GDAŃSK UNIVERSITY OF TECHNOLOGY

TASK Quarterly 29 (4) 2025

Efficiency of an Unsupervised Machine Learning Approach for Assessing Superatom-like Fullerenes

Sümeyye Atmaca [1], **Celina Sikorska** [1,2*]

[1]Faculty of Chemistry, University of Gdańsk, Fahrenheit Union of Universities in Gdańsk, Wita Stwosza 63, Gdańsk, 80-308, Poland
[2]The MacDiarmid Institute for Advanced Materials and Nanotechnology, Department of Physics, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

*celina.sikorska@ug.edu.pl

https://doi.org/10.34808/tq2025/29.4/a

Abstract

Superatoms are of broad interest in materials science due to their high tunability of electronic properties upon structural modification. The quantum-mechanical (QM) descriptors estimated in our previous work (Sikorska C, Puzyn T, Nanotechnology (2015), 26: 455702) have been used to explore structure-property relationships in superatom-like fullerenes. The structural similarity among twenty-six methanofullerenes has been investigated using principal component analysis (PCA) and two-way hierarchical cluster analysis (t-HCA) based on these QM descriptors, demonstrating how electronic structure parameters and geometrical features influence fullerene cluster properties. This unsupervised machine learning approach highlights the fact that descriptors derived from quantum-mechanical calculations enable groups of structurally similar compounds for which we can assume similar values of selected physicochemical properties to be distinguished. In addition, the use of computational methods not only reduces the time and costs of research but also the amount of waste generated during experimental analyses. Hence, the research described has significant social and economic significance. At the same time, our results provide a framework for understanding structure-property relationships in nanomaterials that can be used in the future to define new quantitative structure-property relationship (QSPR) models for predicting physicochemical properties of fullerene-based materials directly from the fullerene structure.

Keywords:

superatoms, fullerene derivatives, molecular descriptors, principal component analysis (PCA), hierarchical cluster analysis (HCA), electronic structure

1. Introduction

A superatom is a cluster of atoms that collectively mimics the electronic structure and chemical behavior of a single atom [1–3]. Its features can be precisely controlled by adding, substituting, or removing a single atom within it [1]. Fullerenes, such as C_{60}, can exhibit a superatom-like character due to their highly symmetric icosahedral structure and closed-shell π-electron configurations (analogous to those of noble gas atoms) [1, 4]. Fullerene molecules have captured the attention of chemists, physicists, and materials scientists worldwide due to their unique combination of structural stability and electronic versatility. However, beyond their well-established applications as electron acceptors and drug carriers, fullerenes can function as superatoms with distinct superatom molecular orbitals (SAMOs), placing them at the forefront of contemporary materials science [5]. The recognition of fullerenes as superatomic systems has opened new paradigms for understanding molecular electronics and designing advanced functional materials [6, 7].

The concept of superatom molecular orbitals in fullerenes is rooted in the distinctive electronic architecture of these organic molecules. Unlike conventional organic compounds, in which frontier orbitals are localized on specific atoms or bonds, fullerenes have delocalized electronic states arising from their symmetrical hollow cage structures [5, 8]. These delocalized states exhibit properties reminiscent of nearly free-electron (NFE) bands, which are typically associated with metallic systems [9]. The SAMOs embedded within fullerene cages interact in ways analogous to atomic orbitals, but with substantially altered characteristics dictated by the carbon framework. This electronic behavior enables fullerenes to transcend traditional molecular chemistry boundaries and exhibit collective properties that bridge the gap between molecular and condensed-matter physics [10, 11].

The main merit of superatomic systems is the tunability of their properties upon structural modification. Superatom features can be precisely controlled by adding, substituting, or removing a single atom within it [1, 2, 12]. Examples include magnesium-based oxyfluoride superatoms, whose electron affinity can be precisely controlled by single-atom substitution [2, 13]. The high tunability of the electronic features in magnesium-based oxyfluoride superatoms arises because modifying their composition (e.g., changing the number and type of ligands like O, F, and OF, or the number of magnesium central atoms) strongly affects key electronic parameters such as electron binding energies, HOMO/LUMO characteristics, and overall stability. Structural modification of magnesium-based oxyfluoride enables systematic control over its properties, such as electron affinity and the HOMO-LUMO gap. Superatomic structure modification allows the superatomic electronic structure to be tailored through targeted ligand and structural changes to achieve desired functionalities [1, 2, 13].

Tuning fullerene electronic properties through strategic chemical modification is a cornerstone of modern materials science. The symmetrical cage structures of pristine C_{60} and C_{70} provide abundant double C=C bonds, enabling versatile chemical modification with a wide range of aliphatic and aromatic groups [14]. Among various fullerenes, the electronic properties of different cage sizes differ markedly. For instance, larger fullerenes, such as C_{82}, have larger superatom molecular orbital energy values than C_{60}, a variation directly attributable to molecular symmetry and cage geometry [10]. Furthermore, endohedral doping, the incorporation of metal atoms or clusters within the fullerene cavity, profoundly influences SAMO energies and wavefunction distributions [15]. Structural modifications enable precise control over the electronic framework, creating a pool of derivatives with tailored properties suitable for chemical applications.

The importance of understanding and exploiting fullerene superatom characteristics extends across multiple frontiers of materials science. Examples include the binary assembly of C_{60} fullerene and metal chalcogenide clusters [i.e., $Co_6Se_8(P(C_2H_5)_3)_6$, $Cr_6Te_8(P(C_2H_5)_3)_6$, and $Ni_9Te_6(P(C_2H_5)_3)_8$] as building blocks [7]. The $[Co_6Se_8(P(C_2H_5)_3)_6][C_{60}]_2$ and $[Cr_6Te_8(P(C_2H_5)_3)_6][C_{60}]_2$ solids are gapped semiconductors, whereas $[Ni_9Te_6(P(C_2H_5)_3)_8][C_{60}]$ material has ferromagnetic characteristics. In these fullerene-based solids, the C_{60} serves as an electron acceptor, and their physicochemical properties can be fine-tuned upon a properly chosen metal cluster (acting as an electron donor) with which C_{60} interacts [11, 16]. Due to their large electron affinity and efficient charge transfer, fullerenes can also serve as acceptor compartments in photovoltaic devices [17–20]. In the fullerene-based solar cell, the open-circuit voltage (V_{oc}, the voltage that one would measure across an open-circuited solar cell with sunlight shining on it) can be enhanced by introducing fullerenes of lowest unoccupied molecular orbital (LUMO) energy above −3 eV [14]. Thus, exploring the electronic properties of superatom-like fullerenes is crucial for understanding donor-acceptor processes that ultimately affect the photovoltaic performance.

The high stability and electronic properties of fullerene-based materials arise from substantial orbital and electrostatic interactions, making them promising building blocks for nanostructured materials [11]. The ability to manipulate the energy levels of superatom molecular orbitals through rational structural design

offers unprecedented opportunities in molecular electronics, enabling the development of highly efficient solar cells. The potential chemical applications underscore the profound influence that superatom concepts exert on contemporary materials design [21]. However, high research costs and time-consuming experimental procedures limit the likelihood of obtaining electronic properties of fullerene clusters. Therefore, computational methods (also known as an in silico approach) are a promising alternative to experimental studies [1, 22–27].

Although the efficiency of computational methods for carbon-based superatoms has been reported [11, 14], there is still a need to investigate quantum-mechanical molecular descriptors, especially if these descriptors are to be used for quantitative structure-property relationship (QSPR) modeling. In our previous work, we systematically investigated a family of fullerene derivatives using an integrated first-principles approach combined with principal component analysis [14]. This study demonstrated that QM molecular descriptors enable the prediction of the open-circuit voltage of organic photovoltaic cells with fullerene as the electron acceptor. Thus, the descriptors derived from quantum-mechanical calculations (particularly frontier orbital energies) can significantly reduce the burden on experimental studies by predicting the efficiency of fullerene-based solar cells before their synthesis.

Building on our work [14], the present study aims to investigate the applicability of quantum-mechanical molecular descriptors for elucidating structure–property relationships in superatom research. Using principal component analysis (PCA) and hierarchical cluster analysis (HCA), the structural similarity of selected fullerenes was investigated in the multidimensional feature space defined by QM descriptors. The purpose of using the PCA and HCA analysis was to indicate groups of structurally similar compounds for which it will be possible to assume similar values of physicochemical properties. The knowledge acquired in this way can be used in the future to define the field of new quantitative structure-property relationship models for predicting the physicochemical properties of fullerene-based materials directly from their structures.

2. Methodology

Quantum-mechanical (QM) descriptors were obtained from recently reported calculations using (i) density functional theory (DFT) with the Becke's parameter hybrid method and the LYP (Lee-Yang-Parr) correlation functional (B3LYP) and (ii) semi-empirical methods (PM6 and PM7) [14]. The use of semi-empirical approaches substantially reduces the computational cost and processor requirements. The PM7-based molecular descriptors were selected due to their favorable balance between accuracy and accessibility, as they can be readily computed using the open-source MOPAC package [28, 29]. The PM7 molecular descriptors considered in this study include: (i) energy of the highest occupied molecular orbital, E_{HOMO}; (ii) energy of the lowest unoccupied molecular orbital, E_{LUMO}; (iii) standard enthalpy of formation, HOF; (iv) dipole moment, μ ; (v) molecular volume, CV; (vi) solvent accessible surface area, CA; (vii) absolute hardness ($\eta = (E_{LUMO} - E_{HOMO})/2$); (viii) absolute electronegativity ($\chi = -(E_{LUMO} + E_{HOMO})/2$); and (ix) reactivity index ($S = \chi^2/2\eta$); and (x) E_{HOMO}/E_{LUMO} ratio.

Table 1: Structure description of the studied C_{60} and C_{70} fullerene superatoms.

FD	C_n	Aryl group	Alkyl group	R group
FD1	C_{60}	phenyl	$-C_3H_6COOR$	Me
FD2	C_{60}	phenyl	$-C_3H_6COOR$	Et
FD3	C_{60}	2-thienyl	$-C_3H_6COOR$	Me
FD4	C_{60}	2-furyl	$-C_3H_6COOR$	Me
FD5	C_{70}	phenyl	$-C_3H_6COOR$	Me
FD6	C_{60}	phenyl	$-C_2H_5COOR$	Me
FD7	C_{60}	phenyl	$-C_2H_5COOR$	Et
FD8	C_{60}	phenyl	$-C_2H_5COOR$	Pr-n
FD9	C_{60}	phenyl	$-C_2H_5COOR$	Pr-i
FD10	C_{60}	phenyl	$-C_2H_5COOR$	Bu-n
FD11	C_{60}	phenyl	$-C_2H_5COOR$	Bn
FD12	C_{60}	4-methoxyphenyl	$-C_2H_5COOR$	Me
FD13	C_{70}	phenyl	$-C_2H_5COOR$	Me
FD14	C_{70}	phenyl	$-C_2H_5COOR$	Et
FD15	C_{70}	phenyl	$-C_2H_5COOR$	Pr-n
FD16	C_{70}	phenyl	$-C_2H_5COOR$	Bu-n
FD17	C_{60}	2-thienyl	$-C_2H_5COOR$	Et
FD18	C_{60}	2-thienyl	$-C_2H_5COOR$	Pr-n
FD19	C_{60}	2-thienyl	$-C_2H_5COOR$	Bu-n
FD20	C_{70}	2-thienyl	$-C_2H_5COOR$	Pr-n
FD21	C_{70}	2-thienyl	$-C_2H_5COOR$	Bu-n
FD22	C_{60}	phenyl	$-C_6H_{13}-n$	H
FD23	C_{60}	phenyl	$-C_4H_9-n$	H
FD24	C_{60}	2-thienyl	$-C_6H_{13}-n$	H
FD25*	C_{60}	NAN	$-C_2H_5COOR$	Me
			$-C_2H_5COOR$	Me
FD26*	C_{60}	NAN	$-COOR$	Et-O-Me
			$-COOR$	Et

* Fullerene derivatives without aromatic substituent

Principal component analysis was applied to achieve dimensionality reduction by transforming the original variables into orthogonal factors, known as principal components (PCs) [30, 31]. The optimal number of principal components was selected based on the analysis of eigenvalues and the cumulative percentage of variance described by subsequent PCs. The eigenvalue analysis was carried out based on the minimum eigenvalue exceeding one (the so-called Kaiser-Guttman criterion) [32]. The optimal number of principal components was selected to explain at least 90% of the data matrix's vari-

Table 2: The constitutional and QM descriptors of fullerene derivatives (FD). The standard enthalpy of formation (HOF, in kcal/mol), dipole moment (μ, in D), energy of the highest occupied molecular orbital (E_{HOMO}, in eV), energy of the lowest unoccupied molecular orbital (E_{LUMO}, in eV), molecular volume (CV, in Å^3); solvent accessible surface area (CA, in Å^2), molecular weight (MW), total number of atoms (nAT), number of carbon atoms (nC), number of hydrogen atoms (nH), number of heteroatoms ($nHet$), absolute hardness ($\eta = (E_{LUMO} - E_{HOMO})/2$, in eV), absolute electronegativity ($\chi = -(E_{LUMO} + E_{HOMO})/2$, in eV), reactivity index ($S = \chi^2/2\eta$), ratio, difference between total molecular weight and molecular weight of isolated unsubstituted carbon cage ($MW - MWcage$), and difference between total molecular volume and molecular volume of isolated unsubstituted carbon cage ($CV - CV_{cage}$).

FD	HoF	μ	E_{HOMO}	E_{LUMO}	MW	CA	CV	nAT	nC	nH	nHet	η	χ	S	MW–MW$_{cage}$	CV–CV$_{cage}$
FD1	740	3.46	-9.27	-2.79	911	573	858	88	72	14	2	6.48	6.03	5.61	190	233
FD2	732	3.94	-9.26	-2.78	925	594	876	91	73	16	2	6.48	6.02	5.59	204	250
FD3	747	3.27	-9.22	-2.80	917	568	853	85	70	12	3	6.41	6.01	5.63	196	228
FD4	715	3.21	-9.31	-2.82	901	562	837	85	70	12	3	6.49	6.07	5.68	180	212
FD5	809	3.66	-8.86	-3.05	1031	629	957	98	82	14	2	5.81	5.95	6.10	190	234
FD6	746	3.42	-9.30	-2.81	897	554	834	85	71	12	2	6.49	6.05	5.65	176	209
FD7	738	3.75	-9.29	-2.80	911	574	854	88	72	14	2	6.49	6.04	5.63	190	229
FD8	730	1.29	-9.30	-2.82	925	567	882	91	73	16	2	6.48	6.06	5.67	204	257
FD9	729	3.90	-9.28	-2.79	925	591	879	91	73	16	2	6.49	6.04	5.62	204	253
FD10	728	3.85	-9.28	-2.80	939	615	899	94	74	18	2	6.49	6.04	5.62	218	274
FD11	764	3.50	-9.25	-2.76	973	573	932	95	77	16	2	6.49	6.00	5.56	252	307
FD12	705	3.98	-9.24	-2.79	927	584	869	89	72	14	3	6.45	6.01	5.61	206	244
FD13	814	3.79	-8.88	-3.06	1017	588	935	95	81	12	2	5.81	5.97	6.13	176	213
FD14	805	4.16	-8.87	-3.06	1031	602	958	98	82	14	2	5.81	5.96	6.12	190	236
FD15	800	3.99	-8.86	-3.05	1045	619	982	101	83	16	2	5.81	5.95	6.10	204	259
FD16	795	4.12	-8.86	-3.04	1059	636	1006	104	84	18	2	5.81	5.95	6.09	218	283
FD17	744	3.83	-9.25	-2.81	917	567	849	85	70	12	3	6.45	6.03	5.64	196	223
FD18	737	1.31	-9.27	-2.83	931	556	868	88	71	14	3	6.43	6.05	5.69	210	243
FD19	734	3.93	-9.25	-2.81	945	609	893	91	72	16	3	6.44	6.03	5.64	224	268
FD20	807	3.88	-8.86	-3.05	1051	613	973	98	81	14	3	5.81	5.96	6.11	210	250
FD21	802	4.02	-8.86	-3.05	1065	635	995	101	82	16	3	5.81	5.95	6.10	224	272
FD22	808	3.47	-9.23	-2.76	895	580	865	91	73	18	0	6.48	5.99	5.55	174	240
FD23	818	3.47	-9.23	-2.76	867	539	819	85	71	14	0	6.47	5.99	5.55	146	194
FD24	814	3.38	-9.19	-2.77	901	574	859	88	71	16	1	6.42	5.98	5.57	180	234
FD25	627	2.41	-9.37	-2.87	907	571	841	87	69	14	4	6.50	6.12	5.75	186	216
FD26	611	1.69	-9.49	-2.97	909	534	834	85	68	12	5	6.52	6.23	5.95	188	209

ability. Additionally, to assign the principal components a more straightforward physical interpretation, the PCs were rotated using the VARIMAX method. The VARIMAX algorithm involves rotating the coordinate system so that the sum of the variances of the components of the directional vectors is maximized, and the new axes are as close as possible to the clusters of directional vectors representing the explanatory variables [30].

Hierarchical cluster analysis (HCA) was performed to assess similarities among compounds based on the molecular descriptors. Prior to HCA analysis, the data were standardized using z-score normalization to eliminate scale differences and ensure comparability among variables [30]. Pairwise distances between samples were calculated using the Euclidean distance metric. Agglomerative clustering was conducted using Ward's linkage, which minimizes the increase in total within-cluster variance at each step, thereby producing compact, internally homogeneous clusters. The results were visualized as a dendrogram, and the optimal number of clusters was determined by inspecting the dendrogram structure and selecting an appropriate cut-off level that maximized between-cluster separation while preserving within-cluster homogeneity.

Chemometric analyses were conducted using MATLAB R2024a and the PLS Toolbox [33].

3. Results and Discussion

This contribution aimed to verify the usefulness of constitutional and quantum-mechanical descriptors for describing superatoms. Fullerene derivatives were selected as representative superatom-like compounds.

3.1. Quantum-mechanical descriptors

The first stage of our research studies involved collecting molecular descriptors for various fullerene derivative (**FD**) superatoms. Examples include [6,6]-phenyl-C_{61}-butyric acid methyl ester (PCBM, **FD1**), [6,6]-thienyl-C_{61}-buteric acid methyl ester (TCBM, **FD3**), and [6,6]-phenyl-C_{61}-propionic butyl ester (PCPB, **FD10**). In total, twenty-six fullerenes were examined, with their structures summarized in Table 1. These compounds are cyclopropanated derivatives of C_{60} and C_{70} fullerenes, commonly referred to as methanofullerenes. The cyclopropane addend bears two distinct substituents that

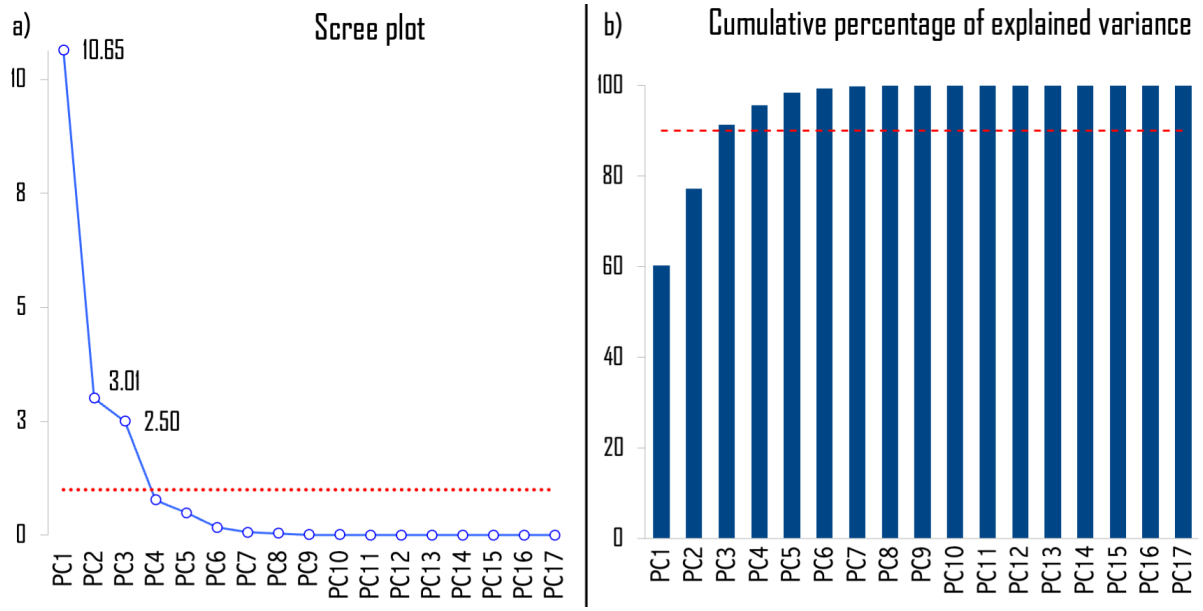

Figure 1: a) Scree (eigenvalue) plot, where the horizontal asymptote corresponds to an eigenvalue of one (red dotted line), (b) cumulative percentage of variance explained by the 17 principal components (PC1-PC17) with dashed red line representing 90% of data matrix variability.

can be independently modified: (i) an aromatic moiety and (ii) an aliphatic chain, enabling systematic tuning of electronic properties [14, 17]. Constitutional descriptors [describing the number of atoms (*nAT*, *nC*, *nH*, and *nHet*) and molecular weight (*MW*)], together with twelve quantum-mechanical (QM) descriptors, are shown in Table 2.

3.2. Principal component analysis

For modeling purposes, a matrix of 17 autoscaled structural descriptors (constitutional and quantum-mechanical) was used. A principal component analysis was performed to transform the original variables into orthogonal factors, known as principal components (PCs). To determine the optimal number of principal components, the PC's eigenvalues and the cumulative percentage of variance explained by subsequent principal components were analyzed. The result of eigenvalue estimation is presented graphically in Figure 1. The number of significant principal components was determined based on the following two criteria: (i) the minimum eigenvalue greater than one (the so-called Kaiser-Guttman criterion, indicated by the red dotted line in Figure 1a) and (ii) the cumulative percentage of variance exceeding 90 % (indicated by the red dashed line in Figure 1b) [32].

Based on the PC's eigenvalues and cumulative variance explained, three principal components were selected, which together explain 91.4% of the variability in the data matrix. The first factor (PC1) differentiates fullerene derivatives mainly based on their size (*MW*, *CV*, *nAT*, *nC*, Figure 2a) and reactivity (η, *S*, E_{LUMO}, E_{HOMO}/E_{LUMO}). The second factor (PC2) describes the carbon chain of the substituent (nH, see Figure 2b). The third factor (PC3) differentiates fullerenes based on the molecular weight of the substituent (MW–MWcage, Figure 2c) and molecular volume of the substituent ($CV - CV_{cage}$, Figure 2c). Features of the whole compound determine the first component, while PC2 and PC3 rely solely on substituent properties. The above observations

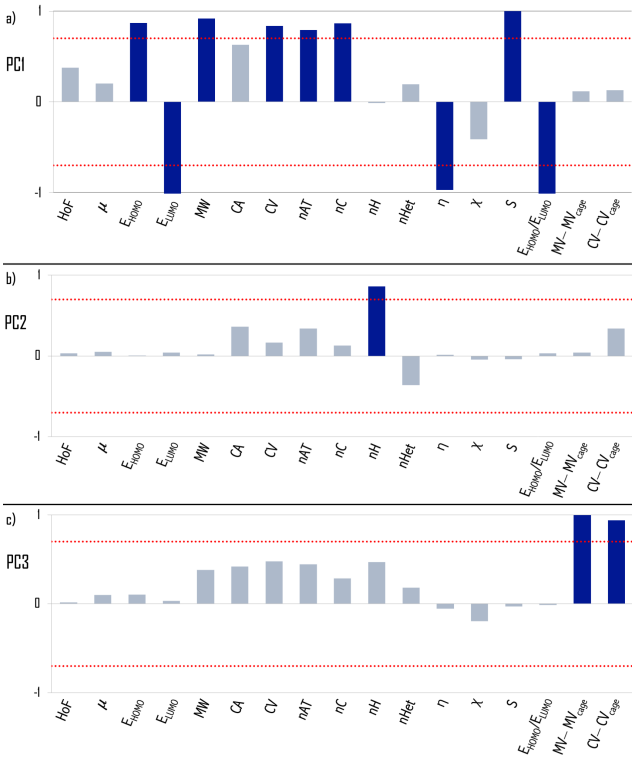

Figure 2: PCA loading factors estimated for (a) PC1, (b) PC2, and (c) PC3. The dotted red lines correspond to an absolute loading factor value of 0.7.

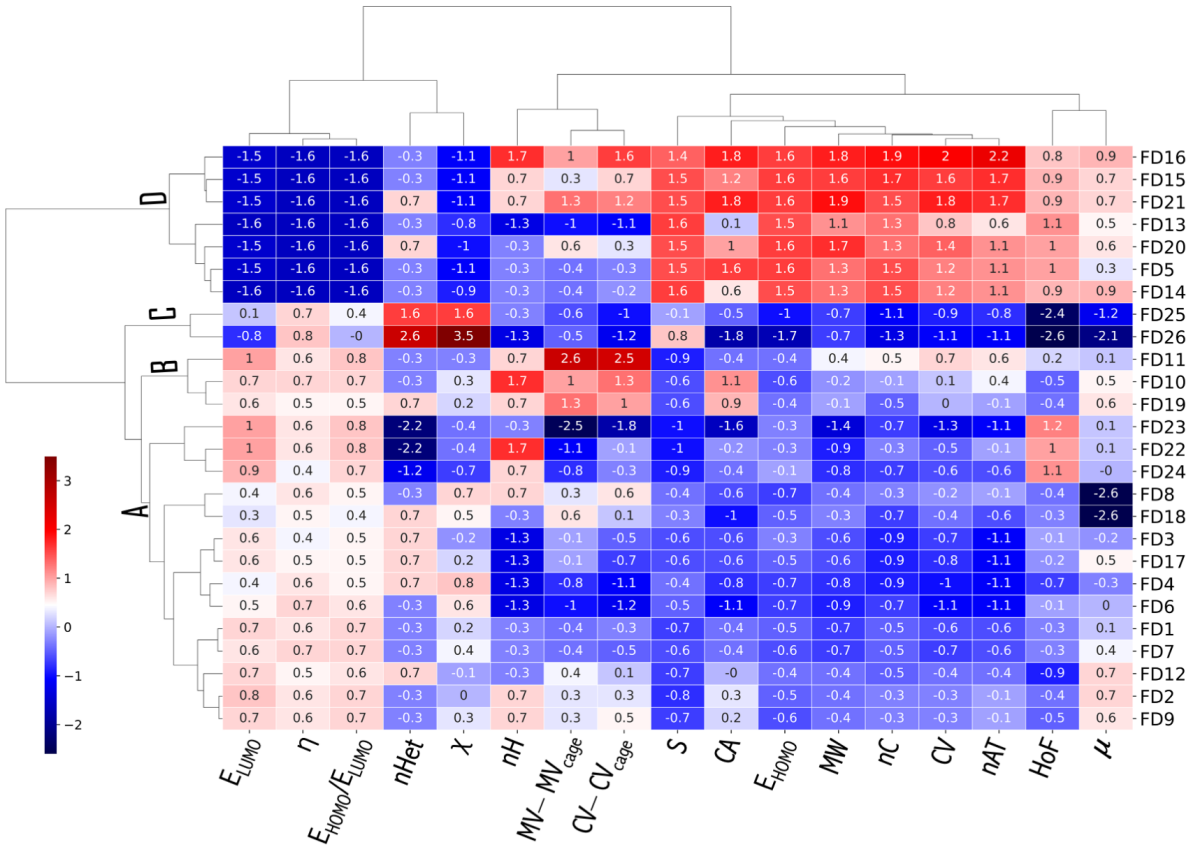

Figure 3: Dendrogram illustrating the result of clustering 26 fullerene derivatives (FDs). A two-way hierarchical cluster analysis was performed using the Euclidean distance metric and the Ward linkage method.

emphasize the strong influence of structure modifications on collective molecular features.

The PC1 is dominated by strong positive loadings for molecular weight, molecular volume, number of atoms, number of carbon atoms, and strong negative loadings for E_{LUMO}, absolute hardness, and E_{HOMO}/E_{LUMO} ratio. Consequently, the PC1 primarily separates molecules according to molecular size (MW, nAT, nC, CV) and is opposed to electron-acceptor character/reactivity descriptors (E_{LUMO}, η, E_{HOMO}/E_{LUMO}). Thus, PC1 can be interpreted as a molecular size and electron-acceptor strength axis, contrasting large, electronically rich molecules (the positive side of PC1) with smaller, harder, more electron-accepting systems (the negative side of PC1).

The PC2 is strongly dominated by a high positive loading of the number of hydrogen atoms (nH, Figure 2b). Therefore, the PC2 can be interpreted as a hydrogen-content axis, distinguishing more saturated/hydrocarbon-like molecules from those with lower hydrogen content. Finally, the PC3 component primarily reflects variability in the size of substituents attached to the fullerene cage (i.e., C_{60} or C_{70}), as captured by substituent molecular weight and volume (see Figure 2c). Molecules with larger, bulkier substituents have higher positive scores on PC3, while molecules with smaller substituents score lower. Thus, the PC3 can be interpreted as a substituent size/volume axis, independent of the global electronic (PC1) and hydrogen content (PC2) effects.

3.3. Hierarchical cluster analysis

The next step in assessing the utility of constitutional and quantum-mechanical descriptors for evaluating the physicochemical properties of fullerenes was to perform a C_{70} hierarchical cluster analysis (t-HCA) for all twenty-six FD superatoms. The t-HCA analysis was performed using a data matrix in which the columns correspond to the calculated QM and constitutional descriptors. The resulting dendrogram is presented in Figure 3. Based on the HCA analysis, the compounds were grouped according to the correlation of FD pairs (compounds strongly correlated with each other formed a specific cluster). As shown in Figure 3, cluster A consists of low-molecular-weight FDs (C_{60} derivatives such as **FD1**; Figure 4a), cluster B comprises C_{60} derivatives with a large ester substituent (where the organyl group, R, is n-butyl or benzyl (**FD11**; Figure 4b), and cluster C is formed by fullerenes lacking aromatic substituents (e.g., **FD25**; Figure 4c). In contrast, cluster D consists of C_{70} fullerene derivatives (e.g., **FD5**; Figure 4d).

The HCA clustering result is consistent with fullerene grouping observed in the score plot of the most

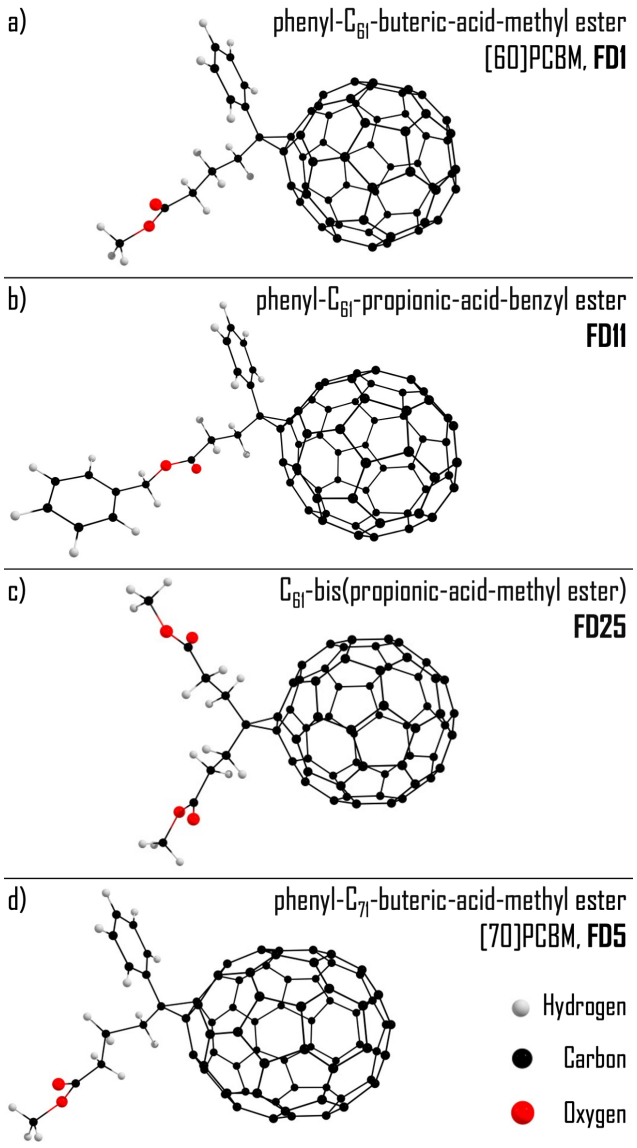

Figure 4: Representative fullerenes corresponding to (a) cluster A, (b) cluster B, (c) cluster C, and (d) cluster D identified by the hierarchical cluster analysis (HCA; Figure 3).

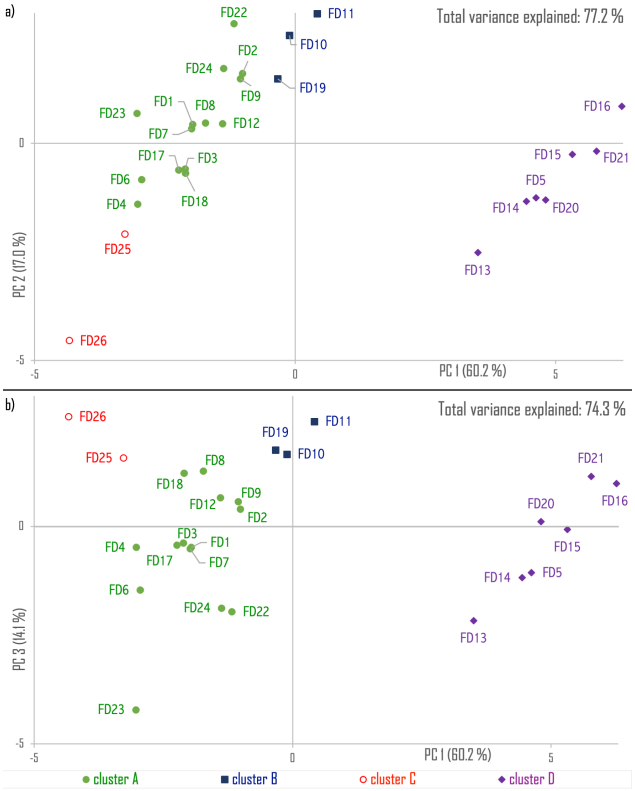

Figure 5: PCA linear maps of (a) PC1 and PC2 and (b) PC1 and PC3 components.

significant principal components. Specifically, in Figure 5 one can clearly distinguish C_{60} and C_{70} fullerenes, as they differ significantly in PC1 values. This observation can be explained by the fact that PC1 encodes molecular size (as reflected in the *MW*, *CV*, *nAT*, and *nC* descriptors). Consequently, the PC1 values are the lowest for non-aromatic substituted C_{60} fullerenes (cluster C) and increase via aromatic substituted C_{61}-acid esters (cluster A) and aryl (hetero-aryl)-C_{61}-propionic acid esters with a large ester substituent R (cluster B) to aryl (hetero-aryl)-C_{71}-acid esters (cluster D, indicated in purple in Figure 5).

The three-dimensional PCA score plot (see Figure S1 of the supplementary material) shows that the first three principal components explain 91.4% of the total variance, with PC1 accounting for 60.2%, PC2 for 17.0%, and PC3 for 14.1%. The distribution of compounds reveals clear clustering, indicating a strong underlying group structure in the dataset. Separation is primarily driven by PC1, which captures most of the variability (60.2%) and distinctly discriminates the classes along its axis. PC2 and PC3 components provide additional discriminatory power, further refining the separation and reducing overlap between groups. The minimal inter-cluster overlap observed in the score space suggests good intrinsic separability of the compounds. It indicates that the studied molecular descriptors contain substantial discriminatory information suitable for subsequent classification analyses.

4. Summary

The usefulness of constitutional and quantum-mechanical descriptors for assessing the physicochemical properties of fullerenes has been demonstrated using unsupervised machine learning methods. The performance of principal component analysis (PCA) and two-way hierarchical cluster analysis (t-HCA) enabled structurally similar compounds for which we can assume similar values of physicochemical properties to be distinguished. The PCA revealed that the first three principal components explain 91.4% of the total variance, with PC1 (60.2%) primarily representing a molecular size and electron-acceptor strength axis, contrasting large, electronically rich systems with smaller, harder, and more electron-

accepting molecules. The PC2 (17.0%) was dominated by hydrogen content, distinguishing more saturated derivatives from less hydrogenated structures, while PC3 (14.1%) reflected substituent size and volume effects independent of global electronic and hydrogen-related contributions. The HCA analysis produced four distinct clusters corresponding to meaningful structural groupings, including low-molecular-weight C_{60} fullerenes, C_{60} fullerenes with bulky ester substituents, non-aromatic substituted C_{60} fullerenes, and C_{70} fullerenes. The HCA clustering pattern was consistent with the PCA score plots, where separation was primarily driven by molecular size (PC1), with additional refinement provided by hydrogen content (PC2) and substituent size (PC3). The clear separation and minimal overlap between clusters confirm that the selected descriptors effectively capture the key structural and electronic differences among the studied superatom-like fullerenes, demonstrating their suitability for classification and further predictive modeling. The main advantage of the developed method of grouping and cross-sectional estimation is that it does not require a large amount of data to define groups of compounds with similar properties. Moreover, the project findings can be used in the future to define the field of new QSPR (quantitative structure-property relationship) models for predicting physicochemical features (such as the efficiency of organic photovoltaic cells) directly from carbon cluster structure.

Data Availability

The data supporting this article have been deposited in the repository (DOI: 10.18150/ASF6SV). Supplementary information: (i) constitutional and quantum-mechanical descriptors of twenty-six fullerenes, (ii) three-dimensional (3D) score plot of three principal components (PC1-PC3, which in total explain 91.4 % of the data set variability), see https://doi.org/10.18150/ASF6SV.

Authors contributions

Sümeyye Atmaca: investigation and writing – review & editing.

Celina Sikorska: conceptualization, supervision, writing – original draft, writing – review & editing, and funding acquisition.

Conflict of Interest

There are no conflicts to declare.

Data Availability

The data supporting this article have been deposited in the repository (DOI: 10.18150/ASF6SV). Supplementary information: (i) constitutional and quantum-mechanical descriptors of twenty-six fullerenes, (ii) three-dimensional (3D) score plot of three principal components (PC1-PC3, which in total explain 91.4% of the data set variability), see https://doi.org/10.18150/ASF6SV.

Acknowledgments

This research is part of the project No. 2022/45/P/ST4/01907 co-funded by the National Science Centre and the European Union Framework Programme for Research and Innovation Horizon 2020 under the Marie Skłodowska-Curie grant agreement No. 945339. CS would like to express the deepest gratitude to Prof. Adam Liwo for their outstanding mentorship and support throughout this Polonez Bis project. For the purpose of open access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission. Calculations were carried out in (a) the Wroclaw Centre for Networking and Supercomputing (http://www.wcss.pl, grant No. 378), (b) the Centre of Informatics-Tricity Academic Supercomputer and Network (CI TASK) in Gdansk (project No. pt01088), and (c) the New Zealand eScience Infrastructure (NeSI) high-performance computing facilities funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure program (https://www.nesi.org.nz, project no. uoa02699).

References

[1] C. Sikorska, "Design and investigation of superatoms for redox applications: First-principles studies," *Micromachines*, vol. 15, no. 1, p. 78, 2024.

[2] C. Sikorska, "Magnesium-based oxyfluoride superatoms: Design, structure, and electronic properties," *Journal of Chemical Information and Modeling*, vol. 59, no. 5, pp. 2175–2189, 2019.

[3] H. Hakkinen, "Atomic and electronic structure of gold clusters: understanding flakes, cages and superatoms from simple concepts," *Chemical Society Reviews*, vol. 37, no. 9, pp. 1847–1859, 2008.

[4] W. Xie, F. Yu, X. Wu, Z. Liu, Q. Yan, and Z. Wang, "Constructing the bonding interactions between endohedral metallofullerene superatoms by embedded atomic regulation," *Physical Chemistry Chemical Physics*, vol. 23, no. 30, pp. 15899–15903, 2021.

[5] Y. Pavlyukh and J. Berakdar, "Superatom molecular orbitals: new types of long-lived electronic states," *Journal of Chemical Physics*, vol. 135, no. 20, p. 201103, 2011.

[6] C. Sikorska, "Mg3f7: a superhalogen with potential for new nanomaterials design," *International Journal of Quantum Chemistry*, vol. 118, no. 21, 2018.

[7] X. Roy *et al.*, "Nanoscale atoms in solid-state chemistry," *Science*, vol. 341, no. 6142, pp. 157–160, 2013.

[8] M. Feng, J. Zhao, T. Huang, X. Zhu, and H. Petek, "The electronic properties of superatom states of hollow molecules," *Accounts of Chemical Research*, vol. 44, no. 5, pp. 360–368, 2011.

[9] X. K. Zhao, Y. Y. Zhang, J. Zhao, H. S. Hu, and J. Li, "Understanding the electronic structure and chemical bonding in the 2d fullerene monolayer," *Inorganic Chemistry*, vol. 63, no. 25, pp. 11572–11582, 2024.

[10] R. Suresh *et al.*, "Superatom molecular orbitals of endohedral c(82)," *Journal of Physical Chemistry A*, vol. 127, no. 39, pp. 8126–8132, 2023.

[11] C. Sikorska and N. Gaston, "Modified lennard-jones potentials for nanoscale atoms," *Journal of Computational Chemistry*, vol. 41, no. 22, pp. 1985–2000, 2020.

[12] J. T. A. Gilmour and N. Gaston, "5-fold symmetry in superatomic scandium clusters," *Physical Chemistry Chemical Physics*, vol. 22, no. 7, pp. 4051–4058, 2020.

[13] C. Sikorska, "Utilizing fluoroxyl groups as ligands in superhalogen anions," *Chemical Physics Letters*, vol. 638, pp. 179–186, 2015.

[14] C. Sikorska and T. Puzyn, "The performance of selected semiempirical and dft methods in studying c60 fullerene derivatives," *Nanotechnology*, vol. 26, no. 45, p. 455702, 2015.

[15] P. Venkatakrishnan *et al.*, "Superatom molecular orbital in c(80)," *Journal of Computational Chemistry*, vol. 45, no. 12, pp. 827–833, 2024.

[16] L. Hammerschmidt, J. Schacht, and N. Gaston, "First-principles calculations of the electronic structure and bonding in metal cluster-fullerene materials considered within the superatomic framework," *Physical Chemistry Chemical Physics*, vol. 18, no. 47, pp. 32541–32550, 2016.

[17] P. A. Troshin *et al.*, "Material solubility-photovoltaic performance relationship in the design of novel fullerene derivatives for bulk heterojunction solar cells," *Advanced Functional Materials*, vol. 19, no. 5, pp. 779–788, 2009.

[18] J. H. Choi, K. I. Son, T. Kim, K. Kim, K. Ohkubo, and S. Fukuzumi, "Thienyl-substituted methanofullerene derivatives for organic photovoltaic cells," *Journal of Materials Chemistry*, vol. 20, no. 3, pp. 475–482, 2010.

[19] M. Jeon *et al.*, "Silyl substituted methanofullerenes as electron acceptors in organic photovoltaic cells," *Molecular Crystals and Liquid Crystals*, vol. 519, pp. 266–275, 2010.

[20] H. U. Kim *et al.*, "Naphthalene-, anthracene-, and pyrene-substituted fullerene derivatives as electron acceptors in polymer-based solar cells," *ACS Applied Materials & Interfaces*, vol. 6, no. 23, pp. 20776–20785, 2014.

[21] C. M. Aikens, R. Jin, X. Roy, and T. Tsukuda, "From atom-precise nanoclusters to superatom materials," *Journal of Chemical Physics*, vol. 156, no. 17, p. 170401, 2022.

[22] E. A. Lubecka and A. Liwo, "New unres force field package with fortran 90," *Task Quarterly*, vol. 20, pp. 399–407, 2016.

[23] R. Ganzynkowicz, A. Liwo, and W. Wiczk, "A fluorescence, 1h nmr spectroscopy and molecular dynamics study of the influence of rotamer population on fluorescence decay of tyrosine, phenylalanine and their derivatives," *Task Quarterly*, vol. 5, pp. 311–316, 2001.

[24] M. A. Mozolewska, P. Krupa, B. Rasulev, A. Liwo, and J. Leszczyński, "Preliminary studies of interaction between nanotubes and toll-like receptors," *Task Quarterly*, vol. 18, pp. 351–355, 2014.

[25] M. A. Mozolewska, "Influence of substitutions of isu1 residues on binding to jac1 protein," *Task Quarterly*, vol. 20, no. 4, pp. 417–423, 2016.

[26] S. Kar, A. Gajewicz, T. Puzyn, and K. Roy, "Nano-quantitative structure-activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells," *Toxicology in Vitro*, vol. 28, no. 4, pp. 600–606, 2014.

[27] A. Gajewicz-Skretna, A. Furuhama, H. Yamamoto, and N. Suzuki, "Generating accurate in silico predictions of acute aquatic toxicity for a range of organic chemicals," *Chemosphere*, vol. 280, p. 130681, 2021.

[28] J. J. Stewart, "Stewart computational chemistry," 2007.

[29] J. E. Moussa and J. J. P. Stewart, "Mopac," 2025.

[30] J. Mazerski, *Podstawy chemometrii.* Wydawnictwo Politechniki Gdańskiej, 2000.

[31] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[32] K. A. Yeomans and P. A. Golder, "The guttman-kaiser criterion as a predictor of the number of common factors," *The Statistician*, vol. 31, no. 3, pp. 221–229, 1982.

[33] The MathWorks Inc., "Matlab r2024a," 2024.