

УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ



Лазар Савић

РАЗВОЈ СОФТВЕРА ЗА ОТКРИВАЊЕ МОЛЕКУЛСКИХ СТРУКТУРА СА ПОТЕНЦИЈАЛНО ЛЕКОВИТИМ ДЕЈСТВОМ

мастер рад

Београд, 2026.

Ментор:

проф. др Јована КОВАЧЕВИЋ, ванредни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

др Мирјана МАЉКОВИЋ РУЖИЧИЋ, доцент
Универзитет у Београду, Математички факултет

др Стефан КАПУНАЦ, асистент
Универзитет у Београду, Математички факултет

Датум одбране: _____

Порогици

Наслов мастер рада: Развој софтвера за откривање молекулских структура са потенцијално лековитим дејством

Резиме: Развој нових лекова представља сложен процес који захтева значајне временске и финансијске ресурсе. Примена метода рачунарске интелигенције омогућава аутоматизацију појединих корака процеса откривања лекова и ефикасније претраживање великог простора потенцијалних молекулских структура.

У овом раду представљен је развој конфигурабилног софтверског система за генерисање, евалуацију и оптимизацију молекулских структура са потенцијално лековитим дејством. Као основни приступ за претраживање простора молекулских структура коришћен је генетски алгоритам, уз подршку различитих репрезентација молекула и генетских оператора прилагођених њиховим карактеристикама. За процену квалитета структура коришћене су фитнес функције засноване на QED коефицијенту и предиктивним моделима за процену биолошке активности молекула.

Предиктивни модели обучени су на подацима из базе ChEMBL за процену pIC₅₀ вредности EGFR инхибитора, при чему је за поделу података коришћен приступ заснован на молекулским скелетима. Експериментално је испитан утицај величине популације, избора генетских оператора и фитнес функције на понашање генетског алгоритма, квалитет добијених структура и време извршавања.

Развијени систем омогућава аутоматизовано генерисање и оптимизацију молекулских структура и представља основу за примену генетских алгоритама у рачунарски потпомогнутом откривању лекова. Добијени резултати показују да избор конфигурације генетског алгоритма значајно утиче на однос квалитета добијених структура и рачунарске захтевности процеса оптимизације.

Кључне речи: генетски алгоритми, рачунарски потпомогнуто откривање лекова, молекулске репрезентације, предвиђање pIC₅₀ вредности, EGFR инхибитори

Садржај

1	Увод	1
2	Репрезентација молекула у програмима	3
2.1	SMILES нотација	3
2.2	SELFIES нотација	5
2.3	Објекат Mol библиотеке RDKit	5
2.4	<i>Morgan fingerprint</i> вектори	6
3	Процена лековитости и биолошке активности	8
3.1	QED коефицијент	8
3.2	Регресиони предиктивни модели	10
4	Генетски алгоритам	17
4.1	Селекција	18
4.2	Укрштање	19
4.3	Мутација	25
5	Имплементација и архитектура система	30
5.1	Архитектура система	31
5.2	Кориснички интерфејс	32
6	Експериментални резултати	39
6.1	Основна експериментална поставка	40
6.2	Утицај величине популације	40
6.3	Поређење генетских оператора	42
6.4	Поређење фитнес функција	47
7	Закључак	50

Глава 1

Увод

Откривање нових лекова захтева испитивање огромног простора хемијских структура, што традиционалним експерименталним приступима представља дуготрајан и скуп процес. Методе рачунарске интелигенције омогућавају да се део тог процеса аутоматизује и да се потенцијално перспективне структуре идентификују пре њиховог експерименталног испитивања. Међу таквим приступима, генетски алгоритми су погодни за претраживање дискретног и комбинаторно огромног простора молекула, јер истовремено омогућавају истраживање нових структура и усмеравање претраге ка решењима бољег квалитета.

Циљ овог рада је развој конфигурабилног софтверског система за генерисање, евалуацију и оптимизацију молекулских структура са потенцијално лековитим дејством. Систем омогућава примену генетског алгоритма за аутоматизовано претраживање простора молекулских структура, при чему су подржани различити начини представљања молекула и генетски оператори прилагођени њиховим карактеристикама. Квалитет генерисаних структура може се процењивати на основу њихових молекулских својстава или помоћу посебно обучених предиктивних модела, који предвиђају биолошку активност молекула. Понашање генетског алгоритма испитано је кроз експерименте који подразумевају утицај различитих конфигурација генетског алгоритма, укључујући величину популације, избор генетских оператора и фитнес функције.

Рад је организован на следећи начин. У поглављу 2 описане су репрезентације молекула коришћене у систему. Поглавље 3 посвећено је методама процене квалитета молекулских структура, укључујући QED коефицијент и регресионе предиктивне моделе. У поглављу 4 представљени су генетски алго-

ритам и имплементирани оператори селекције, укрштања и мутације. Поглавље 5 описује архитектуру система и кориснички интерфејс. Експериментална евалуација и добијени резултати приказани су у поглављу 6, док су закључци и правци даљег рада изнети у поглављу 7.

Глава 2

Репрезентација молекула у програмима

Молекулске структуре могу се представити на више различитих начина, при чему свака репрезентација наглашава одређене аспекте молекула и прилагођена је конкретној намени. Док су неке репрезентације погодне за једноставно записивање и размену података о молекулима, друге омогућавају њихову рачунарску обраду, анализу или примену у алгоритмима машинског учења. Избор одговарајуће репрезентације зависи од проблема који се решава и операција које је потребно извршити над молекулским структурама.

У овом раду коришћене су четири међусобно повезане репрезентације молекула. За текстуални запис молекулских структура коришћене су SMILES нотација (одељак 2.1) и SELFIES нотација (одељак 2.2), за њихову интерну графовску репрезентацију и манипулацију објекат Mol библиотеке RDKit (одељак 2.3), док су за потребе обучавања модела машинског учења и предикције биолошке активности молекули представљени помоћу *Morgan fingerprint* вектора (одељак 2.4). У наставку су описане основне карактеристике сваке од наведених репрезентација и њихова улога у оквиру развијеног система.

2.1 SMILES нотација

SMILES (енг. *Simplified Molecular Input Line Entry System*) [10] представља једноставан и компактан текстуални начин представљања молекулских структура у облику ниске ASCII карактера. Због своје читљивости и мале величине записа, широко се користи за складиштење, размену и обраду по-

датака о молекулима. Иако је реч о једнодимензионом текстуалном запису, SMILES омогућава представљање структуре молекула, укључујући атоме, хемијске везе, гранања и цикличне структуре, а по потреби и информације о стереохемији. Заснива се на неколико једноставних правила записа, која ће у наставку бити описана и илустрована примерима датим у табели 2.1.

Правило 1 - Атоми: Атоми који улазе у састав молекула представљени су својим хемијским симболима, у угластим заградама. Ове заграде се изостављају за „органски” подскуп атома - В, С, N, О, Р, S, F, Cl, Br и I. Ако су заграде изостављене, подразумеван је имплицитан број водоникових атома који задовољава стандардну валенцу атома, те се у тим случајевима H атоми изостављају. Јони се увек пишу у заградама, попут $[Co^{3+}]$ или $[OH^{-}]$.

Правило 2 - Везе: Везе између атома представљају се на различите начине, у зависности од њихове вишеструкости. Везе између суседних атома су подразумевано једноструке, и тада се оне не наводе. Двоструке везе представљене су симболом '=', а троструке са '#'. Ароматичне везе могу се експлицитно представити симболом ':'.

Правило 3 - Гранања: У случају разгранатих молекула, грана која се издваја из родитељског ланца је одвојена обичним заградама '()'

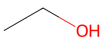
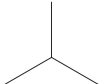
Правило 4 - Циклуси: Циклични молекули се такође записују линеарно, тако што се позиције на којима се прстен отвара и затвара обележе истим бројем. У случају ароматичних прстена, сви атоми који учествују у том прстену обележавају се малим словима, како би се разликовали од одговарајућих алифатичних прстена.

Правило 5 - Неповезане структуре: Уколико се молекул састоји од више неповезаних компоненти, он се може представити тако што се на сваку компоненту примене правила 1 - 4, и добијени записи се повежу симболом '.'.

Важно је напоменути да SMILES запис није јединствен за један молекул. Иста молекулска структура може имати више различитих SMILES записа, у зависности од избора почетног атома и редоследа обиласка структуре. Како би се обезбедила јединствена и стандардизована репрезентација молекула, често се користи канонски SMILES запис (енг. *canonical SMILES*), који алгоритамски одређује један представник за дату молекулску структуру.

2.2 SELFIES нотација

SELFIES (енг. *SELF-Referencing Embedded Strings*) [6] представља начин текстуалне репрезентације молекулских структура, сличан SMILES нотацији. За разлику од SMILES-а, код којег поједини карактери не представљају увек засебне структурне елементе, SELFIES је организован као низ јасно раздвојених токена, при чему се сваки токен записује у угластим заградама. Токени могу представљати атоме, хемијске везе и друге елементе неопходне за опис молекулске структуре. Свака синтаксно исправна SELFIES ниска одговара валидној молекулској структури, и свака молекулска структура може да се представи SELFIES ниском. Ово је значајна предност SELFIES репрезентације, јер се тиме избегавају многи случајеви невалидних структура који се могу јавити приликом директне манипулације SMILES записима. У табели 2.1 дати су примери молекулских структура и одговарајућих SELFIES записа.

Молекул	Структура	SMILES	SELFIES
Етанол		<chem>CCO</chem>	[C] [C] [O]
Етен		<chem>C=C</chem>	[C] [=C]
Изобутан		<chem>CC(C)C</chem>	[C] [C] [Branch1] [C] [C] [C]
Бензен		<chem>c1ccccc1</chem>	[C] [=C] [C] [=C] [C] [=C] [Ring1] [=Branch1]

Табела 2.1: Примери молекулских структура и одговарајућих SMILES и SELFIES записа

2.3 Објекат *Mol* библиотеке *RDKit*

RDKit (енг. *RDKit Cheminformatics Toolkit*) је библиотека отвореног кода за рад са молекулским структурама [7]. Имплементирана је у C++-у, са *Python* интерфејсом, и широко се користи у областима попут хемоинформатике

и машинског учења¹. За разлику од текстуалних формата попут SMILES-a, RDKit пружа богат интерфејс за рад са молекулима, кроз објекат Mol. Објекат Mol интерно представља молекулску структуру као граф чији су чворови атоми, а гране хемијске везе, уз додатне информације о особинама веза и атома. Оваква репрезентација омогућава ефикасно извршавање различитих операција над молекулима, као што су валидација структуре, израчунавање молекулских дескриптора, генерисање *Morgan fingerprint* вектора, али и визуализација структура у корисничком интерфејсу.

2.4 *Morgan fingerprint* вектори

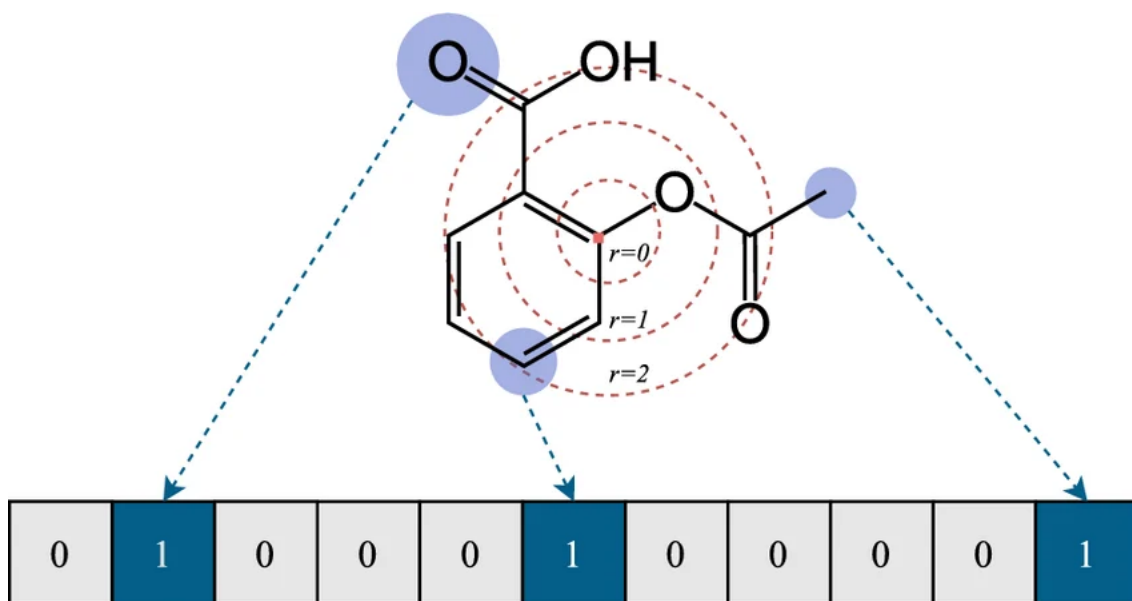
Morgan fingerprint [9] представља нумеричку репрезентацију молекулске структуре у облику бит-вектора фиксне дужине, погодну за примену алгоритама машинског учења. За разлику од SMILES записа и RDKit Mol објекта, који чувају информације о структури молекула, *Morgan fingerprint* издваја присуство одређених локалних структурних карактеристика молекула и кодира их у бинарни вектор. На тај начин сваки бит представља присуство или одсуство одређеног структурног мотива, што омогућава једноставну примену како математичких метода, тако и метода машинског учења над молекулским подацима (слика 2.1).

Генерисање *Morgan fingerprint* вектора заснива се на итеративном обиласку околине сваког атома у молекулу. За сваки атом посматрају се суседни атоми до задатог растојања, дефинисаног параметром *radius*, након чега се добијене локалне структуре хеширају и пресликавају у позиције бит-вектора. Избор вредности овог параметра представља компромис између детаљности и генерализације молекулске репрезентације. Мање вредности описују локалне карактеристике блиске појединачним атомима, док веће омогућавају кодирање сложенијих структурних фрагмената који обухватају већи део молекула.

Оваква репрезентација погодна је за квантификовање мере сличности два молекула. За поређење два молекулска отиска користи се Танимото коефицијент сличности (енг. *Tanimoto coefficient*) [11]. За бинарне отиске молекула *A* и *B* он се може дефинисати као

$$T_{\text{Tanimoto}}(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

¹Поготово у комбинацији ове две области - рачунарски потпомогнуто дизајнирању лекова (енг. *Computer-Aided Drug Design, CADD*), што је и суштина овог рада.



Слика 2.1: Пример *Morgan fingerprint* репрезентације (аспирин): бит-вектор и подструктуре за различите радијусе. Преузето из [8].

где $|A \cap B|$ представља број заједничких молекулских образаца, док $|A \cup B|$ представља број образаца који се појављују у бар једном од посматраних молекула. Пошто посматрамо репрезентацију молекула у виду бит-вектора, ову вредност можемо посматрати и на следећи начин:

$$T_{\text{Tanimoto}}(A, B) = \frac{c}{a + b - c},$$

где c представља број позиција на којима оба бит-вектора имају вредност 1, a број јединица у бит-вектору молекула A , а b број јединица у бит-вектору молекула B . Вредност коефицијента налази се у интервалу $[0, 1]$. Већа вредност овог коефицијента указује на већу сличност молекула. Његов значај у контексту генетског алгорита биће описан у поглављу 5.2.

Глава 3

Процена лековитости и биолошке активности

У оквиру развијеног система потребно је проценити лековитост, односно потенцијалну биолошку активност генерисаних молекулских структура, како би се омогућило њихово рангирање у оквиру генетског алгоритма. Пошто експериментално одређивање потентности сваке новогенерисане структуре није могуће у току процеса оптимизације, искоришћени су другачији приступи који представљају брзе рачунарске апроксимације експерименталних мерења. Са једне стране, то је QED коефицијент (одељак 3.1), а са друге стране, регресиони модели машинског учења (одељак 3.2). У оба случаја, метрика коју рачунамо може да се користи као вредност фитнес функције, која се оптимизује у току генетског алгоритма.

3.1 QED коефицијент

QED (енг. *Quantitative Estimate of Drug-likeness*) развили су Ричард Бикертон и сарадници [3]. Основу ове метрике чине физичко-хемијске карактеристике молекула (тј. *молекулски дескриптори*):

- моларна маса (енг. *molecular weight*, MW),
- октанол-вода партициони коефицијент (енг. *octanol-water partition coefficient*, ALOGP),
- број донора водоничних веза (енг. *hydrogen bond donors*, HBD),

- број акцептора водоничних веза (енг. *hydrogen bond acceptors*, HBA),
- поларна молекулска површина (енг. *molecular polar surface area*, PSA),
- број ротабилних веза (енг. *rotatable bonds*, ROTB),
- број ароматичних прстенова (енг. *aromatic rings*, AROM),
- број структурних упозорења (енг. *structural alerts*, ALERTS).

Рачунањем вредности ових карактеристика, одређује се у којој мери молекул поседује карактеристике типичне за успешне лекове. За сваки дескриптор, дефинисана је функција пожељности (енг. *desirability function*), која пресликава вредност дескриптора у интервал $[0, 1]$. Општи облик те функције дат је једначином (3.1).

$$d(x) = a + \frac{b}{\left[1 + \exp\left(-\frac{x-c+\frac{d}{2}}{e}\right)\right]} \cdot \left[1 - \frac{1}{\left[1 + \exp\left(-\frac{x-c-\frac{d}{2}}{f}\right)\right]}\right]. \quad (3.1)$$

Параметри a , b , c , d , e и f за сваку од функција d_{MW} , d_{ALOGP} , d_{HBD} , d_{HBA} , d_{PSA} , d_{ROTB} , d_{AROM} и d_{ALERTS} дати су у оригиналном раду [3].

Описани молекуларни дескриптори могу имати различиту важност при одређивању QED коефицијента, па се он може израчунати као *тежинска геометријска средина* вредности функција молекуларних дескриптора, што се може описати следећом једначином:

$$\text{QED} = \exp\left(\frac{\sum_{i=1}^n w_i \ln(d_i)}{\sum_{i=1}^n w_i}\right). \quad (3.2)$$

Библиотека RDKit пружа могућност рачунања QED коефицијента за молекулске структуре, при чему тежински коефицијенти w_i из једначине (3.2), уколико корисник не зада другачије, имају подразумеване вредности (према оригиналном раду [3]). Имплементирани систем омогућава ручно подешавање ових коефицијената, што ће бити описано у одељку 5.2.

Добијена вредност QED коефицијента припада интервалу $[0, 1]$, при чему веће вредности указују на већу сличност молекула са познатим лековима, и самим тим, на већу погодност молекула са аспекта општих физичко-хемијских својстава.

3.2 Регресиони предиктивни модели

Основна вредност коју обучени модели у овом раду предвиђају је IC50 (*half maximal inhibitory concentration*), која представља концентрацију супстанце потребну за постизање 50% инхибиције одређеног биолошког процеса. Ниже вредности IC50 указују на већу потентност молекула, па се у циљу погодније примене у моделима машинског учења често користи трансформисана вредност pIC50, при чему је $pIC50 = -\log_{10}(IC50)$.

За потребе обучавања модела коришћени су подаци из јавно доступне базе биоактивних молекула ChEMBL. Изабране су молекулске структуре за које постоје експериментално одређене вредности активности према циљном протеину EGFR (*Epidermal Growth Factor Receptor*). Овај протеин је рецептор тирозинских киназа и учествује у регулацији процеса као што су раст, деоба и преживљавање ћелија. Због његове улоге у развоју и напредовању појединих малигних обољења, EGFR представља значајну мету у развоју лекова.

Припрема скупа података

Релевантни атрибути којима су описане молекулске структуре из (необрађеног) скупа података су њихов идентификатор (*Molecule ChEMBL ID*), SMILES запис, експериментална IC50 вредност (*Standard Value*) са мерном јединицом (*Standard Units*) и типом релације (*Standard Relation*). На слици 3.1 приказан је начин на који је филтриран сирови скуп података, као и број молекула који су остали након сваког корака филтрирања. У наставку су објашњени сви појединачни кораци у овом процесу, који су приказани на слици 3.1.

1. Филтрирање по мерној јединици: Избачени су сви ентитети у којима мерна јединица IC50 вредности није наведена, или је изражена у јединицама различитим од nM.

2. Филтрирање по типу релације: У неким случајевима експериментално мерење не даје тачну вредност IC50, већ само доњу или горњу границу измерене вредности. Из тог разлога, како би се користиле само егзактне нумеричке вредности погодне за регресију, искључени су сви редови у којима тип релације недостаје, или је различит од једнакости.

3. Филтрирање по вредности: Остављени су само они редови у којима је IC50 вредност позитивна.

4. Филтрирање по SMILES запису: Остављени су само они записи у којима је SMILES запис молекула валидан. Исправност SMILES записа проверавамо помоћу библиотеке RDKit, тако што испитујемо да ли је могуће успешно конструисати Mol објекат на основу датог записа. За сваки преостали ред одређена је циљна вредност pIC50 трансформацијом IC50 вредности.

5. Агрегација поновљених мерења: Како би сваки молекул имао једну јединствену циљну вредност, извршено је груписање записа према идентификатору молекула. За сваки молекул задржана је максимална измерена pIC50 вредност, јер она представља највећу забележену потентност тог молекула у доступним експерименталним мерењима. Овакав приступ је у складу са каснијом употребом у генетском алгоритму, где веће pIC50 вредности представљају бољи резултат оптимизације. Иако оправдан са становишта циља рада, овај приступ може довести до благо оптимистичне процене активности у случајевима када постоји већи број различитих мерења за исти молекул.



Слика 3.1: Поступак филтрирања и агрегације скупа података

Након примене свих наведених корака добијен је коначан скуп од 10834 јединствених молекулских структура, који је коришћен за даљу обраду и обу-

чавање предиктивних модела.

Формирање улазне репрезентације молекула

Формирани скуп података садржи молекулске структуре представљене у облику SMILES записа и одговарајуће pIC50 вредности. Међутим, текстуални запис молекула није директно погодан за примену већине алгоритама машинског учења, који захтевају нумеричку репрезентацију улазних података.

За сваки молекул из коначног скупа, SMILES запис је најпре конвертован у Mol објекат библиотеке RDKit (одељак 2.3). Након тога, над добијеним Mol објектима генерисани су *Morgan fingerprint* вектори (одељак 2.4). У овом раду коришћени су бинарни вектори дужине 2048 бита, са параметром *radius* једнаким 2, узимајући притом у обзир хиралност молекула¹. Коришћена вредност параметра *radius* одговара ECFP4 конфигурацији (енг. *Extended-Connectivity Fingerprint 4*) [9]. На овај начин, добијена је бинарна матрица димензија 10834×2048 , чије вредности представљају улазне атрибуте регресионих модела, док pIC50 вредност представља циљну променљиву коју модели уче да предвиде.

Контролом квалитета утврђено је да се 10834 јединствених молекулских структура пресликало у 10674 различита *Morgan fingerprint* вектора. То значи да је приближно 1.5% молекула имало идентичну бинарну репрезентацију. Оваква појава је очекивана и представља последицу фиксне дужине и начина хеширања структурних карактеристика.

Подела на тренинг, валидациони и тест скуп

Након формирања нумеричке репрезентације молекула, скуп података је подељен на тренинг, валидациони и тест скуп. Тренинг скуп коришћен је за обучавање регресионих модела, валидациони скуп за избор одговарајућих хиперпараметара и поређење модела током развоја, док је тест скуп коришћен за коначну процену перформанси одабраних модела.

Уместо случајне поделе молекула, коришћена је подела заснована на Мурковим скелетима (енг. *Murcko scaffold*) [2]. Мурков скелет представља структурно језгро молекула које се добија издвајањем основног система прстенова

¹За молекула кажемо да је *хиралан* ако се не може поклопити са својим ликом у огледалу. За овакав молекул и његов лик у огледалу кажемо да су *енантиомери*, и у овом контексту, правимо разлику између њих.

и веза које их повезују, при чему се бочне функционалне групе и терминални делови молекула уклањају. Молекули са истим или веома сличним структурним језгром на тај начин могу бити сврстани у исту групу. На слици 3.2 приказан је пример два таква молекула, који имају заједнички скелет у виду киназолинског језгра.



Слика 3.2: Молекули А и Б са истим скелетом у виду киназолинског језгра

Приликом поделе, сви молекули који имају исти Мурков скелет задржани су у истом подскупу. На овај начин се спречава да веома сличне молекулске структуре истовремено буду присутне у тренинг и тест скупу, што би код случајне поделе могло довести до прецењивања способности модела да предвиђа активност структурно различитих молекула. Подела према скелетима стога представља реалистичнију процену способности модела да генерализује на молекулске структуре које нису биле присутне у процесу обучавања.

Циљ је био да приближно 80% молекула припада тренинг скупу, а по 10% валидационом и тест скупу. Пошто се додељују целе групе са заједничким скелетом, а не појединачни молекули, коначни број записа у сваком подскупу би могао да незнатно одступа од циљаног односа. У нашем случају, добијене пропорције одговарају циљаним, а конкретне кардиналности ових скупова дате су у табели 3.1.

Скуп	Број молекула	Удео [%]
Тренинг	8 667	80,0
Валидациони	1 083	10,0
Тест	1 084	10,0

Табела 3.1: Подела на тренинг, валидациони и тест скуп

Обучени модели

За предвиђање pIC50 вредности одабрана су три регресиона модела различитих приступа моделовању: *Ridge* регресија, *Random Forest* и *LightGBM*. *Ridge* регресија је укључена као линеарни референтни модел, док *Random Forest* и *LightGBM* представљају нелинеарне приступе засноване на ансамблима стабала одлучивања. На овај начин омогућено је поређење једноставнијег линеарног модела са сложенијим моделима који могу да опишу нелинеарне односе између улазних карактеристика и циљне вредности.

Избор хиперпараметара извршен је коришћењем валидационог скупа, при чему је за сваки модел дефинисан посебан простор могућих конфигурација, као што је приказано у табели 3.2. Као критеријум за избор конфигурације коришћена је RMSE метрика. За *Random Forest* и *LightGBM* примењена је двостепена претрага. У првој фази испитан је шири простор вредности хиперпараметара, док је у другој фази претрага усмерена на околину најбоље конфигурације добијене у првој фази. На тај начин је омогућено да се након грубе процене области у којој се налазе добре конфигурације изврши детаљнија претрага. За *Ridge* регресију коришћена је једностепена претрага. Приликом подешавања *LightGBM* модела коришћено је и рано заустављање (енг. *early stopping*): обука се прекида ако се коренована средњеквадратна грешка (RMSE) на валидационом скупу не побољша у року од 100 узастопних итерација. Укупно је испитано 10 конфигурација за *Ridge* регресију, 48 за *Random Forest* (по 24 у грубој и финој фази) и 59 за *LightGBM* (32 у грубој и 27 у финој фази). Након избора хиперпараметара, коначна верзија сваког модела обучена је на обједињеном тренинг и валидационом скупу. Тест скуп коришћен је искључиво за коначну процену перформанси овако добијених модела.

Резултати и упоредна анализа

Перформансе коначних модела на тест скупу приказане су у табели 3.3. За процену грешке предикције коришћене су RMSE и MAE метрике, изражене у јединицама pIC50, док су R^2 и Пирсонов коефицијент корелације коришћени за процену слагања предвиђених и експериментално одређених вредности.

Најбоље укупне резултате остварио је *Random Forest*, са најнижом вредношћу и RMSE и MAE метрике, као и највишим вредностима R^2 и Пирсоновог коефицијента корелације. *LightGBM* је остварио перформансе веома сличне

Модел	Хиперпараметар	Претражене вредности	Изабрано
<i>Ridge</i>	alpha	0,1; 1; 10; 100; 1000	100
	solver	sparse_cg, lsqr	lsqr
<i>Random Forest</i>	n_estimators	грубо: 500, 800; фино: 700, 800	800
	max_depth	грубо: 40, 60; фино: 50, 60, 70	60
	min_samples_leaf	грубо: 1, 5, 10; фино: 1, 3	1
	max_features	грубо: 0,3, sqrt; фино: 0,3, 0,5	0,3
<i>LightGBM</i>	learning_rate	грубо: 0,05, 0,1; фино: 0,04, 0,05, 0,06	0,04
	num_leaves	грубо: 31, 63; фино: 47, 63, 79	79
	min_child_samples	грубо: 20, 50; фино: 40, 50, 60	50
	reg_alpha	грубо: 0, 1; фино: 0	0
	reg_lambda	грубо: 0, 1; фино: 1	1
	n_estimators	горња граница 5000 (рано заустављање)	983

Табела 3.2: Простор претраге хиперпараметара и изабране вредности

Модел	RMSE	MAE	R^2	Пирсон
<i>Ridge</i>	1,03	0,80	0,43	0,67
<i>Random Forest</i>	0,89	0,65	0,58	0,77
<i>LightGBM</i>	0,91	0,67	0,55	0,75

Табела 3.3: Перформансе модела на тест скупу

Random Forest моделу, што указује да у овом експерименту сложенији приступ градијентног појачавања није донео значајно побољшање у односу на *Random Forest*. Ови резултати указују да линеарна зависност није довољна за адекватно описивање односа између молекулске репрезентације и pIC50 вредности, док нелинеарни модели дају значајно боље резултате.

Резултат *Random Forest* модела (као најбољег од обучених модела) на тест скупу упоређен је са ширим истраживањима која предвиђају pIC50 активност према EGFR на ChEMBL подацима [1, 5]. У табели 3.4 приказана је упоредна анализа квалитета најбољих модела у наведеним радовима са најбољим моделом овог рада.

Резултати приказани у табели показују да су перформансе развијеног *Random Forest* модела у складу са резултатима претходних истраживања за пред-

ГЛАВА 3. ПРОЦЕНА ЛЕКОВИТОСТИ И БИОЛОШКЕ АКТИВНОСТИ

Рад	Репрезентација	Модел	Подела	MAE	RMSE	R^2	Корелација
Gupta et al. [5]	2D дескриптори	<i>Extra Trees</i>	случајна	0,61	0,89	0,67	0,82
Akinnusi et al. [1]	ECFP + дескриптори	<i>Extra Trees</i>	скелет	—	$0,90 \pm 0,02$	$0,55 \pm 0,01$	$0,74 \pm 0,02$
Овај рад	<i>Morgan fingerprint</i>	<i>Random Forest</i>	скелет	0,65	0,89	0,58	0,77

Табела 3.4: Упоредна анализа модела за предвиђање pIC50 активности према EGFR. У нумеричким колонама подебљана је најбоља вредност (ниже MAE и RMSE, виши R^2 и корелација).

виђање pIC50 активности према EGFR. Посебно је значајно поређење са радом Akinnusi et al. [1], у којем је, као и у овом раду, коришћена подела према Мурковим скелетима. Иако се експерименталне поставке и коришћене репрезентације не поклапају у потпуности, добијени резултати су веома слични, при чему наш модел постиже нешто бољу вредност за обе метрике.

Резултати које представљају Gupta et al. [5] извештавају нешто нижу MAE грешку у односу на наш модел, док су RMSE вредности једнаке и износе 0,89. Међутим, битно је нагласити да је њихов модел заснован на случајној подели скупа. Случајна подела може дати оптимистичнију процену перформанси због присуства структурно сличних молекула у тренинг и тест скупу, док подела према скелетима представља строжи и конзервативнији сценарио евалуације.

Имајући то у виду, чињеница да наш модел на подели према скелетима постиже упоредиве резултате са [5], а у односу на [1] и нешто боље, указује да добијене перформансе не одступају од резултата пријављених у литератури. Истовремено, резултат показује да модел задржава релативно добру способност генерализације и у строжем сценарију евалуације, што је посебно значајно, имајући у виду његову каснију примену у генетском алгоритму.

Глава 4

Генетски алгоритам

Генетски алгоритми представљају једну од метахеуристичких метода за решавање оптимизационих проблема. Инспирисани су принципима природне еволуције и Дарвиновом теоријом природне селекције. Основна идеја је да се процес оптимизације посматра као аналоган еволуцији популације живих организама. Уместо једног решења, алгоритам одржава популацију потенцијалних решења. Свако решење представљено је на одговарајући начин, а његов квалитет се процењује помоћу фитнес функције (тј. функције прилагођености, енг. *fitness function*). На основу квалитета решења врши се селекција, при чему се квалитетнијим јединкама даје већа вероватноћа избора за стварање наредне генерације. Над изабраним јединкама примењују се генетски оператори *укрштања* и *мутације*. Укрштањем се комбинују карактеристике постојећих решења, док мутација уводи случајне измене (са одређеном вероватноћом) и тиме одржава разноврсност популације.

Након примене ових оператора формира се нова генерација, која се поново процењује, а поступак се итеративно понавља. На тај начин се популација постепено усмерава ка квалитетнијим решењима. Истовремено, генетски алгоритам настоји да одржи равнотежу између *експлоатације* већ пронађених добрих решења и *исстраживања* нових области простора претраге. Алгоритам се обично зауставља након одређеног броја генерација, достизања задатог квалитета или када више не долази до значајног побољшања решења.

У контексту проблема откривања молекулских структура са потенцијално лековитим дејством, простор могућих молекулских структура је дискретан и комбинаторно огроман, што отежава примену класичних градијентних метода оптимизације. Због тога се еволутивна претрага поставља као природан избор,

па је потребно описати начин на који се општи принципи генетског алгорита прилагођавају овом проблему.

Свака јединка у популацији представљена је помоћу репрезентација описаних у поглављу 2 — пре свега SMILES записом и RDKit Mol објектом, уз *Morgan fingerprint* векторе када је потребна предикција активности. Фитнес сваке јединке процењује се метрикама из поглавља 3. Кориснику је омогућен избор фитнес функције, почетне популације, броја генерација и осталих параметара генетског алгорита (што ће бити описано у одељку 5.2). У наставку поглавља биће објашњени оператори селекције, укрштања и мутације који се користе у оквиру развијеног система.

4.1 Селекција

Селекција одређује које ће јединке из тренутне популације бити изабране за родитеље приликом стварања нових јединки укрштањем. На основу вредности фитнес функције, квалитетније јединке имају већу вероватноћу да буду изабране. Истовремено, стохастичка природа селекције омогућава очување разноврсности популације и смањује могућност њеног прераног усмеравања ка малом броју сличних структура. У развијеном систему подржана су два начина селекције: турнирска (енг. *tournament selection*) и рулетска селекција (енг. *roulette wheel selection*).

Турнирска селекција. Приликом турнирске селекције, из популације се случајно бира k јединки, где k представља *величину турнира* (енг. *tournament size*). Од изабраних јединки за родитеља се бира она са највећом вредношћу фитнес функције. Већа вредност параметра k повећава селекциони притисак, јер квалитетније јединке имају већу предност у односу на остале, док мања вредност омогућава очување веће разноврсности популације.

Рулетска селекција. Код рулетске селекције, вероватноћа избора сваке јединке одређује се на основу њеног удела у укупној вредности фитнес функције популације. Јединке са већим вредностима фитнес функције имају већи удео у укупној вероватноћи избора и самим тим имају већу вероватноћу да буду изабране. Избор се реализује генерисањем случајне вредности и одређи-

вањем јединке на основу кумулативне расподеле вероватноћа, по принципу рулетског точка.

Елитизам. Поред избора родитеља, у свакој генерацији примењује се и *ели-тизам*. Најбољих e јединки из претходне генерације, према вредности фитнес функције, преноси се непосредно у наредну генерацију, без примене укрштања и мутације. На тај начин се гарантује да најбоља до тада пронађена решења неће бити изгубљена услед стохастичке природе генетских оператора. Преостала места у популацији попуњавају се новим јединкама добијеним применом селекције, укрштања и мутације.

4.2 Укрштање

Укрштање је генетски оператор којим се из два родитељска молекула генеришу потомачке структуре комбиновањем њихових делова. У контексту молекулских структура, циљ укрштања није само механичко комбиновање две структуре, већ генерисање хемијски валидних молекула који наслеђују одређене структурне карактеристике родитеља и истовремено проширују простор претраге.

За разлику од класичних генетских алгоритама над нумеричким или бинарним представама, код молекула није могуће произвољно комбиновати делове њихове репрезентације и очекивати валидан резултат. Због тога сваки од имплементираних оператора укрштања укључује поступак провере исправности добијених структура, а у случају неуспешног укрштања омогућава се поновни покушај са другим случајно изабраним деловима родитељских структура.

Након избора родитеља, они се упарују и над сваким паром се примењује одабрани оператор укрштања. У развијеном систему подржана су четири начина укрштања:

1. **SMILES укрштање** — укрштање на нивоу текстуалног SMILES записа;
2. **SELFIES укрштање** — укрштање на нивоу SELFIES токена;
3. **Графовско укрштање** — укрштање директно над молекулским графом;

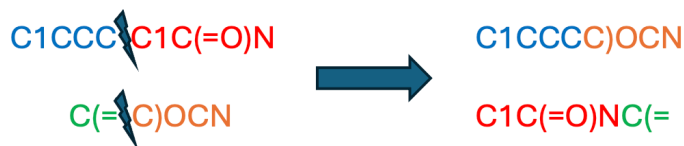
4. BRICS укрштање — укрштање на нивоу хемијски дефинисаних молекулских фрагмената.

Наведени оператори се међусобно разликују пре свега по начину представљања молекула и нивоу на којем се врши њихова рекомбинација. У наставку је за сваки оператор описан поступак избора делова родитељских структура, њиховог комбиновања и провере добијених потомака.

SMILES укрштање

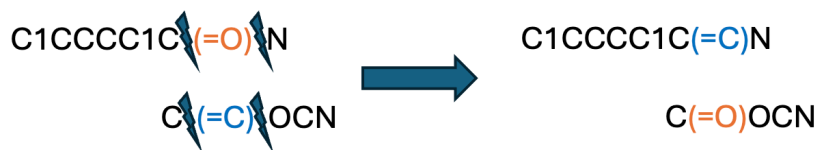
SMILES укрштање се извршава директно над текстуалним записима молекулских структура, без било каквог разматрања детаља саме структуре. Подржане су две врсте укрштања: једнопозиционо и двопозиционо.

Код једнопозиционог укрштања, за сваки од два родитељска SMILES записа насумично се бира по једна позиција пресека, при чему се пресек не поставља на почетак или крај ниске. Ниске се затим раздвајају на два дела и комбинују тако да први потомак садржи почетни део првог родитеља и завршни део другог родитеља, док други потомак настаје обрнутом комбинацијом. На овај начин добијају се две нове SMILES ниске, које се затим проверавају помоћу RDKit библиотеке како би се утврдило да ли представљају валидне молекулске структуре. Поступак избора позиција и формирања потомака приказан је на слици 4.1.



Слика 4.1: Једнопозиционо укрштање SMILES ниски

Уколико након унапред дефинисаног броја покушаја (табела 4.1) није могуће добити два валидна потомка једнопозиционим укрштањем, прелази се на двопозиционо укрштање. У овом случају се у сваком родитељском SMILES запису бирају по две различите позиције пресека. Део ниске између изабраних позиција једног родитеља замењује се одговарајућим делом другог родитеља, чиме се добијају два потомка. И у овом случају валидност добијених SMILES записа проверава се помоћу RDKit библиотеке. Пример двопозиционог укрштања приказан је на слици 4.2.



Слика 4.2: Двопозиционо укрштање SMILES ниски

Иако веома једноставан и брз, овакав поступак има велику ману. Насумичним пресецањем SMILES ниски веома често долази до формирања невалидних SMILES записа. То илуструје слика 4.1, где су добијена два невалидна потомка, док се двопозиционим укрштањем приказаним на слици 4.2 добијају два валидна потомка. Да би се овај негативни ефекат ублажио, поступак се понавља велики број пута (табела 4.1), али чак ни тада нема гаранције да ће се добити валидне структуре. Ако се ни тада не добију валидни потомци, родитељске SMILES ниске се пропуштају у неизмењеном облику у наредну генерацију.

SELFIES укрштање

Укрштање над SELFIES нискама реализовано је на сличан начин као једнопозиционо укрштање над SMILES нискама, при чему се размена делова родитељских структура врши на нивоу SELFIES токена. За разлику од SMILES укрштања, код којег се пресечне позиције бирају међу појединачним карактерима ниске, код SELFIES укрштања пресечне позиције се бирају између токена. На тај начин се спречава раздвајање појединачних SELFIES токена, који могу садржати више карактера. Поступак почиње енкодирањем SMILES записа родитеља у SELFIES ниске, након чега следи избор пресечних позиција. Потомци се потом формирају разменом завршних делова родитељских низова токена. Први потомак добија почетни део првог родитеља и завршни део другог родитеља, док други потомак добија почетни део другог родитеља и завршни део првог родитеља. Добијени низови токена се затим спајају у SELFIES ниске и декодирају у SMILES репрезентацију.

Укрштање графова

Графовско укрштање врши рекомбинацију директно над молекулским графом, односно над Mol објектима библиотеке RDKit. За разлику од укрштања

над текстуалним репрезентацијама, код овог приступа место пресецања одређује се на основу структуре молекула, чиме се омогућава директна манипулација атомима и хемијским везама.

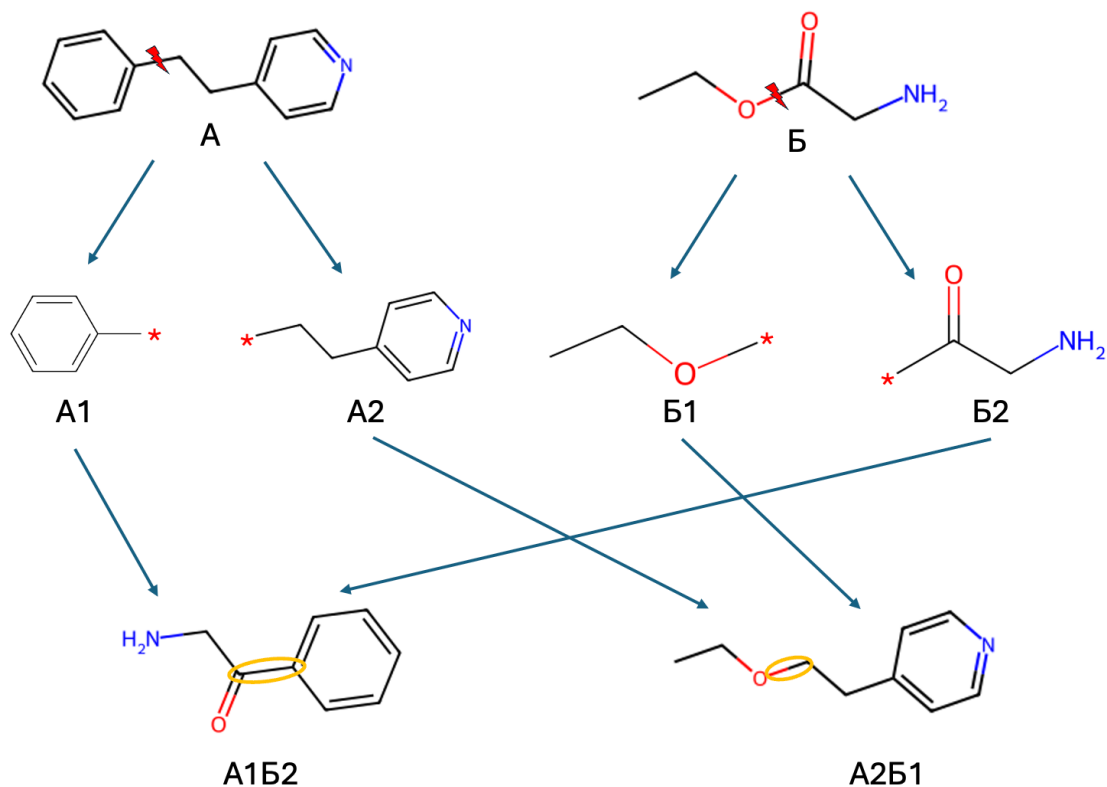
Поступак укрштања започиње случајним избором једне везе из сваког родитељског молекула. Као кандидати за пресецање разматрају се искључиво ацикличне везе, тј. везе које нису део ниједног прстена. Овим ограничењем се избегава директно пресецање цикличних структура и гарантује се раздвајање молекулских графова на два одвојена фрагмента. На слици 4.3 приказан је начин укрштања два родитељска молекула - А и Б. Посебно су означене везе које су изабране за пресецање, а настали фрагменти означени су са А1, А2, Б1 и Б2.

Атоми који су градили пресечену везу се затим повезују помоћним, односно *dummy* атомима. Они означавају места на којима је могуће поново повезати настале фрагменте. На слици 4.3, ти атоми означени су симболом *.

Добијени фрагменти се затим укрштају тако што се по један фрагмент првог родитеља спаја са фрагментом другог родитеља, и обрнуто (на тај начин, спајамо фрагменте А1 и Б2, односно А2 и Б1). Два фрагмента се комбинују тако што се њихови *dummy* атоми уклоне, а атоми који су били њихови суседи повежу се једноструком хемијском везом. Једнострука веза је изабрана као најсигурнија опција за поновно повезивање, јер гарантовано неће нарушити валенце атома који се повезују, а тиме и валидност новодобијених структура. У примеру приказаном на слици, на тај начин добијамо структуре А1Б2 и А2Б1.

BRICS укрштање

BRICS (енг. *Breaking of Retrosynthetically Interesting Chemical Substructures*) [4] представља метод за декомпозицију молекула на мање структурне фрагменте раскидањем оних веза које се често могу формирати у реакцијама синтезе. За разлику од једноставног пресецања SMILES ниске на одређеној позицији, BRICS декомпозиција узима у обзир хемијско окружење атома који учествују у вези. Метод дефинише 16 различитих типова хемијских окружења, који су на слици 4.4 приказани ознакама од L1 до L16. На овој слици, свака линија између фрагмената означава да се они могу повезати хемијском везом током неке познате хемијске реакције, тј. да су *компатибилни*. Приликом декомпозиције, везе између компатибилних типова окружења могу бити



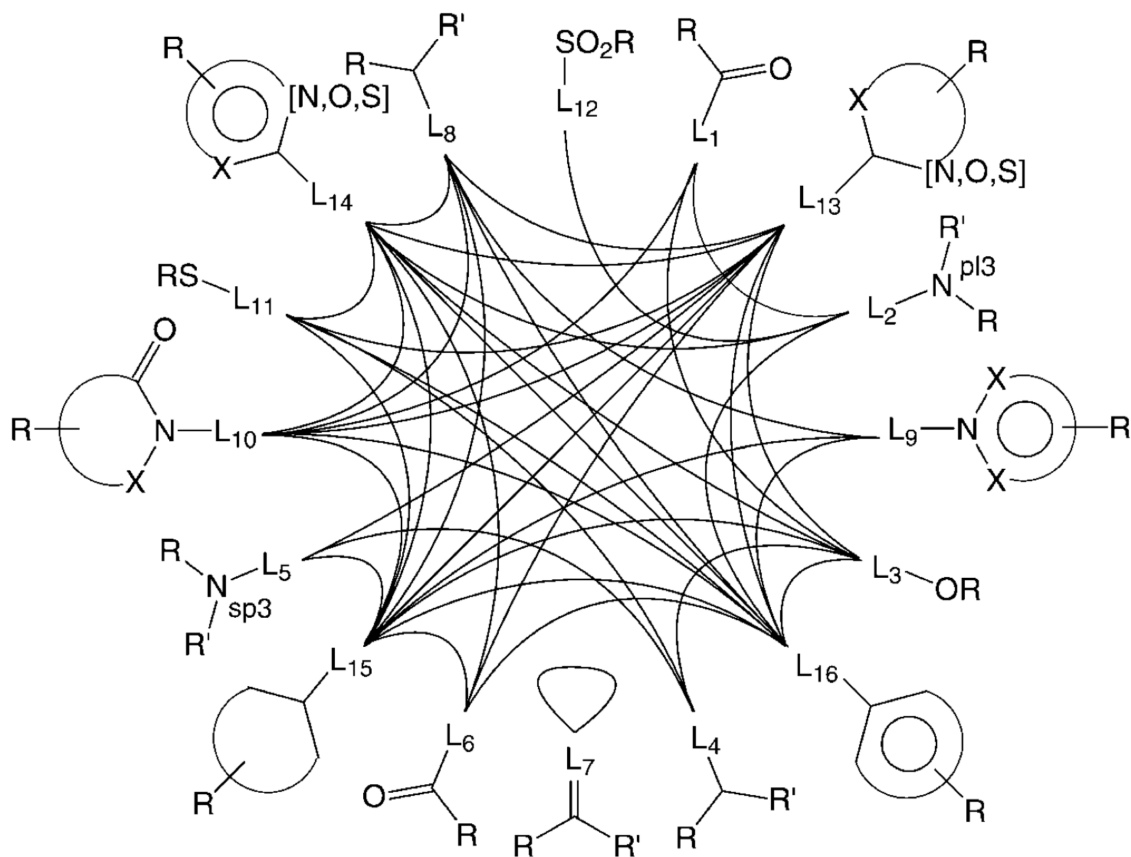
Слика 4.3: Схема графовског укрштања: пресецање ацикличних веза, формирање и рекомбинација фрагмената

прекинуте, чиме се добијају фрагменти који на местима прекида садрже ознаке које одређују њихов BRICS тип. Те ознаке представљају места на којима фрагмент може поново да се повеже са неким другим компатибилним фрагментом.

У оквиру генетског алгорита, BRICS укрштање најпре декомпонује оба родитељска молекула применом BRICS правила. Добијени фрагменти се затим третирају као градивни елементи родитељских структура. Сваки од родитељских скупова фрагмената се партиционише на два подскупа, након чега се изврши рекомбинација ових партиција. Другим речима, ако су фрагменти првог родитеља подељени на скупове F_1^A и F_1^B , а фрагменти другог родитеља на F_2^A и F_2^B , скупови потомачких фрагмената C_1 и C_2 се формирају као:

$$C_1 = F_1^A \cup F_2^B,$$

$$C_2 = F_2^A \cup F_1^B.$$



Слика 4.4: BRICS типови окружења (L1 - L16) и дозвољене везе између компатибилних фрагмената. Преузето из [4].

Тако добијени скупови прослеђују се поступку BRICS реконструкције, којој покушава да повеже компатибилне фрагменте. Пошто се фрагменти једног скупа могу повезивати на различите начине, реконструкција може произвести велики број различитих молекулских структура. У имплементацији у овом раду не врши се потпуна еnumerација свих могућих структура, већ се разматра ограничен број генерисаних кандидата, након чега се један од њих насумично бира као потомак.

Важна особина BRICS укрштања је то што се заснива на хемијски мотивисаним правилима, захваљујући којима се добијају хемијски смислени фрагменти који могу представљати реалне градивне блокове за синтезу. Приликом њихове рекомбинације поштују се правила о међусобно компатибилним типовима фрагмената, чиме се повећава вероватноћа да ће добијене молекулске структуре бити синтетички доступне. Ово је значајна предност у откривању

нових молекула, јер поред њихових жељених својстава треба узети у обзир и практичну изводљивост синтезе.

4.3 Мутација

Мутација је генетски оператор који уводи случајне измене у структуру појединачних јединки и на тај начин доприноси одржавању разноврсности популације. За разлику од укрштања, које комбинује структурне делове два родитеља, мутација делује на једној јединки и омогућава истраживање нових молекулских структура које не морају бити директна комбинација родитељских структура.

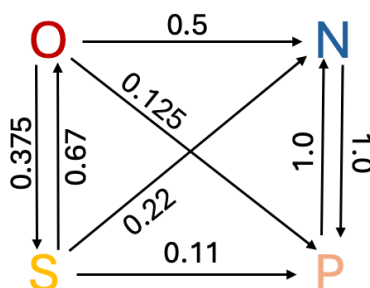
У развијеном систему мутација се примењује на сваког потомка након укрштања, са задатом вероватноћом. Као и у случају укрштања, подржана су четири типа мутација: SMILES, SELFIES, графовска и BRICS мутација. Избор режима мутације врши се независно од режима укрштања, што омогућава произвољно комбиновање различитих приступа, попут SELFIES укрштања са SMILES мутацијом. Сви оператори настоје да очувају хемијску валидност добијених молекула. У имплементираном систему, уколико изабрани тип мутације није могуће применити на конкретну јединку, или пак он не доводи до валидне структуре, примењује се резервна, SMILES мутација. Оператори мутације су на тај начин повезани у ланац резервних поступака, при чему се у крајњем случају користе једноставније SMILES операције, чија је успешност извршавања загарантована.

SMILES мутација

SMILES мутација подразумева директну манипулацију SMILES ниском, изменом њених специфичних делова. Подржана су четири типа SMILES мутација: замена хетероатома, замена функционалних група, делеција и инсерција. У случају да систем одлучи да изврши SMILES мутацију, на случајан начин, са једнаким вероватноћама, бира се један од наведених конкретних типова.

Замена хетероатома (атома различитих од угљеника и водоника) реализована је тако што се пронађу све позиције хетероатома у SMILES ниски, а онда се за произвољно одабрану позицију врши замена тог хетероатома дру-

гим хетероатомом. Скуп ових прелаза је унапред дефинисан, и једноставности ради, ограничен је на атоме кисеоника, азота, сумпора и фосфора. Граф прелаза једног атома у други приказан је на слици 4.5, при чему тежине грана означавају вероватноћу преласка једног атома у други. Битно је нагласити да су прелази дефинисани тако да не нарушавају валидност новодобијеног молекула. Овакав тип мутације није могуће применити на молекуле који не садрже хетероатоме.



Слика 4.5: Граф прелаза између хетероатома и њихове вероватноће

Замена функционалних група заснива се на испитивању присуства одређених функционалних група у молекулу. Функционалне групе су специфични делови молекула који диктирају његово хемијско понашање, тј. типове реакција у које ступа, али и физичка својства. Због тога је смислено мутирати овакве делове молекула. Конкретно, омогућена је замена карбонилне групе карбоксилном, карбоксилне групе амидном, карбонилне групе аминоксидом, као и нитрилне групе алдехидном. Такође, омогућена су формирања етарске или кето функционалне групе између два суседна атома угљеника, али и продужетак угљеничног ланца. Постојањем последње наведене мутације, повећава се могућност генерисања бар једног кандидата за мутацију, чак и у случајевима када молекул не садржи неку од специфичних функционалних група које су предмет осталих дефинисаних замена. За сваку дефинисану промену и свако њено појављивање у SMILES ниски, формира се по један кандидат, при чему се мења само конкретно појављивање функционалне групе. Од свих тако добијених кандидата, један се насумично одабира и користи као резултат мутације.

Делеција се реализује уклањањем одређеног дела SMILES ниске. Најпре се испитује да ли SMILES ниска садржи гранања, при чему се гранање посматра као подниска ограничена одговарајућим упареним заградама. Уколико грана-

ња постоје, насумично се бира једно од њих и покушава се његово уклањање. Уколико молекула не садржи гранања или ниједан покушај њиховог уклањања не резултује валидном SMILES ниском, прелази се на покушај брисања појединачног атома са насумично одабране позиције. Након сваког покушаја проверава се валидност добијене структуре, а поступак се понавља до унапред дефинисаног броја покушаја (табела 4.1)). Уколико ниједан од наведених поступака не буде успешан, као резервна операција примењује се замена атома.

Инсерција подразумева додавање атома или функционалне групе у постојећу SMILES ниску. Најпре се из унапред дефинисаног скупа могућих уметања насумично одабира један атом или функционална група, након чега се насумично бира позиција у SMILES ниски на коју ће бити извршено уметање. Будући да произвољно уметање може довести до невалидне молекулске структуре, добијена SMILES ниска се након сваког покушаја проверава. Поступак се понавља до унапред дефинисаног броја покушаја (табела 4.1), све док се не добије валидна структура. Уколико ниједан покушај не буде успешан, примењује се резервна операција делеције.

SELFIES мутација

SELFIES мутација се реализује над SELFIES репрезентацијом молекула, чиме се омогућава измена молекулске структуре на нивоу њених токена. Најпре се улазна молекулска структура преводи из SMILES у SELFIES запис, након чега се случајно бира једна од три врсте мутације: замена, уметање или брисање токена. Код замене се један токен замењује другим насумично изабраним токеном из дозвољене SELFIES азбуке, док се код уметања нови токен додаје на случајно изабрану позицију. Брисањем се, са друге стране, уклања један постојећи токен. Ова операција се примењује само ако молекула садржи више од једног токена. Измењени низ токена се затим декодира у SMILES запис. Уколико добијена структура није валидна или је идентична почетној структури, покушава се нова мутација, при чему је број покушаја ограничен (табела 4.1). Уколико ниједан покушај не доведе до одговарајуће измене, примењује се резервна мутација - SMILES делеција.

Графовска мутација

Графовска мутација примењује се над графом молекула, описаном у одељку 2.3. Могуће је извршити замену атома (што одговара промени чвора у графу) или замену везе (што одговара промени гране у графу). У случају замене атома, насумично се бира атом и замењује са произвољним атомом, из унапред дефинисаног скупа атома. Замена везе подразумева промену њене вишеструкости - насумично одабрана веза мења се једноструком или двоструком везом. Ниједна од наведених графовских мутација не улаже напор да произведе валидну молекулску структуру, већ измене обавља механички, без провере хемијског контекста. Сваки кандидат се проверава RDKit санитизацијом ¹, а одбацује се ако није валидан или је идентичан полазном молекулу. Због тога се овај поступак понавља до унапред дефинисаног броја покушаја (табела 4.1), а ако ни тада не успемо да мутирамо молекул, примењује се резервна мутација - SMILES делеција.

BRICS мутација

BRICS мутација заснива се на BRICS декомпозицији молекула, замени насумично одабраног фрагмента молекулске структуре другим и поновној BRICS композицији добијеног скупа фрагмената, након чега се насумично бира једна од добијених могућих структура. Фрагмент за замену је на случајан начин одабран из скупа фрагмената који је добијен BRICS декомпозицијом свих молекула из прве генерације генетског алгорита, заједно са неколицином референтних молекула. Резултујући молекул се проверава и санитизује, а поступак се понавља уколико структура није валидна или је идентична полазном молекулу. Уколико ни након ограниченог броја покушаја (табела 4.1) није могуће добити одговарајућу структуру, примењује се SMILES делеција као резервна мутација.

¹RDKit санитизација представља поступак провере и припреме молекулске структуре, који обухвата проверу валентности атома, ароматичности, формалних наелектрисања и других структурних својстава неопходних за исправну интерпретацију молекула.

Репрезентација	Укрштање [макс. покушаја]	Мутација [макс. покушаја]
SMILES	2000 (1-поз.) 2000 (2-поз.)	замена атома / групе: 1 покушај инсерција: 200 делеција: 20 покушаја за гране / 500 покушаја за атоме
SELFIES	20	20
Граф	100	50
BRICS	100	20

Табела 4.1: Максималан број покушаја пре одустајања оператора

Глава 5

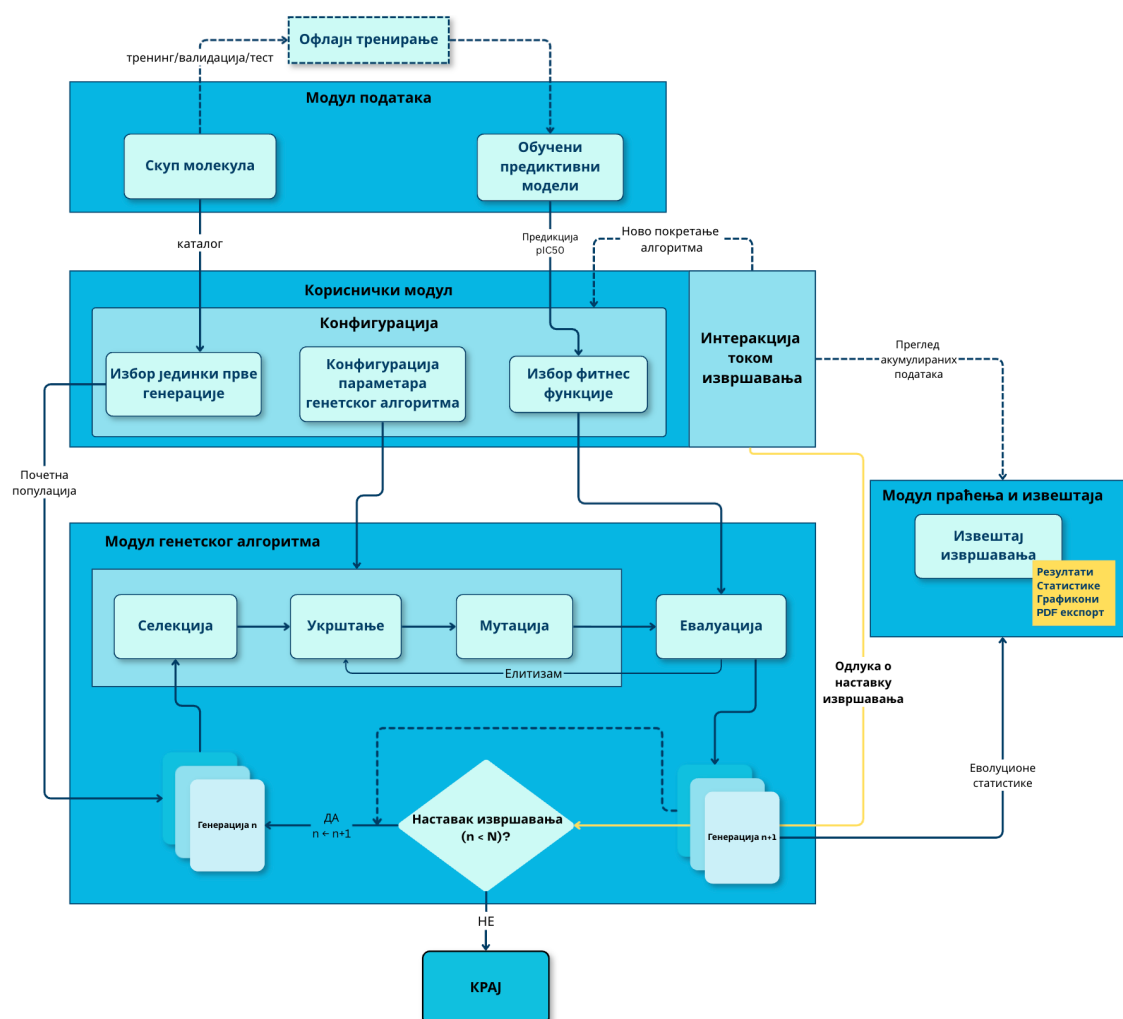
Имплементација и архитектура система

Развијени систем представља десктоп апликацију намењену примени генетског алгоритма у процесу откривања молекулских структура са потенцијално лековитим дејством. Систем обједињује управљање молекулским подацима, извршавање генетског алгоритма, различите начине представљања и измене молекула, евалуацију њихове подобности, као и прикупљање статистичких података о току извршавања. Циљ је био материјализовати концепте објашњене у претходним поглављима и пружити једноставан алат за извођење и поређење различитих експеримената над молекулским структурама. Изворни код апликације, упутство за инсталацију и покретање, као и актуелна верзија рада у PDF формату, доступни су у јавном репозиторијуму пројекта на адреси <https://github.com/killica/DrugDiscovery>.

Систем је имплементиран у програмском језику *Python*. Графички кориснички интерфејс реализован је помоћу библиотеке *PyQt5*, док се за рад са молекулским структурама користи *RDKit*. За рад са *SELFIES* репрезентацијом користи се посебна библиотека *selfies*. Предиктивни модели обучавају се ван апликације коришћењем библиотека *scikit-learn* и *LightGBM*, након чега се серијализовани модели учитавају у апликацију и користе за евалуацију молекулских структура током извршавања генетског алгоритма. Резултати експеримената визуализују се помоћу библиотеке *matplotlib*.

5.1 Архитектура система

Архитектура развијеног система организована је у четири функционалне целине које одговарају блоковима приказаним на слици 5.1: модул података, кориснички модул, модул генетског алгорита и модул праћења и извештаја. Раздвајање ових одговорности омогућава независан развој и замену појединих делова без измене остатка система.



Слика 5.1: Архитектура система

Модул података обухвата скуп молекула из ког се формира почетна популација и скуп претходно обучених предиктивних модела. Обука модела спроводи се офлајн, ван апликације, и није део извршавања генетског алгорита. У фази извршавања, користе се само серијализоване верзије модела за пре-

дикцију pIC50 (одељак 3.2).

Кориснички модул је одговоран за формирање конфигурације експеримента, што подразумева одабир почетне популације, параметара генетског алгорита, као и начин евалуације. Детаљан опис ових корака дат је у одељку 5.2. Такође, овај модул је заслужан за управљање извршавања генетског алгорита - одлучује да ли ће се и колико дуго алгоритам извршавати, а такође омогућава преглед акумулираних података из модула праћења и извештаја током извршавања.

Централни део система чини *модул генетског алгорита*, који на основу примљене конфигурације итеративно формира нове генерације и позива компоненту за евалуацију јединки. Евалуација је издвојена од саме логике алгорита, тако да он зависи само од интерфејса фитнес функције, а не и од конкретне имплементације (QED или предиктивни модел; одељци 3.1 и 3.2).

Током извршавања, свака нова генерација прослеђује еволуционе статистике *модулу праћења и извештавања*, који их прикупља без утицаја на ток алгорита. Након завршетка, прикупљени подаци доступни су кроз извештај о извршавању (описан у одељку 5.2).

5.2 Кориснички интерфејс

Кориснички интерфејс апликације организован је тако да прати основне фазе извршавања генетског алгорита. Рад са апликацијом започиње избором почетне популације и начина евалуације молекула, након чега следи конфигуравање параметара генетског алгорита. По покретању алгорита кориснику је омогућено праћење формирања нових генерација, док се у завршној фази приказују статистички подаци о току извршавања.

Избор молекула и фитнес функције

Почетни екран апликације приказан је на слици 5.2. Његова основна улога је формирање почетне популације над којом ће се извршавати генетски алгоритам, као и избор фитнес функције којом ће се апроксимирати квалитет јединке у популацији. У делу интерфејса предвиђеном за каталог приказан је подскуп молекула из доступног скупа података, при чему су молекули представљени својим структурним формулама, идентификатором из базе

ChEMBL и одговарајућом вредношћу фитнес функције. Одабиром молекула они се додају у почетну популацију, која је приказана у доњем делу екрана.

Поред појединачног избора, интерфејс омогућава аутоматски избор свих приказаних молекула, као и случајно узорковање задатог броја молекула из доступног скупа. Додатно, корисник може да дода произвољан молекул из доступног скупа од приближно 10 хиљада молекула, уносом његовог ChEMBL идентификатора или SMILES репрезентације. Уколико тражени молекул није пронађен у доступном скупу, могуће је додати произвољан молекул уносом његове SMILES репрезентације у форму приказану на дну прозора на слици 5.2. На овај начин, почетна популација не мора бити ограничена искључиво на молекуле приказане у каталогу.

На десној страни приказан је део интерфејса намењен избору фитнес функције. Корисник може изабрати QED (одељак 3.1) или један од претходно обучених модела за предикцију вредности rIC50 (одељак 3.2). У зависности од изабраног начина евалуације, приказују се релевантне информације о њему. Код предиктивних модела приказане су перформансе модела и његови хиперпараметри (што је случај на слици 5.2), док се избором QED режима омогућава подешавање значаја појединачних дескриптора који учествују у израчунавању QED вредности, што су заправо коефицијенти w_i из једначине 3.2 (приказано на слици 5.3).

Конфигурисање генетског алгорита

Након дефинисања почетне популације и начина евалуације, корисник прелази на конфигурисање генетског алгорита, чији се параметри подешавају на засебном екрану, приказаном на слици 5.4. Корисник задаје број генерација, величину турнира у случају турнирске селекције, број јединки које се задржавају применом елитизма и вероватноћу мутације. Поред тога, бирају се и конкретне методе укрштања и мутације (одељци 4.2 и 4.3). Након подешавања параметара, покретањем претраге апликација иницира извршавање генетског алгорита са изабраном конфигурацијом и тиме се креира друга генерација молекула.

The screenshot displays the 'Drug Discovery' application interface. It is divided into several sections:

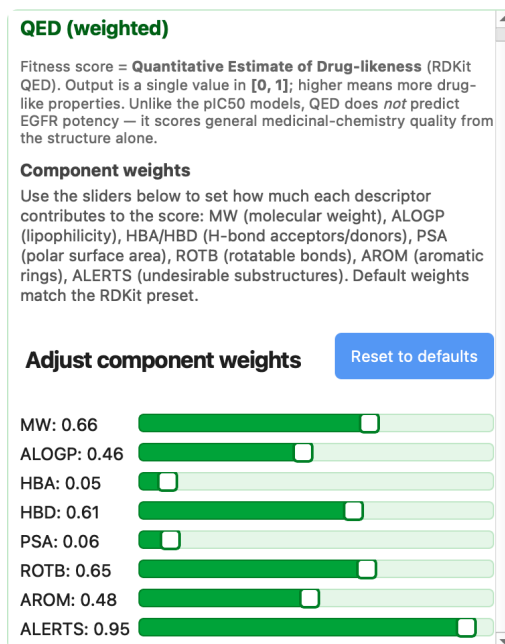
- Catalogue:** A grid of six chemical structures. The first three are highlighted with green borders. Below each structure is its ChemBL ID and pIC50 value:
 - CHEMBL5434495, pIC50: 10.1311
 - CHEMBL5597724, pIC50: 9.2770
 - CHEMBL4279016, pIC50: 9.1798
- Selected for first generation:** A section with two chemical structures, also highlighted with green borders. Below them are their ChemBL IDs and pIC50 values:
 - CHEMBL474141, pIC50: 8.9996
 - CHEMBL3759356, pIC50: 8.0558
- Fitness:** A panel on the right showing model selection and performance metrics.
 - Model Selection:** Radio buttons for QED (weighted), Random Forest (pIC50) (selected), LightGBM (pIC50), and Ridge (pIC50).
 - Random Forest (pIC50) Details:**
 - Fitness score = predicted pIC50 (higher is better). Scaffold-split test metrics below.
 - Test (refit on train+val):** RMSE: 0.8873, MAE: 0.6547, R²: 0.5781, Pearson: 0.7700.
 - Validation (tuning selection):** RMSE: 0.9782, MAE: 0.7173, R²: 0.5025, Pearson: 0.7153.
 - Hyperparameters:** max_depth: 60, max_features: 0.3, min_samples_leaf: 1, n_estimators: 800.
 - Fingerprint:** Morgan r=2, 2048 bits (chirality=on).
 - Model:** random_forest_best.joblib
 - Tuning log:** tuning_history_random_forest.csv
- Add molecule to catalogue:** A form with fields for 'SMILES...' and 'Name or description (optional)', followed by an 'Add to catalogue' button.
- Continue:** A green button at the bottom right.

Слика 5.2: Приказ почетног прозора апликације - избор молекула и фитнес функције

Праћење извршавања генетског алгорита

Током извршавања алгорита интерфејс приказује тренутну популацију и новоформирану генерацију, као што је приказано на слици 5.5. На овај начин корисник може непосредно да прати промене молекулских структура током еволуције популације и да их упореди са структурама из претходне генерације. Поред приказа молекула, интерфејс приказује и тренутни напредак извршавања, укључујући редни број обрађене генерације и текући број успешно формираних јединки.

Извршавање се може наставити корак по корак генерисањем наредне генерације, или се поступак може аутоматски наставити до унапред дефинисаног броја генерација. Кориснику је такође омогућен приступ извештају о тренутном извршавању (што ће бити објашњено у наредном поделу), као



Слика 5.3: Приказ панела за подешавање параметара QED режима

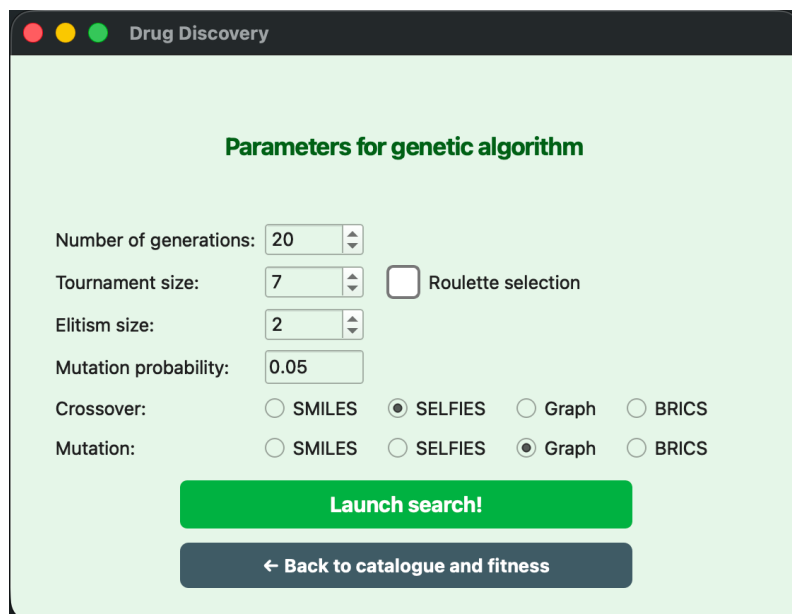
и поновно покретање анализе са новом конфигурацијом.

Статистике извршавања

Резултати извршавања генетског алгорита обједињују се у посебном извештају, на начин приказан на слици 5.6. У њему се приказују параметри коришћени током конкретног извршавања, најбоље пронађена молекулска структура и статистички показатељи процеса еволуције. Поред фитнес вредности најбоље јединке кроз генерације, прате се и просечне вредности фитнес функције у популацији, успешност оператора укрштања и мутације и диверзитет популације.

Диверзитет описује колико се јединке једне популације међусобно разликују. Ова величина је значајна јер превелика сличност јединки може указивати на губитак разноврсности популације и прерано усмеравање претраге ка ограниченом делу простора молекулских структура. Са друге стране, већи диверзитет указује на то да популација обухвата разноврсније молекулске структуре и да генетски алгоритам и даље истражује шири простор могућих решења.

За квантификовање диверзитета популације употребљен је Танимотов ко-



Drug Discovery

Parameters for genetic algorithm

Number of generations: 20

Tournament size: 7

Elitism size: 2

Mutation probability: 0.05

Crossover: ☐ SMILES ☒ SELFIES ☐ Graph ☐ BRICS

Mutation: ☐ SMILES ☐ SELFIES ☒ Graph ☐ BRICS

☐ Roulette selection

Launch search!

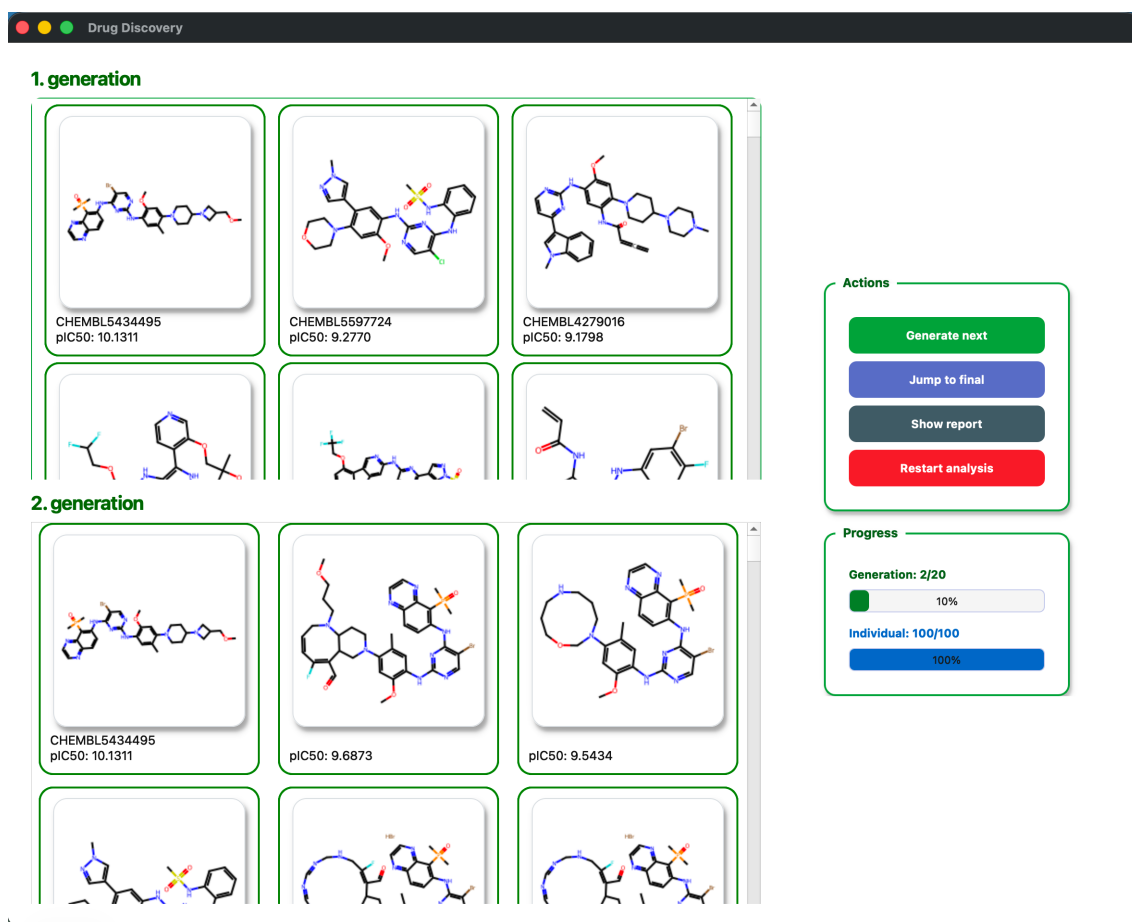
← Back to catalogue and fitness

Слика 5.4: Конфигурисање параметара генетског алгорита

ефицијент сличности, који је описан у одељку 2.4. Као и у случају формирања предиктивних модела (одељак 3.2), и овде су коришћене 2048-битне векторске репрезентације молекула, док је вредност параметра *radius* постављена на 2.

Када се посматра целокупна популација, Танимото коефицијент рачуна се за сваки пар јединки, па се на основу добијених вредности формира хистограм за текућу генерацију, приказан у извештају на слици 5.6. На основу добијених вредности могуће је проценити степен међусобне сличности популације, а самим тим и њен диверзитет. У том контексту, већа просечна сличност указује на мањи диверзитет, док мања просечна сличност указује на већу разноврсност популације. На тај начин, расподела вредности Танимото коефицијента пружа додатну информацију о понашању генетског алгорита која се не може добити само посматрањем вредности фитнес функције. На пример, пораст вредности фитнес функције најбоље јединке праћен истовременим смањењем диверзитета може указивати на конвергенцију популације ка сличним молекулским структурама. Насупрот томе, очување већег диверзитета указује на то да популација и даље садржи различите структуре које могу представљати полазну основу за даље побољшање.

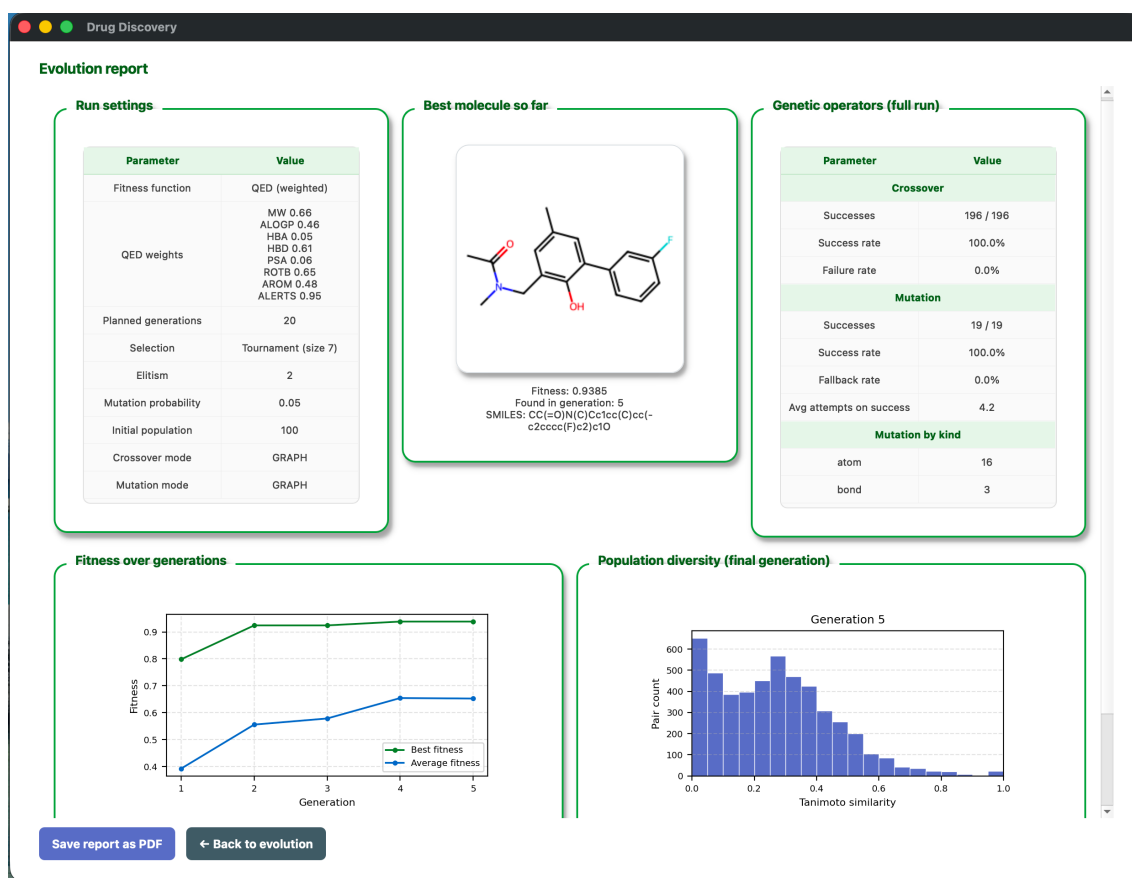
Приказивањем описаних статистика кориснику није омогућено само праћење тренутног стања алгорита, већ и анализа његовог понашања током целокупног извршавања. Извештај је могуће сачувати у PDF формату, чи-



Слика 5.5: Праћење извршавања генетског алгоритма

ме се резултати конкретног експеримента могу архивирати и користити за накнадно поређење.

Напомена о коришћењу алата вештачке интелигенције. Током израде овог рада коришћени су алати вештачке интелигенције - *Cursor* и *ChatGPT*. *Cursor* је коришћен за дебаговање, уобличавање, и генерисање појединих делова програмског кода, док је *ChatGPT* коришћен за израду нацрта и јасније формулисање делова текста рада. Све научне тврдње, резултате експеримента, закључке и коначну верзију рада аутор је самостално проверио, допунио и сноси одговорност за њихову тачност. Без преке потребе за наглашавањем, алати вештачке интелигенције нису коришћени за генерисање експерименталних резултата нити слика приказаних у раду, а сви предлози текста и кода увек су били предмет ауторске ревизије.



Слика 5.6: Статистике извршавања генетског алгоритма

Глава 6

Експериментални резултати

Развијени софтверски систем је прилично конфигурабилан и самодовољан, чиме пружа кориснику слободу да истражи утицаје различитих параметара на исход извршавања генетског алгоритма. Изражена конфигурабилност за последицу има велики број комбинација вредности параметара, што практично искључује њихово потпуно исцрпно испитивање. Наиме, поред параметара самог генетског алгоритма (број генерација, вероватноћа мутације, методе укрштања и мутације и др.), корисник може мењати и почетну популацију, начин евалуације (QED или предиктивни модел), као и тежине појединачних дескриптора у QED режиму. Простор могућих експеримената, стога, експоненцијално расте са бројем слободних параметара, а свако појединачно извршавање алгоритма захтева значајно процесорско време, нарочито када се за фитнес функцију користи предиктивни модел.

С једне стране, ова конфигурабилност омогућава дубљу анализу — корисник може испитивати не само коначни исход, већ и однос између појединих подешавања и понашања популације током еволуције (конвергенција, диверзитет, брзина побољшања). С друге стране, систематично мапирање тих веза захтева велики број поновљених покретања са различитим конфигурацијама, што представља главно ограничење у експерименталном делу рада. Стога су у наредним одељцима приказани репрезентативни експерименти, бирани тако да илуструју утицај најзначајнијих параметара, а не да обухвате цео простор претраге.

6.1 Основна експериментална поставка

Како би резултати различитих експеримената били међусобно упоредиви, за потребе експерименталног испитивања дефинисана је основна конфигурација генетског алгорита. Основна конфигурација приказана је у табели 6.1 и обухвата вредности параметара које су, осим уколико није другачије наведено, коришћене у свим експериментима. У експериментима у којима се испитује утицај неког од наведених параметара, његова вредност се мења у складу са поставком конкретног експеримента, док се остали параметри задржавају на вредностима из основне конфигурације.

Параметар	Вредност
Величина популације	30
Број генерација	80
Метод селекције	Турнирска селекција
Величина турнира	7
Елитизам	2 јединке
Вероватноћа мутације	0,05

Табела 6.1: Основна конфигурација генетског алгорита

6.2 Утицај величине популације

Величина популације представља један од основних параметара генетског алгорита, будући да директно одређује број јединки над којима се генетски оператори примењују у свакој генерацији. Повећањем популације повећава се број молекулских структура које алгоритам истовремено разматра, што може допринети већој разноврсности популације и квалитетнијим решењима. С друге стране, већи број јединки повећава рачунску сложеност извршавања алгорита, па је неопходно испитати однос између квалитета добијених решења и времена потребног за њихово проналажење. Циљ овог експеримента је испитивање утицаја величине популације на перформансе генетског алгорита. Посебно се посматрају квалитет најбољег пронађеног решења, просечан квалитет популације и време извршавања алгорита до последње генерације.

У експериментима су разматране популације величине 15, 30, 50 и 100 јединки. За све конфигурације, број генерација постављен је на 80, коришћен

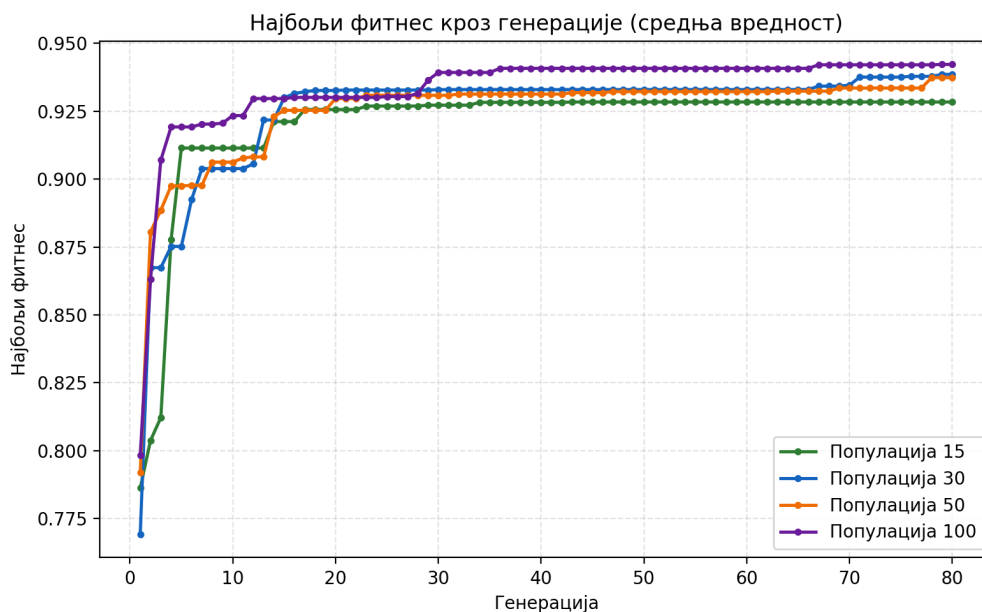
је QED коефицијент као фитнес функција, док су за укрштање и мутацију коришћене SELFIES методе. Због стохастичке природе генетског алгоритма, свака испитивана величина популације тестирана је у три независна покретања. За најбољи и просечни фитнес у последњој генерацији израчунате су средња вредност и стандардна девијација на основу три покретања. За време извршавања израчуната је медијана три измерена времена. Добијени резултати су приказани у табели 6.2, и они омогућавају поређење коначног квалитета решења и процесорског времена за различите величине популације. Поред коначних вредности, за свако извршавање забележене су вредности најбољег и просечног фитнеса у свакој генерацији. На сликама 6.1 и 6.2 приказан је ток ових вредности за различите величине популације. За сваку генерацију и сваку величину популације израчуната је средња вредност одговарајуће метрике на основу три независна извршавања. На тај начин свакој величини популације одговара једна крива на графикону, која представља просечан ток најбољег, односно просечног фитнеса током 80 генерација.

Величина популације	Број понављања	Време [s]	Најбољи фитнес	Просечан фитнес
15	3	393	0.9284 ± 0.009	0.6819 ± 0.029
30	3	622	0.9386 ± 0.005	0.5885 ± 0.077
50	3	1013	0.9374 ± 0.012	0.6125 ± 0.069
100	3	1830	0.9422 ± 0.005	0.5712 ± 0.068

Табела 6.2: Резултати у зависности од величине популације

На слици 6.1 видљиво је да се највећи део побољшања најбољег фитнеса остварује у ранијим генерацијама. Код свих испитиваних величина популације вредност нагло расте у почетним генерацијама и већ око 15. генерације достиже вредност блиску коначној, након чега се криве постепено стабилизују. Ово указује на брзу почетну конвергенцију ка квалитетним решењима, док даљи ток еволуције у мањој мери доприноси побољшању најбоље пронађене јединке. Највиши ниво најбољег фитнеса у каснијим генерацијама остварује популација од 100 јединки, док популација од 30 јединки постиже сличан квалитет уз знатно краће време извршавања, приказано у табели 6.2.

Слика 6.2 показује другачију динамику просечног фитнеса популације. За разлику од најбољег фитнеса, просечни фитнес наставља да расте током већег дела извршавања, при чему се између различитих величина популације



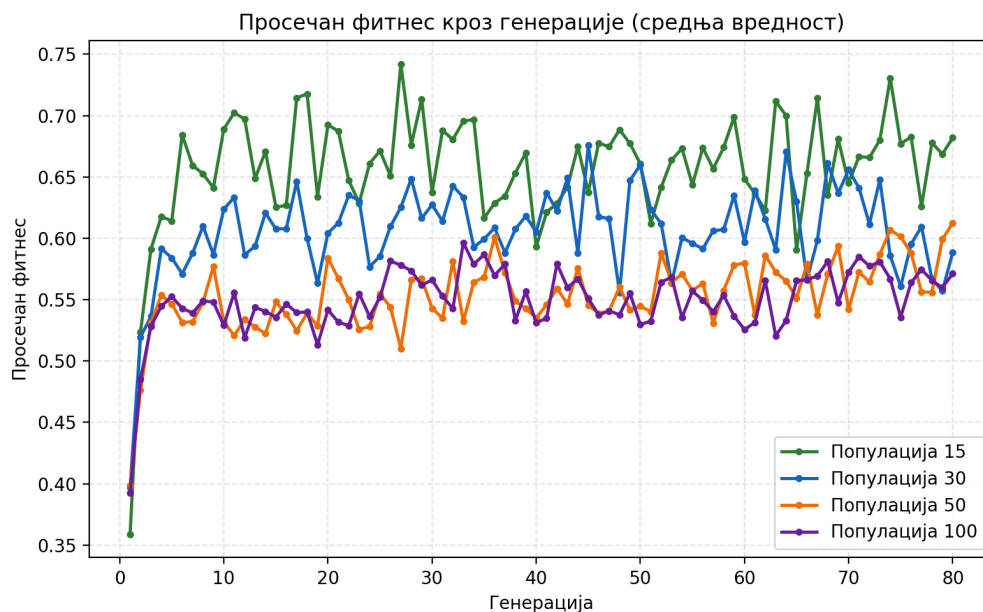
Слика 6.1: Ток еволуције најбољег фитнеса популације

не уочава јасан монотон тренд. Популација од 15 јединки остварује највиши просечни фитнес у последњој генерацији, што показује да већа популација не мора нужно довести до пропорционалног побољшања просечног квалитета свих јединки.

Посматрајући резултате из табеле 6.2 и приказане графиконе заједно, популација од 30 јединки представља одговарајући компромис између квалитета решења и времена извршавања. Иако веће популације у одређеним случајевима могу остварити нешто виши најбољи фитнес, побољшање није довољно велико да оправда додатно време извршавања. Због тога је величина популације од 30 јединки усвојена као стандардна вредност у експериментима описаним у наставку.

6.3 Поређење генетских оператора

Избор метода укрштања и мутације представља један од значајних фактора који утичу на понашање генетског алгоритма. Различити генетски оператори користе различите приступе измени молекулске структуре, па се могу разликовати како по способности генерисања квалитетних потомака, тако и по рачунској захтевности и учесталости успешног извођења. Због тога је



Слика 6.2: Ток еволуције просечног фитнеса популације

извршено експериментално поређење имплементираних метода укрштања и мутације.

Циљ експеримента је да се испита утицај избора генетских оператора на квалитет добијених решења и време извршавања генетског алгорита. Посебна пажња посвећена је и успешности извршавања самих оператора, будући да генерисање невалидних молекулских структура може утицати на ефикасност алгорита.

За потребе поређења разматране су четири конфигурације генетских оператора, при чему су за укрштање и мутацију коришћене исте врсте представљања молекула. Испитиване су комбинације приказане у табели 6.3.

Конфигурација	Укрштање	Мутација
SMILES	SMILES	SMILES
SELFIES	SELFIES	SELFIES
Граф	Граф	Граф
BRICS	BRICS	BRICS

Табела 6.3: Конфигурације генетских оператора коришћене у експерименту

Како би поређење различитих конфигурација било што поузданије, за све експерименте коришћена је иста почетна популација од 30 молекула. Састав

почетне популације фиксиран је пре извођења експеримената и није мењан између различитих конфигурација генетских оператора. Свака конфигурација је испитана у три независна покретања, при чему су за свако покретање забележени време извршавања, најбољи и просечни фитнес у свакој генерацији, као и успешност извршавања генетских оператора.

За најбољи и просечни фитнес у последњој генерацији израчунате су средња вредност и стандардна девијација на основу три независна покретања. За време извршавања израчуната је медијана три измерена времена. Стопа успешности генетских оператора одређена је на нивоу њиховог појединачног извршавања. Приликом једног извршавања оператор може више пута покушати да генерише валидну молекулску структуру. Извршавање се сматра успешним уколико је структура успешно генерисана у оквиру дозвољеног броја унутрашњих покушаја (који су приказани у табели 4.1), док се у супротном сматра неуспешним. Стопа успешности представља однос броја успешних извршавања оператора и укупног броја његових извршавања. Резултати експеримента приказани су у табели 6.4.

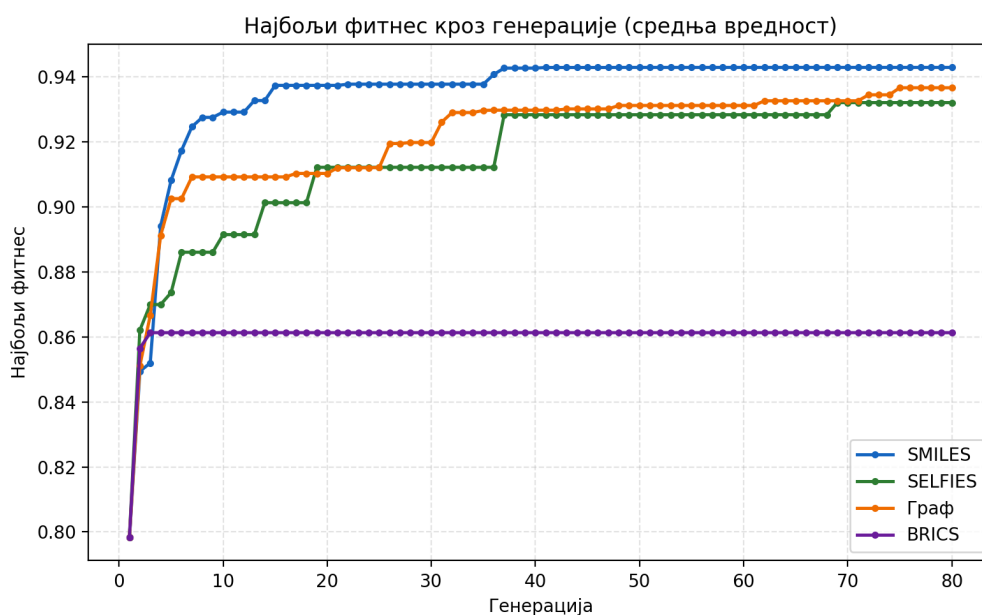
Конфигурација	Време [s]	Најбољи фитнес	Просечни фитнес	Укрштање [%]	Мутација [%]
SMILES	641	0.9429 ± 0.005	0.7055 ± 0.084	100.0	98.7
SELFIES	602	0.9321 ± 0.004	0.5557 ± 0.035	100.0	100.0
Граф	641	0.9367 ± 0.017	0.6775 ± 0.089	100.0	100.0
BRICS	5718	0.8613 ± 0.060	0.7752 ± 0.128	96.5	80.7

Табела 6.4: Резултати поређења генетских оператора

Праг броја унутрашњих покушаја пре проглашења неуспеха разликује се по операторима (табела 4.1), будући да оператори имају различиту вероватноћу генерисања валидних структура по покушају и различит рачунски трошак. Циљ је био да се у разумном времену постигне висока стопа успешног извршавања. Због различитих прагова унутрашњих покушаја, стопе успешности није могуће директно поредити између конфигурација, те оне не служе као критеријум за избор најбоље конфигурације, већ само као потврда да су у оквиру имплементације постављени довољно високи прагови.

Графици приказани на сликама 6.3 и 6.4 добијени су на исти начин као и одговарајући графици из одељка 6.2. У случају тока еволуције најбољег фитнеса, код свих конфигурација највећи део побољшања остварује се у ранијим

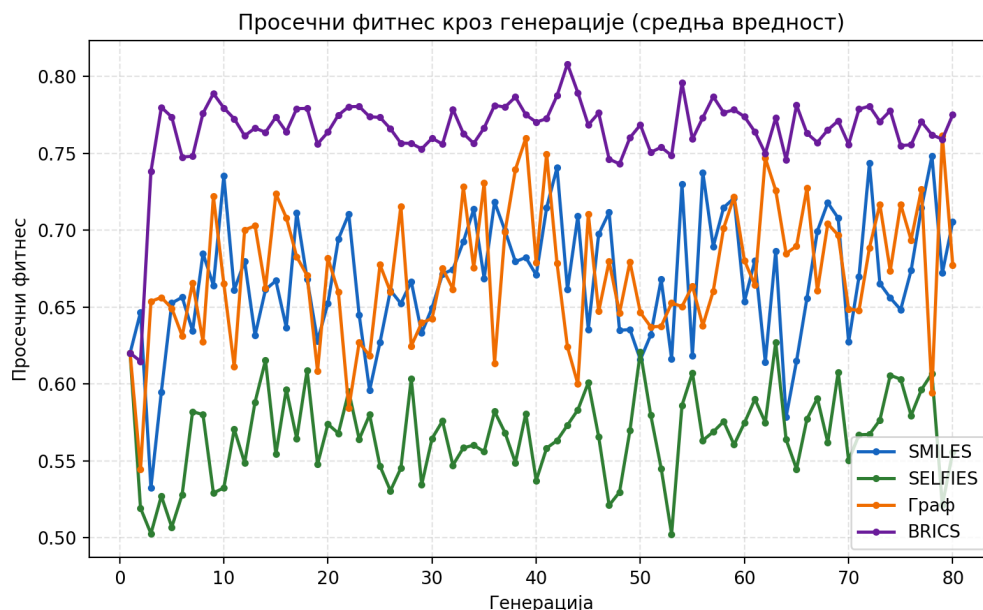
генерацијама, након чега долази до постепене стабилизације. SMILES конфигурација најбрже достиже висок ниво квалитета и остварује највишу коначну вредност најбољег фитнеса. Граф и SELFIES такође проналазе квалитетне јединке, али до њих долазе нешто спорије. Насупрот томе, BRICS конфигурација након почетног побољшања не показује даље напредовање, што указује на ограничену способност даљег истраживања простора решења.



Слика 6.3: Ток еволуције најбољег фитнеса за различите конфигурације генетских оператора

Слика 6.4 показује да динамика просечног фитнеса не прати нужно динамику најбољег фитнеса. Посебно се издваја BRICS конфигурација, код које је просечан фитнес релативно висок, али је најбољи фитнес знатно нижи у поређењу са осталим конфигурацијама. Додатном анализом диверзитета популације утврђено је да код ове конфигурације долази до изражене хомогенизације популације, односно да велики број јединки током еволуције постаје веома сличан. Висока вредност просечног фитнеса се стога не може тумачити као показатељ успешније оптимизације, већ пре као последица конвергенције великог дела популације ка сличним структурама. С друге стране, SELFIES одржава нижи просечни фитнес, што указује на већу хетерогеност популације, иако алгоритам и даље успева да пронађе квалитетне јединке.

Посматрано у целини, резултати показују да за разматрани задатак није



Слика 6.4: Ток еволуције просечног фитнеса за различите конфигурације генетских оператора

довољно посматрати само просечан фитнес популације. Значајнији су квалитет најбоље пронађене јединке, стабилност резултата и време потребно за извршавање алгорита. SMILES конфигурација остварује највишу вредност најбољег фитнеса, док SELFIES постиже нешто нижи коначни квалитет, али уз краће време извршавања и мању варијабилност најбољег резултата између понављања. Квалитет граф конфигурације упоредив је са SELFIES конфигурацијом, али су време извршавања и варијабилност најбољег резултата већи. BRICS конфигурација показује неповољну комбинацију слабог најбољег решења, изражене хомогенизације популације и значајно већег времена извршавања.

На основу наведених резултата, за наредне експерименте усвојена је SELFIES конфигурација, односно SELFIES укрштање и мутација. Она представља погодан компромис између квалитета најбољег пронађеног решења, стабилности извршавања и времена потребног за оптимизацију. Додатна предност ове конфигурације је стопроцентна успешност и укрштања и мутације у испитаним извршавањима, што указује на поуздано генерисање валидних молекулских структура.

6.4 Поређење фитнес функција

Циљ наредних експеримената је да се испита како избор фитнес функције утиче на понашање генетског алгорита, квалитет добијених структура и време извршавања. У експерименту су разматране две конфигурације: QED коефицијент, уз подразумевање тежине молекуларних дескриптора, и *Random Forest* модел за предикцију rIC50 вредности. Изабран је баш овај модел будући да је међу обученим моделима остварио најбоље перформансе на тестном скупу (поделељак 3.2). Сви остали параметри генетског алгорита засновани су на вредностима из табеле 6.1, као и на резултатима претходних експеримената.

За QED конфигурацију коришћени су резултати три независна покретања из одељка 6.3, будући да су генетски оператори и остали параметри алгорита били идентични. За *Random Forest* конфигурацију извршена су три нова независна покретања, над истом почетном популацијом од 30 јединки. Током сваког покретања бележени су време извршавања, најбољи и просечни фитнес по генерацијама. Ради упоредне анализе, на крају сваког покретања, на најбољој пронађеној јединки додатно је израчуната и секундарна мера: предвиђени rIC50 при QED оптимизацији, односно QED коефицијент при оптимизацији према rIC50.

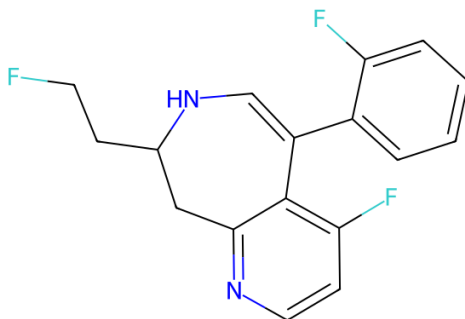
QED коефицијент и rIC50 припадају различитим скалама и представљају различите критеријуме процене, па се њихове апсолутне вредности фитнеса не могу директно поредити. Поређење је зато засновано на понашању алгорита унутар сваке конфигурације, времену извршавања, као и на вредностима секундарних мера израчунатих на најбоље пронађеним структурама.

Резултати приказани у табели 6.5 показују значајне разлике између две конфигурације. Оптимизација према QED доводи до постепеног побољшања квалитета најбоље јединке (што се може видети на слици 6.3, SELFIES конфигурација), док је код *Random Forest* конфигурације најбоља предвиђена rIC50 вредност остала непромењена током извршавања. У свих 80 генерација и у сва три независна покретања није пронађена јединка са већом предвиђеном rIC50 вредношћу од најбоље јединке иницијалне генерације. Поред тога, медијана времена извршавања код *Random Forest* конфигурације била је приближно 2,4 пута већа од медијане QED конфигурације, што је очекивана последица додатног израчунавања *Morgan fingerprint* вектора и примене предиктивног

модела за сваку оцењивану јединку. Секундарне мере на најбољим јединкама показују супротан однос између два критеријума: QED оптимизација доводи до структуре са високим QED коефицијентом и умереном предвиђеном активношћу, док pIC50 оптимизација даје веома високу предвиђену активност, али уз знатно нижи QED коефицијент. Резултати тако указују да различите функције фитнеса усмеравају генетски алгоритам ка структурно различитим решењима. Молекулске структуре глобално најбољих јединки приказане су на сликама 6.5 и 6.6.

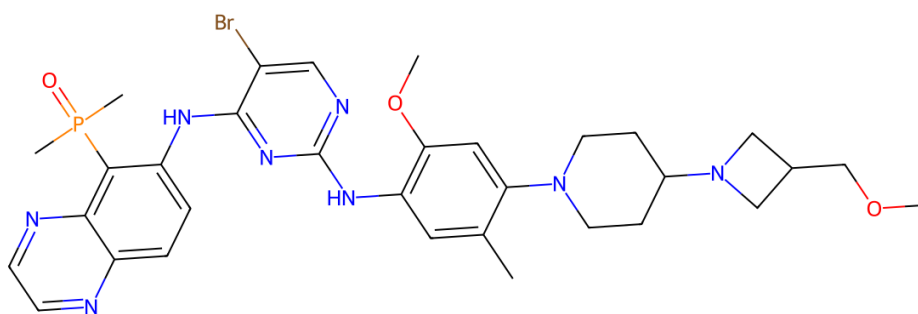
Фитнес	Време [s]	Најбољи фитнес	Просечни фитнес	QED најбоље	Предв. pIC50
QED	602	0,9321 $\pm$ 0,004	0,5557 $\pm$ 0,035	0,9365	6,35
Random Forest	1458	10,1311 $\pm$ 0,000	7,5079 $\pm$ 0,410	0,1968	10,1311

Табела 6.5: Резултати поређења функција фитнеса



Слика 6.5: Молекулска структура најбољег QED фитнеса

Посматрано заједно, резултати показују да избор функције фитнеса представља компромис између брзине оптимизације, QED коефицијента и предвиђене биолошке активности. QED конфигурација омогућава бржу претрагу и повољнији QED профил, док *Random Forest* конфигурација даје вишу предвиђену активност, али уз дуже извршавање и структуре које се јасно разликују од QED оптимума.



Слика 6.6: Молекулска структура најбољег предвиђеног pIC₅₀

Глава 7

Закључак

У овом раду развијен је софтверски систем за генерисање, процену и оптимизацију молекулских структура са потенцијално лековитим дејством применом генетских алгоритама. Систем обједињује више начина представљања, измене и евалуације квалитета молекула, као и графички кориснички интерфејс који омогућава подешавање параметара оптимизације и праћење њеног тока. Посебна пажња посвећена је испитивању различитих величина популације, генетских оператора и функције фитнеса, како би се утврдило на који начин њихов избор утиче на квалитет добијених структура и рачунарску сложеност оптимизације.

Експерименти са различитим величинама популације показали су да повећање популације не доводи нужно до пропорционалног побољшања резултата. Веће популације омогућавају разноврснији простор за претрагу и стабилније понашање алгорита, али истовремено значајно повећавају рачунарско време. На основу добијених резултата, популација величине 30 представља погодан компромис између квалитета оптимизације и времена извршавања, док су веће популације оправдане у ситуацијама у којима је приоритет исцрпнија претрага простора решења.

Испитано је више приступа укрштању и мутацији молекулских структура, укључујући приступе засноване на SMILES и SELFIES репрезентацијама, директној измени молекулског графа и BRICS декомпозицији. Поређење четири конфигурације генетских оператора (одељак 6.3) показало је да избор оператора представља компромис између квалитета добијених решења, поузданости извршавања и рачунарске ефикасности. На основу ових резултата, SELFIES конфигурација је одабрана за даље експерименте као најбољи ком-

промис између наведених аспеката.

Посебно је испитан и утицај функције фитнеса. QED је коришћен као брзо доступна мера општег квалитета структуре, заснована на физичко-хемијским дескрипторима молекула, док је за циљану оптимизацију биолошке активности развијен предиктивни приступ заснован на машинском учењу. За предвиђање pIC50 вредности EGFR инхибитора обучени су *Ridge* регресија, *Random Forest* и *LightGBM* модели. *Random Forest* је постигао најбоље резултате на независном тест скупу. Добијени резултати су упоредиви са резултатима релевантних истраживања из литературе.

Поређење различитих функција фитнеса показало је да избор фитнеса значајно одређује карактер добијене оптимизације. QED омогућава брзо вредновање великог броја структура и погодан је када је циљ генерисање молекула са општим својствима повољним са становишта леколикости. Са друге стране, фитнес заснован на предвиђеној pIC50 вредности омогућава усмеравање претраге ка молекулима за које се очекује већа активност према конкретној биолошкој мети, али уз већи рачунарски трошак и зависност од способности предиктивног модела. Стога се не може издвојити једна функција фитнеса која је оптимална за све сценарије примене, већ њен избор треба прилагодити циљу оптимизације.

Иако резултати показују да развијени систем може успешно да генерише и оптимизује молекулске структуре, постоји више ограничења која треба узети у обзир. Експерименти су спроведени на ограниченом броју конфигурација генетског алгорита и појединачни резултати не представљају исцрпну претрагу целокупног простора могућих параметара. Такође, експерименти са различитим конфигурацијама засновани су на ограниченом броју понављања, па одређени степен случајне варијабилности остаје присутан. Предиктивни модели су обучени за једну биолошку мету, EGFR, и на једном скупу података, па се њихове перформансе не могу аутоматски генерализовати на друге мете или скупове података.

Додатно ограничење представља чињеница да предиктивна вредност модела не представља експериментално потврђену биолошку активност. Молекул који модел процени као перспективан кандидат захтева даљу валидацију, која може обухватати додатне рачунарске методе, као и експериментално испитивање. Због тога резултате генетског алгорита треба посматрати као начин за генерисање и рангирање потенцијално занимљивих структура, а не као

коначну потврду њиховог лековитог дејства.

Развијени систем пружа основу за више праваца даљег развоја. Један од природних праваца је увођење вишекритеријумске функције фитнеса која би истовремено узимала у обзир предвиђену биолошку активност, QED коефицијент и друге релевантне молекулске особине. На тај начин било би могуће смањити ризик да оптимизација доведе до структура које имају високу предвиђену активност, али неповољна друга својства. Систем би се такође могао проширити на више биолошких мета и већи број предиктивних модела, чиме би се омогућила флексибилнија примена у различитим сценаријима откривања лекова.

Даљи рад може обухватити и унапређење самог генетског алгоритма, пре свега испитивањем различитих стратегија селекције, елитизма и контроле разноврсности популације. Посебно би било значајно испитати начине за спречавање преране конвергенције и побољшање истраживања простора молекулских структура. Ефикасност се такође значајно може побољшати паралелизацијом у оквиру процеса селекције, укрштања, мутације и евалуације. На страни предиктивних модела могуће је испитати и сложеније приступе, укључујући моделе који директно користе SMILES репрезентацију или молекулски граф. Посебно су занимљиви модели засновани на трансформер архитектури, који могу директно да оперишу над SMILES или графовском репрезентацијом. На нивоу припреме података, уместо агрегације поновљених мерења максимумом могуће је испитати и робусније статистике, пре свега медијану pIC50 вредности. Поређење модела обучених на овако добијеним циљним вредностима показало би у којој мери избор агрегације утиче на тачност предикције и на понашање генетског алгоритма.

Коначно, један од најважнијих праваца даљег развоја представљала би експериментална валидација најперспективнијих молекула добијених оптимизацијом. На тај начин би се могло испитати у којој мери предвиђене активности и остала својства одговарају стварним експерименталним вредностима и тиме проценити практична вредност развијеног система у процесу откривања нових лекова.

На основу спроведених експеримената може се закључити да је постављени циљ рада остварен: развијен је функционалан и конфигурабилан систем који омогућава примену генетских алгоритама за аутоматизовано генерисање и оптимизацију молекулских структура, уз могућност коришћења како оп-

пште мере леколикости, тако и предиктивног модела специфичног за изабрану биолошку мету. Експериментална анализа је показала да избор представљања молекула, генетских оператора, величине популације и функције фитнеса значајно утиче на понашање алгоритма, при чему не постоји једна конфигурација која је оптимална за све сценарије. Добијени резултати зато представљају основу за даље унапређивање система и његову примену у сложенијим задацима рачунарски потпомогнутог откривања лекова.

Литература

- [1] P. A. Akinnusi, G. D. Egunjobi, A. Akinnusi, T. Oluwadare, F. Ogunbiyi, and S. A. Shodehinde. Machine learning-guided drug repurposing for EGFR inhibition using scaffold-split validation, docking, and molecular dynamics. *Journal of Computer-Aided Molecular Design*, 40(1):142, June 2026.
- [2] Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887--2893, July 1996.
- [3] G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90--98, 2012.
- [4] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using ‘Drug-Like’ chemical fragment spaces. *ChemMedChem*, 3(10):1503--1507, 2008.
- [5] A. Gupta, A. S. Thind, and R. Purohit. EGFRAP: a predictive machine learning model for assessing small molecule activity against the epidermal growth factor receptor. *RSC Medicinal Chemistry*, 16(9):4415--4426, July 2025.
- [6] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, October 2020.
- [7] Gregory Landrum. RDKit: Open-source cheminformatics, May 2006. <https://www.rdkit.org>.

- [8] Thanh-Hoang Nguyen-Vo, Paul Teesdale-Spittle, Joanne Harvey, and Binh Nguyen. Molecular representations in bio-cheminformatics. *Memetic Computing*, 16:519--536, July 2024.
- [9] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742--754, May 2010.
- [10] David Weininger. SMILES, 1. introduction and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31--36, 1988.
- [11] Peter Willett. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, 11(23):1046--1053, 2006.

Биографија аутора

Лазар Савић (*Београд, 19. фебруар 2003.*) завршио је основну школу „Душко Радовић” у Сремчици, а потом и Математичку гимназију у Београду, са просечном оценом 5,00. Као највећи успех средњошколског образовања издвајају се две бронзане и једна сребрна медаља на интернационалним хемијским олимпијадама. Основне студије на Математичком факултету Универзитета у Београду, на студијском програму Информатика, уписао је 2021. године, а завршио 2025, са просечном оценом 10,00. Исте године уписао је и мастер студије, на истом студијском програму. Запослен је у компанији „Cisco” као софтверски инжењер од августа 2025.