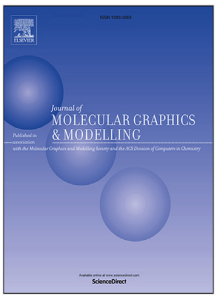
Fragments quantum descriptors in classification of bio-accumulative compounds

Fliszkiewicz, Bartłomiej; Sajdak, Marcin

DOI:
10.1016/j.jmgm.2023.108584

Journal Pre-proof



Fragments quantum descriptors in classification of bio-accumulative compounds

Bartłomiej Fliszkiewicz, Marcin Sajdak

PII: S1093-3263(23)00182-1
DOI: https://doi.org/10.1016/j.jmgm.2023.108584
Reference: JMG 108584

To appear in: *Journal of Molecular Graphics and Modelling*

Received date : 9 May 2023
Revised date : 24 July 2023
Accepted date : 29 July 2023

Please cite this article as: B. Fliszkiewicz and M. Sajdak, Fragments quantum descriptors in classification of bio-accumulative compounds, *Journal of Molecular Graphics and Modelling* (2023), doi: https://doi.org/10.1016/j.jmgm.2023.108584.

Fragments quantum descriptors in classification of bio-accumulative compounds

Bartłomiej Fliszkiewicz[a], Marcin Sajdak[b,c]

[a]*Department of New Technologies and Chemistry, Military University of Technology, Kaliskiego 2, Warsaw, 00-908, Poland*
[b]*Faculty of Energy and Environmental Engineering, Silesian University of Technology, Akademicka 2A, Gliwice, 44-109, Poland*
[c]*School of Chemical Engineering, University of Birmingham, S W Campus, Birmingham, B15 TT, United Kingdom*

Abstract

The aim of the following research is to assess the applicability of calculated quantum properties of molecular fragments as molecular descriptors in machine learning classification task. The research is based on bio-concentration and QM9-extended databases. A number of compounds with results from quantum-chemical calculations conducted with Psi4 quantum chemistry package was also added to the quantum properties database. Classification results are compared with a baseline of random guesses and predictions obtained with the traditional RDKit generated molecular descriptors. Chosen classification metrics show that results obtained with fragments quantum descriptors fall between results from baseline and those provided by molecular descriptors widely applied in cheminformatics. According to the results, the implementation of principal component analysis, causes a drop in categorization metrics.

Keywords: molecular descriptors, fragments quantum descriptors, machine learning, cheminformatics, quantum computing

1. Introduction

Because some xenobiotics can be stored within organisms at higher concentrations than measured in the environment, increasing their concentration across the trophic chain and achieving harmful levels in fish, wildlife, and humans, the bioconcentration factor (BCF) is essential in aquatic environmental assessments [1, 2]. Accumulation can take place at each trophic

level, either through the skin or respiratory surfaces (such as lungs or gills), [3, 4, 5, 6]. which is referred to as bioconcentration, or through the food that an organism consumes, which is referred to as dietary bioaccumulation [7]. This results in a rise in chemical concentration with increasing trophic level, which exposes organisms at high trophic levels, such as humans, to long-term repercussions that are difficult to predict, such as endocrine disruption [8]. Large regulatory efforts are being made to find and get rid of the compounds that bioaccumulate the most. This is because pollution is now seen as a threat to both the environment and human health, which has made people more worried. BCF data are in high demand because they are required by the European Commission's (EC) regulation Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) [9], and they may also be useful in the context of the Globally Harmonized System (GHS). REACH is an initiative that aims to enhance the protection of human and environmental health while simultaneously facilitating the free circulation of chemicals via the early identification of chemical properties. The Globally Harmonized System (GHS) standard is designed to offer a globally consistent framework for the categorization and labelling of chemicals. The rule known as the Classification, Labeling, and Packaging (CLP) regulation is responsible for putting the Globally Harmonized System into effect in Europe. This regulation is a part of REACH. The current European Regulation on the REACH [9] makes it essential, among other things, to estimate the rate of bio-accumulation for chemicals that are manufactured or imported in quantities that are greater than 10 tonnes per year. This regulation was passed in 2007. Expensive and time-consuming bio-concentration measurements are only needed for substances that are made or brought into the country in quantities of more than 100 tonnes per year. BCF is the requested criterion for bioaccumulation assessment in many regulatory frameworks [1], but its determination is very expensive (more than 35,000 euros) and requires the use of more than 100 animals for each standard study, which has led to a general lack of data in the field. Because of this, the development of models to anticipate BCF has been necessitated. In the context of regulation, the primary goals are to choose and use parameters that are fundamental and simple to compute (like log P), and to develop models that can assist in the prediction of the BCF data that is the most accurate possible [10]. In Annex XI of the REACH Act, the prerequisites for the appropriate selection and application of quantitative structure-activity relationship (QSAR) models for regulatory purposes are stated [11]. This sets the framework for the appli-

cation of QSAR models that are faster and more cost-effective in order to evaluate the bioaccumulation potential of chemicals, as well as the creation of predictive models. Over the course of several decades, a great number of conceptual models for predicting BCF have been created [12, 13, 14, 15]. However, in more recent years, there have been publications of a few categorization models. Regression models make up the vast bulk of these models. A normal distribution of the data set is supposed to be the case for regression models, and the vast majority of BCF data sets have a normal distribution. This might be one possible cause.

QSAR makes use of statistical and mathematical techniques to quantitatively connect a biological characteristic to molecular characteristics (such as structural features or physicochemical qualities), which are numerically stored inside so-called molecular descriptors. QSAR has become more significant in international decision-making frameworks [16], and the European REACH legislation [9] promotes its use for prioritising, data gap filling, and the rationalisation of animal testing. The vast majority of BCF QSAR models are constructed using octanol/water partition coefficient (K_{ow}) or other descriptors that are quite comparable to them [17]. In practice, bioconcentration happens predominantly as a thermodynamically driven partitioning between water and the lipid phases of organisms, as shown by K_{ow} [18]. However, other processes, such as metabolism and excretion, as well as specific interactions with tissues other than lipids, can greatly contribute to the apparent quantities that are present within organisms [19]. Chemicals that can be transformed into hydrophilic molecules may be eliminated more quickly, and as a result, their BCF values are lower than what is predicted by K_{ow} [20, 21]. However, substances that form specific connections with non-lipid tissues can have a bigger BCF than one would anticipate. One example of this is methylmercuric chloride, which has a low $\log K_{ow}$ but a very high BCF (up to 1,000,000 in fish) as a result of its linkage with protein sulfhydryl groups. Other examples of this type of compound include benzene, which has a low $\log K_{ow}$ but a very high BCF [22]. In a similar way, K_{ow}-based QSAR models, which vary in how complex they are, can either underestimate or overestimate the true BCF [23].

Many environmental contaminants, such as dichlorodiphenyl trichloroethane (DDT), hexachlorobenzene (HCB), dieldrin (HEOD), and polychlorinated biphenyls (PCBs), generated by industrial activities, pass through food chains, posing a risk of exposure to the general population [24, 25]. Chemicals with both a high lipophilicity and a high environmental persistency should be

thoroughly studied for potential toxicity via bioconcentration and bioaccumulation, both of which should be assessed over lengthy durations of exposure. The bioconcentration and bioaccumulation of chemical compounds in aquatic and terrestrial organisms are significant criteria for ecotoxicological evaluation and hazard assessment [26, 27, 28]. The few studies using QSAR methods also focused on the development of models to determine the BCF specifically for compounds containing a chlorine atom or atoms in their structure. The models developed were based on different quantitative superstructure/activity relationships (QSSAR) [27, 29]. These models have the potential to be valuable in forecasting the bioaccumulation capacity of novel compounds, a crucial aspect in evaluating their ecological implications. The limited number of studies that have explored the inclusion of compounds with chlorine atoms in their structure necessitates further research to enhance the precision and credibility of these models for a broader spectrum of chemical structures. This research uses a QSAR technique to categorise bioaccumulating compounds using a collection of novel molecular descriptors. Two benchmarks are used to assess the new descriptors: a random guess and a classification based on a well-known open source cheminformatics tool.

2. Materials and Methods

The research was conducted with the usage of Python [30], scikit-learn [31], matplotlib [32], Seaborn [33], RDKit[34], XGBoost [35] and Light GBM [36].

2.1. Fragments quantum descriptors (FQDs)

The study involved the creation of a non-traditional set of descriptors based on molecular substructures. Molecular descriptors focused on molecular fragments are widely applied in cheminformatics, some examples may be found in [37, 38, 39]. The process of calculating novel descriptors involved cross-referencing molecules from an experimental database with molecules from a separate database containing quantum properties of small compounds, in order to identify substructure matches. The process of matching substructures was carried out utilising the HasSubstructMatch method in RDKit. Quantitative and qualitative descriptors were computed based on the quantum properties of the detected substructures. Calculated quantum properties of whole molecules have been successfully applied as molecular descriptors [40, 41, 42] but in this research descriptors are derived from molecular

fragments. Broadly speaking, the descriptors comprise of the quantum properties of identified sub-structures that are categorised based on the count of atoms present in them. The values were subjected to averaging in order to render the descriptors unaffected by the quantity of substructures detected. Given that the descriptors are derived from substructures and the substructures database comprises molecules composed of a finite range of chemical elements, it is advisable to limit the experimental database to molecules that are also composed of the same elements.

$$
\begin{aligned}
&FQD^{i}_{prop} = \frac{\sum_{j(i)} prop_j * N_{occ}}{n}, \\
&prop - \text{quantum property of substructure } j \\
&i - \text{number of atoms in j-th substructure} \\
&j(i) - \text{a substructure containing } i \text{ number of atoms} \\
&n - \text{number of detected substructures} \\
&N_{occ} - \text{number of substructure occurences}
\end{aligned}
\tag{1}
$$

Both qualitative and quantitative FQDs were computed. The greatest value of N_{occ} in qualitative descriptors, which are a subclass of quantitative descriptors, was 1. This formulation led to the calculation of descriptors that only considered the existence of such a substructure. Qualitative descriptors also included the frequency of occurrences in the parent molecule. The aforementioned descriptors' per-atom variations were also determined by dividing the value of the descriptor by the number of atoms in the parent molecule. 352 quantum-based descriptors were created as a result, and these were used in the study.

2.2. Quantum properties database

QM9 database [43] contains computed properties of over 134 000 molecules made up of 9 atoms of C, O, N and F. Molecules included in the database were picked from GDB17 chemical universe [44]. The computations were conducted on B3LYP/6-31G(2df,p) level of theory. The source of the QM9 database was MoleculeNet [45] website. Recently, the database was extended by Lim et. al. [46] with compounds containing Cl and S atoms chosen from GDB17. Independently to QM9-extended database we also made an effort to extend the QM9 database. To achieve this objective, a number of SMILES were generated and quantum-chemistry calculations were conducted with

Psi4 [47] open source quantum chemistry calculations software. The process is described in next section. Quantum properties included in the QM9, QM9-extended databases and whether they were calculated with Psi4, are listed in Table 1. The final substructures database was a combination of QM9-extended and our own QM9 extension.

Table 1: Quantum properties in databases.

Property	Unit	Description	A	B	C
A	GHz	Rotational constant A	+	x	+
B	GHz	Rotational constant B	+	x	+
C	GHz	Rotational constant C	+	x	+
mu	Debye	Dipole moment	+	+	+
alpha	$Bohr^3$	Isotropic polarizability	+	+	+
homo	Hartree	Energy of HOMO	+	+	+
lumo	Hartree	Energy of LUMO	+	+	+
gap	Hartree	Gap, LUMO and HOMO difference	+	+	+
r2	$Bohr^2$	Electronic spatial extent	+	+	x
zpve	Hartree	Zero point vibrational energy	+	+	+
U0	Hartree	Internal energy at 0 K	+	+	+
U	Hartree	Internal energy at 298.15 K	+	+	+
H	Hartree	Enthalpy at 298.15 K	+	+	+
G	Hartree	Free energy at 298.15 K	+	+	+
Cv	cal/(mol K)	Heat capacity at 298.15 K	+	+	+

A - QM9; B - QM9-extended; C - calculated for this research with Psi4
+ - feature present in the database, x - feature missing in the database

2.3. Quantum chemistry calculations

In order to conduct the aforementioned calculations, a number of structures were generated from molecules appearing in the original QM9 database. The whole process is visualised in Figure 1. The original database was queried for molecules containing fluorine. Based on SMILES representation of the chosen molecules, new structures were generated by simple replacement of F to Cl. The resulting SMILES were employed to produce a set of input files to conduct the calculations. Initial geometries were generated with RDKit EmbedMolecule method which uses ETKDG method [48]. In compliance with the original QM9 database, B3LYP method was used in these calculations. Although due to initial plan to make calculations for molecules with F

exchanged with not only Cl but also Br and I, the def2-svp basis set was chosen. Dipole moments and polarizabilities of molecules were calculated using CCSD method. In the case of failure, the calculations were firstly repeated with enabled second order SCF. In the third step with enabled cartesian coordinates and enabled back transformation. Then the density of the DFT grid was expanded and the geometry convergence threshold tightened. The final step involved loosing the geometry convergence threshold. Molecules that still failed to yield results were dismissed. The resulting output files were read with AaronTools [49] which applies post calculations RRHO correction.

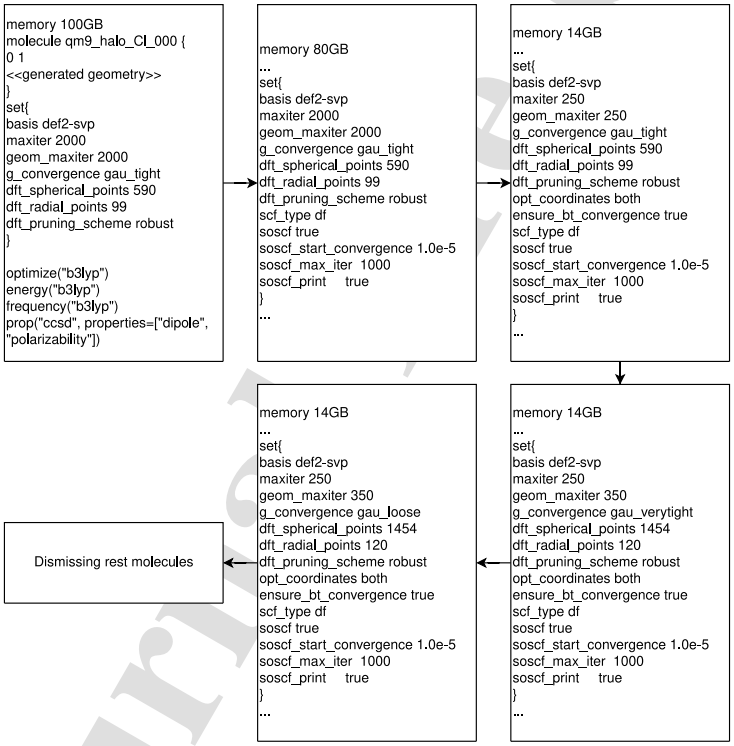


Figure 1: Steps of the process of extending QM9-extended database with chlorinated molecules. In case of failure of calculations in all five steps the molecules were dismissed.

Despite the initial plan to include the Br and I substructures in the research, they were excepted because the calculations were time consuming and their occurence in experimental database is limited.

2.4. Bio-concentration database

The experimental database mentioned in the previous section was the bio-concentration database [50, 51] which contains 1007 compounds that were derived from literature with their corresponding experimental logBCF values. These values were converted to BCF class either as bio-accumulative or non-bio-accumulative compounds with the threshold of logBCF of 3. The source of the database is qsardb.org [52]. A subset of compounds that contained only C, O, N, S, F and Cl atoms was chosen for further research. The original database was reduced from 1007 to 868 compounds, and which composition (as a distribution and elements allocation) is shown in Figure 2. Among these compounds, 681 were classified as non-bio-accumulative and 187 as bio-accumulative.

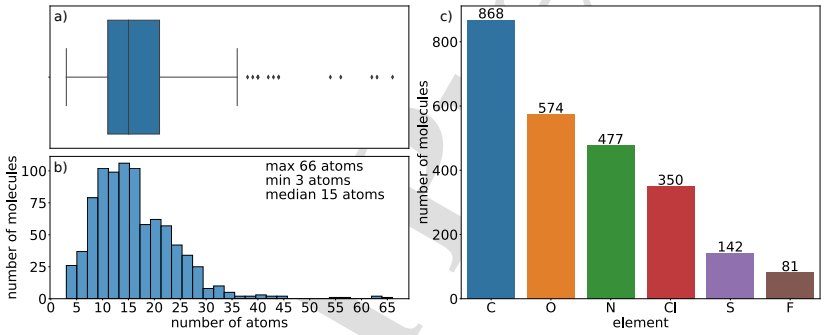


Figure 2: Curated bio-concentration database molecules composition. a) - the overall distribution of number of atoms in molecules in the database; b) - the number of molecules according to the number of atoms they contain; c) - the number of molecules that contain at least one atom of elements from C, O, N, F, Cl, S.

There are more molecules containing chlorine than fluorine.

Since the substructures database is not exhaustive in terms of all possible atoms combinations, it is possible that there might be compounds where no substructures would be found. Such case appeared with 12 compounds, thus the dataset was further reduced to 856 compounds. The compounds that lacked substructures are listed in the supplementary material.

2.5. Data processing

Due to the possibility that some fragment quantum descriptor values may be missing, three types of handling strategies were evaluated using 5-fold

cross validation. Leaving missing values, filling with zeros while adding additional variables containing information on whether the value was missing, and k-nearest neighbours imputation were the strategies evaluated. The bioconcentration classes were encoded with the values 0 and 1, with 1 denoting bioaccumulation. 330 molecules were sampled while maintaining the classes ratio. This procedure generated an out-of-the-bag (OOB) dataset for testing the final model. The remaining samples formed a training set. Since the data is class-imbalanced, random oversampling was used to modify the class distribution in the training set. The training set's values were then subject to standardisation. Two different approaches to data processing were also evaluated - either descriptors were introduced into machine learning algorithms right after standardization or principal component analysis (PCA) was also applied. The PCA was set to retain 95% of variance.

2.6. Model selection

The data was introduced into Logistic Regression, Random Forest, Gradient Boosting, k-nearest neighbours, Support Vector Classification, XGBoost and LightGBM algorithms. Models were checked in five-fold cross validation process. Oversampling, standardization and PCA were applied separately for every training fold of the data in cross validation. The validation part of the cross validation data fold underwent only standardization and PCA with parameters previously set on the training part of the data fold, thus the scoring metrics were evaluated with data with the original class ratio. The testing metrics used to evaluate models were F-score and balanced accuracy. Following good practices in machine learning studies [53, 54], in order to check whether the predictions are better than simple guesses, a random classifier was set as a baseline. One best performing algorithm was selected based on the aforementioned scoring results.

2.7. Hyperparameters tuning and OOB test

The algorithm's parameters were optimised in order to maximize model's F-score. Models with various parameter combinations were tested using 5-fold cross validation along the procedure The range of tested parameters is outlined in Supplementary Material. The selected algorithm with parameters tuned was trained with the whole training dataset and finally tested with OOB dataset. The training dataset underwent oversampling and preprocessing before the final test. The training dataset's preprocessing procedures were

also used for the test dataset. The final model was tested 5 times with different seeds of pseudorandom generators of the machine learning algorithms. The diagram depicting the whole process is shown in Figure 3.

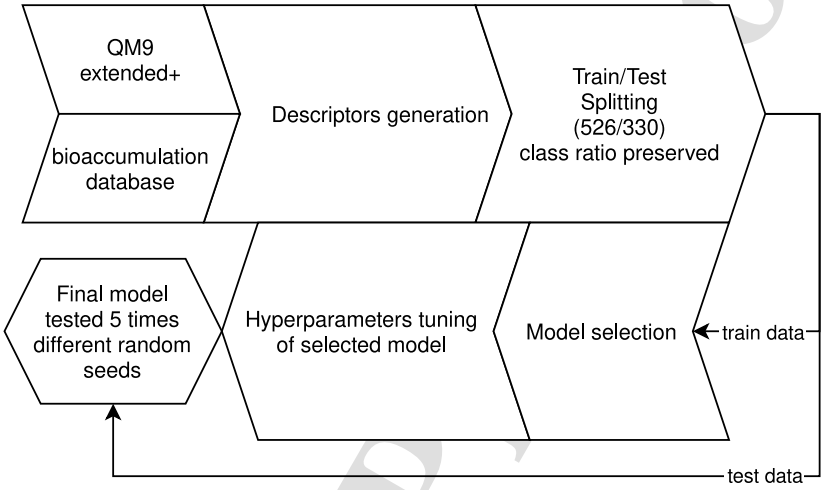


Figure 3: Step by step visualization of the research.

In order to create a full picture of the potential of the novel molecular descriptors, the research process was repeated with 2 other sets of molecular descriptors:

- RDKit generated molecular descriptors (208), MACCS keys (166) and Morgan Fingerprints (2048),
- combination of novel and abovementioned descriptors.

In both cases since MACCS keys and Morgan Fingerprints are binary data, the PCA was applied only to FQDs and RDKit molecular descriptors.

3. Results and discussion

3.1. Fragments quantum descriptors

Due to the fact that two different variants of descriptors were calculated, it was determined by comparing the variance coefficients whether or not the information they contain differs. The findings indicated that there is

no difference in the parameter between the related FQDs, leading to the conclusion that using both qualitative and quantitative descriptors would not have an impact on the study's findings. As a result, the research only used qualitative FQDs.

3.2. Quantum-chemical calculations

In the case of our own QM9 extension, out of 2163 calculations, 99 failed for various reasons. Since QM9-extended database was published during our own calculations related to this article, there appeared to be an overlap of 290 molecules. In these cases, while combining the results of calculations with QM9-extended database, properties from QM9-extended were considered as bearing priority and taken into the research. The resulting database of quantum properties (QM9-extended-plus) contained 155468 compounds with 11 calculated properties.

3.3. Handling of missing data

Leaving missing values limited the number of applicable machine learning algorithms to XGBoost and LightGBM. Such approach to resolving missing data problem was outperformed by other applied techniques. Figure 4 shows the results of F-score and balanced accuracy of handling missing data by filling with zeros and KNN imputation. Although KNN imputation performed better in certain instances, such as for KNN and SVC, other algorithms performed better for replacing missing values with zeros, or both approaches produced results that were similar.

3.4. Most accurate machine learning algorithm

The best performing algorithm was determined to be the Light Gradient Boosting Machine (LGBM) based on the metrics values in cross validation (Figure 5). All models outperformed the baseline. The top performing algorithm for RDKit-generated molecular descriptors was XGBoost, but LGBM performed best when FQDs and RDKit-generated descriptors were combined. The most effective methods for PCA-based preprocessing were Random Forest for both only FQDs and only RDKit descriptors and Gradient Boosting for the combination of FQDs and RDKit descriptors. The cross validation results are demonstrated in the Suppmementary Material.

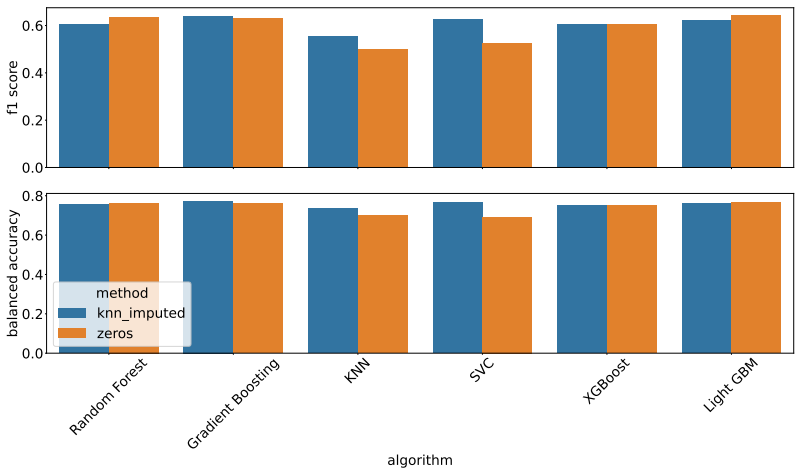


Figure 4: 5-fold cross validation scores of handling missing data strategies.

3.5. Out of bag test

The 330 molecules previously unused in the research that were separated from reduced original dataset were used to test final models. Table 2 and 3 show achieved balanced accuracy, and F-score of baseline and 3 sets of predictors with 5 different seeds of pseudorandom generators of the machine learning algorithms. Molecular descriptors generated by RDKit produce greater results in terms of F-score and balanced accuracy. The usage of FQDs and RDKit descriptors together produced predictions with improved specificity and precision - please refer to Supplementary Material to investigate specificity, sensitivity, precision and accuracy obtained in OOB test. The application of PCA caused a significant drop in the evaluated metrics.

3.6. Principal components analysis

While preserving 95% of variance, PCA reduced the number of machine learning features to 18 components (Figure 6). The most contribution in first principal component (PC) was attributed to quantum properties of 4 atoms fragments detected in parent molecules. Fragments built up of 8 and 7 atoms had biggest impact on second PC. Third component was composed mainly of properties of 2 atoms fragments.

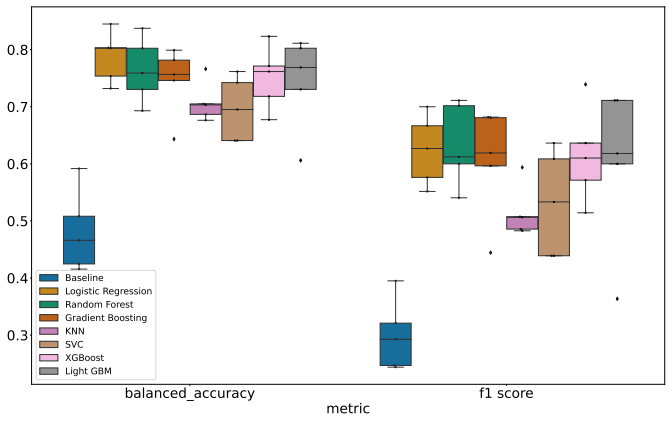


Figure 5: 5-fold cross validation scores regarding FQDs molecular representation. Boxplots and individual points are shown.

Table 2: OOB test results when PCA was part of preprocessing.

metric	test no.	Baseline	molecular representation		
			FQDs	RDKit	combined
F-score	I	0.444	0.594	0.735	0.682
	II	0.444	0.586	0.711	0.672
	III	0.444	0.594	0.725	0.652
	IV	0.444	0.588	0.720	0.657
	V	0.444	0.603	0.715	0.652
balanced accuracy	I	0.458	0.751	0.834	0.789
	II	0.458	0.747	0.829	0.788
	III	0.458	0.751	0.838	0.777
	IV	0.458	0.746	0.836	0.779
	V	0.458	0.759	0.842	0.777

Amongst first 7 PCs the most informative quantum properties were dipole moment, zero point vibrational energy, free energy, internal energy and enthalpy at 298.15K, heat capacity, energy gap between LUMO and HOMO, polarizability.

Since the application of principal component analysis caused a drop in the

Table 3: OOB test results when PCA was not applied

metric	test no.	Baseline	molecular representation		
			FQDs	RDKit	combined
F-score	I	0.444	0.620	0.759	0.729
	II	0.444	0.620	0.759	0.729
	III	0.444	0.620	0.759	0.729
	IV	0.444	0.620	0.759	0.729
	V	0.444	0.620	0.759	0.729
balanced accuracy	I	0.458	0.759	0.875	0.829
	II	0.458	0.759	0.875	0.829
	III	0.458	0.759	0.875	0.829
	IV	0.458	0.759	0.875	0.829
	V	0.458	0.759	0.875	0.829

scoring metrics, the effect was investigated by estimating mutual information between bioaccumulation class and used descriptors.

The analysis proved that, in the dataset, bioaccumulation class is mostly associated to FQDs resulting from fragments containing 7, 8 and 9 atoms. Results from mutual information regarding quantum properties of fragments are more consistent with PCA - the most informative properties are dipole moment, internal energy at 0K, enthalpy and internal energy at 298.15K, energy gap between LUMO and HOMO, zero point vibrational energy, heat capacity, polarizability.

Mutual information between principal components and bioaccumulation class was also estimated. The first two principal components were estimated as most informative only in the case of RDKit generated descriptors, eg. the mutual information score of PC4 from combined descriptors was 3 times higher than PC1.

4. Conclusions

Quantum properties of molecular fragments allowed classification of bioaccumulative compounds with better accuracy than the baseline. However, molecular descriptors generated from quantum features of molecular fragments are inferior to RDKit generated descriptors in classification of bioaccumulative properties. Additionally, generation of the traditional molecular descriptors require less computational resources.

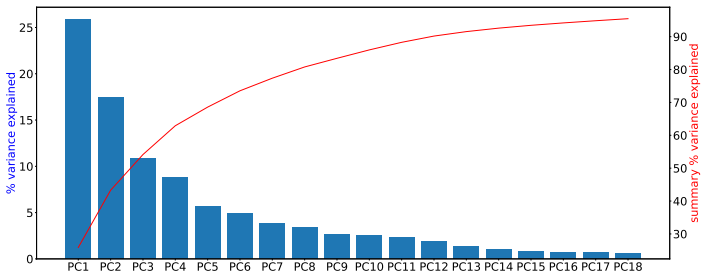


Figure 6: Scree plot of principal components obtained from FQDs. First 6 components account for 80% of variance. The red line is the cumulative sum of principal components' explained variance.

The biggest disadvantage of FQDs is that the database in which molecular fragments are searched for contains only fragments composed of a limited set of chemical elements. For this reason if generated from molecules built of other elements, FQDs could provide less information and fail to form a strong relationship with chemical properties. In the research, quantum-chemical calculations provided a further extension of QM9-extended database, thus enriching the set of possible substructures that could be detected. It is also a contribution to the publicly available quantum calculations databases.

In the research the application of principal component analysis caused a decrease in the evaluated classification metrics. Which means that the variance in the considered groups of molecular descriptors did not highly correspond to the bioaccumulation classes.

Since the research focused on a narrow QSAR application, further research towards applicability of fragments quantum descriptors should be conducted. The problem of missing FQD's values should also be addressed.

5. Data and Software Availability

The code in the form of Jupyter Notebooks is available in supplementary material. The QM9-extended-plus dataset is available at Zenodo [55].

Acknowledgements

Quantum chemistry calculations were conducted using the infrastructure of Center for Advanced Studies Systems Engineering, Cybernetics Faculty, Military University of Technology.

This work was financed by Military University of Technology, Warsaw, Poland under research project UGB 803/2023.

Supplementary Information

Additional figures are available in supplementary material.

References

[1] J. A. Arnot, F. A. Gobas, A review of bioconcentration factor (bcf) and bioaccumulation factor (baf) assessments for organic chemicals in aquatic organisms, Environmental Reviews 14 (4) (2006) 257–297. arXiv:https://doi.org/10.1139/a06-005, doi:10.1139/a06-005.
URL https://doi.org/10.1139/a06-005

[2] Y. Wang, Y. Wen, J. J. Li, J. He, W. C. Qin, L. M. Su, Y. H. Zhao, Investigation on the relationship between bioconcentration factor and distribution coefficient based on class-based compounds: The factors that affect bioconcentration, Environmental toxicology and pharmacology 38 (2) (2014) 388–396.

[3] Y.-C. J. Chen, Y.-L. Guo, C.-C. Hsu, W. J. Rogan, Cognitive Development of Yu-Cheng ('Oil Disease') Children Prenatally Exposed to Heat-Degraded PCBs, JAMA 268 (22) (1992) 3213–3218. arXiv:https://jamanetwork.com/journals/jama/articlepdf/401742/jama_268_22_028.pdf, doi:10.1001/jama.1992.03490220057028.
URL https://doi.org/10.1001/jama.1992.03490220057028

[4] P. M. Cook, J. A. Robbins, D. D. Endicott, K. B. Lodge, P. D. Guiney, M. K. Walker, E. W. Zabel, R. E. Peterson, Effects of aryl hydrocarbon receptor-mediated early life stage toxicity on lake trout populations in lake ontario during the 20th century, Environmental Science & Technology 37 (17) (2003) 3864–3877.

[5] B. C. Gladen, W. J. Rogan, P. Hardy, J. Thullen, J. Tingelstad, M. Tully, Development after exposure to polychlorinated biphenyls and dichlorodiphenyl dichloroethene transplacentally and through human milk, The Journal of pediatrics 113 (6) (1988) 991–995.

[6] D. Ratcliffe, Decrease in eggshell weight in certain birds of prey, Nature 215 (5097) (1967) 208–210.

[7] F. Gobas, H. A. Morrison, Bioconcentration and biomagnification in the aquatic environment, Handbook of property estimation methods for chemicals (2000) 189–231.

[8] H. J. Geyer, G. G. Rimkus, I. Scheunert, A. Kaune, K.-W. Schramm, A. Kettrup, M. Zeeman, D. C. Muir, L. G. Hansen, D. Mackay, Bioaccumulation and occurrence of endocrine-disrupting chemicals (edcs), persistent organic pollutants (pops), and other organic compounds in fish and other organisms including humans, in: Bioaccumulation–New Aspects and Developments, Springer, 2000, pp. 1–166.

[9] Corrigendum to regulation (ec) no 1907/2006 of the european parliament and of the council of 18 december 2006 concerning the registration, evaluation, authorisation and restriction of chemicals (reach), establishing a european chemicals agency, amending directive 1999/45/ec and repealing council regulation (eec) no 793/93 and commission regulation (ec) no 1488/94 as well as council directive 76/769/eec and commission directives 91/155/eec, 93/67/eec, 93/105/ec and 2000/21/ec (oj l 396, 30.12.2006, p. 1) (corrected version in oj l 136, 29.5.2007, p. 3) (May 2008).

[10] R. Garg, C. J. Smith, Predicting the bioconcentration factor of highly hydrophobic organic chemicals, Food and chemical toxicology 69 (2014) 252–259.

[11] R. Cesnaitis, M. A. Sobanska, B. Versonnen, T. Sobanski, V. Bonnomet, J. V. Tarazona, W. De Coen, Analysis of the ecotoxicity data submitted within the framework of the reach regulation. part 3. experimental sediment toxicity assays, Science of the total environment 475 (2014) 116–122.

[12] J. Dearden, Qsar modeling of bioaccumulation, Predicting chemical toxicity and fate (2004) 333–355.

[13] M. Pavan, A. P. Worth, T. I. Netzeva, Review of qsar models for bioconcentration, European Commission, Joint Research Centre: Ispra, Italy (2006).

[14] G. Piir, S. Sild, A. Roncaglioni, E. Benfenati, U. Maran, Qsar model for the prediction of bio-concentration factor using aqueous solubility and descriptors considering various electronic effects, SAR and QSAR in Environmental Research 21 (7-8) (2010) 711–729.

[15] A. P. Toropova, A. A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, Coral: Quantitative models for estimating bioconcentration factor of organic compounds, Chemometrics and Intelligent Laboratory Systems 118 (2012) 70–73. doi:https://doi.org/10.1016/j.chemolab.2012.08.002.
URL https://www.sciencedirect.com/science/article/pii/S0169743912001578

[16] M. T. D. Cronin, J. D. Walker, J. S. Jaworska, M. H. I. Comber, C. D. Watts, A. P. Worth, Use of qsars in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances., Environmental Health Perspectives 111 (10) (2003) 1376–1390. doi:10.1289/ehp.5759.

[17] M. Pavan, T. I. Netzeva, A. P. Worth, Review of literature-based quantitative structure–activity relationship models for bioconcentration, QSAR & Combinatorial Science 27 (1) (2008) 21–31.

[18] D. Mackay, Correlation of bioconcentration factors, Environmental Science & Technology 16 (5) (1982) 274–278, pMID: 22257252. doi:10.1021/es00099a008.

[19] ECETOC, The role of bioaccumulation in environmental risk assessment: The aquatic environment and related food webs (1995).

[20] W. de Wolf, J. H. de Bruijn, W. Seinen, J. L. Hermens, Influence of biotransformation on the relationship between bioconcentration factors and octanol-water partition coefficients, Environmental science & technology 26 (6) (1992) 1197–1201.

[21] D. C. Muir, B. R. Hobden, M. R. Servos, Bioconcentration of pyrethroid insecticides and ddt by rainbow trout: uptake, depuration, and effect

of dissolved organic carbon, Aquatic Toxicology 29 (3) (1994) 223–240. doi:https://doi.org/10.1016/0166-445X(94)90070-1.
URL https://www.sciencedirect.com/science/article/pii/0166445X94900701

[22] R. E. Reinert, L. J. Stone, W. A. Willford, Effect of temperature on accumulation of methylmercuric chloride and p, p' ddt by rainbow trout (salmo gairdneri), Journal of the Fisheries Board of Canada 31 (10) (1974) 1649–1652.

[23] F. Grisoni, M. Cassotti, R. Todeschini, Reshaped sequential replacement for variable selection in qspr: comparison with other reference methods, Journal of Chemometrics 28 (4) (2014) 249–259.

[24] L. Carlsen, J. D. Walker, Qsars for prioritizing pbt substances to promote pollution prevention, QSAR & Combinatorial Science 22 (1) (2003) 49–57.

[25] T. Feijtel, P. Kloepper-Sams, K. Den Haan, R. Van Egmond, M. Comber, R. Heusel, P. Wierich, W. Ten Berge, A. Gard, W. De Wolf, et al., Integration of bioaccumulation in an environmental risk assessment, Chemosphere 34 (11) (1997) 2337–2350.

[26] N. Judd, W. C. Griffith, E. M. Faustman, Contribution of pcb exposure from fish consumption to total dioxin-like dietary exposure, Regulatory Toxicology and Pharmacology 40 (2) (2004) 125–135.

[27] T. Ivanciuc, O. Ivanciuc, D. J. Klein, Modeling the bioconcentration factors and bioaccumulation factors of polychlorinated biphenyls with posetic quantitative super-structure/activity relationships (qssar), Molecular Diversity 10 (2) (2006) 133–145.

[28] L. Carlsen, A qsar approach to physico-chemical data for organophosphates with special focus on known and potential nerve agents, Internet Electronic Journal of Molecular Design 4 (2005) 355–366.

[29] L. Carlsen, P. Sørensen, M. Thomsen, R. Brüggemann, Qsar's based on partial order ranking, SAR and QSAR in Environmental Research 13 (1) (2002) 153–165.

[30] G. Van Rossum, F. L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[32] J. D. Hunter, Matplotlib: A 2d graphics environment, Computing in Science & Engineering 9 (3) (2007) 90–95. doi:10.1109/MCSE.2007.55.

[33] M. L. Waskom, seaborn: statistical data visualization, Journal of Open Source Software 6 (60) (2021) 3021. doi:10.21105/joss.03021.
URL https://doi.org/10.21105/joss.03021

[34] RDKit: Open-source cheminformatics, http://www.rdkit.org, [Online; accessed 11-February-2021].

[35] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, NY, USA, 2016, pp. 785–794. doi:10.1145/2939672.2939785.
URL http://doi.acm.org/10.1145/2939672.2939785

[36] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.
URL https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb

[37] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, Isida property-labelled fragment descriptors, Molecular Informatics 29 (12) (2010) 855–868. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201000099, doi:https://doi.org/10.1002/minf.201000099.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201000099

[38] L. Chao, H. Mei, X. Pan, W. Tan, T. Liu, L. Yang, Combinations of fragment descriptors for improved prediction of cyp2c19 inhibitors, Chemometrics and Intelligent Laboratory Systems 130 (2014) 109–114. doi:https://doi.org/10.1016/j.chemolab.2013.10.013.
URL https://www.sciencedirect.com/science/article/pii/S0169743913001949

[39] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, Universal fragment descriptors for predicting properties of inorganic crystals, Nature Communications 8 (1) (2017) 15679. doi:10.1038/ncomms15679.
URL https://doi.org/10.1038/ncomms15679

[40] M. Karelson, V. S. Lobanov, A. R. Katritzky, Quantum-chemical descriptors in qsar/qspr studies, Chemical Reviews 96 (3) (1996) 1027–1044. doi:10.1021/cr950202r.
URL https://doi.org/10.1021/cr950202r

[41] S. Chtita, A. Belhassan, M. Bakhouch, A. I. Taourati, A. Aouidate, S. Belaidi, M. Moutaabbid, S. Belaaouad, M. Bouachrine, T. Lakhlifi, Qsar study of unsymmetrical aromatic disulfides as potent avian sars-cov main protease inhibitors using quantum chemical descriptors and statistical methods, Chemometrics and Intelligent Laboratory Systems 210 (2021) 104266. doi:https://doi.org/10.1016/j.chemolab.2021.104266.
URL https://www.sciencedirect.com/science/article/pii/S0169743921000344

[42] L. Wang, J. Ding, L. Pan, D. Cao, H. Jiang, X. Ding, Quantum chemical descriptors in quantitative structure–activity relationship models and their applications, Chemometrics and Intelligent Laboratory Systems 217 (2021) 104384. doi:https://doi.org/10.1016/j.chemolab.2021.104384.
URL https://www.sciencedirect.com/science/article/pii/S0169743921001520

[43] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, Scientific Data 1 (1) (2014) 140022. doi:10.1038/sdata.2014.22.
URL https://doi.org/10.1038/sdata.2014.22

[44] L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17, Journal of Chemical Information and Modeling 52 (11) (2012) 2864–2875, pMID: 23088335. arXiv:https://doi.org/10.1021/ci300415d, doi:10.1021/ci300415d.
URL https://doi.org/10.1021/ci300415d

[45] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, Moleculenet: a bench-

mark for molecular machine learning, Chem. Sci. 9 (2018) 513–530. doi:10.1039/C7SC02664A.
URL http://dx.doi.org/10.1039/C7SC02664A

[46] M. A. Lim, S. Yang, H. Mai, A. C. Cheng, Exploring deep learning of quantum chemical properties for absorption, distribution, metabolism, and excretion predictions, Journal of Chemical Information and Modeling (Jun 2022). doi:10.1021/acs.jcim.2c00245.
URL https://doi.org/10.1021/acs.jcim.2c00245

[47] D. G. A. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O'Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer, A. Y. Sokolov, K. Patkowski, A. E. DePrince, U. Bozkaya, R. A. King, F. A. Evangelista, J. M. Turney, T. D. Crawford, C. D. Sherrill, Psi4 1.4: Open-source software for high-throughput quantum chemistry, The Journal of Chemical Physics 152 (18) (2020) 184108. arXiv:https://doi.org/10.1063/5.0006002, doi:10.1063/5.0006002.
URL https://doi.org/10.1063/5.0006002

[48] S. Riniker, G. A. Landrum, Better informed distance geometry: Using what we know to improve conformation generation, Journal of Chemical Information and Modeling 55 (12) (2015) 2562–2574. doi:10.1021/acs.jcim.5b00654.
URL https://doi.org/10.1021/acs.jcim.5b00654

[49] V. M. Ingman, A. J. Schaefer, L. R. Andreola, S. E. Wheeler, Qchasm: Quantum chemistry automation and structure manipulation, WIREs Computational Molecular Science 11 (4) (2021) e1510. arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1510, doi:https://doi.org/10.1002/wcms.1510.
URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1510

[50] G. Piir, S. Sild, U. Maran, Classifying bio-concentration factor with random forest algorithm, influence of the bio-accumulative vs. non-bio-accumulative compound ratio to modelling result, and applicability domain for random forest model, SAR and

QSAR in Environmental Research 25 (12) (2014) 967–981, pMID: 25482723. arXiv:https://doi.org/10.1080/1062936X.2014.969310, doi:10.1080/1062936X.2014.969310.
URL https://doi.org/10.1080/1062936X.2014.969310

[51] G. Piir, S. Sild, U. Maran, Data for: Classifying bio-concentration factor with random forest algorithm, influence of the bio-accumulative vs. non-bio-accumulative compound ratio to modelling result, and applicability domain for random forest model. qsardb repository, qdb.116. 2014. http://dx.doi.org/10.15152/qdb.116.

[52] V. Ruusmann, S. Sild, U. Maran, Qsar databank repository: open and linked qualitative and quantitative structure-activity relationship models, J. Cheminf 7 (2015) 32.

[53] N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, A. Walsh, Best practices in machine learning for chemistry, Nature Chemistry 13 (6) (2021) 505–508. doi:10.1038/s41557-021-00716-z.
URL https://doi.org/10.1038/s41557-021-00716-z

[54] A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, T. D. Sparks, Machine learning for materials scientists: An introductory guide toward best practices, Chemistry of Materials 32 (12) (2020) 4954–4965. arXiv:https://doi.org/10.1021/acs.chemmater.0c01907, doi:10.1021/acs.chemmater.0c01907.
URL https://doi.org/10.1021/acs.chemmater.0c01907

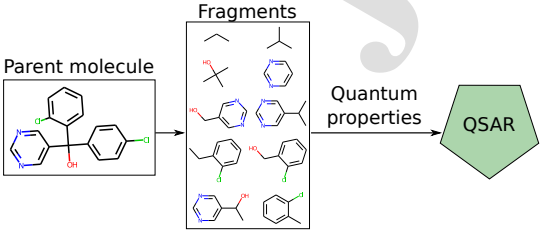
[55] B. Fliszkiewicz, M. Sajdak, Qm9-extended-plus database, License CC-BY-NC-SA (Aug. 2022). doi:10.5281/zenodo.7845173.
URL https://doi.org/10.5281/zenodo.7845173

Highlights

Fragments quantum descriptors in classification of bio-accumulative compounds

Bartłomiej Fliszkiewicz, Marcin Sajdak

- New type of molecular descriptors are proposed.
- Large database of quantum properties has been further extended.
- More than 800 compounds are used to establish prediction models.
- A simple model with strong predictive power is established.
- The overall predictive accuracy is larger than 0.80.



Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: