JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

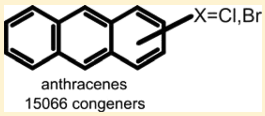
Combinatorial × Computational × Cheminformatics (C^3) Approach to Characterization of Congeneric Libraries of Organic Pollutants

Maciej Haranczyk,*[,†] Piotr Urbaszek,[‡] Esmond G. Ng,[†] and Tomasz Puzyn[‡]

[†]Computational Research Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Mail Stop 50F-1650, Berkeley, California 94720-8139, United States

[‡]Laboratory of Environmental Chemometrics, Department of Chemistry, University of Gdańsk, Sobieskiego 18/19, 80-952 Gdańsk, Poland

ABSTRACT: Congeners are molecules based on the same carbon skeleton but are different by the number of substituents and/or a substitution pattern. Examples are 1-chloronaphthalene, 1,4-dichloronaphthalene, and 1,3,8-trichloronaphthalene. Various persistent organic pollutants (POPs) exist in the environment as families of congeners. Very large numbers of possible congeners make their experimental characterization and risk assessment unfeasible. Computational high-throughput and quantitative structure−property relationship (QSPR) modeling has been limited by the lack of tools and approaches facilitating analysis of such POP families. We present a comprehensive approach that enables modeling of extremely large congeneric libraries. The approach involves three steps: (1) combinatorial generation of a library of congeners, (2) quantum chemical characterization of each structure at the PM6 semiempirical level to obtain molecular descriptors, and (3) analysis of the information generated in step 2. In steps 1−3, we employ combinatorial, computational, and cheminformatics techniques, respectively. Therefore, this hybrid approach is named "Combinatorial × Computational × Cheminformatics", or just abbreviated as C^3 (or C-cubed) approach. We demonstrate the usefulness of this approach by generating and characterizing Br- and Cl-substituted congeneric families of 23 typical POPs. The analysis of the resulting set of 1 840 951 congeners that includes Cl-, Br-, and mixed Br/Cl-substituted species, proves that, based on structural similarities defined by the molecular descriptors' values, the existing QSPR models developed originally for Cl- and Br-substituted congeners can be applied also to mixed Br/Cl-substituted ones. Thus, the C^3 approach may serve as a tool for exploring structural applicability domains of the existing QSPR models for congeneric sets.



INTRODUCTION

Many environmental studies are focused nowadays on groups of chemical pollutants characterized by high persistence (P), bioaccumulation ability (B), and toxicity (T). One of the most significant subset of the PBT chemicals is persistent organic pollutants (POPs).[1] The typical representatives of POPs include polycyclic aromatic hydrocarbons (PAHs) and their halogen-substituted derivatives such as chlorine-containing molecules like polychlorinated dibenzo-*p*-dioxins (PCDDs), polychlorinated dibenzofurans (PCDFs), polychlorinated biphenyls (PCBs), polychlorinated naphthalenes (PCNs), and pesticides like 1,1,1-trichloro-2,2-di(4-chlorophenyl)ethane (DDT), as well as groups of brominated flame retardants (BFRs) such as polybrominated diphenyl ethers (PBDEs), polybrominated biphenyls (PBBs), and others. The widespread application of pesticides, plasticizers, massive disposal of BFR containing materials (e.g., discarded televisions and monitors, computers, paints, furnishings for car interiors), pollution from chemical plants, thermal recycling of waste, and fossil fuel combustion contribute to the total amount of POPs introduced into the environment. Substantial volumes of these compounds are also released in the effect of giant fires, such as the recent fire of the oil spill at the Deepwater Horizon platform in the Gulf of Mexico.[2] Regardless of their source, the exposure to POPs can cause a vast range of acute and chronic health effects, including mutagenic, carcinogenic, and metabolic ones. In addition, as persistent and liphophilic substances, POPs can be bioaccumulated in human and animal tissues,[3] and biomagnified in natural ecosystems.[4]

Many POPs have already been identified as a serious environmental threat. Therefore, there is an urgent need to perform comprehensive studies to determine environmental and health risks coming from all possible POPs, since novel, unstudied structures can be formed in effect of various chemical reactions occurring in the environment. Well-studied POPs become substrates for synthesis of novel species (e.g., mixed Br/Cl-substituted derivatives can be formed in effect of chemical degradation of Br-substituted compounds based on the same carbon skeleton). Risk assessment for the unstudied POPs can be performed through established protocols on the basis of physicochemical properties.[4−6] The major challenge comes from the fact that, in most cases, the POPs of interests exist in large families of congeners. The term "congener" describes an individual member of a chemical family, based on the same carbon skeleton and differing by the number of substituents and/or a substitution pattern, e.g. 1-chloronaphthalene, 1,4-dichloronaphthalene, 1,3,8-trichloronaphthalene, and so on. Recent experimental studies have reported diverse families of POP-like congeners that contain not only rather

Received: June 21, 2012

   dx.doi.org/10.1021/ci300289b | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

typical Cl- or Br- substituents but also F atoms, hydroxyl (HO−) and methoxy (CH_3O−) groups, or mixes thereof.[7−12] The number of congeners in a typical congeneric family created by substitution with only one type of substituent usually reaches a hundred or more.[13] The number of possible congeners obtained by substitution with two types of chemical groups, e.g. Cl- and Br-substituted, is at least 1 order of magnitude larger, and roughly increases by 1 order of magnitude with each additional substituent group type present in a molecule.[13] For such a large number of compounds, empirical measurement of the physicochemical properties is impossible, due to high costs and time limitations of the analytical procedures. An exciting alternative approach for physicochemical characterization of POPs is based on computation, in particular the quantitative structure−property relationships (QSPR) approach. QSPR modeling is based on the assumption that the physicochemical property of interest can be expressed as a mathematical function of a chemical structure, represented by a set of molecular descriptors. This mathematical function is obtained by fitting a mathematical model to the experimental data, available only for some representatives of a group of compounds. It is then possible to interpolate the physicochemical properties for compounds, for which such data is missing, from the calculated molecular descriptors and the trained model.[14]

QSPR techniques have been already used to predict physicochemical properties of congeners in families of small POPs with only one type of substituent, typically chlorine or bromine. The studied properties included *n*-octanol/water (K_{OW}), *n*-octanol/air (K_{OA}), air/water (K_{AW}) partition coefficients, persistence (half-lives) in air, water, soil and sediments, subcooled vapor pressure (P_L), water solubility (S_W), and others.[15−19] However, the application of QSPR techniques to predict physicochemical properties of congeners in large libraries of POPs has been limited by a number of factors: (i) lack of software tools for combinatorial generation of congeneric sets of molecules; (ii) high computational cost of quantum chemical descriptors suitable for QSPR modeling of very similar structures such as congeners; and (iii) scattered or lacking experimental data, in particular for mixed-substituent-type congeners. In the recent years we have begun to address challenges i and ii.

In our previous contributions, we developed a congener generator software package, ConGENER,[20] which was inspired by our earlier tool for combinatorial generation of tautomers.[21] It enabled automatic generation of congeners with one type of halogen substituent. However, its applicability was limited to a small number of POPs due to the lack of symmetry detection routine, which might result in generating duplicate structures. We also investigated molecular descriptors, which are based on results of quantum chemical calculations and can be used to characterize structurally similar congeners of POPs. The most important conclusion of the latter study[22] was that descriptors based on recent, relatively inexpensive semiempirical methods, such as PM6 and RM1, allow building QSPR models with similar quality to that of models based on descriptors obtained by much more expensive, but (at that time) widely used, DFT/B3LYP methods. Combining the latter conclusion with the application of the ConGENER package has allowed us to routinely perform characterization and risk assessment of Cl- or Br-substituted congeneric libraries of small POPs. The total number of congeners involved in those studies was on the order of ca. 1500, which is only a small fraction of congeners that need to be analyzed to enhance our understanding of the health and environmental risks carried by congeners of POPs.

In the current work, we present a comprehensive, high-throughput approach to characterization of extremely large congeneric libraries of POPs. This approach constitutes three components responsible for the following: (i) combinatorial generation of congeneric libraries of POPs; (ii) quantum chemical characterization of each congener in the generated library; and (iii) further analysis of the resulting data. Steps i−iii correspond to the methodology employed, namely combinatorial, computational, and cheminformatics techniques, respectively. Therefore, following the convention established in our work on tautomers,[23] we name this hybrid approach as "Combinatorial × Computational × Cheminformatics", or just abbreviated as the C^3 (or C-cubed) approach. This also includes a new version of our congener generator software package—ConGENER—which allows generation of multi-substituent type congeners at a high level of automation. In the following sections, we will briefly describe the implemented congener library generation procedure and demonstrate its application by generation of possible congeners of 23 organic molecules obtained by substitution of hydrogen atoms with up to two chemical group types (Cl and Br atoms). The selected molecules represent typical organic pollutants: dioxin-like compounds and polyaromatic hydrocarbons. The resulting library of 1 840 951 congeners is characterized to obtain 26 molecular descriptors including 18 quantum chemical descriptors. The cheminformatics part of the C^3 approach has an objective to assess the structural similarities and differences within the resulting multivariate and megaobjective data set (1 840 951 × 26). In particular, we investigate the distribution in the descriptor space of brominated, chlorinated, and mixed congeners from the diverse library (congeners with different number of aromatic rings and their position in respect to each other) and whether this distribution will enable the development of global QSAR models with predictive capacity.

METHODS

Combinatorial Generation of a Library of Congeners for a Given Molecule. Our procedure of generating possible congeners of a given POP molecule consists of three steps: (i) the parent, unsubstituted POP molecule is analyzed to identify which of its atoms are equivalent by symmetry; (ii) all possible substitution patterns are enumerated and analyzed using the information of step i to finally reduce the set to the set of only nonredundant congeners; (iii) structures (Cartesian coordinates of all atoms) are generated and written to files. Steps i and ii were described in our earlier study focused estimation of numbers of possible mixed-substituent-type congeners[13] and, therefore, will be only be briefly described here.

The identification of equivalent atoms of the parent molecule is done using molecular graph representations and Ullmann's subgraph isomorphism algorithm.[24] Typically, it is used to check if a query substructure S can be found within the reference molecule M by finding a match between the atoms of S and atoms of M. In our application, the same algorithm is used to find all possible matches of atoms of query molecule M′ in reference M″. In this case, molecules M′ and M″ are the same chemical structure, M, but may have different locants assigned to the corresponding atoms. With application of the Ullman algorithm, we can identify the pairs of corresponding atoms in M′ and M″ and encode the resulting information in a form of a matching matrix **A**, i.e. the elements of the matching

dx.doi.org/10.1021/ci300289b | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$N_m \times N_m$ matrix take the value "1" if a match is possible between the corresponding pair of atoms and "0" otherwise (N_m is the number of atoms in M).

To begin enumeration of all possible congeners, the parent molecule M has to be provided together with information on the n hydrogen atoms that are to be substituted with any of p chemical group types. In practice, all to-be-substituted hydrogen atoms are marked as dummy atoms, X, and hydrogen is added as one of possible substituting groups. In addition, position numbers (locants) of X atoms have to be provided. The parent molecule is analyzed using the Ullman algorithm to obtain information on the symmetry of the parent molecule stored in a form of valid **A** matrices.

During the enumeration, congeners are represented as vectors, which hold information if a corresponding dummy atom is to be substituted with any of the specified chemical groups. The vectors are n positions long and hold integer numbers from 0 to $p - 1$ to represent any of the provided p substituting groups, respectively. For the provided pair of n and p, all possible vectors, the number of which is given by $(1 + p)^n$, are generated in a loop. Each generated vector **v** is validated as follows: (a) For every matrix **A**, generate variations of **v** that correspond to different but equivalent substitution patterns. (b) Identify the vector that corresponds to the correct International Union of Pure and Applied Chemistry (IUPAC) naming recommendations. (c) Check if the selected vector has been already saved in the library of possible congeners. If no duplicate is found, the vector is saved as a unique congener. Otherwise, the current vector is discarded and the program proceeds to generate another congener. To improve the execution time of step c, a simple modulo-based hashing procedure is employed to prescreen the generated vectors (keys). If a hash collision is detected, all affected keys are shortlisted and compared with the selected vector.

After all possible congener vectors are enumerated, we proceed to generation of structures. Each of substituting groups is provided in a form of Cartesian coordinates of atoms involved, plus a dummy atom, Y, marking the direction of chemical bond that will bind the group to the parent molecule. The coordinates of the dummy atoms Y, the to-be-substituted X atoms, and coordinates of adjacent atoms to Y and X in, respectively, the substituting group and the parent molecule are used to correctly position the group with respect to the parent molecule. The coordinates of atoms constituting the generated congeners are saved into separate files. The filenames include unique IDs corresponding to the position on a list of congeners sorted according to the convention used in studies of POPs. The described procedure has been implemented in version 2.0 of the ConGENER package.

Computational Characterization of Congeners and Generation of Molecular Descriptors. Currently, there are about 5 000 molecular descriptors available.[25] They can be derived from different theories and approaches and are, generally, grouped according to their "dimensionality" that reflects the complexity of molecule representation. The majority of them, however, rely on approximate description of a molecule and are incapable of capturing differences between congeners, which are highly similar in terms of both molecular and electronic structure. In our study, we intended to have possibly the most comprehensive description of congeners with use of reasonable computational resources. Thus, we selected a set of 26 numerical descriptors that consist of few simple structural descriptors and a series of quantum-mechanical descriptors calculated with the semiempirical PM6 method. The selected descriptors enabled the observation of slight differences in spatial and electronic properties between particular congeners as well as their composition.

As demonstrated in our previous study,[22] descriptors calculated at the semiempirical PM6 level are suitable for QSPR modeling for Cl- and Br-containing organic pollutants combining both high accuracy and relatively short calculation time allowing processing of large sets of molecules. The procedure of congener characterization was as follows. For each structure generated with ConGENER, we performed geometry optimization to minimize the gradient of its potential energy. The optimal geometry was then used in three single-point calculations to obtain molecular polarizability, characterization in water simulated by the conductor-like screening model, and a number of additional descriptors (Mulliken's electronegativity, Parr and Pople's absolute hardness, and Schuurmann MO shift alpha). All calculations were performed with the MOPAC 2007 package[26] using the increased criteria of precision (PRECISE keyword). Calculations were automated using a set of C shell, Perl, and AWK scripts responsible for generation of input files, submission of jobs, handling of typical error, and convergence problems as well as final extraction of results. Calculations were executed in parallel on dual Intel Xeon (quad core) servers.

A set of quantum chemical descriptors was supplemented with a number of constitutional descriptors to form the following final set of 26 descriptors: the number of hydrogen atoms (#H), number of chlorine atoms (#Cl), number of bromine atoms (#Br), total number of atoms in a molecule (nAT), heat of formation (HOF), electronic energy (EE), core repulsion energy (Core), total energy of a molecule (TE), energy of the highest occupied molecular orbital (HOMO), energy of the lowest unoccupied molecular orbital (LUMO), heat of formation in water (HOFc), total energy in water (TEc), solvent accessible surface (SAS), molecular volume (MV) molecular weight (MW), dipole moment (D), x, y, and z components of the dipole moments (D_x, D_y, D_z), the lowest negative Mulliken partial charge (q_min), the highest negative Mulliken partial charge (q_max), polarizability derived from the heat of formation (Ahof), polarizability derived from the dipole moment (Ad), Mulliken electronegativity (EN), Parr and Pople's absolute hardness (Hard), and Schurmann MO shift alpha (Shift). This pool of descriptors has relatively simple interpretability.

Cheminformatic Analysis of Congener Libraries. Molecular descriptors calculated for the library of congeners formed a multivariate and megaobjective data matrix (1 840 951 compounds × 26 descriptors). We have initiated the analysis of this data matrix with a data reduction step. We used the Kennard–Stone (KS) algorithm[27,28] to select a representative subset of compounds reflecting the original distribution of objects (compounds) in the variables (descriptors) space. The first element (named as s_1) selected by KS is always the closest object to the mathematical center (arithmetical mean) of the given data matrix **X** ($n \times m$). The second object assigned to the subset (s_2) is the most distant one from the previously selected s_1. The next object is the most distant one from both s_1 and s_2. Third is the most distant one from s_1 and s_2, as well as from s_3 and so on. The result of applying the Kennard–Stone algorithm is a selection of the central object, then outliers and farthest objects situated on the surface of the multidimensional cloud formed by all of the objects and then objects inside the cloud.

dx.doi.org/10.1021/ci300289b | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

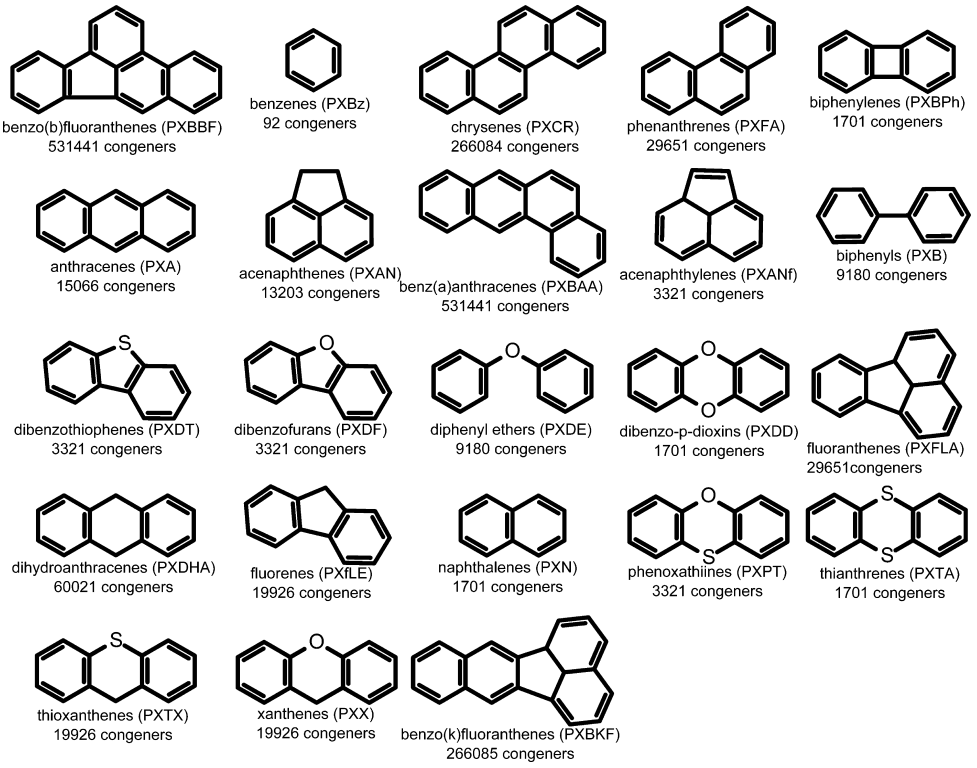


Figure 1. Chemical structures of organic molecules included in the study. Every hydrogen atoms may be substituted by Cl or Br atom in the congener generation procedure.

After selecting the representative set of compounds, the principal components analysis (PCA) was used to reduce the number of variables.[29] In result, the most important information extracted from the calculated descriptors was compressed in a much lower number of variables so-called principal components (PCs). All calculations were carried out in MATLAB 2008 environment. We employed a script file (m-file) decoding the KS algorithm designed by Daszykowski et al.[28] and PLS Toolbox (ver. 6.2) for MATLAB to perform the PCA analysis.

RESULTS

Congener Library Generation. C^3 methodology has been applied in analysis of 1 840 951 chlorinated, brominated, and mixed chlorinated/brominated congeners originating from 23 congener families of POPs, including the following: benzenes (PXBzs), acenaphthenes (PXANs), acenphthylenes (PXANfs), anthracenes (PXAs), biphenyls (PXBs), biphenylenes (PXBPhs), dibenzofurans (PXDFs), diphenyl ethers (PXDEs), dibenzo-*p*-dioxins (PXDDs), dibenzothiophenes (PXDTs), dihydroanthracenes (PXDHAs), fluorenes (PXFLEs), naphthalenes (PXNs), phenanthrenes (PXFAs), phenoxathiines (PXPTs), thianthrenes (PXTAs), thioxanthenes (PXTXs), xanthenes (PXXs), fluoranthenes (PXFLAs), chrysenes (PXCRs), benzo(k)fluoranthenes (PXBKFs), benzo(b)-fluoranthenes (PXBBFs), and benz(a)anthracenes (PXBAAs) (Figure 1).

All generated congeners underwent computational characterization according to the procedure presented in the Methods section. The calculations for all 1 840 951 congeners took roughly 8 months on three eight-core machines (ca. 140 000 CPU h total). Without doubts, there are no empirical methods which can be employed for obtaining data for such a number of compounds in a similar time of measurements.

Processing of the Congener Data Set. We reduced the dimensions of the obtained data matrix by using a combination of the Kennard−Stone algorithm, principal component analysis and, once again, Kennard−Stone algorithm (KS−PCA−KS). The reduction of dimensions and the selection of representative compounds in the initial KS−PCA steps were critical to account for scaling limitations of the computational algorithms, whereas the final KS step was crucial for sensible visual analysis of the results.

Initially, the KS algorithm was used to select approximately 2% from each of 23 families of compounds. This step resulted in 36 811 congeners. The most numerous families, namely, PXCRs, PXBKFs, PXBBFs, and PXBAAs were divided into 10 (PXCRs, PXBKFs) and 20 (PXBBFs, PXBAAs) smaller sets to make the calculations possible with using a standard personal computer (3.16 GHz Intel Core2Duo, 4GB RAM). The large matrices were split with use of the so-called 1:*X* algorithm,[15] in which the compounds are sorted according to one variable and, then, every *X*th compound was taken into a smaller set. Congeners in all matrices were sorted by the number of substituents first, then by the number of Cl and Br atoms, respectively. We have also verified that the sequence of chemical compounds in the largest matrices, and the criterion of splitting into smaller parts did not affect the choice of representatives by the KS algorithm. To do so, we employed the KS algorithm to select matrices from the matrix of whole family of compounds sorted according to the rising values of different parameters (number of Cl atoms, number of Br atoms) as well as the KS algorithm used for matrices selected by using 1:*X* algorithm[15] (respectively, 1:10 and 1:20). We

dx.doi.org/10.1021/ci300289b | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

concluded that the type of splitting does not affect the KS-based selection.

In the next stage, we were able to reduce the number of variables from 26 to 5 (without losing significant information on the structure of studied congeners) by performing PCA on the matrix of previously selected representative compounds (36 811 × 26). Each principal component (PC) with eigenvalue greater than 1 was considered as significant. Finally selected five principal components (being in fact linear combinations of the original 26 descriptors) explain together 82.9% (51.5% + 11.7% + 10.4% + 5.1% + 4.2%) of total variance in the analyzed data set.

Then, to clarify the graphical presentation of the congener distribution along the particular principal components, once again, we performed the KS-based selection. This time, by selecting 2% of compounds from the matrix of 36 811 (compounds) × 5 (PCs), we obtained the final data set, which has dimensions of 184 (compounds) × 5 (PCs). In this way, the application of KS−PCA−KS approach allowed retaining significant information on the internal structure of the cloud representing all 1 840 951 congeners in only five dimensions by using only about 0.1% of the studied compounds, selected as the most representative ones.

Characterization of the Congener Data Set. As mentioned, principal components are mathematically linear combinations of the original variables (in this study: molecular descriptors). Assigning physical interpretation to a particular component is possible by analyzing the combination's coefficients (called loading values). According to Malinowski's rule,[30] important loadings should be considered only those with absolute values higher than 0.7 (Figure 2). Clear physical meaning of PCs helps to understand sources of similarity/dissimilarity occurring between particular congeners. This similarity can be observed in the score plot (Figure 3) in which objects representing similar congeners or groups of congeners are located close to each other.

The first principle component PC1 (Figure 2a) contains information from almost all calculated descriptors. PC1 is a linear combination of molecular weight (MW), number of H atoms (#H), electronic energy (EE), total molecule energy (TE), solvent accessible surface (SAS), volume available for the solvent (MV), the potential energy and polarizability derived from heat of formation and dipole moment, respectively (Ahof, Ad), electron affinity (negative LUMO value), etc. Unsubstituted congeners have the lowest values of PC1 (Figure 3a−d). With increasing values of PC1, the number of substituents in the molecules is also increasing. The second principal component (PC2) separates the congeners substituted by different halogen atoms. Highly chlorinated compounds from all congeneric families are characterized by the largest values of PC2, because of the significant loadings of #Cl and high negative loading of HOF (Figure 2b). It should be noted that the heat of formation is a descriptor closely related to the total number of chemical bonds and their enthalpies. By comparing differences between the enthalpies of the following three bonds C_6H_5−H (H = 112.9 kcal/mol), C_6H_5−Cl (H = 97.1 kcal/mol), and C_6H_5−Br (H = 84 kcal/mol)[31] it becomes clear why HOF is a significant descriptor in PC2 and it is able to emphasize a distinction between chlorinated and brominated congeners. The second principal component separates chlorinated and brominated congeners and places mixed (Br/Cl) ones mostly between those two groups (Figure 3a). The only descriptor that brings significant information to PC3 is the

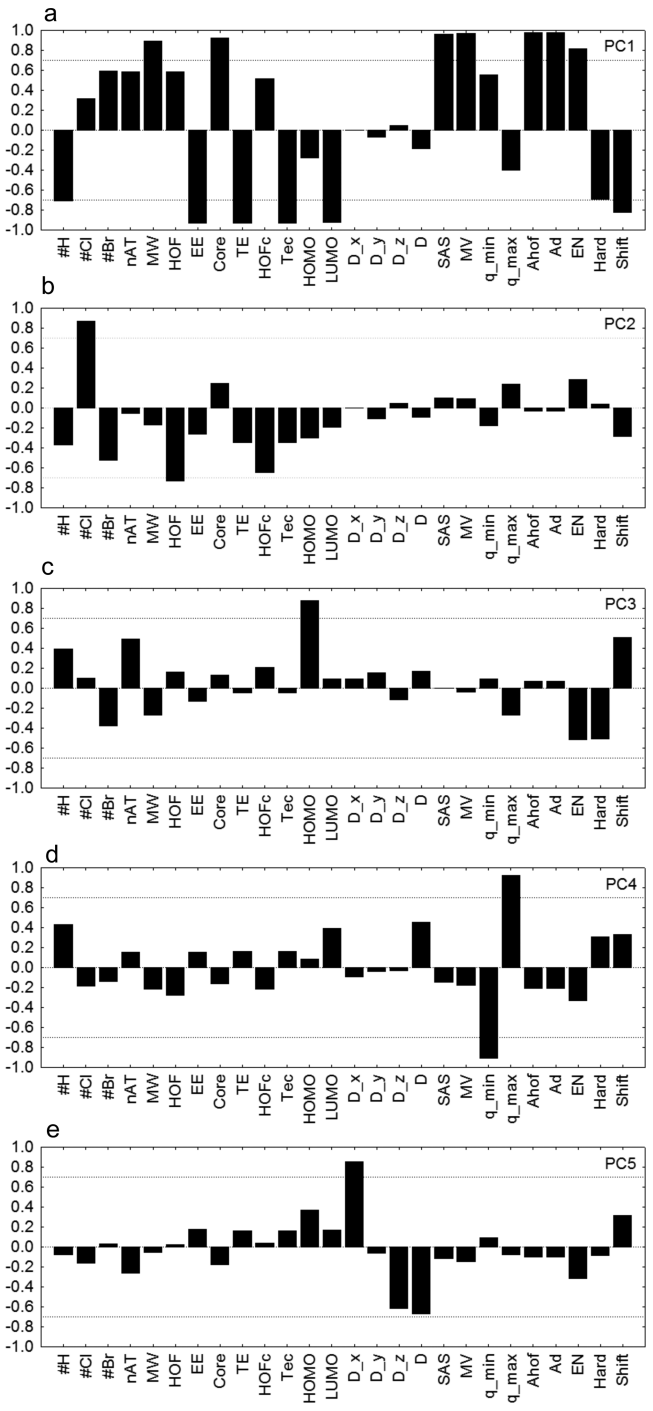


Figure 2. (a−e) Loading values of calculated descriptors into the first five PCs (PC1−PC5). According to Malinowski's rule, loadings with absolute values >0.7 are significant.

energy of the highest occupied molecular orbital (HOMO). According to the Koopman's theorem, a negative value of HOMO can be used as the first approximation of the ionization potential. That explains why the highest PC3 values are observed for congeners that contain the largest numbers of atoms. Smaller and less Cl-substituted congeners are placed at lower values of PC3 (Figure 3b). The main contribution to PC4 (Figure 3d) has the lowest negative Mulliken partial charge (q_min), and the highest negative Mulliken partial charge (q_max) (Figure 2d) gave us information about the electronic structure of the molecule and some overall

dx.doi.org/10.1021/ci300289b | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

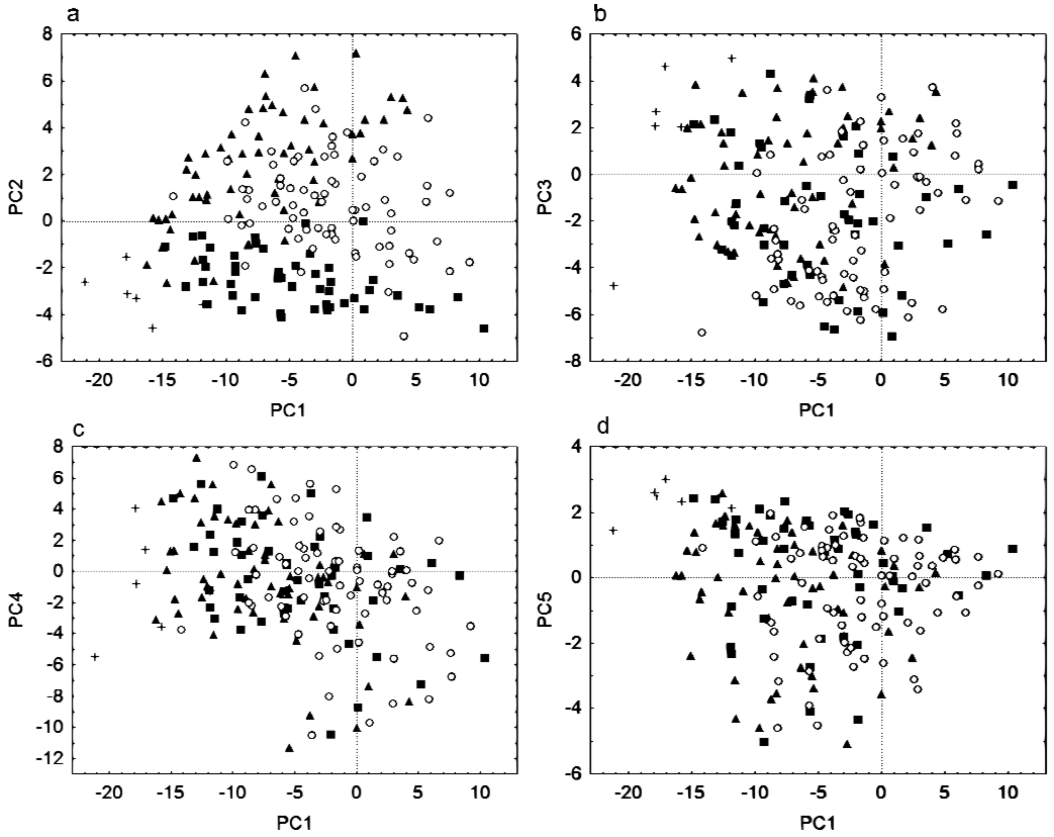


Figure 3. (a–d) Dependencies between principal components for unsubstituted (+), brominated (■), chlorinated (▲), and brominated/chlorinated (○) congeners.

characteristics of the polarity and the partial charge distribution in the molecule, which depends on both the structure of a carbon skeleton and the location of Cl/Br substituents. Because of that, highly halogenated congeners with regular aromatic rings in the structures (chrysenes, anthracenes, naphthalenes, with no oxygen or sulfur atoms in the ring) can be found as a separate group with the lowest PC4 values (Figure 3c). The last selected principal component (PC5) diversifies the congeners using the dipole moment (D) and its distribution along the *x* and *z* axes of the molecule (D_x and D_z). D_z is positively correlated with PC5, whereas D and D_x are negatively correlated (Figure 2e). As a consequence, congeners with zero dipole moment, for instance unsubstituted aromatic molecules, and congeners with very small dipole moments directed parallel to the aromatic skeleton are characterized by high values of PC5 (Figure 3d). On the contrary, congeners with large perpendicular dipole moments exhibit low values of PC5. In the later case, halogen substituents protrude above the carbon skeleton that make the molecule less planar than the other congeners.

The most interesting observation is that the mixed Br/Cl-substituted congeners from every particular family are located always between the regions occupied by purely chlorinated and brominated ones (Figure 3a–d). This has very practical implementation, regarding the applicability of QSPR models currently developed for POPs. Eight models that predict *n*-octanol/water and *n*-octanol/air partition coefficients,[32] water solubility,[17] and vapor pressure,[19] as well as the environmental persistence in air, water, soil, and sediments[16] have been originally calibrated and validated for Cl- and Br-substituted congeners, not mixed Br/Cl-substituted. By demonstrating that the values of structural descriptors for Br/Cl-substituted congeners are located between the corresponding values of their Br- and Cl-substituted counterparts, we have confirmed that the eight models can be also appropriately applied to make predictions for mixed Br/Cl-substituted congeners. As the predicted property is a function of descriptors (which is, in fact, the main assumption of QSPR methodology), the predictions for the mixed congeners should be regarded as interpolations, not extrapolations. Thus, we demonstrate that the applicability domain of the models is much wider than expected. For instance, there are 75 congeners of polychlorinated dibenzo-*p*-dioxins and 75 congeners of polybrominated dibenzo-*p*-dioxins, which gives 150 congeners in total. Taking into account that, additionally, there are 1 700 possible mixed Br/Cl-substituted structures, the applicability domain of the models can be extended from 150 to 1 850 compounds. Thus, the models can be utilized for predicting properties for about 12 times more compounds than previously.

■ CONCLUSIONS

We presented the Combinatorial × Computational × Cheminformatics (C^3) approach, which allows systematic enumeration of congeners and their later characterization using electronic structure methods. We demonstrated the high usefulness of this approach by studying a megaobjective set containing structural data for 1 840 951 compounds, including Cl-, Br-, and mixed Br/Cl-substituted congeners of 23 organic molecules. By a systematic selection of 184 representative compounds, we were able to visualize the internal structure of the whole megaobjective data set of the studied compounds. Structural similarity/dissimilarity analysis of the studied

dx.doi.org/10.1021/ci300289b | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

congeners provided scientifically justified evidence that eight QSPR models, originally developed and validated for chlorinated and brominated POPs can be further applied to predict properties for mixed Br/Cl-substituted congeners, which significantly extends the applicability domains of the models.

An important fact is that by using the C^3 approach we have created one of the world's most extensive database with information about brominated, chlorinated, and mixed Br/Cl-substituted congeners of POPs (1 840 951 533 congeners).

AUTHOR INFORMATION

Corresponding Author

*E-mail: mharanczyk@lbl.gov. Fax. +1 510 486 5812.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank Dr. James Stewart for providing a MOPAC software license. We acknowledge the fruitful discussions with Noriyuki Suzuki. This research was supported in part (to M.H. and E.G.N.) by the U.S. Department of Energy under contract DE-AC02-05CH11231. This work was supported (to P.U. and T.P.) by the Polish Ministry of Science and Higher Education (Grant No. DS/530-8180-D202-12) and the Foundation for Polish Science (Grant No. FOCUS 559-8430-1251-0). This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

(1) Lerche, D.; van de Plassche, E.; Schwegler, A.; Balk, F. Selecting chemical substances for the UN-ECE POP protocol. *Chemosphere* **2002**, *47* (6), 617−630.

(2) Rotkin-Ellman, M.; Navarro, K. M.; Solomon, G. M. Gulf oil spill air quality monitoring: lessons learned to improve emergency response. *Environ. Sci. Technol.* **2010**, *44* (22), 8365−8366.

(3) Muir, D. C.; Howard, P. H. Are there other persistent organic pollutants? A challenge for environmental chemists. *Environ. Sci. Technol.* **2006**, *40* (23), 7157−7166.

(4) United Nations Environment Programme. *Stockholm Convention on Persistent Organic Pollutants*; UNEP: Stockholm, 2009.

(5) United Nations Economic Commission for Europe. *Protocol to the 1979 Convention on Long-Range Transboundary Air Pollution on Persistent Organic Pollutants*; UN-ECE: Aarhus, 1998.

(6) United Nations Environment Programme. *Draft risk profile: pentabromodiphenyl ether*; UNEP: Geneva, 2006.

(7) Luthe, G. M.; Schut, B. G.; Aaseng, J. E. Monofluorinated analogues of polychlorinated biphenyls (F-PCBs): synthesis using the Suzuki-coupling, characterization, specific properties and intended use. *Chemosphere* **2009**, *77* (9), 1242−1248.

(8) Ohta, S.; Tokusawa, H.; Nakao, T.; Aozasa, O.; Miyata, H.; Alaee, M. Global contamination of coplanar polybrominated/chlorinated biphenyls (Co-PXBs) in the market fishes from Japan. *Chemosphere* **2008**, *73* (1 Suppl), S31−38.

(9) Olsman, H.; Engwall, M.; Kammann, U.; Klempt, M.; Otte, J.; Bavel, B.; Hollert, H. Relative differences in aryl hydrocarbon receptor-mediated response for 18 polybrominated and mixed halogenated dibenzo-p-dioxins and -furans in cell lines from four different species. *Environ. Toxicol. Chem.* **2007**, *26* (11), 2448−2454.

(10) Steen, P. O.; Grandbois, M.; McNeill, K.; Arnold, W. A. Photochemical formation of halogenated dioxins from hydroxylated polybrominated diphenyl ethers (OH-PBDEs) and chlorinated derivatives (OH-PBCDEs). *Environ. Sci. Technol.* **2009**, *43* (12), 4405−4411.

(11) Wan, Y.; Wiseman, S.; Chang, H.; Zhang, X.; Jones, P. D.; Hecker, M.; Kannan, K.; Tanabe, S.; Hu, J.; Lam, M. H.; Giesy, J. P. Origin of hydroxylated brominated diphenyl ethers: natural compounds or man-made flame retardants? *Environ. Sci. Technol.* **2009**, *43* (19), 7536−7542.

(12) Weijs, L.; Das, K.; Siebert, U.; van Elk, N.; Jauniaux, T.; Neels, H.; Blust, R.; Covaci, A. Concentrations of chlorinated and brominated contaminants and their metabolites in serum of harbour seals and harbour porpoises. *Environ. Int.* **2009**, *35* (6), 842−850.

(13) Haranczyk, M.; Puzyn, T.; Ng, E. G. On enumeration of congeners of common persistent organic pollutants. *Environ. Pollut.* **2010**, *158* (8), 2786−2789.

(14) Schultz, T. W.; Cronin, M. T. D.; Walker, J. D.; Aptula, A. O. Quantitative structure−activity relationships (QSARs) in toxicology: a historical perspective. *J. Mol. Struct.* **2003**, *622* (1−2), 1−22.

(15) Mostrag, A.; Puzyn, T.; Haranczyk, M. Modeling the overall persistence and environmental mobility of sulfur-containing polychlorinated organic compounds. *Environ. Sci. Pollut. Res. Int.* **2010**, *17* (2), 470−477.

(16) Puzyn, T.; Haranczyk, M.; Suzuki, N.; Sakurai, T. Estimating persistence of brominated and chlorinated organic pollutants in air, water, soil, and sediments with the QSPR-based classification scheme. *Mol. Diversity* **2011**, *15* (1), 173−188.

(17) Puzyn, T.; Gajewicz, A.; Rybacka, A.; Haranczyk, M. Global versus local QSPR models for persistent organic pollutants: balancing between predictivity and economy. *Struct. Chem.* **2001**, *22*, 873−884.

(18) Puzyn, T.; Mostrag, A.; Falandysz, J.; Kholod, Y.; Leszczynski, J. Predicting water solubility of congeners: chloronaphthalenes–a case study. *J. Hazard. Mater.* **2009**, *170* (2−3), 1014−1022.

(19) Gajewicz, A.; Haranczyk, M.; Puzyn, T. Predicting logarithmic values of the subcooled liquid vapor pressure of halogenated persistent organic pollutants with QSPR: How different are chlorinated and brominated congeners? *Atmos. Environ.* **2010**, *44* (11), 1428−1436.

(20) Haranczyk, M.; Puzyn, T.; Sadowski, P. ConGENER - A Tool for Modeling of the Congeneric Sets of Environmental Pollutants. *QSAR Comb. Sci.* **2008**, *27*, 826−833.

(21) Haranczyk, M.; Gutowski, M. Quantum mechanical energy-based screening of combinatorially generated library of tautomers. TauTGen: a tautomer generator program. *J. Chem. Inf. Model.* **2007**, *47* (2), 686−694.

(22) Puzyn, T.; Suzuki, N.; Haranczyk, M.; Rak, J. Calculation of quantum-mechanical descriptors for QSPR at the DFT level: is it necessary? *J. Chem. Inf. Model.* **2008**, *48* (6), 1174−1180.

(23) Haranczyk, M.; Gutowski, M. Combinatorial-computational-chemoinformatics (C3) approach to finding and analyzing low-energy tautomers. *J. Comput.-Aided Mol. Des.* **2010**, *24* (6−7), 627−638.

(24) Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *J. Assoc. Comp. Machinery* **1976**, *23*, 31−42.

(25) Consonni, V.; Todeschini, R. Molecular Descriptors. In *Recent advances in QSAR studies: methods and applications*; Puzyn, T., Leszczynski, J., Cronin, M. T. D., Eds.; Springer: Dordrecht, NY, 2010; Vol. *8*, p 29−102.

(26) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.

(27) Kennard, R. W.; Stone, L. A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137−148.

(28) Daszykowski, M.; Walczak, B.; Massart, D. L. Representative subset selection. *Anal. Chim. Acta* **2002**, *468*, 91−103.

(29) Abdi, H.; Williams, L. J. Principal component analysis. *WIREs: Comp. Stat.* **2010**, *2* (4), 433−459.

(30) Malinowski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*: John Wiley & Sons: New York, 1980.

(31) Blanksby, S. J.; Ellison, G. B. Bond dissociation energies of organic molecules. *Acc. Chem. Res.* **2003**, *36* (4), 255−263.

(32) Puzyn, T.; Suzuki, N.; Haranczyk, M. How do the partitioning properties of polyhalogenated POPs change when chlorine is replaced with bromine? *Environ. Sci. Technol.* **2008**, *42* (14), 5189−5195.