Babeş-Bolyai University

Faculty of Chemistry and Chemical Engineering

Doctoral School of Chemistry

Doctoral Thesis Abstract

Similarity-based estimations and associations of chemical properties/activities

Ph.D. Candidate: **DONATELLA BÁLINT (Căs. NAGY)**
Scientific Supervisor: **PROF. DR. LORENTZ JÄNTSCHI**

CLUJ-NAPOCA
2023

Doctoral Thesis Abstract

Similarity-based estimations and associations of chemical properties/activities

Ph.D. Candidate: **DONATELLA BÁLINT (Căs. NAGY)**

President:

Prof. Univ. Dr. Radu-Lucian SILAGHI-DUMITRESCU – Faculty of Chemistry and Chemical Engineering, Babeş-Bolyai University, Cluj-Napoca

Scientific advisor:

Prof. Dr. Lorentz JÄNTSCHI – Technical University of Cluj-Napoca, Department of Physics and Chemistry

Reviewers:

Conf. Univ. Dr. Réka BARABÁS – Faculty of Chemistry and Chemical Engineering, Babeş-Bolyai University, Cluj-Napoca

Conf. Univ. Dr. Mădălina Ana VĂLEANU – Iuliu Hatieganu University of Medicine and Pharmacy, Cluj-Napoca

C.S. I Dr. Attila BENDE – National Institute for Research and Development of Isotopic and Molecular Technologies, Cluj-Napoca

Date: July 10, 2023
Location: Babeş-Bolyai University, Faculty of Chemistry and Chemical Engineering, Cluj-Napoca

KEYWORDS: molecular modeling, biochemical similarity, geometry optimization, amino acids, gaussian, structure/property relationship

LIST OF PUBLICATIONS

1. 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. Doi: 10.1155/2016/8578156.

2. 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. 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. Joita, 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.

Table of content

Thesis Objectives

The research reported in the present thesis provides an interdisciplinary framework for explaining the molecular behaviour of chemical structures as well as their impacts at the level of biological systems using concepts from chemistry and biology-biochemistry.

The purpose of the study was to explain and comprehend how the molecular structures differ and to classify the examined compounds based on their similarities. To achieve the optimized molecular structures several approaches were considered: multiple linear regression models, the eigenproblem application, factorial study, iterative algorithms, and geometry optimization calculations were performed using a variety of techniques (Hartree-Fock Methods, Semiempirical Methods, Density Functional Theory, Molecular Mechanics).

In order to understand how these strategies, relate to one another and choose which to apply in various situations, the study set out to analyse these methods.

After every part of our research, statistical analysis was performed to validate our results (principal component analysis, cluster analysis, analysis of variance, other data mining methods) and to compare the similarities of various approaches.

Chapter I - Introduction

1. Structure of chemical compounds

A chemical compound's structural formula is a graphical representation of the molecular structure that indicates how the atoms are grouped in three dimensions (Figure 1).

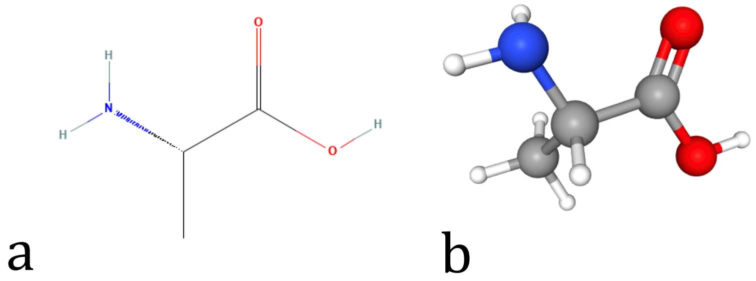


Figure 1. Schematic illustration of the L-Alanine molecule 2D(a) and 3D(b) forms (PubChem Database, Accessed on 17.04.2023)

There is a very clear relationship between the chemical compound properties and their structure, in that the properties are determined by the structure, and certain structural aspects can be inferred from the evaluation and interpretation of the properties (Figure 2).

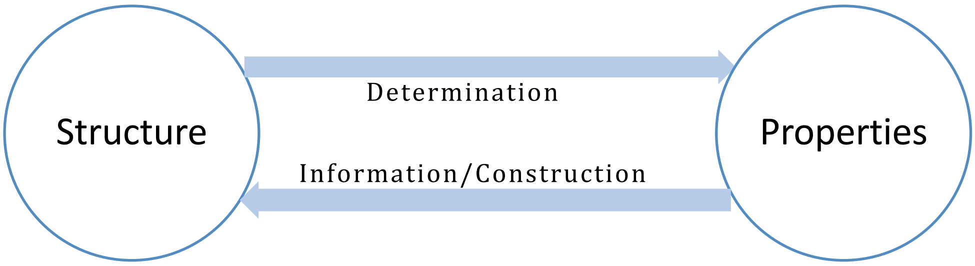


Figure 2. Structure/Property relationship

Different types of geometrical symmetry can be seen in molecules. Geometrical symmetry in molecules refers to the symmetrical arrangement of atoms and bonds in a molecule. It is determined by the presence of symmetry elements such as rotation axes, reflection planes, inversion centres, and improper rotation axes.

A graph, $G = G\ (V, E)$ is a pair of two sets: $V = V(G)$, a finite nonempty set of N points (vertices) and $E = E(G)$, the set of Q unordered pairs of different points of V.

Two vertices are adjacent if they are connected by an edge, and every pair of points represents a line (edge). When two separate edges intersect at a single point, they are said to be adjacent edges. Many times, the hydrogen atoms are left out (Diudea el al., 2002).

There are several different types of graphs, some examples of them are represented below (Figure 3):

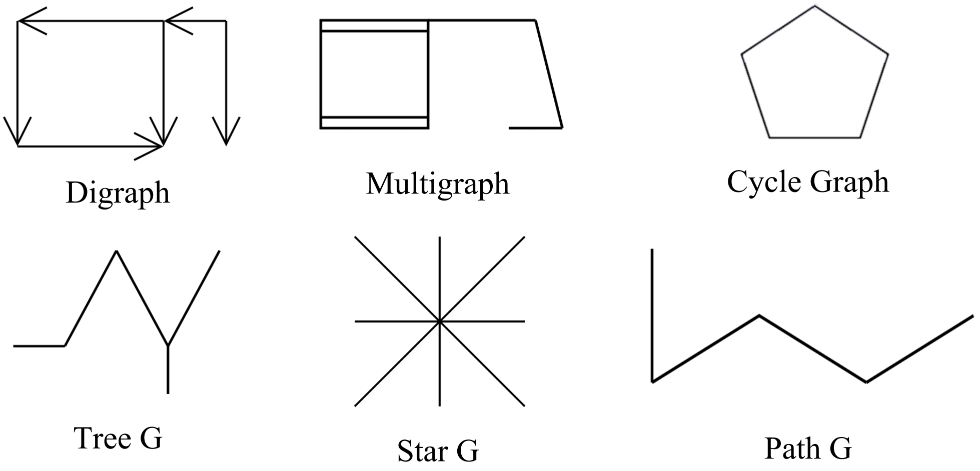


Figure 3. Examples of some graph representations

A molecular graph with the atoms as its vertices and the covalent bonds as its edges can be used to describe the structural formula of a chemical molecule.

2. Molecular similarity

The identification of various models can be facilitated by the similarity of two chemical structures (Doucet and Weber, 1996). Other techniques are needed to calculate how similar different molecular structures are to one another (Bender and Glen, 2004).

The procedure of evaluating molecular similarity is used to evaluate the structural characteristics of two or more molecules. It is a critical stage in the development and design of new drugs since it helps in the identification of possible drug candidates based on similarities to existing active substances.

There are various techniques for evaluating molecular similarity, such as: 2D fingerprint-based methods (uses 2D molecular fingerprints to compare the structural features of molecules); 3D shape-based methods (uses the 3D shape of molecules to compare their structural features); machine learning-based methods (uses machine learning algorithms) (Stumpfe and Bajorath, 2011).

A measure of how much a pair of molecules' properties match is called molecular similarity. Numerous molecular characteristics, such as shape, electron density, electrostatic potential, lipophilicity, and refractivity can all be calculated (Allen et al., 2001).

For already-running programs, searching the database for the needed structure could take many days (Kolodny et al., 2005). Finding more useable answers may be made easiest by using new algorithms (Dong et al., 2018).

This is done using a variety of models, some molecular descriptors (Todeschini and Consonni, 2000), such topological indices, and/or regression analysis (Bolboacă and Jäntschi, 2013).

Chemical structures can be categorized using some similarity criteria thanks to topological characterization. Some fundamental statistics form the basis of the regression analysis.

3. Quantitative Structure-activity/Quantitative Structure-property relationship (QSAR/QSPR)

As a mathematical tool for quantitatively characterizing the relationship between chemical structure and biological activity/property for a specific set of molecules, the concept of QSAR/QSPR arose in 1937. (Reynolds et.al 1992; Hammet 1937).

Studies of the relationships between a structure's properties have a number of advantages. The equations produced from a structure-property investigation, for instance, can be used to estimate the unmeasured properties of related substances. The equations can be used to derive a more fundamental understanding of the roles that particular structural elements play in determining qualities.

After gaining this understanding, the data can be used to create fictitious structures that could have high property values. The structure-property equations can also be used to verify the accuracy of property values that have already been reported in the literature, some of which might have been measured or reported inaccurately (Nelson and Seybold, 2001).

The data collection, the selection of the variables, the building of the model, and the validation assessment are usually the four common stages used in QSAR/QSPR. (Golbraikh and Tropsha, 2000) (Figure 4).

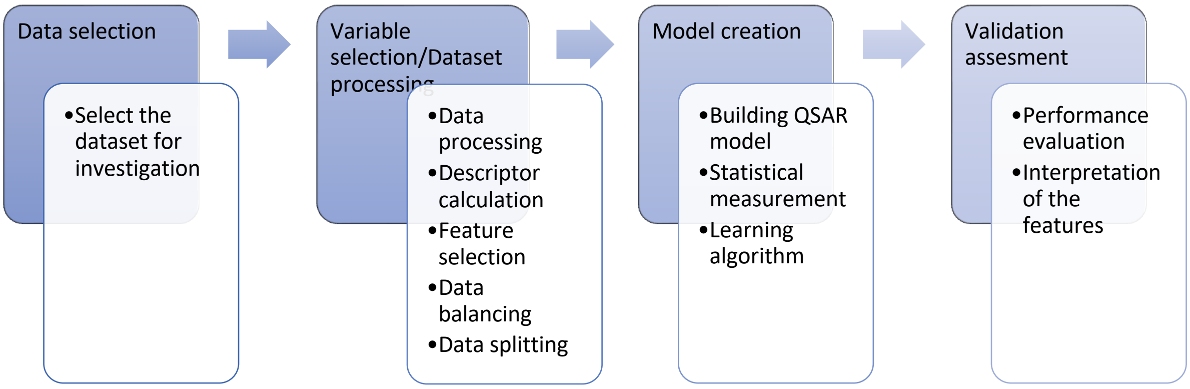


Figure 4. Example of the processes in QSAR/QSPR

Additionally, there are new initiatives in the QSAR literature aimed at the following problems: using conformal predictions, ascertain the degree of trust in predictions made using QSAR models; evaluate the flexibility of data sets to determine whether it is possible to create robust models; creating interpretable QSAR models that biologists and medicinal chemists may employ in practice; ensuring QSAR models can be replicated so that other research teams can use or expand on published models (Nantasenamat, 2020).

Quantitative structure-activity-property linkages, or mathematical approaches able to detect and quantify the relationship between chemical structure and activity/property, are used when the activity or property is a quantitative (linear models) or qualitative (non-linear models) variable (Godarzi et al. 2012). Various molecular descriptors collect the structural information (Jäntschi, 2005).

The process entails defining the peptide structure at the sequence level using amino acid descriptors (AADs) and associating it with observations using machine learning methods (MLMs). The output is a variety of quantitative regression models. These models are used to build new peptides with desired characteristics and to explain the structural elements that generalize known peptide properties to unknown samples (Lin et al., 2023).

Chemical similarity tries to quantify how similar two different molecules are to one another or to a certain feature. The similarity principle, which argues that similar compounds should have similar activities and attributes, is the basis for similarity assessment applications in the toxicology and pharmaceutical domains, which aim to predict the toxicity of chemicals.

4. Molecular optimization

It is possible to optimize the geometry of molecules using *ab initio* methods, semiempirical methods (to solve the Schrödinger equation with some approximations and describe the electron properties of atoms and molecules), and empirical force field methods (molecular mechanics, a quicker but less expensive method that can provide exceptional structural parameters)(Abegg and Ha, 1974).

The use of Slater type bases sets or Gaussian orbitals to represent the wave function was pioneered by John A. People (Pople, 1999). He chose a combination of approaches and data sets, defined models, and contrasted the analyses' experimental findings.

Calculations of atomic and molecular wavefunctions frequently use atomic orbitals of the Gaussian type (Figure 5). They contributed to the development of the Gaussian programs, one of the most widely used computational chemistry software packages.

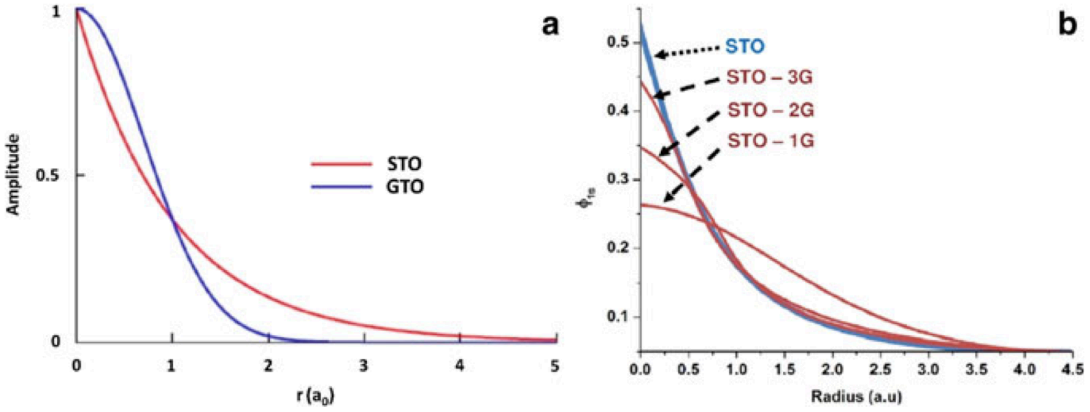


Figure 5. a. Assessment of Slater- and Gaussian-type orbitals; **b.** Schematic evaluation of the STO-1G, STO-2G, and STO-3G levels' results for a 1s Slater function's least-squares fit (Perlt, 2021).

In fact, it takes less time to compute several GTOs and combine them to represent an orbital than it does to compute a single STO. This is the rationale behind the widespread use of GTO combinations to represent STOs, which subsequently explains AOs.

The basis sets indicated by the sign "*" are the polarization basis sets, which contain the d orbitals. The 6-31G** basis is a further improvement, adding a set of p orbitals to each hydrogen in the 6-31G* basis set (Banerjee and Ramalingam, 2015).

If no basis set acronym is provided, STO-3G basis will be used. The basis sets STO-3G, 3-21G, 6-21G, and 6-31G are a few examples. The standard * or ** notation can likewise be used to request single first polarization functions. 6-31G* (or 6-31G(d)) is 6-31G with additional d polarization functions on non-hydrogen atoms; 6-31G** (or 6-31G(d, p)) is 6-

31G* plus p polarization functions for hydrogen. The + and ++ diffuse functions are obtainable with some basis sets. 6-31+G is 6-31G plus diffuse s and p functions for non-hydrogen atoms; 6-31++G also has diffuse functions for hydrogen.

Which basis set to employ depends on the calculation's goal and the molecules being examined. The agreement with experimental data is not always guaranteed, even with a large basis set (Petersson et al., 1998).

Various approaches to the comparison of the basis sets (Zheng et al., 2005; Scuseria, 1992) concur that even though they are comparable, they cannot be generalized. Several suggestions can be found in these publications and simply by reviewing the Gaussian09 lessons, including (Tomberg, 2013; Hill, 2012):

- a larger basis set is not necessarily better (ex: cc-pVQZ is excessive for Hartree-Fock)
- STO-3G should be applied only to very vast systems.
- usually cc-pVDZ is comparable or worse than 6-31G(d,p).
- usually cc-pVTZ is improved than 6-311G(d,p) or alike.
- *Ab initio* approaches settle relatively slowly.

These next bases sets are roughly corresponded to one another:

- 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

Functional cluster analysis (FCA), a different strategy, could be used to analyse multidimensional functional datasets utilizing orthonormalized Gaussian basis functions (Kayano et al., 2010).

The most typical answer is to complement experimental data with the best available *ab initio* data (from molecular orbital or density functional calculations). The ability to compare locations on a PES that are far from symmetry structures by straight calculation as opposed to attempting to understand vibrational spectra is a good aspect of employing theoretical data (Schlegel, 2003).

Personal contributions

Chapter II - Censored data in research calculations

1. Introduction

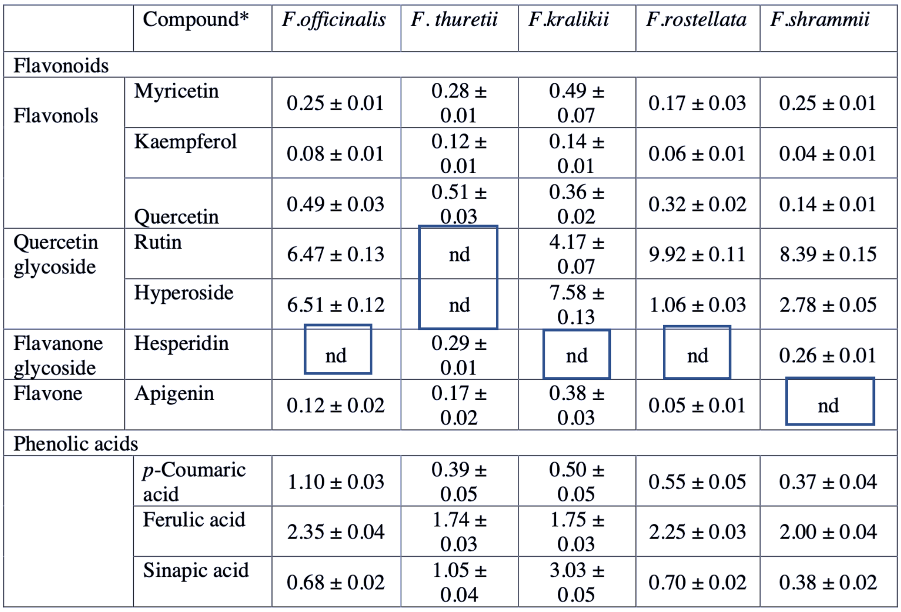
In this research, it was created an iterative method that can locate the most probable phenolic content values that are lacking (predictive power) after conducting the studies.

The investigations of the phenolic and flavonoid compounds due to their antioxidant activity are the subject of many research studies (Some examples are Wojdyło et al., 2007; Yang et al., 2013; Ivanov et al., 2014; Aryal et al., 2019; etc.).

This transparency is important for the interpretation and replication of the results. The next representations are some examples of datasets with missing values:

- HPLC analysis of polyphenols content in ethanol extracts of five Bulgarian *Fumaria* Species the 'nd' abbreviation stands for 'not detected' in Table 1 (Ivanov et al., 2014):

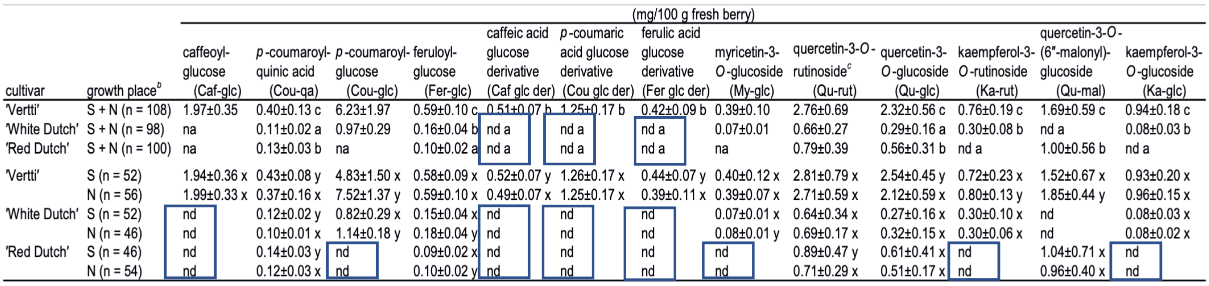
Table 1. Polyphenols content



	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

- Contents of Phenolic Compounds in Currant (Ribes spp.) Berries; the 'nd' abbreviation stands for 'not detected' in Table 2 (Yang et al., 2013):

Table 2. Contents of Phenolic Compounds



		(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

These are just a couple of examples for results with missing data, but several studies face the problem for different reasons.

The results presented by Wojdyło et al., 2007, our input data, with the missing places are given in Table S3 (Supplementary material section).

The outcomes of a quantitative examination of the 32 plants' main phenolic components are shown in Table S1 (Bálint and Jäntschi, 2019). The standard deviations calculated following the study are related to the mean values (Table 3).

Table 3. Outlier and average χ^2 value for the plant *Acorus calamus.*

Acorus calamus	**χ^2 (Outlier) Value**	**χ^2 (Average) Value**
ABTS	4.6788	0.1057
DPPH	4.8017	0.1045
FRAP	4.8000	0.0989

2. Methodology

The Jäntschi (2012) approach was adjusted and changed to fit our experimental results. All calculations were performed using custom *.php programs.

The Chi-square (χ^2) test was used to examine the connection between four chemical compounds (caffeic acid, p-coumaric acid, ferulic acid, and neochlorogenic acid) and their antioxidant activity (Bálint and Jäntschi, 2019).

The working algorithm is represented below (Figure 6):

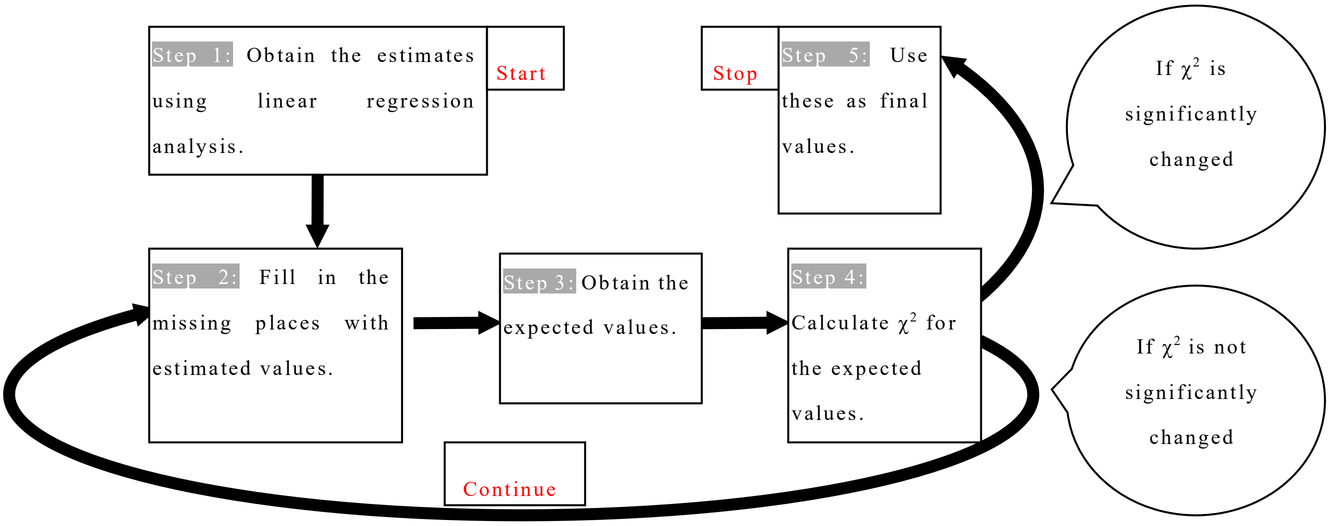

Step 1: Obtain the estimates using linear regression analysis.
Start
Stop
Step 5: Use these as final values.
If χ^2 is significantly changed
Step 2: Fill in the missing places with estimated values.
Step 3: Obtain the expected values.
Step 4: Calculate χ^2 for the expected values.
If χ^2 is not significantly changed
Continue

Figure 6. Working algorithm

The working algorithm:

The procedures used in our examination of the missing data were as follows: (Fig. 6):

- **Step 1**: Check to see if the connection between antioxidant activity and phenol content is linear using the experimental data.
- **Step 2**: Three options were taken into account. The algorithm's first stage included the introduction of the experimental values.
- **Step 3**: Using the coefficients to create estimates in the first cycle of the linear regression analysis.
- **Step 4**: With approximated values, fill in the blanks.
- **Step 5**: Reiteration:
  - Acquire (new) probable values.
  - Estimate χ^2 using observed and expected values.
  - Addition in the absent places the (new) expected values.
- **Step 6**: Until the value of χ^2 is not considerably changed (e.g. convergence)

These steps were taken in order to fill in the gaps in the contingency tables based on the structure of the phenolic component and its antioxidant activity.

The algorithm cycles are represented graphically in Figure 7. The different colour combinations imply that the values in the empty spaces have been changed (in red).

A comprehensive set of estimates, predicted values, and χ^2 computations are also included in each cycle. Changes in data in missing areas are indicated by different colours;

these changes attain their ultimate values in Cycle *n*, after χ^2 has not considerably altered from Cycle *n-1*.

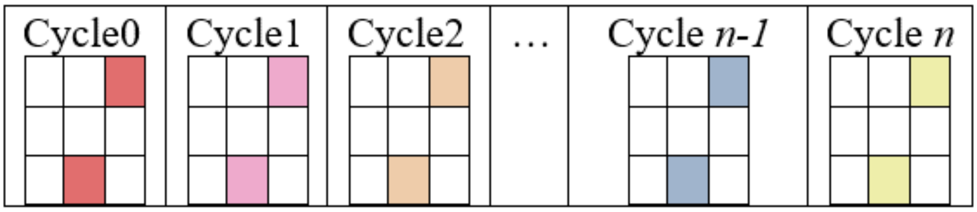


Figure 7. Predicting the values that are absent through time.

Each value in the columns has an impact on each value in the rows, according to the basic premise of the hypothesis (Bálint and Jäntschi, 2019).

Correlation coefficients were determined (Pearson, Spearman, semi-quantitative, see below) after the blank columns were filled in. The Pearson's quantitative and Spearman's rank qualitative coefficients combine to form the semi-quantitative coefficient.

The degree of the monotonic non-linear inference was measured using these coefficients (such as sigmoidal extremal deviations).

3. Results and Discussion

The acquired data were subjected to correlation analysis in order to determine which phenolic compound influences the antioxidant activity after the missing values were filled in the contingency table. The statistical study was necessary due to the fact that each phenolic acid makes a unique contribution to the antioxidant capacity. The following table contains the findings of the correlation analysis.

Because the TEAC (total equivalent antioxidant capacity) experimental values χ^2 were outliers, they were not used in the future calculations. Due to its outlier χ^2 results (Table 3 - see above), the investigational data from the plant *Acorus calamus* was omitted. The link between the four phenolic acids and the antioxidant activities of the remaining plants was examined.

The values for the observed experimental results (obs.), the estimated values that filled in the gaps, and the expected values (exp.) resulting from the regression are shown in Table S4 (Supplementary Material section). The results of the analysis after the data have been adjusted logarithmically are shown by the values (Bálint and Jäntschi, 2019).

Because both the experimental data and the pre-set values are progressing in the same route, it should be noted that the values of each plant are comparable. The findings are inconclusive on whether phenolic acid accurately predicts antioxidant activity. Each substance plays a part in how its effect is delivered.

Correlation analysis was performed after the results were gathered to determine which phenolic component influenced the antioxidant activity.

Table 4 (see below) contains the correlation coefficients.

Table 4. Correlation coefficients were calculated.

		ABTS	**DPPH**	**FRAP**
Pearson's quantitative correlation and significance levels from Student's t	ABTS	-	0.774	0.758
	DPPH	$4.881\cdot10^{-26}$	-	0.669
	FRAP	$1.837\cdot10^{-24}$	$1.858\cdot10^{-17}$	-
		ABTS	DPPH	FRAP
Spearman's qualitative correlation and significance levels from Student's t	ABTS	-	0.774	0.754
	DPPH	$3.333\cdot10^{-26}$	-	0.668
	FRAP	$3.049\cdot10^{-24}$	$1.637\cdot10^{-17}$	-
		ABTS	DPPH	FRAP
Semi-quantitative correlation and significance levels from Student's t	ABTS	-	0.774	0.756
	DPPH	$4.033\cdot10^{-26}$	-	0.669
	FRAP	$2.369\cdot10^{-24}$	$1.744\cdot10^{-17}$	-

They showed that there was little distinction amid Pearson's and Spearman's correlation coefficients. Both are almost equally important. The sole distinction is that Spearman's correlation employs ranks rather than Pearson's x and y values.

The correlation coefficients in the previous table showed a significant link between the outcomes. In this case, the average value of 0.75 showed that there is a linear increase in the relationship between the variables. The coefficients can take values between -1 and +1.

The variables have a statistically significant linear relationship, according to the *Student's t*-test. Each correlation coefficient describes the link between two variables and explains a measure of association between them. Every coefficient change in the same way when the variation between them is not large.

When missing values are not missing at random (MNAR), the χ^2 test can be used to analyse the data. The missing values in this instance can be connected to the value itself or to other dataset variables. The χ^2 test can be used to assess whether there is a statistically meaning connection between the missing data and the other variables by comparing the distribution of missing values across several groups of variables (Agresti, 2007).

Following the algorithm's execution, the relationship between χ^2 and iteration was also examined. This demonstrated, much like the statistical analysis, how closely related the variables are. The following Figure 8 displays the evolution of χ^2 as a function of iteration.

The resulting values of the χ^2 quick met to a minimum after implying the procedure on the experimental data set after different numbers of cycles.

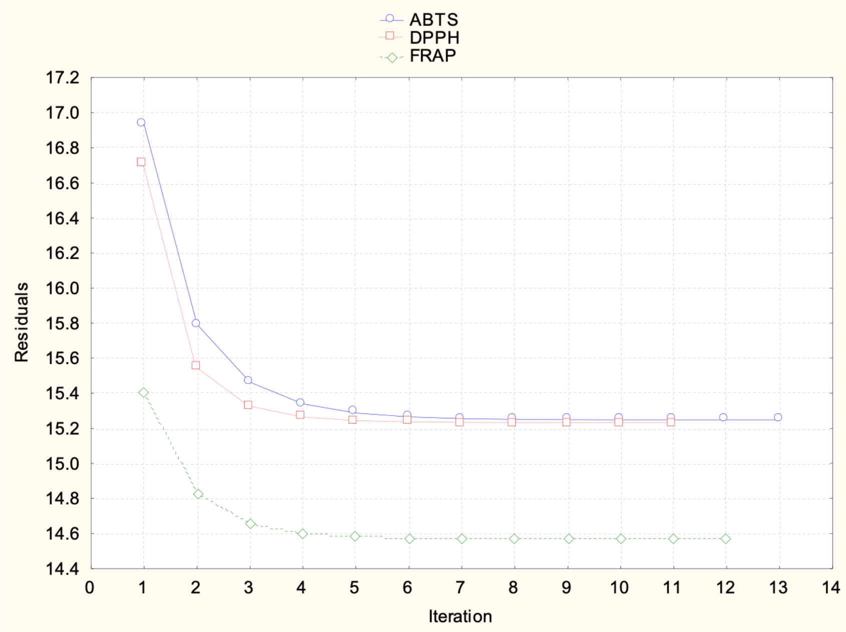


Figure 8. χ^2 as function of iteration

Within a few iterations, the minimum was attained. The procedure was stopped at the 12th iteration for ABTS, the 10th iteration for DPPH, and the 11th iteration for FRAP values utilized in calculations due to a non-significant alteration between subsequent χ^2 results (Bálint and Jäntschi, 2019).

By estimating the predicted frequencies of the missing data using the observed frequencies in the available data, the χ^2 test can also be used to impute missing values. The term "chi-squared imputation" refers to this method. The χ^2 test can then be used to ascertain whether the missing data are connected to the other variable in that situation (McDonald, 2014).

The goodness of fit of computational models to experimental data, such as that from molecular docking and binding investigations, is assessed using the χ^2 test. Additionally, it is employed to determine how similar various molecules are to one another and to optimize the computational model's parameters.

In the scientific literature, it was noted that there were more ways to fill the contingency table (is a specific type of matrix-style table that shows the variables multivariate frequency distribution), such as Monte Carlo techniques. Most Monte Carlo simulations start out by proposing a minor random change to a setup repeatedly (Walter and Barkema, 2015).

The Bootstrap method, one of the Monte Carlo techniques, would work with our collection of data. By selecting a random variety of sets, from the data set, bootstrapping tries to mimic the impact of utilizing a larger data set. Some of the data in each randomly chosen set will appear more than once, while other data will not be present at all (Tropsha, 2006). This method is very similar to our approach.

The results from the Monte Carlo methods, when taken into account, would not provide any new information based on the variables. The Bootstrap approach also has the drawback of ignoring known variables, even when there is a linear association amongst them.

With our proposed method, the contingency table is initially populated, and then the variables are processed by the χ^2. The table is repeatedly populated using the Monte Carlo methods, changing the initial variables.

On the same subject, numerous studies (Wojdyło et al., 2007; Aaby et. al, 2004) produced complex and varied results. The determined compounds differ in terms of geometry and symmetry as well as properties. Finding the best choice is the key to solving the censored data issue. Since it provides the quickest route to the problem's resolution, we employed our methods to fill the contingency table.

Numerous scientific areas use the preceding algorithm when displaying missing data from studies (Ivanova et al., 2005; Luo et al., 2004).

Clinical trials and epidemiological research frequently use censored data. They occur in large-scale research where the occurrence is often linked to an illness, infection, or other failure (Yen et al., 1995; Arnao, 2000). The promptness and quickness of the algorithm are essential to the originality of our work.

4. Conclusions

On a contingency table with breaches (missing data), our algorithm demonstrated its ability to function. The χ^2 statistic is minimized in the procedure.

All the investigated datasets show a linear connection. The results from all three techniques—ABTS, DPPH, and FRAP—used to assess antioxidant capability were equivalent. Based on experimental results and literature reviews, this is what is anticipated. The projected values for the empty spaces match the results of the experiment.

Chapter III - Similarity evaluation using the characteristic polynomial function

1. Introduction

Protein representations either use characterizations that are insensitive to the amino acid ordering that has been chosen or methods that treat any presumed amino acid assignment as strictly equivalent.

The matrix eigenvalues and coefficients of the characteristic polynomial are the two most frequent matrices invariants. The total of the matrix elements above the main diagonal, has been utilized as a molecular descriptor in chemical applications. The characteristic polynomial's eigenvalues and coefficients, both of which naturally form an ordered sequence, result in additional invariants that are provided as an ordered sequence (Randic et al., 2008).

We obtain a pool of polynomials and a pool of compound features following the computations on the 10 amino acids (Supplementary Material - Table S1). The goal is to develop a program that uses polynomials to relate the structures to the properties.

This polynomial can be used to calculate various molecular descriptors, such as the Wiener index (Todeschini and Consonni, 2000) and the Randic index (Randic, 1975), which are used to evaluate the topological similarity between molecules.

2. Methodology

Calculating the eigenvalues of the related matrices is necessary to determine the characteristic polynomial equations for each of the 10 essential amino acids.

The matrices used to represent amino acids can be obtained from a variety of structural or chemical features, such as the atoms' three-dimensional (3D) coordinates or the side chains' electronic structure. Following the creation of the matrices, the characteristic polynomial equation was found using conventional linear algebra methods.

The amino acids presented before are basic compounds in biological systems. The goal was to determine whether the differently labelled chemical structures really differ. All the structures were collected from PubChem databases and choose them based on their complexity and number of isomers (PubChem Database).

Following the selection, the next algorithm was used:

- ➢ **Step 1**: collect of all amino acid structures with the same molecular weight from the PubChem databases.
- ➢ **Step 2**: convert the .sdf structures to .hin files with a home-made *php program.
- ➢ **Step 3**: enter the .hin files to the program from http://l.academicdirect.org to calculate the characteristic polynomial matrixes, after the next path:
    - ➢ enter the site mentioned above → Fundamentals → Graphs → polynomials → a_characteristic_polynomial_in.
- ➢ **Step 4**: collect of the equations from the matrixes given by the program.
- ➢ **Step 5**: analyse the collected equations (sorting, clustering)
- ➢ **Step 6**: discuss of the obtained data.

We selected 10 out of 20 essential amino acids based on the chemical structure: alanine, glycine, valine, leucine, isoleucine, lysine, serine, threonine, aspartate and glutamate.

3. Results and Discussion

The data was sorted after the amino acid analysis and after getting the characteristic polynomial equations. The equations derived from the characteristic polynomial matrices describing the similarity between the molecules are provided in Table 5 below.

Table 5. Characteristic polynomial equations calculated

Amino Acids	***ChP* Equations**
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	$=+1X^{24}-23X^{22}+213X^{20}-1033X^{18}+2868X^{16}-4689X^{14}+4476X^{12}-2388X^{10}+640X^{8}-64X^{6}$

The findings showed that around 95% of the conformers of the various amino acids are identical. Most of the time, the modification happens between the zwitterion and other conformers, but occasionally, different variations can be seen.

Following the data sorting process, the zones "mix" and it becomes apparent that leucine, isoleucine, and a type of glutamate are segregated from other conformers. Their distinctive polynomial equations complement other molecules or conformers more effectively.

The characteristic polynomial is the same for two comparable matrices. Contrarily, this is not always the case: two matrices with the same characteristic polynomial do not necessarily have to be identical.

It's important to remember that the characteristic polynomial might not be the best way to compare amino acids. It is mostly used to calculate eigenvalues, which are mathematical characteristics of a matrix that might not have a clear biological or chemical meaning. For researching amino acid characteristics and interactions, other methods might be more suitable, including comparing amino acid sequences or structure similarity analysis.

Pan (2012) investigated the circumstances under which identical matrices of two characteristic polynomials can resemble one another. He concluded that: if the two n-order matrices (A and B) in the amount ground F have characteristic polynomials that are identical, have n single roots, and are identical to A and B; set A and B as two n-order square matrices in the field F, as a result, A and B are comparable. If the requirement (that the distinctive roots of A and B are in F) can be met.

In the study of Garcia-Planas (2021), due to their scientific applications, pairs of matrices under similarity are taken into consideration. The primary objective of that work is to create connections between the confined geometry surrounding one point and the local geometry surrounding another point, using the characteristic polynomial associated with each matrix of the pair.

4. Conclusions

The study determined the distinctive polynomial equations for the 10 essential amino acids using a matrix formed from the 3D structure of the amino acids.

It was discovered that each amino acid has an own distinctive polynomial equation that could be utilized to distinguish between amino acids.

The study's goal was to compare the structures of the 10 essential amino acids according to their characteristic polynomial function derived.

The characteristic polynomial has been used in the comparison of molecular structures in general, as it provides a unique representation of the structure that is invariant under various transformations. We can state for the fact that there are no appreciable differences among the conformers that were examined.

Chapter IV - Similarity evaluation using the *Gaussian09* Software package.

1. Introduction

The Schrödinger equation's solution is the key issue in electronic structure theory. With a few exceptions, like the hydrogen atom, it is a many-body problem, meaning that numerical solutions must be calculated. The wave function is enlarged in relations of a basis set for this use. A natural place to start in molecular quantum chemistry is with basis sets derived from atom centred orbitals.

Because of this, plane waves are a fairly common choice for these systems' basis functions. However, local basis functions are also commonly employed (Perlt, 2021).

Finding the best optimization strategy for scientific computations is really difficult because there are so many different base sets and optimization techniques.

After consulting the specific literature, the following research questions were formulated:

1. The bigger the bases, the better?
2. Which bases set family is the best? Is it possible to find one?
3. Do we obtain different outcomes if we employ two separate basis sets from different families?
4. Can they be correlated in any way?
5. Does the difference depend on the molecule or the bases set?

The goal of the study was to examine 39 optimization techniques to discover their relationships and choose the best one to apply in various situations (Bálint and Jäntschi, 2021).

To compare the similarity of various approaches, cluster analysis, correlation analysis, statistical analysis (ANOVA), and principal component analysis (PCA) were carried out.

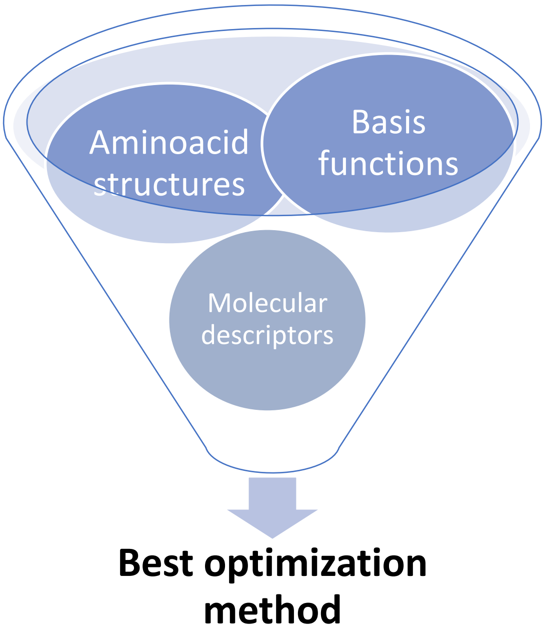


Figure 9. Representation of the main components of the analysis

The computations were performed using a computational chemistry software program called *Gaussian09,* which in general is used to simulate chemical systems and calculate their electrical structures.

Gaussian09 contain various electronic structure models, such as post-Hartree-Fock techniques, Density Functional Theory, and Hartree-Fock approaches; a range of basis sets that let users decide how accurate and expensive their calculations should be; the capacity to carry out transition state searches and geometry optimization and a lot more features (Bálint and Jäntschi, 2021).

2. Methodology

The 20 essential amino acids analysed (Supplementary Material – Table S1 and S2) can be isomers, enantiomers, and conformers among other forms. Most of these chemicals have the same chirality in biological systems, and the majority of amino acids are levorotatory (L) rather than dextrorotatory (D). Geometry optimizations were carried out on the structures using the L conformer of these compounds (Table 6).

Table 6. Geometry optimization methods used in the calculations.

Gaussian Optimization Methods	
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)

A group of molecular descriptors (FMPI- Fragmental Matrix Property Indices) (Jäntschi and Bolboaca, 2016) to assess the level of similarity between the methods were also created following the Gaussian09 program's calculations based on the 39 methods chosen before (Bálint and Jäntschi, 2021).

As shown, after Figure 10, we obtained the three-dimensional (3D) structures of the 20 amino acids (L conformers) from PubChem databases (.sdf files) and used the Gaussian09 program to analyse them after the following protocol:

- **Step 1**: Insert the downloaded PubChem .sdf files to the Gaussian09 package.
- **Step 2**: Convert the file in .gjf file format (the input file format for the database)
- **Step 3**: Examine the amino acids after the next instruction is entered in the program:
  - Calculate → Gaussian09 Calculation Setup → Job type (Optimization) → Method (Ground State) → Set the chosen method (HF, DFT, etc…) → Submit the job (Figure 10)
- **Step 4**: Also, we created a .bcf file to automatize the optimization process.
- **Step 5**: Run the computations by selecting each Gaussian09 Geometry Optimization Method in turn from the Calculation Setup menu (Figure 10)
- **Step 6**: Save the .out file (the output file format for the program) after each calculation.

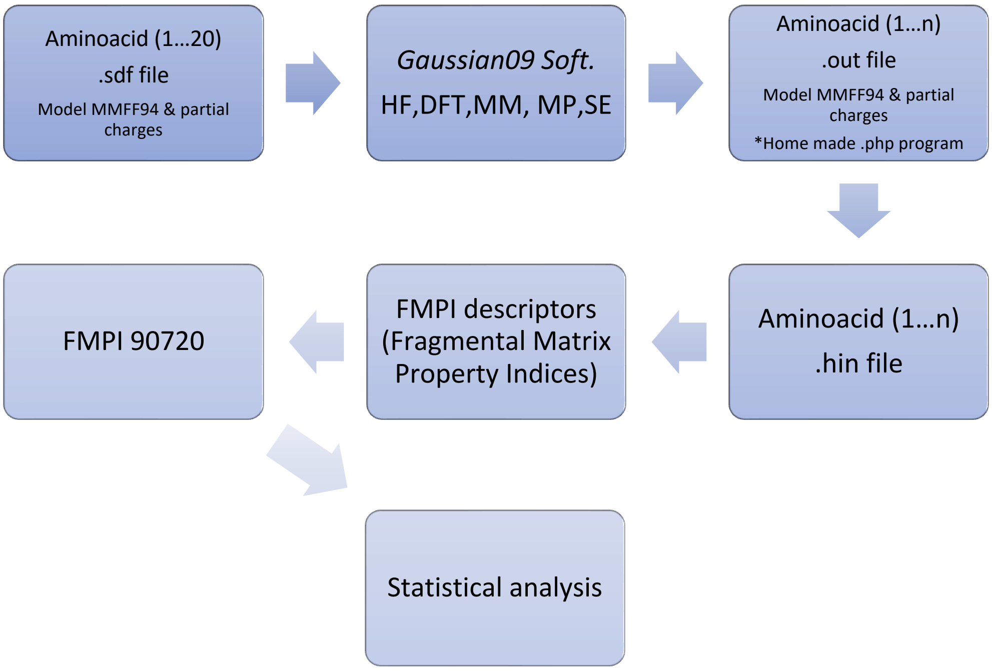


Figure 10. The algorithm of the main workflow

The molecular descriptors (FMPI) were generated and .hin files from the .out and .sdf files using a custom *php software.

The *Statistica* program performed the statistical analysis once we got 90720 descriptors for each amino acid analysed.

A PCA, Clustering analysis, ANOVA and Correlation analysis were carried out. The literature framework provides a thorough description of the PCA.

Each statistical analysis that was used has a common objective from several angles.

The expected results were as its follows:

- To obtain specific results for every optimized molecule.
- To get the molecular descriptor family for every method applied.
- To obtain statistical results for the methods applied.
- To find the greatest Molecular Optimization method to practice in certain conditions based on the molecular descriptors obtained.

3. Results and Discussion

The number of molecules or variables (descriptors) in the initial data set, whichever is lower, is the number of principal components that can be estimated.

To fully describe all of the variance in the data, all of the principal components must typically be taken into account. Due to correlations between the initial variables, many data sets only need a small number of principal components to describe a sizable portion of the variation.

The variables from the entire data set are combined linearly to form each principal component. 3.538.080 descriptors total, 90720 descriptors per component and technique, were examined (Bálint and Jäntschi, 2021).

The results of the PCA examination showed that the principal components (Figure 11) accurately represented the variation in our enormous amount of data by explaining 99.8851% of it.

The first and second components are displayed in the following figure (Figure 11) as the outcome of the PCA analysis.

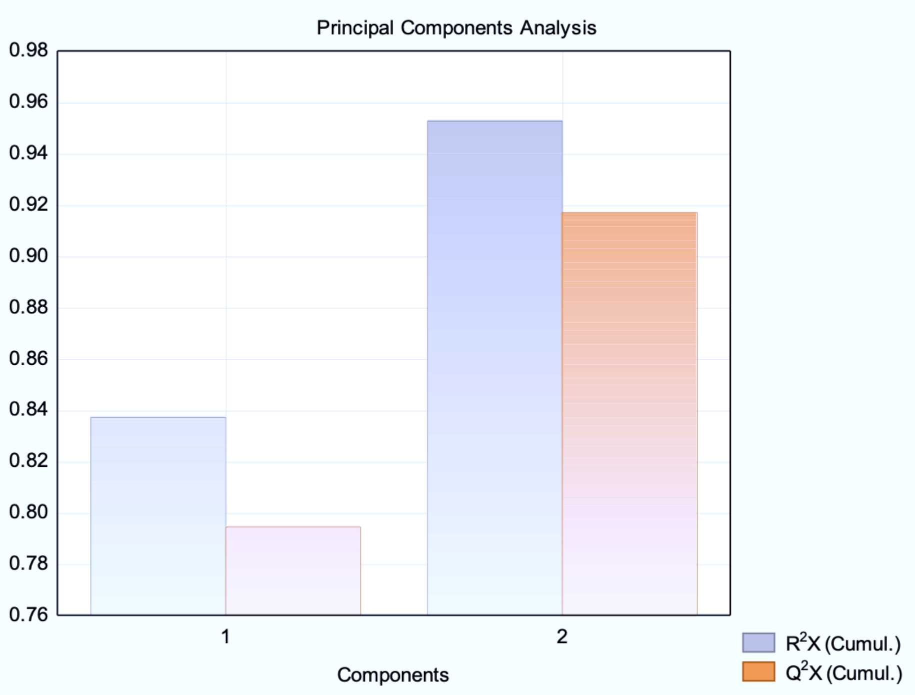


Figure 11. The explained variation (R^2X) and the predictive variation (Q^2X) of the PCA elements

The highest eigenvalue is associated with the first principal component, which accounts for the greatest proportion of the variance; The second most significant eigenvalue is linked to the second principal component, and so on. The eigenvalues display the proportion of variance that each major component contributes to.

The first component explains the most variance overall (71.25%) in the studied data. The maximum variation (14.9%) not covered by the first component is covered by the second component. After the first two, the third constituent also explains the greatest variance (6.51%).

The R^2X uses values between 0 and 1 to describe predictive accuracy. A principal component's R^2X increases with its relevance. Simply put, the explained variance (R^2X_{adj}) is the explained variance R^2X with the degrees of freedom taken into account.

Cross-validation is frequently used to present the quality evaluation, goodness of prediction (Q^2) statistic, which provides a qualitative measure of consistency between the predicted and original data. The value of Q^2 rises as supplementary variables are included in the PCA analysis. Large values of Q^2 suggest that the analysis was meaningful and pertinent.

The following loading plots display the coefficients for each of the descriptors in the various main components. This demonstrates how each descriptor contributes differently to the various main components (Bálint and Jäntschi, 2021).

A loading plot showing the distribution of component 1 versus component 2 is shown in the following figure (Figure 12). The plot suggests that similar techniques are in fact clustered together.

The primary components' spatial orientation is also determined by the loadings. The vectors for loading are p1 and p2.

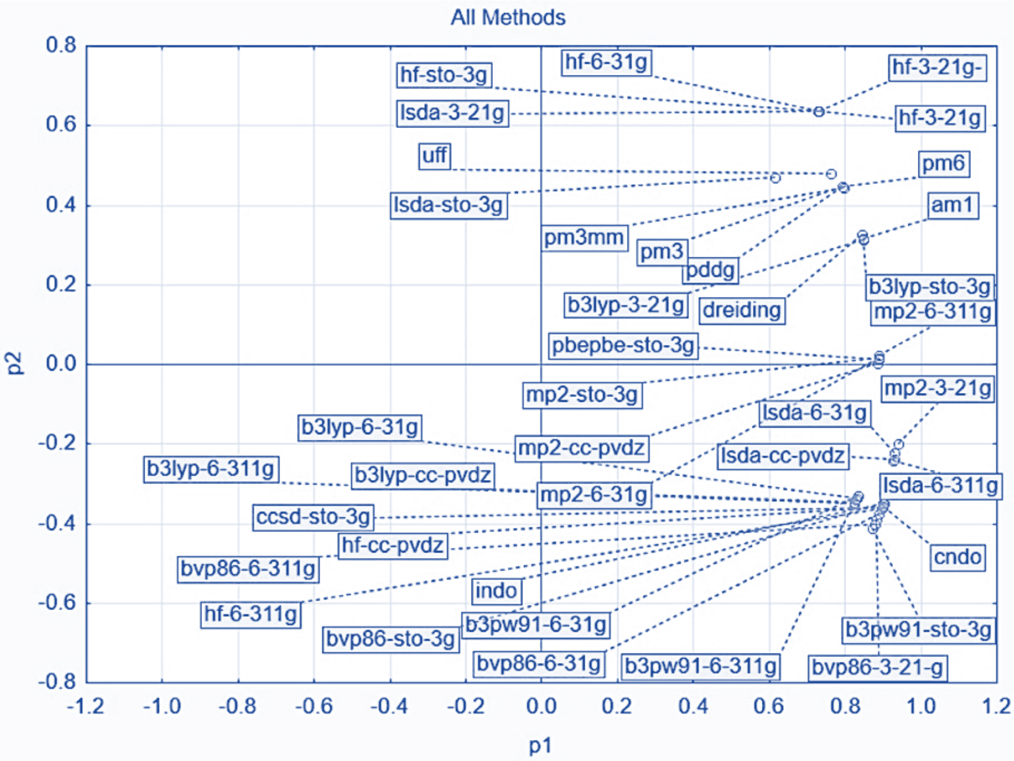


Figure 12. Score plot displaying the distribution of the techniques in the p1 and p2 main components.

In our case, the first three factors accounted for the majority of the data. The previous graphic displays a loading plot that compares the distribution of component 3 to component 2 (Figure 13).

If it’s taken into account how similar they are, the methods classification in various categories—Semiempirical, Density Functional Theory, Molecular Mechanics, Møller-Plesset perturbation theory, Coupled-cluster theory, and Hartree-Fock—differ from the Materials and Methods section.

Based on how closely the approaches resembled one another, the data were divided into various categories. Results from the PCA and cluster analysis were equivalent.

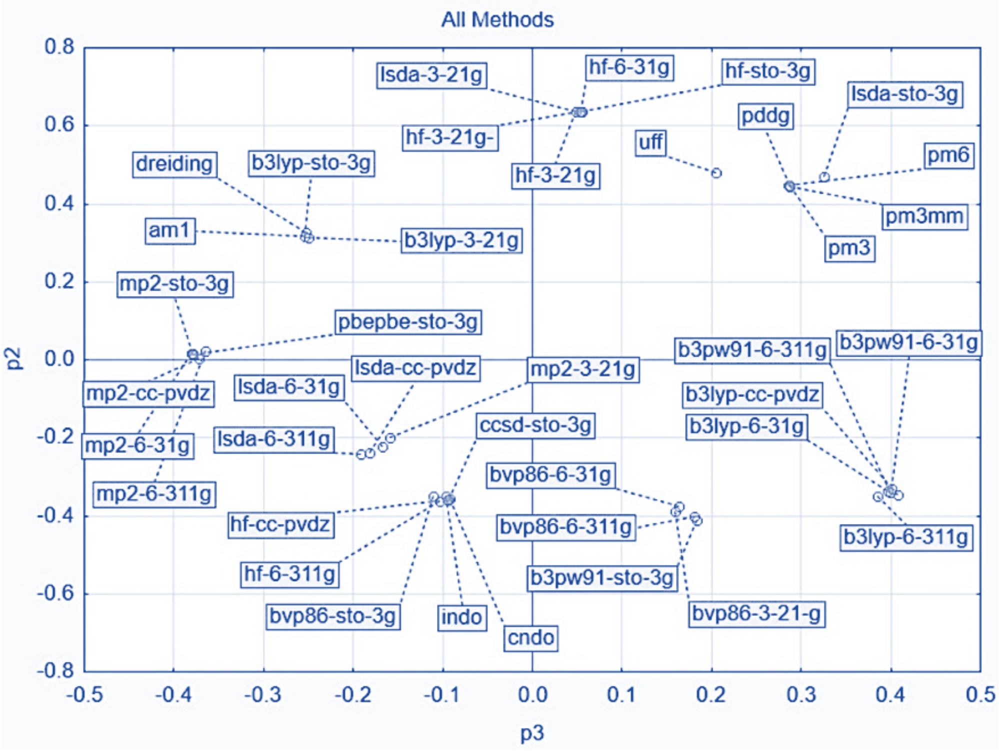


Figure 13. Score plot displaying the distribution of the approaches in the principal components p2 and p3.

The Euclidean distances among the 39 approaches under comparison are displayed in the cluster analysis dendrogram (Figure 14). The closest neighbour or single linkage method is one of the most straightforward hierarchical clustering techniques (Bálint and Jäntschi, 2021).

Due to the size of the data set (3.538.080 variables), the Euclidian distance between the approaches is extremely high. The adjustment of the linkage distance on the Y-axis was picked in order to compare different approaches. The linkage distances (Dlink) divided by the maximum linkage distance (Dmax) are represented by (Dlink/Dmax) *100.

The chosen basis sets affect how comparable the optimization techniques are. A classification into several groups can be made when the PCA and clustering analysis findings have been obtained.

The differences between optimization techniques are negligible, and tree clustering revealed their connection. They ought to be chosen from distinct groups for a thorough investigation in order to obtain various conclusions from various viewpoints.

Hybrid approaches are used by several research for their analysis since they produce better outcomes (Scott and Radom, 1996; Batra et al., 1996; Russ et al., 2004).

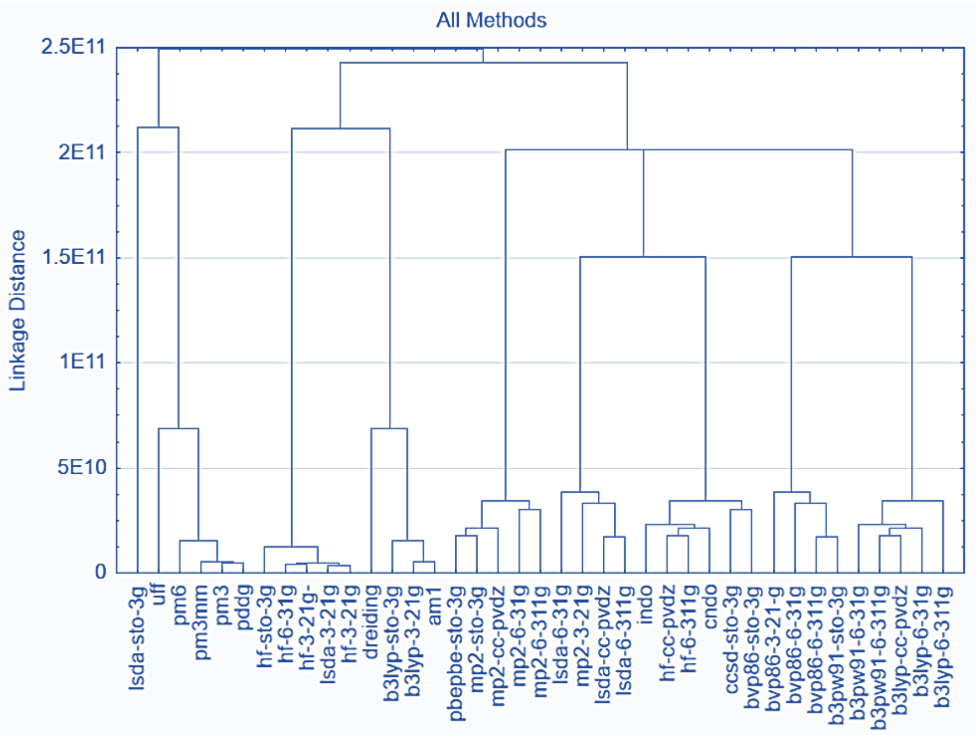


Figure 14. Clustering results

Although many more techniques have subsequently been introduced in computational chemistry, Davidson and Feller (Davidson and Feller, 1988) in 1988 portrayed a few principles on which a collection of the basis sets may be made.

Because different theoretical frameworks and molecular properties have different basis set requirements, various computer architectures and algorithms have different efficiency requirements, and the desired accuracy depends on the application, it is not practical to develop a single "optimal" basis set (Bálint and Jäntschi, 2021).

Cramer (2002) showed how split-valence basis sets evolved over time, starting with the most popular split-valence basis sets like 3-21G, 6-21G, 4-31G, 6-31G, and 6-311G and ending with more recent instances like cc-pCVDZ, cc-pCVTZ, etc.

It is clear from associating all the sets of assessments that it is much harder to predict with precision the geometries of second-row element-containing molecules than it is for simpler organics (Dunning, 1989).

For instance, it was discovered that PM3 outperforms AM1 when applied to these species (Stewart, 1989). The DFT approaches also have inherent drawbacks, including inconsistent trends and high error accuracy (Jensen, 2012).

With the spread of current methodologies, the search for the "optimal" combinations of methods and basis sets that statistically produce positive findings for a certain collection of molecules and attributes has been more pronounced.

In molecular modelling and drug design, energy minimization and geometry optimization are crucial techniques. Inaccurate molecular descriptors directly correlate with ineffective energy minimization and/or geometry optimization (Jäntschi, 2011).

The findings that we got after applying Cluster and PCA on each subgroup are as follows.

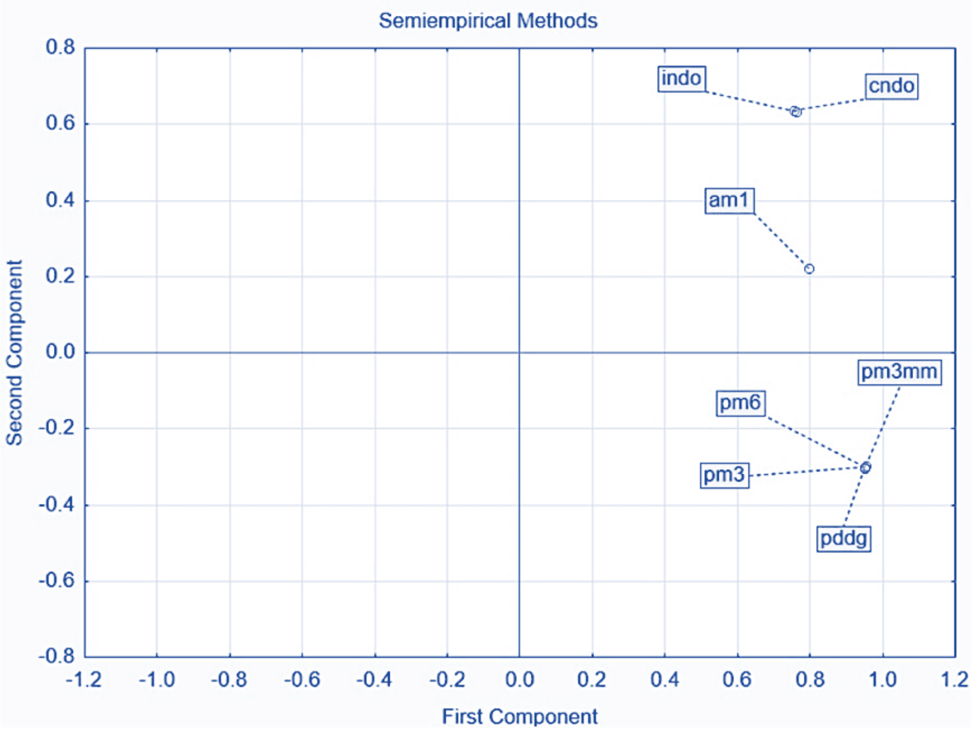


Figure 15. Score plot displaying the distribution of the approaches in the principal components p1 and p2

The results were divided into many subgroups because a relatively large data set was deployed, which makes the results more understandable(Bálint and Jäntschi, 2021).

The results can be classified into three primary groups: pm6, pm3mm, pm3, pddg; am1; indo and cndo. It should be sufficient to describe our data using a single technique from each group, in conclusion.

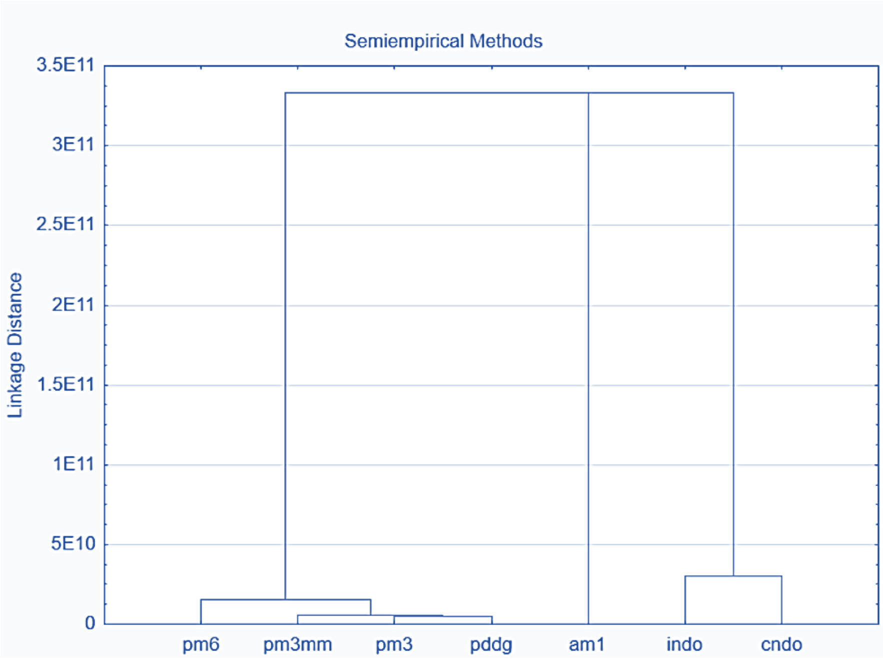


Figure 16. Clustering results

Due to the greater dataset this time, the statistical analysis for the Density Functional Theory approaches appears a little different. Most of the approaches that were examined belonged to this family.

Figures 17 and 18 show how the DFT approaches are comparable to one another while also showing several 'outlier' methods. The methods can be categorized into 4 main groups and 3 smaller groupings of methodologies.

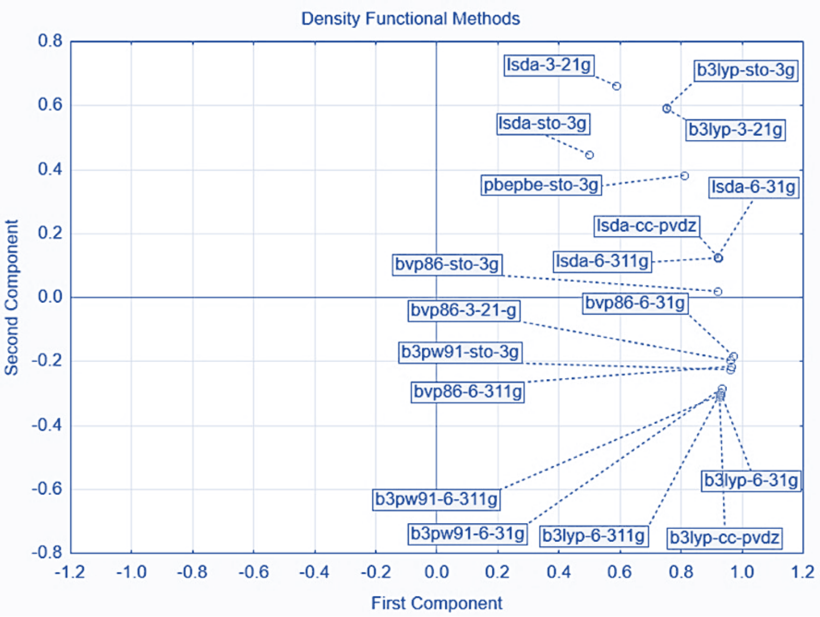


Figure 17. Score plot display of the distribution of the approaches in the principal components p1 and p2

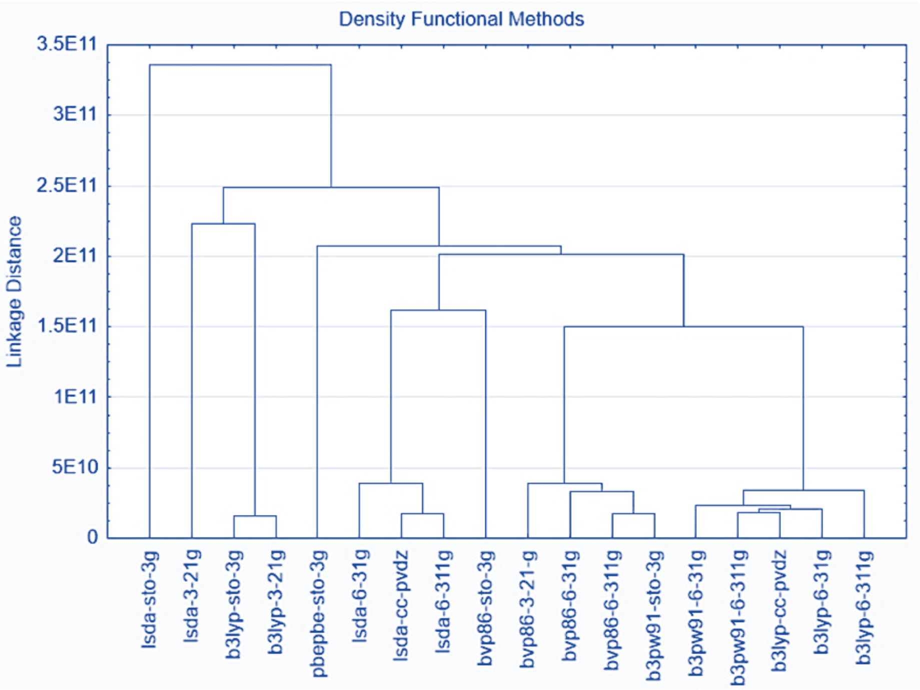


Figure 18. Clustering results

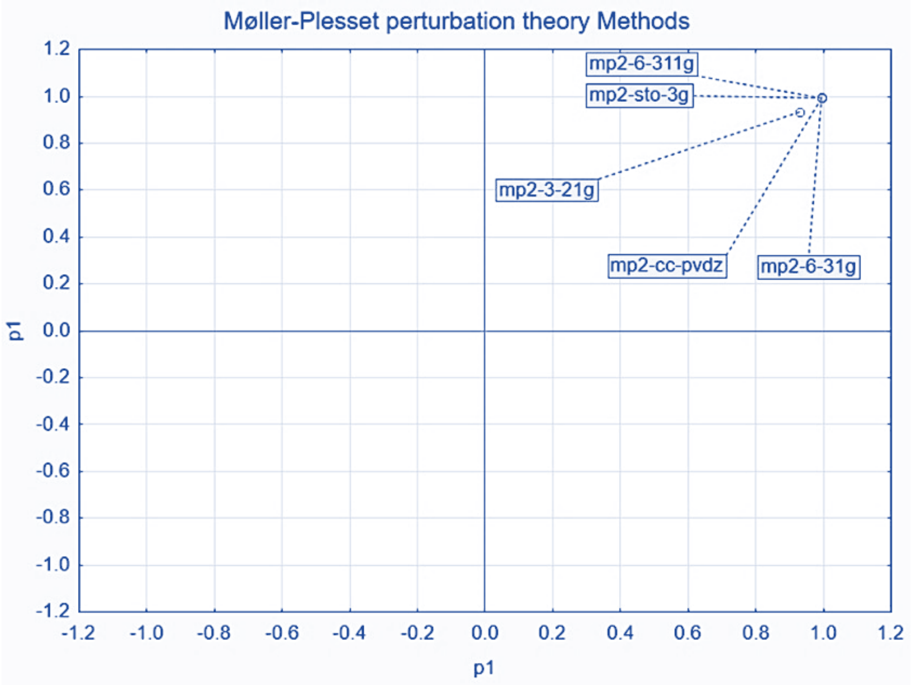


Figure 19. Score plot displaying the distribution of the approaches in the principal component p1

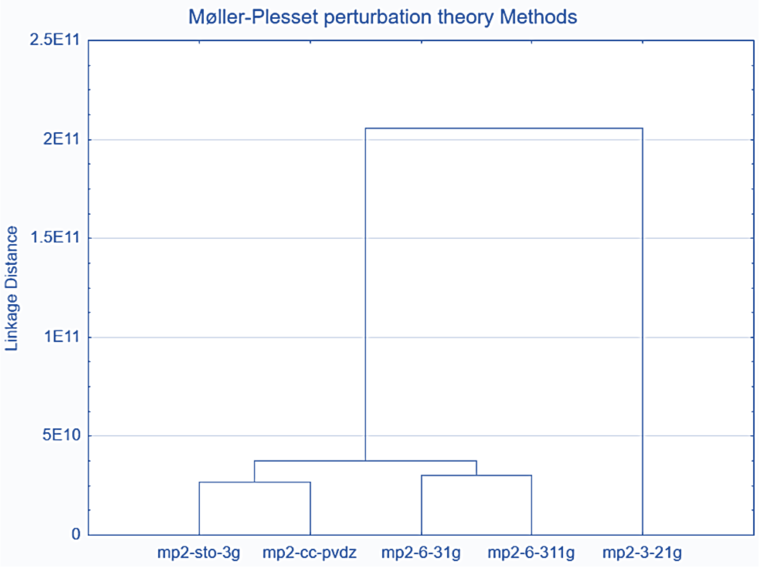


Figure 20. Clustering results

One primary component was found using Møller-Plesset perturbation theory methods (Figure 19). One (mp2-3-21g) and the remaining base sets make up one of the two primary groups into which the methods are separated (Figure 20).

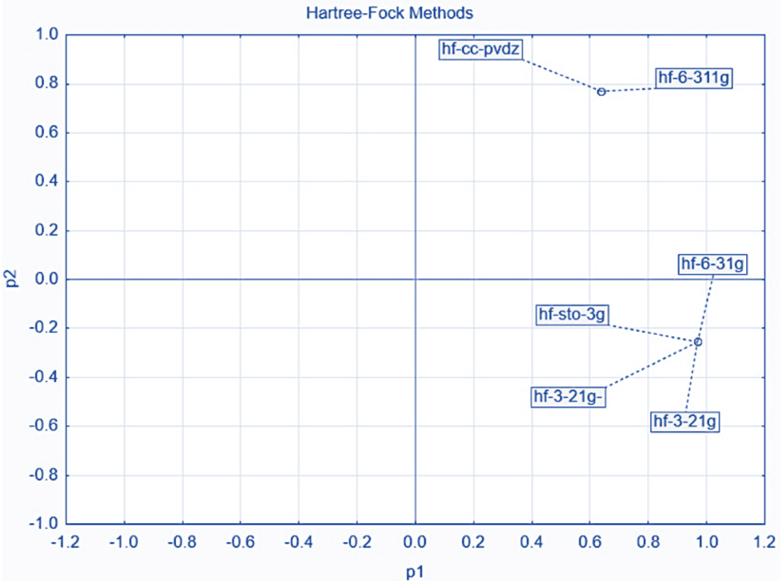


Figure 21. Score plot displaying the distribution of the approaches in the principal components p1 and p2

Hartree-Fock Methods are considered the most widely used among the optimization calculations. Based on our analysis 2 principal components (Figure 21) an also 2 main groups were identified (Figure 22).

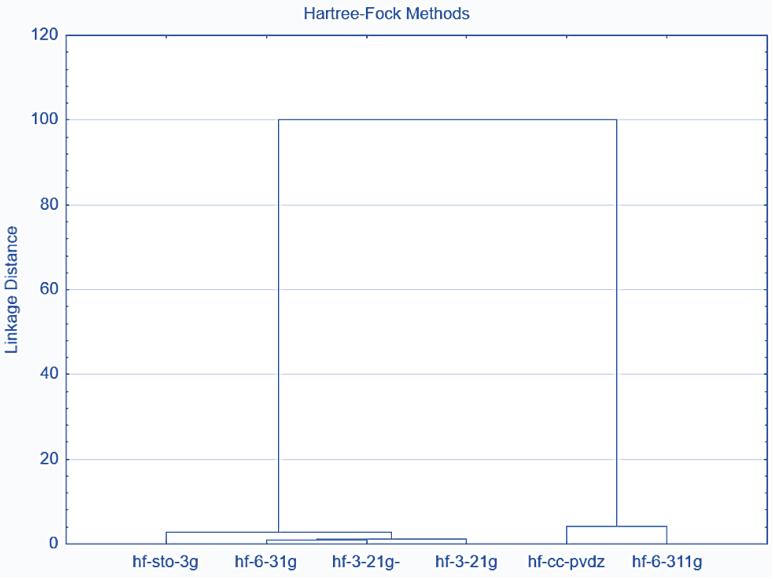


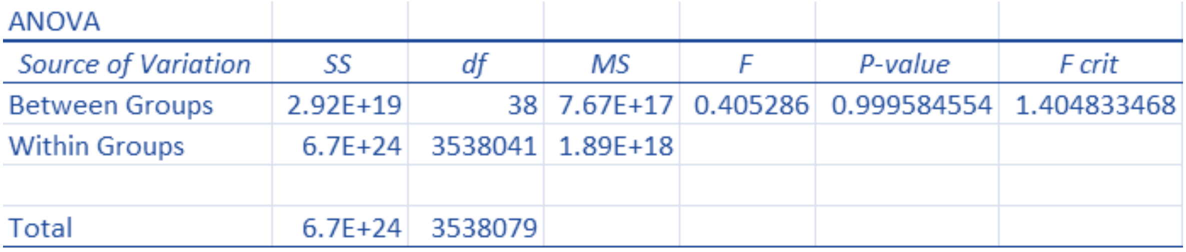
Figure 22. Clustering results

Due to the small dataset, they represented, the other approaches that are a part of the Coupled-cluster theory and Molecular Mechanics could not be individually analysed.

If just one method (CCSD) were examined in Coupled-cluster theory and two methods (UFF, Dreiding) in Molecular Mechanics theory separately, no one revealed a statistically significant result. They are considered in the initial analysis, which examines each and every technique.

It was implemented an alternative statistical examination: The Single Factor ANOVA test (Table 7). The input values for the test were the molecular descriptors obtained for all 39 methods tested (Bálint and Jäntschi, 2021).

Table 7. ANOVA test results



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				

To establish whether there is a variation amid the groups on a certain variable, an ANOVA is conducted. According to the estimated molecular descriptors, the null hypothesis indicates that there is no discernible difference between the procedures analysed.

The F-statistic value (0.405) is the percentage of the variance between groups to the variance within groups, and a larger F-value indicates greater differences between groups. In this case, the variance was indicated within the groups analysed.

If the p-value is 0.9995, it indicates that the possibility of gaining a result as extreme as the observed result, assuming that the null hypothesis is true, is very high. The observed differences between groups are therefore likely to be the result of chance and not statistically significant, to put it another way.

Since the p-value is 0.9995 > 0.05, we accept the null hypothesis and conclude that the differences between the approaches are not appreciably different. According to Table's 9 ANOVA results, the null hypothesis cannot be ruled out.

In correlation analysis, the values of two variables are compared to see if, and if so, how strongly and in which direction, they are related. The correlation coefficient is the statistic used to measure the degree of association between two variables. It ranges from -1 to +1, where a value of -1 indicates a perfect negative correlation (when one variable increases, the other decreases), 0 indicates no correlation, and a value of +1 indicates a perfect positive correlation (when one variable increases, the other also increases).

The correlation coefficients were also calculated between the methods analysed. In the following sections the correlations are presented.

It's essential to keep in mind that correlation does not imply causation in all cases. One variable does not necessarily cause the other just because two variables are connected. The relationship between the two variables may be influenced by additional factors.

The next tables (Table 8-13) present the correlation between the Density Functional Theory methods, Semiempirical Methods, Møller–Plesset perturbation theory, Molecular Mechanics and Hartree-Fock Methods. Correlation between basis sets refers to the degree of agreement between the results obtained from different basis sets.

Due to the extremely large data set, the correlation analysis had to be split into multiple analyses. In each major category previously mentioned, the correlation is done within the sub-methods.

The correlation coefficients calculated for every method indicate a strong relationship between the methods analysed. This is in line with the other statistical results obtained for this data set.

This can provide insights into the underlying mechanisms that govern chemical reactions and can help to design or optimize new materials with desired properties. In addition, correlation analysis can be used to identify important features or variables that contribute to the properties of interest, which can be useful for developing predictive models or designing new experiments.

Table 8. Density functional theory part 1

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

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

The correlation analysis can be applied within the different methods to understand the relationship between different properties of molecules or materials.

Is used to explore the relationship between molecular descriptors, such as electronic energies and atomic charges, and chemical reactivity or other physical or chemical properties.

The bigger the sample size, further reliable and accurate the correlation coefficient will be. A relatively small number of samples may result in a prediction of the actual correlation that is less precise.

This may be seen in comparison between MM methods, where just two methods were compared, and DFT approaches, where a vast dataset was intended to be analysed.

Determining the correlation between Gaussian methods is an important aspect of computational chemistry research. It can help researchers to select the most fitting method for a specific application, as well as to assess the reliability and accuracy of the results obtained from different methods.

However, the correlation between Gaussian methods does not necessarily imply that one method is better than the other, as each technique has its own assets and flaws.

Knowing how one method relates to another is essential for selecting the optimum geometry optimization technique to employ in various circumstances. Due to the fact that they produce results that are almost identical, two procedures that are similar should be excluded from the analysis.

To answer the research questions from the problem statement section, it can be stated the following (Bálint and Jäntschi, 2021):

- The basis set's size does not accurately represent how widely it can be used.
- There are just a few basis sets that fit our dataset; it is impossible to determine the optimal bases set.
- If we use different basis sets from different findings, we will get outcomes that vary, but we must be cautious to choose them correctly.

- It is possible to group and correlate the optimization techniques.

The proper application of various selection and optimization procedures determines the differences in the outcomes.

4. Conclusion

The analysis findings of the study provide insights into the relationships and correlations among different approaches used in the research. We examined a total of 39 methods and reclassified them based on their characteristics and applicability. This reclassification helped in better understanding and selecting the appropriate base sets for different study areas.

By analysing the relationships and correlations among the methods, the researchers gained valuable information about their similarities and differences. This understanding allowed them to identify the strengths and limitations of each method and determine their suitability for specific research areas or applications.

The reclassification of the methods provided a more organized framework for selecting the appropriate base sets. Base sets are important in computational chemistry and quantum mechanics as they form the foundation for calculations and simulations. By matching the characteristics and requirements of the study areas with the specific base sets, researchers can enhance the accuracy and reliability of their results.

Overall, this analysis and reclassification process helped in streamlining the selection of base sets for different study areas, facilitating more efficient and effective research in those fields.

Chapter V – Applications of geometry optimization techniques

Case Study - Factorial analysis of nanostructures

1. Introduction

A dodecahedron-shaped cage made of carbon atoms makes up the four-layered dodecahedrane, a hypothetical carbon allotrope. It has four layers of carbon atoms, each of which has 20 atoms arranged in a regular pentagonal pattern. This is why it is referred to as being "four-layered" (Cao et al., 2020).

Analysis of the relationship between the dodecahedron's atom types and their characteristics can reveal important details about the variety and stability of the resulting structures. This is the reason why they are subjected to numerous studies in this field (Jäntschi et al., 2016).

In this case study, the aim was to examine how various factors and their interactions impact ten calculated properties of a class of dodecahedrane congeners. The molecules were viewed as four-layered structures, as shown in Figure 28.

The study also looked into the effect of forming each layer with either carbon, boron, or nitrogen. To achieve these objectives, a theoretical study was conducted using a full factorial design. The goal was to identify and understand the significant factors that influence the properties of interest (Jäntschi et al., 2016).

2. Methodology

Initially, the cages were created using the software *HyperChem* (in the beginning, PM3 was used to optimize the geometry).

Using *Spartan* software package, the cages geometrical design was improved with the HF method, a 3-21G basis set (Hartree, 1928) and then by MPM (Møller Chr Plesset, 1934) until the next order (MP2) with a 6-31G* basis set (Jäntschi et al., 2016).

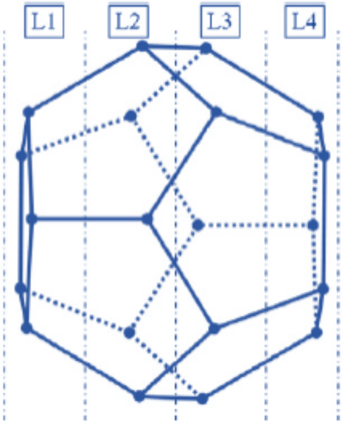


Figure 22. Four layered dodecahedrane

- Geometry optimization: PM3 and Moller-Plesset (MP2) with a 6-31G* basis set.
- Property calculations (volume, surface area, ovality, HOMO, LUMO, polarizability, dipole moment, entropy, enthalpy, energy).
- Full factorial analysis (detect groups of equivalent and irrelevant factors; examine main effects and interaction effects).

Following geometry optimization, MP2 calculations were utilized to determine the features employed for the full factorial analysis.

In the context of exploring the relationship between features taken from compound structures and various qualities, a factor analysis involved testing all possible combinations of the features (independent variables) on the quality measures (response variables).

3. Results and Discussion

A solid explanatory model might have a correlation coefficient (R) close to 0.95, meaning that the factors can account for about 90% of the variance in the relevant attribute. As a result, the models with R > 0.95 underwent additional analysis (Jäntschi et al., 2016).

It was determined which features are most strongly connected with the activity and how they interact by repeatedly altering the levels of each component and monitoring the subsequent activity.

With one exception (energy), the analysis reveals that the number of components varies for various reference atoms. When carbon, boron, and nitrogen were taken into consideration as the reference atoms, the number of components needed to explain the variation in each feature was compared. Volume, surface area, ovality, HOMO, LUMO, polarizability, and entropy were the seven out of ten attributes for which it was discovered that the reference atom carbon required the fewest parameters to be taken into account. However, when carbon was the reference atom, the average number of components needed in the model with a correlation coefficient better than 0.95 was larger (Jäntschi et al., 2016).

The study demonstrated that when characteristics and structural features are intended to be related for corresponding data in a population, data simplicity typically results in simpler models (Kelly, 2011), models with insufficient explanatory power (Kar and Arias-Estrada, 2015). Therefore, in these situations, there is always a trade-off between growing the sample size and reducing model complexity.

The discovery of models with predictive abilities requires the validation of linear models (Gramatica, 2013), but this topic was outside the scope of our investigation and is not addressed in this work.

The work described in this research concentrated on creating a full-factorial analysis for the exploration of the relationship among features taken from compound structures and various qualities (Jäntschi et al., 2016).

Full-factorial strategies can be laborious and resource-intensive, especially when the number of features and levels is large. Yet, they offer several advantages over other experimental designs, such as allowing for the detection of higher-order interactions among the factors and providing a more complete picture of the relationship between the features and the response variables.

4. Conclusion

Despite the fact that 67% of the time, carbon was employed as the reference atom, the models that yielded the highest correlation coefficient were not always the most effective models all around. From the simplest models, which use boron as a reference, to the models that use carbon as a reference, the complexity of the models grows (convoluted models).

Chapter VI – Molecular alignment

Case Study – Biochemical similarity of the selected proteins

1. Introduction

In molecular biology and biological chemistry, determining molecular structure is essential because a molecule's function is highly dependent on its 3D arrangement and geometry of molecules that complement one another. Understanding the nature of the structure and connections of biological networks is essential for arriving at a quantitative description of their functions.

The use of computer tools has become increasingly significant in fields including molecular modelling, docking, and pharmaceutical drug creation. The mathematical computation techniques used to model geometric issues involving molecules as algebraic systems and the algorithms used to solve these systems are frequently examined (Emiris et al., 2005). There are several implementations of well-known and efficient techniques for calculating eigenvalues and eigenvectors available (Murrain and Pan, 2000).

Luo et al. (2006), found that strong interactions between matrix components and eigenvalues cause the eigenvalues to strongly correlate, which results in eigenvalue fluctuations represented by the GOE (Gaussian orthogonal ensemble).

This analysis (Joita et al., 2021) aimed to determine the optimal geometric arrangement of 20 chosen amino acids regarding one another. A development has been made to the earlier study that Jäntschi (Jäntschi, 2019) detailed.

The eigenproblem algorithm aligns the structures, then trilateration is employed to assign all of the previously striped atoms.

2. Methodology

Obtaining the optimal alignment for a molecule implies determining the most favourable alignment of the molecule in space relative to a reference frame or another molecule. The optimal alignment provides a way to compare and analyse the properties and characteristics of molecules based on their spatial arrangement and geometry.

Finding the best alignment is crucial for predicting the characteristics and actions of molecules, such as their reactivity, binding affinity, and stability, in the context of computational chemistry and molecular modelling. In fields like drug development, materials science, and chemical engineering, this knowledge can be used to construct new molecules with the desired features.

The least value of the sum of the squares of the eigenvalues of the Cartesian distance matrix-whose eigenvalues are entirely imaginary since the matrix is antisymmetric-is determined to be a molecule's best alignment.

The working algorithm scheme is presented below (Figure 23).

The amino acids (downloaded from PubChem) are computed. In this case, glycine, which has the least heavy atoms, is used as the reference (Joița et al., 2021).

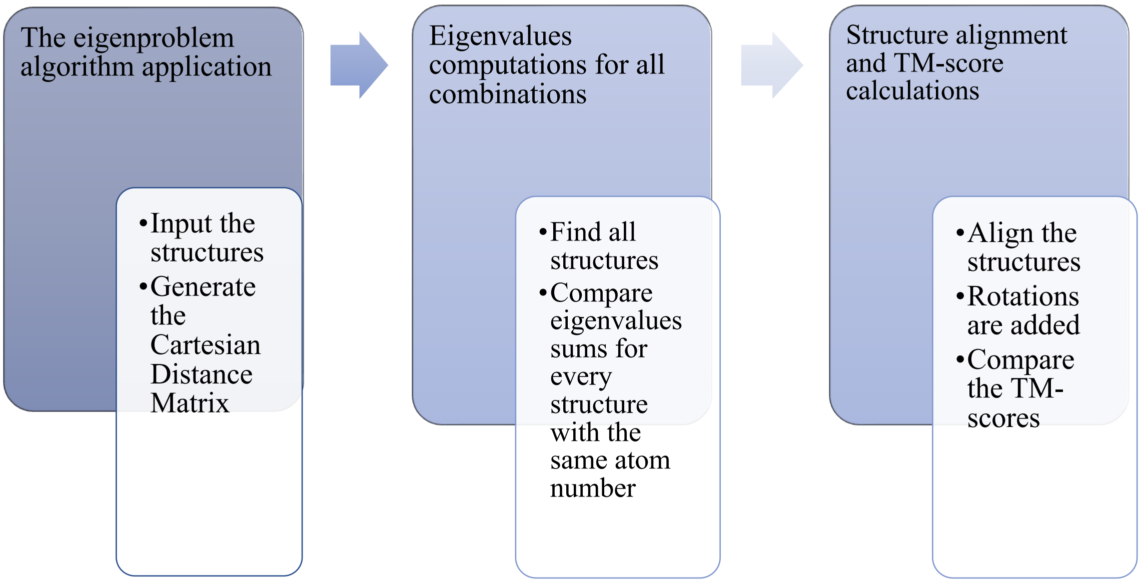

The eigenproblem algorithm application
•Input the structures
•Generate the Cartesian Distance Matrix
Eigenvalues computations for all combinations
•Find all structures
•Compare eigenvalues sums for every structure with the same atom number
Structure alignment and TM-score calculations
•Align the structures
•Rotations are added
•Compare the TM-scores

Figure 23. Working algorithm

Once the conditions have been achieved, the original eigenproblem method is executed to confirm that the program's starting point is a suitable original alignment. Then, all arrangements with fewer than a certain number of atoms are discovered. For each combination, eigenvalues are discovered without rotating the possibilities.

ST sums are also compared up until the input variables are satisfied or all possibilities with at least three atoms are compared. Using the original eigenproblem procedure, structures are aligned, then subjected to trilateration and potentially helpful pi/2 rotations.

The structures coordinates are adjusted such that they align with the coordinates of the reference structure. since the TM-score (Template Modeling score) compares the separations between the atoms in molecules. Final structures who perform well are exported.

3. Results and Discussion

Since one of these rotations may need to produce a good superposition of the two amino acids, the mean values on each axis are calculated for a select few atoms from both structures. Higher scores indicate greater structural similarity. The final score goes from 0 to 1.

The TM score was initially created to aid in the prediction and modelling of protein structures, but it has subsequently been used to address a variety of structural biology issues, such as protein structure comparison, protein-protein docking, and protein-ligand binding investigations. It is especially helpful for comparing structures that, while similar in general, may vary in minute details, such as loop regions or side-chain conformations (Zhang, 2005).

Thirteen findings are single high-confidence alignments of the 19 amino acids that have been aligned to glycine, 11 of which have a high TM-score. The TM-score can be used to determine which of the remaining three (cysteine, lysine, and arginine) offers the greatest results out of two possible good ones (Joita et al., 2021).

For output, a grade of 80% of the maximum score is permitted. This is necessary so that the best alignment, even though it doesn't have the greatest TM-score, is provided as a result.

Viewing the selected structures and eliminating the ones that may have similar mathematical scores, but the incorrect atom types is another simple technique to choose from this group of candidates. Even better outcomes might be achieved by combining other scoring functions or other scoring mechanisms (Joita et al., 2021).

The limiting criteria are relevant for the current comparison because 84% of the best alignments using glycine as a standard can be numerically indicated by a scoring function like the TM-score and 68% of these orientations are exported as single candidates. The scoring function that is currently provided is only useful for 58% of cysteine. A large database would show a reasonable way to select them and aid in machine learning training.

The proper alignment was statistically identified by the TM-score 70% of the cases, on typical, after running the current method using the other amino acids as a reference. By evaluation, 15% more cases with similar scores can be clearly distinguished (Joita et al., 2021).

Full vectorization can speed up the current algorithm. To lessen the effects of partial report abilities and predefined theory-inspired functional shape, machine learning must be introduced to scoring functions.

Instead of enforcing a rigid algorithm, these faults can be rectified by applying machine learning to capture traits that are challenging to predict because there are so many unmeasured/unknown/undiscovered quantitative structure-activity relationships. There is an increasing amount of excellent structural and interaction data in the literature that machine learning can use.

Sequence-based methods, structural alignment techniques, and hybrid techniques that incorporate sequence and structure data are all available for protein alignment in QSAR. ClustalW, MUSCLE, and PyMOL are a few of the software programs that are frequently used for protein alignment in QSAR (Saeys et al., 2007).

4. Conclusion

To establish the optimal geometric alignment of specific amino acids with regard to one another, an application of the eigenproblem was developed.

Therefore, we can say that the ideal alignment is not a straight line. The near results of the same method can be taken into account. Even after a score algorithm has been run, we can infer that the alignment with the highest score is not always the best alignment.

The number of rotations for which a scoring function is executed needs to be reduced with the existing method's parameters. Additionally, integrating various approaches might result in quicker outcomes.

General Conclusions and Future Perspective

Each section's specific and overall conclusions were already provided in the corresponding chapters. Thus, the following paragraphs are devoted to providing a general conclusion to this thesis as well as some ideas for additional study.

The modelling of inorganic and organic chemicals that led in the examination and description of their behaviour at the level of biological systems served as an illustration of their similarities.

The fundamental understanding in the topic of molecular optimization has been developed throughout this thesis. The rationalization of experimental observations with energy calculations has led to an understanding of molecular modelling. On the other hand, reading through previous and current literature on theoretical and experimental studies has given me a broad perspective on this topic of research.

The findings of this thesis demonstrate that computational chemical methods can be used to appropriately define molecular optimization strategies on the molecules under study.

To comprehend the experimental findings and identify the key interactions between molecules, it will be important to evaluate structural models to accurately describe chemical conformation in the near future.

It can be concluded that there is a substantial knowledge breach between the experimental and theoretical domains by working with experimental groups and examining the experimental literature. Particularly, repeated interpretations of experimental findings seem to be lacking theoretical support. To support many of the assertions stated in the literature, a lot of work still has to be done in the fundamental sciences from experimental and theoretical perspectives.

This may influence the choice and selection of methods to be evaluated for optimisation reasons depending on the condition for which method is needed when viewed through the lens of similarity.

The identification of molecular optimization methods that can later assist to better understand the molecular mechanisms is essential for further research.

Finding the best for our needs is both simpler and tougher now that there are so many software programs and algorithms available due to advances in technology. For the objective of defining the concept of similarity, the structure/activity relationship research between the molecules in biological systems is crucial.

Selective Bibliography

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.