БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
Институт по биофизика и биомедицинско инженерство

"Акад. Георги Бончев" бл. 21, 1113 София, България
Телефон: +359-2-8700326, Факс: +359-2-8723787

РЕЦЕНЗИЯ
от доктор на науките **Илза Константинова Пъжева** -
професор в Института по биофизика и биомедицинско инженерство - БАН

на дисертационен труд за присъждане на образователната и научна степен **'доктор'**
в област на висше образование 4. *Природни науки, математика и информатика*
професионално направление 4.2. *Химически науки*
докторска програма *Теоретична химия*

автор: **Веселина Христева Паскалева**
тема: "**КОМПЮТЪРНО ГЕНЕРИРАНЕ НА ТАВТОМЕРИ НА ОРГАНИЧНИ СЪЕДИНЕНИЯ И ПРИЛОЖЕНИЯ В ХИМИЧНАТА ИНФОРМАТИКА**"

Научен ръководител: доц. д-р **Николай Тодоров Кочев** -
катедра "Аналитична химия и компютърна химия" - Химическия факултет,
Пловдивски университет "Паисий Хилендарски"

Със заповед № Р33-2017 от 16.05.2016. г. на Ректора на Пловдивския университет (ПУ) „Паисий Хилендарски" съм определена за член на научното жури за осигуряване на процедура за защита на гореописания дисертационен труд. Авторът на труда е докторант на самостоятелна подготовка към катедра "Аналитична химия и компютърна химия" na Химическия факултет към ПУ. Представеният ми комплект материали на хартиен носител е в съответствие с Чл.36 (1) от Правилника за развитие на академичния състав на ПУ „Паисий Хилендарски" и включва всички изискуеми документи.

Докторант Веселина Христова Паскалева е бакалавър по химия и притежава магистърска степен по "Спектрохимичен анализ" от ПУ „Паисий Хилендарски", с което отговаря на изискванията за подготовка, изпълнение и защита на дисертационен труд за присъждане на научната и образователна степен "доктор".

Тематиката на представения дисертационен труд е в **интердисциплинарното научно поле** на изчислителната химия и хемоинформатиката, като трудът интегрира знания и умения от теоретичната химия, математиката и информатиката. Изследванията се отнасят до генериране на тавтомери на малки органични молекули и оценката на ефекта им при решаване на задачи в химичната информатика с използване на компютърни и

информационни техники за обработка на данни за химически структури. Темата на дисертационния труд е **актуална**, тъй като отчитането на тавтомерията е от **съществено значение** за изследвания в различни научни области като органичната физикохимия, медицинската химия, фармакологията, токсикологията, химията на околната среда и др.

Дисертационният труд е написан на 177 страници, **структуриран е съгласно общоприетите правила за оформяне на дисертационни трудове** и съдържа основните изискуеми раздели: въведение, литературен преглед, цел и задачи на изследването, резултати и дискусия, заключения с изводи, приноси и трудове, свързани с дисертацията, и накрая списък с библиографията и използваните информационни източници. Бих искала да уточня, че към раздела с резултати и дискусия отнасям глави 3, 4 и 5, тъй като те описват основните резултати по разработваните алгоритми за генериране на тавтомери, тяхното тестване и използването им за оценка на влиянието на тавтомерите при решаване на някои задачи в хемоинформатиката, независимо че докторантът е кръстил глава 4 „Тестове. Резултати и дискусия.“ На практика, обсъждане на резултатите има във всяка една от тези глави, което е продиктувано от спецификата на изследванията, без да има нужда от самостоятелно оформен дискусионен раздел. Трудът е онагледен с 27 таблици и 76 фигури, големият им брой е наложил оформянето на самостоятелни списъци в края, което подпомага прочита му. Към дисертацията са приложени също и материали със 7 фигури за допълнително онагледяване на представените резултати.

Литературният преглед е фокусиран върху представянето на основни обекти в химичната информатика, тавтомерията като химичен обект и свързаните с нея проблеми при решаване на задачи в областта на *in silico* изследването на зависимости структура-активност с лиганд- и структура-базирани методи. Последната точка на литературния преглед представя основни софтуерни системи, използвани понастоящем за генериране на тавтомери, придружени с анализ на реализираните алгоритми в случаите, в които има достъпна за тях информация. Цитирани са 157 библиографски и информационни източници, от които над 40 са от последните 5 години. Докторантът успешно се е справил със събирането, анализа, структурирането и представянето в обзора на достъпната в литературата информация по темата на дисертацията.

По литературния преглед имам следните забележки:

- Литературният обзор би спечелил, ако някои от използваните в него термини бяха предварително дефинирани, като „*стандартен тавтомер*“, „*основен тавтомер*“, „*уникална тавтомерна форма*“, „*стабилна тавтомерна форма*“, „*канонична структура*“, „*каноничен тавтомер*“. Още във въведението се споменава, че нормализацията и канонизацията са основни процедури, използвани при обработка на тавтомерни съединения (стр. vii), което предполага тяхното ясно дефиниране още в началото на обзора. Такова определение се дава едва на стр. 25, а термините „*канонична структура*“ и „*каноничен тавтомер*“ се използват в текста и преди това в различни интерпретации (напр., на стр. 20).
- Използвани са редица съкращения, но само част от тях са обяснени в списъка със съкращенията в началото на труда, който е непълен и неподреден. Например, на стр. 4 се спо-

менава за IUPAC и NIST, първата абревиатура е разписана, а втората не е. Същото се отнася и за WHIM описатели, CIP правила и др.

- Литературен източник No. 68 е непълен, а на No. 142 липсва годината.

От литературния преглед следва изводът за „*липсата на надежден, бърз, изчерпателен и разработен под свободен лиценз генератор на тавтомери*“. Този извод обосновава и **целта на дисертацията**, която се състои в разработването на глобален подход за генериране на тавтомери, които да е в състояние да възпроизвежда всички теоретично възможни тавтомерни форми на органични съединения. На практика, обаче, както става ясно от последващото изложение, представените изследвания **надхвърлят тази цел** и това се вижда още от поставените задачи. Те са 12 и, най-общо, могат да бъдат обединени в четири основни групи: (i) разработване на алгоритми за реализация на подход, позволяващ генериране и оценяване (претегляне, ранкиране) на всички възможни тавтомери; (ii) програмиране и имплементиране на алгоритмите в специализиран софтуер; (iii) тестване и оптимизация; (iv) приложение. По такъв начин формулираните задачи следват естествената последователност на етапите в работата по разработване и валидиране на подход за генериране и оценка на тавтомери.

Резултатите са отразени в три глави като първите две описват най-съществената част на дисертационния труд – разработването на алгоритмите и тяхното тестиране; третата глава представя приложения на тези алгоритми. Използваните методи са съобразени с целта и са адекватни на задачите, решени в дисертационния труд Извършена е огромна и нелека работа, която започва с разработване на база днни с тавтомерни правила, преминава през разработване на алгоритмите за генериране на тавтомери, за филтриране и ранкиране на тавтомери, програмна реализация на алгоритмите, тестването им за изчерпателно генериране и еквивалентност, извеждане на канонични тавтомери, оценка на бързодействие и оптимизация на алгоритмите. Тези резултати докторантът обобщава в 10 основни извода, въз основа на които формулира и основните си приноси. Последните са разпределени в две основни групи: (i) научни и (ii) индустриални и приложни приноси. Признавам, че за първи път срещам формулировката „индустриални приноси“. По същество това са приложения на разработените инструменти, които се използват от фирмени организации и проекти, свързани с работа с големи бази данни от химични структури.

Приносите на дисертацията бих определила по следния начин:

Научно-теоретични приноси:

- Предложени са нови алгоритми за генериране на тавтомерни форми на органични съединения: (i) класически комбинаторен, (ii) подобрен комбинаторен, (iii) инкрементен; алгоритмите се базират на оригинални тавтомерни правила за описание на всички видове протропна тавтомерия с отчитане на три нива на миграции на протон;
- Предложени са нов алгоритъм и база данни за ранкиране наа тавтомери;

- Дефинирани са схеми за отчитане на тегла на тавтомери в хемоинформатични приложения.

Научно-приложни приноси:

- Разработен е софтуерен инструмент с отворен код, базиран на предложените алгоритми, който автоматично генерира всички възможни тавтомери на дадено органично съединение;
- Определени са ефектите от приложение на софтуерния инструмент и предложените теглови схеми за решаване на конкретни хемоинформатични задачи.

Бих искала **особено да подчертая** два факта, свързани с горните приноси:

- първо, че разработените алгоритми са сравнени с други подобни алгоритми с широко приложение (*ChemAxon Marvin, Molecular Networks MN.TAUTOMER, ACDLabs, CACTVS, CDK имплементация на алгоритъм, базиран на мобилни H-атоми в InChI*) и сравнението показва, че инкрементният алгоритъм самостоятелно и в комбинация с подобрения комбинаторен алгоритъм **дава сравними и дори по-добри резултати** от тях;

- второ, че разработените алгоритми са **със свободен достъп**, което е от съществено значение за научната и особено за академичната научна общност.

Трудовете, свързани с дисертационния труд, включват три публикации, от които едната в списание с импакт фактор (*Molecular Informatics,* импакт фактор=2.013 за 2013). По тази публикация има забелязани 10 цитирания (две в защитени дисертации, 7 в авторитетни специализирани списания и едно в глава от книга), което показва значимостта на публикуваните в нея резултати. Във всички публикации докторантът е втори автор, като първи автор е научният ѝ ръководител. Докторант Паскалева има и 9 участия в международни и национални научни форуми, от които на 5 форума е била представящ автор. С пълното разбиране, че съвременните научни изследвания са резултат от екипна работа и във връзка с оценката на трудовете и бих искала да попитам докторанта в каква степен формулираните приноси и получени резултати, са негова лична заслуга.

По резултатите от дисертационния труд имам **следните въпроси, коментари и препоръки**:

1. По отношение на предложените алгоритми за генериране на всички възможни тавтомери, ранкиране на тавтомерите и тестване на алгоритмите.

- Кой от полуемпиричните методи в пакета MOPAC е използван за изчисляване на енерийните тегла за ранкиране на тавтомерите? В бъдещи изследвания биха могли да се изчислят и сравнят резултатите от използването на различни полуемпирични методи за определяне на енергийните тела, в т.ч. и такива, които отчитат ефекта на разтворителя.
- На какъв принцип/критерий са избрани и формирани извадките, използвани за тестване и валидиране на алгоритмите (Таблица 4.1, стр. 65 и Таблица 5.2, стр. 113)?

Има ли данни колко от структурите в извадките са уникални за конкретната извадка и колко се припокриват със структури от другите извадки? На какъв принцип се определя за кой тест каква извадка да се използва? Препоръчително е тази информация да бъде представена и да се обоснове изборът на извадка за съответния тест.

- Генерираните тавтомери са сравнени с теоретично получените (т. 4.2.2.1, стр. 69), което е правилен подход, имайки предвид, че целта е генерирането на максимално възможния брой тавтомери. Как е направено това - какво се има предвид под „ръчно определяне на тавтомерите“? Защо именно *Извадка 3* е използвана в този тест?
- За валидирането на генерираните тавтомери би могло също така да се проследи как генерираните тавтомери се съотнасят с тези, за които има налични експериментални данни. Например, в труда се споменава за 64 тавтомерни форми на тетрациклина, докладвани в литература (стр. 11, няма посочен източник), а Ambit-Tautomer дава 99 тавтомера след генериране и филтриране (собствено изследване). Би било интересно да се сравнят генерираните тавтомери с докладваните в литературата. Нещо повече, за същото съединение се споменава, че има докладвани и експериментални спектроскопски данни, които биха могли да се използват, за да се оцени доколко адекватно на експеримента са определени теглата/ ранговете на тавтомерите.

2. По отношение на приложението на разработените алгоритми в тестовете за отчитане на тавтомерията в QSPR/QSAR анализи (т. 5.4).

- В класическата QSPR/QSAR концепция молекулните дескриптори са числа, които характеризират интегрални или локални физикохимични свойства на структурите (хидрофобни, стерични, електронни, вкл. квантовохимични, индикаторни за присъствие/отсъствие на подструктури и др.), а целевото свойство (физикохимичен параметър, активност, токсичност) се изчислява като функция от тях и точно така концепцията е обяснена в началото на раздел 5.4 (стр. 126). В този контекст в дисертационния труд тегловите схеми се прилагат както за модифициране (претегляне) на целево свойство, напр. в тестовете за изчисляване на logP (QSPR), така и на молекулни дескриптори, напр. в тестовете за изчисляване на водна токсичност (QSAR). В тази връзка бих предложила уточняване/специфициране на дефиницията „модифицирани дескриптори“ при бъдещи приложения и публикации с тяхното използване в зависимост от това дали те се използват за изчисляване на независими променливи (молекулни дескрптори в QSAR) или зависима променлива (целево свойство в QSPR).
- Би било полезно да се обосноват избраните температурни стойности в тегловите схеми за модифициране на дескриптори (Табл. 5.9, стр. 130).
- В примера с виолуровата киселина (табл. 5.10, стр. 131) би било полезно да се докладват теглата на генерираните тавтомери и да се проследи съответствието между тегло/ранг на тавтомер и предсказано свойство (мутагенност). Например, какви са теглата на тавтомерите, предсказани като немутагенни и как се съпоставят те с останалите? Тази информация би била ценна по отношение на оценка на риска на химикали.

- При описанието на резултатите от QSAR моделирането на водна токсичност (стр. 133) би било полезно да се опишат използваните 2D дескриптори и доколко те могат да бъдат механистично интерпретирани от гледна точка на моделираната активност.
- В допълнение към насоките за бъдещо развитие на генератора за тавтомерни форми (стр. 142), представен в настоящия труд, бих предложила насочване на усилията към определянето на тавтомера или тавтомерите, които са най-вероятните за условията, при които е изследвано моделираното таргетното свойство в QSPR/QSAR приложения на тавтомерния генератор. Включването на по-малко и най-вероятните тавтомери на органичните съединения би осигурило получаването на по-реалистични QSPR/QSAR модели с тяхно участие.

Заключение

Дисертационният труд на Веселина Христева Паскалева съдържа приноси с научно-теоретичен и научно-приложен характер. Получените резултати представляват оригинален принос в науката и отговарят на изискванията на Закона за развитие на академичния състав в Република България, Правилника за неговото прилагане и съответния Правилник на ПУ „Паисий Хилендарски“. Представените материали и дисертационни резултати напълно съответстват на специфичните изисквания на Химическия факултет, приети във връзка с Правилника на ПУ за приложение на ЗРАСРБ, а публикационната активност на докторанта надхвърля специфичните изисквания за придобиване на образователната и научна степен „доктор“. Дисертационният труд е с актуална тематика, разработен е аналитично и показва добра теоретична и практическа подготовка на докторанта. Извършена е огромна по обем работа, усвоени са редица методи, комбинирани са успешно теоретични знания и практически умения. Въз основа на това давам **положителна оценка** на проведеното изследване и препоръчвам убедено на уважаемите членове на научното жури да присъдят на Веселина Христева Паскалева образователната и научна степен „Доктор“ в област на висше образование 4. *Природни науки, математика и информатика,* професионално направление 4.2. *Химически науки,* докторска програма *Теоретична химия*.

16.08.2016 Рецензент: .....

София (проф. Илза Пъжева)