ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ“

ХИМИЧЕСКИ ФАКУЛТЕТ

КАТЕДРА „АНАЛИТИЧНА ХИМИЯ И КОМПЮТЪРНА ХИМИЯ“

ГЕРГАНА ИЛИЕВА ТАНЧЕВА

ПРИЛОЖЕНИЕ НА МЕТОДИТЕ НА ХИМИЧНАТА ИНФОРМАТИКА ПРИ МУЛТИКОМПОНЕНТНИ СУБСТАНЦИИ И НАНОМАТЕРИАЛИ

АВТОРЕФЕРАТ

НА ДИСЕРТАЦИОНЕН ТРУД

ЗА ПРИСЪЖДАНЕ НА ОБРАЗОВАТЕЛНА И НАУЧНА СТЕПЕН „ДОКТОР“

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

Научен ръководител: доц. д-р Николай Кочев

гр. Пловдив

2024г.

Дисертационният труд е обсъден и насочен за защита на заседание на катедрен съвет на Катедра аналитична химия и компютърна химия на Химически факултет при Пловдивски университет „Паисий Хилендарски“, проведено на 14.10.2024 г.

Дисертационният труд се състои от 196 страници и включва 65 фигури, 6 таблици и 4 приложения, оформени в 7 глави. Цитирани са 246 източника.

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

Научно жури:

Проф. д-р Иванка Милошева Цаковска - Българска академия на науките; Институт по биофизика и биомедицинско инжинерство, *Област на висше образование: 4. Природни науки, математика и информатика; Професионално направление: 4.3 Биологически науки (Фармакология)*

Проф. дхн Ирини Атанас Дойчинова-Цекова, - МУ - София, *Област на висше образование: 7. Здравеопазване и спорт; Професионално направление: 7.3. Фармация (Теоретична химия)*

Проф. дхн Иван Петков Бангов – пенсионер, гр. София, *Област на висше образование: 4. Природни науки, математика и информатика; Професионално направление: 4.2 Химически науки (Теоретична химия)*

Проф. д-р Веселин Петров Баев - ПУ „Паисий Хилендарски“, *Област на висше образование: 4. Природни науки, математика и информатика; Професионално направление: 4.3 Биологически науки (Молекулярна биология)*

Проф. дхн Васил Борисов Делчев - ПУ „Паисий Хилендарски“, *Област на висше образование: 4 Природни науки, математика и информатика; Професионално направление: 4.2 Химически науки (Теоретична химия)*

I. Въведение

Химичната информатика възниква като отговор на нуждата от въвеждане на компютърни и информационни технологии при цялостния цикъл за обработка на генерираните експериментални и/или симулирани данни, тяхното комбиниране с други информационни ресурси и последващото трансформиране на данните в полезна информация за моделиране и развойна дейност, а на следващ етап формализиране на информацията под формата на знание, с крайна цел решаване на практични научно-изследователски проблеми и иновации. Класическите три проблема на химичната информатика са общоизвестни: информационно и компютърно подпомагане при откриване на химични съединения с целеви свойства, откриване на методи за техния синтез и разкриване на структури на неизвестни съединения[1]. Химичната информатика се е развила като интердисциплинарна наука, която обхваща: проектиране, създаване, организация, управление, търсене, анализ, разпространение, визуализация и употреба на химична информация. Методите на химичната информатика са свързани с представяне и съхранение на химичните обекти в химични бази данни, обработка на данните, изчисляване на дескриптори и моделиране на физико-химични свойства и биологична активност. В основата на всички описани компютърни методи е представянето (формализирането) на химичното съединение посредством даден модел за данни.

Приложението на методите на химичната информатика за класическите химични обекти – молекули е добре изследвано в продължение на повече от 40 години и е показало своята полезност във времето. По отношение на експоненциално развиващото се производство на мултикомпонентни субстанции и наноматериали, се очертава нуждата от прилагане на методите на химичната информатика, но тяхната директна употреба е под въпрос. Това предизвикателство мотивира и целта на дисертационния труд, а именно: да се изследват възможностите за приложение на методите на химичната информатика за обработка и съхранение на информацията за мултикомпонентни субстанции, наноматериали и нови (advanced) материали и откриване на перспективи за ефективна обработка на информацията чрез семантичен FAIR модел за данни и приложението му при научни експерименти и моделиране с наноматериали и химични субстанции.

Изводи от литературния обзор

През последното десетилетие производството на наноматериали и химични субстанции расте с експоненциални темпове. Концепцията Safe and Sustainable by Design (SSbD) е обвързана с проектиране на функционални и безопасни химични субстанции и наноматериали още в ранните етапи на тяхното технологично развитие и се насърчава чрез регулаторни инициативи и множество научни проекти. Европейската комисия[2] насърчава, промишлеността, академичните среди и научно-изследователски центрове да гарантират, че методите, моделите и данните, произведени и използвани при прилагането на

европейската рамка, съответстват на ръководните принципи за лесно откриване, достъпност, оперативна съвместимост и повторна използваемост (FAIR). Също така. се насърчава и увеличаването на висококачествени FAIR данни. Насърчава се и разработването на нови методи, модели и инструменти за оценка на риска.

От направения литературен преглед е видно, че отправна точка на всички методи на химичната информатика е представянето на химичното съеинение в машинно четим формат. Съществува ясно дефиниран модел за представяне на молекулите с помощта на три компонента: химична структура, свойства и дескриптори. Този модел е показал своята полезност в академичните среди, но за целите на индустрията и регулаторните агенции е недостатъчен тъй като, в реалността не се работи с чисти вещества, а с многокомпонентни субстанции. Въпреки че различните регулаторни агенции към момента нямат консенсус за дефиниция на химична субстанция и за дефиниция на наноматериал, техните предложения са обединени около едно общо схващане че химичната субстанция не е изградена от един компонент. От тук може да се направи изводът, че е необходима нова парадигма за представяне на химични субстанции за осъществяване на адекватната обработка на данни от наноматериали, микро и нанопластмаси и advanced материали.

Директното приложение на класическите методи на химичната информатика за химични субстанции представлява предизвикателство поради необходимостта от нов подход за адекватно им представяне в съответствие на принципите FAIR. Именно тези предизвикателства и препоръки, описани от европейската комисия, мотивираха основната цел на дисертационния труд.

II. Цел и задачи

2.1 Цел:

Изследване на възможностите за приложение на методите на химичната информатика за обработка и съхранение на информацията за мултикомпонентни субстанции, наноматериали и нови (advanced) материали и откриване на перспективи за ефективна обработка на информацията чрез семантичен FAIR модел за данни и приложението му при научни експерименти и моделиране на свойства на наноматериали и химични субстанции.

2.2 Задачи:

1. Разучаване на съществуващи софтуерни системи, публикувани алгоритми и технологии за представяне на химични обекти и обработка на химична информация.
2. Разучаване на принципите FAIR и възможностите за тяхното прилагане върху данни за мултикомпонентни субстанции и наноматериали.
3. Разучаване на съществуващи семантични модели за представяне на мултикомпонентни субстанции и професионални научни формати за сериализиране на данни.
4. Избор на семантичен модел за представяне (формализиране) на информацията (данни и мета данни) за химични субстанции и наноматериали.

5. Разучаване на съществуващи онтологии за химични субстанции и наноматериали.
6. Адаптиране и прилагане на алгоритмите за обработка и моделиране в химичната информатика за химични субстанции и наноматериали.
7. Създаване на алгоритми за FAIR-фикация на не-FAIR експериментални данни.
8. Приложение на алгоритмите за FAIR-фикация върху данни за наноматериали и химични субстанции от предходни и настоящи научни проекти на Европейската комисия.
9. Създаване на алгоритми и електронни тефтери за сериализация на информация от базовия семантичен модел за данни.
10. Създаване на алгоритми и електронни тефтери за оценка на пълнотата, качеството и степента на FAIR-фикация на химичните обекти представени чрез избрания семантичен модел.
11. Тестване на възможностите на избрания семантичния модел за приложение върху данни от високо производителен скрининг от биологични експерименти (HTS).
12. Създаване на подобрени и автоматизирани методологии за обработка на HTS данни и разработка на методология за интегриране на FAIR принципите за HTS данни.
13. Разработка на прототипи на уникални идентификатори за мултикомпонентни субстанции и наноматериали.
14. Тестване на софтуерни аналитични платформи по отношение на ефективност и бързодействие за предварителна обработка на данни и моделиране.
15. Тестване на софтуерните модули, създадените стратегии и правила по отношение на ефективност, бързодействие и коректност на генерираната химична информация и създадените модели.
16. Сравняване на разработените от нас подходи и алгоритми с други софтуерни системи.

III. Собствени изследвания

3. FAIR принципи при управление на данни.

Ефективното управление и агрегиране на експериментални данни от различни източници е една от основните цели на автоматизираната обработка на данни. Базово изискване за осъществяването на тази цел е данните от оригиналните изпитвания на химичните субстанции да са съчетани с богати метаданни. Един от стълбовете на научния метод е възможността за независимо потвърждение и повтерение на получените резултати и именно тук са необходими метаданни, които описват всички възможни аспекти на проведените експерименти.

През 2016 г. са публикувани четири основни принципа за управление на научни данни, демонстрирани на фигура 1, според които данните трябва да са: откриваеми, достъпни, оперативно съвместими и преизползваеми[3] (FAIR). Тези ръководни принципи насочват изследователите и институциите, които генерират данни как да увеличат ползите от техните данни. Следването на FAIR принципите е важно и в контекста на прилаганите алгоритми,

инструменти и работни процеси, използвани за самото генериране на данните. Инициативата GO-FAIR[4] и европейската комисия[2] в изпълнение на концепцията Safe and Sustainable by Design насърчават приложението на FAIR принципите.

Фигура 1: *Ръководни принципи FAIR*

Първата стъпка при повторното използване на данни е те да бъдат откриваеми. Машинночетимите данни и метаданни са от съществено значение за автоматичното откриване на набори от данни и онлайн услуги. За да бъде реализиран всеки от принципите, трябва да са изпълнени някои изисквания към данните.

FAIR принципите позволяват експерименталните данни да се използват извън техния произход, за решаване на научни проблеми, запълване на липсващи данни, повторно използване на данните в приложения, моделиране и предоставяне на инструменти за други нужди на науката, промишлеността и регулаторите. Принципите наблягат на възможността за машинна обработка (т.е. способността на изчислителните системи да намират, осъществяват достъп, взаимодействат и използват повторно данни без никаква или минимална човешка намеса), тъй като хората все повече разчитат на изчислителна поддръжка за обработка на данни в резултат на увеличаването на обема, сложността и скоростта на генериране на данни.

GO-FAIR препоръчва работен процес от седем основни етапа за трансформиране на не-FAIR данни в FAIR, показан на фигура 2.

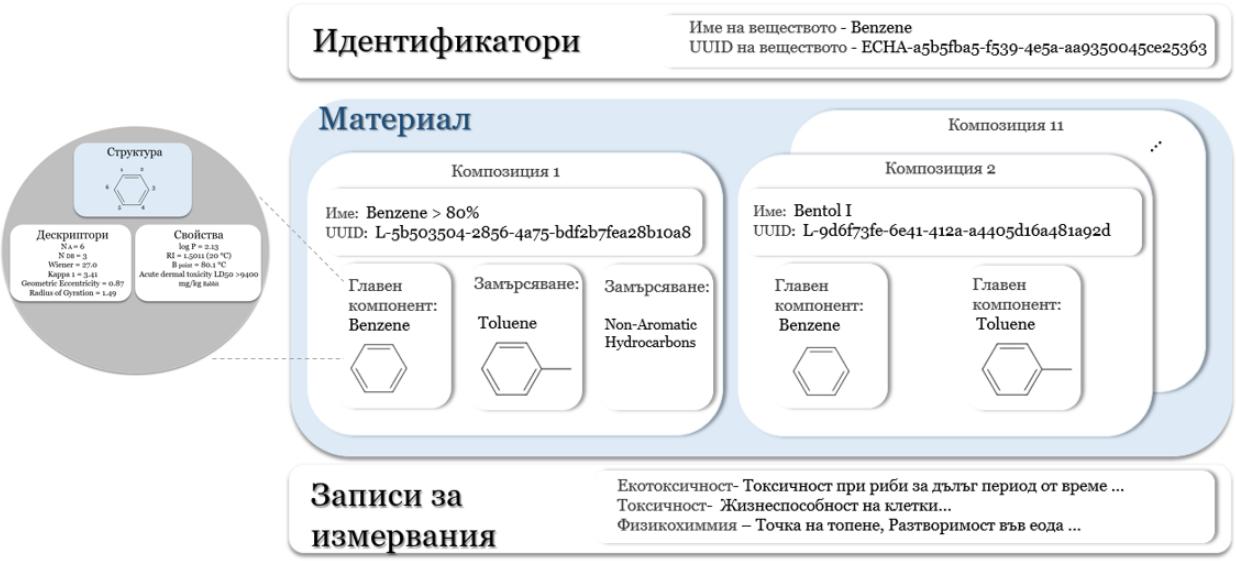
***Фигура 2**: Работен процес за FAIR-ификация на не- FAIR данни*

Семантичният модел за данни включва обекти, представящи аспекти от реалността и техните взаимоотношения. Изследователите рутинно описват експерименталните обекти в научната литература като материали, методи и резултати. Компютърното представяне на експерименталната система (т.е. материалите и методите) изисква дефиниране на елементи от данни, техните връзки и ограниченията между тях. В компютърните науки това е известно като **модел за данни** и служи като план за това как данните ще бъдат съхранявани, достъпвани и манипулирани. Моделът на данни представлява принципна (и абстрактна) логика и е нещо различно от използвания формат за описание на данните, тъй като един и същ модел за данни може да се съхранява в различни формати.

Дефинирането на семантичен модел за данни за представяне на химични обекти е най-важната стъпка от работния процес за FAIR-ификация. В този смисъл усилията за изясняване на модела за данни за химичните субстанции са също и усилия за осъществяването на принципите FAIR. Други етапи от първостепенно значение за FAIR-ификация на данни са включването на богати метаданни (стъпка 6), онтологични анотации и свързване на данни с глобални уникални идентификатори (стъпка 4).

4. Модел за данни за представяне и обработка на химични субстанции и наноматериали

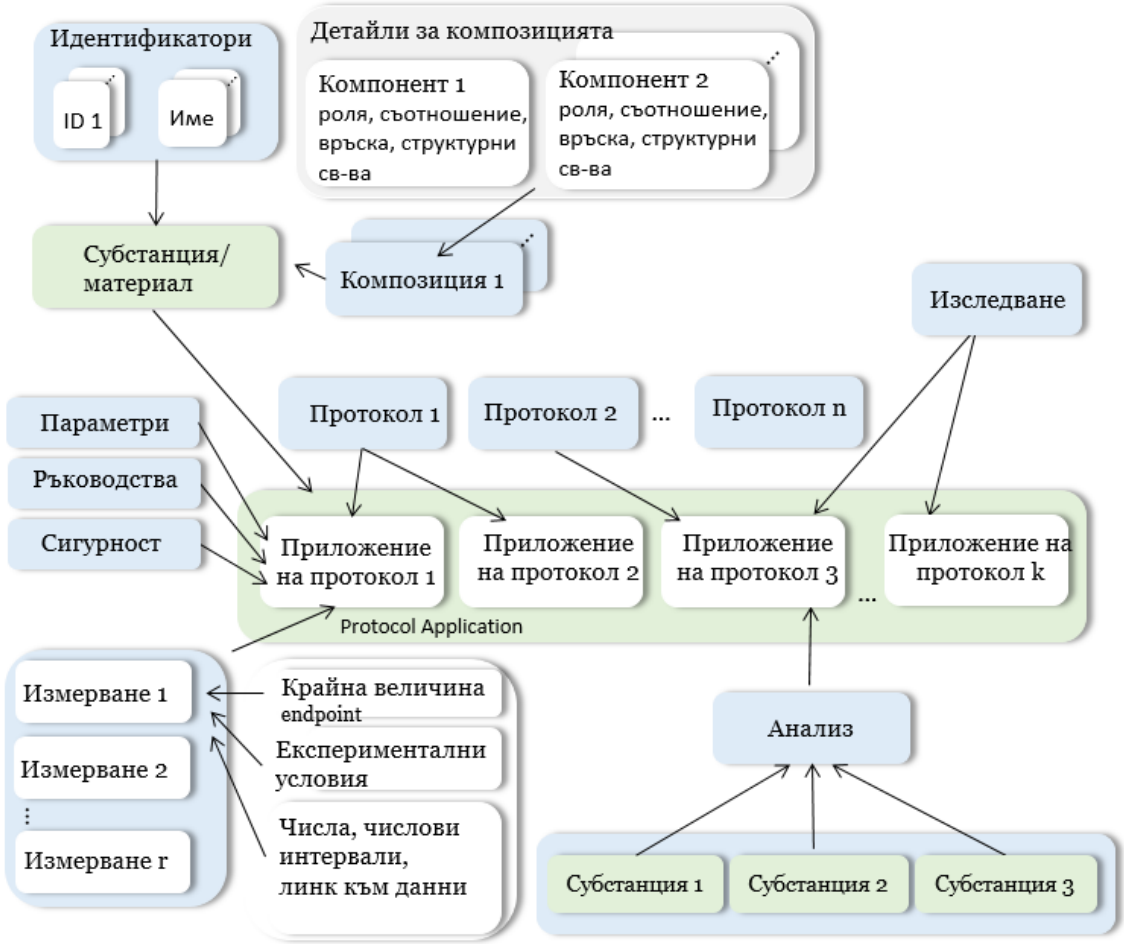
В настоящата дисертация бяха проучени различни подходи за описание на информацията за мултикомпонентни субстанции и наноматериали. Като резултат са оформени три основни слоя с метаданни за FAIR описание на мултикомпонентни субстанции, визуализирани на фигура 3 като: (i) идентификация на веществото, (ii) описание на композицията и (iii) записи за измервания с богато множество метаданни.



Фигура 3: *Химична субстанция "бензен" с няколко различни състава и информация, групирани в три слоя: идентификатори, материал, записи за измервания (примерът е взет от публичните записи на досиетата на ECHA и също е достъпен чрез уеб интерфейса на базата данни Ambit-LRI).*

Конкретният пример на фигура 3 демонстрира, че за химичната субстанция бензен, има 11 регистрирани композиции, като две от тях са: бензен с чистота над 80 % и бентол, където съединението бензен участва като главен компонент на субстанцията. Трябва да подчертаем отново, че в терминологията на ECHA, както и в настоящата дисертяция, се прави значителна разлика между понятията **химично съединение** и **химична субстация.**

Моделът за данни Ambit/eNanoMapper (фигура 4) е концептуално представяне на химичните субстанции и може да бъде реализиран с различни технологии, осигурявайки оперативна съвместимост и свързване на данни. Той съдържа различни компоненти от данни или обекти, изпълняващи определени роли за представяне на елементи от информацията за химични субстанции и измервания. Обектите могат да имат различни реализации в различните етапи от работния поток за обработка на данни: JSON, RDF или HDF5 формати, Java и Python класове, SQL таблици.



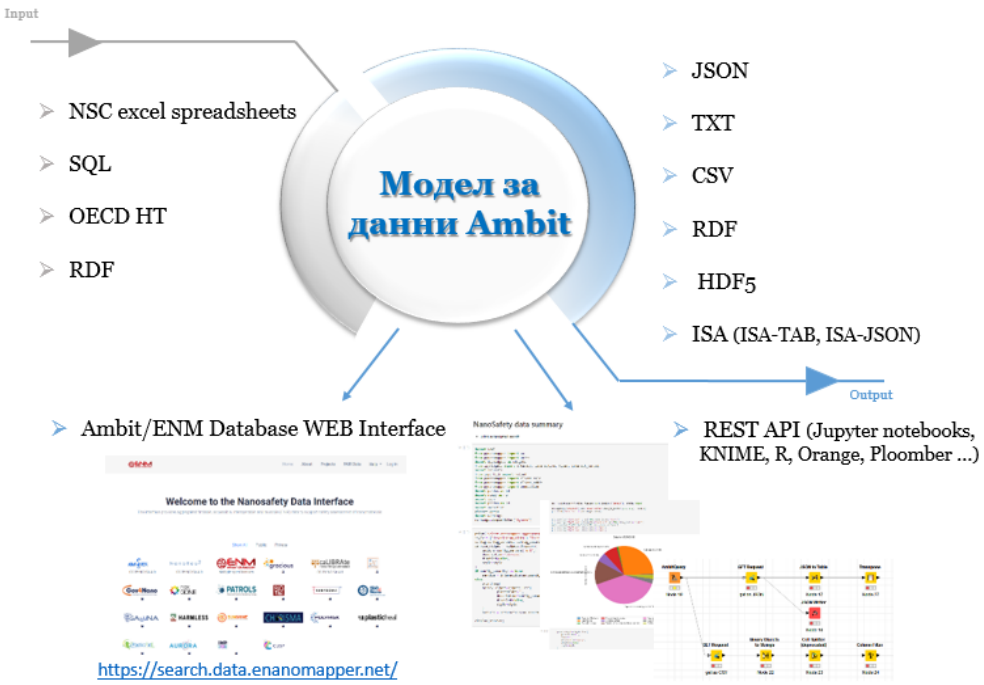
Фигура 4: *Схема на семантичния модел за данни Ambit/eNanoMapper*

В модела за данни, субстанциите се характеризират със своя състав за всеки компонент и се идентифицират с имена и идентификатори. Моделът поддържа множество композиции с един или повече компонента, всеки с определена роля като: главна съставка, добавка, замърсяване, ядро, покритие на наноматериал и т.н. Всеки компонент се представя чрез класическия модел за химично съединение. Резултатите от физикохимичните и биологичните измервания се третират като свойства на цялата химична субстанция и се управляват информационно чрез обекти наречени „приложения на протокол“. Ефективното описание на проведения експеримент в протокола е от решаващо значение за правилната комуникация на научните резултати и за създаване на FAIR ресурси от данни. Последното се осъществява с помощта на богат набор от параметри (метаданни) с гъвкава логическа организация. Всяко приложение на протокол се състои от набор измервания за определена крайна величина при дадени експериментални условия. Резултатът от измерването може да се представи като числова стойност, интервал, текст или връзка към файл с данни (напр. инфрачервен спектър, микроскопско изображение, HTS

данни и др.). Има гъвкавост при съхранението на параметри от метаданните. Всяко измерване е свързано с динамичен списък от експериментални условия (като концентрация, време, реплика и др.), считани за параметри от „по-ниско“ ниво. Метаданните от „високо“ ниво включват параметри, ръководства и връзки към стандартни процедури, публикации и др. Данните за едно вещество могат да съдържат множество „приложения на протоколи“.

5. Платформа за FAIR-ификация на данни за химични обекти – JSON конфигурации, хармонизирани шаблони, онтология

Семантичният модел за данни позволява интегрирането на данни от различни източници, като хармонизираните шаблони на OECD, персонализирани шаблони за електронни таблици, SQL изходни данни и др. След като данните са импортирани в базата данни на eNanoMapper, са налични различни опции за извличане, анализ на данните и конвертиране в други формати (фигура 5). Моделът за данни е гъвкав и позволява различни персонализирани методи за достъп до данните чрез REST API интерфейс и python библиотеката pynanomapper, с помощта на външни инструменти, платформи за машинно обучение и анализ на данни (KNIME, Orange), работни потоци в платформата Ploomber и др.



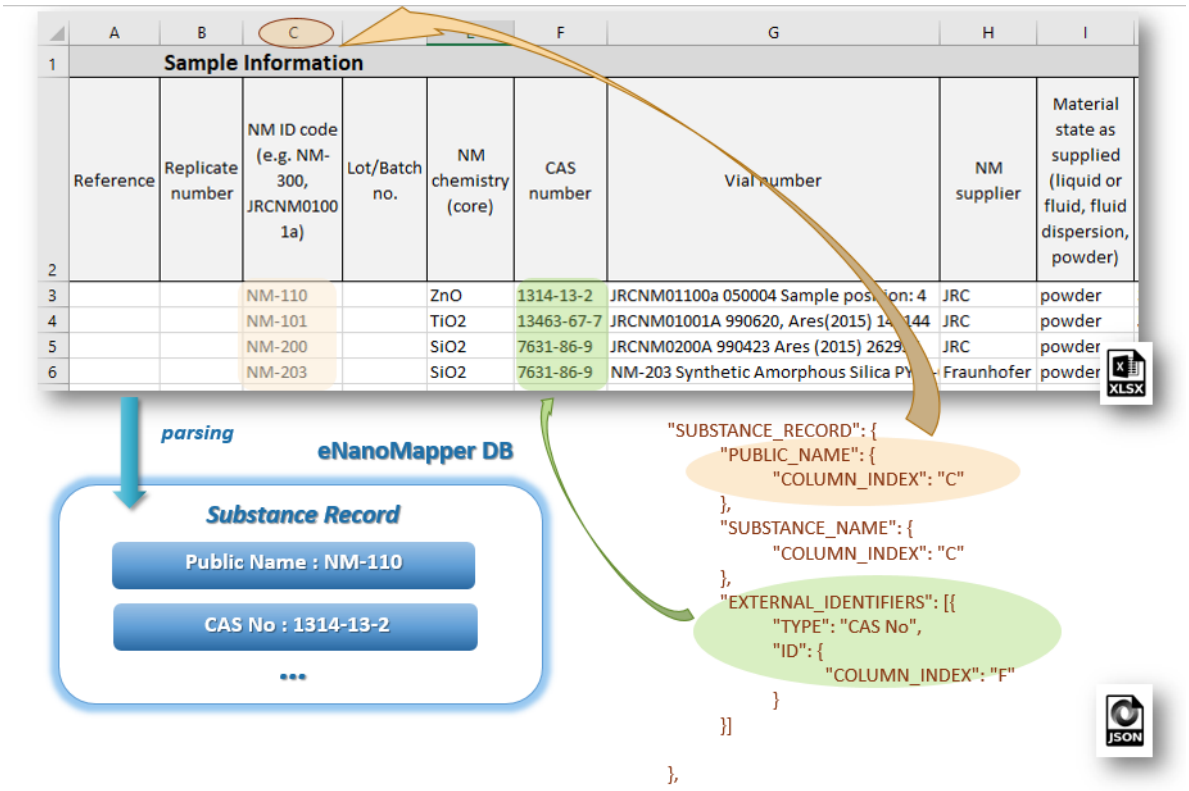
***Фигура 5**: Интегриране и експорт на данни в/от базата данни на eNanoMapper*

Най-предпочитаният начин от изследователите за въвеждане на данни е чрез шаблони за електронни таблици, които в общия случай представляват не-FAIR ресурси за данни и

възниква нуждата от процес на FAIR-фикация на тези EXCEL файлове. За целта е разработен конфигурируем инструмент NMDataParser[5], който улеснява подготовката на данни и качването им в базата данни eNanoMapper, базиран на модела за данни на Ambit/eNanoMapper. Поддържат се различни форми на организация на данните в таблиците – по редове, по колони или по блокове. Поради голямото разнообразие от EXCEL таблични файлове е необходимо конфигуриране на конвертирането чрез отделен JSON файл. Инструментът NMDataParser работи с два входни файла: електронна таблица (*.xlsx файл) и JSON конфигурационен файл, а като резултат се връща итератор към списък със записи за химични субстанции.

5.1 Конфигурационни JSON файлове за импортиране на данни в eNanoMapper

Конфигурационният JSON файл свързва отделните елементи от електронната таблица с компонентите на семантичния модел за данни и анотира данните със специфична за домейна онтология. Синтаксисът на JSON конфигурационния файл за NMDataParser включва набор от ключови думи за прочитане на данните от електронните таблици и тяхното конвертиране в модела за данни на Ambit/eNanoMapper. Ключовите думи определят различни стратегии за прочитане на данните, от един или множество Excel листове, както и позволяват различни комбинации на данни от различни листове, редове, колони, блокове от колони и редове. JSON конфигурационният файл се състои от няколко основни секции, които представляват обекти на първото ниво в JSON схемата.
Фигура 6 илюстрира конфигуриране на JSON файл за прочитане на името на наноматериал и CAS номера като външен идентификатор от EXCEL файл, и конвертиране на данните в модела на Ambit/eNanoMapper.



Фигура 6: *Отнасяне на данни от EXCEL в режим на итерация "ROW_SINGLE" и итериране на SUBSTANCE_RECORD.*

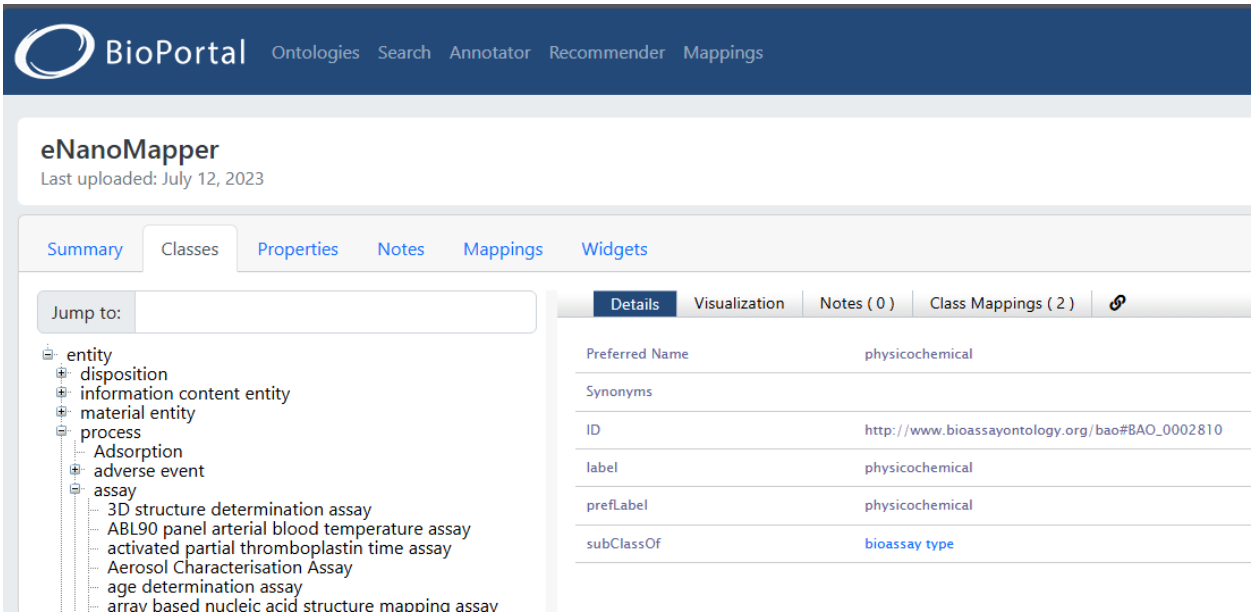
Описаният процес за FAIR-ификация приложихме за 1400 EXCEL файла с експериментални данни. Конвертираните данни съдържаха информация от няколко европейски проекта: NanoTest, NanoReg, NanoReg2, MARINA, ENPRA. Добавените данни в базата данни на eNanoMapper включват голямо разнообразие от физикохимични и биологични анализи като: клетъчна жизнеспособност, оксидативен стрес, имунотоксичност, in vivo/in vitro токсичност, екотоксичност, физикохимично охарактеризиране и други, за множество наноматериали като: въглеродни нанотръби, сребърни наночастици, цинкови и титаниеви диоксиди, железни и цериеви оксиди, такива с различно покритие на ядрото и без покритие.

5.3 Онтология eNanoMapper

Проектът eNanoMapper[6] създава общоевропейска изчислителна инфраструктура за управление на данни за наноматериали, базирана на семантични уеб стандарти и онтологии. По проект eNanoMapper се разработва цялостна онтология[7] и анотирана база данни за безопастността на наноматериалите, които са важна стъпка за справяне с предизвикателството за унифицираната анотация на наноматериали и техните съответни биологични свойства, експериментални модели системи, условия, протоколи и данни за

тяхното въздействие върху околната среда[8]. Описанието на информацията, касаеща безопасността на наноматериалите, включва разнообразни подобласти като биологични анализи, определяне на физикохимични и екологични характеристики и в тази връзка разработването на онтология „от нулата“ би било доста трудоемко и времеемко[9].

Първоначално направихме преглед на термините от онтологията MESOCOSM и ги сравнихме със съществуващи онтологии, като главната цел бе да се изберат, тези с най-близка дефиниция до целевия домейн на онтологията MESOCOSM. Претърсването направихме през уеб-сайта на BioPortal - най-голямото хранилище на биомедицински и природо-математични онтологии. Втората стъпка към интегрирането на онтологията MESOCOSM, бе добавянето на новите термини към онтологията на eNanoMapper. Добавянето на термини към онтологията на eNanoMapper се осъществи посредством библиотеката „slimmer“ и специализирани конфигурационни файлове (.props и .iris). Фигура 7 демонстрира добавения термин „physicochemical“ към онтологията на eNanoMapper, достъпна през платформата BioPortal от линка: https://bioportal.bioontology.org/ontologies/ENM?p=classes&conceptid=http%3A%2F%2Fwww.bioassayontology.org%2Fbao%23BAO_0002810



Фигура 7: *Екран от онтологията eNanoMapper през BioPortal с добавен термин physicochemical.*

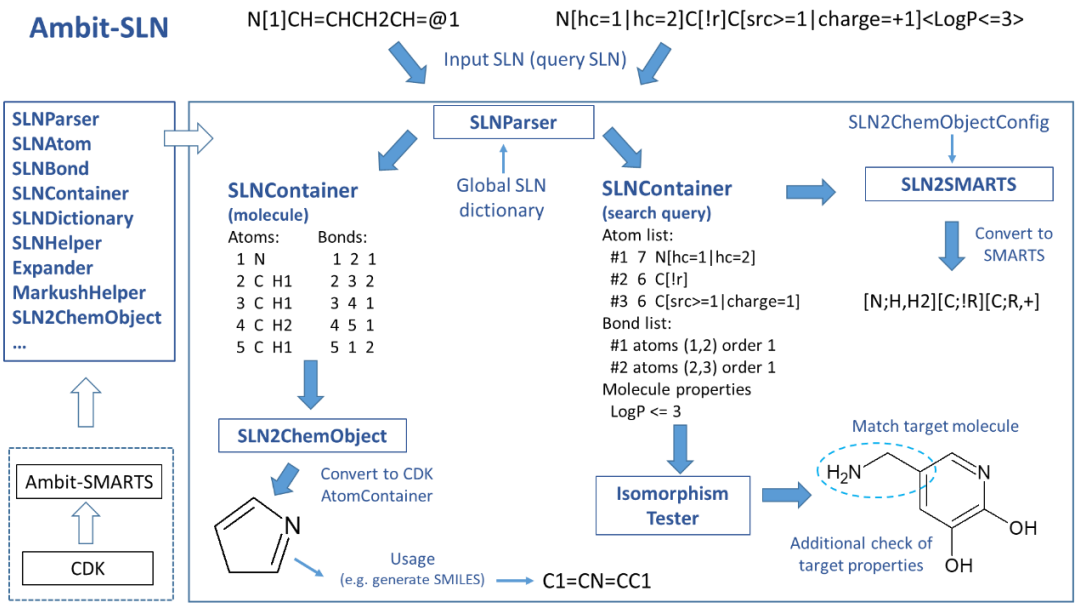
Качването на конфигурационни файлове в Git хранилището на онтологията, не означава, че терминът се добавя автоматично към онтологията на eNanoMapper. Предложените термини са обект на разглеждане, обсъждане и консенсус от екип експерти, които работят по подобряването на онтологията и за всеки нов термин това може да отнеме определено време, дори значително време, ако няма консенсус по дефинията на термина.

6 Представяне (сериализация) на мултикомпонентни субстанции чрез линейни нотации

Разработването на идентификатори за мултикомпонентни субстанции представлява голямо предизвикателство, тъй като подходите, прилагани при традиционните идентификатори за химични съединения, както и най-популярните линейни нотации SMILES и InChI не са директно приложими за описание на мултикомпонентни субстанции. Един от ключовите FAIR принципи е използването на глобално уникални и устойчиви идентификатори.

SYBYL Line Notation[10,11] (SLN) е недвусмислена, неуникална линейна нотация, която поддържа синтаксис за спецификация на молекули, заявки за субструктурно търсене и реакции, които покриват възможностите на стандатните нотации SMILES, SMARTS и SMIRKS. В допълнение, синтаксисът на SLN включва и други ефективни средства за спецификация на дефинирани от потребителя атрибути на атоми, връзки, структури и реакции, макро и Markush атоми за гъвкаво дефиниране на молекулни фрагменти, заявки за търсене и представяне на структурни библиотеки, както и описание на 2D и 3D атомни координати.

Разработена е софтуерна библиотека с отворен код Ambit-SLN за обработка на информация за химични обекти чрез линейната нотация SLN. Ambit-SLN е отделен софтуерен модул в платформата AMBIT. Библиотеката Ambit-SLN включва няколко базови функционалности за вътрешно представяне на информацията от SLN, парсър за пълния SLN синтаксис на заявки за субструктурно търсене, поддръжка за макро и Markush атоми, глобални и локални речници, и дефинирани от потребителя свойства, чрез които могат да се съхраняват и използват основните елементи от модела за данни на Ambit/eNanoMapper. Фигура 8 илюстрира основния работен процес на библиотеката Ambit-SLN и два основни случая на използване на линейната нотация SLN.



***Фигура 8**: Основен работен процес на Ambit-SLN, приложен за стандартна таблица на свързаност и заявка за субструктурно търсене.*

Ambit-SLN покрива голяма част от синтаксиса на SLN. Тествахме възможностите на библиотеката Ambit-SLN за описание на субстанции и наноматериали. Прототипният Ambit-SLN идентификатор за материал Fe_3O_4 с d=38nm и глициново покритие от 2 nm е използван за демонстрация на Nano-SLN прототип:

O[1]Fe[2]OFeOFe@1O@2 <role=core;size=38nm>
CH2(C(=O)OH)NH2 <role=coating;size=2nm>

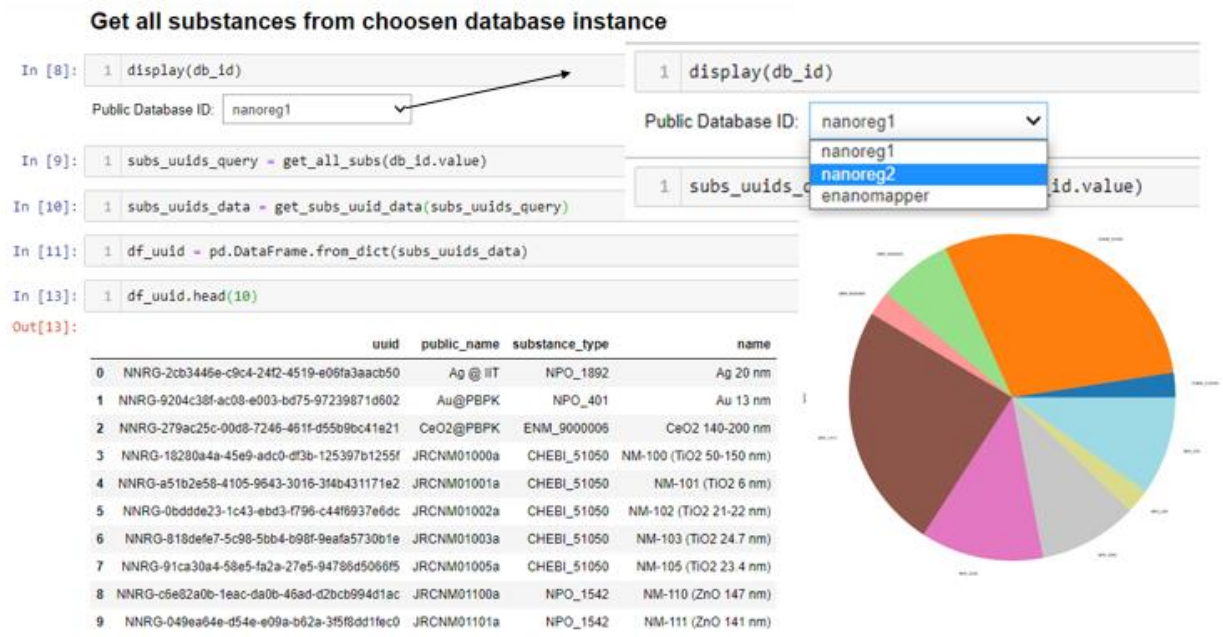
7 Обработка на данни

Експерименталните данни са в основата на работните потоци за обработка на информация при оценка на риска от употребата на наноматериали. Голяма част от времето и усилията при разработка на модели за физикохимични свойства и биологична активност, се изразходват за анализ, събиране, филтриране и подготовка на подходяща извадка от данни. Ефективното извличане на данни позволява на учените да формулират правилни хипотези и да създават по-добри модели.

7.1 Електронен бележник за извличане, обработка и визуализация на данни за нанобезопастност, от базата данни на eNanoMapper чрез REST API заявки.

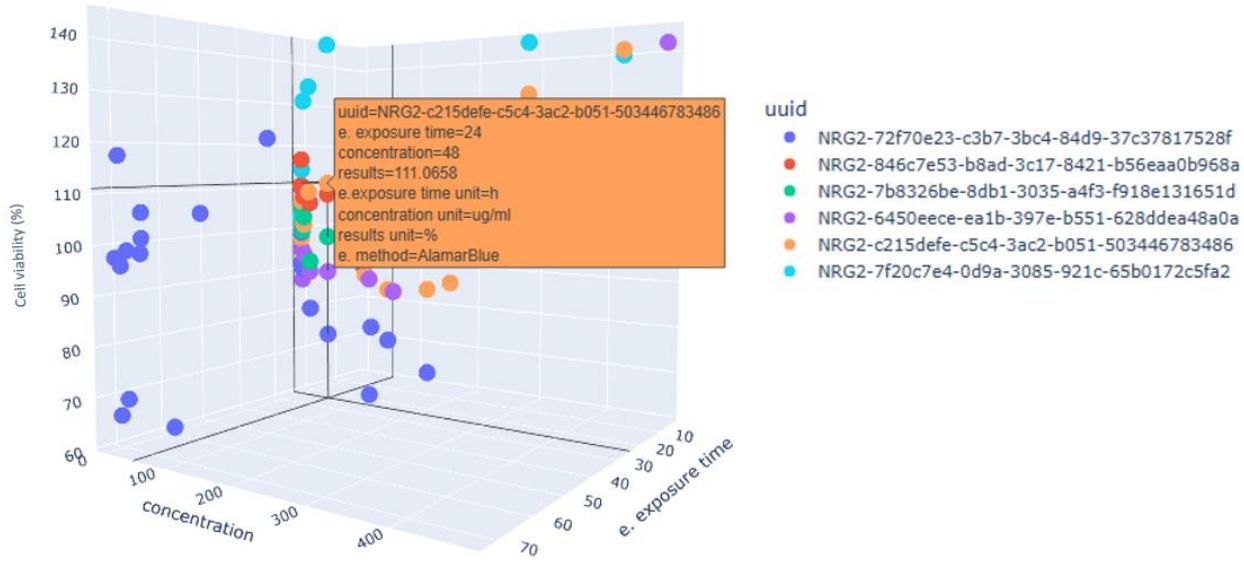
Електронният бележник дава възможност за динамична селекция на екземпляри от базата данни eNanoMapper (екземпляр: обособено подмножество от химични субстанции, генерирани в рамките на даден научен проект) и таблична визуализация на метаданни за

всички химични субстанции с техните уникални идентификатори (uuid), онтологични (substance type) и тривиални (pyblic name) наименования. Фигура 9 показва избор на конкретен екземпляр - Nanoreg2, от базата данни, заедно с таблица на всички налични субстанции и кръгова диаграма от тип „пай“ с наличността на данните, агрегирани по критерий „substance type“. Субстанциите са филтрирани според „substance type“ CHEBI 51050[12] – титаниев диоксид. Фигура 10 показва последваща филтрация по избрана измервана величина „жизнеспособност на клетките“ с налични експериментални данни за 6 наноформи на филтрираната субстанция – титаниев диоксид.



	uuid	public_name	substance_type	name
0	NNRG-2cb3446e-c9c4-24f2-4519-e06fa3aacb50	Ag @ IIT	NPO_1892	Ag 20 nm
1	NNRG-9204c38f-ac08-e003-bd75-97239871d602	Au@PBPK	NPO_401	Au 13 nm
2	NNRG-279ac25c-00d8-7246-461f-d55b9bc41e21	CeO2@PBPK	ENM_9000006	CeO2 140-200 nm
3	NNRG-18280a4a-45e9-adc0-df3b-125397b1255f	JRCNM01000a	CHEBI_51050	NM-100 (TiO2 50-150 nm)
4	NNRG-a51b2e58-4105-9643-3016-3f4b431171e2	JRCNM01001a	CHEBI_51050	NM-101 (TiO2 6 nm)
5	NNRG-0bddde23-1c43-ebd3-f796-c44f6937e6dc	JRCNM01002a	CHEBI_51050	NM-102 (TiO2 21-22 nm)
6	NNRG-818defe7-5c98-5bb4-b98f-9eafa5730b1e	JRCNM01003a	CHEBI_51050	NM-103 (TiO2 24.7 nm)
7	NNRG-91ca30a4-58e5-fa2a-27e5-94786d5066f5	JRCNM01005a	CHEBI_51050	NM-105 (TiO2 23.4 nm)
8	NNRG-c6e82a0b-1eac-da0b-46ad-d2bcb994d1ac	JRCNM01100a	NPO_1542	NM-110 (ZnO 147 nm)
9	NNRG-049ea64e-d54e-e09a-b62a-3f5f8dd1fec0	JRCNM01101a	NPO_1542	NM-111 (ZnO 141 nm)

Фигура 9: *Персонализирано филтриране на данни според наличните екземпляри в базата данни и обобщена визуализация на наличните субстанции в избрания екземпляр.*



Фигура 10: *3D интерактивна графика, показваща разпределението на субстанциите в пространството (доза, ефект, време за третиране) с цветово разделение на материалите и допълнителна информация за всеки от тях.*

7.2 Електронен бележник за проверка на пълнотата на данните

Разработихме електронен бележник за оценяване пълнотата на данните в рамките на използваната терминология в хармонизираните шаблони и наличната информация в базата данни на eNanoMapper. Термините, използвани за описание на данните от избран екземпляр на базата данни, се сравняват с тези от хармонизираните шаблони, като се оценява тяхното подобие с изчисляване на разстоянието на Левенщайн[13], което е минималният брой редакции на символи (вмъквания, изтривания или заместявания), необходими за промяна на една дума в друга. В таблица 1 е демонстриран конкретен пример от работата на бележника.

Таблица 1: *Сравняване на част от термини, използвани за описание на данни за крайна измерена величина "специфична повърхност по BET".*

Термини в базата данни	Термини в шаблон	Резултат от сравнението
MATERIAL_STATE,	MATERIAL_TATE,	Сгрешен термин в шаблона
INSTRUMENT_TYPE, OUTGASSING_TIME, OUTGASSING_TEMP,	INSTRUMENT_TYPE, OUTGASSING_TIME, OUTGASSING_TEMP,	Напълно съвпадащи термини
END_RELATIVE_PRESSURE START_RELATIVE_PRESSURE	-	Липсващи термини в шаблона

-	SAMPLE_DENSITY	Липсващи термини в базата данни

8 Обработка на данни от високо производителен скрининг от биологични експерименти и последващо моделиране

Прилагането на FAIR принципите в контекста на HTS данните представлява предизвикателство заради необходимостта от автоматично свързване на голям обем експериментални данни с описателни метаданни, хармонизиране на използваната терминология и преобразуване в машинно четим формат. Традиционните подходи за съхранение и документиране на резултатите от HTS, като използването на електронни таблици, отнемат време и предразполагат към технически грешки. Интегрирането на външен софтуер, като Toxicological Prioritization Index (ToxPi)[14,15], в работния процес въвежда допълнителна сложност, особено поради необходимостта от прехвърляне на значителни набори от данни от един софтуер към друг. Въпреки че ToxPi служи като ценен софтуерен инструмент за групиране на материалите и визуализация, неговите възможности са ограничени от липсата на функции за предварителна обработка на данни и ограничените опции за представяне и съхранение на получените резултати.

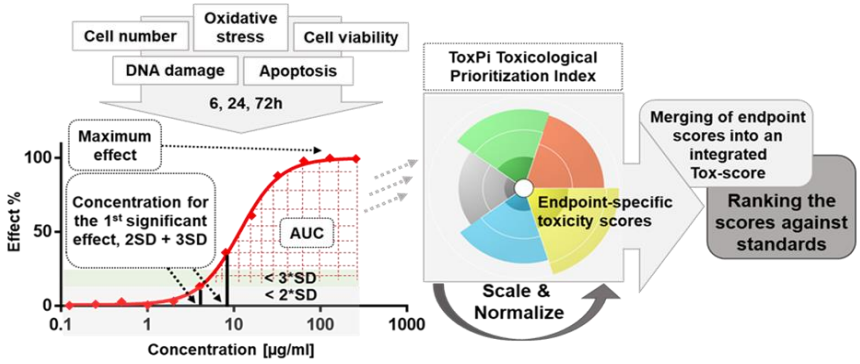
8.1 Клетъчно базиран високопроизводителен скрининг и методология Tox5-Score за групиране на in- vitro данни за токсичност на химични субстанции и наноматериали.

Методологията за високопроизводителен скрининг, разработена в отдела по токсикология към лабораторията на Misvik Biology, позволява бърза оценка на токсичността на множество материали, използвайки набор от пет добре установени анализа за токсичност:

- CellTiter-Glo ® за определяне на клетъчна жизнеспособност чрез луминисцентно измерване на четец за плаки;
- Dapi за определяне на брой живи клетки чрез флуоресцентна микроскопия;
- gammaH2AX за установяване на увреждане на ДНК чрез флуоресцентна микроскопия;
- 8OHG за отчитане на оксидативен стрес на нуклеинова киселина чрез флуоресцентна микроскопия;
- Caspase -Glo 3 за определяне на апоптоза чрез флуоресцентна микроскопия и специфичен анализ Caspase -Glo ® 3/7, но чрез луминесцентно измерване на четец за плаки.

Анализите се осъществяват с вариране на няколко времеви точки, различни концентрационни нива и различни клетъчни линии, адаптирани от описаните методологии[16,17]. Традиционното изпитване за токсичност се основава на определяне на концентрацията на дадено вещество, при която растежът на клетките е инхибиран с 50%

(GI_{50}), въз основа на която се извършват последователни анализи. Комбинирането на резултати за токсичността от няколко времеви точки и няколко анализа може да осигури по-чувствителни и специфични оценки на токсичността. GI_{50} не може да бъде изчислен или не е оптимален за някои от избраните анализи, използвани в методологията на Misvik. Поради тази причина е разработена и използвана GI_{50} – независима точкова система Tox5-score (фигура 11). Подходът Tox5-score[18], се основава на изчислени параметри от доза-ефект кривата, интегриране на тези показатели в един общ резултат за токсичност, изразен като токсикологичен приоритизиращ индекс, който запазва прозрачността спрямо приноса на всеки конкретен анализ.



Фигура 11: *Подход Tox5-score за групиране на in-vitro данни за токсичност.*

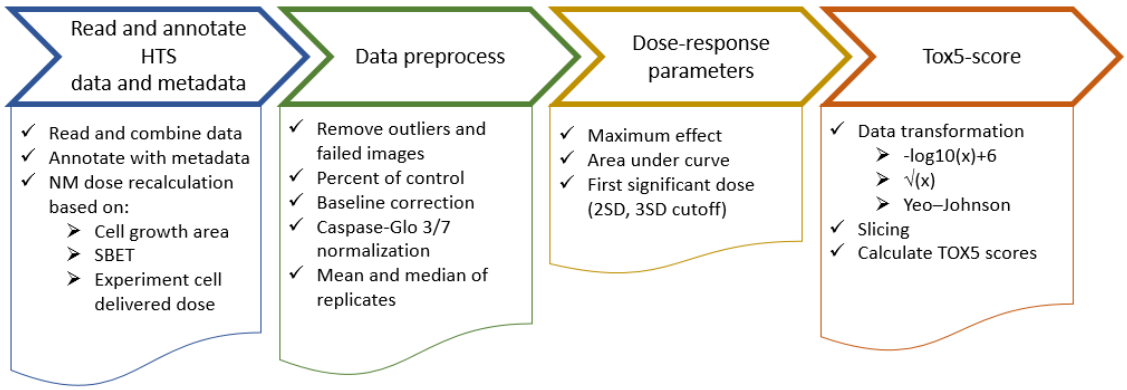
Профилът на токсичност е представен с графика от тип „пай“, която представя сходството в реакциите на токсичност. Сложността и визуализацията варира в зависимост от количеството включена информация. Подходът позволява ясна визуализация на цялостната оценка, позволявайки материалите да бъдат класирани и сравнени с контролни вещества с известна токсичност. При Tox5-score може да се комбинират данни от други експерименти с подобни параметри, което разширява и приложимостта му.

На базата на разработената методология за високопроизводителен скрининг и подхода за изчисляване на токсикологичен приоритизиращ индекс Tox5-score, ние въведохме обща концепция за предварителна обработка и групиране на материали. Въз основа на общата концепция разработихме софтуерна библиотека ToxFAIRy и потребителски интерфейс Orange3-ToxFAIRy.

8.2 Обща концепция за предварителна обработка на данни от високопроизводителен скрининг и Tox5-Score групиране

Общата концепцията за предварителна обработка на данни от високопроизводителен скрининг и изчисляване на токсикологичен индекс чрез Tox5-Score е показана на фигура 12. Тя може да се реализира с различни изчислителни техники, като MS Excel или друг софтуер.

Първата стъпка е свързана със събиране и анотиране на експериментални данни с метаданни, както и преизчисляване на прилаганите дози според специфичната повърхност на наноматериали и спрямо площта на растеж на клетките. Втората стъпка включва нормализация и предварителна обработка на данните, с цел минимизиране на въздействието на систематичните грешки, породени от човешки, биологични или технически фактори по време на експерименталния процес. Така се намалява нежеланата вариация и шумът в данните. Третата стъпка включва изчисляване на параметрите от доза – ефект кривата: максимален ефект, площ под кривата и първа значима доза. Последната стъпка е изчисляване на токсикологичен приоритизиращ индекс и групиране на материалите.



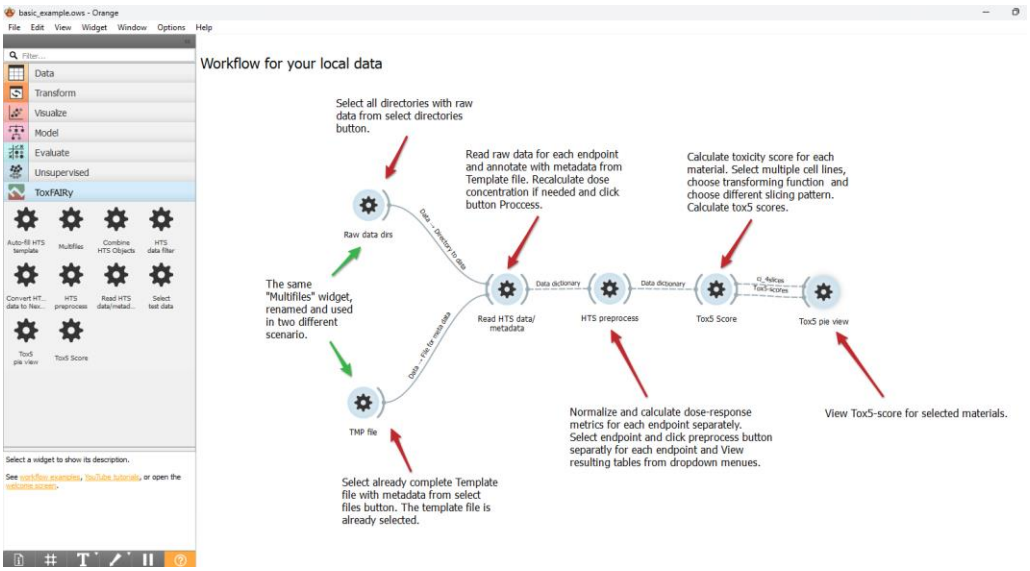
Фигура 12: *Общи компоненти и стъпки в работния процес, изграждащ общата концепция за предварителна обработка на HTS данни и оценка на токсичността.*

8.3 Разработка на софтуерна библиотека ToxFAIRy

Разработихме python библиотека ToxFAIRy за: (i) прочитане и анотиране на HTS данни, (ii) предварителната обработка, (iii) FAIR-ификация и (iv) оценяване на токсичност. Библиотеката ToxFAIRy е достъпна в хранилището orange3-toxfairy на платофрмата GitHub чрез линка: https://github.com/ideaconsult/orange3-toxfairy.

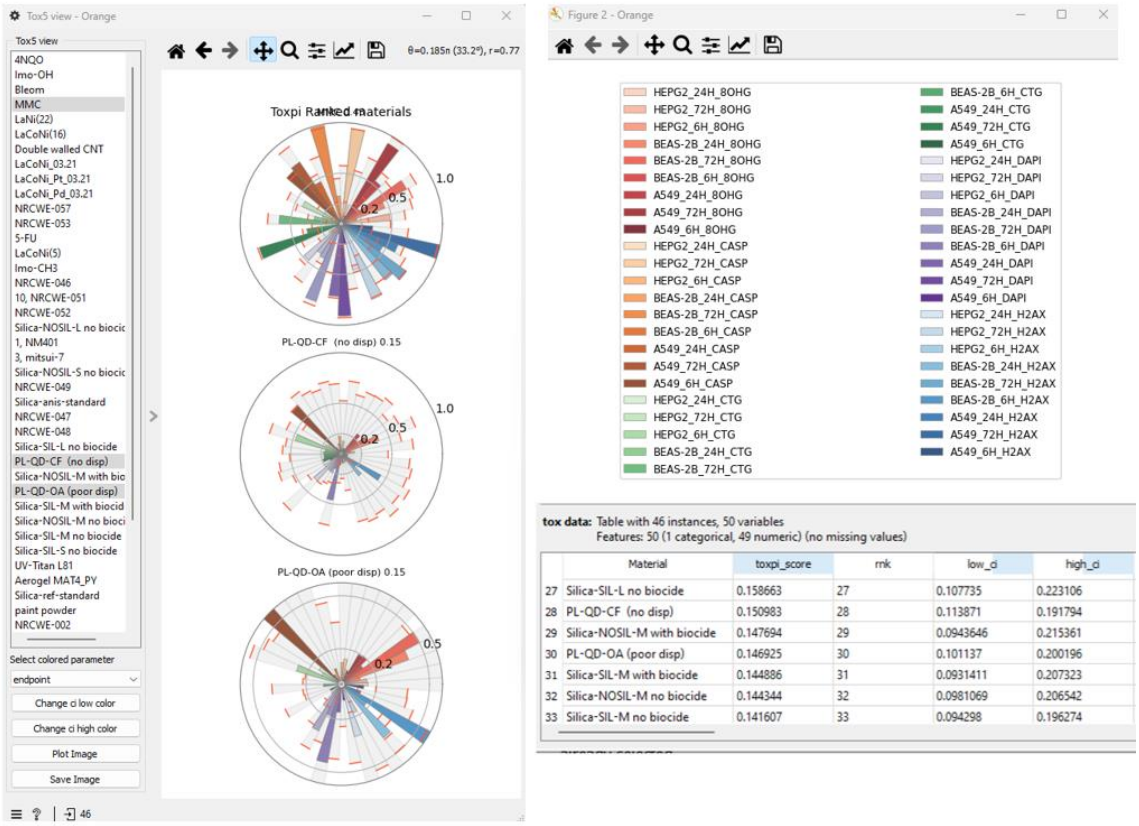
8.4 Разработка на потребителски интерфейс Orange3-ToxFAIRy

Разработихме плъгин (add-on) Orange3-ToxFAIRy към софтуера за машинно обучение и анализ на данни Orange. Orange3-ToxFAIRy може да се използва като потребителски интерфейс към ToxFAIRy библиотеката. Отделните софтуерни компоненти изграждат пълен автоматизиран работен процес (фигура 13). Основното предимство е, че може да се създават сложни работни процеси (workflows) в стил визуално програмиране и да бъдат запазени като OWS файлове и заредени за използване отново и/или споделяни с други потребители. Към потребителския интерфейс разработихме и ръководство за употреба, достъпно от линка: https://doi.org/10.5281/zenodo.13685297.



Фигура 13: *Примерен автоматизиран работен процес (workflow) за обработка на HTS данни и изчисляване на токсикологичен приоритизиращ индекс Tox5-score.*

Профилите на токсичност се визуализират за всеки материал по-отделно чрез софтуерния компонент Tox5 view (фигура 14). Изчислените доверителни интервали са показани в сиво за всеки конкретен срез. За удобство и по-ефективна работа на потребителите, разработихме и автоматизиран подход за оцветяване на срезовете според анализи, клетъчни линии или време за третиране. В показания пример срезовете са оцветени въз основа на различните анализи. Освен графичната визуализация на резултатите за токсичност на материалите, информацията е налична и в табличен вид.



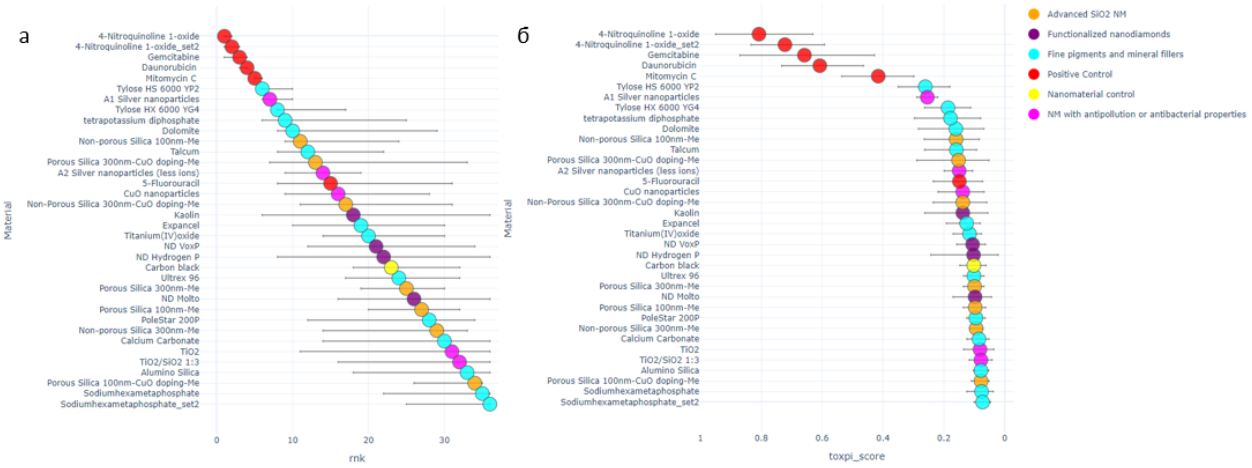
Фигура 14: *Визуализация на изчислените индекси Tox5-score, представени в графика от тип „пай“.*

8.5 Приложение на автоматизирания работен процес към данни от проектите caLIBRAte и HARMLESS.

Всички данни, използвани за разтаботка и тестване, са публично достъпни от линка към Zenodo: https://doi.org/10.5281/zenodo.13683162

Библиотеката ToxFAIRy беше използвана за изграждане на скриптов автоматизиран работен процес с библиотеката ploomber. Работният процес е изграден от конкретни задачи (https://github.com/ideaconsult/orange3-toxfairy/tree/main/toxfairy_workflow/tasks) с добре установени връзки по между си, представени на фигура 15. Цялостната конфигурация на работния процес се задава чрез JSON файл с конкретни инструкции.

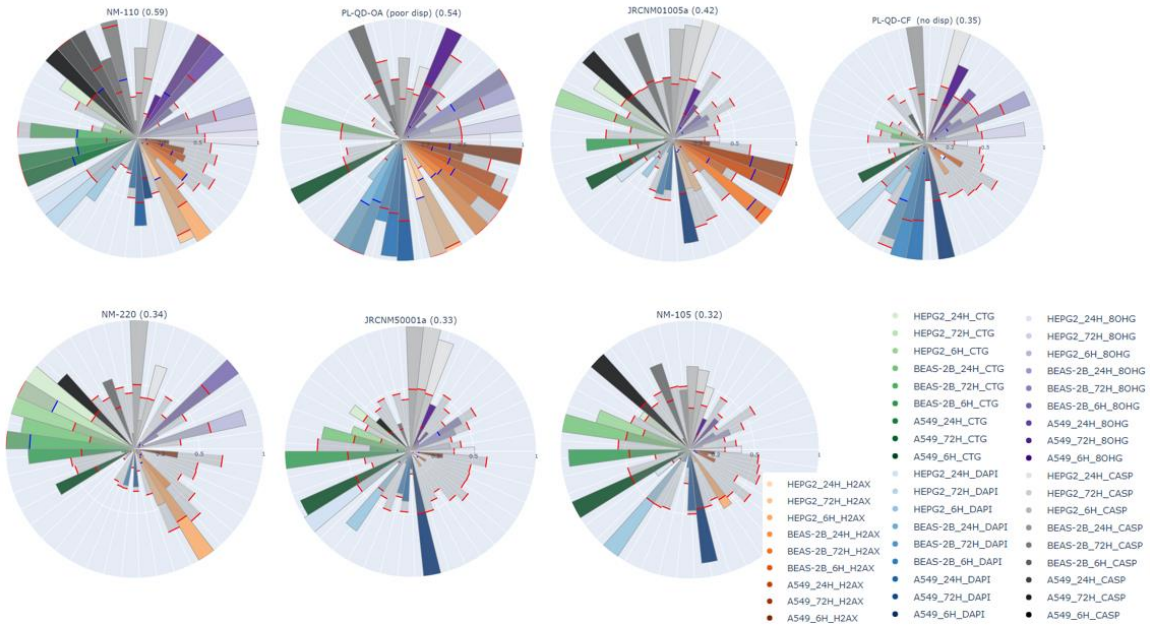
***Фигура 15**: Автоматизиран работен процес за обработка, FAIR-ификация и групиране на HTS данни.*



***Фигура 16:** Класирани калибрирани материали и контроли с начален доверителен интервал а) по рангове и б) по Tox5-резултат.*

На фигура 16 са представени резултатите от автоматичната обработка и класиране на набор материали според данни от проекта caLIBRAte. Доверителните интервали са изчислени за ранговете (рангове по токсичност, като нарастват с нарастване на токсичността) и за Tox5-score индексите и са представени като error bars. Групите материали са цветово разграничени. Включените положителни контроли, отбелязани в черен цвят, и контролите от наноматериали, отбелязани в жълто, създават референтно поле за относителните токсични класации на материалите и позволяват сравнение с нови набори от данни със сходни контроли. Tylose HX 6000 YG4 (химически модифицирана

хидроксиетил целулоза, от групата на пигменти и минерални филтри), отбелязани със син цвят, и TiO2/SiO_2 (от групата наноматериали с противозамърсяващи и антибактериални свойства), отбелязани със светло лилаво, са класирани съответно като най-токсичният и най-слабо токсичният тестван наноматериал.

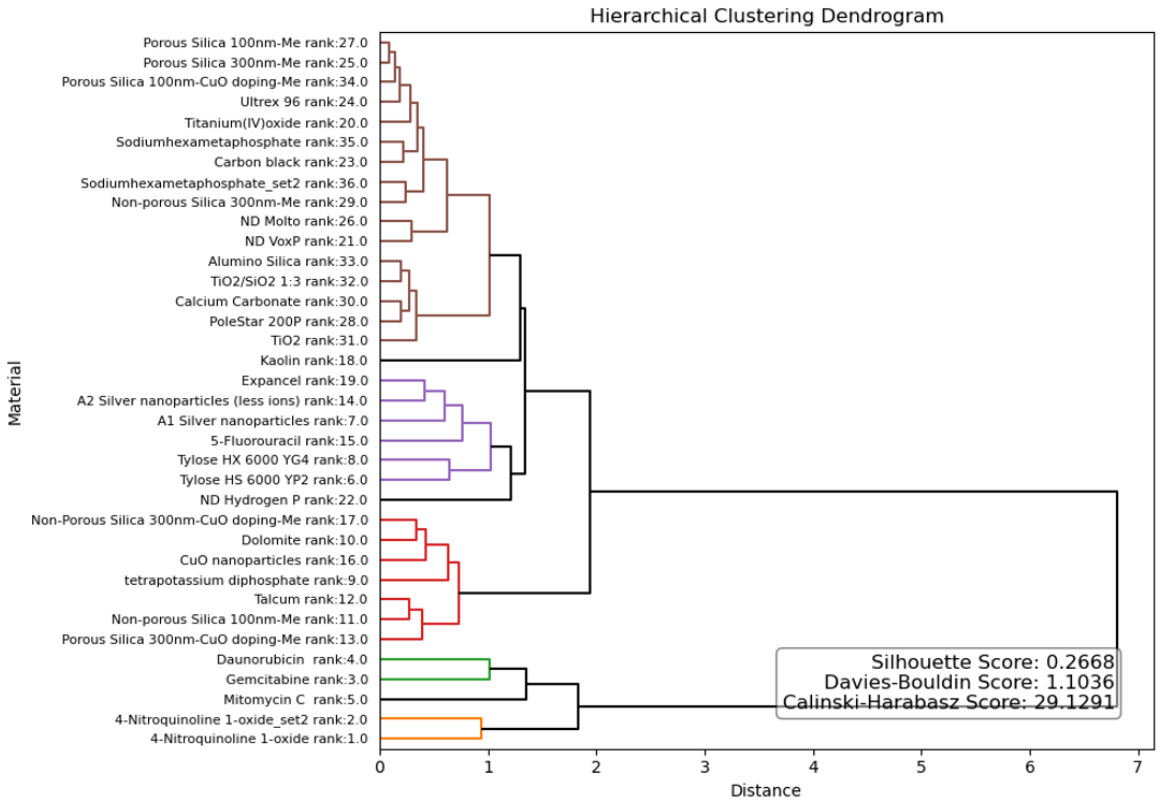


Фигура 17: *Графично представяне на профилите на токсичност за квантови точки и контролни наноматериали.*

Фигура 17 демонстрира групирани квантови точки (ZnCuInS core/ZnS shell (PL-QD-CF) и ZnCdSeS no shell (PL-QD-OA)) и наноматериали като позитивни и негативни контроли чрез графика от тип „пай“, където всеки „пай“ представлява общия Tox5-score за всеки материал. Всеки срез съответства на конкретен резултат за група параметри и е оцветен според анализа. Например, групите на анализа на CTG са оцветени в зелено с градиент, указващ специфичните клетъчни линии и времеви точки. Освен това е изобразен и доверителният интервал за всеки срез, като горната граница е показана в червено, а долната граница в синьо. Квантовите точки намаляват броя жизнеспособни клетките (сини сектори) и предизвикват оксидативен стрес на нуклеиновите киселини (лилави сектори), докато цинковият оксид (NM-110), класиран като най-токсичен, е по-склонен да предизвика апоптоза (сиви сектори) и загуба на клетъчна жизнеспособност (зелени сектори).

Фигура 18 илюстрира йерархично клъстериране, за данните от caLIBRAte, с евклидово разстояние като метрика и метод за групиране - Ward[19]. Оптималният брой клъстери е определен с помощта на метода Elbow, който демонстрира по-добро статистическо представяне по отношение на показателите за значимост на клъстера (Silhouette[20], Davies-

Bouldin[21] and Calinski-Harabasz[22] scores), в сравнение с метода Silhouette. Тези показатели за оценка на клъстерирането са автоматично изчислени и визуализирани на дендрограмата.



Фигура 18: *Йерархично групиране на материалите от caLIBRAte.*

Въпреки малкия и хетерогенен набор от данни, се образуват статистически значими клъстери, например между нанодиамантите VoxP и Molto и порьозни силициеви частици.

Методът TOPSIS беше приложен за извадката от данни с квантови точки и беше съпоставен с резултатите от Tox5-score подхода (Вж. таблица 2). Висока стойност за TOPSIS preferences показва колко близо е дадената алтернатива до по-малко токсичен материал, докато високият Tox5-score представлява потенциално силно токсичен материал. Двата метода показват доста сходно поведение при групирането на материалите с изключение на квантова точка ZnCuInS core/ZnS shell и JRCNM01005a, които са с разменени позиции. Избрахме TOPSIS метода като алтернатива и проверка на Tox5-score подхода, също така TOPSIS е отбелязан като един от петте най-популярни метода, прилагани при вземане на решения според множество критерии, описани в последното ръководства за SSbD (Safe and Sustainable by Design) на JRC[23].

***Таблица 2**: Съпоставка на методите за групиране TOPSIS и Tox5-score, приложени за квантови точки.*

Материал	TOPSIS оценка на предпочитания ефект	Tox5-score	TOPSIS ранг	Tox5-score ранг
NM-110	0.336	0.591	1	1
ZnCdSeS no shell	0.441	0.535	2	2
ZnCuInS core/ ZnS shell	0.555	0.351	3	4
JRCNM01005a	0.631	0.425	4	3
NM220	0.635	0.343	5	5
JRCNM50001a	0.702	0.330	6	6
NM-105	0.807	0.320	7	7

8.6 FAIR-ификация на данни от високопроизводителен скрининг (HTS)

След предварителната обработка и изчисляване на параметрите доза – ефект, структурите от данни се преобразуват в модела за данни на Ambit/eNanoMapper с помощта на библиотеката pynanomapper и се сериализират като NeXus файл. Това позволява импортиране в базата данни и достъп до качествени FAIR данни от високопроизводителен скрининг (HTS). Полученият NeXus файл съдържа всички материали, анализи, включително необработени и обработени данни. Данните се съхраняват като йерархична дървовидна структура от многомерни матрици. Фигура 19 представя Nexus файл за материал: Porous Silica 300nm-CuO doping-Me, анализ: 8OHG с клетъчна линия: BEAS-2B. Фигура 65 представлява графика от тип „heatmap“ за нормализирани данни със селектиран поднабор от данни от реплики D2.

***Фигура 19**: HTS данните на Misvik, преобразувани във формат NeXus*

IV Заключения

9.1 Научни приноси

1. Разработена е концепция за FAIR-ификация на експериментални данни за наноматериали, базирана на представяне на информацията за мултикомпонентни субтсанции чрез семантичен модел за данни Ambit/eNanoMapper с помощта на конфигурируем софтуерен инструмент NMDataParser.
2. Създадохме прототип за идентификатор за наноматериали, който демонстрира възможностите на линейната нотация SLN за представяне на обекти от химичната информатика и наноинформатиката с потенциал за генериране на глобален уникален идентификатор. Създаденият прототип позволява да се кодират богати метаданни заедно със структурната информация, което важно в контекста на управление на химична информация за академични, регулаторни и индустриални нужди.
3. Разработихме обща концепция за анотиране на HTS данни с метаданни, предварителна обработка и изчисляване на токсикологичен приоритизиращ индекс чрез Tox5-Score подхода. Общата концепция дава възможност за разработка на софтуерни инструменти за автоматизирана обработка на данни от високо производителен скриниг, което е в съответствие с регулаторните препоръки и широко изразените нужди на индустрията.

9.2 Научно-приложни приноси

1. Приложихме процеса по FAIR-ификация за 1400 EXCEL файла, чрез което обогатихме базата данни на eNanoMapper с висококачествени FAIR данни, с информация за безопастността на наноматериали от няколко големи европейски проекта. Данните

включват голямо разнообразие от физикохимични и биологични анализи за множество наноматериали.
2. Обогатихме онтологията eNanoMapper в областта на екотоксичност и опазване на околната среда, като добавихме 15 нови термина.
3. Въз основа на разработената обща концепция за анотиране и обработка на HTS данни, разработихме софтуерна библиотека ToxFAIRy, за анотация и обработка на данните, изчисляване на приоритизиращият индекс и FAIR-ификация на HTS данни.
4. Разработихме модул Orange3-ToxFAIRy към аналитична платформа Orange, който може да се ползва като потребителски интерфейс към софтуерната библиотека ToxFAIRy.
5. Разработихме автоматизиран работен процес на платформата Ploomber, който напълно обхваща функционалносттите на библиотеката ToxFAIRy и позволява бърза едновременна обработка на голям обем от данни. Автоматизираният работен процес е приложен за HTS данни за наноматериали и advanced материали, генерирани по европейските проекти caLIBRAte и HARMLESS .

9.3 Насоки за бъдещо развитие

Резултатите, описани в настоящата дисертация, очертават някои основни насоки за бъдещо развитие:

1. Разработване на софтуерни инструменти за изчисляване на дескриптори за наноматериали, advanced материали и мултикомпоненти субстанции въз основа на представянето на обектите чрез FAIR семантичен модел за данни.
2. Приложение на нови модерни алгоритми и векторни преставяния (AI embedings), базирани на изкуствен интелект, за моделиране, QSPR/QSAR анализ, readacross методологии и свързване на модела за данни на Ambit/eNanoMapper със системи базирани на семантично знание.
3. Усъвършенстване и развитие на предложения прототипен SLN идентификатор за мултикомпонентни субстанции и наноматериали. Разработка на алгоритми, които да направят SLN идентификатора уникален. Това би улеснило не само научната общност, но и регулаторните агенции и индустрията в управлението на данни и съхранението на информация за наноматериали.
4. Разширяване на финкционалностите на софтуерната библиотека ToxFAIRy в няколко направления:

- разработка на нови токсикологични индекси включващи физикохимични показатели и резултати от in-vivo анализи, интергирани от базата данни на eNanoMapper.
- разработка на подходи за интегриране на информация от графови бази данни, базирани на информация от AOPs (Adverce Outcome Pathways)[24], която да служи за валидация на получените резултати от приоритизиране на субстанциите.

V Научни съобщения по дисертацията

10.1 Публикации

A1 Kochev, N.; Jeliazkova, N.; Paskaleva, V.; Tancheva, G.; Iliev, L.; Ritchie, P.; Jeliazkov, V. "Your Spreadsheets Can Be FAIR: A Tool and FAIRification Workflow for the eNanoMapper Database". Nanomaterials 2020, 10, 1908. https://doi.org/10.3390/nano10101908

със забелязани 16 цитата

A2 Kochev N, Jeliazkova N, Tancheva G. "Ambit-SLN: an Open Source Software Library for Processing of Chemical Objects via SLN Linear Notation". Mol Inform. 2021 Nov;40(11):e2100027. doi: 10.1002/minf.202100027. Epub 2021 Aug 3. PMID: 34342942.

със забелязан 1 цитат

A3 Jeliazkova N, Kochev N, Tancheva G. "FAIR Data Model for Chemical Substances: Development Challenges, Management Strategies, and Applications". Data Integrity and Data Governance. IntechOpen; 2023. Available from: http://dx.doi.org/10.5772/intechopen.110248

10.2 Постери и доклади

П1 G. Tancheva, N. Kochev , N. Jeliazkova , V. Paskaleva, *DATA PROCESSING FOR CHEMICAL SUBSTANCES AND NANOMATERIALS*, научна конференция Актуални регулативни изисквания към химичния анализ и модерни инструментални технологии за тяхното покриване, 06.2019, Пловдив

П2 G. Tancheva, N. Kochev , V. Paskaleva, *QSAR modeling of melting points of organic compounds. Methods comparison.* Пета научна конференция за студенти, докторанти и млади учени "Предизвикателства в химията", октомври 2019, Пловдив.

П3 G. Tancheva, N. Kochev, N.Jeliazkova, V. Paskaleva, L. Iliev, P. Ritchie, V.Jeliazkov, *FAIRification Workflow for Handling Nano Safety Excel Spreadsheet templates in eNanoMapper database*. International FAIR Convergence Symposium,12.2020

П4 G. Tancheva, N. Kochev, V. Jeliazkov, L. Iliev, N. Jeliazkova, *Fairification workflow for integrating nanosafety data: Enanomapper database*, ACM2 семинар 06.2022, Пловдив

П5 G. Tancheva, N. Kochev , N. Jeliazkova *Electronic notebook for nanosafety data interactive preprocessing analysis and visualization*. 6 Научна конференция за студенти, докторанти и млади учени „Предизвикателства в Химията", 10. 2022, Пловдив

П6 G. Tancheva, P. Nymark, V. Hongisto , N. Kochev, N. Jeliazkova, *Automatic workflow for HTS data FAIRification, preprocessing and Tox5 in-vitro toxicity scoring.* QSAR - 2023, 06. 2023, Копенхаген

П7 G. Tancheva, P. Nymark, V. Hongisto , N. Kochev, N. Jeliazkova, *Automatic workflow for in vitro high-throughput screening data FAIRification, preprocessing and scoring*, 12 Химична конференция, 11. 2023, Пловдив

П8 G. Tancheva, V. Hongisto, K. Patyra, L. Iliev, N. Kochev, P. Nymark, R. Grafström, N. Jeliazkova*, Automatic workflow for in vitro high-throughput screening data FAIRification, preprocessing and scoring*, General Assembly HARMLESS, 10-11. 01.2024, Лудвигсхафен, https://www.harmless-project.eu/harmless-general-assembly-2024/

П9 G. Tancheva, V. Hongisto, K. Patyra, L. Iliev, N. Kochev, P. Nymark, R. Grafström, N. Jeliazkova, *Automatic workflow for in vitro high-throughput screening data FAIRification, preprocessing and scoring*, NANOTOX 2024, 09.2024, Венеция

Д1 *Fairification workflow for integrating nanosafety data: Enanomapper database.* SciDataCon-IDW Seoul 06. 2022. Устен доклад онлайн форма.

Д2 *Automatic workflow for HTS data FAIRification, preprocessing and Tox5 in-vitro toxicity scoring*, M30 General Assembly, HARMLESS project, Turku (Finland) 13 – 14. 06. 2023. Устен доклад онлайн форма, https://www.harmless-project.eu/general-assembly-m30-in-turku/

Д3 *Automatic workflow for HTS data FAIRification, preprocessing and toxicity scoring. Case study: Quantum dots*, Обобщаващо събрание на обединени в сътрудничество проекти по програмата HORIZON 2020 (DIAGONAL, HARMLESS, SUNSHINE) NMBP-16 ambassadors, 18. 06. 2024, Устен доклад онлайн форма.

10.3 Проекти

Проект 1 Международно сътрудничество с университета в Маастрихт по програмата **Transnational Access** (TA) към проекта NanoCommons (HORIZON 2020 EU, agreement #731032), https://www.nanocommons.eu/e-infrastructure/awarded-ta-projects/

Проект 2 **HARMLESS**-Advanced High Aspect Ratio and Multicomponent materials: towards comprehensive intelLigent tEsting and Safe by design Strategies, (HORIZON 2020 EU agreement ID: 953183) https://www.harmless-project.eu/project-summary/

Проект 3 **NanoReg2**-Development & Implementation of Grouping & Safe-by-Design Approaches within Regulatory Frameworks, (Grant agreement ID: 646221)

Проект 4 **NanoinformaTIX**-Development and Implementation of a Sustainable Modelling Platform for NanoInformatics (Grant agreement No 814426) https://www.nanoinformatix.eu/

Проект 5 **POLYRISK**- Understanding human exposure and health hazard of micro- and nanoplastic contaminants in our environment (Grant agreement No 964766) https://polyrisk.science/

VI Библиография

1. Engel, T; Gadteiger J. *Chemoinformatics: A Textbook*.; 2003. doi:10.1002/3527601643
2. Комисия на ЕС. *ПРЕПОРЪКА (ЕС) 2022/2510 НА КОМИСИЯТА*.; 2022. doi:10.1890/0012-9623(2004)85[163:po]2.0.co;2

3. Mark D. Wilkinson; et all; Comment: The FAIR Guiding Principles for scientific data management and stewardship. 2016:1-9.
4. MESOCOSM database. https://aliayadi.github.io/MESOCOSM-database/.
5. Kochev N, Jeliazkova N, Paskaleva V, et al. Your spreadsheets can be fair: A tool and fairification workflow for the enanomapper database. *Nanomaterials*. 2020;10(10):1-23. doi:10.3390/nano10101908
6. eNanoMapper. https://www.enanomapper.net/.
7. Онтология eNanoMapper. https://github.com/enanomapper/ontologies.
8. Janna Hastings (EMBL-EBI) EW (UM). Deliverable Report D2.1 Framework and Infrastructure for ontology development, versioning and dissemination. 2014;(February 2014):1-30.
9. Hastings J, Jeliazkova N, Owen G, et al. eNanoMapper : harnessing ontologies to enable data integration for nanomaterial risk assessment. 2015:1-15. doi:10.1186/s13326-015-0005-5
10. Ash S, Cline MA, Homer RW, Hurst T, Smith GB. SYBYL Line Notation (SLN): A versatile language for chemical structure representation. *J Chem Inf Comput Sci*. 1997;37(1):71-79. doi:10.1021/ci960109j
11. Homer RW, Swanson J, Jilek RJ, Hurst T, Clark RD. SYBYL line notation (SLN): A single notation to represent chemical structures, queries, reactions, and virtual libraries. *J Chem Inf Model*. 2008;48(12):2294-2307. doi:10.1021/ci7004687
12. CHEBI 51050. https://www.ebi.ac.uk/chebi/searchId.do?chebiId=51050.
13. Levenshtein distance. https://en.wikipedia.org/wiki/Levenshtein_distance#cite_note-1.
14. To KT, Fry RC, Reif DM. Characterizing the effects of missing data and evaluating imputation methods for chemical prioritization applications using ToxPi. *BioData Min*. 2018;11(1):1-12. doi:10.1186/s13040-018-0169-5
15. Marvel SW, To K, Grimm FA, Wright FA, Rusyn I, Reif DM. ToxPi Graphical User Interface 2.0: Dynamic exploration, visualization, and sharing of integrated data models. *BMC Bioinformatics*. 2018;19(1):1-7. doi:10.1186/s12859-018-2089-2
16. Kohonen P, Ceder R, Smit I, et al. Cancer biology, toxicology and alternative methods development go hand-in-hand. *Basic Clin Pharmacol Toxicol*. 2014;115(1):50-58. doi:10.1111/bcpt.12257
17. Grafström RC, Nymark P, Hongisto V, et al. Toward the replacement of animal experiments through the bioinformatics-driven analysis of “omics” data from human cell cultures. *ATLA Altern to Lab Anim*. 2015;43(5):325-332. doi:10.1177/026119291504300506
18. Hongisto V, Nymark P. Systems toxicology to support development of adverse outcome pathways, Abstracts of the 55th Congress of the European Societies of Toxicology (EUROTOX 2019) TOXICOLOGY SCIENCE PROVIDING SOLUTIONS. *Toxicol Lett*. 2019;314(October):S25. doi:10.1016/j.toxlet.2019.09.002
19. Ward JH. Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc*. 1963;58(301):236-244. doi:10.1080/01621459.1963.10500845
20. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20(C):53-65. doi:10.1016/0377-0427(87)90125-7
21. Davies DL, Bouldin DW. A Cluster Separation Measure. *IEEE Trans Pattern Anal Mach Intell*. 1979;PAMI-1(2):224-227. doi:10.1109/TPAMI.1979.4766909

22. Calinski T, Harabasz J. Communications in Statistics A dendrite method for cluster analysis. *Commun Stat*. 1974;3(1):1-27.
23. Abbate E, Garmendia Aguirre I, Bracalente G, et al. *Safe and Sustainable by Design Chemicals and Materials - Methodological Guidance*.; 2024. doi:10.2760/28450
24. Adverse Outcome Pathway. https://aopwiki.org/.