ПЛОВДИВСКИ УНИВЕРСИТЕТ “ПАИСИЙ ХИЛЕНДАРСКИ”
ХИМИЧЕСКИ ФАКУЛТЕТ
КАТЕДРА АНАЛИТИЧНА ХИМИЯ И КОМПЮТЪРНА ХИМИЯ

Веселина Христева Паскалева

Компютърно генериране на тавтомери на органични съединения и приложения в химичната информатика

А В Т О Р Е Ф Е Р А Т

на дисертационен труд
за присъждане на образователната и научна степен
“ДОКТОР”

Област на висше образование: 4. Природни науки, математика и информатика
Професионално направление: 4.2. Химически науки
Докторска програма: Теоретична химия

Научен ръководител:
Доц. д-р. Николай Т. Кочев

Рецензенти:
1. …
2. ….

Пловдив 2016

Дисертационният труд е обсъден и насрочен за защита от катедрения съвет на катедра „Аналитична химия и компютърна химия“ на Химическия факултет при Пловдивския университет „Паисий Хилендарски“ на 09.05.2016г.

Дисертационният труд съдържа *177* страници (формат А4), в които са включени *76* фигури и *27* таблици. Цитирани са *157* източника.

Материалите по защитата са на разположение на интересуващите се в отдел „Развитие на академичния състав и докторантури“ към ПУ „Паисий Хилендарски“ и Националния център за информация и документация към Министерството на образованието, младежта и науката.

Научно жури:

Член-кор. проф. дбн Илза Пъжева

Проф. дхн Иван Бангов

Доц. д-р Иванка Цаковска

Доц. д-р Пламен Пенчев

Доц. д-р Николай Кочев

Защитата на дисертационния труд ще се състои на ……. от …… часа в зала ……. на ПУ „Паисий Хилендарски“, на заседание на Научно жури. Материалите по защитата са на разположение на интересуващите се в Централната библиотека на ПУ „Паисий Хилендарски“.

1. Въведение

Тавтомерните форми на химичните съединения предизвикват интерес в различни области на науката. Третирането на тавтомерията е предизвикателство както в експерименталните методи за анаилиз, така и при компютърната обработка на структурна информация. Проблемите в Химичната информатика предизвикани от тавтомерията се дължат на различното структурно представяне на отделните тавтомерни форми, което е следствие на различната им топология. В много от случаите тавтомерите се съхраняват в компютърните системи като отделни структури, което води до изчисляване на различни стойности на молекулните дескриптори и фингърпринти за отделните тавтомери. Тавтомерията влияе върху прилагането на методи за отчитане на структурно подобие, регистриране и търсене в химични бази данни, прилагане и построяване на методите за QSAR/QSPR.

Понастоящем няма разработени стандартни процедури за отчитане и обхващане на тавтомерната информация при различните методи и процедури прилагани в Химичната информатика. Това налага вземането на индивидуални решения дали тавтомерията на ползваните съединения да бъде отчетена или пренебрегната при извършваните изследвания и по какъв начин да стане това. Популярни процедури прилагани за обработка на тавтомерни съединения са нормализация или канонизация. При методите за канонизиране обикновено се използват алгоритми за генериране на всичките тавтомерни форми. Често поради технически затруднения или липса на софтуерен генератор за тавтомери, тавтомерията се пренебрегва при прилагането на методи за обработка на структурна информация.

Въпреки напредъка в областта изследването на тавтомерните съединения, се повдигат въпроси, които чакат своето решение. В тази насока особено полезни са алгоритмите за изчерпателно генериране на тавтомерни форми. В настоящата дисертация сравнихме 5 известни нам и достъпни тавтомерни генератори. Наблюденията ни показват, че повечето софтуери пропускат значителен брой тавтомери, като дори водещите според нас софтуерни системи (CACTVS и ChemAxon) имат някои проблеми. Горните факти мотивираха целта на настоящата дисертация.

2. Цел и задачи на изследването

Целта на настоящия дисертационен труд е *да се разработят подходи за компютърно генериране на всички теоретично възможни тавтомерни форми на органични съединения и прилагането им при решаване на основни задачи в химичната информатика*. За постигане на целта бяха формулирани следните основни задачи:

1. Разучаване на съществуващи софтуерни генератори и/или публикувани алгоритми за генериране на тавтомери;
2. Създаване на основна база данни с тавтомерни правила покриващи основните типове тавтомерия;
3. Реализиране на класически комбинаторен подход за генериране на тавтомери. Тестване на подхода. Откриване и анализ на проблемни случаи;
4. Разработване на нови алгоритми за генериране на тавтомери;
5. Разработване на алгоритми и правила за филтриране на получените тавтомери;
6. Разработване на алгоритми за ранкиране на тавтомерите и база данни с ранкиращи правила;
7. Програмна реализация в системата по химична информатика Ambit;
8. Оптимизиране на създадените алгоритми за генериране, филтриране и ранкиране по отношение на коректност на химичната информация и бързодействие;
9. Сравняване на разработените алгоритми с програми за генериране на тавтомери от литературната справка;
10. Реализация на стратегия за оптимално автоматично генериране на тавтомери базирана на комбинация от разработените алгоритми, подобрения на базата данни с правилата, филтрите и методите за ранкиране;
11. Да се изследва влиянието на тавтомерията при решаване на основни задачи в химичната информатика;
12. Да се предложат стратегии за оптимално включване на тавтомерната информации при решаване на задачи в химичната информатика.

3. Резултати

3.1 Разработване на алгоритми за генериране на тавтомери

3.1.1 База с тавтомерни правила

Създадохме база данни с предварително дефинирани тавтомерни правила, които се основават на дефиницията за тавтомерия по IUPAC. За описанието на правилата въведохме синтаксис основан на употребата на ключови думи (Таблица 1).

Таблица 1 Ключови думи използвани за дефиниране на тавтомерно правило

Ключова дума	Описание
$$ NAME	име на тавтомерното правило
$$ TYPE	тип тавтомерия: MOBILE_GROUP или RING_CHAIN
$$ GROUP	тип на мобилната група
$$ STATES	възможни състояния на фрагмента, в следствие на тавтомерната промяна
$$ GROUP_POS	позиция на мобилната група, т.е. номера на атома, към който е прикачена
$$ INFO	допълнителна информация

Правилата са дефинирани с две възможни състояния, описани в ключова дума **STATES**. Възможните състояния на фрагмента, засегнат от тавтомерната трансформация, отговарящи на състоянията на правилата, се описват чрез **SMARTS** заявки. Пример за дефинирано правило с две възможни състояния е представен на Фигура 1.

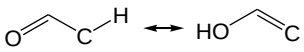
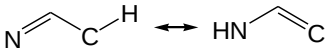
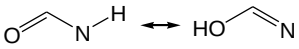
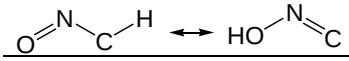
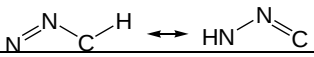
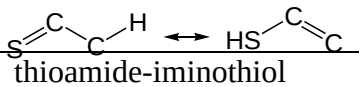
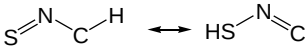
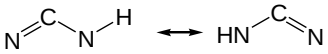
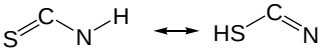
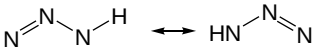
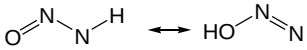
$$ NAME: keto/enol
$$ TYPE: MOBILE_GROUP
$$ GROUP: H
$$ STATES: [#8]=[#6][#6] [#8][#6]=[#6]
$$ GROUP_POS: 3,1
$$ INFO: O=CC

Фигура 1 Описание на keto/enol-но правило

Правилото описва прототропна тавтомерия, характеризирана с кето-енолна трансформация и миграция на мобилна група на 1-3 разстояние.

Списък с предварително дефинираните основни тавтомерни правила описващи 1-3 миграции на протон е даден в Таблица 2.

Таблица 2 Списък с предварително дефинираните 1-3 тавтомерни правила

1-3 миграция на протон (proton shift)		
keto-enol	imin-amin	amide-imid
O=CC ↔ O-C=C	N=CC ↔ N-C=C	O=CN ↔ O-C=N
		
nitroso-oxime	azo-hydrazone	thioketo-thioenol
O=NC ↔ O-N=C	N=NC ↔ N-N=C	S=CC ↔ S-C=C
		
thionitroso-thiooxime	amidine-imidine	thioamide-iminothiol
S=NC ↔ S-N=C	N=CN ↔ N-C=N	S=CN ↔ S-C=N
		
diazoamino-diazoamino	nitrosamine-diazohydroxide	
N=NN ↔ N-N=N	O=NN ↔ O-N=N	
		

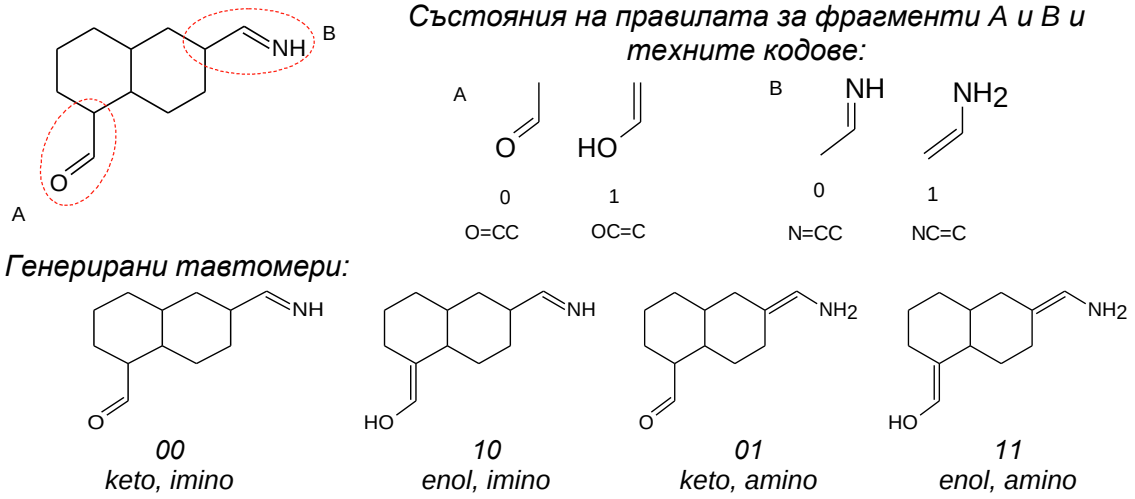
Допълнително бяха създадени правила описващи миграция на по-голямо топологично разстояние: 1-5 и 1-7 тавтомерни правила.

3.1.2 Реализация на Класически комбинаторен подход (СМ) за генериране на тавтомери

Разработихме класически комбинаторен подход (СМ) за генериране на тавтомерните форми. Той се основава на две основни стъпки. При *първата стъпка* се намират всички състояния на правилата, които присъстват в структурата, за която искаме да генерираме тавтомерите, като използваме алгоритъм за субструктурно търсене. Намерените структурни фрагменти съответстващи на състояния на правилата се наричат инстанции (rule instance). Състоянията на всяко правило са кодирани с 0 и 1, като всяка една намерена инстанция съответства на едно от състоянията на правилото (0 или 1). Цялостното състояние на молекулата се формализира като бинарно число с *n* на брой двоични цифри $b_1 \quad b_2 \ldots \quad b_n$, *където* $b_k = 0$ *или* 1, *за* $k = 1, 2, \ldots, n$. Като *втора стъпка* се обхождат всички възможни комбинации от състоянията на намерените инстанции в първа стъпка. Броят на възможните тавтомери може да се оцени като 2^n, където *n* е броят намерени тавтомерни правила. Генерирането на възможните комбинации от състоянията на намерени инстанции, се извършва

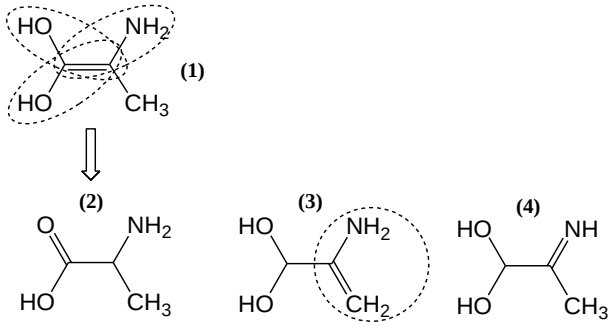
като частта от структурата, която не участва в тавтомерните региони, се запазва постоянна. Промяна се прави само в тавтомерните региони при трансформация на едно състояние в друго. Обхождайки всички възможни комбинации от състоянията на всяка инстанция получаваме крайната извадка от тавтомери за целевата структура.

На Фигура 2 е даден пример с молекулата на 6-(iminomethyl) decahydronaphthalene-1-carbaldehyde. Като резултат от субструктурното търсене са намерени две инстанции - фрагменти **A** и **B**. Фрагмент **A** (инстанция на keto-enol правило) е намерен в състояние 0 - "keto". Фрагмент **B** (инстанция на правило amino-imino) е намерен в състояние 0 - "imino".



Фигура 2 Приложение на СМ за генериране тавтомерите на молекулата на 6-(iminomethyl) decahydronaphthalene-1-carbaldehyde.

В случаите, когато има силно припокриване на намерените инстанции прилагането на комбинаторния метод е затруднено. Такъв е примерът с alanine (Вж. Фигура 3). Трите намерени инстанции на правила споделят два атома и двойна връзка. Като резултат, алгоритъмът генерира тавтомери **(2)** и **(4)**, но пропуска тавтомер **(3).** Ако за стартова структура е избрана структура **(3)** се намира само едно правило (imine/amine) и като резултат се генерира един тавтомер - структура **(4)**, а се пропускат тавтомери **(1)** и **(2)**.



Фигура 3 Генериране на тавтомерите за структурата на alanine със силно припокриване на правилата.

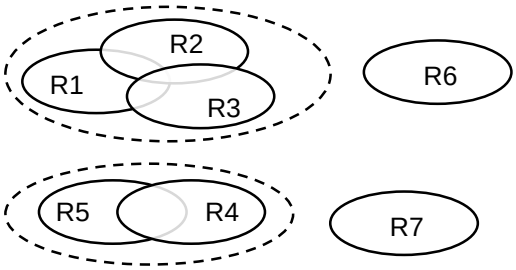
3.1.3 Разработване на нови алгоритми за генериране на тавтомери

3.1.3.1 Подобрен комбинаторен метод (CMI)

Вторият разработен подход е подобреният комбинаторен метод (CMI). Той също се основава на две стъпки – (1) формиране на кластери от инстанции на правила и (2) формиране на подкомбинации (Фигура 4):

(1) Множеството от всички инстанции на правилата, които се припокриват се групират в кластер. Оформят се няколко кластера, като инстанциите на правилата от различните кластери не се припокриват;

(2) Дадена под-комбинация се получава по следния начин: избира се по една инстанция на правило от всеки кластер. По този начин всяка под-комбинация съдържа само неприпокриващи се правила. Всяка под-комбинация съдържа толкова на брой правила, колкото е броят на кластерите. Броят под-комбинации е равен на произведението от размерите на всички кластери.



Формиране на кластери:
{R1, R2, R3}
{R4, R5}
{R6}
{R7}

Формиране на подкомбинации:
R1, R4, R6, R7
R1, R5, R6, R7
R2, R4, R6, R7
R2, R5, R6, R7
R3, R4, R6, R7
R3, R5, R6, R7

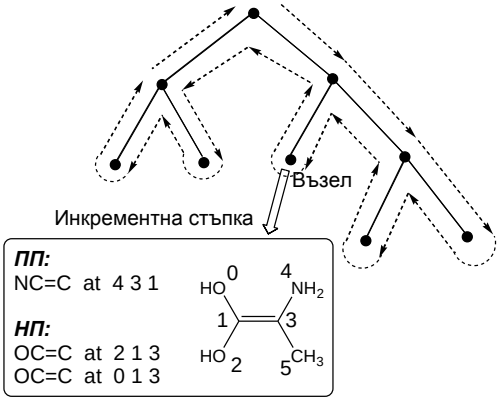
Фигура 4 Основни стъпки при приложение на CMI

След формиране на подкомбинациите се прилага CM метода за всяка от тях. По този начин CM се прилага многократно върху неприпокриващи се инстанции на

правила. Този подход води до значително подобрение на класическия комбинаторен метод.

3.1.3.2 Инкрементен подход (IA-DFS)

Третият разработен подход за генериране на тавтомери е базиран на алгоритъм за търсене в дълбочина (Depth First Search). Той може да се визуализира, като обхождане на дървовиден граф (Вж. Фигура 5).



Фигура 5 Визуализация на алгоритъма за търсене в дълбочина

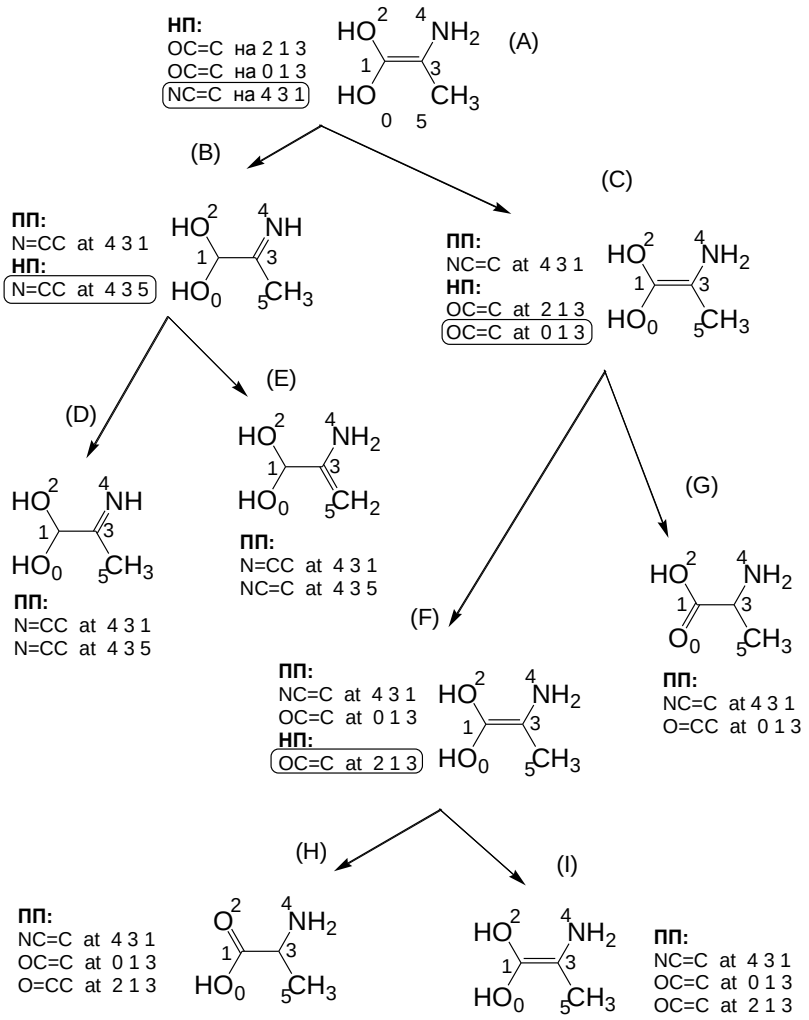
Чрез обхождане на дървото се претърсва пространството на всички възможни тавтомерни форми дефинирани чрез приложение на правилата от базата знание. Всеки възел от дървото представено на Фигура 5 представлява конкретна конфигурация на целевата структура (позиция на двойните връзки и мобилните групи). Вътрешните възли от дървото представляват междинни структури използвани за изпълнение на следващите инкрементни стъпки. Крайните възли от дървото отговарят на инкрементни стъпки, които нямат наследници, т.е. това са генерираните тавтомери, тъй като всички правила вече са били приложени. Даден възел се получава итеративно по време на претърсваща процедура и се нарича "Инкрементна стъпка", която се дефинира чрез следните компоненти:

- структурна конфигурация
- списък с инстанциите на **приложените правила** (ПП)
- списък с инстанциите на **неприложените** правила (НП)

Генерирането на тавтомери, чрез Инкрементния подход (IA-DFS), се основава на няколко основни принципа:

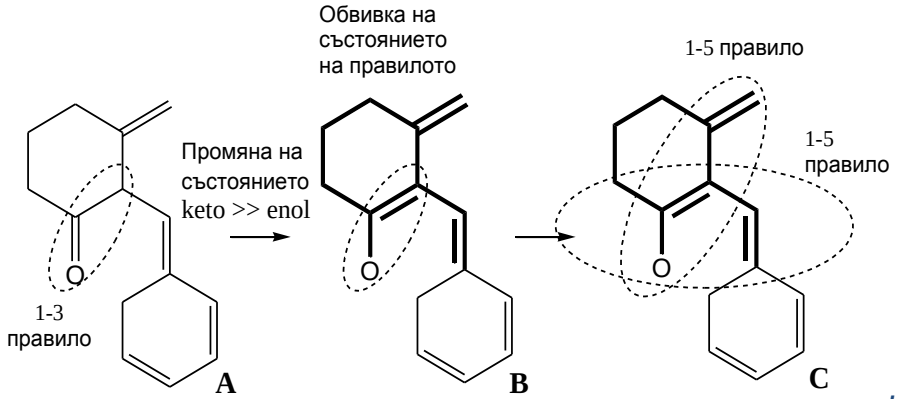
- Новите възли (Инкрементни стъпки) от дървото на търсене се генерират чрез избор на едно от неприложените правила (НП) и реализиране на всичките му възможни състояния. Използваната инстанция се премества от списъка с НП в този с ПП;
- Нов тавтомер се получава от определен възел, в който няма неприложени инстанции на правила (списъка с НП е празен);
- След прилагане на правило за дадена инстанция от списъка с НП, този списък се обновява, с цел обхващане промените в молекулната конфигурация;
- Обновяването става с претърсване, около приложеното правило, в топологична околност с определен размер. Тази околност се нарича "обвивка" на инстанцията.

На Фигура 6 е илюстрирано пълното дърво за генериране на тавтомери, обходено чрез прилагане на Инкрементния подход, за структурата на alanine. Коренът на дървото, възел **(A)**, има първоначално намерени три инстанции на правила: keto-enol (състояние OC=C) при атоми 2 1 3, keto-enol (състояние OC=C) при атоми 0 1 3 и imin-amin (състояние NC=C) при атоми 4 3 1. От правилото amin-imin са генерирани два дъщерни възела **(B)** и **(C)**, които съответстват на следните състояния на това правило: N=CC и NC=C. Съответно и в двете инкрементни стъпки (възли **(B)** и **(C)**) правилото imin-amin е преместено в списъка с приложени правила (ПП). Във възел **(B)**, списъкът с НП правила е обновен, което анулира валидността на keto-enol-ното правило (то се премахва от списъка с НП правила). В същото време е намерено ново правило, неприсъстващо в първоначалния списък, това е правилото imin-amin (състояние N=CC) при атоми 4 3 5. Наследниците на възел **(B)** са генерирани чрез прилагане на последното правило от списъка с НП правила в този етап, като резултат се получават възли **(D)** и **(E)**. За последните възли, списъкът с неприложени правила е празен, затова тези крайни възли представляват резултатни тавтомери. По аналогичен начин възел **(C)** е претърсен в дълбочина генерирайки възли **(F)**, **(G)**, **(H)**, **(I)**. Всички крайни възли са резултатни тавтомерни форми. Това са възли **(D)**, **(E)**, **(G) (H)** и **(I)**. Възел **(G)** е изоморфен на възел **(H)**, поради което само един от тях е запазен в крайния резултат.



Фигура 6 Дърво на решенията за alanine при прилагане на IA-DFS

Обновяването на списъка с правилата става чрез претърсване на топологична обвивка от ***k***-ти порядък за дадена инстанция на правило. Тази топологична обвивка се дефинира като всички атоми, които се намират на топологично разстояние **≤*k***, измерено за всеки атом от това правило. Понастоящем в Инкрементния подход се използват топологични обвивки от втори порядък (Вж. Фигура 7). На фигурата с удебеляване е посочена дефинираната топологична обвивка, в която е извършено ново търсене на валидни инстанции на правилата.



Фигура 7 Дефиниране на топологична обвивка на инстанцията на правило 1-3 и обновяване на правилата в обвивката

Изчерпателността на Инкрементния подход се обуславя от това, че на всяка инкрементна стъпка, може да се премине от едно тавтомерно „пространство“ в друго. Поради природата на алгоритъма за търсене в дълбочина, напълно са проверени всички тавтомерни пространства, т.е. всеки възел е обходен до най-дълбокото му ниво. Това позволява намиране на „скрити“ тавтомери и разрешава проблема с припокритите правила.

3.2 Разработване на алгоритми и правила за филтриране на получените тавтомери

Разработени са три филтъра за изчистване на дублиращи се тавтомерни структури: (1) базиран на вътрешна тавтомерно кодиране, (2) базиран на изоморфизъм и (3) базиран на InChI хеширани кодове.

Създадени са списък с предупредителни правила (*Warning List*) и списък с изключващи правила (*Exclusion List*). Те са формулирани като правила, описани с единични SMARTS заявки. Те служат за премахване на структури от генерирания набор от тавтомери отговарящи на предварително зададени критерии.

3.3 Разработване на алгоритми за ранкиране на тавтомери и база данни с ранкиращи правила

За изчисляване ранга на генерираните тавтомери беше реализирана адитивна схема, която се основава на емпирично изведени правила. Всяко едно от ранкиращите правила съдържа енергийни тегла за двете си състояния - ES_i^0, ES_i^1. Ранга на всеки тавтомер може да се изчисли по формулата:

$$\text{Енергийно тегло} = \sum_{i}^{N_{\text{правила}}} ES_i^{\text{състояние}(i)} + N_{\text{ароматниАтома}} C_{\text{аром.}}$$

където $ES_i^{\text{състояние}(i)}$ са енергийните тегла за правило в състояние *състояние(i)*, $N_{\text{ароматниАтома}}$ е броят на ароматните атоми, $C_{\text{аром.}}$ е корекцията за ароматен атом, емпирически определена, като -0.1eV. Рангът на даден тавтомер се изчислява като сума от всички енергийни тегла за всички 1-3 състояния на правилата присъстващи в тавтомера.

Описаните алгоритми за генериране на тавтомери са имплементирани като модул Ambit-Tautomer в софтуерната система Ambit.

4. Тестове. Резултати и дискусия

Тестовете бяха разделени на две категории (Вътрешни тестове и Външни тестове), всяка съдържаща по няколко различни теста. Вътрешните тестове имаха за цел да проверят ефективността на разработените алгоритми и да се сравнят резултатите получени от тях; да се провери как правилата за миграция на по-големи разстояния оказват влияние върху получаваните резултати; и до каква степен разработените подходи ще възпроизведат теоретично изчислени резултати. Външните тестове имаха за цел да се сравнят разработените подходи с резултатите на други софтуерни системи, включително тестове бързодействие. Извършените тестове са систематизирани както следва:

- Сравняване на резултатите получени от алгоритмите CMI и IA-DFS;
- Анализ на влиянието на 1-3, 1-5 и 1-7 правилата;
- Тестване еквивалентността на всички генерирани тавтомерни форми чрез IA-DFS алгоритъма;
- Сравняване на IA-DFS алгоритъма с няколко други софтуерни пакета за генериране на тавтомерни форми;
- Сравняване на резултатите на алгоритъма IA-DFS с теоретично изчислени такива;
- Сравнение на каноничните тавтомери генерирани чрез Ambit IA-DFS и тези генерирани чрез други софтуерни системи;
- Сравняване на скоростта на генериране на тавтомерни форми от алгоритмите на Ambit-Tautomer и други софтуерни системи;

За да осъществим тестовете бяха избрани и подготвени няколко структурни извадки. Най-голямата от които съдържа 204 114 структури част от базата с отворен достъп на NCI.

Обобщената процедура използвана за сравнение между два алгоритъма за генериране на тавтомерни форми (по-надолу отбелязани като A1 и A2) приложени върху *Извадка 1* се състои от следните стъпки:

i. алгоритмите се отбелязват като първи (или още основен алгоритъм, A1) и втори алгоритъм (A2);
ii. За всяка една от структурите в тестовата извадка получаваме броя генерирани тавтомери чрез алгоритъм A1, обозначен с **N1**, и броя генерирани тавтомери чрез алгоритъм A2, обозначен с **N2** ;
iii. за всяка структура изчисляваме разликата **Diff=N1-N2** и относителната разлика **relDiff=2Diff/(N1+N2)**;
iv. получаване на хистограми за разликата (**Diff**) и относителната разлика (**relDiff**).

Анализът на данните от стъпка (iv) е базиран изцяло на броя тавтомерни форми и не включва анализ на качеството им. Стъпка (iv) е груб критерий и със сигурност не дава пълна представа дали един алгоритъм е по-добър от друг. За голяма структурна извадка броят тавтомери е практически количествен индикатор. Сравнителни индикатори за качество са възможни за тестови извадки с малки структури, където резултатите получени от алгоритъма са сравнени с тези получени от експерт (или извлечени от литературни източници).

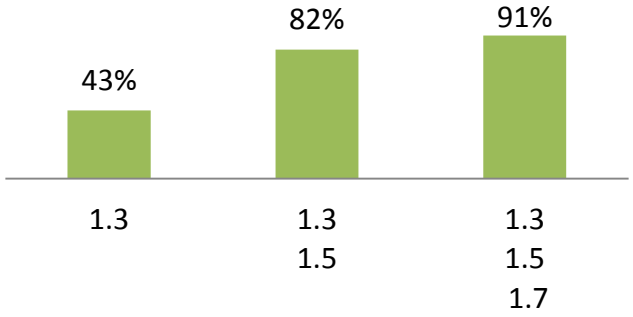
4.1 Вътрешни тестове

4.1.1 Сравнение с теоретично получени резултати. Анализ на резултатите при ползване на 1-3, 1-5 и 1-7 правила

За извършване на тези тестове беше използвана тестова *Извадка 3 (113 съединения).* Теоретичните резултати са получени чрез ръчно определяне на тавтомерите. Тавтомерните форми са генерирани, чрез приложение на Инкрементния подход, съответно (1) при включени само 1-3 правилата, (2) при използвани 1-3 и 1-5 правила и (3) при използване на 1-3, 1-5 и 1-7 правила.

Резултатите от направеното сравнение са обобщени на Фигура 8. При използване само на 1-3 правилтата при генерирането на тавтомери са възпроизведени 43% от теоретичните тавтомерни форми. Включването на 1-5

правилата подобрява възпроизводимостта до 82%. А трите вида правила в комбинация дават 91% от теоретичните тавтомери.



Фигура 8 Възпроизводимост на теоретично изброените тавтомери в %, при използване 1-3, 1-5 и 1-7 правилата

4.1.2 Еквивалентност на генерираните тавтомери

Важна характеристика на софтуерите за генериране на тавтомерни форми е еквивалентността на генерираните структури или така наречения тест за непротиворечивост. Тавтомерите се считат за еквивалентни, когато се генерира един и същ набор от тавтомерни форми, без значение кой тавтомер се взима за начална (целева) структура. Това не е тривиална задача, тъй като алгоритъмът за генериране на тавтомери не следва един и същ път на претърсване, при което за различните стартови структури се получават различни DFS дървета. DFS дърветата зависят също и от номерацията на атомите, която е произволна. Но крайният резултат се очаква да бъде един и същ, без значение атомното номериране и стартова структура.

За да проверим еквивалентността на тавтомерите генерирани чрез Инкрементния подход извършихме автоматични тестове за структурите от тестова извадка с 114 000 структури. На Фигура 9 е дадено разпределението на абсолютния брой нееквивалентни тавтомери.

Фигура 9 Разпределение на абсолютния брой не еквивалентни тавтомери

Получените резултати показват, че при повече от 82% от тестваните целеви структури имат напълно еквивалентни тавтомерни множества. В останалите тестови случаи разликата обикновенно е в порядъка на няколко тавтомерни форми (по-често една или две).

Тестовете за еквивалентност използвани за охарактеризирането на непротиворечивостта на генериране бяха особено полезни във фазата на разработване на софтуера. Анализирането на случаите, в които тестът за еквивалентност даваше пропуски, доведе до подобрение в разработения подход. Резултата от 82% непротиворечивост е получен чрез използване на Инкрементния алгоритъм без рекурсия и без промяна на първоначалните настройки и базата с правилата.

Фигура 10 Разпределение на средната разлика на генерираните не еквивалентни тавтомери

Анализът на резултатите показа, че една причина за пропуските в тестовете, в някои от случаите, е прилагане на Инкрементния алгоритъм със зададена забрана за връщане на правила от ПП списъка в НП списъка.

4.1.3 Сравнение на подобрения комбинаторен метод и инкрементния подход

За сравнение между Инкрементния подход (IA-DFS) и Подобрения комбинаторен (CMI) използвахме тестова *Извадка 1*. За нея генерирахме тавтомерните форми и чрез двата подхода, след което от броя генерирани тавтомери при единия подход е изваден броя генерирани тавтомери чрез другия подход (наречен за целите на теста *втори*). Разпределението на относителната разлика между броя генерирани тавтомери чрез IA-DFS (взет за основен алгоритъм) и съответно CMI е посочено на Фигура 11. Положителните бинове в хистограмата отговарят на структури, за които IA-DFS е генерирал повече тавтомерни форми от CMI (вторият алгоритъм в това сравнение). Съответно отрицателните бинове отговарят на структури, за които CMI е генерирал повече тавтомерни форми от IA-DFS.

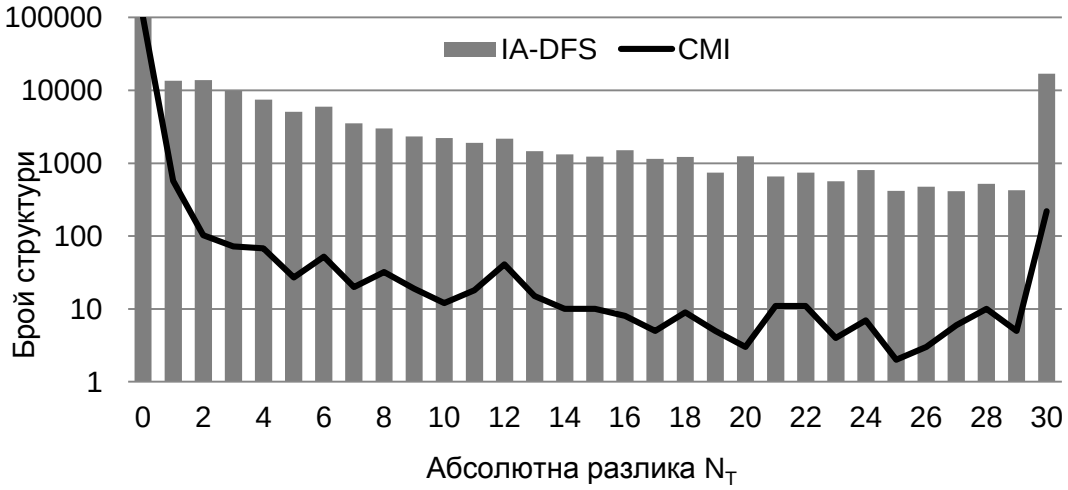
Фигура 11 Разпределение на относителната разлика между броя генерирани тавтомери чрез алгоритите IA-DFS и CMI

Хистограмата от Фигура 11 ясно показва, че IA-DFS генерира значително повече тавтомерни форми от IA-DFS. Много са структурите, за които IA-DFS генерира с 50% повече тавтомери.

CMI е полезен при работа с малки структури или ако е необходимо по-бързо генериране на тавтомерите за сметка на получаване по-малко химична

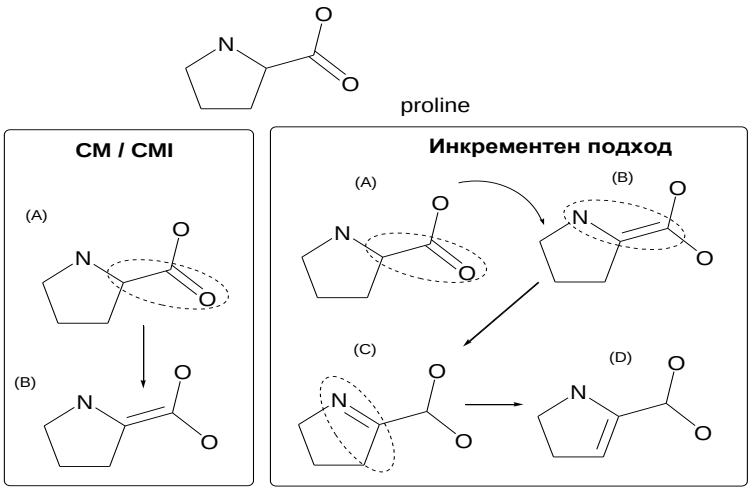
информация. Също така по време на тестовете забелязахме, че в някои случаи на пропуснати тавтомерни форми от IA-DFS същите се получават при използване на CMI.

На Фигура 12 е показана честотата на разпределение на абсолютната разлика между броя генерирани тавтомери с IA-DFS и тези с CMI. Фигурата показва, че като цяло IA-DFS генерира значително повече тавтомерни форми в сравнение с CMI. Черната линия отговаря на отрицателните бинове, т.е. броя на структурите за които IA-DFS генерира по-малко тавтомерни форми от CMI. Със стълбове са отбелязани случаите при които IA-DFS е по-изчерпателен.



Фигура 12 Разпределение на абсолютната разлика между броя генерирани тавтомери, чрез алгоритъма IA-DFS и CMI

Структурата на proline (Фигура 13) илюстрира приложението на разработените подходи за генериране на тавтомери.



Фигура 13 Генерирани тавтомери за молекулата на proline при приложение на CM/CMI (в ляво) и Инкрементния подход (в дясно)

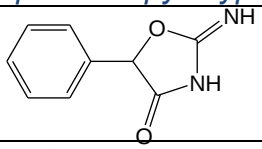
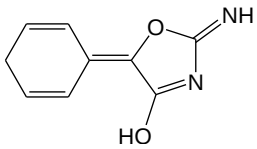
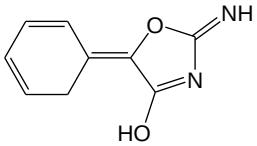
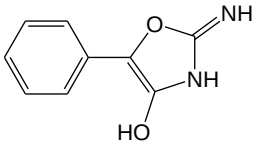
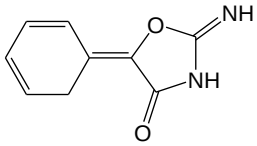
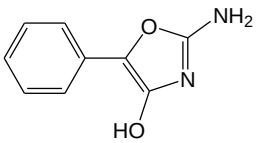
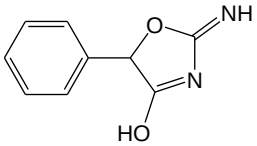
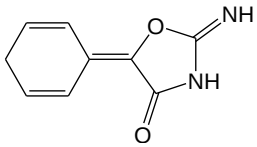
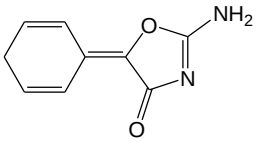
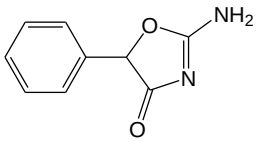
По-изчерпателният резултат даден от Инкрементния подход се дължи на опресняването на списъка с неприложени правила на всяка стъпка от процеса на генериране – при структури (B), (C) и (D).

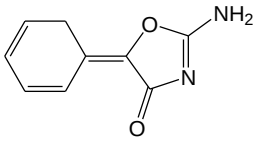
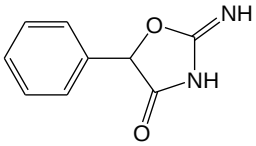
4.2 Външни тестове

4.2.1 Тестове за изчерпателно генериране. Сравнение с чужди подходи

Външните тестове представляват сравняване на резултатите между разработените тук алгоритми и външни програми за генериране на тавтомери. В Таблица 3 са представени 11-те генерирани тавтомерни форми за структурата на pemoline чрез IA-DFS, както и структурите получени чрез прилагане на ChemAxon Marvin, Cactvs, Mn.Tautomer, InChI/CDK и ACD Labs. Под всяка структурна диаграма от резултатите получени чрез IA-DFS, са посочени рангът на структурата и буквена абревиатура. Буквената абревиатура посочва чрез кой софтуер е получена тавтомерната форма - означенията са **M** - *Chem Axon Marvin*, **C** - *Cactvs*, с **A** е означен *ACD Labs*, с **T** е кодирано онлайн приложението за генериране на тавтомери *Mn.Tautomer* и с **I** алгоритъма базиран на линейна нотация *InChI* и реализиран в CDK. Структурата с най-ниско енергийно тегло получават ранг 1, тази с второ най-ниско тегло получава ранг 2 и т.н.

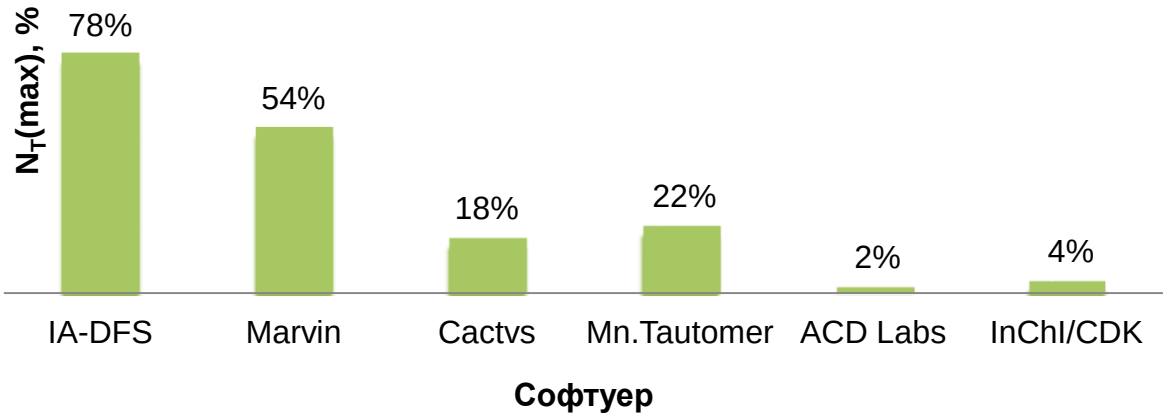
Таблица 3 Сравнение на резултатите на различни алгоритми за генериране на тавтомери за структурата на Pemoline

Целева структура: *Pemoline* 		
Инкрементен подход (IA-DFS)		
 ранг: 8	 ранг: M	 ранг: 4 M, C
 ранг: 6 M	 ранг: 1 M, C, A	 ранг: 7 M, C, T, A
		

ранг: 6	ранг: 5	ранг: 2	C, T, A, I
		*Marvin*	*8 т.ф.*
		Cactvs	*5 т.ф.*
		Mn.Tautomer	*3 т.ф*
		InChI/CDK	*3 т.ф*
ранг: 5 M	ранг: 3 M	*ACD Labs*	*3 т.ф*

Резултатите показват, че разработеният Инкрементен подход генерира три тавтомерни форми, които не се получават при използване на нито един от останалите софтуера. Също така IA-DFS генерира най-голям брой структури - 11, в сравнение с получените резултати от другите софтуери. Каноничният тавтомер според системата Cactvs съвпада с този получен чрез нашия метод.

Сравняването на резултатите между алгоритмите беше извършено върху структурите от тестова *Извадка 2*, съдържаща 139 широко използвани молекули. Резултатите са обобщени на Фигура 14. Сравнението е направено за максимален брой генерирани тавтомерни форми.



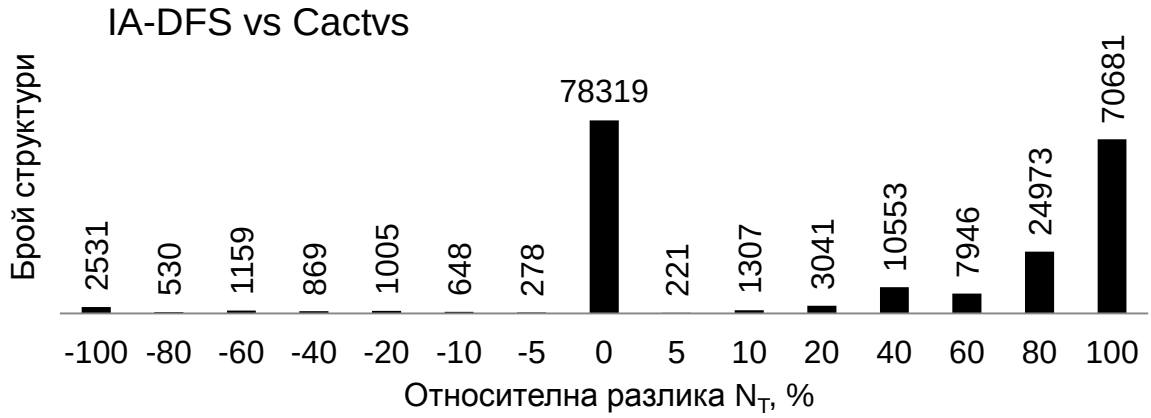
Фигура 14 Процентно разпределение на максималния брой генерирани тавтомери от тестваните шест алгоритъма за тестова **Извадка 2**

В 44% от случаите IA-DFS е единственият алгоритъм генериращ максимален брой тавтомерни форми. В 20% от случаите максималният брой тавтомери е генериран само от Marvin. От тестващите структури няма такива, за които останалите тествани софтуерни продукти да генерират максимален брой тавтомери самостоятелно. Случаи, в които максималният брой тавтомери е достигнат от повече от един алгоритъм са както следва: в 12% от случаите максимален брой тавтомери са генерирани от IA-DFS и Marvin. IA-DFS, Marvin,

Cactvs и Mn.Tautomer дават едновременно максимален брой тавтомери в 9% от случаите.

Резултатите открояват софтуерите Cactvs и Marvin като водещи и затова в тестовете с по-големи извадки сравненията са извършени само с тези два алгоритъма. Автоматичните тестове за сравнение на получаваните резултати от различни софтуерни пакети бяха извършени за голямата тестова *Извадка 1*, съдържащи 204 114 структури от базата данни със свободен достъп на NCI. За целите на тестовете използвахме Cactvs версия 3.393 и Marvin версия 5.10.2. Автоматичните тестове са базирани на броя генерирани тавтомерни форми, като резултатите са обобщени на Фигура 15, Фигура 16, Фигура 17 и Фигура 18.

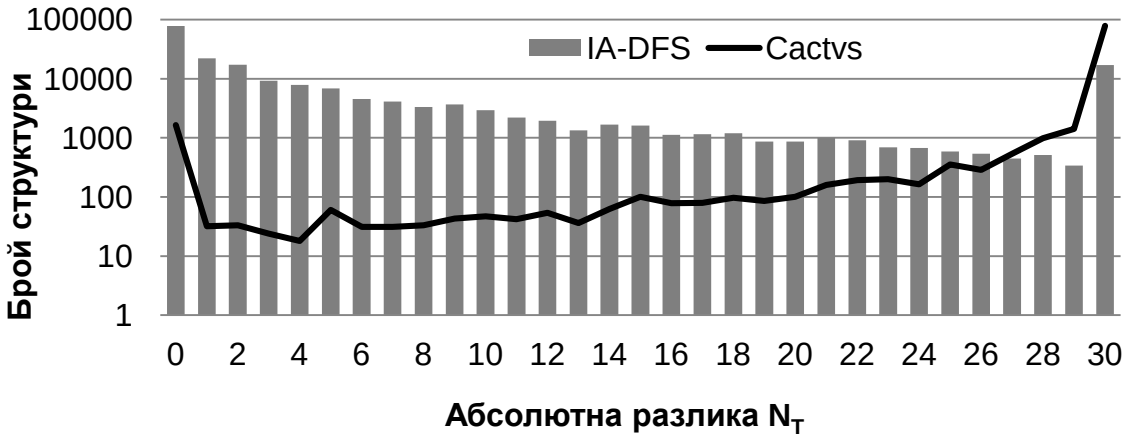
Фигура 15 представя разпределението на относителната разлика между броя генерирани тавтомери чрез IA-DFS и броя генерирани тавтомери чрез системата Cactvs. Положителните бинове представляват случаите, в които чрез IA-DFS са генерирани повече тавтомери, в сравнение с тези генерирани чрез системата Cactvs. Отрицателните бинове показват случаите, когато Cactvs генерира повече тавтомерни форми от IA-DFS. Както се вижда от фигурата значително по-голям е броят на случаите, в които Инкрементния подход генерира повече тавтомери. При 78 319 структури двата алгоритъма дават еднакъв брой тавтомери. При 3041 структури IA-DFS с 20% повече от Cactvs. А при 1005 структури Cactvs генерира с 20% повече тавтомерни форми.



Фигура 15 Разпределение на относителната разлика между броя генерирани тавтомери чрез IA-DFS и броя генерирани тавтомери чрез CACTVS

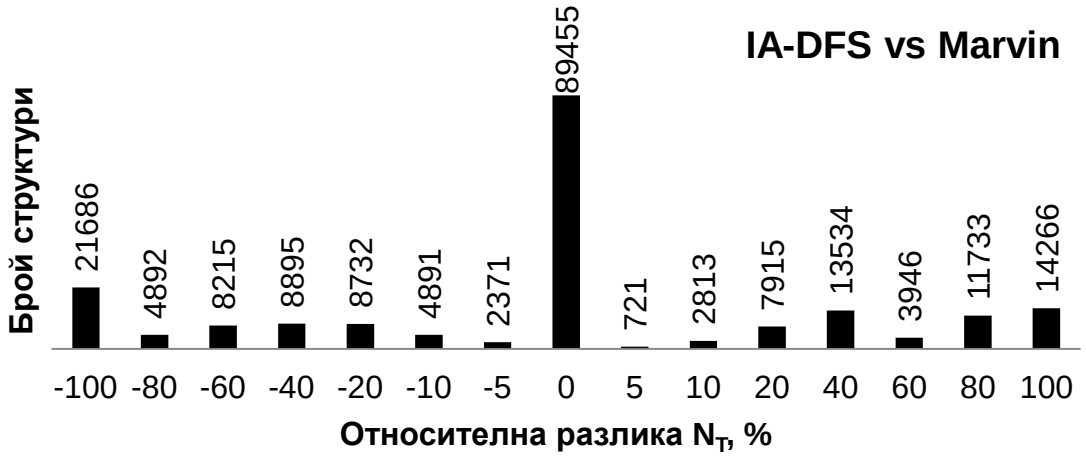
На Фигура 16 е показано разпределение на абсолютната разлика между броя генерирани тавтомери чрез IA-DFS и броя генерирани тавтомери чрез CACTVS. Отрицателните бинове, представени с черна линия на графиката,

показва разпределението на структурите, за които алгоритъмът реализиран в Cactvs е по-изчерпателен от IA-DFS. За структури, при които наблюдаваната разлика в броя генерирани тавтомери е между 0 и 26 IA-DFS е генерирал по-голям брой тавтомери. Cactvs показва по-добри резултати за структури с голям брой тавтомерни структури. Предполагаме, че по-малкият брой генерирани тавтомерни структури от разработения от нас подхос в тези случаи се дължи на избраните параметри за генериране на тавтомери, а именно работа при включени 1-3 и 1-5 правила и ограничение в броя стъпки-назад разрешени на алгоритъма.



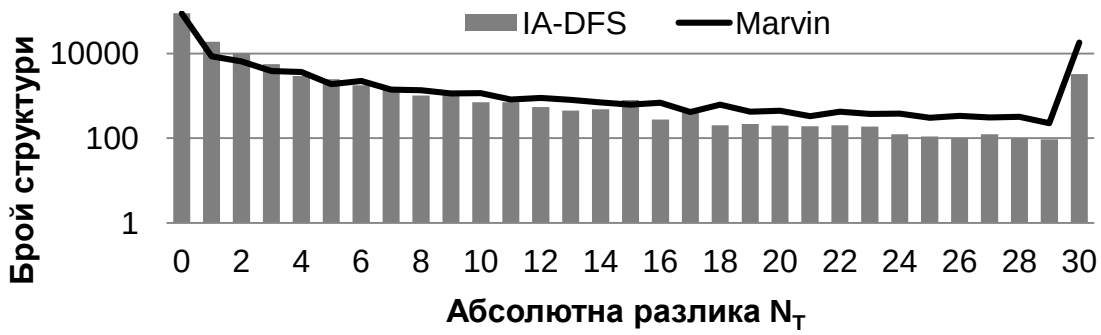
Фигура 16 Разпределение на абсолютната разлика между броя генерирани тавтомери чрез Ambit IA-DFS и броя генерирани тавтомери чрез CACTVS

Аналогични сравнителни тестове бяха направени и между IA-DFS и алгоритъма реализиран в Marvin (Фигура 17 и Фигура 18).



Фигура 17 Разпределение на относителната разлика между броя генерирани тавтомери чрез IA-DFS и броя генерирани тавтомери чрез Marvin

Резултатите от Фигура 17 и Фигура 18 показват, че IA-DSF и Marvin са сравнително близки. В някои от случаите Marvin има малко по-добри резултати. Обикновено по-големи разлики се получават за структури с много голям брой тавтомерни форми.

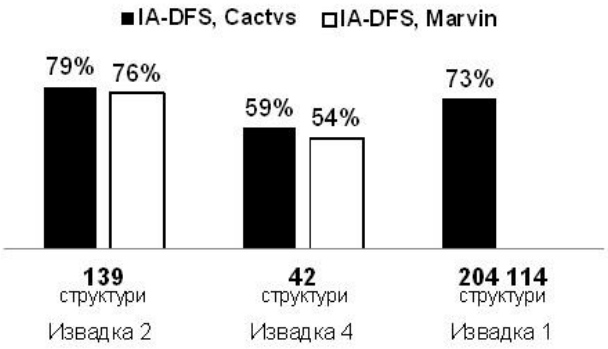


Фигура 18 Разпределение на абсолютната разлика между броя генерирани тавтомери чрез IA-DFS и броя генерирани тавтомери чрез Marvin

Системата на Cactvs беше сравнена и с Подобрения комбинаторен метод (CMI). Резултатите показаха, че двата подхода са съпоставими по брой генерирани тавтомерни структури.

4.2.2 Сравнение на каноничните тавтомери

Под каноничен тавтомер, в разработената от нас система, се разбира този с най-ниско енергийно тегло. Стратегията за намиране на канонична форма понастоящем изисква предварително генериране на всички тавтомерни форми. Проверихме до каква степен генерираните канонични тавтомери от IA-DFS съвпадат с тези получени чрез приложение на Cactvs и Marvin. Сравнението беше извършено върху три извадки с тестови структури - *Извадки 1* (204114 структури), *2* (139 структури) и *4* (42 структури). Резултатите са обобщени на Фигура 19.



Фигура 19 Сравнение на канонична тавтомерна форма между IA-DFS, Cactvs и Marvin за три структурни извадки

Сравнението на каноничния тавтомер за структурите от тестова *Извадка 4* показва, че в 59% от случаите IA-DFS и Cactvs генерират един и същ тавтомер, докато каноничният тавтомер на IA-DFS и Marvin съвпадат в 54% от случаите. За *Извадка 1* софтуерите IA-DFS и Cactvs дават един и същ каноничен тавтомер за приблизително 150 000 структури. Разлика се наблюдаваше за около 50 000 съединения. За много от тези случаи Инкрементният подход определи повече от една структура с най-ниско енергийно тегло като, след тези структури се намираше структурата, която е каноничната според Cactvs.

4.2.3 Тестове за бързодействие на софтуера

Тези тестове бяха осъществени върху структурна извадка от 10 682 структури. Изчисленията бяха извършени на компютър със следните характеристики: **CPU**: Intel Core i3-2330M 2.2 GHz; **RAM**: 8 GB. Резултатите са представени в Таблица 4. Отчитано е общото време за работа на софтуера включващо генерирането на тавтомерите и четене/писане от/във файл.

Таблица 4 Резултат от сравнителните тестове за бързодействие, получени при анализ на тестова Извадка 5

Алгоритъм (софтуер)	Време (час:мин:сек)
ChemAxon Marvin	01:34:23
Cactvs	00:39:30
Ambit-IA-DFS	01:01:42
Ambit-CMI	00:19:35
Ambit-CM	00:20:23

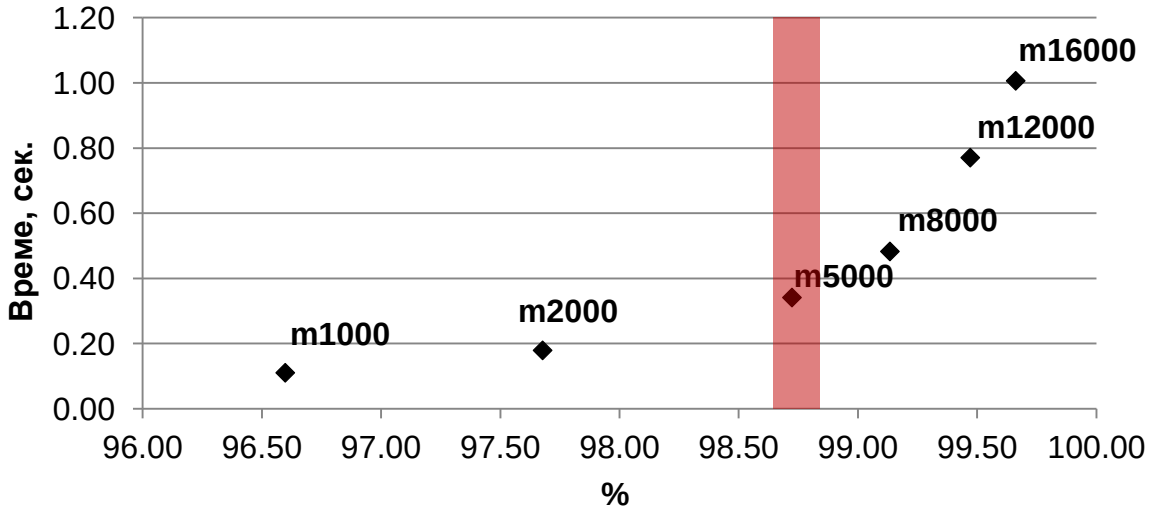
Резултатите показват, че софтуерът на Cactvs е най-ефективен от гледна точка на бързодействие. Скоростта на Инкрементния подход е достатъчна за работа с големи структурни извадки за разумен интервал време. Подобреният комбинаторен метод е почти три пъти по-бърз от Инкрементия, но „цената“ за тази скорост е пропускането на някои тавтомерни форми.

5. Оптимизация на алгоритмите за генериране на тавтомери, филтриране и бързодействие

5.1 Оптимизиране стойността на параметъра *m* (максимален брой „стъпки-назад“)

Стойността на параметъра **m** е от значение. При ниска стойност, процесът на генериране на тавтомерни форми ще бъде прекъснат рано и ще бъдат генерирани ограничен брой структури. Това може да доведе до пропускане на важни структури, но процесът на генериране ще приключи по-бързо. При големи стойности на *m* времето необходимо за генериране се увеличава, но се генерират по-голям брой тавтомерни форми. Тестовете извършени за определяне на оптимална стойност на този парамтър са извършени на лаптоп Lenovo IdeaPad Z570, Операционна система Windows 7 64bits, Intel(R) Core(TM) i3-2330M CPU @ 2.20GHz, RAM 8.00 GB.

За тестовете бяха използвани публично достъпни структури част от TOX21S. Общият брой използвани структури е 5292 молекули. Генерирането на тавтомерите беше извършено при максимален брой "стъпки-назад" равен на *m* = 1000, 2000, 5000, 8000, 12000, 16000 и 20000. При извършения тест отчитахме както броят на генерираните структури така и времето, за което беше извършено генерирането (това е чистото време на генериране, без включване на времето необходимо за четене на входящия файл и запис на резултатите в изходящия файл). Условно приехме, че при стойност на параметъра от *m* = 20000 се генерират практически всички възможни тавтомери (~ 100% от тях). Резултатите от направените тестове за различен брой максимални "стъпки-назад" е представени на Фигура 20 като зависимост между процента генерирани тавтомерни форми и средното време необходимо за генерирането им.



Фигура 20 Избор на оптимална стойност на параметъра "стъпки-назад", m

Процентното съотношение е изчислено като брой генерирани тавтомери към броя генерирани такива при "стъпки-назад" (*m*)=20 000. За оптимална стойност избрахме m=5000, тъй като се получават над 98% от тавтомерите при подобряване на бързодействието с приблизително 70%.

5.2 Оптимизация на филтриращата процедура

Тестовете показаха, че филтрирането на дублиращите се структури въз основа на изоморфизъм е факторът, който забавя значително действието на софтуера. Въведохме нов филтър базиран на линейна нотация InChI. С цел да се увеличи бързодействието без загуба на генерирани тавтомери. Тестовете бяха извършени върху подмножество от публични структури, които са част от база данни на ChEMBL. Времето за изпълнение беше отчетено по три параметъра - общо време, време за генериране и време за четене и писане от/във файл. Тестовете бяха извършени чрез прилагане на филтър за премахване на дублиращи се структури базиран на изоморфизъм и базиран на InChI. Получените резултати представени в Таблица 5.

Таблица 5 Тестване на бързодействието при употреба на InChI филтър и Изоморфизъм филтър

	Използван филтър:	
	Изоморфизъм, s [%]	*InChI, s [%]*
Общо време	313	179
Време за генериране	261 [83%]	124 [69%]
Време за IO/convert	53 [17%]	55 [31%]

От Таблица 5 ред 2, се вижда че филтърът базиран на InChI е повече от два пъти по-бърз, като практически запазва резултатите на филтъра чрез структурен изоморфизъм.

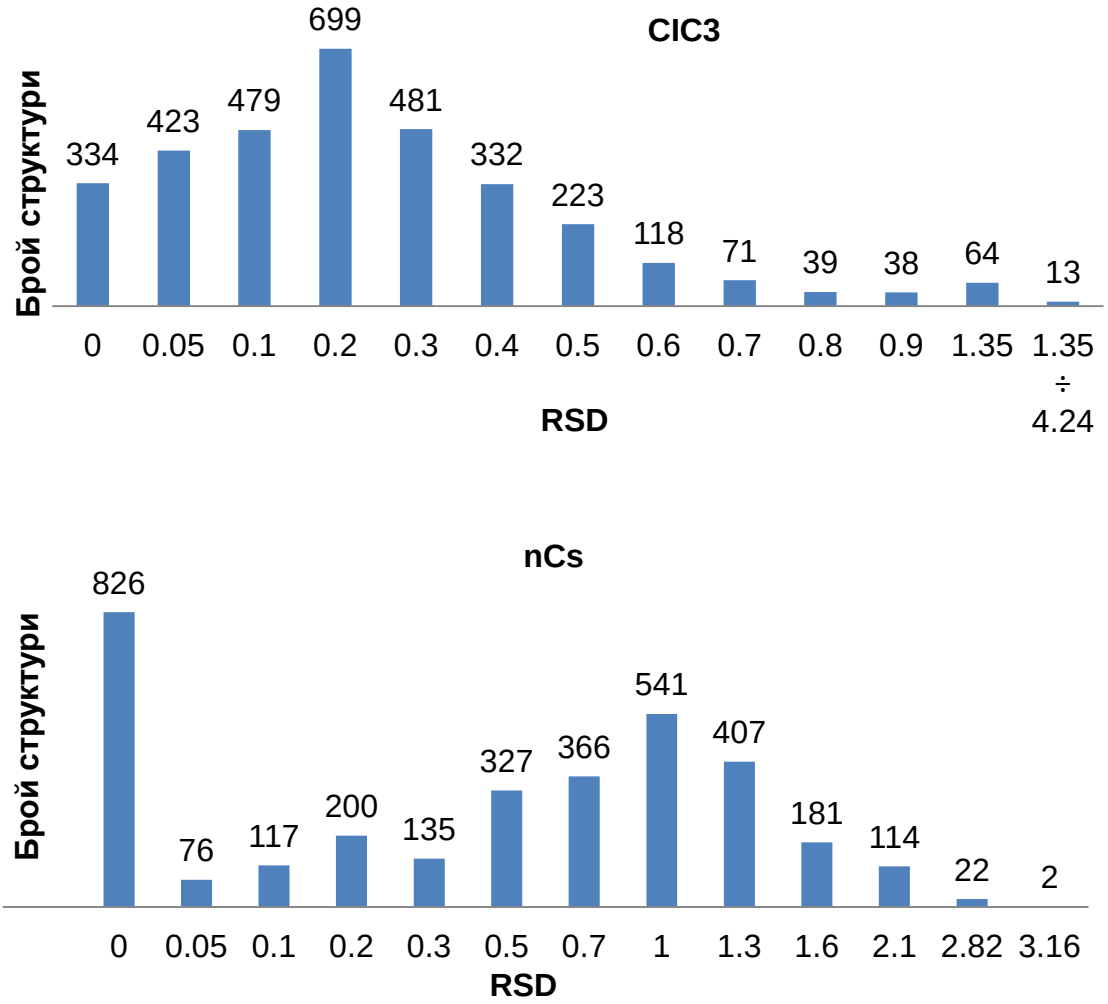
6. Изследване влиянието на тавтомерите при решаване на основни проблеми в химичната информатика

6.1 Влияние на тавтомерията при изчисляване на молекулни дескриптори

За извършване на тези тестове бяха подготвени четири структурни извадки. Извадките съдържат общо 126 542 молекули, за които бяха генерирани 1 773 781 тавтомерни форми чрез прилагане на разработения Инкрементен подход. За

структурите от всички извадки бяха изчислени дескрипторите реализирани в системата на Dragon (923 дескриптора) и тези на PaDEL (728 дескриптора). За всяка структура беше изчислена вариацията (RSD) на стойностите за всеки отделен дескриптор дължаща се на тавтомерните форми на структурата. След това беше изчислена средната вариация за дескриптор върху цялата извадка от съединения - meanRSD, чрез която охарактеризирахме тавтомерното влияние върху даден дескриптор.

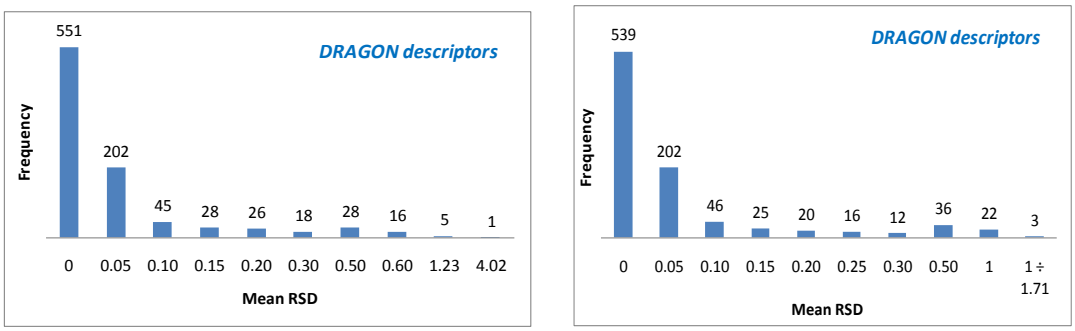
Например meanRSD за дескрипторите CIS3 (complementary information content /симетрия на съседите от трети порядък/) и nCs (общ брой вторични въглеродни атоми) изчислена за *Извадка 9* е 0.23 и 0.57 съответно. На Фигура 21 е показано разпределението на относителното стандартно отклонение за дескрипторите CIS3 и nCs за същата извадка.



Фигура 21 Разпределение на стойностите на относителното стандартно отклонение изчислено за 3314 структури за дескрипторите CIC3 и nCs

От фигурата се вижда, че 27% от структурите в извадката показват вариация в изчислените стойности за тавтомерите си по-голяма от 0.30 за дескриптор CIS3. За дескриптор nCs 25% от структурите не показват никаква вариация в стойността на дескриптора за тавтомерните си форми, но при 59% от тях вариацията е над 0.30.

На Фигура 22 е представено разпределението на средното относително стандартно отклонение (meanRSD) за дескрипторите от софтуера Dragon, изчислени за структурните извадки на NCI и DrugBank. За двете извадки се получиха близки резултати.



Фигура 22 Разпределение на mean RSD за дескрипторите изчислени с DRAGON за структурните извадки на NCI (в ляво) и DrugBunk (в дясно)

Около 60% от дескрипторите изчислени за извадките NCI и DrugBank не се влияят от тавтомерната информация. Малка вариация на стойностите на дескрипторите се наблюдава за около 30% (meanRSD в интервала [0.05,0.15]). Вариация в изчислените стойности на дескрипторите meanRSD > 0.20 се наблюдава при 10% от дескрипторите. Някои дескриптори проявяват вариация в RSD по-голямо от 1.00. Такива дескриптори са хидрофобните и хидрофилните фактори, донорите на Н-връзка и други.

6.2 Влияние на тавтомерията върху изчисляване на молекулни фингърпринти

Осъществихме предварителни тестове с три малки молекули с широко практическо приложение - methimazole, violuric acid и pemoline. Генерирахме тавтомерните им форми чрез прилагане на подхода IA-DFS. За всички получени тавтомери бяха изчислени две групи фингърпринти: CDK Fingerprinter (1024 бита) и PubChem fingerprints (881 бита), реализирани в системата PaDEL. За тях

определихме броя битове имащи различна стойност при отделните тавтомерни форми. Резултатите са посочени в Таблица 6.

Таблица 6 Брой битове на фингърпринти повлияни от тавтомерните форми

Структура	CDK FP (1024)	PubChem (881)
methimazole	145	56
violuric acid	545	106
pemoline	666	132

Резултатите показват, че за CDK Fingerprinter тавтомерните форми на pemoline повлияват 65% (666 от общо 1024) от битовете и 15% от битовете на PubChem. За структурата на methimazole 14% от битовете на CDK Fingerprinter и 6% от тези на PubChem са повлияни от трите генерирани тавтомера на структурата. Съответно при violuric acid битовете на CDK Fingerprinter (53%) се повлияват в по-голяма степен в сравнение с тези на PubChem (12%).

За отчитане влиянието на тавтомерната информация беше въведена оценка, подобна на относителното стандартно отклонение, която нарекохме ***Относително битово отклонение (RBD)***. Тя показва вариацията в изчислените стойности на конкретен бит от фингърпринта за всяка от тавтомерните форми на едно съединение. За к-ти бит, RBD_k се дефинира като относителния брой тавтомери, при които k-ти бит има стойност различна от тази при оригиналната структура.

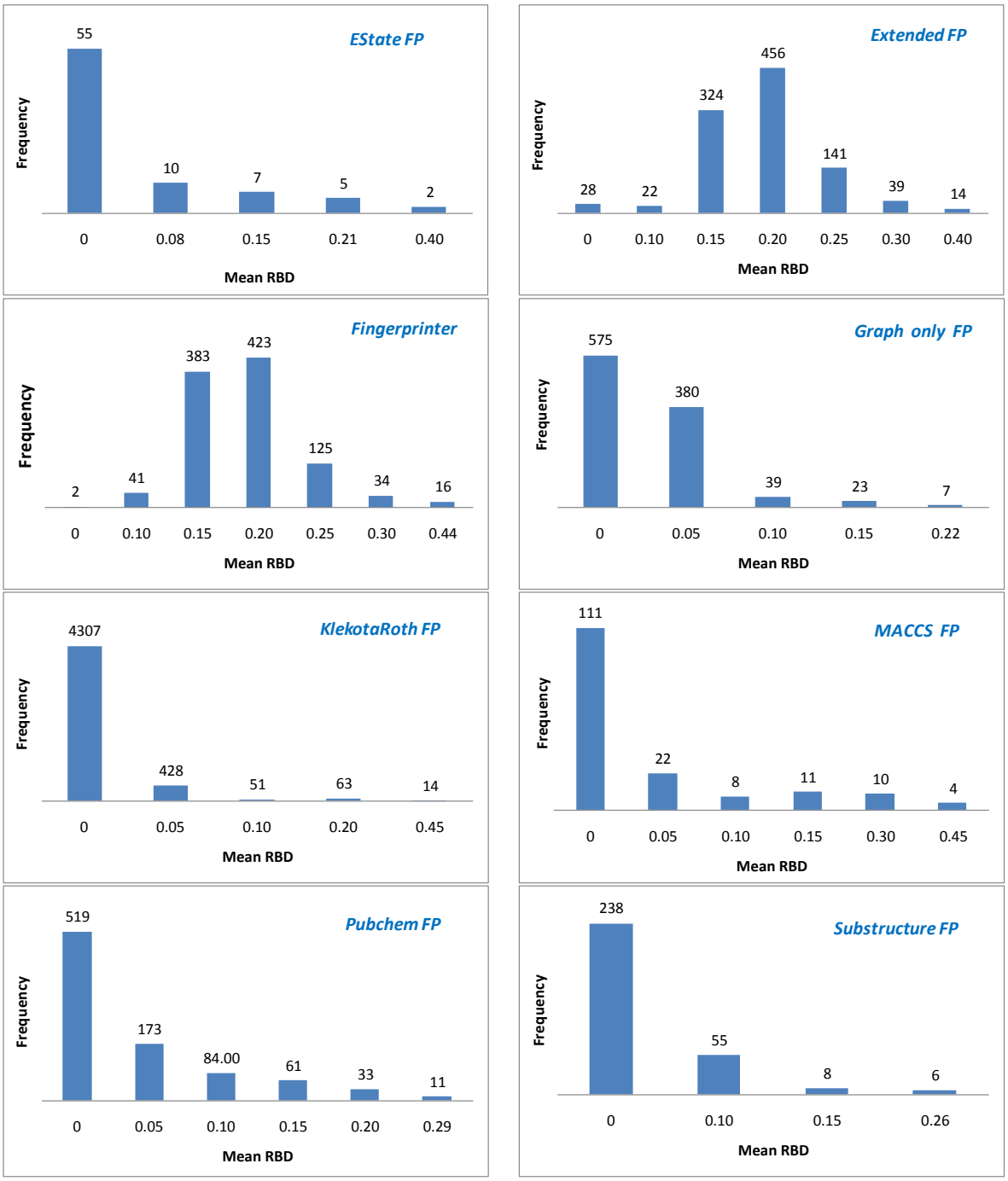
Средното относително битово отклонение (meanRDB) е оценка на вариацията на бит от фингърпринта за всички структури от извадката:

$$meanRBD_k = \frac{1}{N_{(str.)}} \sum_{i=1}^{N_{(str.)}} RBD_k^{(i)}$$

където $N_{(str.)}$ представлява броя структури в извадката, а $RBD_k^{(i)}$ е относителната битова вариация за бит k за съединение (i).

Тестовете бяха извършени върху осем типа фингърпринти реализирани в PaDEL - Fingerprinter (1024), Extended Fingerprinter (1024), EState Fingerprinter (79), Graph Only Fingerprinter (1024), MACCS Fingerprinter (166), Pubchem Fingerprinter (881), Substructure Fingerprinter (307), KlekotaRoth Fingerprinter (4 860). Използвахме 4 структурни извадки: Извадка 9 (Ames) с 3 365 структури (и 67 577 тавтомера), Извадка 10 (DrugBank) с 4 570 (и 169 227 тавтомера), Извадка

11 (LogP) с 8 327 (и 133 768 тавтомера), Извадка 12 (NCI) с 71 854 (и 1 312 621 тавтомера). Изчислихме $meanRBD_{\text{к}}$ за всички извадки. Разпределението на $meanRBD_{\text{к}}$ за различните фингърпринти за *Извадка 10* (DrugBank) е представено на Фигура 23. За останалите извадки се наблюдава подобно разпределение.



Фигура 23 Разпределение на средната относителна битова вариация за Извадка 10 - DrugBank

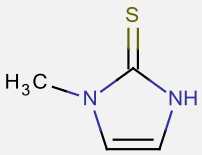
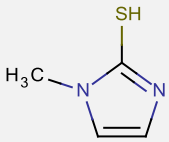
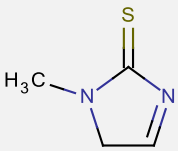
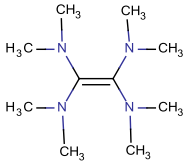
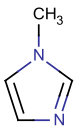
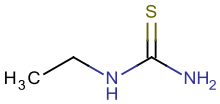
Резултатите, които получихме за средното относително битово отклонение за **Fingerprinter** са близки с тези получени за **Extended Fingerprinter**, което се

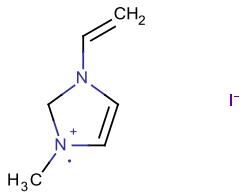
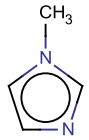
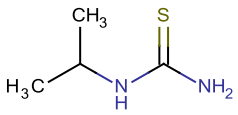
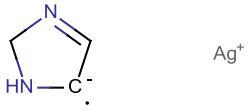
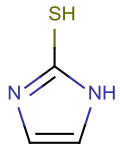
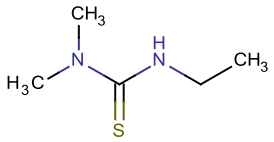
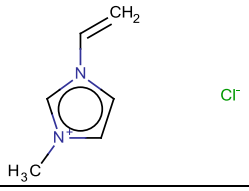
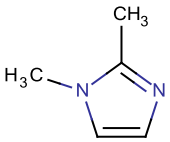
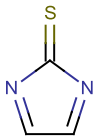
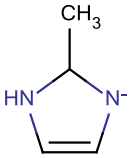
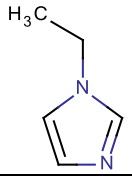
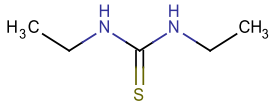
наблюдава и при четирите използвани извадки. Тези фингърпринти са най-силно повлияни от тавтомерната информация - само 2-3 бита за **Fingerprinter** и 28 за **Extended Fingerprinter** не са повлияни. Най-голям брой битове показват вариация между 0.10 и 0.20 за двата типа фингърпринти. Най-малък процент повлияни битове се наблюдава за фингърпринта KlekotaRoth. Около 97% от битовете не са повлияни от тавтомери или показват вариация до 0.05. Едва 63 бита (1%) показват вариация между 0.10 и 0.20. По-голяма вариация в стойността на битовете се наблюдава при 14 бита (по-малко от 0.5%).

6.3 Изследване влиянието на тавтомерията върху алгоритмите базирани на структурно подобие

Във връзка с това изследване извършихме търсене по подобие в базата данни на Ambit, съдържаща 600 000 структури. Метриката за структурно подобие беше дефинирана чрез коефициента на Танимото. Търсенето беше осъществено за трите тавтомера на структурата на methimazole. Резултатите са представени в Таблица 7 Резултати от търсенето по подобие за трите тавтомерни форми на methimazole съдържат петте най-подобни структури за всяка от трите целеви структури и стойностите на коефициента на подобие подредени намалящ ред. За трите тавтомера ползвани като целеви структури, се получават доста различни резултати. Друго, което прави впечатление е, че при нито една от целевите структури в резултатите не се среща нито един от останалите два тавтомера.

Таблица 7 Резултати от търсенето по подобие за трите тавтомерни форми на methimazole

		Similarity		Similarity		Similarity
1.		0.62		0.71		0.47

2.		0.6		0.71		0.45
3.		0.59		0.64		0.44
4.		0.58		0.57		0.44
5.		0.54		0.57		0.43

Този резултат се дължи на влиянието на тавтомерната информация върху молекулните дескриптори и/или пръстови отпечатъци, върху които е изчислено подобието при извършване на търсенето в базата данни.

6.4 Тавтомерия при методите QSAR/QSPR

За да проверим как тавтомерната информация влияе върху резултата получен от прилагането на външни QSAR/QSPR модели генерирахме тавтомрните форми на violuric acid. За всяка от тях приложихме модел за предсказване на мутагенност и модел за изчисляване на коефициент на разпределение в системата октанол-вода XLogP. Резултатите от двата модела са представени в Таблица 8. За петнадесетте тавтомерни форми на violuric acid получихме стойности за XLogP вариращи в интервала [-1.26, 1.23], което означава, че използваният модел определя някои от тавтомерите като хидрофобни, а други като хидрофилни. По-голяма част от тавтомерните форми бяха класифицирани като мутагенни, а два от тавтомерите като немутагенни.

Таблица 8 Резултати за Ames-mutagenicity и XlogP за всички тавтомерни форми на violuric acid

Тавтомери на Violuric acid (SMILES)	Ames Mutagenicity	XlogP
O=C1NC(=O)C(=NO)C(=O)N1	1	0.135
O=C1N=C(O)N=C(O)C1(=NO)	0	-0.086
O=C1N=C(O)C(=NO)C(O)=N1	0	0.267
O=C1N=C(O)C(=NO)C(=O)N1	1	0.041
O=C1N=C(O)NC(=O)C1(=NO)	1	0.361
O=NC1=C(O)N=C(O)N=C1(O)	1	-0.102
O=NC=1C(=O)NC(O)=NC=1(O)	1	-0.084
O=NC=1C(=O)N=C(O)NC=1(O)	1	1.23
O=NC=1C(O)=NC(=O)NC=1(O)	1	0.689
O=NC=1C(=O)NC(=O)NC=1(O)	1	0.363
O=NC1C(O)=NC(=O)N=C1(O)	1	-0.277
O=NC1C(=O)N=C(O)N=C1(O)	1	-1.056
O=NC1C(=O)NC(=O)N=C1(O)	1	-0.932
O=NC1C(=O)N=C(O)NC1(=O)	1	-1.038

За тестовете подготвихме извадка от 8327 структури. За всяка структура от извадката генерирахме възможните тавтомерни форми чрез приложение на разработения Инкрементален подход. За тестовете използвахме реализираните в системата PaDEL – XlogP, Crippen LogP, AlogP и Mannhold LogP; както и реализираните в системата Dragon – AlogP и MlogP. За изчисляване на целевото свойство (logP) за структура отчитайки тавтомерната информация беше използвана формулата:

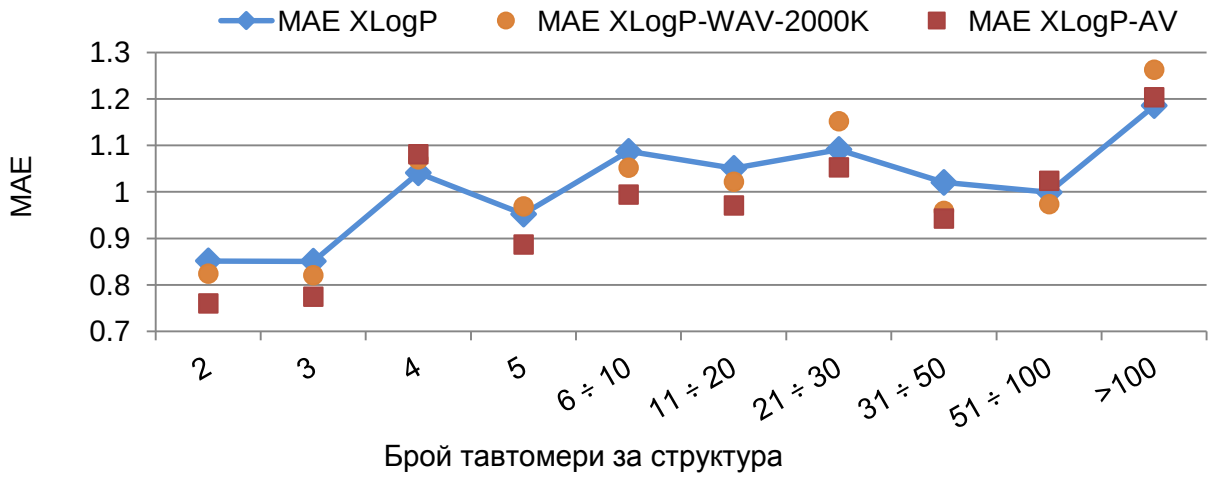
$$P_i^w = \sum_{k=1}^{n_i} w_i^k F\left(T_i^k\right),$$

където w_i^k е теглова схема, която отчита енергийното тегло на тавтомерите, T_i^k е k-тия тавтомер на структура i, F(S) е функцията на модела за LogP, реализирана в съответния софтуер (Dragon/PaDEL).

При изчисляването на модифицираното целево свойство P_i^w бяха използвани всички теглови схеми (означени с w_i^k във формулата) и резултатите бяха сравнени. Получените модифицирани резултати сравнихме с експерименталните стойности. Като параметър за оценяване на резултатите от тестващата извадка използвахме средната абсолютна грешка MAE, изчислена по формулата:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i^w - F(S_i)|$$

На Фигура 24 е даден резултатът за модела XLogP, където с плътна линия е представена абсолютната разлика между предсказаното свойство чрез Конвенционалния модел (модел, при който тавтомерната информация не е отчитана) и експерименталните стойности. С различни маркери са представени абсолютните разлики получени чрез приложение на две от тегловите схеми (WAV-2000K и AV) използвани за включване на тавтомерната информация в модела. При този метод беше наблюдавано най-голямо подобрение на предсказаните стойности след включване на тавтомерната информация.



Фигура 24 Разпределение на средната абсолютна грешка (MAE) за модел XlogP реализиран в PaDEL спрямо броя на генерираните тавтомери

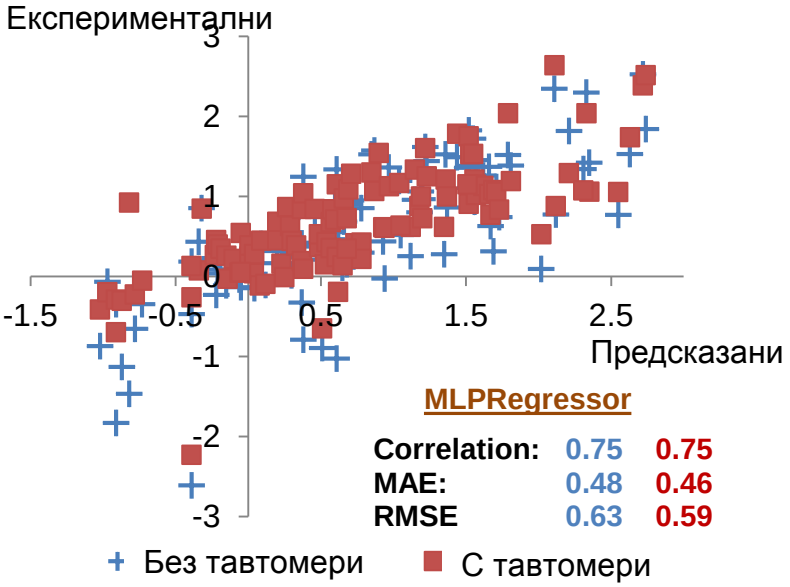
Не се наблюдава влияние на тавтомерната информация при изчисляване на LogP по метода на Mannhold, което се дължи на това, че при него се отчитат само химичните елементи в структурата. При този метод не е от значение тавтомерната форма.

Моделът AlogP реализиран в PaDEL е с по-лоша предсказваща способност (средната абсолютна грешка е над 1.4) от всички тествани методи. При използването на модифициращи схеми за включване на тавтомерите получихме подобряване на резултатите и намаляне на средната абсолютна грешка. Това се наблюдава при всички използвани теглови схеми, с малко изключение при структури с много тавтомерни форми, при които се наблюдава слабо влошаване на резултата за някои от тегловите схеми.

Проверихме как влияят тавтомерите върху построяването на модели *за предсказване на водна токсичност*. Използвахме три стратегии:

конвенционален, модифициран и смесен подход. При конвенционалния модел не се отчита тавтомерната информация нито в обучителната извадка, нито в извадката за външно валидиране. При модифицирания подход, тавтомерната информация се включва и в обучителната извадка и в извадката за външно валидиране. При смесения подход, моделът се построява върху конвенционална обучителна извадка и се валидира с модифицирана извадка или се построява върху модифицирана обучителна извадка и се валидира с конвенционална извадка.

За построяване на моделите бяха използвани едни и същи параметри. За избор на дескриптори използвахме методите CfsSubsetEval и BestFirst реализирани в Weka. При прилагане на метод за избор на дескриптори върху модифицираната обучителна извадка бяха избрани 24 дескриптора, докато при конвенциолания подход избраните дескриптори са 27. При сравнението между избраните дескриптори установихме 20 дескриптора еднакви при двете приложения на селектиращата процедура, и 7 различни дескриптора за конвенционалния метод и 4 при модифицираните дескриптори. На Фигура 25 са представени резултатите от приложение на моделите върху съответните извадки за външно моделиране, като със син цвят са отбелязани резултатите получени при конвенционалният модел, а с червен цвят са отбелязани резултатите получени чрез модифицирания подход.



Фигура 25 Сравнение на резултатите получени при Конвенционалния и Модифицирания подходи за предсказване на водна токсичност

На Фигура 25 са представени някои от статистическите характеристики получени при валидирането. За представения подход не се наблюдава подобрение в коефициента на корелация, но се наблюдава малко подобрение в стойностите за средната абсолютна грешка (MAE) – от 0.48 на 0.46 – и RMSE от 0.63 на 0.59.

7. Обобщение на получените резултати

Резултатите, описани в настоящата дисертация, могат да се обобщят както следва:

1. Създадени са три алгоритъма за генериране на тавтомерни форми на органични съединения:
    - Комбинаторен метод (CM)
    - Подобрен комбинаторен метод (CMI)
    - Инкрементен подход (IA-DFS)
2. Разработен е набор от филтри за премахване на дублиращи се тавтомери, базирани на:
    - вътрешна тавтомерна номерация;
    - изоморфизъм;
    - линейна нотация InChI.
3. Разработени са два филтъра за премахване на тавтомери, отговарящи на предварително за дадени критерии:
    - Предупредителен списък (Warning Filter) – премахващ структури съдържащи аленов атом в ацикличната част на структурата
    - Изключващ списък (Exclude Filter) – премахващ структури съдържащи аленов атом в цикличната част на структурата
4. Създадена е ранкираща система базирана на емпирично изведени енергийни тегла. Тя е използвана за извеждане на канонична тавтомерна форма.
5. Извършена е оптимизация на ранкиращата система, филтрите и бързодействието на алгоритмите
6. Извършени са вътрешни сравнителни тестове на разработените алгоритми:
    - Сравнени са резултатите, получени при използване само на 1-3 тавтомерните правила, при използване на 1-3 и 1-5 тавтомерни

правила, както и при използване на 1-3, 1-5 и 1-7 правилата. Резултатите са сравнени с теоретични резултати

- Проверена е еквивалентността на генерираните тавтомери, при използване на Инкрементния подход
- Сравнени са резултатите получени чрез Подобрения комбинаторен метод (CMI) и Инкрементния подход (IA-DFS)

7. Извършени са сравнителни тестове на резултатите получени чрез разработените алгоритми за генериране на тавтомери и тези получени чрез използване на външни софтуерни продукти и алгоритми за генериране.
    - IA-DFS и Cactvs,Marvin,ACDLabs,(InChI/CDK),Mn.Tautomer
    - CMI и Cactvs,Marvin
    - Тестове за бързодействие
    - Сравнение на каноничните тавтомери
8. Направен е предварителен анализ на влиянието на тавтомерната информация върху основни задачи използвани в химичната информатика:
    - Влияние върху изчисляване стойностите на молекулните дескриптори и фингърпринти.
    - Влияние на тавтомерната информация при прилагане на QSAR/QSPR модели.
    - Влияние на тавтомерната инфромация при построяване на модели.
9. Въведени са няколко теглови схеми за обхващане и оптимално включване на тавтомерната информация в изброените процеси.
10. Предложена е стратегия за генериране на тавтомерните форми чрез комбиниране на разработените подходи.

8. Научни приноси

Принос
Създадени са три алгоритъма за изчерпателно генериране на тавтомерни форми на органични съединения
Създадена е ранкираща система базирана на емпирично изведени тегла, използвана за избор на канонична тавтомерна форма
Предложени са теглови схеми за включване на тавтомерната информация при изчисляване на молеулни дескриптори.
Предложени са стратегии за включване на тавтомерната информация при QSAR/QSPR моделиране

9. Индустриални и приложни приноси

Тук представените алгоритми за генериране на тавтомерни форми намират следните приложения:

- Алгоритмите са имплементирани в системата за химична информатика Ambit, която се използва от академични организации, регулаторни агенции и компании от химичната и фармацевтичната индустрии
- Алгоритмите се използват от международна компания за производство на агрохимикали
- Включени са в софтуера за изчисляване на молекулни дескриптори и бит-вектори PaDEL
- В системата KNIME е включен елемент (node) позволяващ ползването на създадените алгоритми за генериране на тавтомери
- Международен проект ExCAPE – с участници AstraZeneca, Janssen, Intel, IdeaConsult и други организации. В проекта ExCAPE алгоритмите са използвани за стандартизация на над 89 милиона структури от базата данни на PubChem и няколко други големи бази данни.

10. Научни съобщения по темата на дисертацията

Публикации:

1. Kochev, N. T., **Paskaleva, V. H**. and Jeliazkova, N. (2013), Ambit-Tautomer: An Open Source Tool for Tautomer Generation. Mol. Inf., 32: 481–504. doi: 10.1002/minf.201200133
2. N. Kochev, **V. Paskaleva**, N. Jeliazkova, Chemical Information Processing for QSAR/ QSPR Modeling, Scientific Researches of the Union of Scientists in Bulgaria -Plovdiv, Series C. Natural Sciences and Humanities, vol. XVI, 2013, 223-226
3. N. Kochev, **V. Paskaleva**, N. Jeliazkova, Automatic Generation of Tautomers, University of Plovdiv - Scientific Papers, 38 (5), 2011, 65-70

Постери:

1. Nikolay T. Kochev, Ognyan Pukalov, **Vesselina H. Paskaleva**, Nina Jeliazkova, Ambit-Tautomer - a software tool for exhaustive tautomer generation. Applications in QSAR/ QSPR modeling, Quantitative Structure-Activity Relationships in Environmental and Health Sciences (QSAR2014) at the IRCCS -Istituto di Ricerche Farmacologiche ”Mario Negri”, Milan, Italy, June 16-20, 2014
2. **V.Paskaleva**, N.Kochev, O.Pukalov, N.Jeliazkova, Exploratory data analysis on QSAR/QSPR modeling and tautomerism, Union of Scientists in Bulgaria, 31 October -01 November 2014
3. Nikolay T. Kochev, **Vesselina H. Paskaleva**, Nina Jeliazkova, A set of predictive modles for Ames mutagenicit, Union of Scientists in Bulgaria, 31 October - 01 November
4. **V.Paskaleva**, N.Kochev, O.Pukalov, N.Jeliazkova, Exploratory data analysis on QSAR/QSPR modeling and tautomerism, 2nd Student Chemistry Conference, 21 – 22 November 2014
5. Nikolay T. Kochev, **Vesselina H. Paskaleva**, Nina Jeliazkova, Algorithms for automatic tautomer generation and their applications, OpenTox Euro 2013 at Johannes Gutenberg University of Mainz, Mainz, Germany, 30 September - 2 October 2013
6. Nikolay T. Kochev, **Vesselina H. Paskaleva**, Nina Jeliazkova, Algorithms for automatic tautomer generation and their applications, 6th Sheffield Conference on Chemoinformatics at Sheffield, England, 22-24 July 2013
7. N. Kochev, **V. Paskaleva**, N. Jeliazkova, Chemoinformatics Structural Data Processing for QSAR/QSPR modeling, INTERNATIONAL CONFERENCE OF YOUNG SCIENTISTS – PLOVDIV’2013, 13 June 2013
8. N. Kochev, **V. Paskaleva**, N. Jeliazkova, Ambit-Tautomer: An open source tool for tautomer generation, 19th EuroQSAR at Vienna, Austria, August 26-30, 2012
9. N. Kochev, **V. Paskaleva**, N. Jeliazkova, Automatic generation of tautomers, 9th Chemistry conference of University of Plovdiv, October 2011

11. Забелязани цитати

Публикация 1 е цитирана в:

1. Marta Marín-Luna, Ibon Alkorta, José Elguero, Interaction of beryllium derivatives with N-methylated DNA bases: 9-methylguanine and 1-methylcytosine, Tetrahedron, 2016, 72, 16, 1978
2. Albert J. Kooistra, Georgi K. Kanev, Oscar P.J. van Linden, Rob Leurs, Iwan J.P. de Esch, Chris de Graaf, KLIFS: a structural kinase-ligand interaction database, Nucleic Acids Research, 2016, 44, D1, D365
3. Marta Marín-Luna, Ibon Alkorta, José Elguero, The effect of cytosine methylation on its halogen-bonding properties, Computational and Theoretical Chemistry, 2016, 1076, 101
4. Torsten Thalheim, Barbara Wagner, Ralph Kühne, Martin Middendorf, Gerrit Schüürmann, A Branch-and-Bound Approach for Tautomer Enumeration, Molecular Informatics, 2015, 34, 5
5. Wendy A. Warr, Many InChIs and quite some feat, Journal of Computer-Aided Molecular Design, 2015, 29, 8, 681
6. Marta Marín-Luna, Ibon Alkorta, José Elguero, The influence of intermolecular halogen bonds on the tautomerism of nucleobases. I. Guanine, Tetrahedron, 2015, 71, 33, 5260
7. Sascha Urbaczek, Adrian Kolodzik, Matthias Rarey, The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States, Journal of Chemical Information and Modeling, 2014, 54, 3, 756
8. Elguero, J. (2016) Tautomerism: A Historical Perspective, in Tautomerism: Concepts and Applications in Science and Technology (ed L. Antonov), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany. doi: 10.1002/9783527695713.ch1

Както и в две защитени дисертации:

1. May, J. W. (2015). *Cheminformatics for genome-scale metabolic reconstructions* (Doctoral dissertation, University of Cambridge).
2. Kolodzik, A. (2014). In silico modeling of small molecules and design of eIF-5A activation inhibitors.