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

Завідувач кафедри супрамолеколярної хімії
проф. Сергій Вікторович Рябухін
Протокол №____ засідання кафедри
від “____” _____20___ р.

Шлях ло De novo генерації селективних лігандів для аденозин-рецептора A2B

Випускна кваліфікаційна робота магістра
студентки спеціальності
102 Хімія
ОП «Хемоінформатика»
Денисенко Олени Вікторівни

Науковий керівник від кафедри
професор кафедри
супрамолекулярної хімії
д.х.н. **Комаров Ігор Володимирович**

Робота виконана у відділі розробки та безпечності лікарських засобів, Лейденський академічний центр пошуку лікарських засобів, Лейденський університет
під керівництвом Prof. **Gerard JP van Westen**

Оцінка захисту роботи

_____

Київ − 2022 р.

АНОТАЦІЯ

Денисенко О.В. Шлях до De novo генерації селективних лігандів для аденозин А2В рецептора**.** -- Випускна кваліфікаційна робота магістра за спеціальністю 102 Хімія ОП «Хемоінформатика».

У роботі наведено результати побудування моделей QSAR («кількісних співвідношень структура-властивість») для активних лігандів аденозин А2В рецептора. Цей метод застосовує методи математичної статистики і машинного навчання для побудови моделей, що дозволяють за описом структур хімічних сполук передбачати їх властивості (фізичні, хімічні, біологічну активність). Отримані результатистворені QSAR моделі) можуть бути використані для подальшої генерації нових, потенційно активних та селективних антагоністів А2В рецептора.

Ключові слова: аденозин, рецептор, QSAR, SBDD, LBDD, машинне навчання (machine learning, ML), лікарський засіб

ANNOTATION

This work presents the results of building Quantitative structure–activity relationship (QSAR) models for active ligands of the adenosine A2B receptor. This method uses mathematical statistics and machine learning (ML) algoritms to build models that allow to predict the properties of chemical compounds (physical, chemical, and biological activity) using their described structures. These results (created QSAR models) can be used as an environment for further generation of new, potentially active and selective A2B receptor antagonists.

Key words: adenosine, receptor, QSAR, SBDD, LBDD, machine learning (ML), drug

ЗМІСТ

ВСТУП

Віртуальний скринінг є дуже потужним сучасним методом в розробці лікарських засобів для скринінгу багатомільйонних бібліотек малих молекул на наявність нових препаратів із бажаними властивостями, які потім можна перевірити експериментально. Подібно до інших обчислювальних підходів, віртуальний скринінг має на меті не замінити аналізи *in vitro* чи *in vivo*, а прискорити процес розробки лікарського засобу, зменшити кількість потенційних кандидатів, які підлягають експериментальній перевірці та раціоналізувати їх вибір. Крім того, обчислювальні методи набули популярності у фармацевтичних компаніях та наукових організаціях завдяки економії часу, коштів, ресурсів. Серед підходів віртуального скринінгу, пошук кількісних співвідношень структура-властивість (Quantitative Structure-Activity Relationship, QSAR) є найпотужнішим методом завдяки швидкості застосування та високої точності.

У даній роботі для побудови QSAR моделей як мішень було обрано A2B рецептор аденозину, що належить до групи G-білок-спряжених рецепторів, котрі складають 12% з усіх білкових мішеней в людей [1, 2]. A2B рецептор відіграє важливу роль у проліферації пухлин, ангіогенезі, метастазуванні та пригніченні імунітету [3]. Отже, можна зробити висновок, що антагоністи A2B рецепторів є новими, потенційно привабливими протираковими препаратами. Натхненні попередніми дослідженнями, ми вирішили сфокусуватися на дизайні селективних антагоністів A2B аденозинових рецепторів, використовуючи методи комп'ютерного моделювання та машинного навчання, за допомогою створення QSAR моделей.

1. Літературний огляд мішені

1.1 G-білок-спряжені рецептори (G protein-coupled receptor)

G-білок-спряжені рецептори (GPCR) є найбільшою групою клітинних трансмембранних рецепторів, що містить понад 800 представників [1]. GPCR складають 12% усіх потенційних білкових мішеней лікарських засобів і беруть участь у багатьох важливих біологічних процесах. Їх активація відбувається при дії різноманітних молекул та чинників (пептидів, білків, іонів, фотонів тощо). Більше того, GPCRs є цільовими білками для приблизно 35% схвалених FDA препаратів (Рис.1.1) [2]. Одним зі специфічних сімейств GPCR є сімейство аденозинових рецепторів, причетних до багатьох захворювань, серед яких хвороба Паркінсона та Альцгеймера.

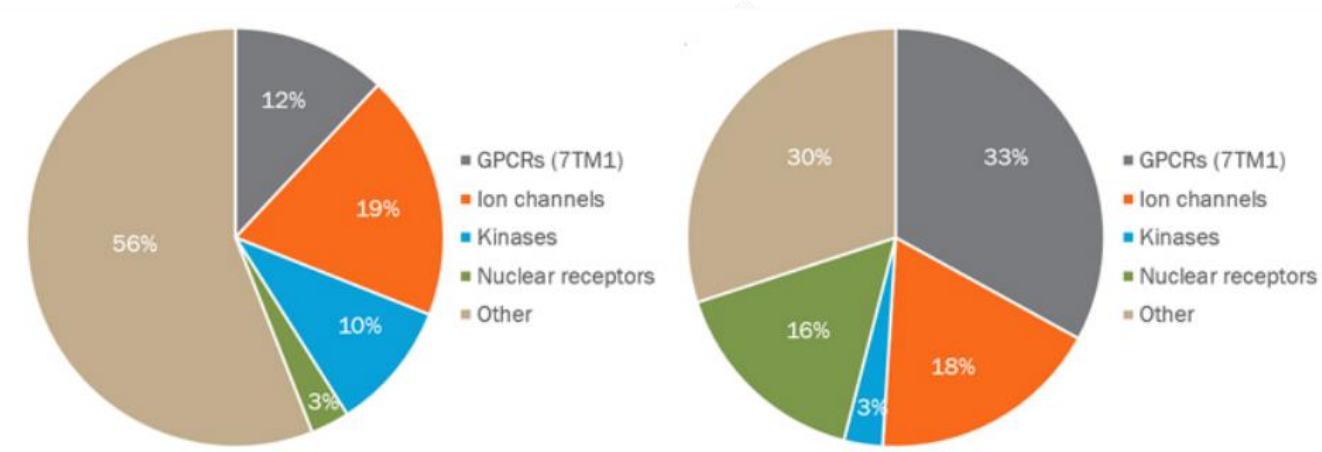

Рис.1.1: Основні сімейства білків як мішені ліків [2].

1.1.1 Фізіологічна роль

GPCR беруть участь у широкому спектрі фізіологічних процесів. Деякі приклади їх фізіологічної ролі включають:

1. Візуальне сприйняття: опсини використовують реакцію фотоізомеризації, щоб перевести електромагнітне випромінювання в клітинні сигнали.
2. Відчуття смаку: GPCR у смакових клітинах опосередковують вивільнення густуцину у відповідь на гіркі, умамі та солодкі речовини.
3. Нюх: рецептори нюхового епітелію зв'язують запахи (нюхові рецептори) і феромони (вомероназальні рецептори).
4. Регуляція поведінки та настрою: рецептори в мозку ссавців зв’язують кілька різних нейромедіаторів, включаючи серотонін, дофамін, гістамін, ГАМК (*гамма*- аміномасляна кислота) і глутамат.
5. Регуляція активності імунної системи та запалення: хемокінові рецептори зв’язують ліганди, які опосередковують зв’язок між клітинами імунної системи; рецептори, такі як рецептори гістаміну, зв’язують медіатори запалення і залучають типи клітин-мішеней у запальну відповідь [3].
6. Передача сигналів нервової системи: як симпатична, так і парасимпатична нервові системи регулюються GPCR, відповідальними за контроль багатьох автономних функцій організму, таких як кров'яний тиск, частота серцевих скорочень і процеси травлення.
7. Зондування щільності клітин: нова роль GPCR в регуляції зондування щільності клітин.
8. Модуляція гомеостазу (наприклад, водний баланс).
9. Беруть участь у зростанні та метастазуванні деяких видів пухлин [4].
10. Використовуються ендокринною системою для пептидних та амінокислотних похідних гормонів, які зв’язуються з GCPR на клітинній мембрані клітини-мішені.

1.1.2 Будова та механізм

GPCR складаються з одного поліпептиду, який згорнутий у глобулярну форму та вбудований у плазматичну мембрану клітини. Сім сегментів цієї молекули охоплюють всю ширину мембрани, що пояснює, чому GPCR іноді називають **семи-трансмембранними рецепторами**, а проміжні частини опиняються як всередині, так і зовні клітини. Позаклітинні петлі утворюють частину кишень, в яких сигнальні молекули зв'язуються з GPCR.

Як випливає з їх назви, GPCR взаємодіють з G-білками в плазматичній мембрані. Коли зовнішня сигнальна молекула зв'язується з GPCR, це викликає конформаційні зміни в GPCR. Ця зміна потім запускає взаємодію між GPCR і сусіднім G- білком.

G-білки — це спеціалізовані білки зі здатністю зв'язувати нуклеотиди гуанозинтрифосфат (GTP) і гуанозиндифосфат (GDP). Деякі G-білки, такі як сигнальний білок Ras, є невеликими білками з однією субодиницею. Також G-білки, які асоціюються з GPCR, є **гетеротримерними**, що означає, що вони мають три різні субодиниці: альфа-субодиницю, бета-субодиницю та гамма-субодиницю. Дві з них — альфа і гамма — прикріплені до плазматичної мембрани за допомогою ліпідних якорів (Рис. 1.2).

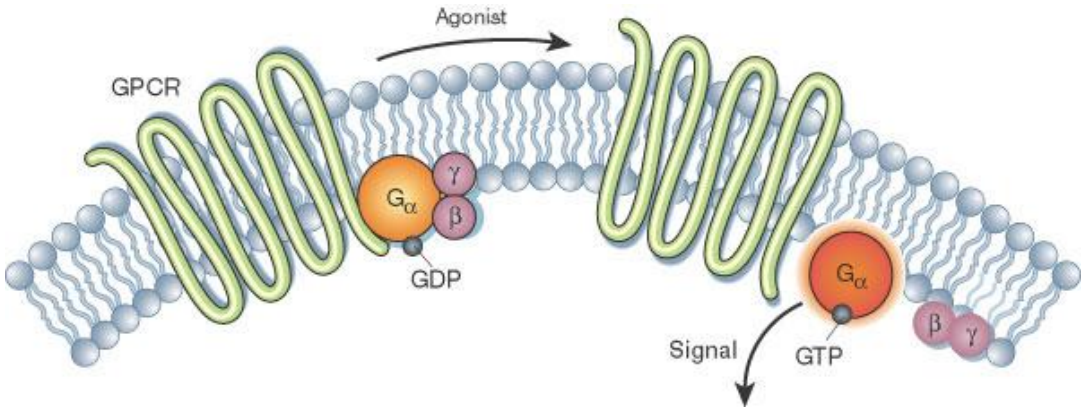

Рис. 1.2 : Активація G-альфа-субодиниці рецептора, зв'язаного з G-білком[5].

У нестимульованих клітинах стан G-альфа (помаранчеві кола) визначається його взаємодією з GDP (гуанозиндифосфатом), G-бета-гамма (фіолетові кружечки) та G-білок-спряженим рецептором (GPCR; світло-зелені петлі). Після стимуляції рецептора лігандом, який називається агоністом, стан рецептора змінюється. G-альфа дисоціює від рецептора, G-бета-гамма обмінюють GDP на GTP (гуанозинтрифосфат), що призводить до G-альфа-активації. Потім G-альфа активує інші молекули в клітині [5].

Альфа-субодиниця G-білка зв’язує або GTP, або GDP, в залежності від того, чи активний білок (GTP), чи неактивний (GDP). Якщо сигнал відсутній, GDP приєднується до альфа-субодиниці, і весь комплекс G-білок-GDP зв’язується з сусіднім GPCR. Така конформація зберігається до тих пір, поки сигнальна молекула не приєднається до GPCR. У цей момент зміна конформації GPCR активує G-білок, а GTP замінює GDP, зв’язаний з альфа-субодиницею. В результаті субодиниці G- білка розпадаються на дві частини: альфа-субодиницю, зв’язану з GTP, і бета-гамма-димер. Обидві частини залишаються закріпленими на плазматичній мембрані, але вони більше не зв’язані з GPCR, тому тепер можуть дифундувати для взаємодії з іншими білками мембрани. G-білки залишаються активними до тих пір, поки їх альфа-субодиниці з'єднані з GTP. Однак, коли цей GTP

гідролізується з утворенням GDP, субодиниці знову набувають форми неактивного гетеротримера, і весь G-білок знову асоціюється з тепер неактивним GPCR. Таким чином, G-білки працюють як перемикач — вмикаються або вимикаються взаємодією сигнал-рецептор на поверхні клітини.

Коли G-білок активний, його GTP-зв’язана альфа-субодиниця та відповідний бета-гамма-димер можуть передавати сигнали в клітині, взаємодіючи з іншими білками мембрани, які беруть участь у його передачі. Специфічні мішені активованих G-білків включають різні ферменти, які виробляють вторинні месенджери, а також певні йонні канали, які дозволяють йонам діяти як вторинні месенджери. Деякі G-білки стимулюють активність цих мішеней, тоді як інші є інгібуючими. Геноми хребетних містять кілька генів, які кодують альфа-, бета- і гамма-субодиниці G-білків. Багато різних субодиниць, які кодуються цими генами, поєднуються різними способами, щоб продукувати різноманітне сімейство G-білків (Рис. 1.3).

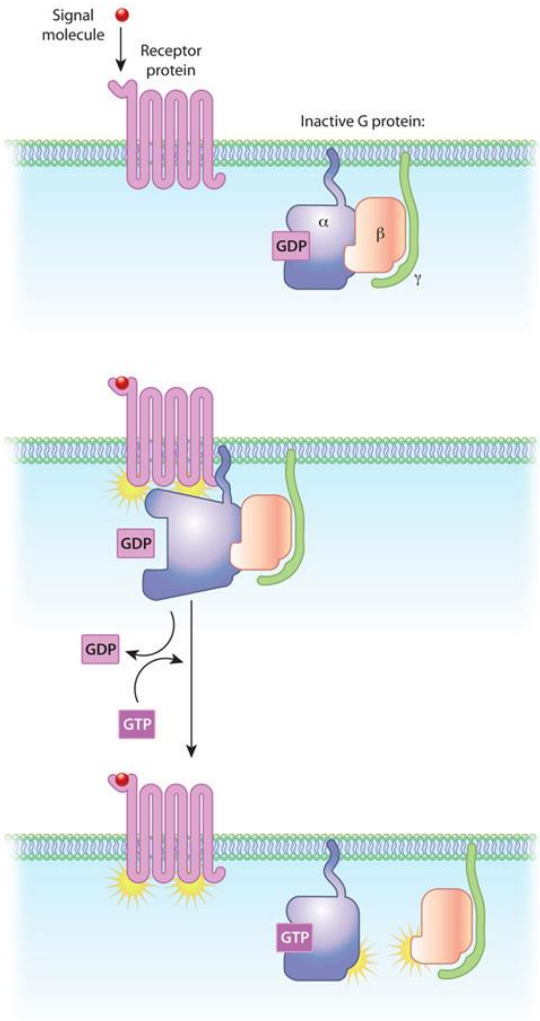

Рис. 1.3: Зв'язок G-білків з плазматичною мембраною[6].

На цій діаграмі активації рецептора, пов’язаного з G-білком, альфа-, бета- та гамма-субодиниці показані з чіткими взаємними зв’язками і асоційовані з плазматичною мембраною. Після обміну GDP з GTP на альфа-субодиниці, альфа-субодиниця та бета-гамма-комплекс можуть взаємодіяти з іншими молекулами для стимулювання сигнальних каскадів. Важливо, що альфа-субодиниця і бета-гамма-комплекс залишаються прив’язаними до плазматичної мембрани, допоки вони активовані. Ці активовані субодиниці можуть діяти на йонні канали в клітинній мембрані, а також на клітинні ензими та вторинні месенджери, що є в клітині [6].

Завдяки описаній вище послідовності подій, GPCR регулюють неймовірний діапазон функцій організму, від відчуттів до гормональних реакцій [7].

1.2 Аденозин

Аденозин (символ A або Ado) — це органічна сполука, яка широко зустрічається в природі у вигляді різноманітних похідних. Молекула складається з аденіну, приєднаного до рибози за допомогою β-N9-глікозидного зв'язку (Рис. 1.4). Аденозин є одним з чотирьох нуклеозидних будівельних блоків для ДНК і РНК, які необхідні для всього живого. Його похідні включають носіїв енергії, таких як аденозин моно-, ди- і трифосфат, також відомий як АМФ/АДФ/АТФ.

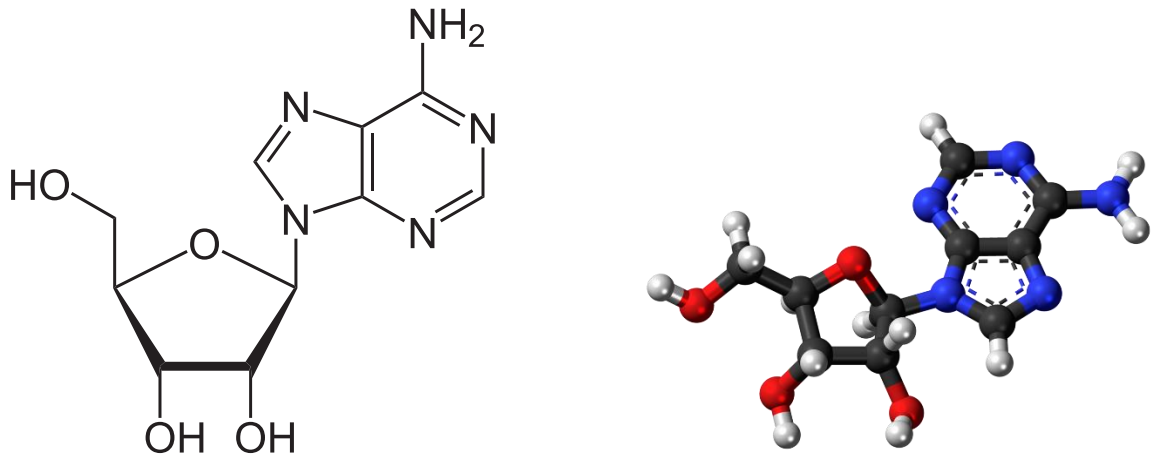

Рис. 1.4: структура аденозину

Перші докази ролі аденозину в клітинній фізіології датуються 1927 роком, коли в екстрактах із серцевих тканин було виявлено наявність сполуки аденіну, здатної уповільнювати серцевий ритм та його частоту [8]. Відтоді вчені з різних галузей — фізіології, біохімії, фармакології, хімії та імунології — зосереджували свої зусилля на дослідженні багатьох ролей

аденозину в медицині та хворобах, створюючи таким чином нову область досліджень.

Аденозин, пуриновий нуклеозид, був описаний як «метаболіт у відповідь» (retaliatory metabolite) завдяки його здатності функціонувати аутокринним способом і змінювати активність ряду клітин в процесі його позаклітинного накопичення під час клітинного стресу або пошкодження [9]. Ці ефекти є значною мірою захисними і викликаються зв'язуванням аденозину з будь-яким із чотирьох підтипів аденозинових рецепторів, а саме A1, A2A, A2B, A3, які експресуються в більшості органів. Кожен з них кодується окремим геном і виконує різні функції, хоча іноді ці функції можуть збігатися. Наприклад, рецептори A1 і A2A відіграють важливу роль у регуляції споживання кисню та коронарного кровотоку. Доведено, що аденозин відіграє ключову роль у різних фізіологічних функціях, таких як індукція сну, нейропротекція та захист від окисного стресу. Зараз з'являється все більше доказів того, що аденозинові рецептори можуть бути перспективними терапевтичними цілями в широкому діапазоні станів, включаючи серцеві, легеневі, імунологічні та запальні захворювання [10] .

З філогенетичної точки зору, найперші докази ролі аденозину як життєво необхідної молекули, було опубліковано в 1981 році, коли виділений аденозин був ідентифікований як клітинний сигнал в бактерії *Myxococcus xanthus* [11]. Згодом він був пов'язаний з енергетичним метаболізмом, завдяки фізіологічним доказам збільшення вироблення аденозину в лейкоцитах і клітинах серця під час катаболізму АТФ. Також було помічено, що аденозин відіграє роль «помічника» у захисті таких клітин, як нейрони та кардіоміоцити від стресових умов, дозволяючи їм регулювати споживання енергії та адаптувати свою активність для зменшення потреби в АТФ. Цей ефект в основному зумовлений зменшенням енерговитратних видів діяльності, а також збільшенням

концентрації поживних речовин/кисню в тканинах через розширення судин (Рис. 1.5). Це спростувало існуючу гіпотезу про його роль як другого посередника шляху цАМФ, а пізніше спонукало до введення терміну «метаболіт у відповідь» для опису цього корисного нуклеозиду. За нормальних фізіологічних умов рівень позаклітинного аденозину становить від 20 до 300 нМ , наближаючись до низького мікромолярного діапазону в екстремальних фізіологічних ситуаціях, таких як інтенсивні фізичні навантаження або низький рівень кисню в атмосфері (наприклад, на великій висоті) і мікромолярних концентраціях (~30 мкМ) та при таких патологічних станах, як ішемія [12].

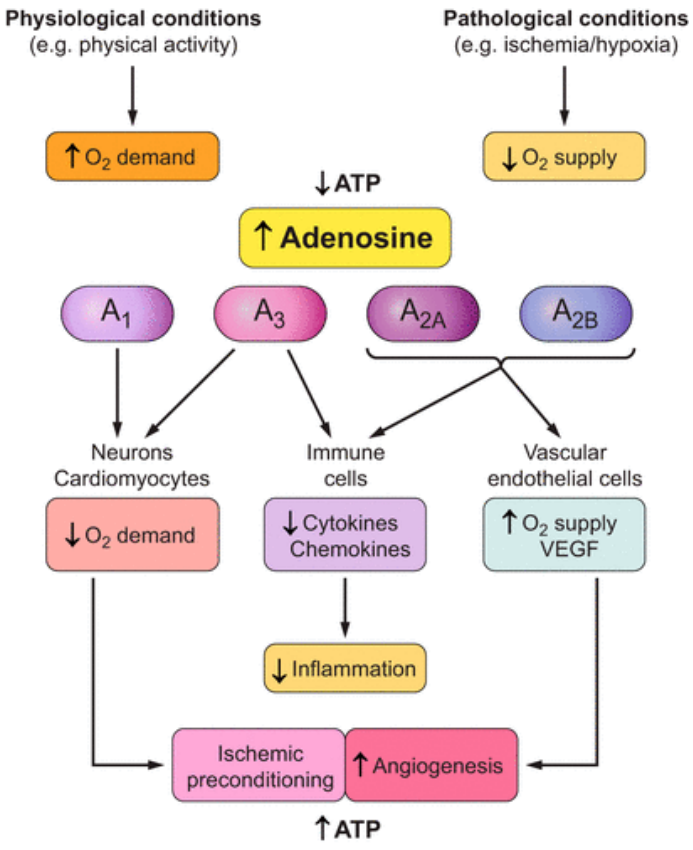

Рис. 1.5**:** Фізіологічна роль аденозину, що реалізується шляхом його взаємодії з рецепторами аденозину A_1, A_{2A}, A_{2B} і A_3 [13].

Отже, аденозин є ендогенним медіатором, рівень якого сильно підвищується після гіпоксії, ішемії або фізичної активності через споживання АТФ [13]. Він захищає організм за допомогою різних механізмів, які запускаються активацією аденозинових рецепторів, що приводить до зниження потреби в кисні та зменшення запалення, збільшення постачання кисню та ангіогенезу, а також до ішемічного прекондиціонування.

1.3 A2B рецептор

Ген *a2в* кодує рецептор аденозину (A2BAR), який є членом надсімейства рецепторів, зв'язаних з G-білком. Цей інтегральний мембранний білок стимулює активність аденілатциклази в присутності аденозину [14].

З'являється все більше доказів того, що A2BAR відіграє важливу роль у проліферації пухлинних клітин, ангіогенезі, метастазуванні та пригніченні імунітету [15] (Рис 1.6). Таким чином, антагоністи A2BAR є новими, потенційно привабливими протипухлинними агентами [16, 17, 18, 19, 20, 21, 22, 23]. Кілька антагоністів, націлених на A2BAR, зараз проходять клінічні випробування для терапії різних видів раку.

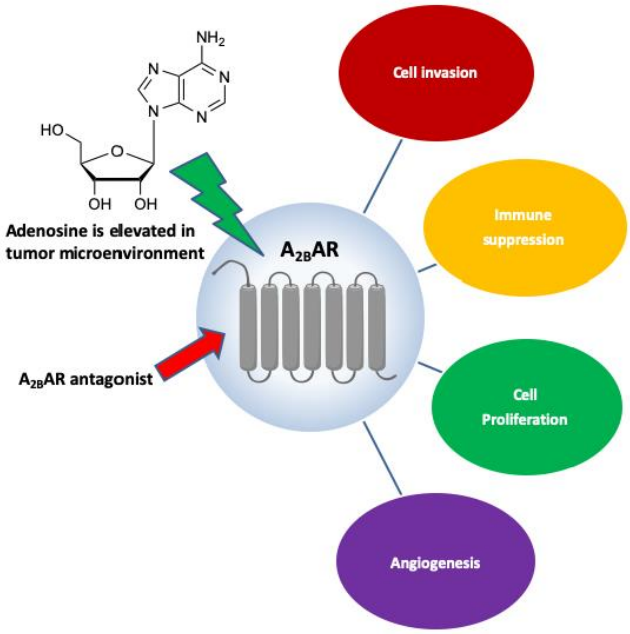

Рис 1.6: Роль A2B рецептору [15].

Хоча ефекти A2BAR у зрізах мозку були охарактеризовані на початку 1980-х років [24], донедавна A2BAR був недостатньо охарактеризований у порівнянні з іншими трьома рецепторами. Це частково пов’язано з тим, що A2BAR має низьку спорідненість до ендогенного агоніста аденозину. Таким чином, вважалося, що A2BAR має незначне фізіологічне значення. Проте все більше доказів демонструють, що спостерігається різке збільшення позаклітинної концентрації аденозину та значне підвищення експресії A2BAR при багатьох патологічних станах [25, 26, 27], таких як гіпоксія, запалення та рак, що може свідчити про вирішальну роль A2BAR при багатьох захворюваннях. Тому ми вирішили, що нові селективні антагоністи A2BAR можуть стати перспективними та цікавими біологічно активними речовинами.

Розділ 2. Використання технологій у дизайні лікарських засобів

2.1 Комп'ютерне проектування у дизайні лікарських засобів

Комп'ютерне моделювання у драг дизайні (англ. CADD -- Computer-Aided Drug Design) дає можливість значно збільшити кількість нових потенційних лікарських сполук. Воно використовується не тільки для описування терапевтичної активності молекул, але й передбачення можливих похідних цих молекул, які можуть мати значно підвищену активність. У кампанії з виявлення ліків CADD зазвичай використовується для трьох основних цілей:

(1) фільтрація великих бібліотек сполук у менші набори передбачених активних сполук, які можна дослідити експериментально;

(2) керування оптимізацією лідів, щоб підвищити їх спорідненість з мішенями або оптимізувати метаболізм і інші фармакокінетичні властивості, включаючи всмоктування, розподіл, метаболізм, виведення та токсичність;

(3) моделювання нових сполук, шляхом «нарощування» вихідних молекул по одній функціональній групі за раз, або шляхом об’єднання фрагментів у нові хемотипи. Малюнок ілюструє положення CADD в області дизайну лікарських засобів (Рис. 2.1).

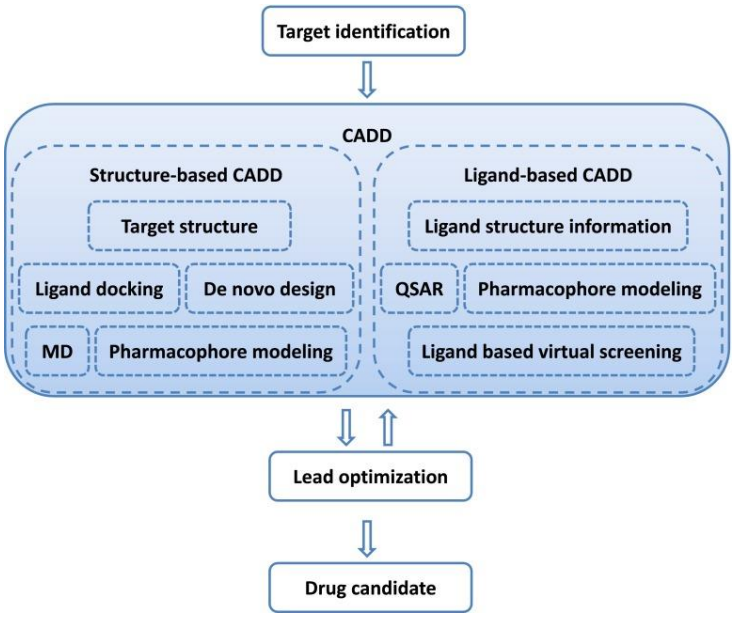

Рис. 2.1: Комп'ютерне моделювання у дизайні лікарських засобів

Визначається терапевтична мішень, проти якої необхідно розробити препарат. Залежно від наявності інформації про структуру використовується підхід на основі структури або підхід на основі ліганду. Успішне використання CADD дозволяє згенерувати декілька потенйіно активних сполук. Далі одержані молекули синтезуються, тестуються *in vivo* для виявлення кандидатів лікарських засобів.

2.2 Структура- та ліганд-базовані підходи

В даний час багато фармацевтичних компаній і науково-дослідних установ по всьому світу створили власні відділи CADD, і постійні зусилля були спрямовані на розробку та оптимізацію методологій і програмного забезпечення для розробки лікарських засобів на основі швидкого розвитку кристалографії та успішного застосування гомологічного моделювання та структурно-базованого віртуального скринінгу (англ. Structure-based virtual

screen -- SBVS), що виявився корисним методом для швидкого виявлення біоактивних молекул.

Розробка ліків на основі структури (англ. **Structure-based drug design – SBDD**) базується на інформації, одержаної з тривимірної структури біологічної мішені, отриманої за допомогою таких методів, як рентгенівська кристалографія або ЯМР-спектроскопія [28]. Якщо експериментальна структура мішені недоступна, можна створити гомологічну модель мішені на основі експериментальної структури спорідненого білка. Використовуючи структуру біологічної мішені, можна розробити лікарські засоби-кандидати, які, як передбачається, з високою спорідненістю та селективністю зв'язуються з мішенню, використовуючи інтерактивну графіку та компетенції спеціалістів в галузі медичної хімії (Рис.2.2). В якості альтернативи можна використовувати різні автоматизовані обчислювальні процедури, щоб запропонувати нові ліки-кандидати [29].

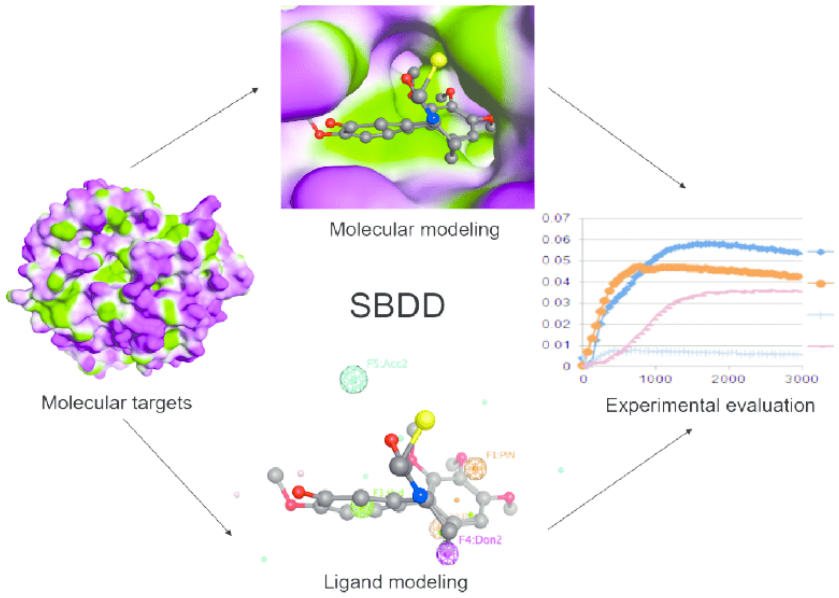

Рис.2.2: Схема структура-орієнтованого дизайну лікарських засобів [29].

Сучасні методи розробки ліків на основі структури мішені можна приблизно розділити на три основні категорії [30]. Перший метод полягає в ідентифікації нових лігандів для даної мішені шляхом пошуку у великих базах даних тривимірних структур малих молекул, щоб знайти ті, що відповідають сайту зв’язування мішені за допомогою комп’ютерних програм. Цей метод відомий як віртуальний скринінг. Друга категорія — це розробка нових лігандів *de novo*. У цьому методі молекули ліганду створюються в межах сайту зв’язування шляхом поетапного збирання малих частин. Ці частини можуть бути як окремими атомами, так і молекулярними фрагментами. Ключова перевага такого методу полягає в тому, що можна запропонувати нові структури, які не містяться в жодній базі даних [31, 32, 33]. Третій метод — оптимізація відомих лігандів шляхом оцінки запропонованих аналогів у сайті зв'язування.

Дуже часто використовують перший підхід – молекулярний докінг. Його можна використовувати для моделювання взаємодії між малою молекулою та білком на атомному рівні, що дозволяє охарактеризувати поведінку малих молекул у місці зв’язування цільових білків, а також з’ясувати фундаментальні біохімічні процеси [34]. Процес докінгу включає два основних етапи: передбачення конформації ліганду, а також його положення та орієнтації всередині цих сайтів (зазвичай називають *пози*) та оцінку спорідненості зв’язування (Рис.2.3) [35].

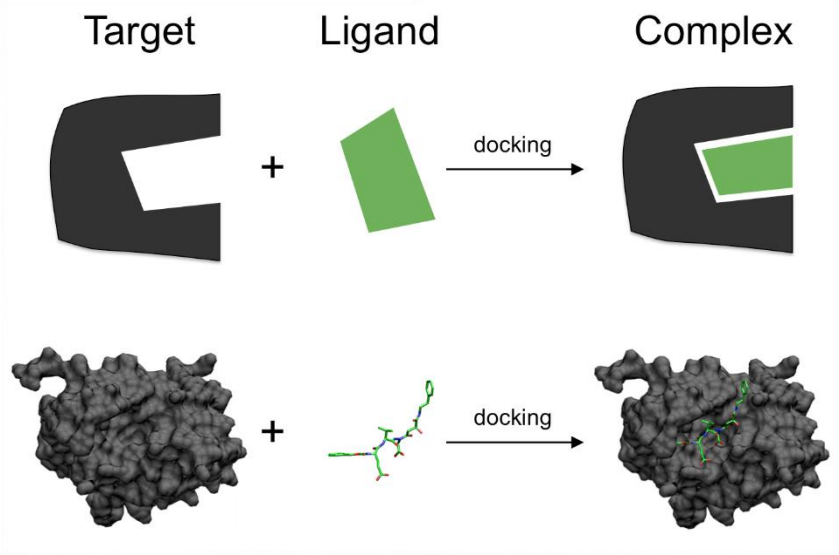

Рис.2.3: Схематична ілюстрація приєднання невеликої молекули ліганду (зелений) білковою мішенню (чорний), утворюючи комплекс білок-ліганд.

Дизайн ліків на основі ліганду (англ. ligand-based drug design -- LBDD) включає в себе такі методи, як пошук кількісних співвідношень структура-властивість (3D-QSAR), пошук на основі 2D подібності. Використання скафолду та фармакофорні дослідження також є ефективними підходами для підвищення коректності прогнозування активності на основі доступної інформації про молекулу [36].

Найпопулярнішими підходами до розробки ліків на основі лігандів є метод QSAR та фармакофорне моделювання. QSAR – це обчислювальний метод для кількісної оцінки кореляції між хімічними структурами ряду сполук і конкретним хімічним або біологічним процесом. Основна гіпотеза методу QSAR полягає в тому, що подібні структурні або фізико-хімічні властивості дають подібну активність [37, 38]. Спочатку визначається група хімічних речовин або молекул, які мають бажану біологічну активність. Далі встановлюється кількісне співвідношення між фізико - хімічними особливостями активних молекул і біологічною активністю. Розроблена

модель QSAR потім використовується для оптимізації активних сполук, щоб максимізувати відповідну біологічну активність. Далі передбачені сполуки експериментально перевіряють на очікувану активність. Таким чином, метод QSAR може бути використаний як інструмент для ідентифікації модифікацій сполуки з покращеною активністю.

QSAR побудована на серії послідовних кроків: (Рис. 2.4)

(1) Визначнння лігандів з експериментально виміряними значеннями бажаної біологічної активності. В ідеалі ці ліганди мають споріднений ряд, але вони повинні мати широку хімічну різноманітність, щоб мати значну варіацію активності.

(2) Ідентифікація та визначення молекулярних дескрипторів, пов'язаних з різними структурними та фізико - хімічними властивостями досліджуваних молекул.

(3) Визначення кореляції між молекулярними дескрипторами та біологічною активністю, які можуть пояснити різницю активності в наборі даних.

(4) Перевірка статистичної стабільності і передбачуваної сили моделі QSAR.

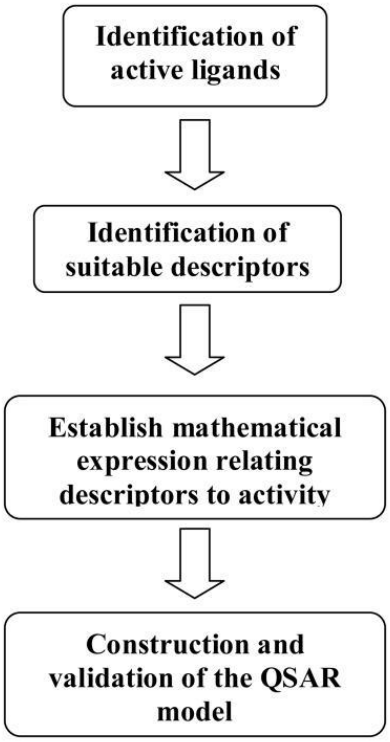

Рис. 2.4: Типовий робочий процес методів QSAR [39].

У поєднанні з наявністю різноманітних баз даних сполук, стратегії на основі структур або лігандів значно підвищують ефективність відкриття ліків і відкривають нові горизонти та перспективні шляхи для подолання небезпечних для життя захворювань (Рис. 2.5) [40].

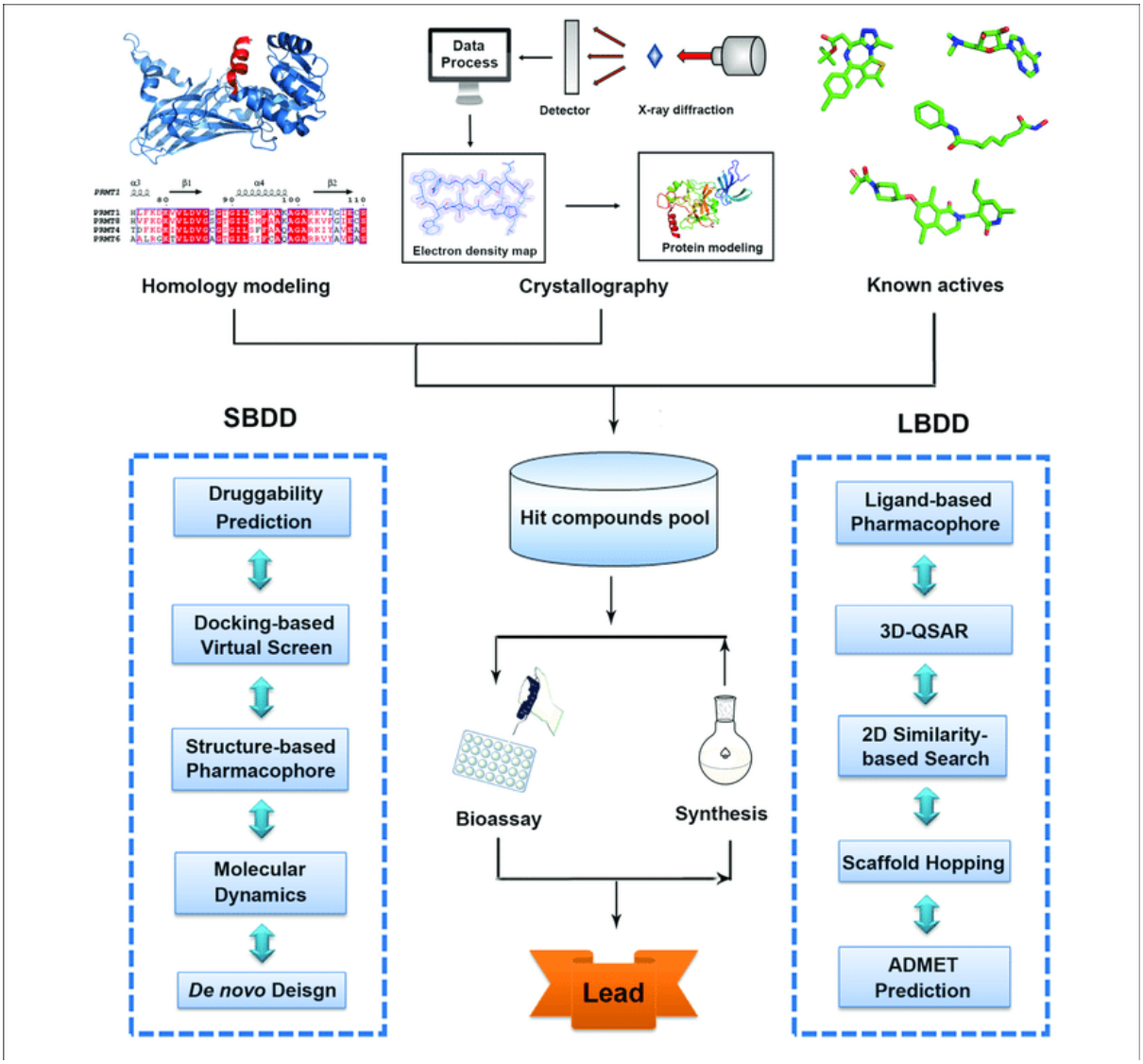

Рис. 2.5: Процес дизайну лікарських засобів на основі структури (SBDD) і дизайну лікарських засобів на основі ліганду (LBDD) [40].

2.3 Штучний інтеллект: мережі та інструменти

За останні кілька років відбулося різке зростання цифровізації даних у фармацевтичному секторі. Однак ця цифровізація супроводжується проблемами фінансування, ретельного вивчення та застосування цих знань для вирішення складних клінічних завдань [41]. Це мотивує вчених

використовувати штучний інтелект (ШІ), оскільки він може обробляти великі обсяги даних із розширеною автоматизацією .

ШІ включає кілька областей та методів, таких як аналіз, представлення знань, пошук рішення і, серед них, фундаментальна парадигма машинного навчання (ML). ML використовує алгоритми, які можуть розпізнавати залежності (патерни) в наборах даних. Підполем ML є глибоке навчання (DL), яке залучає штучні нейронні мережі (ANN). Вони включають набір взаємопов'язаних складних обчислювальних елементів, що включають «перцептони», аналогічні людським біологічним нейронам, що імітують передачу електричних імпульсів у мозку людини [42]. ANN являють собою набір вузлів, кожен з яких отримує окремий вхід, в кінцевому підсумку перетворюючи їх у вихідні сигнали, як окремо, так і багатозв'язно, використовуючи алгоритми для вирішення задач [43] .

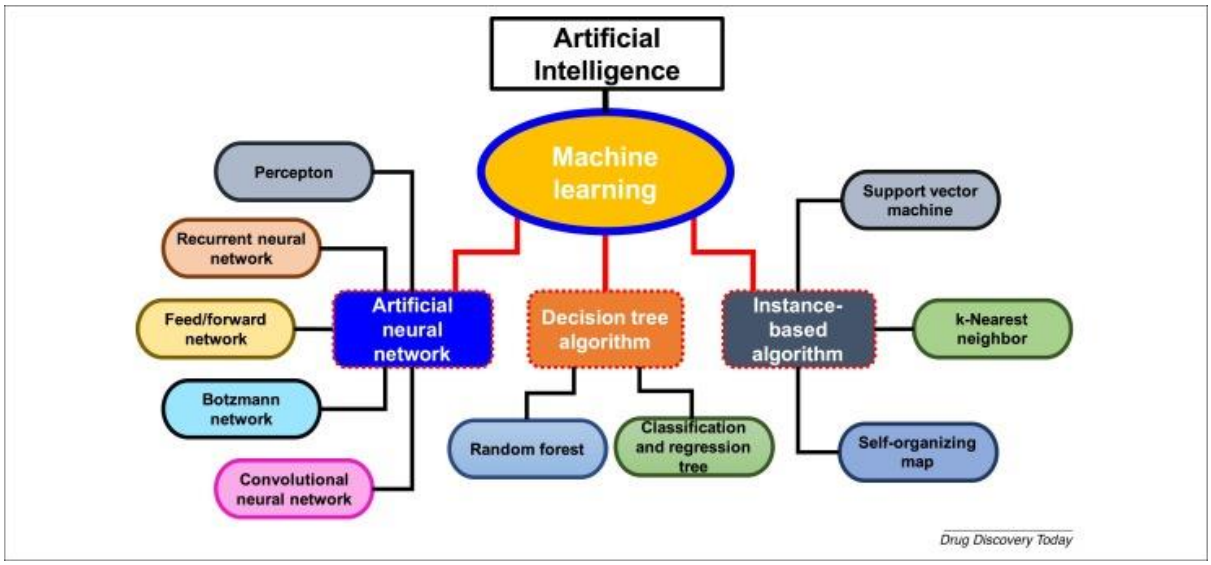

Рис. 2.6: різні класи методів ШІ [44].

На цьому малюнку (Рис. 2.6) показані різні класи методів ШІ разом із їх підкласами, які можуть бути реалізовані в різних сферах виявлення та розробки ліків[44].

Рішення ML засновані на моделюванні та аналізі великих баз даних. Дані можуть походити з різних джерел (наприклад, сховищ даних, внутрішніх експериментів та публікацій) і можуть відрізнятися за форматом, що робить узагальнення, зберігання та підготовку даних для аналізу складним, хоча й необхідним.

ML навчає систему самостійно робити висновки та приймати рішення без будь-якої зовнішньої підтримки. Рішення приймаються, коли система вчиться і вдосконалюється на основі минулого досвіду - вона вчиться з даних, які їй надали, і розшифровує пов'язані патерни, що містяться в ній. Потім за допомогою розпізнавання та аналізу патернів система видає «результат», яким може бути передбачення або класифікація [45]. Завдання ML в цілому поділяються на три категорії: навчання з наглядом (supervised learning), навчання без нагляду (unsupervised learning) та послідовне навчання (sequential learning). Дані в ML можуть бути двох типів – мічені (labeled) та немічені (unlabeled).

Навчання з наглядом (supervised learning) спирається на набір даних, який діє як тренер, навчаючи модель або машину. Після навчання модель може почати робити прогнози та приймати рішення в міру отримання нових даних. DL та SVM, які зазвичай використовуються в біологічних цілях, підпадають під supervised learning. Глибоке навчання (DL) [46] використовує штучні нейронні мережі (ANN) для виявлення дуже складних патернів у великих наборах даних (Рис. 2.7).

Навчання без нагляду (unsupervised learning) встановлює зв’язки або закономірності немічених даних (unlabeled data). Модель навчається

незалежно через спостереження і створює кластери спостережуваних моделей і відносин у наборі даних.

Послідовне навчання (sequential learning) дозволяє агенту, який є цілеспрямованою сутністю, навчатися в інтерактивному середовищі, використовуючи зворотний зв'язок з його власних дій і досвіду. Послідовне навчання ґрунтується на методі проб і помилок для прийняття послідовності рішень.

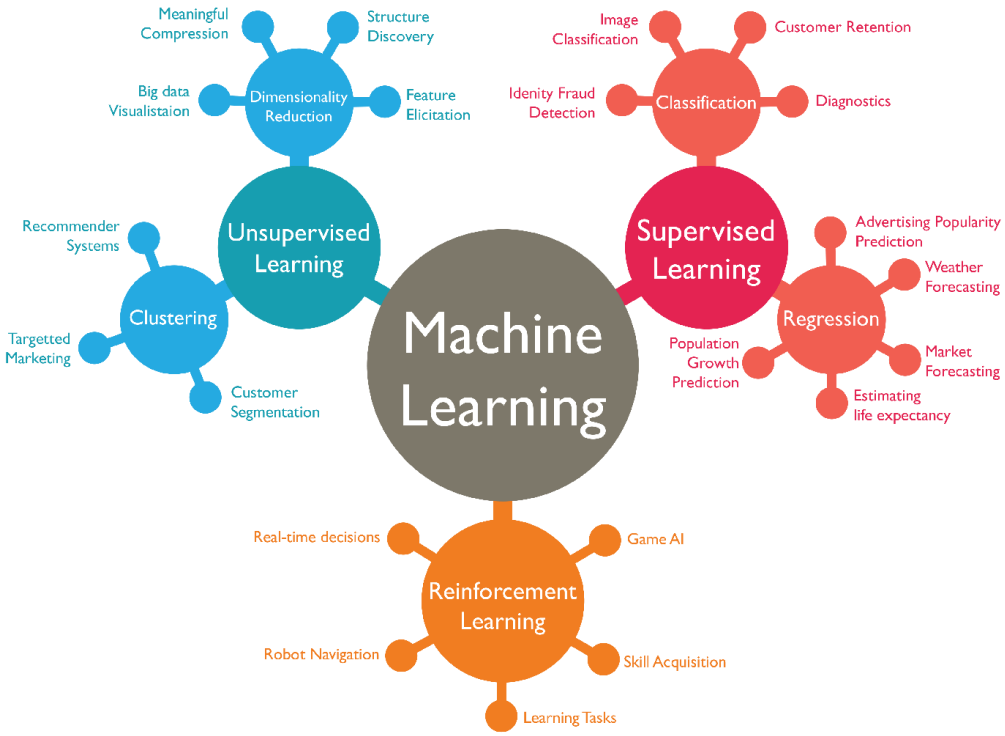

Рис. 2.7: Схема підходів машинного навчання для дослідження лікарських засобів [46].

Підходи ML можна застосовувати на кількох етапах під час раннього виявлення ліків для [47]:

- Спрогнозувати цільову структуру
- Визначити та оптимізувати так звані "хіти"
- Дослідити біологічну активність нових лігандів
- Розробити моделі, які передбачають фармакокінетичні та токсикологічні властивості кандидатів на ліки

Розділ 3 Матеріали та методи

3.1 Класифікаційне та регресійне прогнозувальне моделювання

Прогнозувальне моделювання (predictive modeling) — це проблема розробки моделі з використанням історичних даних для прогнозування нових даних, на які ми не маємо відповіді.

Прогнознувальне моделювання можна описати як математичну задачу апроксимації функції відображення (f) від вхідних змінних (X) до вихідних змінних (y). Це називається проблемою апроксимації функції. Завдання алгоритму моделювання полягає в тому, щоб знайти найкращу функцію відображення, яку ми можемо, враховуючи наявний час і ресурси. Як правило, ми можемо розділити всі завдання наближення функцій на задачі класифікації та задачі регресії.

3.1.1.Класифікація (classification)

Моделювання з прогнозуванням класифікації (classification predictive modeling) — це завдання апроксимації функції відображення (f) від вхідних змінних (X) до дискретних вихідних змінних (y). Вихідні змінні часто називають мітками або категоріями. Функція відображення прогнозує клас або категорію для даного спостереження. Наприклад, електронний лист із текстом можна віднести до одного з двох класів: «спам» і «*не спам*».

- Проблема класифікації вимагає, щоб приклади були класифіковані в один із двох або більше класів.
- Класифікація може мати дійсні або дискретні вхідні змінні.
- Проблему з двома класами часто називають проблемою двокласової або бінарної класифікації.

- Проблему з більш ніж двома класами часто називають проблемою класифікації кількох класів.
- Проблема, де прикладу призначено кілька класів, називається проблемою класифікації з кількома мітками.

Зазвичай моделі класифікації передбачають безперервне значення як ймовірність належності даного прикладу до кожного вихідного класу. Імовірності можна інтерпретувати як імовірність або впевненість даного прикладу, що належить кожному класу. Прогнозовану ймовірність можна перетворити на значення класу, вибравши мітку класу з найбільшою ймовірністю.

Існує багато способів оцінки навичок класифікаційної прогнозної моделі, але, мабуть, найпоширенішим є обчислення точності класифікації. Точність класифікації – це відсоток правильно класифікованих прикладів з усіх зроблених прогнозів. Алгоритм, який здатний вивчати прогнозну модель класифікації, називається алгоритмом класифікації (classification algorithm).

3.1. 2 Регресія (Regression)

Регресійне прогнознувальне моделювання (regression predictive modeling) — це завдання апроксимації функції відображення (f) від вхідних змінних (X) до безперервної вихідної змінної (y). Безперервна вихідна змінна — це реальне значення, наприклад ціле число або значення з плаваючою комою. Часто це величини, такі як суми та розміри. Оскільки модель з прогнозуванням регресії передбачає кількість, навички моделі повинні бути повідомлені як помилка в цих передбаченнях.

Алгоритм, який здатний вивчати модель з прогнозуванням регресії, називається алгоритмом регресії (regression algorithm).

3.1.3 Класифікація у порівнянні з регресією

Classification predictive modeling відрізняються від задач regression predictive modeling.

- Класифікація — це завдання прогнозування дискретної мітки класу.
- Регресія — це завдання прогнозування безперервної величини.

Існує деяке перекриття між алгоритмами класифікації та регресії; наприклад:

- Алгоритм класифікації може передбачити безперервне значення, але безперервне значення має форму ймовірності для мітки класу.
- Алгоритм регресії може передбачити дискретне значення, але дискретне значення у вигляді цілого числа.

Деякі алгоритми можна використовувати як для класифікації, так і для регресії з невеликими модифікаціями, наприклад, дерева рішень і штучні нейронні мережі. Деякі алгоритми не можуть або не можуть легко використовуватися для обох типів проблем, наприклад, лінійна регресія для прогнозного моделювання регресії та логістична регресія для моделювання з прогнозуванням класифікації [48].

Важливо те, що спосіб, яким ми оцінюємо прогнози класифікації та регресії, різниться і не перетинається, наприклад:

- Прогнози класифікації можна оцінити з точністю, тоді як прогнози регресії не можуть.
- Прогнози регресії можна оцінити за допомогою середньоквадратичної помилки, тоді як прогнози класифікації не можуть.

3.2 Алгоритми машинного навчання

- **Random Forest, RF** : Ансамбль дерев рішень. Одне дерево рішень розбиває ознаки вхідного вектора таким чином, що максимізує цільову функцію [49]. У алгоритмі RF отримані дерева декорелюються, оскільки вибір об’єктів для розгалужень вибирається випадковим чином (Рис 3.1)

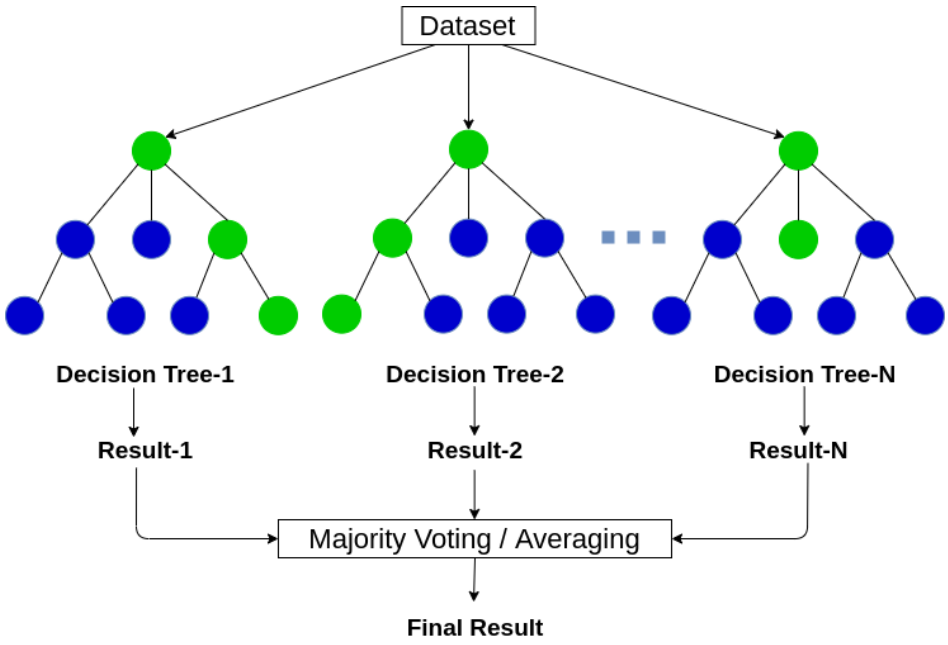

Рис 3.1: схема RF алгоритму [49].

- **Support Vector Machines, SVMs** : SVM можуть ефективно виконувати нелінійну класифікацію, використовуючи те, що називається kernel trick, неявно відображаючи свої вхідні дані у просторі багатовимірних функцій [50]. Класифікатор заснований на ідеї максимізації маржі (margin) як цільової функції (Рис. 3.2).

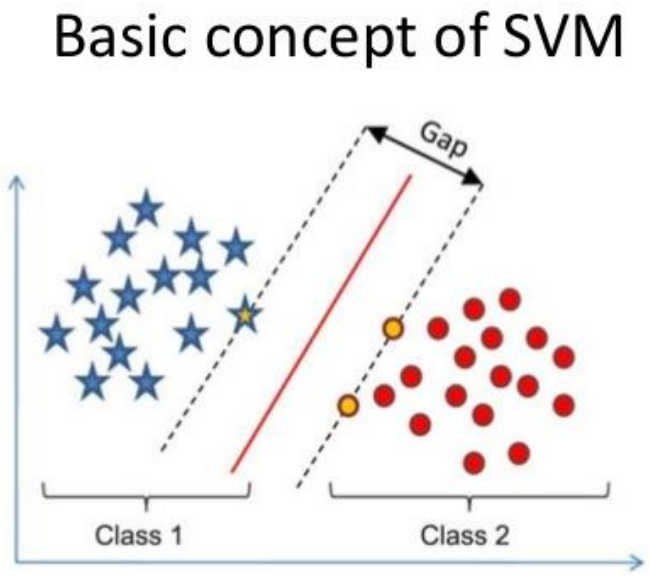

Рис. 3.2: схема SVM алгоритму [50].

- **Artificial neural networks, ANN** : ANN заснована на сукупності пов’язаних одиниць або вузлів, які називаються штучними нейронами, які вільно моделюють нейрони в біологічному мозку. Кожне з’єднання, як і синапси в біологічному мозку, може передавати сигнал від одного штучного нейрона до іншого [51]. Штучний нейрон, який отримує сигнал, може обробляти його, а потім сигналізувати додатковим штучним нейронам, підключеним до нього (Рис. 3.3)

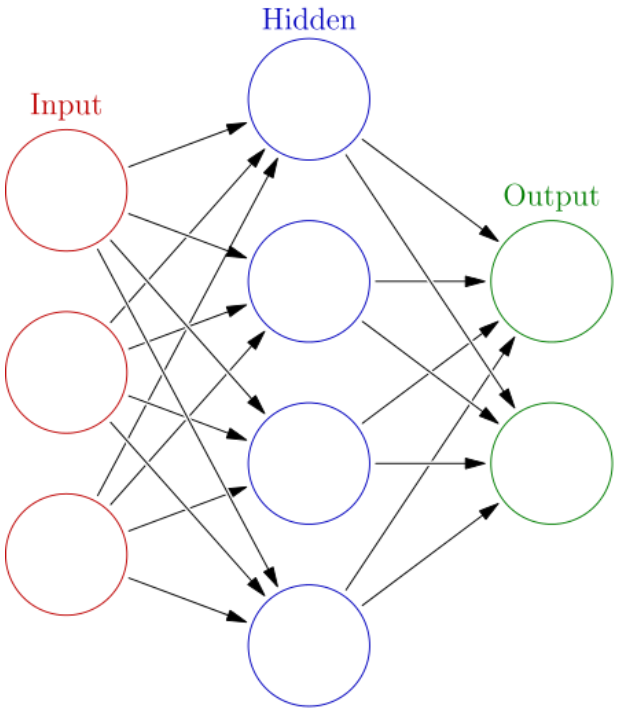

Рис. 3.3: схема ANN алгоритму [51].

3.3 Перевірка та оцінка моделі

Стратегія перевірки: K-кратна перехресна перевірка (K-fold cross validation):

Ця техніка перевірки моделі розбиває набір даних на дві групи в ітераційний спосіб:

- Навчальний набір даних: розглядається як відомий набір даних, на якому навчається модель
- Тестовий набір даних: невідомий набір даних, на якому потім тестується модель
- Процес повторюється k-раз

Мета полягає в тому, щоб перевірити здатність моделі передбачати дані, яких вона ніколи раніше не бачила, щоб позначити проблеми, відомі як перенавчання (over-fitting), і оцінити здатність моделі до узагальнення.

Матриця плутанини

У машинному навчанні та статистиці ми дуже часто використовуємо терміни справжнє позитивне (true positive, TP) та справжнє негативне(true negative, TN).

Позитивні та негативні – це не що інше, як лише два класи, наприклад, вижили/не вижили, рак/не рак, шахрайство з кредитними картками/не шахрайство, спам/не спам тощо. Це не тільки між двома трапляється/не трапляється, але також можна розділити на кішка/собака, самець/жінка. Отже, один клас ми вважаємо позитивним, а інший – негативним. Це довільно або залежить від мети дослідження, до якого ви сприймаєте одне як позитивне, а інше як негативне. У ньому немає хорошого (позитивного) чи поганого (негативного) аспекту.

Коли ми маємо вибіркові дані з деякої сукупності, і ми використовуємо моделювання , за допомогою якого ми можемо передбачити її клас/мітки. «True» представляє записи, які модель змогла ідентифікувати як свій клас, тоді як «false» представляє записи, які модель не змогла ідентифікувати.

Матриця плутанини — це таблиця, яка представляє підсумок результатів прогнозу щодо задачі класифікації. Нижче ми бачимо матрицю плутанини. Позиція передбачуваних значень і фактичних значень змінює положення хибно- негативних (FN) і хибно-позитивних (FP), але істинно

позитивні (TP) і істинно негативні (TN) залишаються на тому ж місці в матриці, розміщеній по діагоналі один до одного (Рис. 3.4).

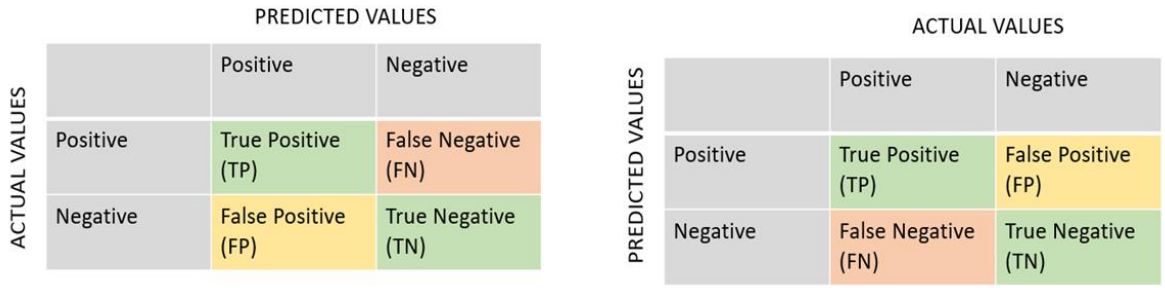

Рис. 3.4: Матриця плутанини

Цей спосіб допомагає зрозуміти продуктивність моделі. За формулами ми можемо обчислити ці показники, а для обчислення ми можемо використовувати матрицю плутанини.

Оцінка якості моделей

Для статистичної оцінки якості та поведінки наших моделей використовувалися дані параметри:

- **Чутливість (Sensitivity)** , також true positive rate
  - TPR = TP/(FN + TP)
  - *Інтуїтивно*: скільки з усіх фактичних позитивних результатів було передбачено як позитивне?
- **Специфічність (Specificity)**, також true negative rate
  - TNR = TN/(FP + TN)
  - *Інтуїтивно*: скільки з усіх фактичних негативних значень було передбачено негативними?
- **Точність (Accuracy)** , а також trueness
  - ACC = (TP + TN)/(TP + TN + FP + FN)

- *Інтуїтивно*: частка правильних прогнозів.
- **ROC-крива** , крива receiver operating characteristic curve Графік, який ілюструє діагностичну здатність нашого класифікатора
- Показує чутливість проти специфічності (sensitivity against the specificity)

- **AUC** , площа під кривою ROC (the area under the ROC curve):
  - Описує ймовірність того, що класифікатор оцінить випадково вибраний позитивний приклад вище, ніж негативний.
  - Значення від 0 до 1, чим вище, тим краще
- **MCC** (коефіцієнт кореляції Метьюза)
  - міра якості бінарних і багатокласових класифікацій. Він враховує істинні та хибні позитивні та негативні значення і, як правило, розглядається як збалансований показник, який можна використовувати, навіть якщо класи дуже різного розміру.
  - діапазон значень MCC лежить від -1 до **+1** . Модель з оцінкою +1 - ідеальна модель, а -1 – погана

Розділ 4 Обговорення результатів

Завдяки більшим доступним джерелам даних, машинне навчання (ML) набуло обертів у відкритті ліків і особливо у віртуальному скринінгу на основі лігандів. У даній роботі ми націлилися на використання різних алгоритмів ML для прогнозування активності нових сполук проти нашої мішені – A2BAR.

Створення вихідних даних про сполуки та їхню активність.

Першим кроком було завантаження наших даних, що сфокусовані на аденозинових рецепторах (A1, A2A, A2B та A3). Файл csv з Papyrus dataset [52] було завантажено у набір вихідних даних (датафрейм)з наступними стовбцями:

- SMILES: структура відповідної сполуки
- pchembl_value_Mea: виміряна афінність
- Year : рік, коли дана сполука була описана
- type_Ki, type_KD, type_IC50, type_EC50: як додаткові показники ефективності молекул

Одержаний датафрейм наведено на рисунку (Рис. 4.1) .

	canonical_SMILES	Year	pchembl_value_Mean	type_Ki	type_KD	type_IC50	type_EC50
10	CCCCn1c(=O)[nH]c2[nH]c(-c3ccc(OCC(=O)N4CCN(Cc5...	2002.0	8.89	1	0	0	0
18	N#Cc1c(-c2ccccc2)cc(-c2ccco2)nc1N	2008.0	6.77	1	0	0	0
20	CCCn1c(=O)c2nc(-c3ccccc3)[nH]c2n(CCCOC)c1=O	2009.0	6.49	1	0	0	0
23	Nc1nc(N)n2nc(-c3ccco3)nc2n1	2011.0	4.96	0	0	1	0
25	CCNC(=O)C1OC(n2cnc3c(NCC)nc(C#CCCCc4ccccc4)nc3...	2006.0	4.62	1	0	0	0
...	...	...	...	...	...	...	...
11734	CCCn1c(=O)c2nc(-c3cc(OCC(=O)Nc4ccc(OC)c(OC)c4)...	2004.0	7.29	1	0	0	0
11737	CCOC(=O)c1cnc(NCC(C)C)n2nc(-c3ccco3)nc12	2014.0	5.15	1	0	0	0
11740	COc1cc(-c2cc3c([nH]2)c(=O)n(C)c(=O)n3C)ccc1OCC...	2006.0	6.57	1	0	0	0
11742	CCCn1cc2c(nc(NC(=O)Nc3ccccc3OC)n3nc(-c4ccco4)n...	2002.0	6.78	1	0	0	0
11745	CCCn1c(=O)c2[nH]c(-c3ccc(OCC(=O)Nc4ccc(F)cc4)c...	2006.0	8.48	1	0	0	0

Рис. 4.1: Вихідний датафрейм

Наступним кроком було створення додаткової колонки «активність». В нашій роботі для сортування молекул на активні або неактивні ми використовували експериментально виміряну афінність цих молекул. Тобто якщо значення афінності більше або дорівнює 6.3 – ця молекула активна (active = 1). В усіх інших випадках – неактивна (active = 0) (Рис. 4.2).

	canonical_SMILES	Year	pchembl_value_Mean	type_Ki	type_KD	type_IC50	type_EC50	active
10	CCCCn1c(=O)[nH]c2[nH]c(-c3ccc(OCC(=O)N4CCN(Cc5...	2002.0	8.89	1	0	0	0	1.0
18	N#Cc1c(-c2ccccc2)cc(-c2ccco2)nc1N	2008.0	6.77	1	0	0	0	1.0
20	CCCn1c(=O)c2nc(-c3ccccc3)[nH]c2n(CCCOC)c1=O	2009.0	6.49	1	0	0	0	0.0
23	Nc1nc(N)n2nc(-c3ccco3)nc2n1	2011.0	4.96	0	0	1	0	0.0
25	CCNC(=O)C1OC(n2cnc3c(NCC)nc(C#CCCCc4ccccc4)nc3...	2006.0	4.62	1	0	0	0	0.0

Рис. 4.2: Датафрейм з доданою графою «активність»

В результаті такого сортування, ми одержали таку інформацію:

Кількість активних речовин: 1071

Кількість неактивних сполук: 600

Кодування молекул

Для машинного навчання молекули потрібно перетворити в список функцій, так званий molecular fingerprint (fp). FP - це спосіб кодування хімічної структури молекули (в нашому випадку SMILES) та представлення її у вигляді, який розуміє комп’ютер. Конвертація була реалізована використанням алгоритмів бібліотеки rdkit **(**Morgan fingerprints та ECFP).

Для одержаного відбитка було створено окрему колонку «fp» (Рис. 4.3)

	canonical_SMILES	Year	pchembl_value_Mean	type_Ki	type_KD	type_IC50	type_EC50	active	fp
10	CCCCn1c(=O)[nH]c2[nH]c(-c3ccc(OCC(=O)N4CCN(Cc5...	2002.0	8.89	1	0	0	0	1.0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
18	N#Cc1c(-c2ccccc2)cc(-c2ccco2)nc1N	2008.0	6.77	1	0	0	0	1.0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
20	CCCn1c(=O)c2nc(-c3ccccc3)[nH]c2n(CCCOC)c1=O	2009.0	6.49	1	0	0	0	0.0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...

Рис. 4.3: Датафрейм з доданим fp.

Щоб підвищити точність наших моделей, було вирішено створити ще один, додатковий new_fp, що включає у себе додаткову інформацію про властивості молекул і включає в себе дані стовпців type_Ki, type_KD, type_IC50, type_EC50 (Рис 4.4).

	canonical_SMILES	Year	pchembl_value_Mean	type_Ki	type_KD	type_IC50	type_EC50	active	fp	new_fp
10	CCCCn1c(=O)[nH]c2[nH]c(-c3ccc(OCC(=O)N4CCN(Cc5...	2002.0	8.89	1	0	0	0	1.0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
18	N#Cc1c(-c2ccccc2)cc(-c2ccco2)nc1N	2008.0	6.77	1	0	0	0	1.0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
20	CCCn1c(=O)c2nc(-c3ccccc3)[nH]c2n(CCCOC)c1=O	2009.0	6.49	1	0	0	0	0.0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
23	Nc1nc(N)n2nc(-c3ccco3)nc2n1	2011.0	4.96	0	0	1	0	0.0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
25	CCNC(=O)C1OC(n2cnc3c(NCC)nc(C#CCCCc4ccccc4)nc3...	2006.0	4.62	1	0	0	0	0.0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...

Рис 4.4: Датафрейм з додатаковим new_fp

В даній роботі ми використовували такі алгоритми машинного навчання:

- Random Forest (RF)
- Support Vector Machine (SVM)

- Artificial Neural Network (ANN)

Основна задача полягає в тому, щоб перевірити здатність моделі передбачати дані, які вона ніколи раніше не бачила та уникнути проблем, таких як перенавчання (over fitting), і оцінити здатність одержаної моделі до генералізації.

Для цього ми використовували функції model_training_and_validation, яка відповідає моделі з заданим розподілом даних (так званий temporal split) і повертає такі показники, як точність, чутливість, специфічність і AUC (accuracy, sensitivity, specificity and AUC відвовідно), що оцінюються на тестовому сеті. (Рис. 4.5)

В нашому випадку розділення відбувалося таким чином:

Тестовий сет(test set): сполуки після 2015 року.

Тренувальний сет(training set): сполуки до 2015 року.

	RF temporal	SVM temporal	ANN temporal
A2B	Sensitivity: 0.77	Sensitivity: 0.79	
	Specificity: 0.78	Specificity: 0.74	
	AUC: 0.88	AUC: 0.86	
	updated dataset		
	Sensitivity: 0.75	Mean accuracy: 0.84	Sensitivity: 0.54
	Specificity: 0.81	Mean sensitivity: 0.92	Specificity: 0.81
	AUC: 0.88	Mean specificity: 0.67	AUC: 0.80
		Mean AUC: 0.91	
	with additional fp		
	Sensitivity: 0.74	Mean accuracy: 0.82	Sensitivity: 0.74
	Specificity: 0.67	Mean sensitivity: 0.89	Specificity: 0.53
	AUC: 0.82	Mean specificity: 0.69	AUC: 0.66
		Mean AUC: 0.88	

Рис. 4.5: Результат моделей RF, SVM та ANN з використанням заданого розподілу даних (temporal split).

На наступному етапі ми визначаємо функцію cross_validation, яка виконує процедуру кросвалідації 5 разів та виводить статистичні результати, що описують якість наших моделей. (Рис. 4.6)

	RF-cv	SVM-cv	ANN temporal
A2B	Mean accuracy:	Mean accuracy: 0.83	
	Mean sensitivity:	Mean sensitivity: 0.93	
	Mean specificity: 0.67	Mean specificity: 0.63	
	Mean AUC: 0.90	Mean AUC: 0.90	
	updated dataset		
	Mean accuracy: 0.84	Mean accuracy: 0.84	Sensitivity: 0.54
	Mean sensitivity: 0.91	Mean sensitivity: 0.92	Specificity: 0.81
	Mean specificity: 0.68	Mean specificity: 0.67	AUC: 0.80
	Mean AUC: 0.90	Mean AUC: 0.91	
	with additional fp		
	Mean accuracy: 0.82	Mean accuracy: 0.82	Sensitivity: 0.74
	Mean sensitivity: 0.88	Mean sensitivity: 0.89	Specificity: 0.53
	Mean specificity: 0.70	Mean specificity: 0.69	AUC: 0.66
	Mean AUC: 0.89	Mean AUC: 0.88	

Рис. 4.6: Результат моделей RF, SVM та ANN з використанням кросвалідації(cv)

Нажаль, як видно з рис. 4.5 та рис. 4.6, додавання нового new_fp, котрий включає у себе дані про властивості молекул type_Ki, type_KD, type_IC50, type_EC50 значно не покращив характеристики наших моделей.

Тому було вирішено оновити та розширити наш набір даних за допомогою додавання до нього так званої low quality data з Papyrus dataset, який містить близько 60 мільйонів записів [52]. Він складається з кількох великих загальнодоступних наборів даних, таких як ChEMBL і ExCAPE-DB, у поєднанні з кількома меншими наборами даних. Нові 4 моделі знову

пройшли навчання. Крім того, було розраховано коефіцієнт кореляції Метьюза (MCC).

Результати розрахунків якості наших моделей після оновлення датасету продемонстровано на Рис. 4.7

	RF cv	RF temporal	SVM cv	SVM temporal	ANN temporal	ANN cv
	low quality data added					
A2B	Mean accuracy: 0.87		Mean accuracy: 0.86			Mean accuracy: 0.82
	Mean sensitivity: 0.84	Sensitivity: 0.74	Mean sensitivity: 0.84	Sensitivity: 0.51	Sensitivity: 0.64	Mean sensitivity: 0.79
	Mean specificity: 0.89	Specificity: 0.67	Mean specificity: 0.88	Specificity: 0.83	Specificity: 0.71	Mean specificity: 0.85
	Mean AUC: 0.94	AUC: 0.82	Mean AUC: 0.93	AUC: 0.75	AUC: 0.72	Mean AUC: 0.90
	MCC: 0.76	MCC: 0.39	MCC: 0.67	MCC: 0.34	MCC: 0.32	MCC: 0.72

Рис. 4.7: Результати розрахунків після оновлення датасету

Отже, ми порівняли кількісні характеристики наших 22 моделей (чутливість, специфічність, точність, площа під кривою ROC та MCC). Табличні дані усіх QSAR моделей (Рис. 4.8) демонструють, що результати моделі з використанням оновленого та розширеного набору даних, що включає low quality data, методу машинного навчання RF та кросвалідації (CV) показали кращий результат в порівнянні з іншими моделями, адже усі кількісні показники більше наближені до 1, в порівнянні з іншими моделями. Наступним кроком може бути використання цих моделей як середовища для генерації нових молекул

	RF cv	RF temporal	SVM cv	SVM temporal	ANN temporal	ANN cv
A2B	Mean accuracy:	Sensitivity: 0.77	Mean accuracy: 0.83	Sensitivity: 0.79		
	Mean sensitivity:	Specificity: 0.78	Mean sensitivity: 0.93	Specificity: 0.74		
	Mean specificity: 0.67	AUC: 0.88	Mean specificity: 0.63	AUC: 0.86		
	Mean AUC: 0.90		Mean AUC: 0.90			
	updated dataset					
	Mean accuracy: 0.84	Sensitivity: 0.75	Mean accuracy: 0.84	Sensitivity: 0.80	Sensitivity: 0.54	Mean accuracy: 0.82
	Mean sensitivity: 0.91	Specificity: 0.81	Mean sensitivity: 0.92	Specificity: 0.74	Specificity: 0.81	Mean sensitivity: 0.91
	Mean specificity: 0.68	AUC: 0.88	Mean specificity: 0.67	AUC: 0.86	AUC: 0.80	Mean specificity: 0.63
	Mean AUC: 0.90		Mean AUC: 0.91			Mean AUC: 0.86
	with additional fp					
	Mean accuracy: 0.82	Sensitivity: 0.74	Mean accuracy: 0.82	Sensitivity: 0.75	Sensitivity: 0.74	Mean accuracy: 0.79
	Mean sensitivity: 0.88	Specificity: 0.67	Mean sensitivity: 0.89	Specificity: 0.59	Specificity: 0.53	Mean sensitivity: 0.87
	Mean specificity: 0.70	AUC: 0.82	Mean specificity: 0.69	AUC: 0.80	AUC: 0.66	Mean specificity: 0.66
	Mean AUC: 0.89		Mean AUC: 0.88			Mean AUC: 0.85
	low quality data added					
	Mean accuracy: 0.87		Mean accuracy: 0.86			Mean accuracy: 0.82
	Mean sensitivity: 0.84	Sensitivity: 0.74	Mean sensitivity: 0.84	Sensitivity: 0.51	Sensitivity: 0.64	Mean sensitivity: 0.79
	Mean specificity: 0.89	Specificity: 0.67	Mean specificity: 0.88	Specificity: 0.83	Specificity: 0.71	Mean specificity: 0.85
	Mean AUC: 0.94	AUC: 0.82	Mean AUC: 0.93	AUC: 0.75	AUC: 0.72	Mean AUC: 0.90
	MCC: 0.76	MCC: 0.39	MCC: 0.67	MCC: 0.34	MCC: 0.32	MCC: 0.72

Рис. 4.8: Порівняльна таблиця усіх 22 QSAR моделей.

Для більш зручного порівняння моделей, нижче наведено дані у вигляді графіка на рис. 4.9.

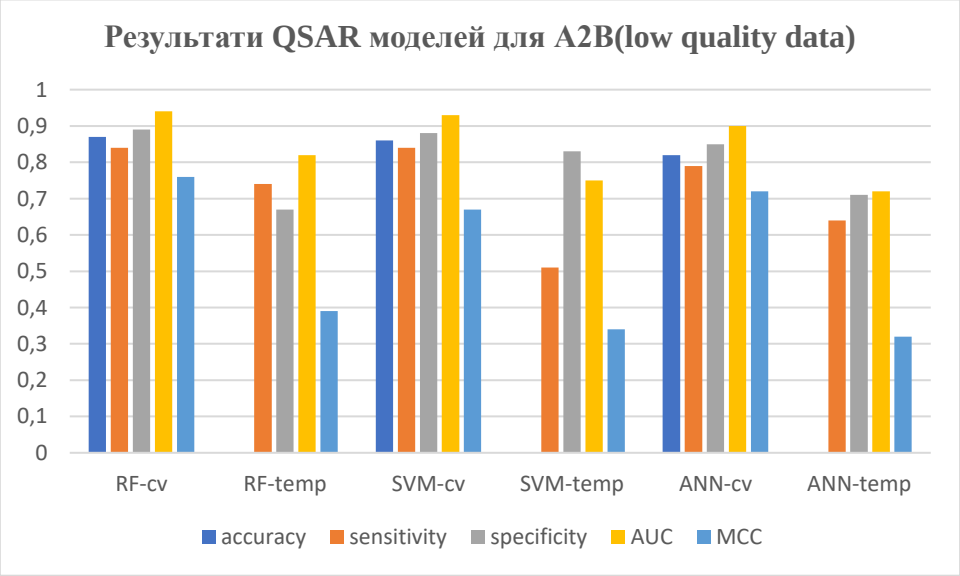

Рис. 4.9: Графік з результатами QSAR моделей для A2B (low quality data)

ВИСНОВКИ

Було успішно створено та обучено 22 моделі машинного навчання, спрямованих на активність хімічних сполук проти аденозин-рецептора A2B. Аналіз та порівняння кількісних характеристик показали, що моделі з використанням оновленого та розширеного набору даних, що включають low quality data, методу машинного навчання RF та кросвалідації (CV) мають найкращу передбачувальну здатність.

Побудовані QSAR моделі в подальшому можуть бути використані для генерації нових молекул: селективних лігандів аденозин-рецептора A2B.

Список використаних джерел

1. Lv, X., et al., In vitro expression and analysis of the 826 human G protein-coupled receptors.

Protein Cell, 2016. 7(5): p. 325-37.

2. Santos, R., et al., A comprehensive map of molecular drug targets. Nat Rev Drug Discov, 2017.p. 19-34

3. Saroz Y, Kho DT, Glass M, Graham ES, Grimsey NL (December 2019). "Cannabinoid Receptor 2 (CB2) Signals via G-alpha-s and Induces IL-6 and IL-10 Cytokine Secretion in Human Primary Leukocytes". ACS Pharmacology & Translational Science. 2 (6): 414–428.

4. Sharma N, Akhade AS, Qadri A (April 2013). "Sphingosine-1-phosphate suppresses TLR-induced CXCL8 secretion from human T cells". Journal of Leukocyte Biology. 93 (4): 521–8

5.Li J, Ning Y, Hedley W, Saunders B, Chen Y, Tindill N, Hannay T, Subramaniam S. The Molecule Pages database. Nature. 2002 Dec 12;420(6916):716-7

6. Brian T. Layden, M.D., Ph.D., Vivek Durai & William L. Lowe, Jr., M.D. (Division of Endocrinology, Metabolism and Molecular Medicine, Northwestern University) © 2010 Nature Education

7.Leurs, R., Bakker, R., Timmerman, H. et al. The histamine H3 receptor: from gene cloning to H3 receptor drugs. Nat Rev Drug Discov 4, 107–120 (2005).

8. Drury AN, Szent-Györgyi A. The physiological activity of adenine compounds with special reference to their action upon the mammalian heart. J Physiol 68: 213–237, 1929.

9. Borea PA, Gessi S, Merighi S, Varani K. Adenosine as a Multi-Signalling Guardian Angel in Human Diseases: When, Where and How Does it Exert its Protective Effects? Trends Pharmacol Sci 37: 419–434, 2016.

10. Manjunath S, Sakhare PM. Adenosine and adenosine receptors: Newer therapeutic perspective. Indian J Pharmacol. 2009

11. Borea PA, Gessi S, Merighi S, Vincenzi F, Varani K. Pharmacology of Adenosine Receptors: The State of the Art. Physiol Rev. 2018 Jul 1;98(3):1591-1625.

12. Newby AC. Adenosine and the concept of "retaliatory metabolites." Trends Biochem Sci 9: 42–44, 1984

13. Fredholm BB, Chen JF, Masino SA, Vaugeois JM. Actions of adenosine at its receptors in the CNS: insights from knockouts and drugs. Annu Rev Pharmacol Toxicol 2005

14. Kasama H, Sakamoto Y, Kasamatsu A, Okamoto A, Koyama T, Minakawa Y, Ogawara K, Yokoe H, Shiiba M, Tanzawa H, Uzawa K. Adenosine A2b receptor promotes progression of human oral cancer. BMC Cancer. 2015 Jul 31;15:563.

15. Gao ZG, Jacobson KA. A2B Adenosine Receptor and Cancer. Int J Mol Sci. 2019;20(20):5139. Published 2019 Oct 17. doi:10.3390/ijms20205139

16. 3. Seitz L., Jin L., Leleti M., Ashok D., Jeffrey J., Rieger A., Tiessen R.G., Arold G., Tan J.B.L., Powers J.P., et al. Safety, tolerability, and pharmacology of AB928, a novel dual adenosine receptor antagonist, in a randomized, phase 1 study in healthy volunteers. Invest. New Drugs. 2019

17. 4. Wei Q., Costanzi S., Balasubramanian R., Gao Z.G., Jacobson K.A. A2B adenosine receptor blockade inhibits growth of prostate cancer cells. Purinergic Signal. 2013;9:271–280.

18. Cekic C., Sag D., Li Y., Theodorescu D., Strieter R.M., Linden J. Adenosine A2B receptor blockade slows growth of bladder and breast tumors. J. Immunol. 2012;188:198–205

19. Kasama H., Sakamoto Y., Kasamatsu A., Okamoto A., Koyama T., Minakawa Y., Ogawara K., Yokoe H., Shiiba M., Tanzawa H., et al. Adenosine A2b receptor promotes progression of human oral cancer. Clin. Cancer Res. 2016;22:158–166.

20. Mittal D., Sinha D., Barkauskas D., Young A., Kalimutho M., Stannard K., Caramia F., Haibe-Kains B., Stagg J., Khanna K.K., et al. Adenosine 2B Receptor Expression on Cancer Cells Promotes Metastasis. Cancer Res. 2016;76:4372–4382.

21. Sepúlveda C., Palomo I., Fuentes E. Role of adenosine A2b receptor overexpression in tumor progression. Life Sci. 2016;166:92–99.

22. Ryzhov S., Novitskiy S.V., Zaynagetdinov R., Goldstein A.E., Carbone D.P., Biaggioni I., Dikov M.M., Feoktistov I. Host A2B adenosine receptors promote carcinoma growth. Neoplasia. 2008;10:987–995.

23. Iannone R., Miele L., Maiolino P., Pinto A., Morello S. Blockade of A2b adenosine receptor reduces tumor growth and immune suppression mediated by myeloid-derived suppressor cells in a mouse model of melanoma. Neoplasia. 2013;15:1400–1409.

24. Daly, J.W., Butts-Lamb, P. & Padgett, W. Subclasses of adenosine receptors in the central nervous system: Interaction with caffeine and related methylxanthines. Cell Mol Neurobiol 3, 69–80 (1983).

25. Borea P.A., Gessi S., Merighi S., Varani K. Adenosine as a Multi-Signalling Guardian Angel in Human Diseases: When, Where and How does it Exert its Protective Effects? Trends Pharmacol. Sci. 2016;37:419–434.

26. Borea P.A., Gessi S., Merighi S., Vincenzi F., Varani K. Pharmacology of Adenosine Receptors: The State of the Art. Physiol. Rev. 2018;98:1591–1625.

27. Cekic C., Linden J. Purinergic regulation of the immune system. Nat. Rev. Immunol. 2016;16:177–192.

28. Leach AR, Harren J (2007). Structure-based Drug Discovery. Berlin: Springer. ISBN 978-1-4020-4406-9.

29. Mauser H, Guba W (May 2008). "Recent developments in de novo design and scaffold hopping". Current Opinion in Drug Discovery & Development. 11 (3): 365–74.

30. Klebe G (2000). "Recent developments in structure-based drug design". Journal of Molecular Medicine. 78 (5): 269–81

31. Wang R, Gao Y, Lai L (2000). "LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design". Journal of Molecular Modeling. 6 (7–8): 498–516.

32. Schneider G, Fechner U (Aug 2005). "Computer-based de novo design of drug-like molecules". Nature Reviews. Drug Discovery.

33. Jorgensen WL (Mar 2004). "The many roles of computation in drug discovery". Science. 303 (5665): 1813–8.

34. McConkey BJ, Sobolev V, Edelman M. The performance of current methods in ligand-protein docking. Current Science. 2002

35. Meng XY, Zhang HX, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des. 2011;7(2):146-157. doi:10.2174/157340911795677602

36. Lu W, Zhang R, Jiang H, Zhang H, Luo C. Computer-Aided Drug Design in Epigenetics. Front Chem. 2018;6:57. Published 2018 Mar 12.

37. Akamatsu M. Current State and Perspectives of 3D-QSAR. Curr. Top. Med. Chem. 2002

38. Verma RP, Hansch C. Camptothecins: A SAR/QSAR Study. Chem. Rev. 2009;109:213–235.

39. Acharya, Chayan et al. “Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach.” Current computer-aided drug design vol. 7,1 (2011): 10-22. doi:10.2174/157340911793743547

40. Lu W, Zhang R, Jiang H, Zhang H, Luo C. Computer-Aided Drug Design in Epigenetics. Front Chem. 2018 Mar 12;6:57.

41. Ramesh A. Artificial intelligence in medicine. Ann. R. Coll. Surg. Engl. 2004;86:334–338

42. Beneke F., Mackenrodt M.-O. Artificial intelligence and collusion. IIC Int. Rev. Intellectual Property Competition Law. 2019.

43. Steels L., Brooks R. Routledge; 2018. The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents.

44. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021;26(1):80-93.

45. Vamathevan, J., Clark, D., Czodrowski, P. et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18, 463–477 (2019).

46. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018 Jun;23(6):1241-1250.

47. Lima AN, Philot EA, Trossini GH, Scott LP, Maltarollo VG, Honorio KM. Use of machine learning approaches for novel drug discovery. Expert Opin Drug Discov. 2016

48. Breiman. Manual on setting up, using, and understanding random forests v3.1, 2002.

49. Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).

50. Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995).

51. Van Gerven, Marcel, and Sander Bohte. "Artificial neural networks as models of neural information processing." Frontiers in Computational Neuroscience 11 (2017): 114.

52. Olivier J. M. Béquignon, Brandon J. Bongers . ‘’ Papyrus - A large scale curated dataset aimed at bioactivity predictions’’10.26434/chemrxiv-2021-1rxhk