JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

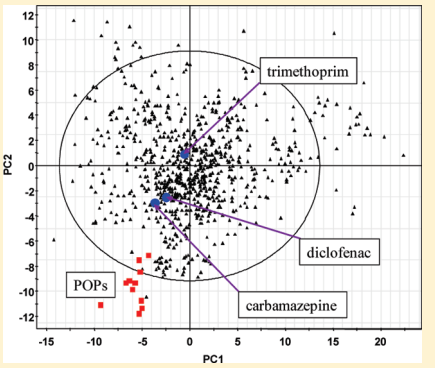
A Multivariate Chemical Similarity Approach to Search for Drugs of Potential Environmental Concern

Patrik L. Andersson,* Jerker Fick, and Stefan Rännar

Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden

Supporting Information

ABSTRACT: A structural similarity tool was developed and aimed to search for environmentally persistent drugs. The basis for the tool was a selection of so-called anchor molecules and a multidimensional chemical map of drugs. The map was constructed using principal component analysis covering 899 drugs described by 67 diverse calculated chemical descriptors. The anchor molecules (diclofenac, trimethoprim, and carbamazepine) were selected to represent drugs of known environmental concern. In addition 12 chemicals listed by the Stockholm Convention on persistent organic pollutants were used representing typical environmental pollutants. Chemical similarity was quantified by measuring relative Euclidean distances in the five-dimensional chemical map, and more than 100 nearest neighbors (kNNs) were found within a relative distance of less than 10% from each drug anchor. The developed chemical similarity approach not only identified persistent or semipersistent drugs but also a large number of potentially persistent drugs lacking environmental fate data.



INTRODUCTION

Drugs for human use can reach the environment via excretion in urine or feces, inappropriate disposal routines, or direct release from production plants. Large volumes of drugs reach sewage treatment plants and some, such as trimethoprim and carbamazepine, are essentially unaffected by sewage treatment.[1] To date, more than 150 drugs have been found in the environment.[2,3] Investigations of the environmental impact of these substances have only recently started, but it has long been known that synthetic estrogens can have endocrine disruptive effects, and other environmental effects have been discussed including the development of antibiotic resistance. The number of active pharmaceutical ingredients in use globally is very large. In Sweden alone, about 1200 individual compounds are registered, and analyzing the environmental fate and risk for all of these substances would be virtually impossible. Hence, strategic selection strategies are warranted to enhance the risk assessment process of these emerging potential pollutants.

A common strategy for identifying potential environmental pollutants from inventories of commercial chemicals is to apply various filter approaches.[4,5] These filters may include limit values for physicochemical characteristics, such as the octanol−water partition coefficient, vapor pressure, Henry's law constant, and water solubility. Other measures include predicted environmental fate characteristics, e.g., biodegradability, bioconcentration, and measures of long-range transport potential. An approach suggested by Huggett et al.[6] to search for environmentally problematic drugs is to use the ratio between known therapeutic plasma concentrations and predicted steady-state plasma concentrations. Ranking and prioritization efforts of drugs have also been undertaken using combinations of in silico tools[7] or stepwise classification using, e.g., exposure and mechanism of action data.[8] An alternative approach that has been used for screening substance databases is multivariate chemical characterization of the substances using calculated molecular descriptors followed by various selections or sorting procedures. This approach has been applied in both environmental and drug research to search for molecules with similar structural features to molecules with known characteristics.[9−11] These methodo logies can basically be divided into similarity and dissimilarity approaches. Similarity approaches aim to identify structurally similar compounds that are assumed to have similar physicochemical or biological properties to well characterized molecule(s), which is useful in applications, such as searches for new candidate drugs or environmental pollutants.[5,12,13] Dissimilarity approaches are used to select training sets for the development of in silico models, such as quantitative structure−activity relationship models,[11,14−16] and to search for potential drugs with similar activity, but different molecular structures, among chemicals in large domains.[17]

The chemical similarity search tool presented here was developed to facilitate the identification of drugs that potentially are persistent in the environment. Essentially, it involves multivariate chemical characterization using a diverse set of calculated chemical descriptors followed by principal component analysis (PCA). The multidimensional chemical domain of the drugs is mapped, and the *k*-nearest neighbors (kNN) of individual compounds and/or groups of compounds are identified based on Euclidean distances. Key components in the approach are the

Received: March 3, 2011
Published: July 17, 2011

so-called anchor molecules. These are selected to represent drugs with distinct, known environmental fate-related properties. Here we have used as anchor drugs trimethoprim, diclofenac, and carbamazepine. In addition 12 chemicals defined as persistent organic pollutants (POPs) by the Stockholm Convention (http://chm.pops.int/) were used as anchors representing internationally recognized persistent chemicals. The approach is developed on a data set of 899 drugs used in Sweden but represents the most widely used drugs worldwide. The hypothesis addressed was that the developed methodology can identify environmentally persistent drugs with discrete underlying structure-fate properties.

MATERIAL AND METHODS

Drug Database. A list of more than 900 drugs was compiled from http://www.fass.se/LIF/home/index.jsp (2005) that was later merged with data from sales statistics (2007) provided by LIF, the research-based pharmaceutical industry in Sweden, (http://www.lif.se/) including 7113 entries. This merge resulted in a database of 1400 pharmaceuticals that were included in the first phase of the study. These represent the most widely used pharmaceuticals worldwide. Proteins, polymers, vitamins, and minerals were excluded following the guidelines of the European Medicines Agency.[18] All covered compounds were treated in their neutral form, and counterions of salts were replaced by hydrogen. All chemical structures (SMILES) were checked using a structural formula register, and the molecules included in the database were restricted to those with molecular weights between 100 and 1600 Da and to compounds with Chemical Abstracts Service (CAS) registry numbers and anatomical therapeutic chemical (ATC) codes. In total, the database includes 899 organic substances.

Anchor Molecules. The nonsteroidal anti-inflammatory drug diclofenac (CAS 15307-86-5), the antibiotic trimethoprim (CAS 738-70-5), the anticonvulsant carbamazepine (CAS 298-46-4), and 12 chemicals listed by the Stockholm Convention on persistent organic pollutants as POPs[19] were used as anchor molecules and representatives of chemicals with distinct environmental fate properties. The POPs included representative pesticides, industrial chemicals, and byproducts, some of which—toxaphene, polychlorinated biphenyls (PCBs), and polychlorinated dibenzo-*p*-dioxins/polychlorinated dibenzofurans (PCDD/PCDF)—are classes of chemicals encompassing a large number of congeners. For this analysis 22′44′55′-hexachlorobiphenyl (PCB153, CAS 35065-27-1), 2378-tetrachlorodibenzo-*p*-dioxin (2378-TCDD, CAS 1746-01-6), 2378-tetrachlorodibenzofuran (2378-TCDF, CAS 51207-31-9), and toxaphene (2,2,5-*endo*,6-*exo*, 8,8,9,10-octachlorobornane, CAS 58002-18-9) were selected as representatives of their respective classes. The other POPs were *p*,*p*′-dichlorodiphenyltrichloroethane (DDT, CAS 50-29-3), hexachlorobenzene (HxCBz, CAS 118-74-1), aldrin (CAS 309-00-2), mirex (CAS 2385-85-5), chlordane (CAS 57-74-9), dieldrin (CAS 60-57-1), endrin (CAS 72-20-8), and heptachlor (CAS 76-44-8). In the chemical similarity analysis the POPs were treated as a group to illustrate how the chemical characteristics of these substances relate to the chemical space covered by the drugs.

Chemical Descriptors. The substances were characterized using 67 chemical descriptors calculated in MOE (chemcomp.com) from their structures (represented by SMILES codes). These descriptors were selected for their chemical relevance, interpretability, and (hence) utility for classifying the major chemical variation within and among sets of compounds. A complete list with brief explanations can be found in Table S1, Supporting Information. The descriptor set includes the logarithm of the octanol−water partition coefficient (log K_{ow}), molecular polarizability, and van der Waals volume in combination with selected flexibility, shape, and connectivity indices. Molecular surface characteristics were represented by 16 "partial equalization or orbital electronegativity" (PEOE) descriptors derived in MOE. Counts of selected atom types and single and aromatic bonds were also used, together with some count ratios.

Chemical descriptors were logarithmically transformed (if not already log transformed) if the calculated skewness exceeded 2 prior to the analysis, to improve the normality of the descriptors' distributions and minimize the impact of extreme values. Descriptors reflecting environmentally relevant properties, e.g., biodegradability, atmospheric oxidation half-life, and bioconcentration were calculated using the BIOWIN, AOPWIN, and BCFWIN modules included in the EPISuite (http://www.epa.gov/opptintr/exposure/pubs/episuite.htm).

Principal Component Analysis. In order to obtain an overview of large data matrices in which many objects, in our case molecules, are characterized by many different measurements, one needs to use an appropriate visualization tool. One such tool is PCA, a latent vector-based method that compresses data into a few orthogonal vectors summarizing the variation and the correlation patterns in the data. PCA is a very powerful tool, and typically 2 to 5 principal components (PCs) are sufficient to represent the structured information in a multivariate table. Since each PC consists of two vectors, one cannot only look for similarities among objects but also interpret the correlation patterns among the descriptors. The score vector visualizes the similarities among the substances (or observations), and the loading vector shows the correlation pattern among the descriptors (variables). In this study we have used the software SIMCA-P+ v12 for the multivariate analysis (http://umetrics.com/).

Similarity Measurement. As a measure of structural similarity between compounds, we used the Euclidean distance in the latent variables from each anchor molecule. However, in an effort to make the measurements more illustrative, instead of using absolute distances we converted them to relative distances (summed distances between pairs of objects divided by the range spanned by each principal component of the PCA model in the five-dimensional chemical space). The formula used to calculate the distance corresponding to a 5% relative distance is shown in eq 1. In our case, the Euclidean distance corresponding to 5% of the principal component range was 2.63:

$$distance_{5\%} = \sqrt{\sum_{a=1}^{A} (0.05R_a)^2} \quad (1)$$

where $distance_{5\%}$ is the limit corresponding to 5% of the model range, A is total number of principal components (in our case $A = 5$), and R_a is the range for principal component a.

RESULTS AND DISCUSSION

Chemical Map of Drugs. The chemical variation of the 899 drugs was analyzed using PCA based on 67 chemical descriptors. Five PCs were calculated that condense 82% of the variation in the descriptor set. The first two PCs are shown in Figure 1. The first component is heavily influenced by chemical descriptors,

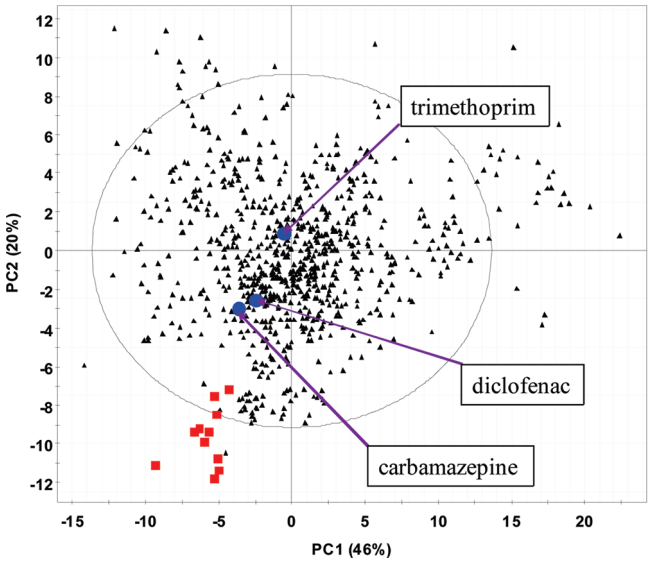


Figure 1. PCA score plot of the 899 studied drugs: the first PC (PC1) versus the second (PC2), explaining 46 and 20% of the variation in the 68 chemical descriptors, respectively. The 3 selected anchor pharmaceuticals and 12 compounds listed as persistent organic pollutants by the UNEP Stockholm Convention are marked as blue circles and red squares, respectively.

such as molecular weight, van der Waals volume and area, and other measures of molecular size (Figure S1, Supporting Information). The second is related to the chemicals' hydrophobicity (log K_{ow}) and surface-describing descriptors, such as fractions of hydrophobic and polar van der Waals surface areas. Large chemicals have positive score values in the first PC, relatively polar chemicals have high scores in PC2, and as shown in Figure 1, a few of the drugs have extreme values in these dimensions. Examples of large molecules with high PC1 values include ganirelix (CAS 124904-93-4), a gonadotropin-releasing hormone antagonist, with a molecular weight of 1570, and a number of drugs used for treating malignant neoplastic diseases, such as bleomycin (CAS 9060-10-0). The compound with the lowest PC1 value is halothane (CAS 151-67-7), a halogenated ethane with a molecular weight of just 197, used as an anesthetic agent, substituted with three fluorine atoms, one chlorine, and one bromine. The compound with the lowest PC2 value is mitotane (CAS 53-19-0), and an example of a very polar drug with a high PC2 value is calcium glubionate (CAS 12569-38-9). The third dimension of the chemical map is influenced by aromaticity, as described by descriptors such as relative numbers of aromatic atoms and bonds. The fourth dimension is related to the number of halogens (e.g., chloro- and fluoro-substituted anesthetics) and the fifth dimension to bond types, such as number of rotatable, single, and double bonds (Figures S2−S5, Supporting Information). Examples of drugs with relatively high numbers of aromatic bonds are the immunosuppressive drug mercaptopurine (CAS 50-44-2) and allopurinol (CAS 315-30-0), which is primarily used to treat hyperuricemia. Drugs with large numbers of halogens are the anesthetics isoflurane (CAS 26675-46-7) and enflurane (CAS 13838-16-9).

Notably, 26% of the drugs in the database violate one or more of Lipinski's rule of 5, i.e., have more than 5 hydrogen-bond donors and 10 acceptors, molecular weights greater than 500, and calculated log K_{ow} values greater than 5.[20] The drugs show a large overlap in physicochemical properties with widely used industrial chemicals, more specifically the 6400 European high and low production volume chemicals (H/LPVC) recently analyzed by Rännar and Andersson.[16] For example 37% of the drugs and 42% of the H/LPVCs have log K_{ow} values between 3 and 8, and the median log K_{ow} values for the drugs and H/LPVC are very similar (2.3 and 2.5, respectively). A slightly higher proportion of the drugs is halogenated (26% versus 23% of the H/LPVC), but the median numbers of hydrogen-bond donors and acceptors are higher for drugs (two donors and four acceptors) than for the bulk industrial chemicals (no donors and two acceptors). Unexpectedly, a lower proportion (15%) of the HPVC chemicals (1340 in total) violate Lipinski's rule of five.

Similarity Search Tool. Three pharmaceuticals (carbamazepine, diclofenac, and trimethoprim) were selected as anchor compounds to represent drugs with well characterized environmental persistency properties. In addition, 12 halogenated compounds, defined as POPs by the Stockholm Convention,[19] were included as reference compounds of internationally recognized persistent pollutants. The Euclidean distances from each anchor in the five-dimensional map of drugs were calculated and the 15 closest drugs (kNN 15) to each anchor were identified (Table 1). These drugs have a similar pattern of chemical descriptors as each anchor and should hypothetically display similar environmental fate characteristics.

Each anchor was found to form distinct kNN groups with very few overlaps between each group of 15 core kNN compounds. Except the POPs, each anchor has a clear-cut core but more than 100 drugs within a relative distance of 10%. Notably, carbamazepine was found at a relative distance of only 9.8% from the POP group and as a member of the POPs' 15 kNN set. The most structurally similar groups seem to be the carbamazepine and diclofenac clusters, for which four shared substances were recorded: benzoylperoxide (CAS 94-36-0), nitroscanate (CAS 19881-18-6), oxazepam (CAS 604-75-1), and pyrimethamine (CAS 58-14-0). Carbamazepine and diclofenac are also the most similar anchor molecules according to the two-dimensional plot of the chemical map of drugs, as shown in Figure 1, hence the high degree of overlap of their nearest-neighbor sets is consistent with expectations. Notably, the trimethoprim cluster shows no overlaps with any other anchors, whereas the kNN groups of carbamazepine and diclofenac have three and two shared substances with the POP group, respectively. Benzoylperoxide (CAS 94-36-0) is a member of the POP, carbamazepine, and diclofenac kNNs; benzylbenzoate (CAS 120-51-4) and diazepam (CAS 439-14-5) are members of the POP and carbamazepine kNNs; and tolfenamicacid (CAS 13710-19-5) is a member of the POP and diclofenac kNNs. Information is very limited on the environmental fate properties of the mentioned kNNs, but clearly benzoylperoxide is one example of a false positive chemical that is not persistent in the environment. The compound is a strong oxidizing agent and illustrates the challenges associated with predicting environmental persistence and calculating chemical descriptors reflecting this property. Both EpiSuite modules BIOWIN and AOPWIN indicate that this compound has rather high persistence, e.g., an atmospheric half-life (reactions with hydroxyl radicals) of more than three days and an estimated ultimate biodegradation of weeks to months. This weakness of the EpiSuite programs was recently discussed by Howard and Muir in 2010,[21] and they propose expert judgments to deal with easily oxidized or hydrolyzed chemicals. To obtain a crude estimate of the amount of knowledge available on the environmental

Table 1. CAS Registry Numbers, Substance Names, Relative Distances in the Chemical Map of Drugs to Selected Anchors, and Number of Literature Hits for Members of kNN Sets for Each of the Anchor Compounds

CAS[a]	substance name	POP[b,c]	carbamazepine[c]	diclofenac[c]	trimethoprim[c]	hits[d]
53-19-0	mitotane	2,9				0
94-36-0	benzoylperoxide	9,9	3,9	3,7		0
120-51-4	benzylbenzoate	7,8	5,0			0
130-26-7	clioquinol	7,7				0
298-46-4	carbamazepine[e]	9,8				107
439-14-5	diazepam	9,5	3,3			9
1744-22-5	riluzole	9,5				1
13710-19-5	tolfenamicacid	10		2,5		0
23593-75-1	clotrimazole	9,2				2
28911-01-5	triazolam	7,0				0
28981-97-7	alprazolam	7,9				0
59467-70-8	midazolam	7,8				0
75706-12-6	leflunomide	9,9		2,6		0
79617-96-2	sertraline	8,7				1
100643-71-8	desloratadine	9,6				1
28721-07-5	oxcarbazepine		3,4			1
57-41-0	phenytoin		3,5			3
58-00-4	apomorphine		3,9			0
22316-47-8	clobazam		4,0			0
3902-71-4	trioxysalen		4,0			0
58-25-3	chlordiazepoxide		4,5			0
19881-18-6	nitroscanate		4,6	5,7		0
604-75-1	oxazepam		4,7	5,8		3
435-97-2	phenprocoumon		4,8			0
59803-98-4	brimonidine		4,9			0
58-14-0	pyrimethamine		5,1	5,4		0
80012-43-7	epinastine		5,1			0
38677-85-9	flunixin			3,2		1
22071-15-4	ketoprofen			4,4		31
50-65-7	niclosamide			5,2		1
846-49-1	lorazepam			5,2		0
68693-11-8	modafinil			5,3		0
22494-42-4	diflunisal			5,8		1
89796-99-6	aceclofenac			6,2		0
53230-10-7	mefloquine			6,3		0
84057-84-1	lamotrigine			6,3		0
7683-36-5	obidoxime				4,3	0
73590-58-6	omeprazole				4,7	0
104227-87-4	famciclovir				5,0	0
136470-78-5	abacavir				5,2	0
117976-89-3	rabeprazole				5,8	2
13392-18-2	fenoterol				6,0	0
51481-61-9	cimetidine				6,1	1
73-31-4	melatonin				6,3	0
106941-25-7	adefovir				6,6	0
26839-75-8	timolol				6,6	1
103628-46-2	sumatriptan				6,7	0
76824-35-6	famotidine				6,8	0
15676-16-1	sulpiride				6,8	1
56211-40-6	torasemide				6,9	0
69655-05-6	didanosine				6,9	0

[a] Chemical Abstracts Service (CAS) registry number. [b] Persistent organic pollutant (POP), note the distance is calculated from the closest anchor molecule. [c] Relative distance (%) to anchor molecule. [d] Number of hits (accessed September 10, 2010) in Web of Science using search strings: "compound" and "environment" and "fate". [e] Compound is also an anchor molecule.

dx.doi.org/10.1021/ci200107b |J. Chem. Inf. Model. 2011, 51, 1788–1794

characteristics of the studied drugs, the Web of Science database (http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/; accessed September 10, 2010) was searched for literature references to the 51 identified kNN drugs using the search terms "compound", "environment", and "fate" (Table 1).

Anchor: Carbamazepine. Carbamazepine, an anticonvulsant and mood-stabilizing drug, is one of the most well studied pharmaceuticals in relation to its environmental properties. It has been identified in sewage and natural waters[22−25] and fish tissues,[26] and it is very resistance to degradation in sewage treatment works.[25,27−29] The compound has been detected in drinking water and is not fully removed by either ozone or chlorine treatment.[30] Carbamazepine shows moderate sorption to sediment[31] but only weak sorption to soil.[32,33]

One of the most thoroughly studied drugs among the kNNs of carbamazepine is the tranquilizer diazepam. In the study by Löffler et al.,[31] diazepam exhibited high persistence and oxazepam moderate persistence in a water−sediment system. Diazepam has also been shown to be moderately persistent during anaerobic sludge digestion[27] and in passage through a complete sewage treatment works.[34] Similarly, the waste water treatment removal efficiency of another member of the carbamazepine cluster, the anticonvulsant phenytoin (57-41-0), is reportedly less than 50%.[35] The entire cluster has calculated log K_{ow} values between 0.6 and 5.2, generally low calculated bioconcentration potential, all the compounds have phenyl rings, and many have amino groups. Among the carbamazepine kNNs, six drugs are used for treating nervous system-related indications, one of which, oxcarbazepine (CAS 28721-07-5), was recently detected in environmental samples for the first time, and removal rates between 24 and 73% in sewage treatment plants were found for the substance.[36] The cluster is very tight, and among the 15 kNNs, 12 were found within a relative distance less than 5% (Table 1).

Anchor: Diclofenac. Much attention has been paid to diclofenac in recent years due to the decline of the vulture populations in Pakistan and India,[37] which has been related to renal failure caused by diclofenac residues in livestock. Diclofenac has been found in sewage and surface waters[25,38] and is also resistant to degradation in sewage treatment works, where removal efficiencies are in the 20−40% range.[25] Diclofenac shows weak sorption to both sediment and soil.[32,39]

In general very little information on environmental characteristics of the diclofenac kNNs is available, except for ketoprofen (CAS 22071-15-4) and (to a lesser degree) oxazepam, which is also a member of the carbamazepine cluster (see above). Ketoprofen is found at a relative distance of 4.4% from diclofenac, it has been detected in sewage influent and effluent waters[38] and has shown to affect the reproductive system in fathead minnow.[40] All members of the cluster have low calculated bioconcentration factors, except nitroscanate (19881-18-6) which has a bioconcentration factor >2000. Two of the 15 kNNs are potentially recalcitrant in the environment, as estimated using BIOWIN: the fluorinated compounds flunixin (CAS 38677-85-9) and mefloquine (CAS 53230-10-7). All compounds in the cluster have aromatic moieties, and many are chloro- or fluoro-substituted. Few of the drugs in the group share ATC classifications, the largest group being anti-inflammatory drugs (as is diclofenac).

Anchor: Trimethoprim. Trimethoprim, an antibiotic that is mainly used in the treatment of urinary tract infections, has been shown to resist degradation in sewage treatment plants, with removal efficiencies in the 0−40% range.[1,28,29] In a study by Thomas et al.,[41] median levels of trimethoprim in effluent waters were found to be higher than influent levels in a Norwegian sewage treatment plant, and the highest levels found in effluents by these authors were in the μg range. The compound has also been detected in surface waters at ng/L levels[42] and in drinking water, although it is effectively removed by ozone or chlorine.[30]

Two members of the trimethoprim cluster, the antihistamines famotidine (CAS 76824-35-6) and cimetidine (CAS 51481-61-9), are only partly removed in sewage treatment plants.[28,43] Cimetidine has been found at μg/L levels in both influents and effluents and detected in surface waters.[43] None of the compounds in the trimethoprim kNN group are predicted to bioconcentrate, but almost half (7/15) have ultimate biodegradation estimates in the range of months. Four of the compounds are antivirals, and four are prescribed for acid-related disorders (e.g., gastric ulcers). The compounds have diverse molecular structures but share a few characters; none of the kNNs are halogenated, most have aromatic heterocyclic entities, such as pyridine, pyrimidine, or pyrrole rings, and ether groups are well represented. Five of the 15 kNNs and trimethoprim are aryl ethers.

Anchor: POPs. In comparison with the chemically very diverse drugs, the 12 studied POPs are relatively similar and homogeneous. These chemicals each have at least five chlorines, they are relatively hydrophobic, and all, except toxaphene, have phenyl rings. The POPs form a cluster with relatively low PC1 and PC2 values, hence they are positioned below and to the left of the major drug cluster in the PC1 versus PC2 plot (Figure 1). Using the five-component PCA model and the kNN-based methodology, the chemical homogeneity of the POPs was analyzed, and among the 12 POPs, 11 were found within a relative distance of 10% from at least one other POP substance. The exception was hexachlorobenzene, for which PCB 153 is the closest neighbor, at a relative distance of 11.5%. Since all the UNEP chemicals have similar environmental fate characteristics, it seems reasonable to assume that any compound within a relative distance of <10% from these anchor molecules would have some structural similarities to them. However, the overlap with the drugs is limited, although 15 substances are positioned within 10% relative distance to at least one of the POPs (Table 1). The most POP-like drug is mitotane (used in the treatment of adrenocortical carcinoma), which was found at a distance of less than 10% from three POPs.

Among the 15 drugs found close (<10% relative distance) to the POPs, 7 are routinely prescribed for indications related to the nervous system. Most of the compounds have high hydrophobicity, as indicated by calculated log K_{ow} values ranging from 2.2 to 6.3. Ten of the compounds are halogenated, and all have phenyl rings. The three compounds mitotane, clotrimazole (CAS 23593-75-1), and sertraline (CAS 79617-96-2) would qualify as bioaccumulative according to European chemicals legislation REACH,[44] i.e., having calculated bioconcentration factors above 2000. In addition, the antidepressant sertraline has recently been detected at high ng/g levels in livers of fish sampled in various rivers in the United States.[26] According to the literature survey, only carbamazepine and diazepam have been recorded more than twice (Table 1). In the study by Löffler et al.,[31] these compounds were defined as persistent in a water−sediment system, with relatively high sorption to sediment. The antifungal clotrimazole has been detected in natural waters and strongly sorbs to particulate matter.[45,46] For the other compounds in this

group, there is virtually no information on their environmental fate, and screening studies are warranted.

CONCUDING REMARKS

The selected anchor molecules, carbamazepine, diclofenac, and trimethoprim, are known to be persistent in sewage treatment plants and the environment. Our analysis showed that only carbamazepine has chemical features close to the traditional environmental pollutants, which indicates that classical filter approaches may not be appropriate to discover drugs of environmental concern. Identification of potentially persistent drugs is challenging since classical attributes of pollutants may not be relevant, known, or calculable by current methods. The novel kNN-based multivariate read-across methodology presented here has not only identified false positives (e.g., benzoylperoxide) but also true positives, such as ketoprofen, oxazepam, phenytoin, cimetidine, and diazepam, that are known be persistent or semipersistent in sewage treatment plants. We also identified a large number of drugs that show structural resemblance with drugs of environmental concern but that lack data on environmental fate and occurrence. Validation of presented methodology is warranted, and we suggest including identified drugs in future persistency testing and environmental monitoring programs.

ASSOCIATED CONTENT

Supporting Information. Table S1, information on the chemical descriptors, Figures S1-S5, and Score and loading plots. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: patrik.andersson@chem.umu.se. Telephone: +46907865266.

ACKNOWLEDGMENT

Anna Vesterlund is acknowledged for her efforts with the compilation of the database and calculation of chemical descriptors. The Swedish Research Council for Environment, Agricultural Sciences, and Spatial Planning (Formas) is acknowledged for funding the project.

REFERENCES

(1) Lindberg, R. H.; Olofsson, U.; Rendahl, P.; Johansson, M. I.; Tysklind, M.; Andersson, B. A. V. Behavior of fluoroquinolones and trimethoprim during mechanical, chemical, and active sludge treatment of sewage water and digestion of sludge. *Environ. Sci. Technol.* **2006**, *40* (3), 1042–1048.

(2) Kümmerer, K. The presence of pharmaceuticals in the environment due to human use - present knowledge and future challenges. *J. Environ. Manage.* **2009**, *90* (8), 2354–2366.

(3) Segura, P. A.; Francois, M.; Gagnon, C.; Sauve, S. Review of the occurrence of anti-infectives in contaminated wastewaters and natural and drinking waters. *Environ. Health Perspect.* **2009**, *117* (5), 675–684.

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

(5) Brown, T. N.; Wania, F. Screening chemicals for the potential to be persistent organic pollutants: A case study of Arctic contaminants. *Environ. Sci. Technol.* **2008**, *42* (14), 5202–5209.

(6) Huggett, D. B.; Cook, J. C.; Ericson, J. F.; Williams, R. T. A theoretical model for utilizing mammalian pharmacology and safety data to prioritize potential impacts of human pharmaceuticals to fish. *Hum. Ecol. Risk Assess.* **2003**, *9* (7), 1789–1799.

(7) Sanderson, H.; Johnson, D. J.; Reitsma, T.; Brain, R. A.; Wilson, C. J.; Solomon, K. R. Ranking and prioritization of environmental risks of pharmaceuticals in surface waters. *Regul. Toxicol. Pharmacol.* **2004**, *39* (2), 158–183.

(8) Besse, J. P.; Garric, J. Human pharmaceuticals in surface waters implementation of a prioritization methodology and application to the French situation. *Toxicol. Lett.* **2008**, *176* (2), 104–123.

(9) Andersson, P. M.; Lundstedt, T. Hierarchical experimental design exemplified by QSAR evaluation of a chemical library directed towards the melanocortin 4 receptor. *J. Chemom.* **2002**, *16* (8−10), 490–496.

(10) Linusson, A.; Gottfries, J.; Olsson, T.; Ornskov, E.; Folestad, S.; Norden, B.; Wold, S. Statistical molecular design, parallel synthesis, and biological evaluation of a library of thrombin inhibitors. *J. Med. Chem.* **2001**, *44* (21), 3424–3439.

(11) Stenberg, M.; Andersson, P. L. Selection of non-dioxin-like PCBs for *in vitro* testing on the basis of environmental abundance and molecular structure. *Chemosphere* **2008**, *71* (10), 1909–1915.

(12) Auer, J.; Bajorath, J. Distinguishing between bioactive and modeled compound conformations through mining of emerging chemical patterns. *J. Chem. Inf. Model.* **2008**, *48* (9), 1747–1753.

(13) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996.

(14) Harju, M.; Hamers, T.; Kamstra, J. H.; Sonneveld, E.; Boon, J. P.; Tysklind, M.; Andersson, P. L. Quantitative structure-activity relationship modeling on *in vitro* endocrine effects and metabolic stability involving 26 selected brominated flame retardants. *Environ. Toxicol. Chem.* **2007**, *26* (4), 816–826.

(15) Papa, E.; Fick, J.; Lindberg, R.; Johansson, M.; Gramatica, P.; Andersson, P. L. Multivariate chemical mapping of antibiotics and identification of structurally representative substances. *Environ. Sci. Technol.* **2007**, *41* (5), 1653–1661.

(16) Rännar, S.; Andersson, P. L. A novel approach using hierarchical clustering to select industrial chemicals for environmental impact assessment. *J. Chem. Inf. Model.* **2010**, *50* (1), 30–36.

(17) Taylor, R. Simulation analysis of experimental-design strategies for screening random compounds as potential new drugs and agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (1), 59–67.

(18) *EMEA 2006 Guideline on the Environmental Risk Assessment of Medicinal Products for Human Use*, EMEA/CHMP/SWP/4447/00; The European Agency for the Evaluation of Medicinal Products; London, U.K. Accessed June 1,2006.

(19) *UNEP. Final Act of the Conference of Plenipotentiaries on The Stockholm Convention On Persistent Organic Pollutants*; United Nations Environment Program: Geneva, Switzerland, 2001, pp 44.

(20) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23* (1−3), 3–25.

(21) Howard, P. H.; Muir, D. C. G. Identifying new persistent and bioaccumulative organics among chemicals in commerce. *Environ. Sci. Technol.* **2010**, *44* (7), 2277–2285.

(22) Coetsier, C. M.; Spinelli, S.; Lin, L.; Roig, B.; Touraud, E. Discharge of pharmaceutical products (PPs) through a conventional biological sewage treatment plant: MECs vs PECs? *Environ Int.* **2009**, *35* (5), 787–792.

(23) Ternes, T. A. Occurrence of drugs in German sewage treatment plants and rivers. *Water Res.* **1998**, *32* (11), 3245–3260.

(24) Vieno, N. M.; Tuhkanen, T.; Kronberg, L. Analysis of neutral and basic pharmaceuticals in sewage treatment plants and in recipient rivers using solid phase extraction and liquid chromatography-tandem mass spectrometry detection. *J. Chromatogr., A* **2006**, *1134* (1−2), 101–111.

(25) Zhang, Y. J.; Geissen, S. U.; Gal, C. Carbamazepine and diclofenac: Removal in wastewater treatment plants and occurrence in water bodies. *Chemosphere* **2008**, *73* (8), 1151–1161.

(26) Ramirez, A. J.; Brain, R. A.; Usenko, S.; Mottaleb, M. A.; O'Donnell, J. G.; Stahl, L. L.; Wathen, J. B.; Snyder, B. D.; Pitt, J. L.; Perez-Hurtado, P.; Dobbins, L. L.; Brooks, B. W.; Chambliss, C. K. Occurrence of pharmaceuticals and personal care products in fish: Results of a national pilot study in the United States. *Environ. Toxicol. Chem.* **2009**, *28* (12), 2587–2597.

(27) Carballa, M.; Omil, F.; Ternes, T.; Lema, J. M. Fate of pharmaceutical and personal care products (PPCPs) during anaerobic digestion of sewage sludge. *Water Res.* **2007**, *41* (10), 2139–2150.

(28) Radjenovic, J.; Petrovic, M.; Barcelo, D. Fate and distribution of pharmaceuticals in wastewater and sewage sludge of the conventional activated sludge (CAS) and advanced membrane bioreactor (MBR) treatment. *Water Res.* **2009**, *43* (3), 831–841.

(29) Suarez, S.; Lema, J. M.; Omil, S. Removal of pharmaceuticals and personal care products (PPCPs) under nitrifying and denitrifying conditions. *Water Res.* **2010**, *44*, 3214–3224.

(30) Benotti, M. J.; Trenholm, R. A.; Vanderford, B. J.; Holady, J. C.; Stanford, B. D.; Snyder, S. A. Pharmaceuticals and endocrine disrupting compounds in US drinking water. *Environ. Sci. Technol.* **2009**, *43* (3), 597–603.

(31) Löffler, D.; Rombke, J.; Meller, M.; Ternes, T. A. Environmental fate of pharmaceuticals in water/sediment systems. *Environ. Sci. Technol.* **2005**, *39* (14), 5209–5218.

(32) Drillia, P.; Stamatelatou, K.; Lyberatos, G. Fate and mobility of pharmaceuticals in solid matrices. *Chemosphere* **2005**, *60* (8), 1034–1044.

(33) Yu, L.; Fink, G.; Wintgens, T.; Melin, T.; Ternes, T. A. Sorption behavior of potential organic wastewater indicators with soils. *Water Res.* **2009**, *43* (4), 951–960.

(34) Kreuzinger, N.; Clara, M.; Strenn, B.; Vogel, B. Investigation on the behaviour of selected pharmaceuticals in the groundwater after infiltration of treated wastewater. *Water Sci. Technol.* **2004**, *50* (2), 221–228.

(35) Yu, J. T.; Bouwer, E. J.; Coelhan, M. Occurrence and biodegradability studies of selected pharmaceuticals and personal care products in sewage effluent. *Agric. Water Manage.* **2006**, *86* (1−2), 72–80.

(36) Leclercq, M.; Mathieu, O.; Gomez, E.; Casellas, C.; Fenet, H.; Hillaire-Buys, D. Presence and fate of carbamazepine, oxcarbazepine, and seven of their metabolites at wastewater treatment plants. *Arch. Environ. Contam. Toxicol.* **2009**, *56* (3), 408–415.

(37) Oaks, J. L.; Gilbert, M.; Virani, M. Z.; Watson, R. T.; Meteyer, C. U.; Rideout, B. A.; Shivaprasad, H. L.; Ahmed, S.; Chaudhry, M. J.; Arshad, M.; Mahmood, S.; Ali, A.; Khan, A. A. Diclofenac residues as the cause of vulture population decline in Pakistan. *Nature* **2004**, *427* (6975), 630–633.

(38) Vieno, N. M.; Tuhkanen, T.; Kronberg, L. Seasonal variation in the occurrence of pharmaceuticals in effluents from a sewage treatment plant and in the recipient water. *Environ. Sci. Technol.* **2005**, *39* (21), 8220–8226.

(39) Ternes, T. A.; Herrmann, N.; Bonerz, M.; Knacker, T.; Siegrist, H.; Joss, A. A rapid method to measure the solid-water distribution coefficient (Kd) for pharmaceuticals and musk fragrances in sewage sludge. *Water Res.* **2004**, *38* (19), 4075–4084.

(40) Ankley, G. T.; Jensen, K. M.; Kahl, M. D.; Makynen, E. A.; Blake, L. S.; Greene, K. J.; Johnson, R. D.; Villeneuve, D. L. Ketoconazole in the fathead minnow (*Pimephales promelas*): Reproductive toxicity and biological compensation. *Environ. Toxicol. Chem.* **2007**, *26* (6), 1214–1223.

(41) Thomas, K. V.; Dye, C.; Schlabach, M.; Langford, K. H. Source to sink tracking of selected human pharmaceuticals from two Oslo city hospitals and a wastewater treatment works. *J. Environ. Monit.* **2007**, *9* (12), 1410–1418.

(42) Segura, P. A.; Francois, M.; Gagnon, C.; Sauve, S. Review of the occurrence of anti-infectives in contaminated wastewaters and natural and drinking waters. *Environ. Health. Perspect.* **2009**, *117* (5), 675–684.

(43) Choi, K.; Kim, Y.; Park, J.; Park, C. K.; Kim, M.; Kim, H. S.; Kim, P. Seasonal variations of several pharmaceutical residues in surface water and sewage treatment plants of Han River, Korea. *Sci. Total Environ.* **2008**, *405* (1−3), 120–128.

(44) Regulation (EC) no. 1907/2006 of the European Parliament and of the Council of December 18, 2006 concerning the registration, evaluation, authorization 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. Official Journal of the European Communities L163/3.

(45) Lindberg, R. H.; Fick, J.; Tysklind, M. Screening of antimycotics in Swedish sewage treatment plants - Waters and sludge. *Water Res.* **2010**, *44* (2), 649–657.

(46) Peschka, M.; Roberts, P. H.; Knepper, T. P. Analysis, fate studies and monitoring of the antifungal agent clotrimazole in the aquatic environment. *Anal. Bioanal. Chem.* **2007**, *389* (3), 959–968.