Academy of Sciences*, 2008, 105, 1369–1373. <https://doi.org/10.1073/pnas.0708977105>.
- (51) Liantonio, A.; Imbrici, P.; Camerino, G. M.; Fracchiolla, G.; Carbonara, G.; Giannico, D.; Gradogna, A.; Mangiatordi, G. F.; Nicolotti, O.; Tricarico, D.; Pusch, M.; Camerino, D. C. Kidney CLC-K Chloride Channels Inhibitors. *Journal of Hypertension*, 2016, 34, 981–992. <https://doi.org/10.1097/hjh.0000000000000876>.
- (52) Koster, A. K.; Wood, C. A. P.; Thomas-Tran, R.; Chavan, T. S.; Almqvist, J.; Choi, K.-H.; Du Bois, J.; Maduke, M. A Selective Class of Inhibitors for the CLC-Ka Chloride Ion Channel. *Proceedings of the National Academy of Sciences*, 2018, 115. <https://doi.org/10.1073/pnas.1720584115>.
- (53) Bongers, B. J.; IJzerman, Adriaan. P.; Van Westen, G. J. P. Proteochemometrics – Recent Developments in Bioactivity and Selectivity Modeling. *Drug Discovery Today: Technologies*, 2019, 32–33, 89–98. <https://doi.org/10.1016/j.ddtec.2020.08.003>.
- (54) Moore, A.; Lee, M. S. Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets. *Journal of Artificial Intelligence Research*, 1998, 8, 67–91. <https://doi.org/10.1613/jair.453>.
- (55) Tyagi, S.; Mittal, S. Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning. *Lecture Notes in Electrical Engineering*, 2019, 209–221. https://doi.org/10.1007/978-3-030-29407-6_17.

- (56) Kuhn, M.; Szklarczyk, D.; Franceschini, A.; von Mering, C.; Jensen, L. J.; Bork, P. STITCH 3: Zooming in on Protein-Chemical Interactions. *Nucleic Acids Research*, 2011, 40, D876–D880. <https://doi.org/10.1093/nar/gkr1011>.
- (57) Li, X.; Wang, X.; Snyder, M. Systematic Investigation of Protein-Small Molecule Interactions. *IUBMB Life*, 2012, 65, 2–8. <https://doi.org/10.1002/iub.1111>.
- (58) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods*, 2015, 71, 58–63. <https://doi.org/10.1016/j.ymeth.2014.08.005>.
- (59) Li, J.; Fu, A.; Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdisciplinary Sciences: Computational Life Sciences*, 2019, 11, 320–328. <https://doi.org/10.1007/s12539-019-00327-w>.
- (60) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 2010, 50, 742–754. <https://doi.org/10.1021/ci100050t>.
- (61) Gobbi, A.; Poppinger, D. Genetic Optimization of Combinatorial Libraries. *Biotechnology and Bioengineering*, 1998, 61, 47–54. [https://doi.org/10.1002/\(sici\)1097-0290\(199824\)61:1<47::aid-bit9>3.0.co;2-z](https://doi.org/10.1002/(sici)1097-0290(199824)61:1<47::aid-bit9>3.0.co;2-z).
- (62) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*, 2021, 596, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.

- (63) Berman, H.; Henrick, K.; Nakamura, H. Announcing the Worldwide Protein Data Bank. *Nature Structural & Molecular Biology*, 2003, 10, 980–980. <https://doi.org/10.1038/nsb1203-980>.
- (64) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold - Making Protein Folding Accessible to All, 2021. <https://doi.org/10.1101/2021.08.15.456425>.
- (65) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics*, 2009, 25, 1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>.
- (66) Yesylevskyy, S. O. Pteros: Fast and Easy to Use Open-Source C++ Library for Molecular Analysis. *Journal of Computational Chemistry*, 2012, 33, 1632–1636. <https://doi.org/10.1002/jcc.22989>.
- (67) Yesylevskyy, S. O. Pteros 2.0: Evolution of the Fast Parallel Molecular Analysis Library for C++ and Python. *Journal of Computational Chemistry*, 2015, 36, 1480–1488. <https://doi.org/10.1002/jcc.23943>.
- (68) Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *Journal of Cheminformatics*, 2015, 7. <https://doi.org/10.1186/s13321-015-0078-2>.
- (69) Meiler, J.; Zeidler, A.; Schm◊schke, F.; M◊ller, M. Generation and Evaluation of Dimension-Reduced Amino Acid Parameter Representations by Artificial Neural Networks. *Journal of Molecular Modeling*, 2001, 7, 360–369. <https://doi.org/10.1007/s008940100038>.
- (70) Mount, D. W. Using BLOSUM in Sequence Alignments. *Cold Spring Harbor Protocols*, 2008, 2008, pdb.top39. <https://doi.org/10.1101/pdb.top39>.
- (71) Chen, J.; Zheng, S.; Zhao, H.; Yang, Y. Structure-Aware Protein Solubility Prediction from Sequence through Graph Convolutional Network and Predicted

Contact Map. *Journal of Cheminformatics*, 2021, 13.

<https://doi.org/10.1186/s13321-021-00488-1>.

- (72) Gardini, S.; Cheli, S.; Baroni, S.; Di Lascio, G.; Mangiavacchi, G.; Micheletti, N.; Monaco, C. L.; Savini, L.; Alocci, D.; Mangani, S.; Niccolai, N. On Nature's Strategy for Assigning Genetic Code Multiplicity. *PLOS ONE*, 2016, 11, e0148174. <https://doi.org/10.1371/journal.pone.0148174>.
- (73) Jolliffe, I. Principal Component Analysis. *International Encyclopedia of Statistical Science*, 2011, 1094–1096. https://doi.org/10.1007/978-3-642-04898-2_455.
- (74) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. arXiv 2017. <https://doi.org/10.48550/ARXIV.1710.10903>.
- (75) Brody, S.; Alon, U.; Yahav, E. How Attentive Are Graph Attention Networks? arXiv 2021. <https://doi.org/10.48550/ARXIV.2105.14491>.
- (76) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. arXiv 2016. <https://doi.org/10.48550/ARXIV.1609.02907>.
- (77) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks? arXiv 2018. <https://doi.org/10.48550/ARXIV.1810.00826>.
- (78) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-Training Graph Neural Networks. arXiv 2019. <https://doi.org/10.48550/ARXIV.1905.12265>.
- (79) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. arXiv 2017. <https://doi.org/10.48550/ARXIV.1704.01212>.
- (80) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. arXiv 2015. <https://doi.org/10.48550/ARXIV.1509.09292>.

- (81) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv 2019. <https://doi.org/10.48550/ARXIV.1912.01703>.
- (82) Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. arXiv 2019. <https://doi.org/10.48550/ARXIV.1903.02428>.
- (83) Toropova, A. P.; Toropov, A. A. CORAL: Binary Classifications (Active/Inactive) for Drug-Induced Liver Injury. *Toxicology Letters*, 2017, 268, 51–57. <https://doi.org/10.1016/j.toxlet.2017.01.011>.
- (84) Jin Huang; Ling, C. X. Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17, 299–310. <https://doi.org/10.1109/tkde.2005.50>.
- (85) Sofaer, H. R.; Hoeting, J. A.; Jarnevich, C. S. The Area under the Precision-recall Curve as a Performance Metric for Rare Binary Events. *Methods in Ecology and Evolution*, 2019, 10, 565–577. <https://doi.org/10.1111/2041-210x.13140>.
- (86) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. arXiv 2019. <https://doi.org/10.48550/ARXIV.1907.10902>.
- (87) Sehnal, D.; Bittrich, S.; Deshpande, M.; Svobodová, R.; Berka, K.; Bazgier, V.; Velankar, S.; Burley, S. K.; Koča, J.; Rose, A. S. Mol* Viewer: Modern Web App for 3D Visualization and Analysis of Large Biomolecular Structures. *Nucleic Acids Research*, 2021, 49, W431–W437. <https://doi.org/10.1093/nar/gkab314>.

