Universitatea Babeș-Bolyai

Facultatea de Chimie și Inginerie Chimică

Școala Doctorală de Chimie

Rezumatul tezei de doctorat

Estimări și asocieri de proprietăți/activități chimice bazate pe similaritate

Doctorand: **DONATELLA BÁLINT (Căs. NAGY)**
Coordonator științific: **PROF. DR. LORENTZ JÄNTSCHI**

CLUJ-NAPOCA
2023

Rezumatul tezei de doctorat

Estimări și asocieri de proprietăți/activități chimice bazate pe similaritate

Doctorand: **DONATELLA BÁLINT (Căs. NAGY)**

Președinte:

Prof. Univ. Dr. Radu-Lucian SILAGHI-DUMITRESCU – Facultatea de Chimie și Inginerie Chimică, Universitatea Babeș-Bolyai, Cluj-Napoca

Coordonator științific:

Prof. Dr. Lorentz JÄNTSCHI – Universitatea Tehnică din Cluj-Napoca, Departamentul de Fizică și Chimie

Referenți:

Conf. Univ. Dr. Réka BARABÁS – Facultatea de Chimie și Inginerie Chimică, Universitatea Babeș-Bolyai, Cluj-Napoca

Conf. Univ. Dr. Mădălina Ana VĂLEANU – Universitatea de Medicină și Farmacie Iuliu Hațieganu, Cluj-Napoca

C.S. I Dr. Attila BENDE – Institutul Național de Cercetare-Dezvoltare pentru Tehnologii Izotopice și Moleculare, Cluj-Napoca

Data susținerii publice: Iulie 10 , 2023
Locația: Facultatea de Chimie și Inginerie Chimică, Universitatea Babeș-Bolyai, Cluj-Napoca

CUVINTE CHEIE: modelare moleculară, similaritate biochimică, optimizare geometrică, aminoacizi, gaussian, relații structură/proprietate.

LISTA DE PUBLICAȚII

1. Jäntschi, L.; **Bálint**, **D**.; Bolboacă, S.D. 2016. Multiple linear regressions by maximizing the likelihood under assumption of generalized Gauss-Laplace distribution of the error. Computational and Mathematical Methods in Medicine. Doi: 10.1155/2016/8578156.

2. Jäntschi, L.; **Bálint**, **D**.; Pruteanu, L.L.; Bolboacă, S.D. 2016. Elemental factorial study on one-cage pentagonal faces nanostructure congeners. Materials Discovery, 5, pp. 14 - 21. Doi: 10.1016/j.md.2016.12.001.

3. **Bálint**, **D**.; Jäntschi, L. 2019. Missing data calculation using the antioxidant activity in selected herbs. Symmetry, 11(6). Doi: 10.3390/sym11060779.

4. **Bálint**, **D**.; Jäntschi, L. 2021. Comparison of molecular geometry optimization methods based on molecular descriptors. Mathematics, 9(22). Doi: 10.3390/math9222855.

5. Joița, D.M.; Tomescu, M.A.; **Bálint**, **D**.; Jäntschi, L. 2021. An application of the eigenproblem for biochemical similarity. Symmetry, 13(10). Doi: 10.3390/sym13101849.

CUPRINS

Lista de publicații **3**
Cuprins **4**
Obiectivele tezei de doctorat **5**
Capitolul I – Introducere **6**
1. Structura compușilor chimici **6**
2. Similaritatea moleculară **7**
3. Relația cantitativă structură-activitate/structură-proprietate (QSAR/QSPR) **8**
4. Optimizarea moleculară **10**

Contribuții personale
Capitolul II – Date cenzurate în calculele de cercetare **12**
1. Introducere **12**
2. Metodologie **13**
3. Rezultate și discuții **15**
4. Concluzii **18**

Capitolul III – Evaluarea similarității folosind funcția polinomului caracteristic (ChP) **19**
1. Introducere **19**
2. Metodologie **19**
3. Rezultate și discuții **20**
4. Concluzii **23**

Capitolul IV – Evaluarea similarității folosind softul *Gaussian09* **24**
1. Introducere **24**
2. Metodologie **25**
3. Rezultate și discuții **28**
4. Concluzii **42**

Capitolul V – Aplicații ale tehnicilor de optimizare geometrică **43**
Studiu de caz – Analiza factorială al nanostructurilor **43**
1. Introducere **43**
2. Metodologie **43**
3. Rezultate și discuții **44**
4. Concluzii **45**

Capitolul VI – Aplicații ale tehnicilor de aliniere moleculară **46**
Studiu de caz – Similitudine biochimică a proteinelor selectate **46**
1. Introducere **46**
2. Metodologie **46**
3. Rezultate și discuții **48**
4. Concluzii **49**

Concluzii generale și perspective de viitor **50**
Bibliografie selectivă **51**

Obiectivele tezei de doctorat

Scopul general al cercetării actuale este de a oferi un cadru interdisciplinar pentru explicarea comportamentului molecular al structurilor chimice precum și a impactului acestora la nivelul sistemelor biologice folosind concepte din chimie și biologie-biochimie.

Obiectivul studiului a fost de a explica și de a înțelege modul în care structurile moleculare diferă și de a le clasifica pe baza asemănărilor lor. Pentru aceasta au fost luate în considerare mai multe abordări: modele de regresie liniară multiple, aplicarea problemelor proprii, studiu factorial, algoritmi iterativi și calcule de optimizare a geometriei efectuate folosind o varietate de tehnici (Metode Hartree-Fock, Metode Semiempirice, Teoria Funcțională a Densității, Mecanica Moleculară).

Pentru a înțelege modul în care aceste strategii sunt în relație între ele și pentru a alege pe care să se aplice în diverse situații, studiul și-a propus să analizeze aceste metode.

După fiecare parte a cercetării noastre, analiza statistică a fost efectuată pentru a valida rezultatele noastre (analiza componentelor principale, analiza clusterului, analiza variației, alte metode de extragere a datelor) și pentru a compara similitudinile diferitelor abordări.

Capitolul I - Introducere

1. Structura compușilor chimici

Formula structurală a unui compus chimic este o reprezentare grafică a structurii moleculare care indică modul în care atomii sunt grupați în trei dimensiuni (Figura 1).

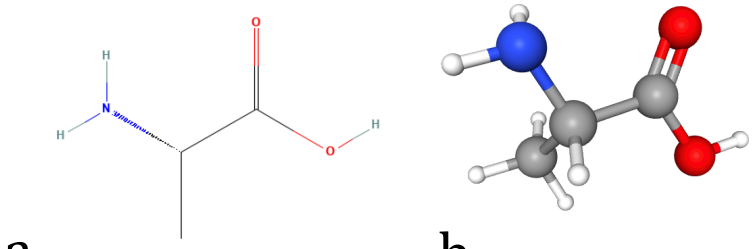

Figura 1. Ilustrarea schematică a moleculei de L-Alanina formelor 2D(a) și 3D(b) (Baza de date PubChem, Accesat la data de 17.04.2023)

Există o relație foarte clară între proprietățile compusului chimic și structura acestora, prin aceea că proprietățile sunt determinate de structură, iar anumite aspecte structurale pot fi deduse din evaluarea și interpretarea proprietăților (Figura 2).

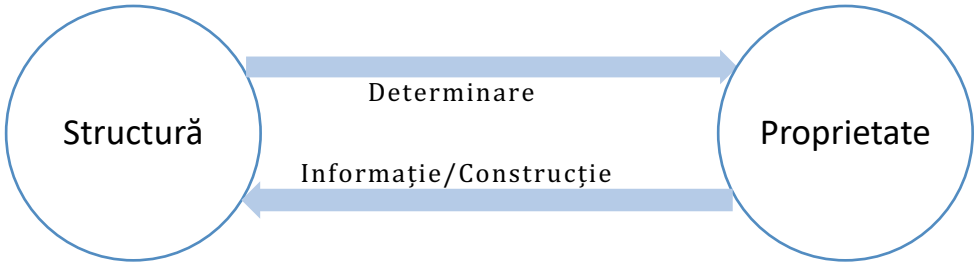

Figura 2. Relație proprietate/structură

În molecule pot fi observate diferite tipuri de simetrie geometrică. Simetria geometrică în molecule se referă la aranjarea simetrică a atomilor și a legăturilor dintr-o moleculă. Este determinată de prezența elementelor de simetrie, cum ar fi axele de rotație, planurile de reflexie, centrele de inversare și axele de rotație necorespunzătoare.

Un graf, $G = G\ (V, E)$ se poate defini prin: $V = V(G)$, o mulțime finită (nu goală) de N puncte (vârfuri) și $E = E(G)$, mulțimea de Q perechi neordonate de diferite puncte ale lui V.

Două vârfuri sunt adiacente dacă sunt conectate printr-o muchie, iar fiecare pereche de puncte reprezintă o linie (margine). Când două muchii separate se intersectează într-un

singur punct, se spune că sunt muchii adiacente. De multe ori, atomii de hidrogen sunt omiși (Diudea și colab., 2002).

Există mai multe tipuri diferite de grafuri, câteva exemple ale acestora sunt reprezentate mai jos (Figura 3):

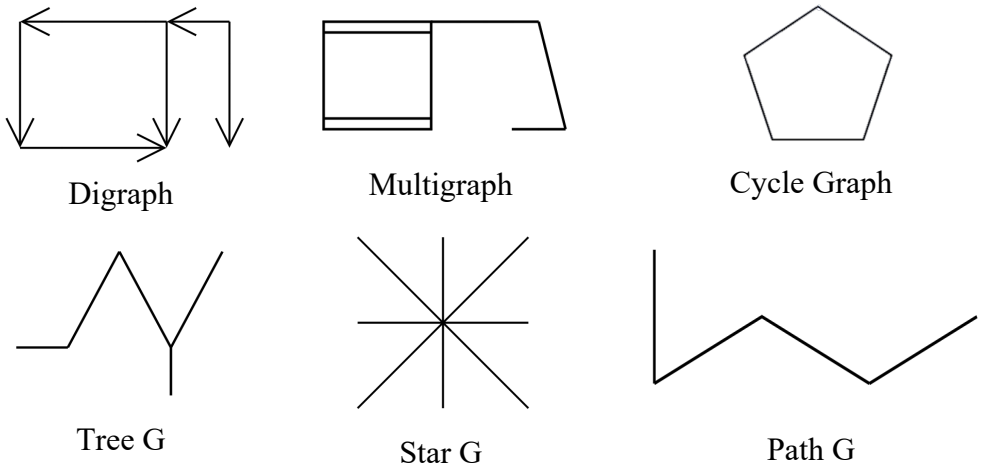

Figura 3. Exemple de grafuri

Un graf molecular cu atomii ca vârfuri și legăturile covalente ca margini poate fi folosit pentru a descrie formula structurală a unei molecule chimice.

2. Similaritatea moleculară

Identificarea diferitelor modele poate fi facilitată de asemănarea a două structuri chimice (Doucet și Weber, 1996). Sunt necesare alte tehnici pentru a calcula cât de asemănătoare sunt structurile moleculare diferite între ele (Bender și Glen, 2004).

Procedura de evaluare a similitudinii moleculare este utilizată pentru a evalua caracteristicile structurale a două sau mai multor molecule. Este o etapă critică în dezvoltarea și proiectarea de noi medicamente, deoarece ajută la identificarea posibililor candidați la medicamente pe baza asemănărilor cu substanțele active existente.

Există diverse tehnici de evaluare a similitudinii moleculare, cum ar fi: metode bazate pe amprenta 2D (folosește amprentele moleculare 2D pentru a compara caracteristicile structurale ale moleculelor); metode bazate pe forme 3D (folosește forma 3D a moleculelor pentru a compara caracteristicile lor structurale); metode bazate pe învățare automată (folosește algoritmi de învățare automată) (Stumpfe și Bajorath, 2011).

O măsură a cât de mult se potrivesc proprietățile unei perechi de molecule se numește similaritate moleculară. Pot fi calculate numeroase caracteristici moleculare, cum ar fi

forma, densitatea electronică, potențialul electrostatic, lipofilitatea și refracția (Allen și colab., 2001).

Pentru programele care rulează deja, căutarea în baza de date a structurii necesare ar putea dura multe zile (Kolodny și colab., 2005). Găsirea mai multor răspunsuri utilizabile poate fi ușoară prin folosirea noilor algoritmi (Dong și colab., 2018).

Acest lucru se realizează folosind o varietate de modele, niște descriptori moleculari (Todeschini și Consonni, 2000), astfel de indici topologici și/sau analize de regresie (Bolboacă și Jäntschi, 2013).

Structurile chimice pot fi clasificate folosind unele criterii de similaritate datorită caracterizării topologice. Unele statistici fundamentale stau la baza analizei de regresie.

3. Relația cantitativă structură-activitate/structură-proprietate (QSAR/QSPR)

Ca instrument matematic pentru caracterizarea cantitativă a relației dintre structura chimică și activitatea/proprietatea biologică pentru un set specific de molecule, conceptul de QSAR/QSPR a apărut în 1937 (Reynolds și colab., 1992; Hammet, 1937).

Studiile relațiilor dintre proprietățile unei structuri au o serie de avantaje. Ecuațiile produse dintr-o investigație de structură-proprietate, de exemplu, pot fi utilizate pentru a estima proprietățile nemăsurate ale substanțelor înrudite. Ecuațiile pot fi folosite pentru a obține o înțelegere mai fundamentală a rolurilor pe care anumite elemente structurale le joacă în determinarea calităților.

După obținerea acestei înțelegeri, datele pot fi utilizate pentru a crea structuri fictive care ar putea avea valori ridicate de proprietate. Ecuațiile structură-proprietate pot fi, de asemenea, utilizate pentru a verifica acuratețea valorilor proprietăților care au fost deja raportate în literatură, dintre care unele ar fi putut fi măsurate sau raportate incorect (Nelson și Seybold, 2001).

Colectarea datelor, selectarea variabilelor, construirea modelului și evaluarea validării sunt de obicei cele patru etape comune utilizate în QSAR/QSPR (Golbraikh și Tropsha, 2000) (Figura 4).

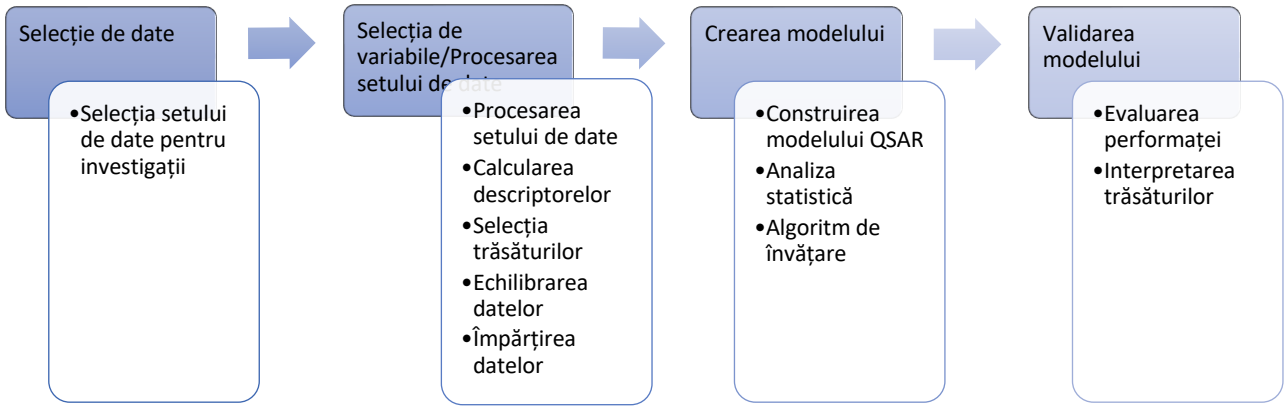

Figura 4. Exemplu de procese în QSAR/QSPR

În plus, există inițiative noi în literatura de specialitate QSAR care vizează următoarele probleme: utilizarea predicțiilor conforme, stabilirea gradului de încredere în predicțiile realizate folosind modele QSAR; să evalueze flexibilitatea seturilor de date pentru a determina dacă este posibil să se creeze modele robuste; crearea de modele QSAR interpretabile pe care biologii și chimiștii medicinali le pot folosi în practică; asigurând că modelele QSAR pot fi replicate, astfel încât alte echipe de cercetare să poată utiliza sau extinde modelele publicate (Nantasenamat, 2020).

Legăturile cantitative structură-activitate-proprietate sau abordări matematice capabile să detecteze și să cuantifice relația dintre structura chimică și activitate/proprietate sunt utilizate atunci când activitatea sau proprietatea este o variabilă cantitativă (modele liniare) sau calitativă (modele neliniare) (Godarzi și colab. 2012). Diferiți descriptori moleculari colectează informațiile structurale (Jäntschi, 2005).

Procesul presupune definirea structurii peptidei la nivel de secvență folosind descriptori de aminoacizi (AAD) și asocierea acesteia cu observații folosind metode de învățare automată (MLM). Rezultatul este o varietate de modele de regresie cantitativă. Aceste modele sunt folosite pentru a construi noi peptide cu caracteristicile dorite și pentru a explica elementele structurale care generalizează proprietățile cunoscute ale peptidelor la mostre necunoscute (Lin și colab., 2023).

Similitudinea chimică încearcă să cuantifice cât de asemănătoare sunt două molecule diferite una cu cealaltă sau cu o anumită caracteristică. Principiul similarității, care susține că compușii similari ar trebui să aibă activități și atribute similare, stă la baza aplicațiilor de evaluare a similarității în domeniile toxicologie și farmaceutice, care urmăresc să prezică toxicitatea substanțelor chimice.

4. Optimizarea moleculară

Este posibil să se optimizeze geometria moleculelor folosind metode *ab initio*, metode semiempirice (pentru a rezolva ecuația Schrödinger cu unele aproximări și pentru a descrie proprietățile electronilor atomilor și moleculelor) și metode empirice ale câmpului de forță (mecanica moleculară, o metodă mai rapidă, dar mai puțin costisitoare) metodă care poate oferi parametri structurali excepționali) (Abegg și Ha, 1974).

Utilizarea seturilor de baze de tip Slater sau a orbitalilor gaussieni pentru a reprezenta funcția de undă a fost pionierat de John A. People (Pople, 1999). El a ales o combinație de abordări și seturi de date, a definit modele și a contrastat rezultatele experimentale ale analizelor.

Calculele funcțiilor de undă atomice și moleculare folosesc frecvent orbitali atomici de tip Gaussian (Figura 5). Ei au contribuit la dezvoltarea programelor gaussiene, unul dintre cele mai utilizate pachete software de chimie computațională.

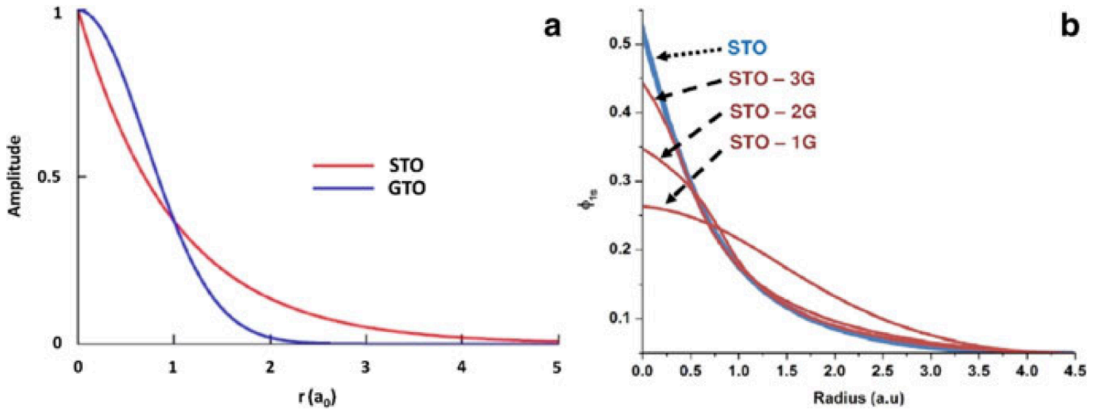

Figura 5. a. Evaluarea orbitalilor de tip Slater și Gaussian; b. Evaluarea schematică a rezultatelor nivelurilor STO-1G, STO-2G și STO-3G pentru potrivirea cu cele mai mici pătrate a funcției 1s Slater (Perlt, 2021).

De fapt, este nevoie de mai puțin timp pentru a calcula mai multe GTO și a le combina pentru a reprezenta un orbital decât pentru a calcula un singur STO. Acesta este motivul din spatele utilizării pe scară largă a combinațiilor GTO pentru a reprezenta STO, care explică ulterior AO.

Seturile de baze indicate prin semnul „*” sunt seturile de baze de polarizare, care conțin orbitalii d. Baza 6-31G** este o îmbunătățire suplimentară, adăugând un set de orbitali p la fiecare hidrogen din setul de baze 6-31G* (Banerjee și Ramalingam, 2015).

Dacă nu este furnizat un acronim al setului de baze, se va folosi baza STO-3G. Seturile de bază STO-3G, 3-21G, 6-21G și 6-31G sunt câteva exemple. Notația standard * sau ** poate fi, de asemenea, utilizată pentru a solicita funcții de prima polarizare unice. 6-31G* (sau 6-31G(d)) este 6-31G cu funcții de polarizare d suplimentare pe atomi non-hidrogen; 6-31G** (sau 6-31G(d, p)) este 6-31G* plus p funcții de polarizare pentru hidrogen. Funcțiile difuze + și ++ pot fi obținute cu unele seturi de bază. 6-31+G este 6-31G plus funcții difuze s și p pentru atomii non-hidrogen; 6-31++G are și funcții difuze pentru hidrogen.

Ce bază stabilită pentru a fi folosită depinde de scopul calculului și de moleculele examinate. Concordanța cu datele experimentale nu este întotdeauna garantată, chiar și cu un set mare de baze (Petersson și colab., 1998).

Diverse abordări ale comparației seturilor de baze (Zheng și colab., 2005; Scuseria, 1992) sunt de acord că, deși sunt comparabile, ele nu pot fi generalizate. Mai multe sugestii pot fi găsite în aceste publicații și pur și simplu prin revizuirea lecțiilor Gaussian09, inclusiv (Tomberg, 2013; Hill, 2012):

- Un set de baze mai mare nu este neapărat mai bun (ex: cc-pVQZ este excesiv pentru Hartree-Fock)
- STO-3G ar trebui aplicat numai sistemelor foarte vaste.
- De obicei cc-pVDZ este comparabil sau mai rău decât 6-31G(d,p).
- De obicei cc-pVTZ este îmbunătățit decât 6-311G(d,p) sau similar.
- Abordările ab initio se stabilesc relativ lent.

Următoarele seturi de baze corespund aproximativ unele cu altele:

- 6-31G ≈ cc-pVDZ
- 6-311G ≈ aug-cc-pVDZ
- 6-31+G(d) ≈ cc-pVTZ
- 6-311+G(d) ≈ aug-cc-pVTZ
- 6-31++G(d,p) ≈ cc-pVQZ
- 6-311++G(d,p) ≈ aug-cc-pVQZ

Analiza clusterului funcțional (FCA), o strategie diferită, ar putea fi utilizată pentru a analiza seturi de date funcționale multidimensionale utilizând funcții de bază gaussiene ortonormalizate (Kayano și colab., 2010). Răspunsul cel mai tipic este de a completa datele experimentale cu cele mai bune date disponibile *ab initio* (din calculele orbitale moleculare sau ale densității funcționale). Abilitatea de a compara locații pe un PES care sunt departe de structurile de simetrie prin calcul drept, spre deosebire de încercarea de a înțelege spectrele vibraționale, este un aspect bun al utilizării datelor teoretice (Schlegel, 2003).

Contribuții personale

Capitolul II - Date cenzurate în calculele de cercetare

1. Introducere

În această cercetare, a fost creată o metodă iterativă care poate localiza cele mai probabile valori ale conținutului fenolic care lipsesc (putere predictivă) după efectuarea studiilor.

Investigațiile compușilor fenolici și flavonoizi datorită activității lor antioxidante fac obiectul multor studii de cercetare (unele exemple sunt Wojdyło și colab., 2007; Yang și colab., 2013; Ivanov și colab., 2014; Aryal și colab., 2014; 2019; etc.). Această transparență este importantă pentru interpretarea și replicarea rezultatelor. Următoarele reprezentări sunt câteva exemple de seturi de date cu valori lipsă:

- Analiza HPLC a conținutului de polifenoli din extractele de etanol din cinci specii de *Fumaria* bulgărească, abrevierea „nd" înseamnă „nedetectat" în Tabelul 1 (Ivanov și colab., 2014):

Tabel 1. Conținut de polifenoli

	Compound*	*F.officinalis*	*F. thuretii*	*F.kralikii*	*F.rostellata*	*F.shrammii*
Flavonoids						
Flavonols	Myricetin	0.25 ± 0.01	0.28 ± 0.01	0.49 ± 0.07	0.17 ± 0.03	0.25 ± 0.01
	Kaempferol	0.08 ± 0.01	0.12 ± 0.01	0.14 ± 0.01	0.06 ± 0.01	0.04 ± 0.01
	Quercetin	0.49 ± 0.03	0.51 ± 0.03	0.36 ± 0.02	0.32 ± 0.02	0.14 ± 0.01
Quercetin glycoside	Rutin	6.47 ± 0.13	nd	4.17 ± 0.07	9.92 ± 0.11	8.39 ± 0.15
	Hyperoside	6.51 ± 0.12	nd	7.58 ± 0.13	1.06 ± 0.03	2.78 ± 0.05
Flavanone glycoside	Hesperidin	nd	0.29 ± 0.01	nd	nd	0.26 ± 0.01
Flavone	Apigenin	0.12 ± 0.02	0.17 ± 0.02	0.38 ± 0.03	0.05 ± 0.01	nd
Phenolic acids						
	p-Coumaric acid	1.10 ± 0.03	0.39 ± 0.05	0.50 ± 0.05	0.55 ± 0.05	0.37 ± 0.04
	Ferulic acid	2.35 ± 0.04	1.74 ± 0.03	1.75 ± 0.03	2.25 ± 0.03	2.00 ± 0.04
	Sinapic acid	0.68 ± 0.02	1.05 ± 0.04	3.03 ± 0.05	0.70 ± 0.02	0.38 ± 0.02

- Conținutul de compuși fenolici în boabe de coacăz (Ribes spp.); abrevierea „nd" înseamnă „nedetectat" în Tabelul 2 (Yang et al., 2013):

Tabel 2. Conținut de compuși fenolici

(mg/100 g fresh berry)

cultivar	growth place[b]	caffeoyl-glucose (Caf-glc)	*p*-coumaroyl-quinic acid (Cou-qa)	*p*-coumaroyl-glucose (Cou-glc)	feruloyl-glucose (Fer-glc)	caffeic acid glucose derivative (Caf glc der)	*p*-coumaric acid glucose derivative (Cou glc der)	ferulic acid glucose derivative (Fer glc der)	myricetin-3-*O*-glucoside (My-glc)	quercetin-3-*O*-rutinoside[c] (Qu-rut)	quercetin-3-*O*-glucoside (Qu-glc)	kaempferol-3-*O*-rutinoside (Ka-rut)	quercetin-3-*O*-(6"-malonyl)-glucoside (Qu-mal)	kaempferol-3-*O*-glucoside (Ka-glc)
'Vertti'	S + N (n = 108)	1.97±0.35	0.40±0.13 c	6.23±1.97	0.59±0.10 c	0.51±0.07 b	1.25±0.17 b	0.42±0.09 b	0.39±0.10	2.76±0.69	2.32±0.56 c	0.76±0.19 c	1.69±0.59 c	0.94±0.18 c
'White Dutch'	S + N (n = 98)	na	0.11±0.02 a	0.97±0.29	0.16±0.04 b	nd a	nd a	nd a	0.07±0.01	0.66±0.27	0.29±0.16 a	0.30±0.08 b	nd a	0.08±0.03 b
'Red Dutch'	S + N (n = 100)	na	0.13±0.03 b	na	0.10±0.02 a	nd a	nd a	nd a	na	0.79±0.39	0.56±0.31 b	nd a	1.00±0.56 b	nd a
'Vertti'	S (n = 52)	1.94±0.36 x	0.43±0.08 y	4.83±1.50 x	0.58±0.09 x	0.52±0.07 y	1.26±0.17 x	0.44±0.07 y	0.40±0.12 x	2.81±0.79 x	2.54±0.45 y	0.72±0.23 x	1.52±0.67 x	0.93±0.20 x
	N (n = 56)	1.99±0.33 x	0.37±0.16 x	7.52±1.37 y	0.59±0.10 x	0.49±0.07 x	1.25±0.17 x	0.39±0.11 x	0.39±0.07 x	2.71±0.59 x	2.12±0.59 x	0.80±0.13 y	1.85±0.44 y	0.96±0.15 x
'White Dutch'	S (n = 52)	nd	0.12±0.02 y	0.82±0.29 x	0.15±0.04 x	nd	nd	nd	0.07±0.01 x	0.64±0.34 x	0.27±0.16 x	0.30±0.10 x	nd	0.08±0.03 x
	N (n = 46)	nd	0.10±0.01 x	1.14±0.18 y	0.18±0.04 y	nd	nd	nd	0.08±0.01 y	0.69±0.17 x	0.32±0.15 x	0.30±0.06 x	nd	0.08±0.02 x
'Red Dutch'	S (n = 46)	nd	0.14±0.03 y	nd	0.09±0.02 x	nd	nd	nd	nd	0.89±0.47 y	0.61±0.41 x	nd	1.04±0.71 x	nd
	N (n = 54)	nd	0.12±0.03 x	nd	0.10±0.02 y	nd	nd	nd	nd	0.71±0.29 x	0.51±0.17 x	nd	0.96±0.40 x	nd

Acestea sunt doar câteva exemple de rezultate cu date lipsă, dar mai multe studii se confruntă cu problema din diferite motive. Rezultatele prezentate de Wojdyło și colab., 2007, datele noastre de intrare, cu locurile lipsă sunt prezentate în Tabelul S3 (secțiunea Material suplimentar).

Rezultatele unei examinări cantitative a principalelor componente fenolice ale celor 32 de plante sunt prezentate în Tabelul S1 (Bálint și Jäntschi, 2019). Abaterile standard calculate în urma studiului sunt legate de valorile medii (Tabelul 3).

Tabel 3. Valorile medii χ^2 și abaterile pentru planta *Acorus calamus.*

Acorus calamus	**χ^2 (Abatere)**	**χ^2 (Medie)**
ABTS	4.6788	0.1057
DPPH	4.8017	0.1045
FRAP	4.8000	0.0989

2. Metodologie

Abordarea Jäntschi (2012) a fost ajustată și schimbată pentru a se potrivi cu rezultatele noastre experimentale. Toate calculele au fost efectuate folosind programe *.php personalizate.

Testul Chi-pătrat (χ^2) a fost folosit pentru a examina legătura dintre patru compuși chimici (acid cafeic, acid p-cumaric, acid ferulic și acid neoclorogenic) și activitatea lor antioxidantă (Bálint și Jäntschi, 2019).

Algoritmul de lucru este reprezentat mai jos (Figura 6):

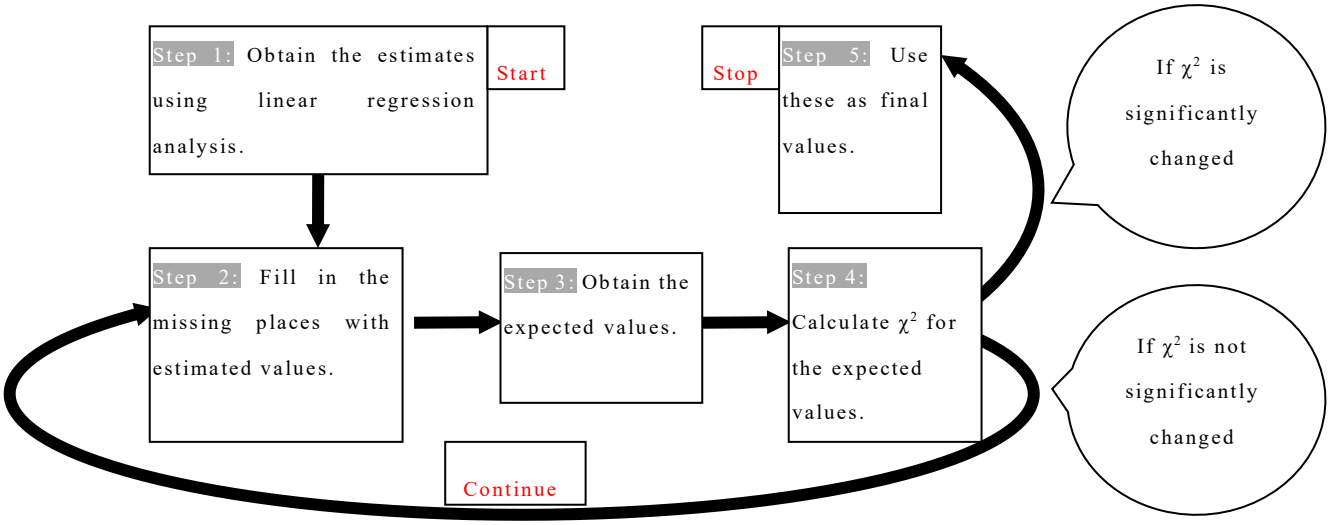

Figura 6. Algoritm de lucru

Algoritmul de lucru:

Procedurile utilizate în examinarea datelor lipsă au fost următoarele (Fig. 6):

- **Pasul 1**: Verificare pentru a vedea dacă legătura dintre activitatea antioxidantă și conținutul de fenol este liniară folosind datele experimentale.
- **Pasul 2**: Au fost luate în considerare trei opțiuni. Prima etapă a algoritmului a inclus introducerea valorilor experimentale.
- **Pasul 3**: Utilizarea coeficienților pentru a crea estimări în primul ciclu al analizei de regresie liniară.
- **Pasul 4**: Cu valori aproximative, se completează spațiile libere.
- **Pasul 5**: Reiterare:
  - Dobândirea (noi) valori probabile.
  - Estimarea χ^2 folosind valorile observate și așteptate.
  - Adăugarea în locurile absente a valorilor (noi) așteptate.
- **Pasul 6**: Până când valoarea lui χ^2 nu este modificată considerabil (de exemplu, convergența)

Acești pași au fost întreprinși pentru a completa golurile din tabelele de contingență pe baza structurii componentei fenolice și a activității sale antioxidante. Ciclurile algoritmului sunt reprezentate grafic în Figura 7. Diferitele combinații de culori implică faptul că valorile din spațiile goale au fost modificate (în roșu).

Un set cuprinzător de estimări, valori prezise și calcule χ^2 sunt, de asemenea, incluse în fiecare ciclu. Modificările datelor din zonele lipsă sunt indicate prin culori diferite; aceste

modificări își ating valorile finale în Ciclul n, după ce χ^2 nu s-a modificat considerabil față de Ciclul n-1.

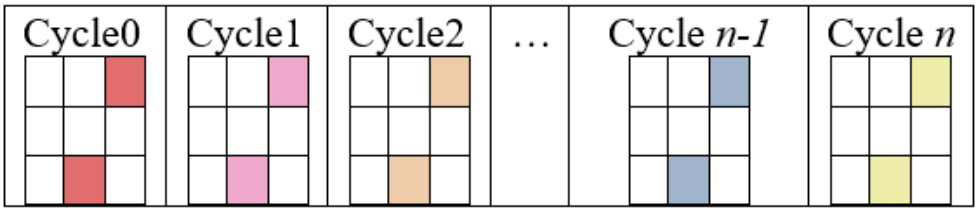

Figura 7. Obținerea valorilor care lipsesc de-a lungul timpului.

Fiecare valoare din coloane are un impact asupra fiecărei valori din rânduri, conform premisei de bază a ipotezei (Bálint și Jäntschi, 2019).

Coeficienții de corelație au fost determinați (Pearson, Spearman, semicantitativ, vezi mai jos) după completarea coloanelor goale. Coeficienții cantitativi ai lui Pearson și ai coeficienților calitativi de rangul lui Spearman se combină pentru a forma coeficientul semicantitativ.

Gradul de inferență monotonă neliniară a fost măsurat folosind acești coeficienți (cum ar fi deviațiile extreme sigmoidale).

3. Rezultate și discuții

Datele dobândite au fost supuse analizei de corelație pentru a determina ce compus fenolic influențează activitatea antioxidantă după ce valorile lipsă au fost completate în tabelul de contingență. Studiul statistic a fost necesar datorită faptului că fiecare acid fenolic aduce o contribuție unică la capacitatea antioxidantă. Următorul tabel conține constatările analizei de corelație.

Deoarece valorile experimentale TEAC (capacitate antioxidantă echivalentă totală) χ^2 au fost valori anormale, acestea nu au fost utilizate în calculele viitoare. Datorită rezultatelor sale aberante χ^2 (Tabelul 3 - vezi mai sus), datele de investigație de la planta *Acorus calamus* au fost omise. A fost examinată legătura dintre cei patru acizi fenolici și activitățile antioxidante ale plantelor rămase.

Valorile rezultatelor experimentale observate (obs.), valorile estimate care au completat golurile și valorile așteptate (exp.) rezultate din regresie sunt prezentate în Tabelul S4 (secțiunea Material suplimentar). Rezultatele analizei după ce datele au fost ajustate logaritmic sunt prezentate prin valori (Bálint și Jäntschi, 2019).

Deoarece atât datele experimentale, cât și valorile prestabilite progresează pe același traseu, trebuie remarcat faptul că valorile fiecărei plante sunt comparabile. Descoperirile sunt neconcludente dacă acidul fenolic prezice cu exactitate activitatea antioxidantă. Fiecare substanță joacă un rol în modul în care este determinat efectul său.

Analiza corelației a fost efectuată după ce rezultatele au fost colectate pentru a determina ce componentă fenolică a influențat activitatea antioxidantă.

Tabelul 4 (vezi mai jos) conține coeficienții de corelație care au fost determinați:

Tabel 4. Coeficienți de corelare

		ABTS	**DPPH**	**FRAP**
Corelația cantitativă a lui Pearson și nivelurile de semnificație din testul Student's t	ABTS	-	0.774	0.758
	DPPH	$4.881 \cdot 10^{-26}$	-	0.669
	FRAP	$1.837 \cdot 10^{-24}$	$1.858 \cdot 10^{-17}$	-
Corelația cantitativă a lui Spearman și nivelurile de semnificație din testul Student's t		ABTS	DPPH	FRAP
	ABTS	-	0.774	0.754
	DPPH	$3.333 \cdot 10^{-26}$	-	0.668
	FRAP	$3.049 \cdot 10^{-24}$	$1.637 \cdot 10^{-17}$	-
Corelația semi-cantitativă și nivelurile de semnificație din testul Student's t		ABTS	DPPH	FRAP
	ABTS	-	0.774	0.756
	DPPH	$4.033 \cdot 10^{-26}$	-	0.669
	FRAP	$2.369 \cdot 10^{-24}$	$1.744 \cdot 10^{-17}$	-

Ei au arătat că există puține distincții între coeficienții de corelație ai lui Pearson și Spearman. Ambele sunt aproape la fel de importante. Singura distincție este că corelația lui Spearman folosește ranguri, decât valorile x și y ale lui Pearson.

Coeficienții de corelație din tabelul anterior au arătat o legătură semnificativă între rezultate. În acest caz, valoarea medie de 0,75 a arătat că există o creștere liniară a relației dintre variabile. Coeficienții pot lua valori între -1 și +1.

Variabilele au o relație liniară semnificativă statistic, conform testului t Student. Fiecare coeficient de corelație descrie legătura dintre două variabile și explică o măsură a asocierii dintre ele. Fiecare coeficient se modifică în același mod atunci când variația dintre ele nu este mare.

Când valorile lipsă nu lipsesc la întâmplare (MNAR), testul χ^2 poate fi utilizat pentru a analiza datele. Valorile lipsă în această instanță pot fi conectate la valoarea în sine sau la alte variabile ale setului de date. Testul χ^2 poate fi utilizat pentru a evalua dacă există o legătură semnificativă statistic între datele lipsă și celelalte variabile prin compararea distribuției valorilor lipsă în mai multe grupuri de variabile (Agresti, 2007).

În urma execuției algoritmului, a fost examinată și relația dintre χ^2 și iterație. Acest lucru a demonstrat, la fel ca analiza statistică, cât de strâns legate sunt variabilele. Următoarea Figura 8 prezintă evoluția lui χ^2 în funcție de iterație.

Valorile rezultate ale χ^2 rapid s-au atins la minim după implicarea procedurii pe setul de date experimentale după un număr diferit de cicluri.

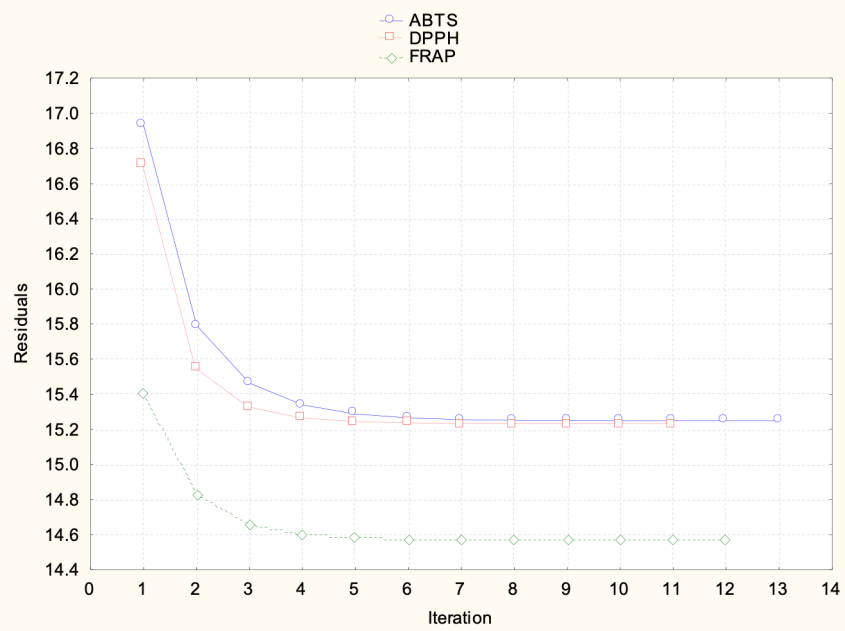

Figura 8. χ^2 ca și funcție de iterație

În câteva iterații, minimul a fost atins. Procedura a fost oprită la a 12-a iterație pentru ABTS, a 10-a iterație pentru DPPH și a 11-a iterație pentru valorile FRAP utilizate în calcule din cauza unei modificări nesemnificative între rezultatele χ^2 ulterioare (Bálint și Jäntschi, 2019).

Estimând frecvențele prezise ale datelor lipsă folosind frecvențele observate în datele disponibile, testul χ^2 poate fi folosit și pentru a imputa valorile lipsă. Termenul „Imputare Chi-pătrat" se referă la această metodă. Testul χ^2 poate fi apoi utilizat pentru a stabili dacă datele lipsă sunt conectate la celelalte variabile în acea situație (McDonald, 2014).

Potrivirea modelelor de calcul la datele experimentale, cum ar fi cele din investigațiile de andocare și legare moleculară, este evaluată folosind testul χ^2. În plus, este folosit pentru a determina cât de asemănătoare sunt diferitele molecule între ele și pentru a optimiza parametrii modelului de calcul.

În literatura științifică, s-a remarcat că există mai multe modalități de a completa tabelul de contingență (este un tip specific de tabel, tip matrice care arată distribuția de frecvență multivariată a variabilelor), cum ar fi tehnicile Monte Carlo. Majoritatea simulărilor Monte Carlo încep prin a propune în mod repetat o modificare aleatorie minoră a unei configurații (Walter și Barkema, 2015).

Metoda Bootstrap, una dintre tehnicile Monte Carlo, ar funcționa cu colecția noastră de date. Selectând o varietate aleatorie de seturi, din setul de date, bootstrapping-ul încearcă să imite impactul utilizării unui set de date mai mare. Unele dintre datele din fiecare set ales aleatoriu vor apărea de mai multe ori, în timp ce alte date nu vor fi prezente deloc (Tropsha, 2006). Această metodă este foarte asemănătoare cu abordarea noastră.

Rezultatele metodelor Monte Carlo, atunci când sunt luate în considerare, nu ar furniza informații noi bazate pe variabile. Abordarea Bootstrap are, de asemenea, dezavantajul de a ignora variabilele cunoscute, chiar și atunci când există o asociere liniară între ele.

Cu metoda propusă, tabelul de contingență este inițial populat, iar apoi variabilele sunt procesate de χ^2. Tabelul este populat în mod repetat folosind metodele Monte Carlo, modificând variabilele inițiale.

Pe același subiect, numeroase studii (Wojdyło și colab., 2007; Aaby și colab., 2004) au produs rezultate complexe și variate. Compușii determinați diferă în ceea ce privește geometria și simetria, precum și proprietățile. Găsirea celei mai bune alegeri este cheia pentru rezolvarea problemei datelor cenzurate. Deoarece oferă cea mai rapidă cale de rezolvare a problemei, am folosit metodele noastre pentru a completa tabelul de contingență.

Numeroase domenii științifice folosesc algoritmul precedent atunci când afișează datele lipsă din studii (Ivanova și colab., 2005; Luo și colab., 2004).

Studiile clinice și cercetările epidemiologice folosesc frecvent date cenzurate. Acestea apar în cercetările la scară largă, unde apariția este adesea legată de o boală, infecție sau alt eșec (Yen și colab., 1995; Arnao, 2000). Promptitudinea și rapiditatea algoritmului sunt calități esențiale pentru originalitatea studiului noastre.

4. Concluzii

Pe un tabel de contingență cu date lipsă, algoritmul nostru și-a demonstrat capacitatea de a funcționa. Statistica χ^2 este minimizată în timpul procesului.

Toate seturile de date investigate arată o conexiune liniară. Rezultatele tuturor celor trei tehnici - ABTS, DPPH și FRAP - utilizate pentru a evalua capacitatea antioxidantă au fost echivalente. Pe baza rezultatelor experimentale și a recenziilor literaturii, aceasta este ceea ce se anticipează. Valorile proiectate pentru spațiile goale se potrivesc cu rezultatele experimentului.

Capitolul III - Evaluarea similarității folosind funcția polinomului caracteristic (ChP)

1. Introducere

Reprezentările proteinelor folosesc fie caracterizări care sunt insensibile la ordinea de aminoacizi care a fost aleasă, fie metode care tratează orice presupusă atribuire a aminoacizilor ca fiind strict echivalentă.

Valorile proprii ale matricei și coeficienții polinomului caracteristic sunt cele mai frecvente două invariante ale matricelor. Totalul elementelor matricei de deasupra diagonalei principale au fost utilizați ca descriptori moleculari în aplicații chimice. Valorile proprii și coeficienții polinomului caracteristic, ambii formând în mod natural o secvență ordonată, au ca rezultat invarianți suplimentari care sunt furnizați ca o secvență ordonată (Randic și colab., 2008).

Obținem un grup de polinoame și un grup de caracteristici compuse în urma calculelor pe cei 10 aminoacizi (Material suplimentar - Tabelul S1).

Acest polinom poate fi folosit pentru a calcula diferiți descriptori moleculari, cum ar fi indicele Wiener (Todeschini și Consonni, 2000) și indicele Randic (Randic, 1975), care sunt utilizați pentru a evalua similaritatea topologică dintre molecule.

2. Metodologie

Calcularea valorilor proprii ale matricelor aferente este necesară pentru a determina ecuațiile polinomiale caracteristice pentru fiecare dintre cei 10 aminoacizi esențiali.

Matricele utilizate pentru a reprezenta aminoacizii pot fi obținute dintr-o varietate de caracteristici structurale sau chimice, cum ar fi coordonatele tridimensionale (3D) ale atomilor sau structura electronică a lanțurilor laterale. În urma creării matricelor, ecuația polinomială caracteristică a fost găsită folosind metode convenționale de algebră liniară.

Aminoacizii prezentați anterior sunt compuși bazici în sistemele biologice. Scopul a fost de a determina dacă structurile chimice etichetate diferit, diferă într-adevăr sau nu. Toate structurile au fost colectate din bazele de date PubChem și alese în funcție de complexitatea și numărul de izomeri (PubChem Database).

În urma selecției, a fost utilizat următorul algoritm:

- **Pasul 1**: colectarea tuturor structurilor de aminoacizi cu aceeași greutate moleculară din bazele de date PubChem.

- **Pasul 2**: convertirea structurilor .sdf în fișiere .hin cu un program *php de casă.
- **Pasul 3**: introducerea fișierelor .hin în programul de pe http://l.academicdirect.org, pentru a calcula ecuațiile polinomiale caracteristice, după următoarea cale:
  - Introducerea structurilor pe site → Fundamentals → Graphs → polynomials → a_characteristic_polynomial_in.
- **Pasul 4**: colectarea ecuațiilor din matricele date de program.
- **Pasul 5**: analiza ecuațiilor colectate (sortare, grupare).
- **Pasul 6**: discuții referitor la datele obținute.

Am selectat 10 din 20 de aminoacizi esențiali pe baza structurii chimice: alanină, glicină, valină, leucină, izoleucină, lizină, serină, treonină, aspartat și glutamat.

3. Rezultate și discuții

Datele au fost sortate după analiza aminoacizilor și după obținerea ecuațiilor polinomiale caracteristice. Ecuațiile derivate din matricele polinomiale caracteristice care descriu asemănarea dintre molecule sunt prezentate în Tabelul 5 de mai jos.

Tabel 5. Ecuațiile obținute din ChP

Aminoacizi	***ChP Ecuații***
Glycine_ZW_5257127	$=+1X^{10}-9X^{8}+21X^{6}-12X^{4}$
Glycine_750	$=+1X^{10}-9X^{8}+23X^{6}-19X^{4}+4X^{2}$
Alanin_D_71080	$=+1X^{13}-12X^{11}+47X^{9}-73X^{7}+40X^{5}-6X^{3}$
Alanin_DL_602	$=+1X^{13}-12X^{11}+47X^{9}-73X^{7}+40X^{5}-6X^{3}$
Alanin_L_5950	$=+1X^{13}-12X^{11}+47X^{9}-73X^{7}+40X^{5}-6X^{3}$
Alanin_Beta_239	$=+1X^{13}-12X^{11}+47X^{9}-73X^{7}+44X^{5}-8X^{3}$
Serine_D_ZW_6857549	$=+1X^{14}-13X^{12}+56X^{10}-97X^{8}+62X^{6}-12X^{4}$
Serine_ZW_6857552	$=+1X^{14}-13X^{12}+56X^{10}-97X^{8}+62X^{6}-12X^{4}$
Serine_D_71077	$=+1X^{14}-13X^{12}+58X^{10}-112X^{8}+95X^{6}-34X^{4}+4X^{2}$
Serine_DL_617	$=+1X^{14}-13X^{12}+58X^{10}-112X^{8}+95X^{6}-34X^{4}+4X^{2}$
Serine_L_5951	$=+1X^{14}-13X^{12}+58X^{10}-112X^{8}+95X^{6}-34X^{4}+4X^{2}$
Aspartate_D_83887	$=+1X^{16}-15X^{14}+82X^{12}-209X^{10}+262X^{8}-157X^{6}+42X^{4}-4X^{2}$
Aspartate_DL_424	$=+1X^{16}-15X^{14}+82X^{12}-209X^{10}+262X^{8}-157X^{6}+42X^{4}-4X^{2}$
Aspartate_L_5960	$=+1X^{16}-15X^{14}+82X^{12}-209X^{10}+262X^{8}-157X^{6}+42X^{4}-4X^{2}$
Threonine_D_ZW_6995277	$=+1X^{17}-16X^{15}+92X^{13}-238X^{11}+281X^{9}-132X^{7}+18X^{5}$
Threonine_L_ZW_6971019	$=+1X^{17}-16X^{15}+92X^{13}-238X^{11}+281X^{9}-132X^{7}+18X^{5}$
Threonine_D_69435	$=+1X^{17}-16X^{15}+94X^{13}-259X^{11}+353X^{9}-229X^{7}+64X^{5}-6X^{3}$

Threonine_D_allo_90624	$=+1X^{17}-16X^{15}+94X^{13}-259X^{11}+353X^{9}-229X^{7}+64X^{5}-6X^{3}$
Threonine_DL_205	$=+1X^{17}-16X^{15}+94X^{13}-259X^{11}+353X^{9}-229X^{7}+64X^{5}-6X^{3}$
Threonine_L_6288	$=+1X^{17}-16X^{15}+94X^{13}-259X^{11}+353X^{9}-229X^{7}+64X^{5}-6X^{3}$
Threonine_L_allo_99289	$=+1X^{17}-16X^{15}+94X^{13}-259X^{11}+353X^{9}-229X^{7}+64X^{5}-6X^{3}$
Glutamate_Hy_4525487	$=+1X^{18}-17X^{16}+106X^{14}-305X^{12}+418X^{10}-248X^{8}+48X^{6}$
Valine_2S_6971018	$=+1X^{19}-18X^{17}+120X^{15}-372X^{13}+543X^{11}-324X^{9}+54X^{7}$
Valine_D_ZW_6971095	$=+1X^{19}-18X^{17}+120X^{15}-372X^{13}+543X^{11}-324X^{9}+54X^{7}$
Valine_3amino_2760933	$=+1X^{19}-18X^{17}+122X^{15}-397X^{13}+641X^{11}-461X^{9}+96X^{7}$
Valine_D_iso_6971276	$=+1X^{19}-18X^{17}+122X^{15}-397X^{13}+646X^{11}-483X^{9}+120X^{7}$
Valine_L_iso_2724877	$=+1X^{19}-18X^{17}+122X^{15}-397X^{13}+646X^{11}-483X^{9}+120X^{7}$
Valine_D_71563	$=+1X^{19}-18X^{17}+122X^{15}-397X^{13}+649X^{11}-507X^{9}+168X^{7}-18X^{5}$
	$=+1X^{19}-18X^{17}+122X^{15}-397X^{13}+649X^{11}-507X^{9}+168X^{7}-18X^{5}$
Valine_DL_1182	$=+1X^{19}-18X^{17}+122X^{15}-397X^{13}+649X^{11}-507X^{9}+168X^{7}-18X^{5}$
Valine_L_6287	
Glutamate_D_23327	$=+1X^{19}-18X^{17}+124X^{15}-422X^{13}+766X^{11}-746X^{9}+376X^{7}-90X^{5}+8X^{3}$
Glutamate_DL_611	$=+1X^{19}-18X^{17}+124X^{15}-422X^{13}+766X^{11}-746X^{9}+376X^{7}-90X^{5}+8X^{3}$
Glutamate_L_33032	$=+1X^{19}-18X^{17}+124X^{15}-422X^{13}+766X^{11}-746X^{9}+376X^{7}-90X^{5}+8X^{3}$
Leucine_ZW_7045798	$=+1X^{22}-21X^{20}+171X^{18}-690X^{16}+1458X^{14}-1545X^{12}+702X^{10}-108X^{8}$
Isoleucine_L_zw_7043901	$=+1X^{22}-21X^{20}+171X^{18}-690X^{16}+1458X^{14}-1560X^{12}+738X^{10}-108X^{8}$
Leucine_D_tert_6950340	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1615X^{14}-1878X^{12}+981X^{10}-162X^{8}$
Leucine_DL_tert_306131	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1615X^{14}-1878X^{12}+981X^{10}-162X^{8}$
Leucine_L_tert_164608	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1615X^{14}-1878X^{12}+981X^{10}-162X^{8}$
Leucine_D_439524	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1633X^{14}-1998X^{12}+1239X^{10}-354X^{8}+36X^{6}$
Leucine_DL_857	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1633X^{14}-1998X^{12}+1239X^{10}-354X^{8}+36X^{6}$
Leucine_L_6106	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1633X^{14}-1998X^{12}+1239X^{10}-354X^{8}+36X^{6}$
Isoleucine_D_76551	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1633X^{14}-2010X^{12}+1272X^{10}-366X^{8}+36X^{6}$
Isoleucine_D_alloiso_94206	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1633X^{14}-2010X^{12}+1272X^{10}-366X^{8}+36X^{6}$

Isoleucine_DL_791	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1633X^{14}-2010X^{12}+1272X^{10}-366X^{8}+36X^{6}$
Isoleucine_L_6306	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1633X^{14}-2010X^{12}+1272X^{10}-366X^{8}+36X^{6}$
Isoleucine_L_alloiso_99288	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1633X^{14}-2010X^{12}+1272X^{10}-366X^{8}+36X^{6}$
Isoleucine_poly_5351546	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1633X^{14}-2010X^{12}+1272X^{10}-366X^{8}+36X^{6}$
Leucine_D_nor_456468	$=+1X^{22}-21X^{20}+173X^{18}-721X^{16}+1642X^{14}-2061X^{12}+1368X^{10}-432X^{8}+48X^{6}$
Lysine_beta_392	$=+1X^{24}-23X^{22}+213X^{20}-1033X^{18}+2866X^{16}-4675X^{14}+4442X^{12}-2348X^{10}+624X^{8}-64X^{6}$
Lysine_D_beta_10931575	$=+1X^{24}-23X^{22}+213X^{20}-1033X^{18}+2866X^{16}-4675X^{14}+4442X^{12}-2348X^{10}+624X^{8}-64X^{6}$
Lysine_L_beta_439417	$=+1X^{24}-23X^{22}+213X^{20}-1033X^{18}+2866X^{16}-4675X^{14}+4442X^{12}-2348X^{10}+624X^{8}-64X^{6}$
Lysine_D_57449	$=+1X^{24}-23X^{22}+213X^{20}-1033X^{18}+2868X^{16}-4689X^{14}+4476X^{12}-2388X^{10}+640X^{8}-64X^{6}$
Lysine_DL_866	$=+1X^{24}-23X^{22}+213X^{20}-1033X^{18}+2868X^{16}-4689X^{14}+4476X^{12}-2388X^{10}+640X^{8}-64X^{6}$
Lysine_L_5962	$=+1X^{24}-23X^{22}+213X^{20}-1033X^{18}+2868X^{16}-4689X^{14}+4476X^{12}-2388X^{10}+640X^{8}-64X^{6}$
Lysine_L_ZW_5962	$=+1X^{24}-23X^{22}+213X^{20}-1033X^{18}+2868X^{16}-4689X^{14}+4476X^{12}-2388X^{10}+640X^{8}-64X^{6}$

Descoperirile au arătat că aproximativ 95% dintre conformerii diferiților aminoacizi sunt identici. De cele mai multe ori, modificarea are loc între zwitterion și alți conformeri, dar ocazional pot fi observate diferite variații.

În urma procesului de sortare a datelor, zonele „se amestecă” și devine evident că leucina, izoleucina și un tip de glutamat sunt separate de alți conformeri. Ecuațiile lor polinomiale distincte completează mai eficient alte molecule sau conformeri.

Polinomul caracteristic este același pentru două matrici comparabile. Dimpotrivă, nu este întotdeauna cazul: două matrice cu același polinom caracteristic nu trebuie neapărat să fie identice.

Este important să precizăm, că polinomul caracteristic ar putea să nu fie cel mai bun mod de a compara aminoacizii. Este folosit mai ales pentru a calcula valorile proprii, care sunt caracteristici matematice ale unei matrice care ar putea să nu aibă o semnificație biologică sau chimică clară. Pentru cercetarea caracteristicilor și interacțiunilor aminoacizilor, alte

metode ar putea fi mai potrivite, inclusiv compararea secvențelor de aminoacizi sau analiza similarității structurii.

Pan (2012) a investigat condițiile în care matricele identice ale polinoamelor caracteristice pot asemăna între ele. El a concluzionat că: dacă cele două matrice de ordinul n (A și B) în cantitatea câmpul F au polinoame caracteristice care sunt identice, au n rădăcini simple și sunt identice cu A și B; setați A și B ca două matrici pătrate de ordin n în câmpul F, ca rezultat, A și B sunt comparabile. Dacă cerința (ca rădăcinile distinctive ale lui A și B să fie în F) poate fi îndeplinită.

În studiul lui Garcia-Planas (2021), datorită aplicațiilor lor științifice, sunt luate în considerare perechile de matrice aflate sub similitudine. Obiectivul principal al acestei lucrări este de a crea conexiuni între geometria limitată care înconjoară un punct și geometria locală din jurul altui punct, folosind polinomul caracteristic asociat cu fiecare matrice a perechii.

4. Concluzii

Studiul a determinat ecuațiile polinomiale distinctive pentru cei 10 aminoacizi esențiali folosind o matrice formată din structura 3D a aminoacizilor.

S-a descoperit că fiecare aminoacid are o ecuație polinomială distinctă care ar putea fi utilizată pentru a face o distingere între aminoacizi.

Scopul studiului a fost de a compara structurile celor 10 aminoacizi esențiali în funcție de funcția lor polinomială caracteristică derivată.

Polinomul caracteristic a fost utilizat în compararea structurilor moleculare în general, deoarece oferă o reprezentare unică a structurii care este invariantă sub diferite transformări. Putem afirma pentru faptul că nu există diferențe apreciabile între conformerii care au fost examinați.

Capitolul IV – Evaluarea similarității folosind softul *Gaussian09*

1. Introducere

Rezolvarea ecuației Schrödinger este problema cheie în teoria structurii electronice. Cu câteva excepții, cum ar fi atomul de hidrogen, este o problemă cu mai multe straturi, ceea ce înseamnă că trebuie calculate soluții numerice. Funcția de undă este mărită în relațiile unui set de baze pentru această utilizare. Un punct esențial pentru a începe în chimia cuantică moleculară este cu seturile de baze derivate din orbitali centrați pe atom.

Din acest motiv, undele plane sunt o alegere destul de comună pentru funcțiile de bază ale acestor sisteme. Cu toate acestea, funcțiile de bază locale sunt, de asemenea, utilizate în mod obișnuit (Perlt, 2021).

Găsirea celei mai bune strategii de optimizare pentru calculele științifice este cu adevărat dificilă, deoarece există atât de multe seturi de baze și tehnici de optimizare diferite.

După consultarea literaturii de specialitate, au fost formulate următoarele întrebări de cercetare:

1. Cu cât seturile de baze sunt mai mari, cu atât mai bine?
2. Care familie de set de baze este cea mai bună? Este posibil sa găsim unul?
3. Obținem rezultate diferite dacă folosim două seturi de baze separate din familii diferite?
4. Pot fi corelate în orice fel?
5. Diferența depinde de moleculă sau de setul de baze?

Scopul studiului a fost de a examina 39 de tehnici de optimizare pentru a le descoperi relațiile și pentru a alege cea mai bună pe care să o aplice în diferite situații (Bálint și Jäntschi, 2021).

Pentru a compara similitudinea diferitelor abordări, au fost efectuate analize cluster, analiza corelației, analiza statistică (ANOVA) și analiza componentelor principale (PCA).

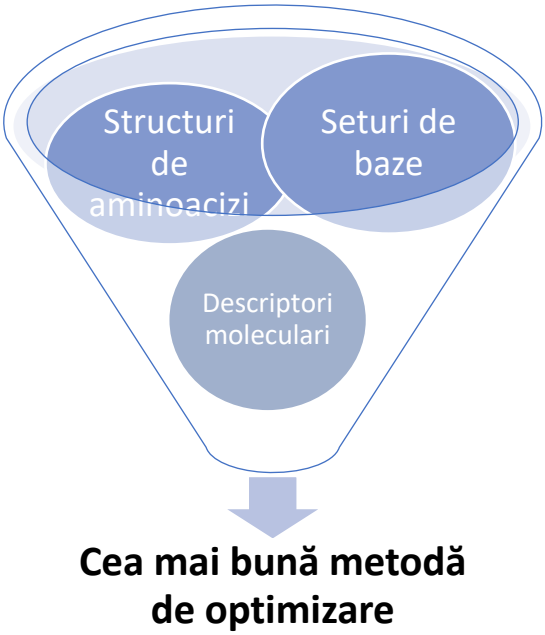

Figura 9. Reprezentarea principalelor componente ale analizei

Calculele au fost efectuate folosind un program software de chimie computațională numit *Gaussian09*, care în general este folosit pentru a simula sisteme chimice și a calcula structurile electrice ale acestora.

Gaussian09 conține diverse modele de structură electronică, cum ar fi tehnici post-Hartree-Fock, Teoria funcțională a densității și abordările Hartree-Fock; o serie de seturi de baze care permit utilizatorilor să decidă cât de precise și de costisitoare ar trebui să fie calculele lor; capacitatea de a efectua căutări de stare de tranziție și optimizare a geometriei și mult mai multe caracteristici (Bálint și Jäntschi, 2021).

2. Metodologie

Cei 20 de aminoacizi esențiali analizați (material suplimentar – tabelul S1 și S2) pot fi izomeri, enantiomeri și conformeri printre alte forme. Cele mai multe dintre aceste substanțe chimice au aceeași chiralități în sistemele biologice, iar majoritatea aminoacizilor sunt levogiri (L) mai degrabă decât dextrogiri(D). Optimizările geometriei au fost efectuate asupra structurilor folosind conformerul L al acestor compuși (Tabelul 6).

Tabel 6. Metode de optimizare a geometriei utilizate în calcule.

Optimizări moleculare Gaussian	
1. Semiempirical Methods (Default Spin)	➢ Parameterized Model 6 - PM6 (opt-pm6) ➢ Austin Model 1 - AM1 (opt-am1) ➢ Parameterized Model 3 - PM3 (opt-pm3) ➢ Parameterized Model 3 (Molecular Mechanics correction) - PM3MM (opt-pm3mm) ➢ Pairwise Distance Directed Gaussian function - PDDG (opt-pddg) ➢ Complete Neglect of Differential Overlap – CNDO (opt-cndo) ➢ Intermediate Neglect of Differential Overlap – INDO (opt-indo)
2. Density Functional Theory (Default Spin)	➢ Becke(three-parameter)-Lee-Yang-Parr (functional) - B3LYP (opt-b3lyp-sto-3g; opt-b3lyp-3-21g; opt-b3lyp-6-31g; opt-b3lyp-6-311g; opt-b3lyp-cc-pvdz;) ➢ Local Spin Density Approximation - LSDA (opt-lsda-3-21g; opt-lsda-sto-3g; opt-lsda-cc-pvdz; opt-lsda-6-311g; opt-lsda-6-31g;) ➢ Perdew–Burke-Ernzerhof (functional) – PBEPBE (opt-pbepbe-sto-3g) ➢ BVP86 (opt-bvp86-sto-3g; opt-bvp86-3-21g; opt-bvp86-6-31g; opt-bvp86-6-311g;) ➢ B3PW91 (opt-b3pw91-sto-3g; opt-b3pw91-6-31g; opt-b3pw91-6-311g;)
3. Møller–Plesset perturbation theory	➢ MP2 (opt-mp2-sto-3g; opt-mp2-3-21g; opt-mp2-6-31g; opt-mp2-6-311g; opt-mp2-cc-pvdz;)
4. Coupled-cluster theory	➢ Coupled Cluster single-double – CCSD (opt-ccsd-sto-3g)
5. Molecular Mechanics (Default Spin)	➢ Universal Force Field - UFF (opt-uff) ➢ Dreiding (opt-dreiding)
6. Hartree-Fock (Default Spin)	➢ STO-3G (opt-hf-sto-3g) ➢ 3-21G (opt-hf-3-21g) ➢ 3-21G* (opt-hf-3-21g*) ➢ 6-31G (opt-hf-6-31g) ➢ 6-311G (opt-hf-6-311g) ➢ CC-pvdz (opt-hf-cc-pvdz)

Un grup de descriptori moleculari (FMPI- Fragmental Matrix Property Indices) (Jäntschi și Bolboaca, 2016) pentru a evalua nivelul de similitudine dintre metode au fost, de asemenea, creat în urma calculelor programului *Gaussian09* pe baza celor 39 de metode alese anterior (Bálint și Jäntschi, 2021).

După cum se arată, după Figura 10, am obținut structurile tridimensionale (3D) ale celor 20 de aminoacizi (conformeri L) din bazele de date PubChem (fișiere .sdf) și am folosit programul *Gaussian09* pentru a le analiza după următorul protocol:

- **Pasul 1**: Introducerea fișierelor .sdf , descărcate din PubChem în *Gaussian09*.
- **Pasul 2**: Convertirea structurii în format .gjf (formatul de intrare pentru baza de date)
- **Pasul 3**: Analiza structurilor după următoarea instrucțiune:
  - Calculate → *Gaussian09* Calculation Setup → Job type (Optimization) → Method (Ground State) → Set the chosen method (HF, DFT, etc…) → Submit the job (Figure 10)
- **Pasul 4**: De asemenea, am creat un fișier .bcf pentru automatizarea procesului de optimizare.
- **Pasul 5**: Rularea calculelor selectând pe rând fiecare metodă de optimizare a geometriei *Gaussian09* din meniul Configurare calcul.
- **Pasul 6**: Salvarea fișierului .out (formatul fișierului de ieșire pentru program) după fiecare calcul.

Următoarea figură ilustrează fluxul de lucru (Figura 10) care a fost urmărit.

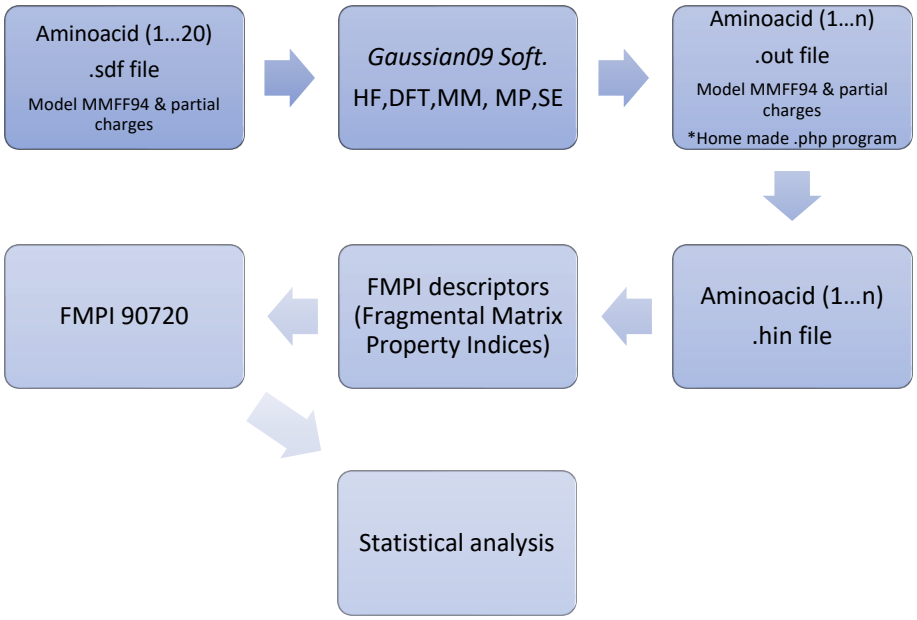

Figura 10. Algoritmul fluxului de lucru principal

Descriptorii moleculari (FMPI) au fost generați și fișierele .hin din fișierele .out și .sdf folosind un software *php personalizat.

Programul *Statistica* a efectuat analiza statistică odată ce am obținut cele 90720 de descriptori pentru fiecare aminoacid analizat.

Au fost efectuate o analiză PCA, grupare, ANOVA și corelare. Cadrul literaturii oferă o descriere amănunțită a PCA. Fiecare analiză statistică care a fost utilizată are un obiectiv comun din mai multe unghiuri.

Rezultatele așteptate au fost următoarele:

- Pentru a obține rezultate specifice pentru fiecare moleculă optimizată.
- Pentru a obține familia de descriptori moleculari pentru fiecare metodă aplicată.
- Să se obțină rezultate statistice pentru metodele aplicate.
- Să se găsească cea mai mare metodă de Optimizare Moleculară de practicat în anumite condiții pe baza descriptorilor moleculari obținuți.

3. Rezultate și discuții

Numărul de molecule sau variabile (descriptori) din setul de date inițial, oricare dintre acestea este mai mic, este numărul de componente principale care pot fi estimate.

Pentru a descrie pe deplin întreaga variație a datelor, trebuie de obicei luate în considerare toate componentele principale. Datorită corelațiilor dintre variabilele inițiale, multe seturi de date au nevoie doar de un număr mic de componente principale pentru a descrie o parte considerabilă a variației.

Variabilele din întregul set de date sunt combinate liniar pentru a forma fiecare componentă principală. Au fost examinați 3.538.080 de descriptori în total, 90720 descriptori per componentă și tehnică (Bálint și Jäntschi, 2021).

Rezultatele examinării PCA au arătat că componentele principale (Figura 11) au reprezentat cu acuratețe variația cantității noastre enorme de date, explicând 99,8851% din aceasta.

Prima și a doua componentă sunt afișate în figura următoare (Figura 11) ca rezultat al analizei PCA.

Cea mai mare valoare proprie este asociată cu prima componentă principală, care reprezintă cea mai mare proporție a variației; a doua cea mai semnificativă valoare proprie este legată de a doua componentă principală și așa mai departe. Valorile proprii afișează proporția de variație la care contribuie fiecare componentă majoră.

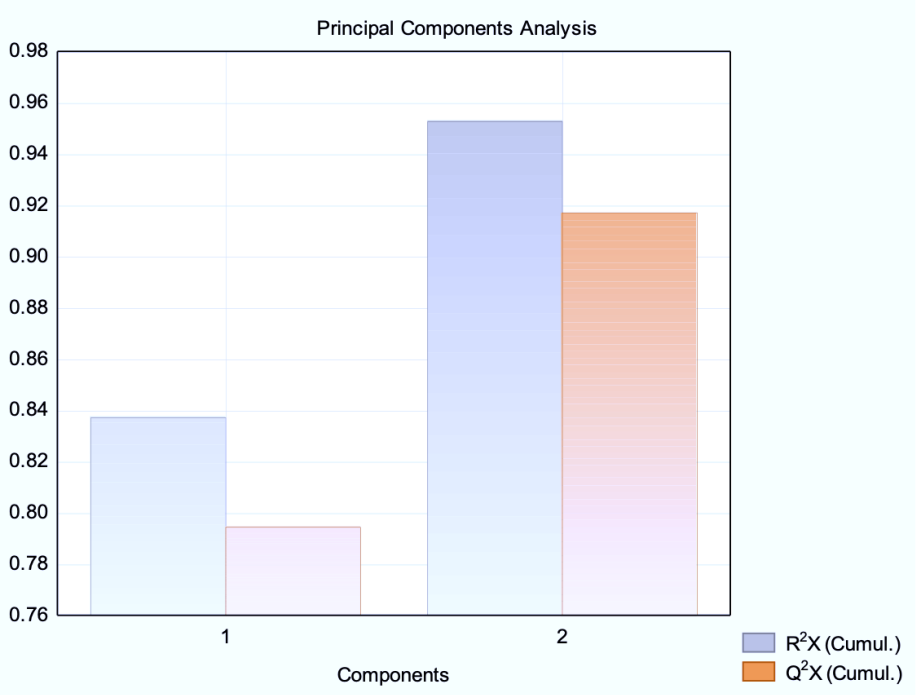

Figura 11. Variația explicată (R^2X) și variația predictivă (Q^2X) a elementelor PCA

Prima componentă explică în cea mai mare parte varianța în general (71,25%) în datele studiate. Variația maximă (14,9%) neacoperită de prima componentă este acoperită de a doua componentă. După primele două, al treilea constituent explică și cea mai mare variață (6,51%).

R^2X folosește valori între 0 și 1 pentru a descrie acuratețea predictivă. R^2X al unei componente principale crește odată cu relevanța sa. Mai simplu spus, varianța explicată (R^2X_{adj}) este varianța explicată R^2X cu gradele de libertate luate în considerare.

Validarea încrucișată este frecvent utilizată pentru a prezenta statistica de evaluare a calității, bunătatea predicției (Q^2), care oferă o măsură calitativă a coerenței între datele prezise și cele originale. Valoarea Q^2 crește pe măsură ce variabilele suplimentare sunt incluse în analiza PCA. Valorile mari ale Q^2 sugerează că analiza a fost semnificativă și pertinentă.

Următoarele diagrame prezentate afișează coeficienții pentru fiecare dintre descriptori din diferitele componente principale. Acest lucru demonstrează modul în care fiecare descriptor contribuie diferit la diferitele componente principale (Bálint și Jäntschi, 2021).

O diagramă care arată distribuția componentei 1 față de componenta 2 este prezentată în figura următoare (Figura 12). Distribuția sugerează că tehnicile similare sunt de fapt grupate împreună.

Orientarea spațială a componentelor primare este determinată și de încărcări. Vectorii sunt p1 și p2.

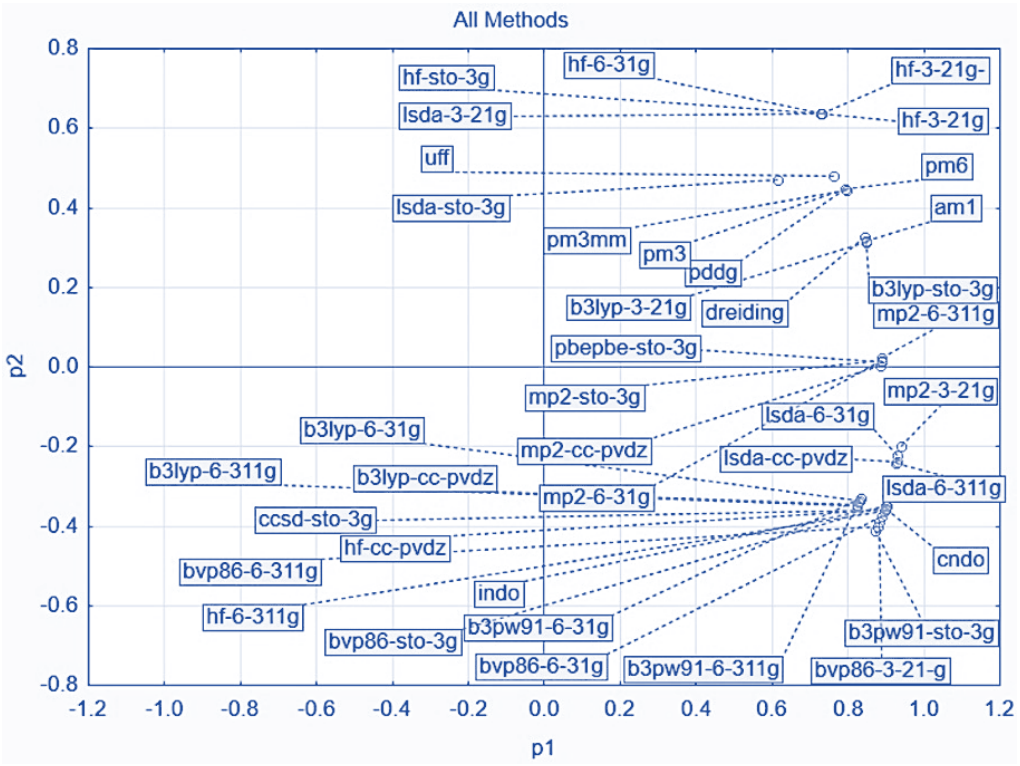

Figura 12. Graficul prezintă distribuția tehnicilor în componentele principale p1 și p2.

În cazul nostru, primele trei componente au reprezentat majoritatea datelor. Graficul anterior afișează distribuția componentei 3 cu componenta 2 (Figura 13).

Dacă se ține cont de cât de asemănătoare sunt, clasificarea metodelor în diferite categorii - Semiempirică, Teoria funcțională a densității, Mecanica moleculară, Teoria perturbației Møller-Plesset, Teoria clusterelor cuplate și Hartree-Fock – diferă în câteva puncte de secțiunea Materiale și Metode.

Pe baza cât de mult semănau abordările între ele, datele au fost împărțite în diferite categorii. Rezultatele analizei PCA și ale clusterelor au fost echivalente.

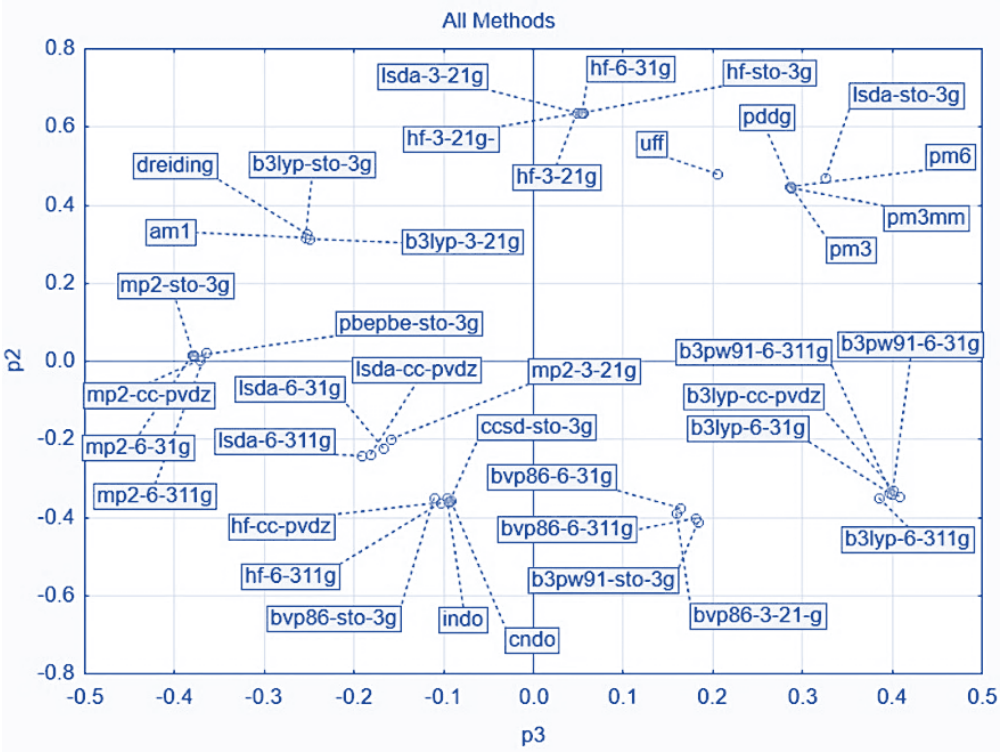

Figura 13. Graficul prezintă distribuția tehnicilor în componentele principale p2 și p3.

Distanțele euclidiene dintre cele 39 de abordări comparate sunt afișate în dendrograma analizei cluster (Figura 14). Metoda celui mai apropiat vecin sau o singură legătură este una dintre cele mai simple tehnici de grupare ierarhică (Bálint și Jäntschi, 2021).

Datorită dimensiunii setului de date (3.538.080 variabile), distanța euclidiană dintre abordări este extrem de mare. Ajustarea distanței de legătură pe axa Y a fost selectată pentru a compara diferite abordări. Distanțele de legătură (Dlink) împărțite la distanța maximă de legătură (Dmax) sunt reprezentate de (Dlink/Dmax) *100.

Seturile de baze alese afectează cât de comparabile sunt tehnicile de optimizare. O clasificare în mai multe grupuri poate fi făcută atunci când au fost obținute rezultatele analizei PCA și grupare (clustering).

Diferențele dintre tehnicile de optimizare sunt neglijabile, iar gruparea arborilor a dezvăluit conexiunea lor. Aceștia ar trebui să fie aleși din grupuri distincte pentru o investigație amănunțită pentru a obține diverse concluzii din diferite puncte de vedere.

Abordările hibride sunt folosite de mai multe cercetări pentru analiza lor, deoarece produc rezultate mai bune (Scott și Radom, 1996; Batra și colab., 1996; Russ și colab., 2004).

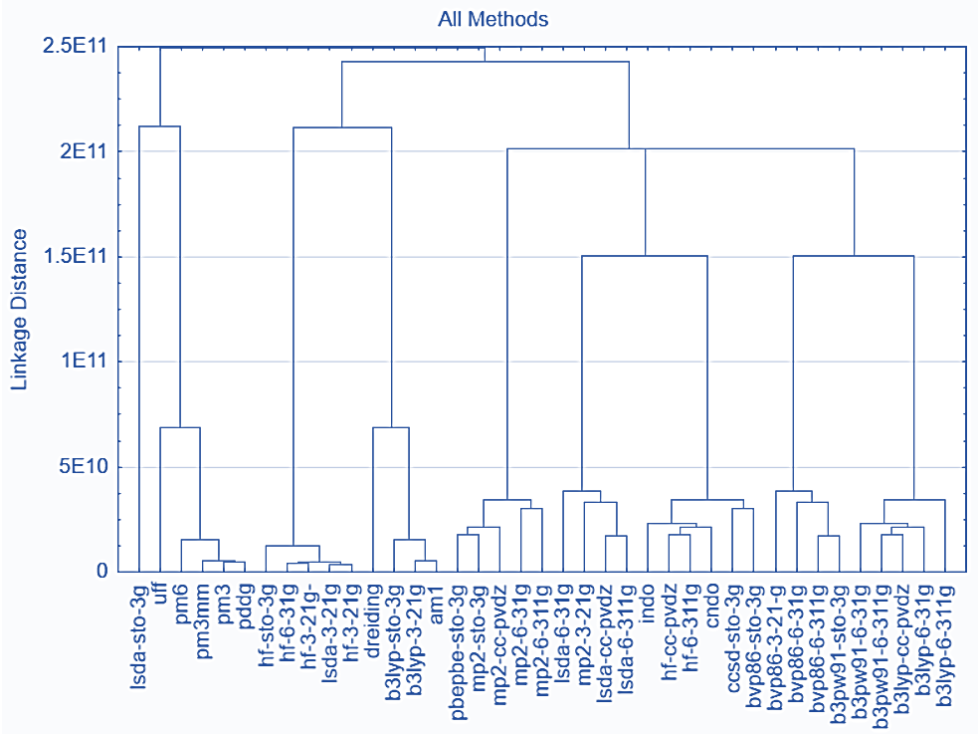

Figura 14. Rezultatele analizei grupării (clustering)

Deși mai multe tehnici au fost introduse ulterior în chimia computațională, Davidson și Feller (Davidson și Feller, 1988) au descris în 1988 câteva principii pe baza cărora se poate face o colecție de seturi de baze.

Deoarece cadrele teoretice și proprietăți moleculare diferite au cerințe diferite ale setului de baze, diferite arhitecturi de analiză și algoritmi au cerințe de eficiență diferite, iar acuratețea dorită depinde de aplicație, nu este practic să se dezvolte un singur set de baze „optimal” (Bálint și Jäntschi, 2021).

Cramer (2002) a arătat cum au evoluat de-a lungul timpului seturile de bază de valență divizată, începând cu cele mai populare seturi de bază de valență divizată, cum ar fi 3-21G, 6-21G, 4-31G, 6-31G și 6-311G și terminând cu mai multe instanțe recente precum cc-pCVDZ, cc-pCVTZ etc.

Este clar din asocierea tuturor seturilor de evaluări că este mult mai greu de prezis cu precizie geometriile moleculelor din a doua ordine, care conțin elemente, decât este pentru substanțele organice mai simple (Dunning, 1989).

De exemplu, s-a descoperit că PM3 depășește AM1 atunci când este aplicat acestor structuri (Stewart, 1989). Abordările DFT au, de asemenea, dezavantaje inerente, inclusiv tendințe inconsecvente și acuratețe ridicată a erorilor (Jensen, 2012).

Odată cu răspândirea metodologiilor actuale, căutarea combinațiilor „optime” de metode și seturi de baze care produc statistic constatări pozitive pentru o anumită colecție de molecule și atribute a fost mai pronunțată.

În modelarea moleculară și proiectarea medicamentelor, minimizarea energiei și optimizarea geometriei sunt tehnici cruciale. Descriptorii moleculari inexacți se corelează direct cu minimizarea ineficientă a energiei și/sau optimizarea geometriei (Jäntschi, 2011).

Descoperirile pe care le-am obținut după aplicarea Cluster și PCA pe fiecare subgrup sunt următoarele.

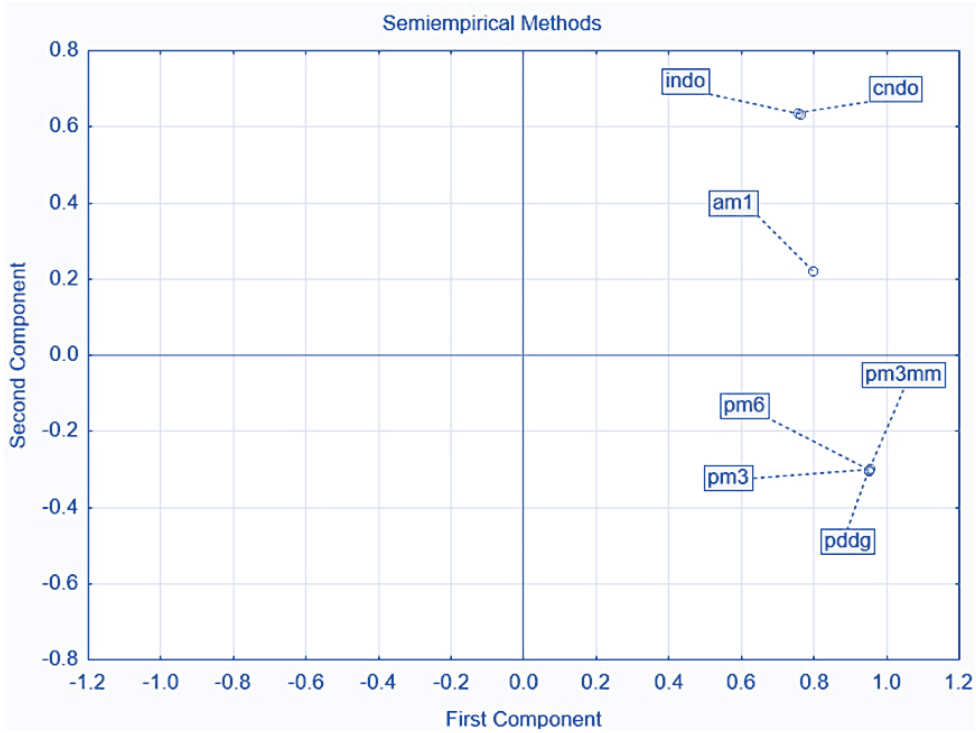

Figura 15. Graficul prezintă distribuția tehnicilor în componentele principale p1 și p2.

Rezultatele au fost împărțite în mai multe subgrupuri, deoarece a fost implementat un set de date relativ mare, ceea ce face rezultatele mai ușor de înțeles (Bálint și Jäntschi, 2021).

Rezultatele pot fi clasificate în trei grupe primare: pm6, pm3mm, pm3, pddg; am1; indo si cndo. Ar trebui să fie suficient să descriem datele noastre folosind o singură tehnică din fiecare grup.

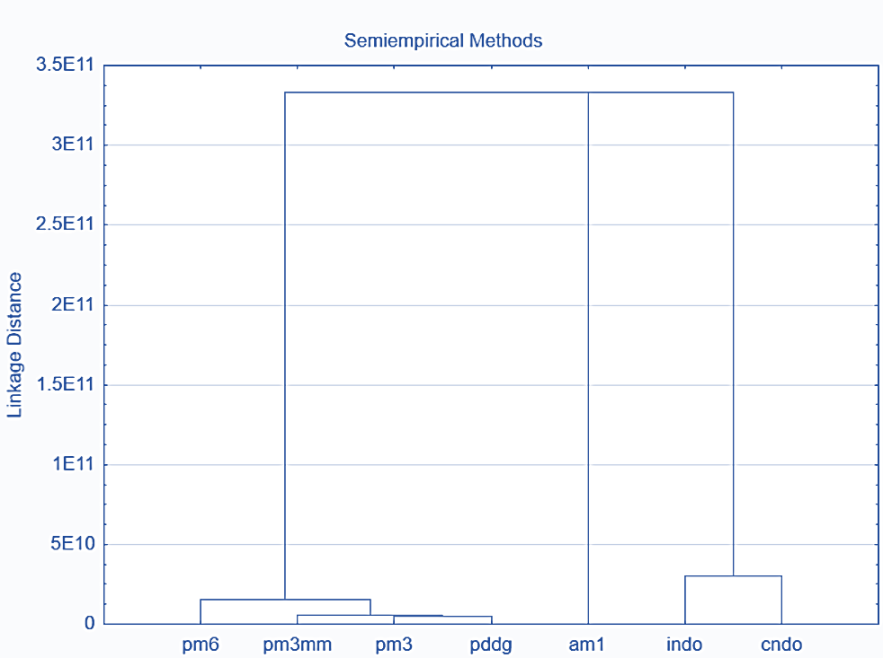

Figura 16. Rezultatele analizei grupării (clustering)

Datorită setului de date mai mare de această dată, analiza statistică pentru abordările DFT pare puțin diferită. Majoritatea abordărilor care au fost examinate au aparținut acestei familii.

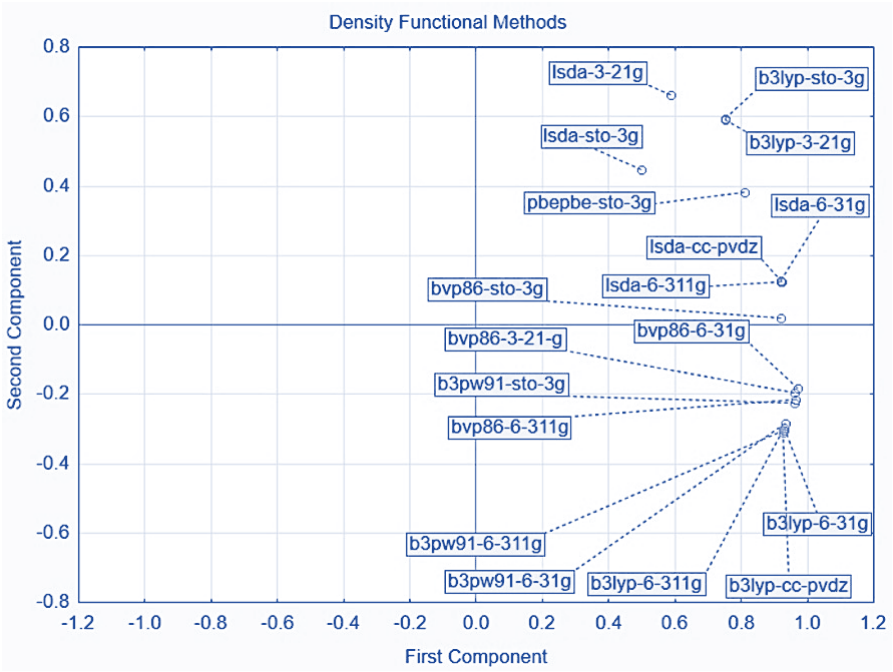

Figura 17. Graficul prezintă distribuția tehnicilor în componentele principale p1 și p2.

Figurile 17 și 18 arată modul în care abordările DFT sunt comparabile între ele, arătând, de asemenea, câteva metode „outlier”. Metodele pot fi clasificate în 4 grupuri principale și 3 grupări mai mici de metodologii.

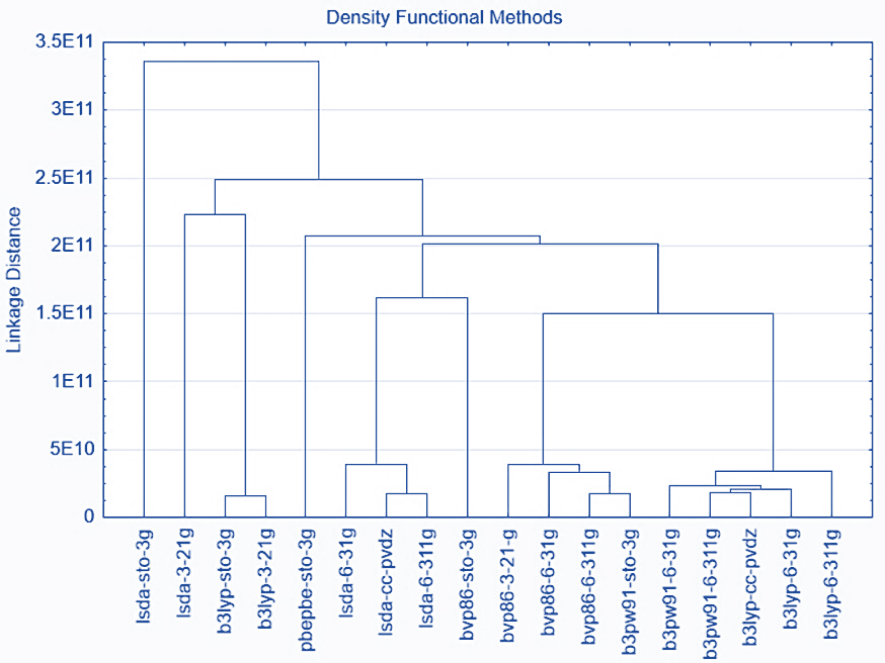

Figura 18. Rezultatele analizei grupării (clustering)

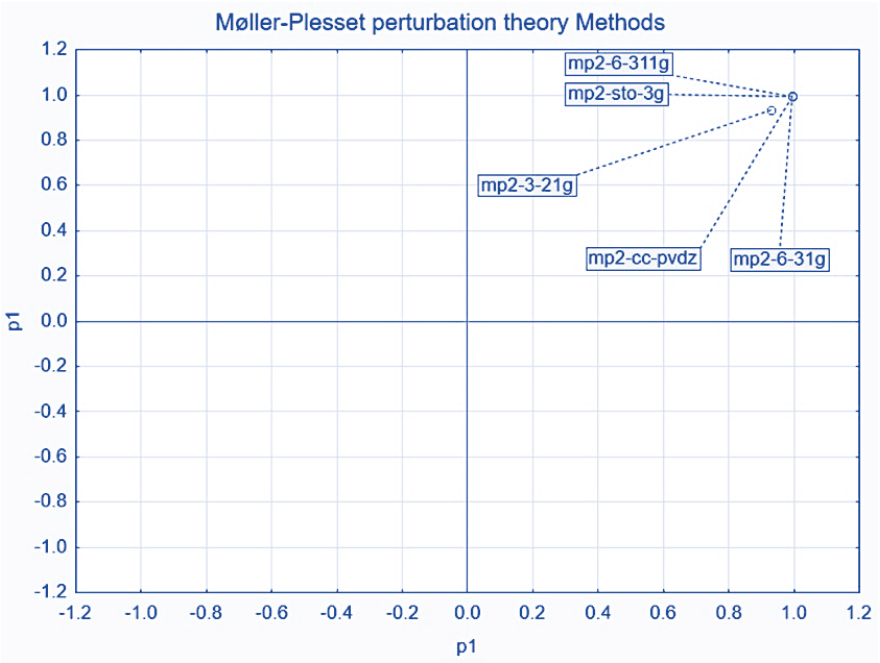

Figura 19. Graficul prezintă distribuția tehnicilor în componentele principale p1 și p2.

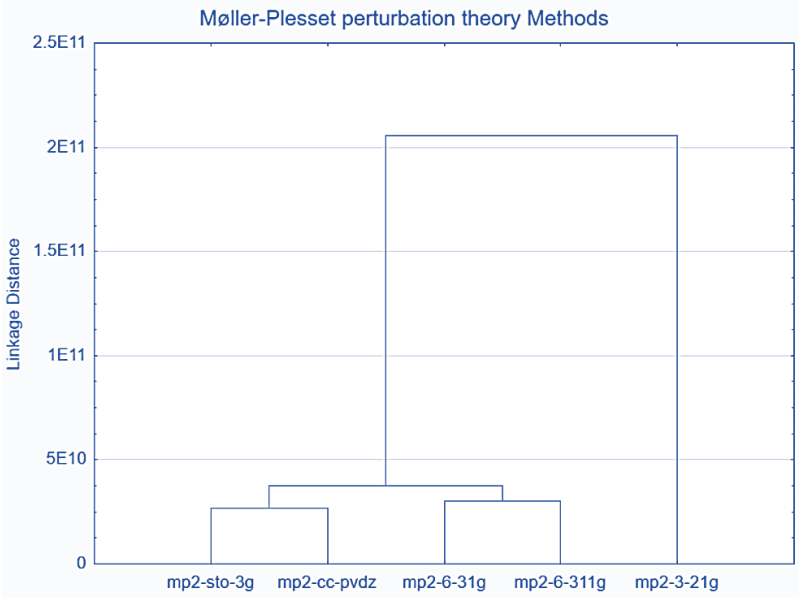

Figura 20. Rezultatele analizei grupării (clustering)

O componentă principală a fost găsită folosind metodele teoriei perturbațiilor Møller-Plesset (Figura 19). Unul (mp2-3-21g) și seturile de baze rămase alcătuiesc unul dintre cele două grupuri primare în care sunt separate metodele (Figura 20).

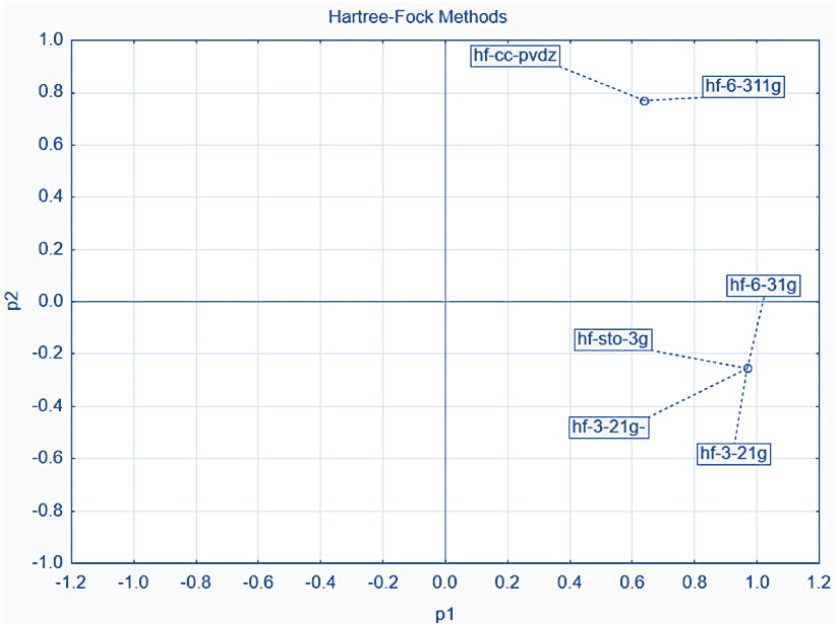

Figura 21. Graficul prezintă distribuția tehnicilor în componentele principale p1 și p2.

Metodele Hartree-Fock sunt considerate cele mai utilizate pe scară largă dintre calculele de optimizare. Pe baza analizei noastre au fost identificate 2 componente principale (Figura 21) și 2 grupuri principale (Figura 22).

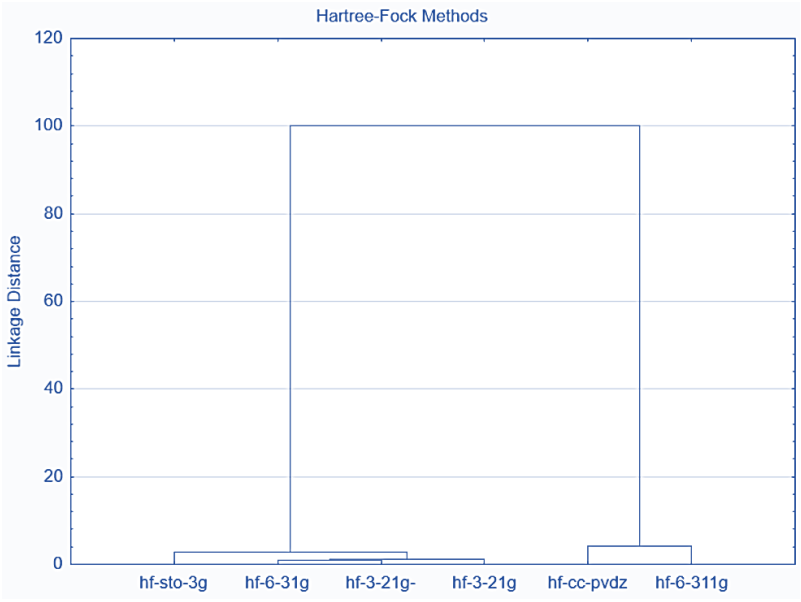

Figura 22. Rezultatele analizei grupării (clustering)

Datorită setului de date mic pe care l-au reprezentat, celelalte abordări care fac parte din teoria clusterelor cuplate și mecanica moleculară nu au putut fi analizate individual.

Dacă doar o metodă (CCSD) ar fi examinată în teoria clusterelor cuplate și două metode (UFF, Dreiding) în teoria mecanicii moleculare separat, nimeni nu a dezvăluit un rezultat semnificativ statistic. Ele sunt luate în considerare în analiza inițială, care examinează fiecare tehnică.

Sa implementat o examinare statistică alternativă: testul ANOVA cu un singur factor (Tabelul 7). Valorile de intrare pentru test au fost descriptorii moleculari obținuți pentru toate cele 39 de metode testate (Bálint și Jäntschi, 2021).

Tabel 7. Rezultate testului ANOVA

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2.92E+19	38	7.67E+17	0.405286	0.999584554	1.404833468
Within Groups	6.7E+24	3538041	1.89E+18			
Total	6.7E+24	3538079				

Pentru a stabili dacă există o variație între grupuri pe o anumită variabilă, se efectuează o ANOVA. Conform descriptorilor moleculari estimați, ipoteza nulă indică faptul că nu există nicio diferență vizibilă între procedurile analizate.

Valoarea statistică F (0.405) este procentul dintre variația dintre grupuri față de variația din cadrul grupurilor, iar o valoare F mai mare indică diferențe mai mari între grupuri. În acest caz, varianța a fost indicată în cadrul grupurilor analizate.

Dacă valoarea p este 0.9995, aceasta indică faptul că posibilitatea de a obține un rezultat la fel de extrem ca rezultatul observat, presupunând că ipoteza nulă este adevărată, este foarte mare. Prin urmare, diferențele observate între grupuri sunt probabil rezultatul întâmplării și nu sunt semnificative statistic, altfel spus.

Deoarece valoarea p este 0.9995 > 0.05, acceptăm ipoteza nulă și concluzionăm că diferențele dintre abordări nu sunt semnificativ diferite. Conform rezultatelor ANOVA din Tabelul 9, ipoteza nulă nu poate fi exclusă.

În analiza corelației, valorile a două variabile sunt comparate pentru a vedea dacă și dacă da, cât de puternic și în ce direcție sunt legate. Coeficientul de corelație este statistica utilizată pentru a măsura gradul de asociere între două variabile. Acesta variază de la -1 la +1, unde o valoare de -1 indică o corelație negativă perfectă (când o variabilă crește, cealaltă scade), 0 indică nicio corelație și o valoare de +1 indică o corelație pozitivă perfectă (când variabila creste, creste si cealaltă).

Au fost calculați și coeficienții de corelație între metodele analizate. În secțiunile următoare sunt prezentate corelațiile.

Este esențial să reținem că corelația nu implică cauzalitate în toate cazurile. O variabilă nu o provoacă neapărat pe cealaltă doar pentru că două variabile sunt conectate. Relația dintre cele două variabile poate fi influențată de factori suplimentari.

Următoarele tabele (Tabelul 8-13) prezintă corelația dintre metodele Teoriei Funcționale a Densității, Metodele Semiempirice, Teoria perturbației Møller–Plesset, Mecanica Moleculară și Metodele Hartree-Fock. Corelația între seturi de baze se referă la gradul de acord între rezultatele obținute din diferite seturi de baze.

Datorită setului de date extrem de mare, analiza corelației a trebuit să fie împărțită în mai multe analize. În fiecare categorie majoră menționată anterior, corelarea se face în cadrul sub-metodelor.

Coeficienții de corelație calculați pentru fiecare metodă indică o relație puternică între metodele analizate. Acest lucru este în concordanță cu celelalte rezultate statistice obținute pentru acest set de date.

Tabel 8. Density functional theory partea 1

	b3lyp-6-311g	b3lyp-6-31g	b3lyp-cc-pvdz	b3pw91-6-311g	b3pw91-6-31g	b3pw91-sto-3g	bvp86-3-21-g	bvp86-6-311g	bvp86-6-31g	bvp86-sto-3g
b3lyp-6-311g	1									
b3lyp-6-31g	0,995	1								
b3lyp-cc-pvdz	0,995	0,997	1							
b3pw91-6-311g	0,996	0,998	0,999	1						
b3pw91-6-31g	0,995	0,997	0,997	0,998	1					
b3pw91-sto-3g	0,942	0,943	0,944	0,944	0,946	1				
bvp86-3-21-g	0,933	0,937	0,935	0,936	0,936	0,995	1			
bvp86-6-311g	0,944	0,946	0,945	0,946	0,947	0,999	0,996	1		
bvp86-6-31g	0,942	0,944	0,944	0,945	0,946	0,996	0,994	0,997	1	
bvp86-sto-3g	0,777	0,782	0,780	0,780	0,785	0,891	0,892	0,890	0,894	1

Table 9. Density functional theory part 2

	lsda-6-311g	lsda-6-31g	lsda-cc-pvdz	pbepbe-sto-3g	b3lyp-3-21g	b3lyp-sto-3g	lsda-3-21g	lsda-sto-3g
lsda-6-311g	1							
lsda-6-31g	0,995	1						
lsda-cc-pvdz	0,999	0,996	1					
pbepbe-sto-3g	0,891	0,885	0,890	1				
b3lyp-3-21g	0,713	0,710	0,713	0,806	1			
b3lyp-sto-3g	0,714	0,711	0,714	0,807	0,999	1		
lsda-3-21g	0,504	0,518	0,508	0,600	0,857	0,856	1	
lsda-sto-3g	0,400	0,418	0,411	0,426	0,534	0,537	0,689	1

Table 10. Semiempirical methods

	cndo	indo	am1	pddg	pm3	pm3mm	pm6
cndo	1						
indo	0,997	1					
am1	0,652	0,660	1				
pddg	0,534	0,539	0,656	1			
pm3	0,536	0,541	0,657	0,999	1		
pm3mm	0,539	0,545	0,657	0,999	0,999	1	
pm6	0,535	0,541	0,659	0,999	0,999	0,999	1

Table 11. Møller–Plesset Methods

	mp2-3-21g	mp2-6-311g	mp2-6-31g	mp2-cc-pvdz	mp2-sto-3g
mp2-3-21g	1				
mp2-6-311g	0,895	1			
mp2-6-31g	0,897	0,997	1		
mp2-cc-pvdz	0,892	0,994	0,995	1	
mp2-sto-3g	0,893	0,992	0,995	0,997	1

Table 12. Molecular Mechanics methods

	dreiding	uff
dreiding	1	
uff	0,634	1

Table 13. Hartree-Fock Methods

	hf-6-311g	**hf-cc-pvdz**	**hf-3-21g**	**hf-3-21g-**	**hf-6-31g**	**hf-sto-3g**
hf-6-311g	1					
hf-cc-pvdz	0,999	1				
hf-3-21g	0,423	0,424	1			
hf-3-21g-	0,421	0,423	0,999	1		
hf-6-31g	0,422	0,423	0,999	0,999	1	
hf-sto-3g	0,421	0,422	0,999	0,999	0,999	1

Această analiză poate oferi informații despre mecanismele de bază care guvernează reacțiile chimice și poate ajuta la proiectarea sau optimizarea de noi materiale cu proprietățile dorite.

În plus, analiza corelației poate fi utilizată pentru a identifica caracteristici sau variabile importante care contribuie la proprietățile de interes, care pot fi utile pentru dezvoltarea modelelor predictive sau proiectarea de noi experimente.

Analiza corelației poate fi aplicată în cadrul diferitelor metode pentru a înțelege relația dintre diferitele proprietăți ale moleculelor sau materialelor.

Este folosit pentru a explora relația dintre descriptorii moleculari, cum ar fi energiile electronice și sarcinile atomice, și reactivitatea chimică sau alte proprietăți fizice sau chimice.

Cu cât dimensiunea eșantionului este mai mare, mai fiabil și precis va fi coeficientul de corelație. Un număr relativ mic de eșantioane poate duce la o predicție a corelației reale care este mai puțin precisă.

Acest lucru poate fi văzut în comparație cu metodele MM, în care au fost comparate doar două metode, și abordările DFT, în care s-a intenționat să fie analizat un set vast de date.

Determinarea corelației dintre metodele gaussiene este un aspect important al cercetării în chimie computațională. Acesta poate ajuta cercetătorii să selecteze cea mai potrivită metodă pentru o anumită aplicație, precum și să evalueze fiabilitatea și acuratețea rezultatelor obținute din diferite metode.

Cu toate acestea, corelația dintre metodele gaussiene nu implică neapărat că o metodă este mai bună decât cealaltă, deoarece fiecare tehnică are propriile sale avantaje și defecte.

Cunoașterea modului în care o metodă se raportează la alta este esențială pentru selectarea tehnicii optime de optimizare a geometriei care să fie utilizată în diferite circumstanțe. Datorită faptului că produc rezultate aproape identice, două proceduri similare ar trebui excluse din analiză.

Pentru a răspunde la întrebările de cercetare din secțiunea enunțarea problemei, se poate afirma următoarele (Bálint și Jäntschi, 2021):

- Mărimea setului de bază nu reprezintă cu exactitate cât de larg poate fi utilizat.
- Există doar câteva seturi de bază care se potrivesc setului nostru de date; este imposibil de determinat setul optim de baze.
- Dacă folosim seturi de baze diferite din grupări diferite, vom obține rezultate care variază, dar trebuie să fim precauți să le alegem corect.
- Este posibilă gruparea și corelarea tehnicilor de optimizare.

Aplicarea corectă a diferitelor proceduri de selecție și optimizare determină diferențele între rezultate.

4. Concluzii

Rezultatele analizei studiului oferă perspective asupra relațiilor și corelațiilor dintre diferitele abordări utilizate în cercetare. Cercetătorii au examinat un total de 39 de metode și le-au reclasificat în funcție de caracteristicile și aplicabilitatea lor. Această reclasificare a ajutat la o mai bună înțelegere și selectare a seturilor de bază adecvate pentru diferite domenii de studiu.

Analizând relațiile și corelațiile dintre metode, am obținut informații valoroase despre asemănările și diferențele lor. Această înțelegere le-a permis să identifice punctele forte și limitările fiecărei metode și să determine potrivirea lor pentru domenii sau aplicații specifice de cercetare.

Reclasificarea metodelor a oferit un cadru mai organizat pentru selectarea seturilor de bază adecvate. Seturile de baze sunt importante în chimia computațională și mecanica cuantică, deoarece formează baza pentru calcule și simulări. Prin potrivirea caracteristicilor și cerințelor zonelor de studiu cu seturile de bază specifice, cercetătorii pot îmbunătăți acuratețea și fiabilitatea rezultatelor lor.

În general, acest proces de analiză și reclasificare a ajutat la eficientizarea selecției seturi de bază pentru diferite domenii de studiu, facilitând o cercetare mai eficientă și mai eficientă în acele domenii.

Capitolul V – Aplicații ale tehnicilor de optimizare geometrică

Studiu de caz – Analiza factorială al nanostructurilor

1. Introducere

O structură în formă de dodecaedru alcătuită din atomi de carbon formează dodecaedrul cu patru straturi, un ipotetic alotrop de carbon. Are patru straturi de atomi de carbon, fiecare dintre ele având 20 de atomi aranjați într-un model pentagonal regulat. Acesta este motivul pentru care este menționat ca fiind „cu patru straturi” (Cao și colab., 2020).

Analiza relației dintre tipurile de atomi ale dodecaedrului și caracteristicile acestora poate dezvălui detalii importante despre varietatea și stabilitatea structurilor rezultate. Acesta este motivul pentru care sunt supuși a numeroase studii în acest domeniu (Jäntschi și colab., 2016).

În acest studiu de caz, scopul a fost să examineze modul în care diverși factori și interacțiunile lor influențează zece proprietăți calculate ale unei clase de congeneri dodecaedru. Moleculele au fost privite ca structuri cu patru straturi, așa cum se arată în Figura 28.

Studiul a analizat, de asemenea, efectul formării fiecărui strat cu carbon, bor sau azot. Pentru atingerea acestor obiective a fost realizat un studiu teoretic folosind un design factorial complet. Scopul a fost identificarea și înțelegerea factorilor semnificativi care influențează proprietățile de interes (Jäntschi și colab., 2016).

2. Metodologie

Inițial, structurile au fost create folosind software-ul *HyperChem* (la început, PM3 a fost folosit pentru optimizarea geometriei).

Folosind pachetul software *Spartan*, designul geometric al cuștilor a fost îmbunătățit cu metoda HF, un set de bază 3-21G (Hartree, 1928) și apoi prin MPM (Møller Chr Plesset, 1934) până la următoarea comandă (MP2) cu un 6-31G. * set de baze (Jäntschi și colab., 2016).

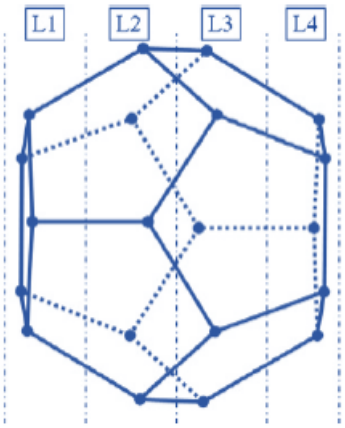

Figura 22. Structura de dodecaedru

- Optimizare geometrică: PM3 și Moller-Plesset (MP2) cu setul de baze 6-31G*.
- Calcularea proprietăților (volum, suprafață, ovalitate, HOMO, LUMO, polarizabilitate, moment dipol, entropie, entalpie, energie).
- Analiză factorială completă (detectarea grupurilor de factori echivalenți și irelevanți; examinarea efectelor principale și a efectelor interacțiunii).

În urma optimizării geometriei, calculele MP2 au fost utilizate pentru a determina caracteristicile utilizate pentru analiza factorială completă.

În contextul explorării relației dintre caracteristicile preluate din structuri compuse și diverse calități, o analiză factorială a presupus testarea tuturor combinațiilor posibile ale caracteristicilor (variabile independente) asupra măsurilor de calitate (variabile de răspuns).

3. Rezultate și discuții

Un model explicativ solid ar putea avea un coeficient de corelație (R) apropiat de 0.95, ceea ce înseamnă că factorii pot reprezenta aproximativ 90% din variația atributului relevant. Drept urmare, modelele cu R > 0.95 au fost supuse unor analize suplimentare (Jäntschi și colab., 2016).

S-a determinat care caracteristici sunt cele mai strâns legate de activitate și modul în care interacționează prin modificarea în mod repetat a nivelurilor fiecărei componente și monitorizarea activității ulterioare.

Cu o singură excepție (energie), analiza arată că numărul de componente variază pentru diferiți atomi de referință. Când carbonul, borul și azotul au fost luați în considerare ca atomi de referință, a fost comparat numărul de componente necesare pentru a explica variația fiecărei caracteristici.

Volumul, suprafața, ovalitatea, HOMO, LUMO, polarizabilitatea și entropia au fost cele șapte din zece atribute pentru care s-a descoperit că atomul de referință de carbon necesită cei mai puțini parametri să fie luați în considerare. Cu toate acestea, atunci când carbonul era atomul de referință, numărul mediu de componente necesare în modelul cu un coeficient de corelație mai bun de 0.95 a fost mai mare (Jäntschi și colab., 2016).

Studiul a demonstrat că atunci când caracteristicile și trăsăturile structurale sunt intenționate a fi legate pentru datele corespunzătoare dintr-o populație, simplitatea datelor are ca rezultat, de obicei, modele mai simple (Kelly, 2011), modele cu putere explicativă insuficientă (Kar și Arias-Estrada, 2015). Prin urmare, în aceste situații, există întotdeauna un compromis între creșterea dimensiunii eșantionului și reducerea complexității modelului.

Descoperirea modelelor cu abilități predictive necesită validarea modelelor liniare (Gramatica, 2013), dar acest subiect a fost în afara domeniului investigației noastre și nu este abordat în această lucrare.

Lucrarea descrisă în această cercetare s-a concentrat pe crearea unei analize factoriale complete pentru explorarea relației dintre caracteristicile preluate din structuri compuse și diverse calități (Jäntschi și colab., 2016).

Strategiile full-factoriale pot fi laborioase și consumatoare de resurse, mai ales atunci când numărul de caracteristici și niveluri este mare. Cu toate acestea, ele oferă mai multe avantaje față de alte modele experimentale, cum ar fi să permită detectarea interacțiunilor de ordin superior între factori și să ofere o imagine mai completă a relației dintre caracteristici și variabilele de răspuns.

4. Concluzii

În ciuda faptului că 67% din timp, carbonul a fost folosit ca atom de referință, modelele care au dat cel mai mare coeficient de corelație nu au fost întotdeauna cele mai eficiente modele din jur. De la cele mai simple modele, care folosesc borul ca referință, până la modelele care folosesc carbonul ca referință, complexitatea modelelor crește (modele contorsionate).

Capitolul VI – Aplicații ale tehnicilor de aliniere moleculară

Studiu de caz – Similitudine biochimică a proteinelor selectate

1. Introducere

În biologia moleculară și chimia biologică, determinarea structurii moleculare este esențială, deoarece funcția unei molecule este foarte dependentă de aranjamentul 3D și geometria moleculelor care se completează reciproc. Înțelegerea naturii structurii și conexiunilor rețelelor biologice este esențială pentru a ajunge la o descriere cantitativă a funcțiilor acestora.

Utilizarea instrumentelor informatice a devenit din ce în ce mai importantă în domenii precum modelarea moleculară, andocare și crearea de medicamente farmaceutice. Tehnicile de calcul matematic utilizate pentru modelarea problemelor geometrice care implică molecule ca sisteme algebrice și algoritmii utilizați pentru rezolvarea acestor sisteme sunt frecvent examinate (Emiris și colab., 2005). Există mai multe implementări ale tehnicilor bine-cunoscute și eficiente pentru calcularea valorilor proprii și vectorilor proprii disponibile (Murrain și Pan, 2000).

Luo și colab. (2006), au descoperit că interacțiunile puternice dintre componentele matricei și valorile proprii fac ca valorile proprii să se coreleze puternic, ceea ce are ca rezultat fluctuații ale valorilor proprii reprezentate de GOE (ansamblu ortogonal gaussian).

Această analiză (Joita și colab., 2021) și-a propus să determine aranjarea geometrică optimă a 20 de aminoacizi aleși unul față de celălalt. S-a făcut o dezvoltare a studiului anterior pe care Jäntschi (Jäntschi, 2019) a detaliat.

Algoritmul „eigenproblem” aliniază structurile, apoi metoda trilateralii este folosită pentru a atribui toți atomii în grupe anterioare.

2. Metodologie

Obținerea alinierii optime pentru o moleculă implică determinarea celei mai favorabile alinieri a moleculei în spațiu în raport cu un cadru de referință sau o altă moleculă. Alinierea optimă oferă o modalitate de a compara și analiza proprietățile și caracteristicile moleculelor pe baza aranjamentului și geometriei lor spațiale.

Găsirea celei mai bune alinieri este crucială pentru prezicerea caracteristicilor și acțiunilor moleculelor, cum ar fi reactivitatea, afinitatea de legare și stabilitatea acestora, în contextul chimiei computaționale și al modelării moleculare. În domenii precum dezvoltarea

medicamentelor, știința materialelor și ingineria chimică, aceste cunoștințe pot fi folosite pentru a construi noi molecule cu caracteristicile dorite.

Cea mai mică valoare a sumei pătratelor valorilor proprii ale matricei de distanță carteziană - ale cărei valori proprii sunt în întregime imaginare, deoarece matricea este antisimetrică - este determinată a fi cea mai bună aliniere a unei molecule.

Schema algoritmului de lucru este prezentată mai jos (Figura 23).

Se analizează aminoacizii (descărcați de pe PubChem). În acest caz, glicina, care are cei mai puțini atomi grei, este folosită ca referință (Joița și colab., 2021).

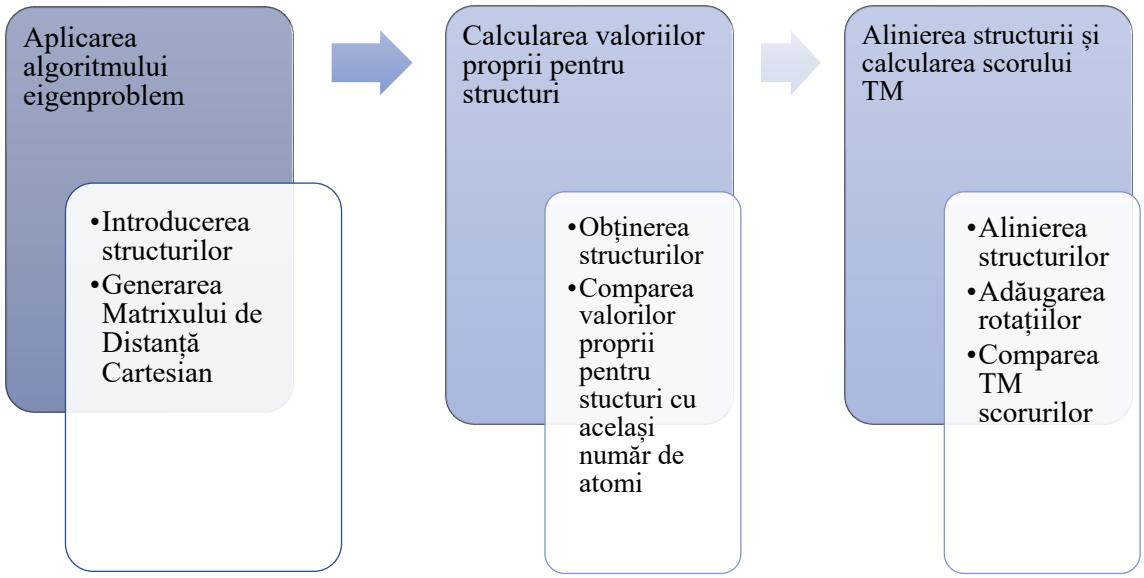

Figura 23. Algoritmul de lucru

Odată ce condițiile au fost îndeplinite, metoda originală a problemelor proprii este executată pentru a confirma că punctul de pornire al programului este o aliniere originală adecvată. Apoi, sunt descoperite toate aranjamentele cu mai puțin de un anumit număr de atomi. Pentru fiecare combinație, valorile proprii sunt descoperite fără a roti posibilitățile.

Sumele ST sunt de asemenea comparate până când variabilele de intrare sunt satisfăcute sau sunt comparate toate posibilitățile cu cel puțin trei atomi. Folosind procedura originală cu probleme proprii, structurile sunt aliniate, apoi supuse trilaterării și rotațiilor potențial utile pi/2.

Coordonatele structurilor sunt ajustate astfel încât să se alinieze cu coordonatele structurii de referință. Întrucât TM scorul (Scorul de modelare șablon) compară separările dintre atomi în molecule, structurile finale care au rezultate bune sunt exportate.

3. Rezultate și discuții

Deoarece una dintre aceste rotații poate avea nevoie să producă o bună suprapunere a celor doi aminoacizi, valorile medii pe fiecare axă sunt calculate pentru câțiva atomi selectați din ambele structuri. Scorurile mai mari indică o asemănare structurală mai mare. Scorul final merge de la 0 la 1.

Scorul TM a fost creat inițial pentru a ajuta la predicția și modelarea structurilor proteinelor, dar ulterior a fost utilizat pentru a aborda o varietate de probleme de biologie structurală, cum ar fi compararea structurii proteinelor, andocarea proteinei-proteine și investigațiile legate de proteină-ligand. Este util în special pentru compararea structurilor care, deși similare în general, pot varia în detalii minuscule, cum ar fi regiunile buclei sau conformațiile lanțului lateral (Zhang, 2005).

Treisprezece constatări sunt aliniamente unice cu încredere ridicată a celor 19 aminoacizi care au fost aliniați la glicină, dintre care 11 au un scor TM ridicat. Scorul TM poate fi utilizat pentru a determina care dintre celelalte trei (cisteină, lizină și arginină) oferă cele mai mari rezultate dintre două posibile bune (Joița și colab., 2021).

Pentru rezultate, este permisă o notă de 80% din punctajul maxim. Acest lucru este necesar pentru ca rezultatul să fie cel mai bun aliniere, chiar dacă nu are cel mai mare scor TM.

Vizualizarea structurilor selectate și eliminarea celor care pot avea scoruri matematice similare, dar tipurile de atom incorecte este o altă tehnică simplă de a alege din acest grup de candidați. Rezultate și mai bune ar putea fi obținute prin combinarea altor funcții de notare sau alte mecanisme de notare (Joița și colab., 2021).

Criteriile de limitare sunt relevante pentru comparația actuală, deoarece 84% dintre cele mai bune aliniamente care utilizează glicina ca standard pot fi indicate numeric printr-o funcție de punctare precum scorul TM și 68% dintre aceste orientări sunt exportate ca candidați unici. Funcția de scor furnizată în prezent este utilă doar pentru 58% din cisteină. O bază de date mare ar arăta o modalitate rezonabilă de a le selecta și ar ajuta la instruirea învățării automate.

Alinierea corectă a fost identificată statistic prin scorul TM în 70% din cazuri, în mod tipic, după rularea metodei curente folosind ceilalți aminoacizi ca referință. Prin evaluare, cu 15% mai multe cazuri cu scoruri similare pot fi distinse clar (Joița et al., 2021).

Vectorizarea completă poate accelera algoritmul curent. Pentru a reduce efectele abilităților de raportare parțială și a formei funcționale predefinite inspirate de teorie, învățarea automată trebuie introdusă în funcțiile de notare.

În loc să impună un algoritm rigid, aceste defecte pot fi remediate prin aplicarea învățării automate pentru a captura trăsături care sunt dificil de prezis, deoarece există atât de multe relații cantitative structură-activitate nemăsurate/necunoscute/nedescoperite. Există o cantitate tot mai mare de date structurale și de interacțiune excelente în literatura de specialitate pe care le poate folosi învățarea automată.

Metodele bazate pe secvențe, tehnicile de aliniere structurală și tehnicile hibride care încorporează date de secvență și structură sunt toate disponibile pentru alinierea proteinelor în QSAR. ClustalW, MUSCLE și PyMOL sunt câteva dintre programele software care sunt utilizate frecvent pentru alinierea proteinelor în QSAR (Saeys și colab., 2007).

4. Concluzii

Pentru a stabili alinierea geometrică optimă a aminoacizilor specifici unul față de celălalt, a fost dezvoltată o aplicație a problemei proprii.

Prin urmare, putem spune că alinierea ideală nu este o linie dreaptă. Rezultatele apropiate ale aceleiași metode pot fi luate în considerare. Chiar și după ce a fost rulat un algoritm de scor, putem deduce că alinierea cu cel mai mare scor nu este întotdeauna cea mai bună aliniere.

Numărul de rotații pentru care se execută o funcție de scor trebuie redus cu parametrii metodei existente. În plus, integrarea diferitelor abordări poate duce la rezultate mai rapide.

Concluzii generale și perspective de viitor

Concluziile specifice și generale ale fiecărei secțiuni au fost deja furnizate în capitolele corespunzătoare. Astfel, următoarele paragrafe sunt dedicate furnizării unei concluzii generale acestei teze, precum și a unor idei pentru un studiu suplimentar.

Modelarea substanțelor chimice anorganice și organice care a condus la examinarea și descrierea comportamentului lor la nivelul sistemelor biologice a servit drept ilustrare a asemănărilor lor.

Înțelegerea fundamentală în tema optimizării moleculare a fost dezvoltată pe parcursul acestei teze. Raționalizarea observațiilor experimentale cu calcule energetice a condus la înțelegerea modelării moleculare. Pe de altă parte, citirea literaturii anterioare și actuale despre studii teoretice și experimentale mi-a oferit o perspectivă largă asupra acestui subiect de cercetare.

Concluziile acestei teze demonstrează că metodele chimice computaționale pot fi utilizate pentru a defini în mod adecvat strategiile de optimizare moleculară pe moleculele studiate.

Pentru a înțelege descoperirile experimentale și a identifica interacțiunile cheie dintre molecule, va fi important să se evalueze modele structurale pentru a descrie cu exactitate conformația chimică în viitorul apropiat.

Se poate concluziona că există o breșă substanțială a cunoștințelor între domeniile experimental și teoretic prin lucrul cu grupuri experimentale și examinarea literaturii experimentale. În special, interpretările repetate ale constatărilor experimentale par să nu aibă suport teoretic. Pentru a susține multe dintre afirmațiile enunțate în literatura de specialitate, mai trebuie făcută multă muncă în științele fundamentale din perspective experimentale și teoretice.

Acest lucru poate influența alegerea și selecția metodelor care trebuie evaluate din motive de optimizare, în funcție de condiția pentru care metoda este necesară atunci când este privită prin prisma asemănării.

Identificarea metodelor de optimizare moleculară care pot ajuta ulterior la înțelegerea mai bună a mecanismelor moleculare este esențială pentru cercetările ulterioare.

Găsirea celui mai bun pentru nevoile noastre este atât mai simplă, cât și mai dificilă acum, deoarece există atât de multe programe software și algoritmi disponibili datorită progreselor tehnologiei. Pentru obiectivul definirii conceptului de similaritate, cercetarea relației structură/activitate între moleculele din sistemele biologice este crucială.

Bibliografie selectivă

Aaby, K.; Hvattum, E.; Skrede, G. 2004. Analysis of flavonoids and other phenolic compounds using high-performance liquid chromatography with coulometric array detection: Relationship to antioxidant activity. Journal of Agricultural and Food Chemistry, 52, pp. 4595–4603.

Abegg, P.W.; Ha, T.K. 1974. Ab initio calculation of spin-orbit-coupling constant from Gaussian lobe SCF molecular wavefunctions. Molecular Physics, 27, pp. 763-67.

Agresti, A. 2007. An introduction to categorical data analysis. John Wiley & Sons.

Allen, B.C.P.; Grant, G.H.; Richards, W.G. 2001. Similarity calculations using two dimensional molecular representations. Journal of Chemical Information and Computer Sciences, 41, pp. 330-337.

Aryal, S.; Baniya,M. K.; Danekhu, K.; Kunwar, P.; Gurung, R.; Koirala, N. 2019. Total Phenolic Content, Flavonoid Content and Antioxidant Potential of Wild Vegetables from Western Nepal. Plants, 8(4), pp. 96.

Arnao, M.B. 2000. Some methodological problems in the determination of antioxidant activity using chromogen radicals: A practical case. Trends in Food Science and Technology, 11, pp. 419–421.

Bálint, D.; Jäntschi, L. 2019. Missing data calculation using the antioxidant activity in selected herbs. Symmetry ,11(6).

Bálint, D.; Jäntschi, L. 2021. Comparison of molecular geometry optimization methods based on molecular descriptors. Mathematics, 9(22).

Banerjee T.; Ramalingam A. 2015. Desulphurization and Denitrification of Diesel Oil Using Ionic Liquids. Experiments and Quantum Chemical Predictions, Elsevier.

Batra P.; Bernd G.; Gescheidt M.S.G.; Houk, K.N. 1996. Calculations of Isotropic Hyperfine Coupling Constants of Organic Radicals. An Evaluation of Semiempirical, Hartree-Fock, and Density Functional Methods. Journal of Physical Chemistry, 100, pp. 18371-18379.

Bender, A.; Glen, R.C. 2004. Molecular similarity: a key technique in molecular informatics. Organic and Biomolecular Chemistry, 2, pp. 3204-3218.

Bolboacă, S.D., Jäntschi, L. 2013. Linear regression modelling and validation strategies for structure-activity relationships, in: BIOMATH, International Conference on Mathematical Methods and Models in Biosciences, pp. 18-21.

Cao, Y.; Yang, Q.Z.; Lu, X. 2020. Theoretical exploration of the stability and mechanical properties of four-layered dodecahedrane. The Journal of Physical Chemistry C, 124(15), pp. 8403-8410.

Cramer, C.J. 2002. Essentials of Computational Chemistry: Theories and Models. John Wiley and Sons, Ltd: West Sussex.

Davidson E.R.; Feller D. 1988. Basis Set Selection for Molecular Calculations. Chemical Reviews, 86, pp. 661-696.

Diudea, M.V.; Gutman, I.; Jäntschi, L. 2002. Molecular Topology; Nova Science: Huntington, New York.

Dong, R.; Peng, Z.; Zhang, Y.; Yang, J. 2018. MTM-Align: An algorithm for fast and accurate multiple protein structure alignment. Bioinformatics, 34, pp. 1719–1725.

Doucet, J.P.; Weber, J. 1996. Molecular similarity. Computer-Aided Molecular Design, pp. 328-362.

Dunning, T.H. 1989. Gaussian basis sets for use in correlated molecular calculations. The atoms boron through neon and hydrogen. The journal of chemical physics, 90, pp. 1007.

Emiris, I.Z.; Fritzilas, E.D.; Manocha, D. 2005. Algebraic algorithms for structure determination in biological chemistry. International Journal of Quantum Chemistry, 106(1), pp. 190–210.

Garcia-Planas, M.I. 2021. Geometric Structure of the Set of Pairs of Matrices under Simultaneous Similarity. Universal Journal of Mathematics and Applications, 4(4), pp. 147-153.

Golbraikh, A.; Tropsha, A. 2000. Predictive QSAR modelling based on diversity sampling of experimental datasets for the training and test set selection. Molecular Diversity, 5, pp. 231-243.

Gramatica, P. 2013. On the development and validation of QSAR models. Methods in Molecular Biology, 930, pp. 499–526.

Hammett, L.P. 1937. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. Journal of American Chemical Society, 59, pp. 96-103.

Hartree, D.R. 1928. The wave mechanics of an atom with a non-coulomb central field: Part I. Theory and methods. Proceedings of Cambridge Philosophical Society, 24, pp. 89–110.

Hill, J.G. 2012. Gaussian basis sets for molecular applications. International Journal of Quantum Chemistry, 113, pp. 21–34.

Ivanov, I.G.; Vrancheva, R.Z.; Marchev, A.S.; Petkova, N.T.; Aneva, I.Y.; Denev, P.P.; Georgiev, G.C.; Pavlov, A.I. 2014. Antioxidant activities and phenolic compounds in Bulgarian *Fumaria* species. International Journal of Current Microbiology and Applied Sciences, 3(2), pp. 296-306.

Ivanova, D.; Gerova, D.; Chervenkov, T.; Yankova, T. 2005. Polyphenols and antioxidant capacity of Bulgarian medicinal plants. Journal of Ethnopharmacology, 96, pp. 145–150.

Jäntschi L. 2011. Computer assisted geometry optimization for in silico modeling. Applied Medical Informatics, 29(3), pp. 11-18.

Jäntschi, L. 2005. Molecular descriptors family on structure activity relationships 1. Review of the methodology. Leonardo Electronic Journal of Practices and Technology, 6, pp. 76-98.

Jäntschi, L. 2012. Distribution Fitting 16. How Many Colours are in the Field? Bulletin of University of Agricultural Sciences and Veterinary Medicine, Horticulture, pp. 69.

Jäntschi, L. 2019. The eigenproblem translated for alignment of molecules. Symmetry, 11, pp. 1027.

Jäntschi, L.; Bálint, D.; Bolboaca, S.D. 2016. Multiple linear regressions by maximizing the likelihood under assumption of generalized Gauss-Laplace distribution of the error. Computational and Mathematical Methods in Medicine.

Jäntschi, L.; Bálint, D.; Pruteanu, L.L.; Bolboaca, S.D. 2016. Elemental factorial study on one-cage pentagonal faces nanostructure congeners. Materials Discovery, 5, pp. 14 - 21.

Jäntschi, L.; Bolboacă, S. 2016. Molecular modelling in compounds series with descriptors families. Anual University Oradea Fascicula Chimie, 23, pp. 5-14.

Jäntschi, L.; Pruteanu, L.L.; Cozma, A.C.; Bolboaca, S.D. 2015. Inside of the linear relation between dependent and independent variables. Computational and Mathematical Methods in Medicine, pp.11.

Jensen F. 2012. Atomic orbital bases sets. WIREs Computational Molecular Science, 3(3), pp. 273-295.

Joita, D.M.; Tomescu, M.A.; Bálint, D.; Jäntschi, L. 2021. An application of the eigenproblem for biochemical similarity. Symmetry, 13(10).

Kar, K.; Arias-Estrada, S. 2015. How to judge predictive quality of classification and regression based QSAR models? in: Z. Ul-Haq, J.D. Madura (Eds.), Frontiers in Computational Chemistry, Volume 2: Computer Applications for Drug Design and Biomolecular Systems, Bentham Science Publishers Ltd., pp. 71–120.

Kayano, M.; Dozono, K.; Konishi, S. 2010. Functional Cluster Analysis via Orthonormalized Gaussian Basis Expansions and Its Application. Journal of Classification, 27, pp. 211–230.

Kelly, K.T. 2011. Philosophy of statistics, in: P.S. Bandyopadhyay, M.R. Forster (Eds.), Handbook of the Philosophy of Science, vol. 7, Elsevier, pp. 983–1024.

Kolodny, R.; Koehl, P.; Levitt, M. 2005. Comprehensive evaluation of protein structure alignment methods: Scoring by geometric measures. Journal of Molecular Biology, 346, pp. 1173–1188.

Lin, J.; Wen, L.; Zhou, Y.; Wang, S.; Ye, H.; Su, J.; Li, J.; Shu, J.; Huang, J.; Zhou, P. 2023. PepQSAR: a comprehensive data source and information platform for peptide quantitative structure–activity relationships. Amino Acids, 55, pp. 235–242.

Luo, Y.; Cai, Q.; Sun, M.; Corke, H. 2004. Antioxidant activity and phenolic compounds of 112 traditional Chinese medicinal plants associated with anticancer. Life Sciences, 74, pp. 2157–2184.

Luo, F.; Zhong, J.; Yang, Y.; Scheuermann, R.H.; Zhou, J. 2006. Application of random matrix theory to biological networks. Physics Letters A, 357(6), pp. 420–423.

McDonald, J.H. 2014. Handbook of biological statistics (3rd ed.). Sparky House Publishing.

Møller Chr Plesset, M.S. 1934. Note on an approximation treatment form many-electron system. Physical Reviews, 46, pp. 618–622.

Nantasenamat, C. 2020. Best Practices for Constructing Reproducible QSAR Models, in: Roy, K. (Ed.), Ecotoxicological QSARs. Methods in Pharmacology and Toxicology, pp. 55-76.

Nelson, S.D.; Seybold, P.G. 2001. Molecular structure-property relationship for alkenes. Journal of Molecular graphics and modelling, 20, pp.36-53.

Pan, J.S. 2012. Properties and Application of Characteristic Polynomial. Advanced Materials Research, 490-495, pp. 3516–3521.

Perlt, E. 2021. Basis Sets in Computational Chemistry. Lecture Notes in Chemistry, Springer.

Petersson G.A.; Malick D.K.; Wilson W.G.; Ochterski J.W.; Montgomery Jr J.A.; Frisch M.J. 1998. Calibration and comparison of the Gaussian-2, complete basis set, and density functional methods for computational thermochemistry. The Journal of Chemical Physics, 109, pp. 10570-10579.

Pople, J.A. 1999. Quantum Chemical Models. Angewandte Chemie, 38, pp. 13–14.

PubChem Database Access: https://pubchem.ncbi.nlm.nih.gov/, Accessed 17 April 2023.

Randić, M. 1975. Characterization of molecular branching. Journal of the American Chemical Society, 97 (23), pp. 6609–6615.

Randić, M.; Novič, M.; Vračko, M. 2008. On novel representation of proteins based on amino acid adjacency matrix. SAR and QSAR in Environmental Research, 19(3-4), pp. 339–349.

Reynolds, C.A.; Burt, C.; Graham Richards, W. 1992. A linear molecular similarity index. Quantitative Structure-Activity Relationships, 11, pp. 34-35.

Russ, N.J.; Crawford, T.D.; Tschumper, G.S. 2004. Real versus artefactual symmetry-breaking effects in Hartree–Fock, density-functional, and coupled-cluster methods. The journal of chemical physics, 120, pp. 7298.

Saeys, Y.; Inza, I.; Larrañaga, P. 2007. A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), pp. 2507-2517.

Schlegel, H.B. 2003. Exploring Potential Energy Surfaces for Chemical Reactions: An Overview of Some Practical Methods. Journal of Computational Chemistry, 124, pp. 1514.

Scott, A.P.; Radom, L. 1996. Harmonic Vibrational Frequencies: An Evaluation of Hartree-Fock, Møller-Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. Journal of Physical Chemistry, 100, pp. 16502-16513.

Scuseria, G.E. 1992. Comparison of coupled-cluster results with a hybrid of Hartree-Fock and Density Functional Theory. The Journal of Chemical Physics, 97, pp. 7528-7530.

Stewart, J.J.P. 1989. Optimization of parameters for semiempirical methods. Applications. Journal of Computational Chemistry, 10, pp. 209-221.

Stumpfe, D.; Bajorath, J. 2011. Recent advances in the development of new similarity searching methods and applications. Expert opinion on drug discovery, 6(1), pp. 61-75.

Todeschini, R.; Consonni, V. 2000a. Handbook of molecular descriptors. Wiley-VCH.

Tomberg, A. 2013. Gaussian 09W Tutorial, an Introduction to Computational Chemistry Using G09W and Avogadro Software. pp. 1-34.

Tropsha, A. 2006. Predictive Quantitative Structure–Activity Relationship Modeling. In: Comprehensive Medicinal Chemistry II, 4, Oxford, Elsevier, pp 149–166.

Walter, J.C.; Barkema, G.T. 2015. An introduction to Monte Carlo methods. Journal of Physics A., 418, pp. 78–87.

Wojdyło, A.; Oszmian´ski, J.; Czemerys, R. 2007. Antioxidant activity and phenolic compounds in 32 selected herbs. Food Chemistry, 105, pp. 940–949.

Yang, B.; Zheng, J.; Laaksonen, O.; Tahvonen, R.; Kallio, H. 2013. Effects of Latitude and Weather Conditions on Phenolic Compounds in Currant (Ribes spp.) Cultivars. Journal of Agricultural and Food Chemistry, 61(14), pp. 3517–3532.

Yen, G.C.; Chen, H.Y. 1995. Antioxidant activity of various tea extracts in relation to their antimutagenicity. Journal of Agricultural and Food Chemistry, 43, pp. 27–32.

Zhang, Y. 2005. TM-Align: A protein structure alignment algorithm based on the tm-Score. Nucleic Acids Research, 33, pp. 2302–2309.

Zheng, G.; Irle, S.; Morokuma K. 2005. Performance of the DFTB method in comparison to DFT and semiempirical methods for geometries and energies of C20-C86 fullerene isomers. Chemical Physics Letters, 412, pp. 210-16.