

**КОМПЮТЪРНО МОДЕЛИРАНЕ
НА ХИМИЧНИ РЕАКЦИИ**

Светлана Любенова Аврамова

Автореферат

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: Органична химия

Научни ръководители:

доц. д-р Николай Кочев

доц. д-р Пламен Ангелов

Дисертационният труд е обсъден и насрочен за защита от катедрения съвет на катедра „Органична химия“ на Химическия факултет при Пловдивския университет „Паисий Хилендарски“ на 20.04.2018 г.

Дисертационният труд съдържа 287 страници (формат А4), в които са включени 129 фигури и 14 таблици. Цитирани са 310 литературни и информационни 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:

проф. дхн Иван Бангов

проф. дхн Людмил Антонов

проф. д-р Илиян Иванов

доц. д-р Иванка Цаковска

доц. д-р Пламен Ангелов

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“.

Съдържание

I. Увод	1
II. Цел и задачи	1
III. Собствени изследвания, резултати и дискусия	2
1. Метод за представяне и алгоритъм за прилагане на реакции (Ambit-SMIRKS).....	2
1.1. Софтуерна архитектура	2
1.2. Представяне на химични реакции	3
1.3. Прилагане на реакции (фактическо трансформиране на молекули).....	4
1.4. Софтуерно приложение Ambit-SMIRKS – GUI	9
1.5. Тестване и приложение на Ambit-SMIRKS	10
2. Виртуален химичен реактор (Ambit-Reactor)	11
2.1. Обща характеристика	11
2.2. База данни с реакции	11
2.3. Конфигуриране на виртуалния реактор Ambit-Reactor.....	12
2.4. Софтуерна архитектура и алгоритъм на Ambit-Reactor	12
3. Ретросинтезен реактор.....	14
3.1. Обща характеристика	14
3.2. Алгоритъм на Ретросинтезния реактор	21
3.3. Ретросинтезен реактор – графичен потребителски интерфейс (GUI)	21
3.4. Ретросинтезен реактор – дискусия.....	24
4. Оценка на синтетичната достъпност	26
4.1. Комплексност на циклите	26
4.2. Цикломатично число	27
4.3. Стереохимична комплексност	27
4.4. Молекулна комплексност.....	28
4.5. Софтуерна реализация.....	28
4.6. Оценка на синтетичната достъпност – резултати	29
IV. Заключение	31
1. Обобщение на получените резултати.....	31
2. Научни приноси.....	32
3. Научно-приложни приноси	32
4. Научни съобщения по дисертацията	32
4.1. Публикации	32
4.2. Постери	33
Библиография и информационни източници.....	34

I. Увод

Развитието на компютърните технологии и новите концептуални подходи за прилагането на принципите на теоретичната химия са мощен инструмент за рационализиране и прогнозиране на химичните процеси. „Днес компютърът е също толкова важен инструмент за химиците, колкото и епруветката” – се казва в съобщението на Шведската кралска академия при обявяването на Нобеловата награда за химия за 2013г., връчена на трима американски учени „за развитието на теоретични модели на сложни химични системи“ [1].

В частност, компютърното моделиране на химични реакции намира все по-широко приложение при решаването на редица задачи в органичната химия, като дава възможност за [2]: съхраняване на информация за химични реакции, извличане на информация и знания за химични реакции, сравняване и анализ на набор от реакции, дефиниране на обхват и ограничения на реакциите, разработване на модели за химическа реактивоспособност, предсказване на хода на химичните реакции, анализ на реакционни мрежи, разработване на модели за синтезен дизайн и пр.

Горните факти мотивираха целта на настоящата дисертация, а именно да се разработи подход за компютърно моделиране на химични реакции, който да бъде приложен при решаване на актуални в химичната информатика и органичната химия, като планиране на органичен синтез и ретросинтезен анализ, оценка на синтетичната достъпност на съединенията, предсказване на метаболити на органични съединения и пр.

II. Цел и задачи

1. Цел:

Целта на настоящия дисертационен труд е да се разработи подход за компютърно моделиране на химични реакции, който да се прилага при решаване на различни задачи в химичната информатика и органичната химия: планиране на органичен синтез и ретросинтезен анализ, предсказване на метаболити на органични съединения, оценка на синтетичната достъпност и пр.

2. Задачи:

2.1. Разучаване на съществуващи софтуерни системи и публикувани алгоритми за моделиране на химични реакции при планиране на органичен синтез: ретросинтезен анализ, предсказване на реакции (синтезно направление) в т.ч. биотрансформации (предсказване на метаболизъм), оценка на синтетичната достъпност;

2.2. Оптимизиране на методите за прилагане на химични реакции посредством линейната нотация SMIRKS в софтуерния модул Ambit-SMIRKS;

2.3. Дизайн и създаване на база данни с химични реакции (синтезно направление), представени в компютърно-четим вид чрез нотацията SMIRKS;

2.4. Създаване на база данни с биотрансформации (метаболизъм), представени в компютърно-четим вид чрез нотацията SMIRKS;

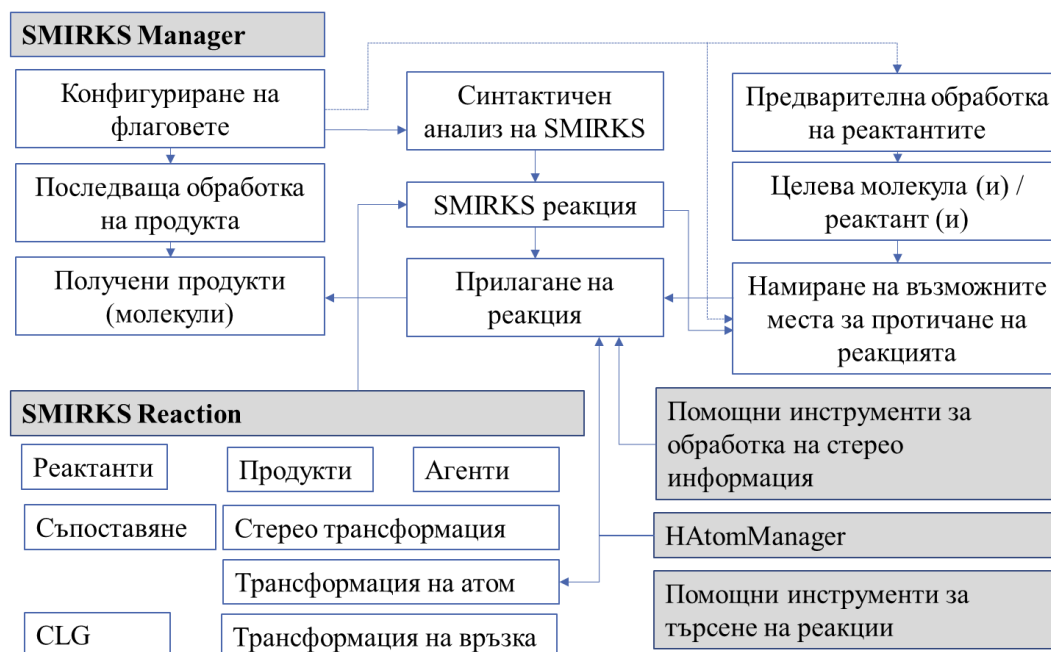
- 2.5. Дизайн и програмна реализация на конфигурируем виртуален химичен реактор;
- 2.6. Създаване на база данни с трансформи (ретросинтезно направление), представени в компютърно-четим вид чрез нотацията SMIRKS;
- 2.7. Създаване на стратегии и правила за компютърно подпомогнат ретросинтез;
- 2.8. Програмна реализация на специализиран софтуерен модул за ретросинтезен анализ (JBSMM-Reactor);
- 2.9. Създаване на модел за оценка на синтетична достъпност;
- 2.10. Тестване на софтуерните модули, създадените стратегии и правила по отношение на ефективност и коректност на генерираната химична информация и създадените модели;
- 2.11. Сравняване на разработените от нас подходи и алгоритми с други софтуерни системи и модели.

III. Собствени изследвания, резултати и дискусия

1. Метод за представяне и алгоритъм за прилагане на реакции (Ambit-SMIRKS)

1.1. Софтуерна архитектура

Ambit-SMIRKS е разширение на библиотеката Ambit-SMARTS [3] и е част от софтуера с отворен код AMBIT [4][5]. Софтуерната архитектура на модула Ambit-SMIRKS е представена на Фигура 1:



Фигура 1: Софтуерна архитектура на Ambit-SMIRKS

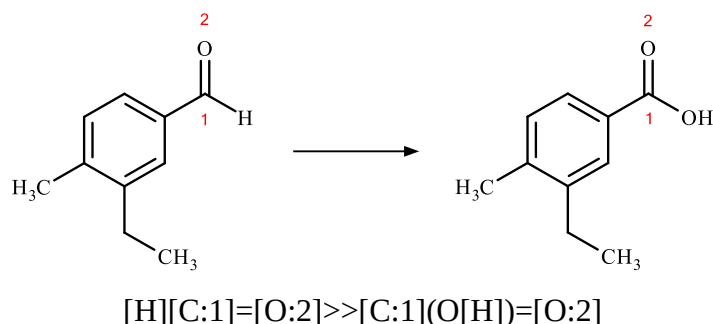
Основната функционалност на Ambit-SMIRKS се изпълнява от 3 основни Java класа: SMIRKSReaction, SMIRKSManager, StereoChemUtils, и от допълнителни помощни средства за обработка на H-атоми и търсене на реакции (Фигура 1).

Класът SMIRKSReaction включва цялата информация, необходима за описание на химическа реакция или молекулна трансформация: реактанти, продукти, съпоставяне на атоми, групиране на компоненти (CLG) и стерео трансформации. С помощта на класа SMIRKSReaction могат да се кодират както обикновени реакции, така и генерични реакции.

Класът SMIRKSManager включва основен програмно-приложен интерфейс (API) за всички функционалности, които обработват информацията в работния поток за прилагането на дадена реакция. Работният поток за прилагането на трансформация се конфигурира от набор флагове (flags), които определят предварителната обработка на целевата молекула, последващата обработка на продукта, режима на търсене, стерео трансформацията, обработката на H-атомите и т.н.

1.2. Представяне на химични реакции

Всички химични реакции в настоящата дисертация са представени чрез линейната нотация SMIRKS, представляваща ефективно кодиране на молекулните графи под формата на символен низ (Фигура 2).



Фигура 2: Окисление на алдехиди, представено чрез линейна нотация SMIRKS

Топологичното представяне на химично съединение (т.е. химичен граф), се имплементира чрез CDK класа AtomContainer, който съдържа списък на атомите, списък на връзките, стандартни операции по тези списъци, свойства на атомите, връзките и молекулата, както и списък със стереоелементите. CDK класът а QueryAtomContainer кодира разширен граф, чрез който се представят заявкита за субструктурно търсене.

За компютърно представяне на информацията за дадена реакция се използва обект от класа SMIRKSReaction, който се получава чрез синтактичен анализ (parsing) на съответната линейна нотация SMIRKS. Класът SMIRKSReaction (Фигура 1) включва два специализирани химични графа – за реактантите и за продуктите. Ако от страната на реактанта или на продукта се съдържа повече от една молекула или фрагмент, съответният граф се разчленява, като се поддържа допълнително номериране на атомите, за да се посочи към кой фрагмент принадлежат атомите.

Много важна част от представянето на една реакция е информацията за съпоставянето на атомите (atom mapping), която свързва атомите на реактанта с атомите на продукта. Въз основа на информацията за съпоставянето на атомите, трансформацията се представя като специализирана структура от данни, описващи промените (обновяване, изтриване или създаване) на молекулни елементи: атоми, атомни свойства, връзки, свойства на връзките и стереоеlementи.

1.3. Прилагане на реакции (фактическо трансформиране на молекули)

Прилагането на дадена реакция се извършва в три основни стъпки:

- (1) идентифициране на местата за протичане на реакцията в целевата молекула;
- (2) действителната трансформация на всички намерени места (или част от тях);
- (3) комбиниране на получените на етап (2) структури в окончателен набор от продукти.

1.3.1. Намиране на местата за протичане на реакции чрез субструктурно търсене

Местата, където протича реакцията, се откриват чрез субструктурно търсене. За тази цел частта от линейната нотация SMIRKS от страната на реактанта се използва като SMARTS дефиниция на заявка за субструктурно търсене.

Субструктурното търсене може да се извърши в няколко режима: единичен (single), неприпокриващ се (non-overlapping), неидентичен (non-identical), нехомоморфен (non-homomorphic) или дефиниран от потребителя. Използването на различни режими при субструктурното търсене позволява ефективен и гъвкав алгоритъм за прилагане на реакции.

1.3.2. Поддръжка на рекурсивни SMARTS-изрази

Една от най-важните характеристики на Ambit-SMIRKS е поддържането на рекурсивните атомни изрази. Рекурсивният SMARTS израз включва в логическите си атомни примитиви друга линейна нотация SMARTS, където рекурсивните изрази се дефинират чрез синтаксис от вида:

[...\$(smarts)...]

Например, изразът [CH3;!\$(C*=O);!\$(C*N)] дефинира метилова група, въглеродният атом на която не е част от карбонилна или аминна група. Поддържането на рекурсивни изрази дава голяма гъвкавост за дефиниране на химическата логика на различни молекулни модели. По този начин Ambit-SMIRKS може точно да дефинира реакционните центрове чрез детайлни субструктури, описващи информацията за атомната околност.

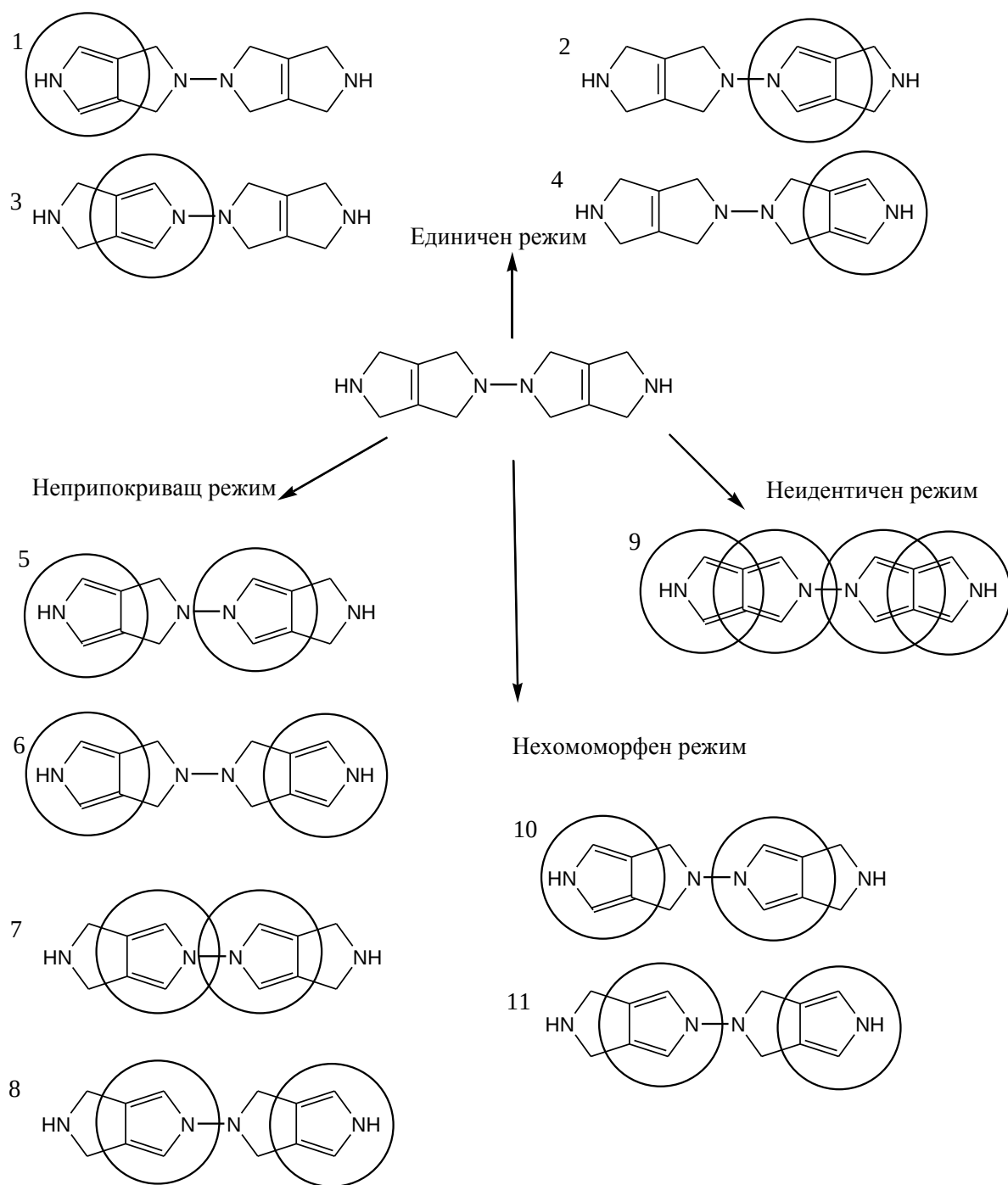
1.3.3. Структурни трансформации

SMIRKS трансформацията на дадена целева молекула може да бъде приложена директно върху целевата молекула чрез трансформиране на нейния обект от клас AtomContainer, като е възможно да се получи повече от един продукт, т.е. обектът AtomContainer ще бъде фрагментиран.

Друга характеристика, поддържана в Ambit-SMIRKS, е генерирането на копия на молекули, съответстващи на конкретни продукти, получени чрез прилагане на трансформацията на определени места. И в двата случая структурните трансформации се основават на режимите за субструктурно търсене. Когато трансформацията се извършва без копиране на молекули, входната целева молекула се модифицира, тъй като трансформацията се прилага на всички намерени места, комбинирани в съответствие с използвания режим.

Например в режим SINGLE ще бъде получен един от продуктите **1, 2, 3** или **4** (Фигура 3). В режим NON_HOMOMORPHIC ще бъде получен един от продуктите **10** или **11**. Прилагането на трансформация директно върху входната молекула в един от режимите ALL или NON_IDENTICAL ще доведе до химически неправилната структура **9**, тъй като всички четири възможни пръстена се трансформират, като по този начин се получават 5 валентни въглероди. Очаква се потребителят да провери химическата коректност на получените реакционни продукти. Неправилната структура, получена в NON_IDENTICAL режим, може да бъде избегната, ако реакционните трансформации се прилагат в каскаден стил, т.е. трансформациите се прилагат на няколко отделни стъпки (една единична трансформация за всеки реакционен обект), докато се проверява дали на останалите места все още е възможно да се приложи реакцията. Директната трансформация на целевата молекула, без копиране, не генерира всички възможни комбинации от продукти, но реакцията се прилага едновременно на всички места, включващи даден режим (ALL, NON_IDENTICAL, NON_OVERLAPPING, NON_HOMOMORPHIC или SINGLE). Например, в режим NON_OVERLAPPING ще бъде получена само една от структурите **5, 6, 7** или **8**.

За да се получат всички възможни продукти в единичен (single) режим (например структури **1, 2, 3** и **4**, Фигура 3), трябва да се приложи трансформация с едно копие за всеки продукт. Така за всяка копирана молекула реакцията ще се извърши на различно място и ще бъдат получени всички възможни продукти. Ambit-SMIRKS поддържа отделна опция за генериране на всички възможни неприпокриващи се комбинации, напр. структури **5, 6, 7** и **8**.



Фигура 3: Примери за трансформации в различните режими

Ако е необходимо, вместо местата за реакция, получени от стандартните режими за търсене, потребителят може да дефинира места за прилагане на реакция чрез IAcceptable интерфейс. Тази характеристика е особено полезна, когато е налице допълнителна информация за местата на протичане на реакция от външни източници (например изчисления от квантова механика, друг метод на молекулно моделиране, избор на експерт / потребител и т.н.).

1.3.4. Атоми, които са съпоставени и обозначени като съответстващи един на друг

Линейната нотация на SMIRKS поддържа дефиниране (обозначаване) на съответстващите атоми със следния синтаксис:

[<atom expression>:<n>]

Съпоставянето на атомите и обозначаването им като съответстващи един на друг се използва, за да се покаже кой от атомите на реактанта съответства на даден атом от продукта. В случаите, когато няколко атома на даден елемент присъстват и от двете страни на реакцията, е необходимо да се прави разграничение между тях по ефективен начин (атомните дефиниции в SMIRKS не са уникални).

Синтаксисът на SMIRKS също така стандартно поддържа необозначени с индекси атоми, напр. трансформацията може да бъде представена по следния начин: C=C>>CC. Софтуерът Ambit-SMIRKS поддържа и двете дефиниции (атоми с и без индекси).

Логиката зад необозначените (без индекси) като съответстващи един на друг атоми е следната:

(1) Атоми от реактанта (намиращи се от лявата страна на SMIRKS линейната нотация), които не са индексирани като съответстващи, се отстраняват (изтриват) в структурите на получените продукти, както и всички връзки, свързани с неиндексирани атоми, се отстраняват от атомните контейнери;

(2) Атоми от продукта (намиращи се от дясната страна на SMIRKS линейната нотация), които не са индексирани като съответстващи, се “създават” и добавят към структурите на получените продукти, където се създават и съответните нови връзки (от неиндексиран атом към другите атоми индексирани или неиндексирани).

Горните точки описват действителните случаи, в които неотбелязаните като съответстващи един на друг атоми трябва да се използват в SMIRKS: изтриване на атоми или добавяне на атоми. Във всички други случаи, когато атомите се „пренареждат“ чрез промяна, добавяне или премахване на връзки, задължителното използване на индекси за обозначаване на съответстващите един на друг атоми се счита за добра практика. Неправилното използване на необозначени атоми води до странични ефекти и „странно“ или неправилно прилагане на реакциите SMIRKS. Дори те да са с правилен синтаксис, ако дефинираната химическа логика е различна, Ambit-SMIRKS ще изпълни точно това, което е определено с тази логиката.

1.3.5. Третиране на водородни атоми

Ambit-SMIRKS, както и по-голямата част от софтуерните системи за химична информатика, обработва водородните атоми по два основни начина: (i) като имплицитни H-атоми (разглеждани по подразбиране) и (ii) експлицитни H-атоми (разглеждани изрично). От гледна точка на прилагането на трансформация с използване на SMIRKS линейна нотация, подходът с използване на експлицитни H-атоми е по-надежден и има

потенциала да опише по-точно химическата реакционна логика. Ambit-SMIRKS ползва три принципни сценария за третиране на Н-атомите:

(1) Експлицитно – счита се за най-добрата практика за описване на химическите реакции, тъй като той определя промените на всички химични връзки, включително тези, свързани с Н-атоми;

(2) Имплицитно или експлицитно с опция за трансформация на Н-атоми – по подразбиране атомните изрази в SMIRKS, които съдържат специфицирани Н-атоми (фактически имплицитни), ще бъдат използвани само за субструктурно търсене на атоми, за да се идентифицират потенциалните места за протичане на SMIRKS реакцията;

(3) Автоматично – основава се на премахване на некоректни Н-атоми (ако са получени такива) след прилагане на трансформацията и автоматично настройване на имплицитни атоми отново (например чрез помощните средства на CDK за добавяне на Н-атоми).

1.3.6. Третиране на ароматност

Информацията за ароматността в рамките на логическите изрази на линейната нотация на SMIRKS се използва главно за дефиниране на заявките за субструктурно търсене, за да се идентифицират местата за протичане на реакцията. Детайлизирани SMIRKS нотации от типа:

[C:1]1[C:2]=[C:3][C:4][C:5][C:6]1>>[c:1]1:[c:2]:[c:3]:[c:4]:[c:5]:[c:6]:1

са коректни от химическа гледна точка, тъй като от страната на продукта целият ароматен цикъл е дефиниран. При получаване на ароматна система от страна на продукта обаче има възможност и за потенциално неправилни резултати, защото например може да не е известно дали има кондензирани цикли и т.н. В рамките на Ambit-SMIRKS считаме за добра практика да се използват Kekule структури при третирането на ароматност, тъй като по този начин всички порядъци на връзките се дефинират експлицитно и SMIRKS трансформацията на връзките също е ясно определена.

След прилагането на реакцията Ambit-SMIRKS изпълнява алгоритъм за откриване на ароматност, за да се установи образуване на ароматна система вследствие на промени във връзките. Използването на този подход може да се счита за недостатъчно надежден, имайки в предвид факта, че е необходимо да се разчита на определен алгоритъм за откриване на ароматност, особено когато в софтуера по химична информатика липсва добър алгоритъм за разпознаване на ароматност. Ambit-SMIRKS разчита на CDK детектор за ароматност, който е значително подобрен в последните версии на CDK [28].

1.3.7. Третиране на стереохимична информация

Стереохимията в системите за химична информатика [6] се представя по два основни начина: абсолютен и относителен. Абсолютният подход описва елементите на стерео групата, като дава приоритет (подредждане) на стерео елементите на базата на

абсолютна химическа логика, която не зависи от номерацията на атомите (обикновено номерацията зависи от алгоритъма за обхождане на графа). В Ambit-SMIRKS стерео обработката се основава на другия подход за представяне на стерео информация – относителния, тъй като и SMIRKS линейните нотации, и вътрешните обекти на CDK са базирани на него.

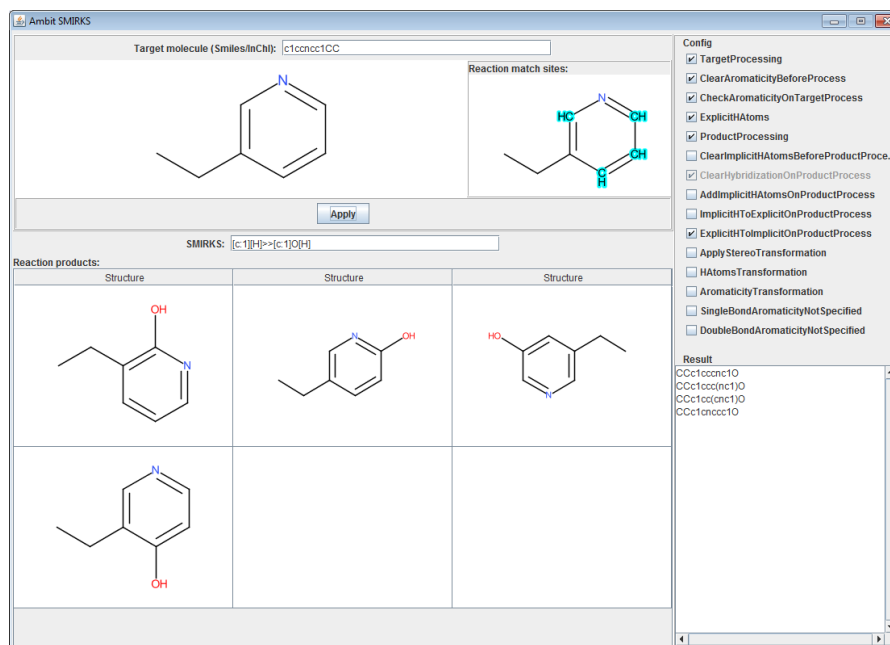
Ambit-SMIRKS поддържа случаи на стерео трансформации, които могат да бъдат обобщени в две основни групи:

- (1) стереотрансформация, която не е пряко посочена от SMIRKS
- (2) стереотрансформация, определена от SMIRKS

1.4. Софтуерно приложение Ambit-SMIRKS – GUI

Разработеното GUI приложение към модула Ambit-SMIRKS е потребителски ориентирано, интуитивно и лесно за употреба. Входната структура, т.е. тази, върху която искаме да приложим дадена реакция, се задава в полето „Target molecule“ чрез линейна нотация SMILES [7], [8] или InChI [9]. В полето „SMIRKS“ се въвежда линейната нотация за реакцията, която искаме да приложим върху въведената входяща структура.

Фигура 4 илюстрира приложението на реакция на ароматно хидроксилиране за молекулата 3-етилпиридин, където се генерират и са показани четири възможни продукта.



Фигура 4: Ароматно хидроксилиране на 3-етилпиридин

Реакцията се прилага с настройките по подразбиране на Ambit-SMIRKS, които в графичния потребителски интерфейс са показани като места за поставяне на отметка в менюто за конфигурация на флагите.

В полето „Reaction match sites“ могат да се видят маркирани със син цвят всички места от входящата структура, на които е възможно да протече зададената реакция. Получените при прилагане на реакцията продукти се представят като SMILES линейни нотации в полето „Results“ и се изобразяват графично в „Reaction products“

Софтуерният модул Ambit-SMIRKS е достъпен като Java библиотека, OpenTox алгоритъм и API съвместима уеб услуга и може да бъде тестван на следната уеб страница: <http://apps.ideaconsult.net:8080/ambit2/depict/reaction>.

1.5. Тестване и приложение на Ambit-SMIRKS

Функционалностите на Ambit-SMIRKS са разработвани, подобрявани и тествани в продължение на няколко години за решаването на различни задачи в химичната информатика. Алгоритъмът за SMIRKS базирани трансформации в модула Ambit-SMIRKS е приложен основен модул по химична информатика в няколко проекта разработени в нашата група и от външни изследователски групи.

Toxtree [10][11] е софтуерното приложение със свободен достъп, което класифицира веществата в няколко категории и предсказва техните токсични ефекти, базирайки се на молекулните им структури. Toxtree се състои от множество модули, в които се прилагат различни модели (напр. правилата на Cramer, схемата на Verhaar, базата данни с правила на Benigni/Bossa за мутагенност и канцерогенност и т.н.). За да се оцени бионаличността, активността и профила на токсичност, трябва да се вземат предвид метаболитните биотрансформации на целевото съединение и няколко от модулите включват правила с трансформации на химичните структури. Затова след прилагане на определено правило, последващата обработка продължава не с входящата молекула, а с набор от реакционни продукти. Тези трансформации се реализират чрез използването на Ambit-SMIRKS.

Ambit-SMIRKS е използван в проекта envPath – система за прилагане на химични реакции, представени като SMIRKS линейни нотации. envPath [12] е база данни и система за предсказване на биотрансформации на органични замърсители в околната среда, която на базата на експериментално доказани биотрансформации предоставя различни модели за предсказване на възможни реакционни пътища. За всяка стъпка на трансформиране на молекулите се използва Ambit-SMIRKS.

AmbitCLI е конзолно приложение [13], част от платформата за химична информатика AMBIT, което включва редица възможности за обработка на химически структури. Приложението AmbitCLI е използвано за стандартизация на ChEMBL, PubChem и други бази данни със свободен достъп, като стандартизираните структури са компилирани в ExCAPE [14] - интегриран мащабен набор от данни, улесняващ Big Data анализа в химичната геномика. Инструментът за стандартизация се използва за обработка на данни и от индустрията.

Библиотеката Ambit-SMIRKS е използвана за прилагане на правила за биотрансформации и генериране на структури в BioTransformer [15] – свободно достъпен софтуерен пакет за предсказване на метаболизъм на малки молекули (тествани са 1,284 биотрансформации).

Ambit-SMIRKS е използван за анализ по метода „дърво на решенията“ на Cramer 3 (разбор на трите основни категории токсичност: ниска, средна и висока – т. нар. класове (съответно I, II и III)). Във връзка с това работихме по 29 правила, съставени от експерти токсиколози с налични знания по отношение на вещества с известна токсичност. Прилагането на такъв анализ изискваше систематизиране на данни за (1) структурните свойства (описани чрез линейни нотации SMARTS) и (2) биотрансформациите на веществата в организма (описани чрез линейни нотации SMIRKS). Изготвените като линейни нотации SMIRKS трансформации бяха тествани в модула Ambit-SMIRKS с 548 съединения, 108 от които налични в храната и 440 – в организма. Токсичността на тези съединения е известна, така че освен тестване на трансформациите, с Toxtree паралелно бяха тествани и новите правила, използвани в Cramer 3, и сравнени с оценките на експертите от Cramer 2.

2. Виртуален химичен реактор (Ambit-Reactor)

2.1. Обща характеристика

В настоящата дисертация е разработен софтуерен модул за симулация на химични реакции Ambit-Reactor [16]. За дадено множество от начални реагенти, Ambit-Reactor прилага изчерпателно всички възможни трансформации от предварително зададен набор от правила за протичане на генерични химични реакции (база данни с трансформации) до достигане на предварително зададени условия, които формално се дефинират чрез логически изрази за стойностите на дескриптори, изчислени за химическите обекти, участващи в реакциите.

2.2. База данни с реакции

При направената справка се установи, че броят на реакциите, включени в публично достъпни бази данни, и същевременно представени в компютърно-четим вид, е много малък. В тази връзка бе изготвена база данни с правила за протичане на генерични химични реакции, представени в компютърно-четим вид, в който могат лесно да се обработват и съхраняват. От литературата [17] бяха избрани 242 генерични реакции, за които бяха изготвени съответстващите им SMIRKS линейни нотации. Прилагането им осигурява възможност за описване на множество синтетични реакции и следователно могат да намерят широко приложение за решаване на различни проблеми в химичната информатика, като например генерирането на комбинаторни библиотеки от съединения.

Допълването на дисертацията и с биотрансформации се счете за необходимо, тъй като при литературния преглед се установи, че към момента същите са актуални и представляват интерес от химична и най-вече от фармацевтична гледна точка – с оглед на факта, че може да се предскаже дали ще повишат биоактивността на дадени съединения (основно ксенобиотици) или ще деактивират терапевтичните му свойства. В тази връзка беше изготвена база данни с биотрансформации, основно такива, чието протичане е катализирано от ензими от семейството на цитохром P450 (CYP450) –

изборът се основа на литературни данни [18], съгласно които въпросните ензими са отговорни за биотрансформациите при метаболизма на повече от 90 % от всички лекарства. Базата данни бе допълнена с биотрансформации, които при литературната справка и при прегледа на бази данни с реакции със свободен достъп [19] се установи, че са най-често срещаните реакции. След като същите бяха представени като линейни нотации SMIRKS, бяха систематизирани в конфигурационен файл, който може да бъде обновяван и допълван с нови трансформации.

2.3. Конфигуриране на виртуалния реактор Ambit-Reactor

Виртуалният химичен реактор се конфигурира чрез JSON конфигурационни файлове, в които се определят правилата за протичане на реакции, стратегията за прилагане на реакции, позволените продукти, забранените продукти, стойностите на параметрите, които определят логическите условия за прилагане на реакции и местата, където реакциите могат да протекат. Всяка реакция е представена като отделна секция в JSON файла в следния формат:

```
{
  "NAME" : "Aldehyde oxidation",
  "CLASS" : "phase1",
  "SMIRKS" : "[H][C:1]=[O:2]>>[C:1](O[H])=[O:2]",
  "USE_CONDITIONS" : ["REACTION_APPLICATIONS_PER_REACTANT < 2"]
}
```

Основното описание на реакцията включва име (Name), реакционен клас (Class), линейна нотация SMIRKS, и условия за прилагане на реакцията (Use Conditions).

Наборът от химически и логически условия се нарича „стратегия на реактора“. Тези условия се определят чрез логически изрази за стойностите на дескрипторите, изчислени за съединенията, участващи в реакциите.

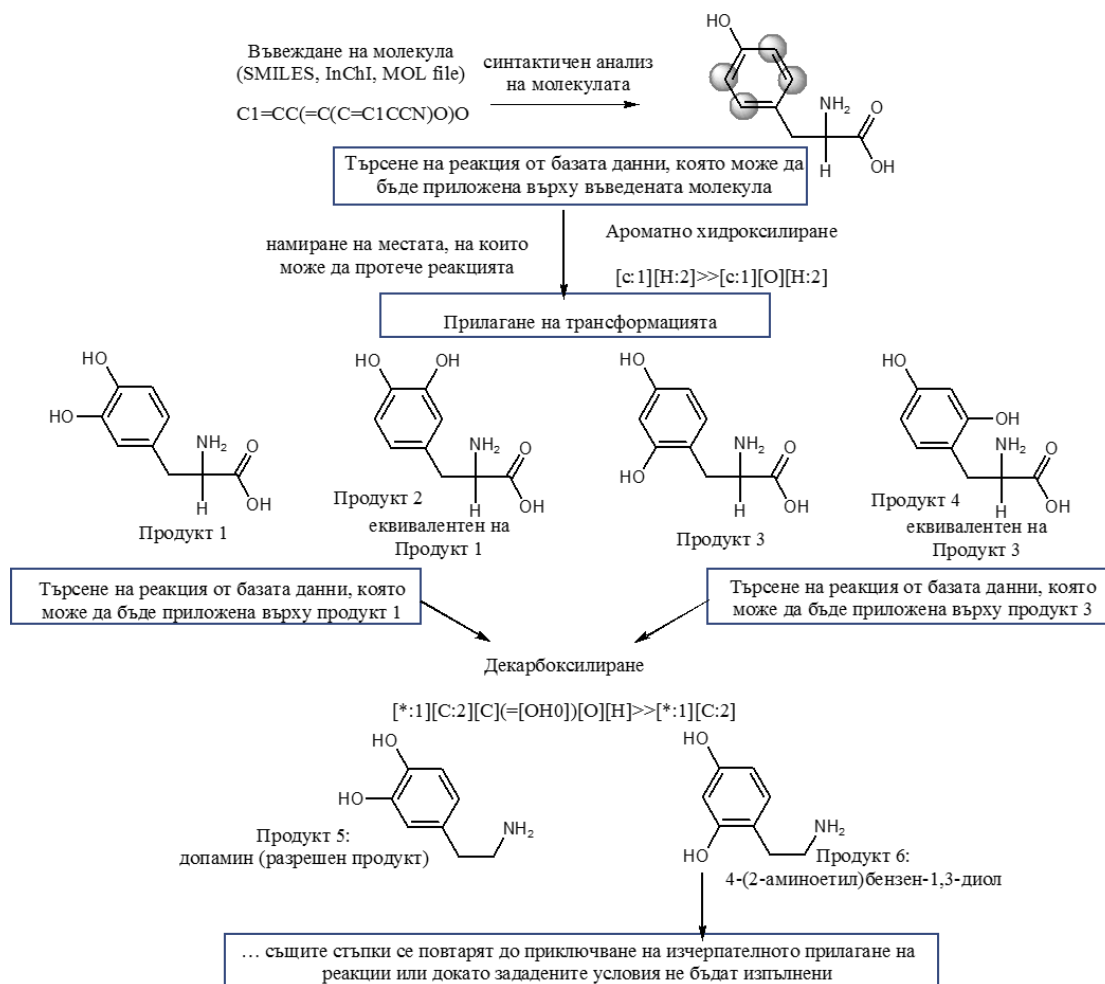
2.4. Софтуерна архитектура и алгоритъм на Ambit-Reactor

Софтуерният модул Ambit-Reactions се характеризира със структурно представяне, вход, изход и обработка на структурна информацията, базирани на CDK вътрешно представяне на химични обекти. Химичните файлови формати, които се поддържат, са SMILES [7], [8], InChI, MOL/SDF [20], CML [21].

Входящата целева структура се задава чрез някой от популярните формати за представяне на структури: SMILES, InChI or *.MOL file. Прави се синтактичен анализ на реактанта и на продукта, и атомите на обектите се съпоставят (mapping). Провежда се субструктурно търсене в базата данни с реакции и за всички открити съответствия, върху зададената молекула се прилагат възможните реакции. За целевата молекула, Ambit-Reactor прилага изчерпателно всички възможни трансформации от предварително дефинираната база данни с генерични реакции. За всеки от продуктите всички възможни трансформации се прилагат отново, за да се получат нови продукти и т.н. За да се контролира комбинаторната експлозия, процесът спира при достигане на определени от

потребителя условия, напр. брой на реакциите, които се извършват, достигане до „позволен“ или „забранен“ продукти и т.н. Симулацията на реакциите може да бъде представена като дървовидна структура от данни, при която всеки възел в дървото (граф тип дърво) представлява конкретните продукти, получени от реакциите, приложени към съединенията от възлите на горното ниво.

Работният процес на алгоритъма на реактора за молекулата на тирозин (Tyrosine) е илюстриран на Фигура 5:



Фигура 5: Работен процес на Виртуалния химичен реактор, демонстриран за молекулата на тирозин

В структурата на тирозина се откриват четири възможни места, на които може да бъде приложена реакция на ароматно хидроксилане, при което се получават четири съединения. Ако разглеждаме множеството като насочен граф, четирите получени продукта представляват наследници на целевата молекула.

В стратегията на реактора е зададено с условие дадена реакция да не се прилага на повече от 2 места за една и съща молекула, затова в процеса, демонстриран на Фигура 5 ароматно хидроксилане не се прилага изчерпателно за продуктите 1 и 3 (които в случая са еквивалентни на 2 и 4). За получените продукти софтуерът прилага реакция на

декарбоксилиране – образуват се допамин (продукт 5) и 4-(2-аминоетил)бензен-1,3-диол (продукт 6). Структурата на допамина е в списъка с позволени продукти и следователно този връх в дървото е терминален (виртуалния химичен реактор преустановява търсенето на реакции, които могат да протекат върху тази молекула). Алгоритъмът на реактора продължава с прилагане на трансформации върху продукт 6. Аналогични стъпки се повтарят за получените продукти докато се приложат изчерпателно всички възможни реакции, или се изпълни някое от предварително зададените условия в стратегията на реактора.

3. Ретросинтезен реактор

Разработеният в настоящата дисертация Ретросинтезен реактор е софтуерен модул, който реализира система за ретросинтезен анализ на базата на дефинирани химически знания и логически условия. Ретросинтезният реактор предлага подход за разчленяване на зададената целева молекула до достигането на достъпни стартови материали.

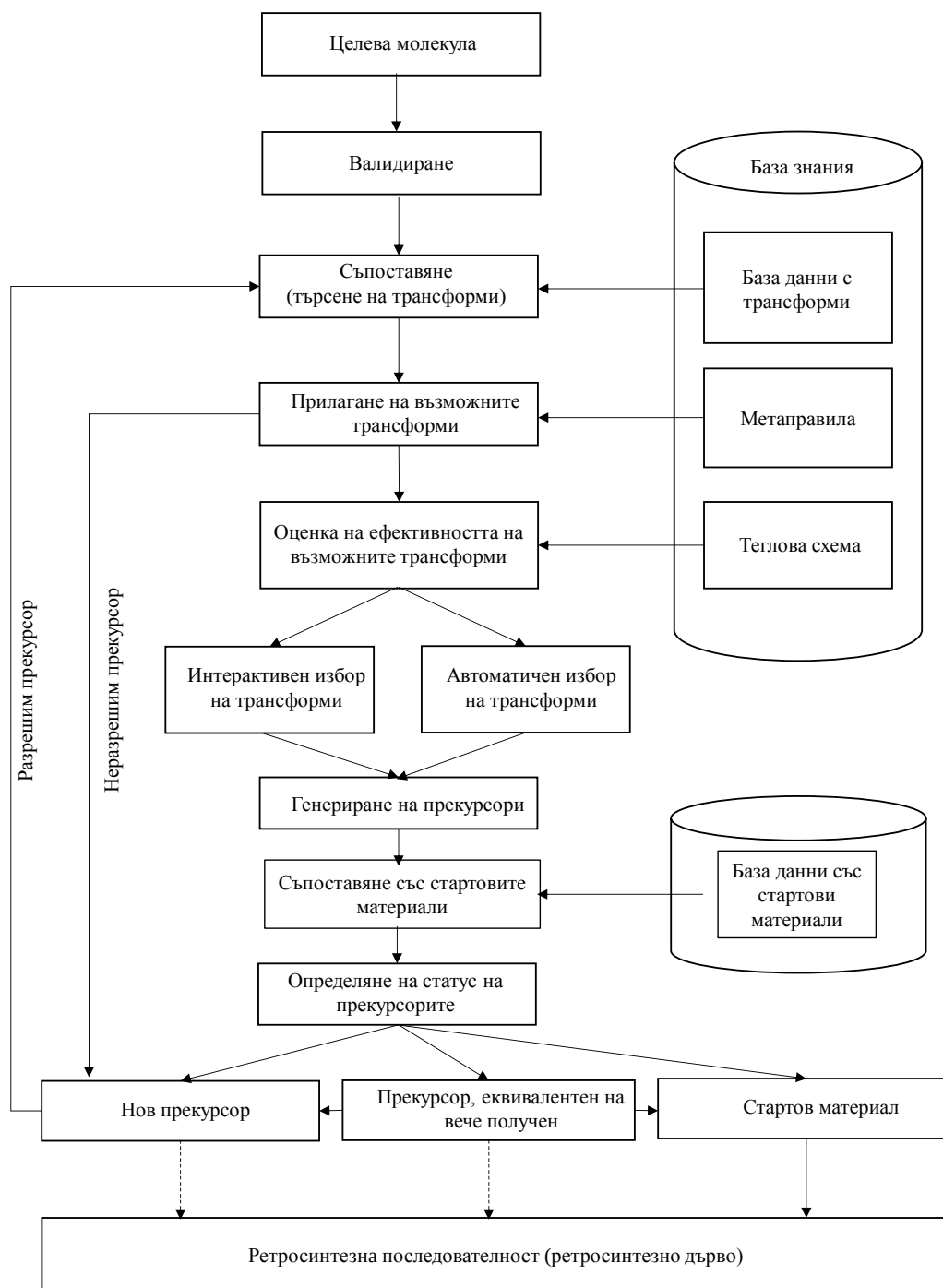
3.1. Обща характеристика

Входящата структура (целевата молекула, на която искаме да направим ретросинтезен анализ), се задава чрез някой от поддържаните стандартни формати, трансформира се във вътрешно представяне като CDK обект и се подлага на валидиране и стандартизация. Процесът на стандартизация включва премахване на зарядите от функционални групи (с някои изключения, като нитрогрупи), проверка и валидиране на типовете връзки, добавяне на експлицитни водородни атоми и др.

За входящата структура, като и за прекурсорите от всяко ниво от ретросинтезното дърво се прилага субструктурно търсене (реализирано чрез модула *Ambit-SMARTS* [3]), за да се установи кои от трансформите, налични в базата данни, е възможно да се приложат.

Ретросинтезният реактор се конфигурира чрез използването на множество логически правила, логически променливи, флагове и теглова схема за описание ретросинтезната стратегия. Формално комбинаторната експлозия се контролира чрез правилата за прилагане на трансформи, зададени чрез логически изрази за стойностите на дескриптори, изчислени за химическите обекти, участващи в реакциите. Процесът се повтаря за получените прекурсори, докато не се стигне до получаването на стартови материали, налични в базата данни.

Основната схема на работния процес на Ретросинтезния реактор е показана на Фигура 6:



Фигура 6: Схема на работен процес на Ретросинтезния реактор

Ретросинтезното дърво се конструира прогресивно, като за всяко ниво от дървото, след прилагане на всички възможни трансформи от базата данни, се оценяват и избират най-оптималните трансформи според параметрите на стратегията и следва конструиране на следващото ниво на ретросинтезното дърво.

Въз основа на зададените критерии се избират най-ефективните пътища за получаване на целевата молекула, които в числов вид могат да бъдат включени в теоретичната оценка на синтетичната достъпност на целевата молекула.

3.1.1. База данни със стартови материали

За създаване на множество със стартови материали, необходима за Ретросинтезния реактор, използвахме ZINC [22] – база данни със свободен достъп, създадена за „виртуален скрининг“. Тъй като един от основните критерии за избор на синтетична схема е участващите в нея стартови материали да са достъпни, използвахме съединения, налични в каталога на Sigma Aldrich, който се предлагат на пазара.

3.1.2. Молекулна комплексност

Ентропията на Shannon (уравнение (1)) играе централна роля в информационната теория като мярка за информация и неопределеност [23].

$$(1) \quad H = -\sum_{i=1}^N p_i \log p_i$$

където p_i е вероятността за принадлежност към клас/група i , а N – броя на възможните класове (инварианти).

Това уравнение може да бъде приложено за измерване на комплексността на дадена молекула, която разглеждаме като граф, чиито връзки и типове атоми дефинират съответните инварианти на молекулната комплексност.

Освен прилагането на концепцията на Shannon за изчисляване на молекулната комплексност, възможно е по същия начин да се дефинира комплексността за всяка отделна атомна околност. В този случай броят и разнообразието от пътища, произтичащи от атома определят комплексността (CA) на всяка атомна околност. В случая p_i е относителната честота на поява (оценена като вероятност) на пътищата от клас i , а N е общият брой на пътищата (с определен брой връзки), излизащи от атома.

$$(2) \quad CA = -\sum_i p_i \log_2 p_i + \log_2 N$$

$$(3) \quad CM = \sum_k CA_k$$

$$(4) \quad CM^* = \log_2 \left(\sum_k 2^{CA_k} \right)$$

Молекулната комплексност (CM) може да бъде дефинирана или като сума от CA_k за всеки атом, както е показано в уравнение (3), или като логаритъм от сумата на експонентите на CA_k за всеки атом, както е показано в уравнение (4).

За всеки атом се установяват всички пътища до определен размер, излизащи от този атом. Относителната честота на поява p_i се определя за всяка група (i) еквивалентни пътища в околността на всеки атом; респективно тези величини се използват във формула (2) за изчисляване съответно на атомната комплексност СА.

СА се променя логически в редица молекулни трансформации и се прилага в ретросинтеза за идентифициране на региони с по-висока комплексност, които могат логично да бъдат приоритизирани за разкъсване на стратегическа връзка.

3.1.3. Ретросинтезна стратегия

Ретросинтезният реактор представлява експертна система, базирана на правила. Трите основни компонента, чрез които тези правила са имплементират, обособяват базата данни със знания на Ретросинтезния реактор, която съдържа (1) база данни с трансформи (ретрореакции); (2) метаправила за тяхното управление, и (3) теглова схема за оценка на ефективността на възможните трансформи.

3.1.3.1. База данни с трансформи (ретро-реакции)

Основата на базата със знания на Ретросинтезния реактор е множеството от 142 трансформа, представени в компютърно-четим формат. За описанието на трансформите е използвана линейната нотация SMIRKS. Основните причини за избора на този формат са, че (1) позволява детайлно описание на реакционите центрове; (2) широко разпространен е и се използва от множество софтуерни пакети и програмни инструменти за молекулно моделиране (свободни и търговски); (3) може да се редактира чрез обикновен текстов редактор.

Трансформите, представени във вид на линейни нотации SMIRKS, се характеризират със следния общ модел:



където А е продуктът на дадена реакция в синтезно направление (целева молекула в ретросинтезния анализ), а В и С са реактантите (прекурсорите в ретросинтезния анализ).

В SMIRKS линейните нотации са описани само субструктурните елементи, които директно участват в протичането на реакцията (т.е. реакционния център) или се считат за съществени за неговата реактивност. Това описание е в основата на прилагането на генерични реакции, широката приложимост на които ни даде основание да ги имплементираме в базата данни с трансформи.

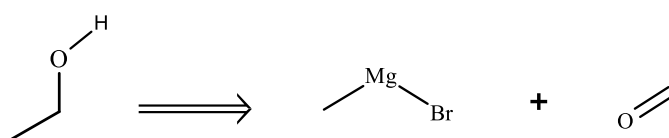
Като основа за компилираното от нас множеството използвахме публикуваните от D'Angelo и Smith 82 реакции, разделени в две основни групи: образуване на въглерод-въглеродна връзка и промяна на функционални групи [24], които от своя страна са разпределени в зависимост от функционалната група, до получаването, на която води прилагането на дадена реакция. С цел изчерпателност на набора от избрани реакции, използвахме и публикациите на Schuerer et al. [25], Hartenfeller et al. [26], Masek et al. [27] и Hartenfeller et al. [28].

Освен представянето на трансформите чрез езика SMIRKS, в базата данни са включени допълнителни атрибути на реакциите като ID/ Name (идентификационен код и наименование на реакцията); FunctionalGroup (до образуването на коя функционална група води прилагането на реакция, съответстваща на дадения трансформ); ExperimentalConditionsDescription (за описание на реакционните условия); Info (допълнителна информация), както и количествени критерии за оценка на приоритета, с който да се приложи даден трансформ.

3.1.3.2. Метаправила за прилагане на трансформи

Ретросинтезният анализ е пряко зависим от решаването на два основни проблема: първият е генерирането на ретросинтезни предложения, а вторият е валидирането на тези предложения като реално приложими синтетични реакции.

Класическото решение на проблема с интерференцията на химическата функционалност е към предложените трансформации да има „изключващи“ правила. За следния пример (Фигура 7) може да бъде зададено такова правило, което да указва, че това присъединяване с участие на органометално съединение може да бъде приложено само ако нито една от участващите молекули не съдържа нитрилна група.



Фигура 7: Генерична реакция

Описанието на тези метаправила е реализирано с помощта на линейни нотации SMARTS. Част от дефинираните метаправила за примера на Фигура 7 са следните SMARTS линейни нотации:

Exclude: [NX1]#[CX2;H1,\$(C([#6]))]

(с нитрили гринярдовият реактив може да образува кетони)

Exclude: [\$([CX3]([#6])([#6])),\$([CX3H]([#6])),\$([CX3H2])]=[\$([NX2]([#6])),\$([NX2H])]

(с имини гринярдовият реактив може да образува амини)

Exclude: [CX4]1[CX4][N]1

(с азиридины гринярдовият реактив може да образува амини)

Трябва да се отбележи обаче, че при някои класове реакции дори подробна спецификация на реакционното „ядро“ понякога е недостатъчна; така например при ароматни системи, в които „отдалечени“ заместители могат да имат решаващо влияние върху протичане на реакциите (електронодонорни и електроноакцепторни заместители).

3.1.3.3. Оценка на ефективността на трансформите

Когато се прилагат множество трансформи на няколко нива, Ретросинтезният реактор, подобно на други методи, базирани на правила, генерира голям брой съединения (прекурсори). И докато един химик може доста успешно да оцени реакциите на база знания и опит, компютърната програма, която се стреми да направи същия избор, трябва да работи по по-формален начин на базата на количествени критерии. За всяка реакция се изчислява така наречения индекс на реакцията (reaction score), с който се отчитат приносите на различни молекулни дескриптори и параметри на ретросинтезната стратегия. За оценка на ефективността на трансформите избрахме следните параметри:

1. Тип на трансформа (S_{type}): Използвана е класификацията, дадена от Angelo et al. [24] и реакциите са разделени условно в две категории: първата съдържа реакции за образуване на въглерод-въглеродна връзка, а втората – за промяна на една функционална група в друга – реакции за промяна на функционални групи (FGI). В ретросинтезно направление разчленителните трансформи са с по-висок приоритет, тъй като водят до получаването на прекурсори с по-ниска молекулна комплексност.

2. Приоритет на трансформа (S_{prior}) – на всеки трансформ се приписва числова стойност, която е показател за това с какъв приоритет да се приложи дадения трансформ: трансформите от тип 1 (разчленителни) ще се приложат приоритетно пред тези от тип 4 (с добавяне на функционална група) например.

3. Основен принос (S_{basic}) – на широко разпространените и добре познати реакции, които често се прилагат при планирания органичен синтез, се приписва по-висок основен принос [29].

4. Принос (score) отчитащ експериментални условия на реакцията (S_{exper}) – отразява разнообразието от реагенти, които могат да бъдат използвани за извършване на определена трансформация. Ако реакцията може да се извърши при различни условия, тя ще има по-висок принос от експериментални условия на реакция, например, която изисква специфичен реагент, катализатор, специални условия и т.н. Оценката на експерименталните условия е съобразена с критериите, публикувани от Cooper et al. [30].

5. Надеждността на даден трансформ (S_{rel}) е мярка за вероятността добивът да попадне в типичния диапазон за голямо разнообразие от реактанти. Стереоселективността, където е фактор, може да бъде включена в тази категория.

6. Принос (score) от интервал на добив (S_{yield}) – информация, получена за добив на реакцията по литературни данни.

7. Принос от комплексността на реакционния център (S_{rcc}) – за да се изчисли приноса на комплексността на реакционния център (S_{rcc}) се използва следната формула:

$$(5) \quad S_{rcc} = k_{rcc} * \left(\sum_{i=1}^n CA_i \right) + m_{rcc}$$

където k_{rcc} и m_{rcc} са коефициенти на линейно преобразуване на комплексността в принос, нормиран в проценти от 0 до 100.

8. Принос от комплексността на продуктите (S_{prod}) – приносът на комплексността на продуктите (прекурсорите в ретросинтезно направление) се изчислява като:

$$(6) \quad S_{prod} = k_{prod} * \left(\sum_{i=1}^n CM_i \right) + m_{prod}$$

където k_{prod} и m_{prod} са коефициенти на линейно преобразуване на комплексността в принос, нормиран в проценти от 0 до 100. Ако в хода на ретросинтезния анализ се получи молекула, която съответства на някое от съединенията в базата данни със стартови материали, стойността за приноса на комплексност (CM_i), която му се приписва, е 0.

За изчисляване на ефективността на трансформите като формален индекс на реакцията S_R (описващ приложението на реакцията върху конкретно място за дадена целева молекула), е приложена адитивна схема. Приносите от отделните компоненти, изразени като числови стойности, нормирани в проценти от 0 до 100, се умножават с различни тегла за всяка индивидуална категория и се сумират. Полученият резултат S_R (индекс на инстанцията на реакцията) се използва като количествен показател, указващ колко ефективен е дадения трансформ по отношение на целевата молекула:

$$(7) \quad S_R = \sum_k w_k * S_k$$

където $k \in \{basic, type, prior, exper, rel, yield, rcc, prod\}$

Колкото по-висока стойност за S_R има един трансформ, толкова по-вероятно е същият да бъде предложен с предимство пред останалите при конструирането на ретросинтезното дърво.

3.1.4. Ретросинтезно дърво

Когато се намери съвпадение между зададената целева молекула и трансформ от базата данни, Ретросинтезният реактор прилага този трансформ и получаването на новата структура представлява стъпка, която води до обособяване на следващо ниво в ретросинтезното дърво. Този процес се повтаря за всички трансформи, които е възможно да бъдат приложени. Поредицата от последователни предложени трансформации по този начин е ретросинтезна последователност, която от информационна гледна точка представлява дървовидна структура от данни – конкретно ретросинтезно дърво. Съвкупността от всички възможни ретросинтезни последователности е пълното ретросинтезно дърво или така нареченото дърво на решенията, в което отделните възли представляват частични или пълни (за крайните елементи на дървото) ретросинтетични последователности.

3.1.5. Режими на работа

Ретросинтезният реактор може да бъде използван в два режима: (i) автоматичен, при който оценката на ефективността на трансформите, които могат да бъдат приложени към целевата молекула, се прави с помощта на предварително дефинирана база данни със знания, и системата прилага трансформите, оценени като най-ефективни върху целевата молекула; и (ii) интерактивен, при който потребителят може сам да определи кой трансформ иска да бъде приложен и така да насочи хода на ретросинтезния анализ в зависимост от собствените си знания и опит (освен върху избора на трансформ, интерактивния модел предоставя възможност и за вариране на теглата, влияещи върху приоритизирането на най-ефективните реакции).

3.2. Алгоритъм на Ретросинтезния реактор

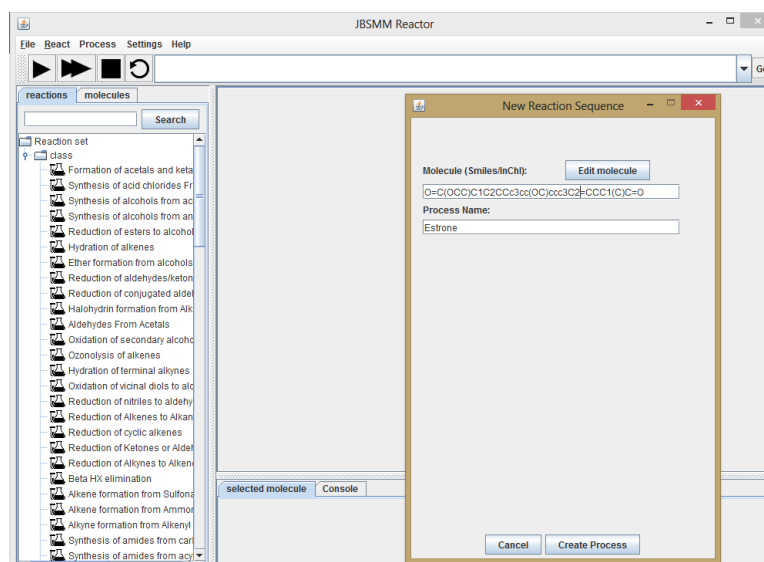
Алгоритъмът за ретросинтезен анализ, разработен като част от модула *Ambit Reactions*, ползва два основни пакета: *retrosynth* (съдържа основните класове, които служат за конструиране на ретросинтезно дърво) и *reactions.rules.scores* (съдържа няколко класа, служещи за изчисляване на молекулни дескриптори).

Ретросинтезният реактор прилага алгоритъм тип „обхождане в дълбочина“ (*depth-first-search*). За реализация на алгоритъма се избира даден връх (възел) от ретросинтезното дърво, който се обозначава като корен (връх без предшественици) и обхождането стартира от него. Последователно се посещават всички следващи върхове до достигането на терминален връх без наследници (съответстващ на стартов материал или неразрешим прекурсор), след което се осъществява търсене с връщане назад (*backtracking*) до достигане на нова крайна точка или при цялостно реализирано обхождане - към корена.

3.3. Ретросинтезен реактор – графичен потребителски интерфейс (GUI)

JBSMM-Reactor е новоразработен софтуерен модул за симулиране на химични реакции, изграден върху *java* библиотеките *Ambit-SMARTS* и *Ambit-Reactions* на платформата за химична информатика *Ambit*. *JBSMM-Reactor* включва функционални модули и графичен потребителски интерфейс (GUI), в който е имплементиран като компонент и Ретросинтезния реактор.

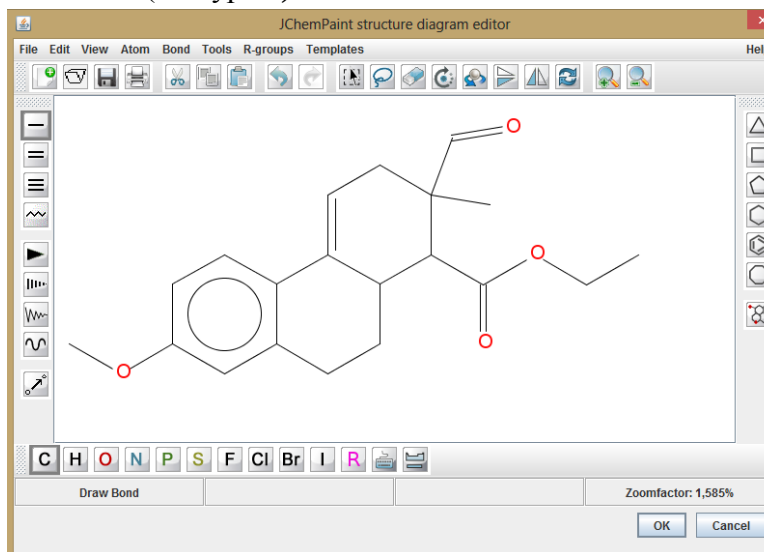
Входящата структура (целевата молекула, на която искаме да направим ретросинтезен анализ), се въвежда от менюто *Process>>New Reaction Sequence* (Фигура 8):



Фигура 8: Въвеждане на целева молекула

В полето „Process name“ потребителят може да зададе желано име на процеса (което би било полезно в случаите, когато се работи с много процеси едновременно).

Съединението може да бъде представено във вид на линейна нотация (SMILES или InChI), или да бъде зададено с помощта на графичния редактор, достъпен при избор на бутона „Edit molecule“ (Фигура 9).



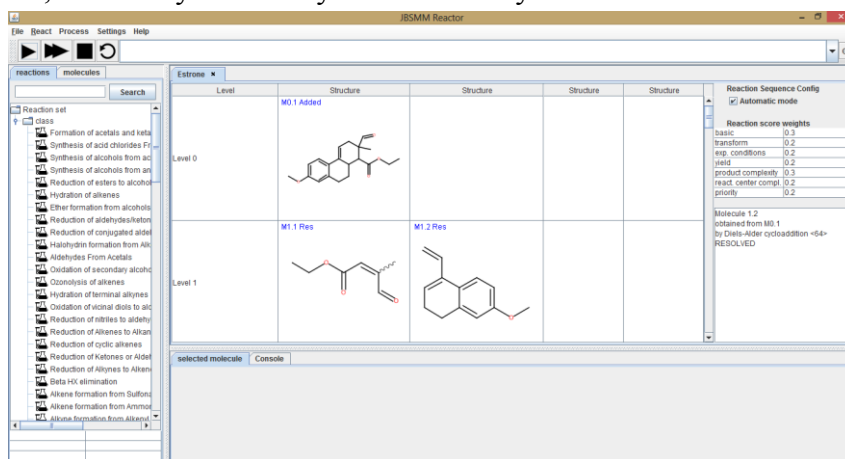
Фигура 9: Графичен редактор на JBSMM-Reactor

Бутонът “Create process” зарежда зададената молекула, която се визуализира като Level 0. Всяко следващо съединение, получено от тази молекула, се обособява на следващи нива – 1, 2, 3 и т.н., в зависимост от броя на приложените възможни трансформи.

От потребителя се очаква да избере в какъв режим иска да работи: интерактивен или автоматичен. Потребителският интерфейс позволява лесно превключване от единия

в другия режим с поставяне на отметка в полето „Automatic mode“. От същия блок, потребителят може да конфигурира условията на стратегията, като варира теглата, участващи в адитивната схема за оценка на ефективността на трансформите, които е възможно да бъдат приложени върху зададената целева молекула.

При избор на автоматичен режим, ретросинтезният анализ се стартира с помощта на бутона „►“ или от менюто React>>Run. На екрана се визуализират получените прекурсори, подредени по нива. Приложението дава информация за трансформа, който е бил приложен, и за статуса на получените молекули.



Фигура 10: JBSMM Reactor: модул за ретросинтез (автоматичен режим)

При избор на интерактивен режим, потребителят има възможност да насочи хода на ретросинтезния анализ в желана от него посока. За всяка целева молекула, Ретросинтезния реактор извежда списък с трансформите, които може да приложи, подредени в низходящ ред в зависимост от изчислените оценки за ефективностите на реакциите (reactions scores) оценени спрямо целевата молекула от конкретното ниво:

No	Reaction info	Target	Products	Score
1	Diels-Alder cycloaddition[C:1][C:2][C:3][C:4]			78.4027786441427
2	Williamson ether synthesis[C:1][C:2][C:4][C:5]			77.38862330318648
3	Williamson ether synthesis[C:1][C:2][C:4][C:5]			75.82363231572681

Фигура 11: Избор на трансформ в интерактивен режим

Изборът на трансформи, които могат да се приложат, се прави по същия начин и за останалите нива: така потребителят може да конструира ретросинтезна последователност в зависимост от собствените си критерии, знания и опит.

3.4. Ретросинтезен реактор – дискусия

Разработената стратегия за ретросинтезен анализ осигурява ефективно решение за селекция на трансформи, които са най-подходящи за зададената целева структура. Логическата последователност на алгоритъма следва принципите на класическия ретросинтез [31] по формален начин, като за целта в разработения модел са имплементирани множество логически и химически условия, които да насочват хода на процеса.

Един от основните принципи в ретросинтеза е молекулите да се охарактеризират по отношение на елементите на тяхната комплексност, и след това се идентифицират трансформите, прилагането на които би довело до намаляване на тази комплексност. В Ретросинтезния реактор този принцип се реализира с помощта на адитивната схема за оценка на възможните трансформи. Избрани са такива компоненти, които да отразяват химическата логика при ретросинтезния анализ. Теглото, приписано на всеки компонент, играе основна роля в оценката на ефективността на трансформите, които е възможно да бъдат приложени. И макар всички компоненти на адитивната схема да носят съответния принос за оценка трансформите, задаването на по-голямо тегло на някои от тях е от съществено значение при конструиране на ретросинтезното дърво.

Молекулната комплексност е основен фактор при провеждане на ретросинтезния анализ и имплементирането на такъв дескриптор в стратегията на Ретросинтезния реактор е обусловено от основната цел на анализа, а именно получаване на целевата молекула от по-прости молекули (прекурсори). Следва да се отбележи, че използването на молекулната комплексност като дескриптор доказва възможността количествен параметър (дескриптор) да бъде приложен за реализиране на химическа логика. Последното е основен критерий при дизайна и разработването на стратегия за компютърно-подпомогнат органичен синтез, при който значителен обем от знания трябва да се формализират в компютърно-четим вид и да се приложат по последователен и ефективен начин.

Друг компонент в адитивната схема с първостепенна важност, това е систематизирането на трансформите, налични в базата данни със знания, в различни категории. Логично, при ретросинтезния анализ ще бъдат приложени с предимство трансформите, дефинирани като силно опростяващи, тъй като, както вече е отбелязано, водят до получаване на прекурсори с по-ниска молекулна комплексност.

Трансформите, които са включени в базата знания на Ретросинтезния реактор, са избрани така, че да покриват обхвата на най-често използваните и добре познати реакции в органичния синтез. Изборът на трансформи може да бъде обоснован с множество публикации от литературата, в които се обобщава информация за най-широко разпространените реакции. В Ретросинтезния реактор на такива реакции се приписва висока стойност за основен принос, тъй като на практика това са реакциите, прилагани

най-често в реални лабораторни условия. Този фактор е изключително важен, тъй като е пряко свързан с възможността конструираното ретросинтезно дърво да бъде реализирано в синтезно направление, като за целта се използват, ако е приложимо за дадената целева молекула, надеждни, добре познати, често използвани реакции.

Тъй като в зададената целева молекула може да не присъства пълният ретрон, нужен за прилагането на конкретен трансформ, Ретросинтезния реактор има капацитета да приложи множество други трансформи, които сами по себе си не са опростяващи, но водят до получаването на структури, съдържащи в себе си необходимия ретрон. Систематизирането на възможните трансформи в класове и задаването на компонент като приоритет на трансформата с по-голямо тегло в адитивната схема осигурява прилагането на силно опростяващи трансформи по време на целия ретросинтезен анализ: по този начин се контролира и комбинаторната експлозия, тъй като неконтролираното и изчерпателно прилагане на всички възможни трансформи не е желателно и не може по никакъв начин да бъде отнесено към концепцията за логичен и планиран органичен синтез.

В стратегията на Ретросинтезния реактор са включени и параметри, установени на база на експериментални изследвания, публикувани в литературата: това са приносът от експериментални условия и интервал на добив от реакцията. Тези критерии са значими от гледна точка на практическата приложимост на реакциите, съответстващи на трансформите от конструираното ретросинтезно дърво: като по-ефективни са оценени реакциите, които дават по-висок добив и/или се извършват без да е необходимо да се осигуряват специфични условия като много ниски/ високи температури, налягане, специфични катализатори и пр.

Прилагането на топологичната стратегия в ретросинтезния анализ сме реализирали чрез метрика, състояща се в определяне на комплексността на реакционния център. Изследването на комплексностите на различните атомни околности при зададена молекула може да се използва като показател за мястото на протичане на реакцията. Така се идентифицират стратегическите връзки, по които се прилагат трансформите, и съответно други връзки, които трябва да се запазят в хода на ретросинтезния анализ. По този начин ретросинтезното разкъсване на стратегически връзки, имплементирано в адитивната схема за оценка на трансформите като принос от комплексността на реакционния център, може да бъде предложено както за ациклични вериги, така и за връзки в кондензирани, спиро, или мостови циклични системи. Така най-ефективното опростяване на целевата молекула се постига на базата на количествен параметър, без да се налага прилагането на различни критерии за всеки от изброените топологични типове системи, каквито някои системи за ретросинтезен анализ ползват.

В базата данни на Ретросинтезния реактор за всеки от трансформите е въведена информация за това до образуването на каква функционална група води прилагането на реакцията в синтезно направление, съответстваща на дадения трансформ. Това позволява гъвкавост при използването на различни трансформи, тъй като може да се модифицират селективно определени функционални групи, налични в целевата молекула.

Както при всяка експертна система, така и за Ретросинтезния реактор може да се отбележи като недостатък невъзможността да предскаже прилагането на реакция извън

собствената си база данни със знания. Това обаче не рефлектира върху качеството на получените резултати в значителна степен, тъй като Ретросинтезният реактор е създаден по-скоро за да даде ефективна оценка и практическа насока при провеждането на ретросинтезен анализ с цел планиран органичния синтез, отколкото да обезпечава рядко използвани или „екзотични“ реакции.

4. Оценка на синтетичната достъпност

Оценката на синтетичната достъпност представлява числена стойност, нормирана в интервала от 0 до 100, получена на база приносите на четири молекулни дескриптора: комплексност на циклите (S_{RC}), цикломатично число (S_{μ}), стереохимична комплексност (S_{SC}) и молекулна комплексност (S_{CM}), умножени със съответните тегла (w_i) :

$$(8) \quad SA = w_{RC} * S_{RC} + w_{\mu} * S_{\mu} + w_{WSC} * S_{WSC} + w_{CM} * S_{CM}$$

За изчисляване на стойността на всеки компонент е използван отделен алгоритъм: първият алгоритъм изчислява комплексността на циклите в молекулата (S_{RC}) на базата на множеството на най-малките прости цикли (SSSR, smallest set of smallest rings); вторият алгоритъм представлява просто определяне на цикломатичното число μ ; третият алгоритъм дава информация за стереохимичната комплексност въз основа на установения брой на претеглените хирални центрове в молекулата (N_{WSC}). Четвъртият алгоритъм изчислява молекулната комплексност (S_{CM}) въз основа на информационната теория и разглеждането на молекулата като граф.

За преобразуване на даден молекулен дескриптор d_i в принос (S_i), нормиран в проценти от 0 до 100, който да се използва в адитивната схема за оценка на синтетичната достъпност, са използвани линейно преобразование от вида $k_i \cdot d_i + m_i$.

4.1. Комплексност на циклите

Комплексността на циклите се счита за важна при оценяване на синтетичната достъпност, тъй като тя е показател за наличието на спиро, мостови и кондензирани системи, които биха могли да създадат трудности при синтеза на съединението [32].

За изчисляване на комплексността на циклите (RC) сме приложили подхода, описан от Gasteiger и Jochum [33], като сме използвали следната формула:

$$(9) \quad RC = \frac{RSS}{nRA}$$

където RSS (Ring Size Sum) е сумата от размерите на всички цикли от SSSR (множеството на най-малките прости цикли), а nRA е броят на всички атоми в молекулата, участващи в поне един цикъл. За намиране на SSSR е използван алгоритъма, реализиран в CDK.

RC се нормира в проценти от 0 до 100, чрез следната формула:

$$(10) \quad S_{RC} = k_{RC} * RC + m_{RC}$$

където преобразуващите коефициенти са съответно: $k_{RC} = -100$, $m_{RC} = 200$. RC типично заема стойности между 1 и 1.5, които са нормирани съответно от 100 до 0 %.

По-високата стойност на комплексността на циклите (RC) предполага по-нисък принос за общата синтетична достъпност, което се вижда от отрицателната стойност на коефициента k_{RC} .

4.2. Цикломатично число

Цикломатичното число представлява минималният брой връзки, които трябва да се премахнат, за да може от цикличен граф да се получи ацикличен такъв. За всеки един граф цикломатичното число е по-малко или равно на броя на циклите в графа.

Цикломатичното число, като елемент във функцията за оценка на синтетичната достъпност, се пресмята по формулата:

$$(11) \quad \mu = m - n + 1$$

където m е броят на връзките, а n е броят на атомите.

Определеното по (11) цикломатично число се изразява като принос към синтетичната достъпност на съединението с помощта на следната формула:

$$(12) \quad S_{\mu} = k_{\mu} * \mu + m_{\mu}$$

с нормиращи коефициенти: $k_{\mu} = -10$, $m_{\mu} = 100$, чрез които се дефинира „наказателна“ функция даваща -10 % принос за всеки цикъл в структурата.

4.3. Стереохимична комплексност

За оценка на стереохимичната комплексност сме използвали подхода, описан от Cahn et al. [34]. За всеки атом в молекулата се проверява дали не представлява стереоцентър: откриват се въглеродните атоми, които не са свързани с двойни или тройни връзки, и се изследват топологичните им околности. Приносът на стереохимичната комплексност към синтетичната достъпност на дадено съединение се изчислява по следната формула:

$$(13) \quad S_{WSC} = k_{WSC} * N_{WSC} + m_{WSC}$$

където N_{WSC} е броят на претеглените хирални центрове в молекулата, като на хиралните атоми се дава тегло 1.0, докато стерео двойните връзки са с тегло 0.3. Приносът S_{WSC} се получава с помощта на коефициентите на линейното преобразование със стойности по подразбиране: $k_{WSC} = -30$, $m_{WSC} = 100$.

По-голям брой стереоцентрове съответства на по-нисък принос към синтетичната достъпност на молекулата, като коефициентите дефинират наказателна функция -30% за всеки хирален център.

4.4. Молекулна комплексност

За изчисляване на молекулната комплексност сме използвали метода, описан от Proudfoot [35]. Приносът на молекулната комплексност се изчислява със следната формула:

$$(14) \quad S_{CM} = k_{CM} * CM + m_{CM}$$

където коефициентите на линейното преобразование са със стойности по подразбиране: $k_{CM} = -100/150$, $m_{CM} = 100$, нормирайки молекулната комплексност 150 към 0% принос и съответно молекулна комплексност 0 към 100% принос.

4.5. Софтуерна реализация

Описаният метод за предсказване на синтетична достъпност е реализиран като конзолно приложение SyntheticAccessibilityCli (Ambit-SA). Ambit-SA е разработено в рамките на софтуерния модул ambit2-reactions, част от платформата за химична информатика Ambit [5], [13]. Ambit-SA е програма с отворен код под LGPL лиценз и може да се използва свободно за академични, регулаторни и комерсиални цели. Изпълним jar файл може да бъде изтеглен от следния уеб адрес: <http://ambit.sourceforge.net/reactor.html>

Ambit-SA може да се стартира от командния ред, като се зададе командата:

```
java -jar SyntheticAccessibilityCli.jar option1 option2 ...
```

Изчисляването на синтетичната достъпност за единична молекула може да се извърши чрез директно въвеждане на молекулата SMILES от командния ред с опцията "-s", например:

```
java -jar SyntheticAccessibilityCli.jar -s
"FC(F)(F)c1cc(ccc1)N5CCN(CCc2nnc3[C@H]4CCC[C@H]4Cn23)CC5"
Calculating SA for: FC(F)(F)c1cc(ccc1)N5CCN(CCc2nnc3[C@H]4CCC[C@H]4Cn23)CC5
SA = 54.116
```

Опцията '-v' може да се използва за генериране на подробни данни на изход:

```
java -jar SyntheticAccessibilityCli.jar -s
"FC(F)(F)c1cc(ccc1)N5CCN(CCc2nnc3[C@H]4CCC[C@H]4Cn23)CC5" -v
Calculating SA for: FC(F)(F)c1cc(ccc1)N5CCN(CCc2nnc3[C@H]4CCC[C@H]4Cn23)CC5
SA = 54.116
SA details:
MOL_COMPLEXITY_01 88.751 score = 40.833
WEIGHTED_NUMBER_OF_STEREO_ELEMENTS 2.000 score = 40.000
CYCLOMATIC_NUMBER 5.000 score = 50.000
RING_COMPLEXITY 1.174 score = 82.609
```

Опцията '-i' може да се използва за задаване на входен файл с набор от структури за последователно изчисление:

```
java -jar SyntheticAccessibilityCli.jar -i sa-mol-set-01-b.smi
Calculating SA for molecule set: sa-mol-set-01-b.smi
Reading D:\ChemSoft\JBSMM-Reactor\sa-mol-set-01-b.smi
#      smiles      NumAtoms      SA
1      CCOP(=S)(OCC)Oc1cc(C)nc(n1)N(C)C      19      85.622
2      OOC1CCOP(=O)(N1)N(CCC1)CCC1      16      81.760
3      O=C1Cc2c(N1)ccc3OCC(CNCc4ccccc4)Oc23      23      64.931
4      CC1OC(=NC1CCOc2ccc(CC3C(=O)NOC3=O)cc2)c4ccccc4      29      72.725
5      O=C(NN1CCCCC1)c2nn(-c3ccc(Cl)cc3Cl)c(-c4ccc(Cl)cc4)c2C      30
      72.241
6      CC(C)C(=O)Oc1ccc2CC(CCC2c1OC(=O)C(C)C)NC      24      66.674
```

4.6. Оценка на синтетичната достъпност – резултати

Основният проблем при разработването на методи за оценка на синтетична достъпност е валидирането на резултатите. В тази връзка, за да оценим ефективността на разработения от нас алгоритъм, сме използвали данни от литературата [36] за съединения със синтетична достъпност, определена от експерти.

За всяка молекула от тестовата ни извадка са изчислени приносите на отделните компоненти към крайната оценка на синтетичната достъпност, а именно: молекулна комплексност, стереохимична комплексност, цикломатично число и комплексност на циклите. Данните са обобщени в Таблица 1, която дава информация за коефициентите на корелация на Пиърсън (Pearson correlation coefficient) r между приносите от отделните компоненти, и стойностите за синтетична достъпност, оценена от експертите химици:

Таблица 1: Коефициенти на корелация за приносите на отделните компоненти

Принос	r
S_{CM}	0.867
S_{μ}	0.675
S_{RC}	0.565

Наблюдава се висока корелация между изчислената молекулна комплексност и данните, дадени от химиците. Това може да бъде обяснено с факта, че от една страна, моделът използва количествена оценка за информационното съдържание (молекулна комплексност) на база топологията на молекулата, и от друга страна, експертите правят същата оценка на база на знания и опит: така например по-големи и сложни молекули, съдържащи кондензирани или мостови системи, се оценяват като трудни за синтезиране.

Умерената корелация на приноса от цикломатичното число с данните на химиците е разбираема, предвид факта че играе ролята на коригиращ фактор на дескриптора, описващ приноса от комплексността на циклите. Освен това че голяма част от тестовата извадка е представена от съединения, които не съдържат кондензирани

цикли, които биха довели до получаване на по-високи стойности за цикломатичното число.

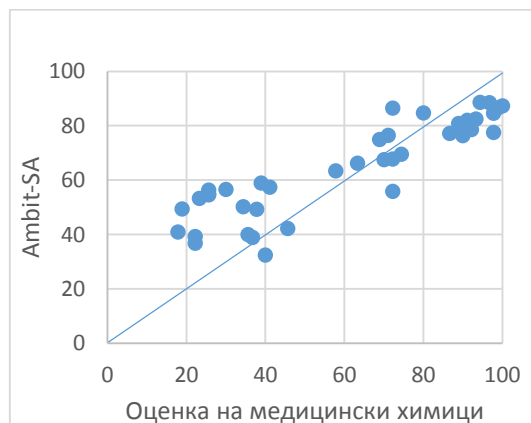
Коефициентът на корелация на Пиърсън е най-нисък между данните за синтетичната достъпност, дадени от химици, и приноса от комплексността на циклите. Въпреки това, този компонент е от важност за разработения модел за изчисляване на синтетичната достъпност.

За изчисляване на синтетичната достъпност е използвана формула (8), където приносите от отделните компоненти са умножени със съответните тегла (w_i). Оптималните тегла са определени като са сравнени получените коефициенти на корелация между нашия модел за SA и резултатите на експертите химици за десетки комбинации на теглата. Установените оптимални стойности за теглата са представени в Таблица 2:

Таблица 2: Оптимални тегла на отделните компоненти на модела за SA

w_i	Тегло
WMC	0.4
WSC	0.2
WCN	0.1
WRC	0.3

Резултатите за теоретично изчислената синтетична достъпност с нашия модел (Ambit-SA) са сравнени с резултатите, дадени от химици (Фигура 12); получен е коефициент на корелация (Pearson) 0.896.



Фигура 12: Сравнение между синтетичната достъпност, изчислена от Ambit (Ambit SA) и синтетичната достъпност, оценена от химици

Данните за синтетична достъпност, изчислени от Ambit-SA, са сравнени с други модели, използвани за предсказване на синтетична достъпност. Резултатите са дадени в Таблица 3:

Таблица 3: Коефициенти на корелация между Ambit-SA и други модели

Модел	r
SYLVIA	0.938
PE	0.890
FA4	0.634

Високата корелация между резултатите от Ambit-SA и данните, получени за тестовата извадка, изчислени от SYLVIA [37] може да бъде обяснена със сходния подход, използван в двата метода: комбинирана оценка на молекулната комплексност, броя на хиралните центрове и отчитане на комплексността на циклите. И в двата модела приносът на молекулната комплексност при изчисляване на синтетичната достъпност е с голямо тегло. За разлика от Ambit-SA обаче, в SYLVIA се използват още два компонента, които имат значителен принос в оценката на синтетичната достъпност: подобие с стартови материали и ретросинтезния анализ.

При сравнението между Ambit-SA и модел, базиран на фрагментиране и подобие със стартови материали (Ertl et al. [36]) също се наблюдава висока корелация. Предвид факта, че нашият модел не отчита такъв тип принос, високата корелация е показател, че Ambit-SA е достатъчно надежден и съдържа компоненти, базирани на структурна информация, значими за оценка на синтетичната достъпност.

За оценка на Ambit-SA е направено сравнение между получените от нас резултати и модел, базиран на стартовите материали (FAs4) [38]. За 130 съединения от литературата [39] корелацията, получена между двата метода, е 0,634.

IV. Заключение

1. Обобщение на получените резултати

1.1. Методите за прилагане на химически реакции посредством линейната нотация SMIRKS в софтуерния модул Ambit-SMIRKS са оптимизирани и модулът е използван за:

- прилагане на набор от химични трансформации, използвани от метода на Cramer за теоретично предсказване на токсичност на органични съединения;
- прилагане на набор от химични трансформации за стандартизация на химичните структури на органични съединения в големи бази данни

1.2. Създадени са три бази с данни с химични реакции, представени в компютърно-четим вид (SMIRKS):

- бази с данни с химични реакции (синтезно направление);
- база данни с биотрансформации (предсказване на метаболизъм);
- база данни с трансформи (ретросинтезно направление).

1.3. Създаден е конфигурируем виртуален химичен реактор, базиран на база данни с реакции, база данни със стартови материали и стратегия на реактора (правила), реализиран в софтуерната системата Ambit като отделен модул - Ambit-Reactor;

1.4. Създадена е стратегия и са конфигурирани правила за компютърно подпомогнат ретросинтез, реализирани в специализиран софтуерен модул за ретросинтезен анализ (JBSMM-Reactor);

1.5. Софтуерните модули, създадените стратегии и метаправила са тествани по отношение на ефективност и коректност на генерираната химична информация;

1.6. Създадена е стратегия и е реализиран модел за оценка на синтетичната достъпност на органични съединения;

1.7. Моделът за оценка на синтетичната достъпност е сравнен с други софтуерни системи и модели (Peter Ertl, SYLVIA, FA4)

2. Научни приноси

2.1. Създадени са три бази с данни с химични реакции (синтезно направление, ретросинтезно направление, биотрансформации), представени в компютърно-четим вид (SMIRKS);

2.2. Създадена е експертна система за ретросинтезен анализ – Ретросинтезен реактор, имплементиран в модула за симулация на химични реакции JBSMM-Reactor;

2.3. Реализиран е модел за теоретична оценка на синтетичната достъпност на органични съединения.

3. Научно-приложни приноси

Създаден е софтуерен модул Ambit-SMIRKS с отворен код, който поддържа пълния SMIRKS синтаксис, и може да се използва в научни и индустриални проекти, изискващи обработка на структурна и химична информация, свързана с химическите реакции. До момента е намерил приложение в няколко проекта: Toxtree, enviPath, ExCAPE, Biotransformer, Cramer 3.

На база разработените в дисертацията методи, алгоритми и стратегии са създадени софтуерни инструменти като конзолни приложения (command line tools - cli) и приложения с графичен интерфейс (GUI), които са достъпни на публичните хранилища sourceforge.net и github.com.

Създадените софтуерни приложения AmbitSmirksCli, AmbitReactorCli, MolecularComplexityCli, SyntheticAccessibilityCli, AmbitSMIRKSGUI и JBSMM-Reactor могат да се използват в проекти по химична информатика и компютърно моделиране на реакции от академични институции, частни компании и регулаторни агенции.

4. Научни съобщения по дисертацията

4.1. Публикации

4.1.1. Kochev, N.; Avramova, Sv.; Jeliaskova, N. “Combinatorial generation of molecules by virtual software reactor” *Sci. Work. Union Sci. Bulg. - Plovdiv*, vol. XV, pp. 214–219, 2017

4.1.2. Avramova, Sv.; Kochev, N.; Angelov, P. "RetroTransformDB: a dataset of generic transforms for retrosynthetic analysis" *Data*, 2018 (accepted for publication)

4.1.3. Kochev, N.; Avramova, Sv.; Angelov, P.; Jeliaskova, N. „Computational prediction of synthetic accessibility of organic molecules with Ambit-SyntheticAccessibility tool“ *Organic Chemistry: An Indian Journal*, 2018 (IF 0.36) (accepted for publication)

4.1.4. Kochev, N.; Avramova, Sv.; Jeliaskova, N. "Ambit-SMIRKS: a Software Module for Reaction Representation, Reaction Search and Structure Transformation" *Journal of Cheminformatics*, 2018 (submitted)

4.2. Постери

4.2.1. Nina Jeliaskova, Nikolay T. Kochev, Patrik Rydberg, Svetlana Avramova. Reaction Representation and Structure Transformation with Ambit-SMIRKS. Application in Metabolite Prediction. OpenTox Euro, Mon, 30. Sep. 2013 to Wed, 02. Oct. 2013, Mainz, Germany (best poster award)

4.2.2. Kochev N. T., Sv. L. Avramova, N. G. Jeliaskova. Reaction Representation and Structure Transformation with Ambit-SMIRKS. Application in Metabolite Prediction. 10th International Conference on Chemical Structures and 10th German Conference on Chemoinformatics June 1-5 2014, Noordwijkerhout, The Netherlands

4.2.3. Н. Кочев, Св. Аврамова, Н. Желязкова. Компютърно представяне на химични реакции и структурни трансформации с Ambit-SMIRKS. Приложения при предсказване на метаболити. II Научна конференция за студенти и докторанти „Предизвикателства в химията“, 21-22 ноември 2014 г., Пловдив, България

4.2.4. Jeliaskova N., N. Kochev, V. Jeliaskov, V. Paskaleva, S. Avramova, V. Chupakhin, J. F. Golib Dzib, Hongming Chen. Standardization of large structure databases with Ambit chemoinformatics platform. 10th Chemistry conference (10CC), 9-11 October 2016, Plovdiv

4.2.5. Kochev N., S. Avramova, N. Jeliaskova. Software simulation of chemical reactions with Ambit-Reactor. 10th Chemistry conference (10CC), October 9-11, 2016, Plovdiv

4.2.6. N. Kochev, S. Avramova, Georgi Dimitrov, Plamen Angelov, N. Jeliaskova. JBSMM Reactor: a Software Tool for Simulation of Chemical Reactions. IV Научна конференция за студенти, докторанти и млади учени „Предизвикателства в химията“ (Пловдив, България), 10-11 ноември 2017 г.

Библиография и информационни източници

- [1] “The Nobel Prize in Chemistry 2013 - Prize Announcement.” [Online]. Available: https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/announcement.html. [Accessed: 16-Apr-2018].
- [2] J. Gasteiger, *Handbook of chemoinformatics*. 2003.
- [3] N. Jeliaskova and N. Kochev, “AMBIT-SMARTS: Efficient Searching of Chemical Structures and Fragments,” *Mol. Inform.*, vol. 30, no. 8, pp. 707–720, 2011.
- [4] N. Jeliaskova and V. Jeliaskov, “AMBIT RESTful web services: an implementation of the OpenTox application programming interface,” *J. Cheminform.*, vol. 3, no. 1, p. 18, Jan. 2011.
- [5] Ideaconsult Ltd., “AMBIT.” [Online]. Available: <http://ambit.sourceforge.net/>.
- [6] J. Gasteiger and T. E. Eds, *Chemoinformatics*. 2003.
- [7] D. Weininger, “SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules,” *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.
- [8] “Daylight Theory: SMILES.” [Online]. Available: <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>. [Accessed: 10-Feb-2018].
- [9] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, *InChI, the IUPAC International Chemical Identifier*, vol. 7, no. 1. Journal of Cheminformatics, 2015.
- [10] G. Patlewicz, N. Jeliaskova, R. J. Safford, A. P. Worth, and B. Aleksiev, “An evaluation of the implementation of the Cramer classification scheme in the Toxtree software,” *SAR QSAR Environ. Res.*, vol. 19, no. 5–6, pp. 495–524, 2008.
- [11] Toxtree, “Toxic Hazard Estimation by decision tree approach.”
- [12] “The environmental contaminant biotransformation pathway resource.” [Online]. Available: <https://envipath.org/>. [Accessed: 20-Dec-2017].
- [13] N. Jeliaskova, N. Kochev, and V. Jeliaskov, “ambitcli-3.0.2,” 14-Apr-2016. [Online]. Available: <https://zenodo.org/record/173560#.WjlcRyvfvHVq>. [Accessed: 19-Dec-2017].
- [14] J. Sun, N. Jeliaskova, V. Chupakhin, J.-F. Golib-Dzib, O. Engkvist, L. Carlsson, J. Wegner, H. Ceulemans, I. Georgiev, V. Jeliaskov, N. Kochev, T. J. Ashby, and H. Chen, “ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics,” *J. Cheminform.*, vol. 9, no. 1, p. 17, Mar. 2017.
- [15] Y. D. Feunang, “Cheminformatics Tools for Enabling Metabolomics,” 2017.
- [16] N. Kochev, Nikolay; Avramova, Svetlana; Jeliaskova, “Combinatorial generation of molecules by virtual software reactor,” *Sci. Work. Union Sci. Bulg. - Plovdiv*, vol. XV, pp. 214–219, 2017.
- [17] G. Pirok, N. Maté, J. Varga, J. Szegezdi, M. Vargyas, S. Dórant, and F. Csizmadia, “Making ‘real’ molecules in virtual space,” *J. Chem. Inf. Model.*, vol. 46, no. 2, pp. 563–8, 2006.
- [18] M. J. de Groot, “Designing better drugs: predicting cytochrome P450 metabolism,” *Drug Discov. Today*, vol. 11, no. 13–14, pp. 601–606, 2006.
- [19] J. Gao, L. B. M. Ellis, and L. P. Wackett, “The University of Minnesota Biocatalysis/Biodegradation Database: Improving public access,” *Nucleic Acids Res.*, vol. 38, no. SUPPL.1, pp. 488–491, 2009.
- [20] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer, “Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited,” *J. Chem. Inf. Comput. Sci.*, vol. 32, no. 3, pp. 244–255, 1992.

- [21] “Chemical Markup Language | CML.” [Online]. Available: <http://www.xml-cml.org/>. [Accessed: 18-Dec-2017].
- [22] T. Sterling and J. J. Irwin, “ZINC 15 – Ligand Discovery for Everyone,” *J. Chem. Inf. Model.*, vol. 55, no. 11, pp. 2324–2337, 2015.
- [23] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. July 1928, pp. 379–423, 1948.
- [24] J. D. Angelo and M. B. Smith, *Hybrid Retrosynthesis*. Elsevier, 2015.
- [25] S. C. Schürer, P. Tyagi, and S. M. Muskal, “Prospective exploration of synthetically feasible, medicinally relevant chemical space,” *J. Chem. Inf. Model.*, vol. 45, no. 2, pp. 239–248, 2005.
- [26] M. Hartenfeller, H. Zettl, M. Walter, M. Rupp, F. Reisen, E. Proschak, S. Weggen, H. Stark, and G. Schneider, “Dogs: Reaction-driven de novo design of bioactive compounds,” *PLoS Comput. Biol.*, vol. 8, no. 2, 2012.
- [27] B. B. Masek, D. S. Baker, R. J. Dorfman, K. Dubrucq, V. C. Francis, S. Nagy, B. L. Richey, and F. Soltanshahi, “Multistep Reaction Based de Novo Drug Design: Generating Synthetically Feasible Design Ideas,” *J. Chem. Inf. Model.*, vol. 56, no. 4, pp. 605–620, 2016.
- [28] M. Hartenfeller, M. Eberle, P. Meier, C. Nieto-Oberhuber, K.-H. Altmann, G. Schneider, E. Jacoby, and S. Renner, “A Collection of Robust Organic Synthesis Reactions for *In Silico* Molecule Design,” *J. Chem. Inf. Model.*, vol. 51, no. 12, pp. 3093–3098, 2011.
- [29] S. D. Roughley and A. M. Jordan, “The medicinal chemist’s toolbox: An analysis of reactions used in the pursuit of drug candidates,” *J. Med. Chem.*, vol. 54, no. 10, pp. 3451–3479, 2011.
- [30] T. W. J. Cooper, I. B. Campbell, and S. J. F. MacDonald, “Factors determining the selection of organic reactions by medicinal chemists and the use of these reactions in arrays (Small Focused Libraries),” *Angew. Chemie - Int. Ed.*, vol. 49, no. 44, pp. 8082–8091, 2010.
- [31] E. J. Corey, *The logic of chemical synthesis*. 1989.
- [32] K. Boda, T. Seidel, and J. Gasteiger, “Structure and reaction based evaluation of synthetic accessibility,” *J. Comput. Aided. Mol. Des.*, vol. 21, no. 6, pp. 311–325, 2007.
- [33] J. Gasteiger and C. Jochum, “An Algorithm for the Perception of Synthetically Important Rings,” *J. Chem. Inf. Comput. Sci.*, vol. 19, no. 1, pp. 43–48, 1979.
- [34] R. S. Cahn, C. Ingold, and V. Prelog, “Specification of Molecular Chirality,” *Angew. Chemie Int. Ed. English*, vol. 5, no. 4, pp. 385–415, Apr. 1966.
- [35] J. R. Proudfoot, “Molecular Complexity and Retrosynthesis,” *J. Org. Chem.*, vol. 82, no. 13, pp. 6968–6971, 2017.
- [36] P. Ertl and A. Schuffenhauer, “Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions,” *J. Cheminform.*, vol. 1, no. 1, pp. 1–11, 2009.
- [37] Molecular Networks, “SYLVIA - Estimation of the Synthetic Accessibility of Organic Compounds.” [Online]. Available: <https://www.mn-am.com/products/sylvia>. [Accessed: 15-Jan-2018].
- [38] Y. Fukunishi, T. Kurosawa, Y. Mikami, and H. Nakamura, “Prediction of synthetic accessibility based on commercially available compound databases,” *J. Chem. Inf. Model.*, vol. 54, no. 12, pp. 3259–3267, 2014.
- [39] Q. Huang, L. L. Li, and S. Y. Yang, “RASA: A rapid retrosynthesis-based scoring method for the assessment of synthetic accessibility of drug-like molecules,” *J. Chem. Inf. Model.*, vol. 51, no. 10, pp. 2768–2777, 2011.