Can Descriptors of the Electron Density Distribution Help To Distinguish Functional Groups?

Julien Burton,* Nathalie Meurice,[†] Laurence Leherte, and Daniel P. Vercauteren

Laboratoire de Physico-Chimie Informatique, Groupe de Chimie Physique, Théorique et Structurale, University of Namur (FUNDP), 61 rue de Bruxelles, B-5000 Namur, Belgium

Received April 29, 2008

Our study is aimed at understanding the characteristics of functional group descriptors based on peaks of the electronic density distribution $\rho_{(\vec{r})}$. The descriptors calculated are the $\rho_{(\vec{r})}$ value at peak location, volume, ellipticity, curvatures of $\rho_{(\vec{r})}$, and the peak-functional group distance. By the implementation of an automated and global process for large-scale calculation of the descriptors, we generated a statistically meaningful data set focusing on the association between peaks and 77 types of functional groups extracted from 62,936 organic molecules issued from the Cambridge Structural Database. Statistical analyses demonstrated that selected descriptors are capable of discriminating subtypes of functional groups. A projection in a principal component space coupled to a hierarchical clustering confirmed the suitability of the descriptors to provide an appropriate description of the functional groups. The results indicated that functional similarity or dissimilarity could be quantified based on electron density descriptors.

INTRODUCTION

In silico prediction of pharmacological properties is more and more considered as a necessary step in the early stages of the ligand design and lead identification processes.[1] For that purpose, theoretical studies rely on a description of the molecules, usually under the form of a list of measured or calculated physicochemical properties (also simply called descriptors). For example, molecular size, count of atoms, molecular electrostatic potential, molecular lipophilicity, and other physicochemical properties are well-established and extensively used as descriptors in quantitative structure–activity relationships (QSAR).[2] In addition, the description of molecules in terms of their structural features has emerged as a standard to guide fragment-based syntheses of druglike molecules. Reasoning in terms of global molecular properties has, however, often been proven to be valid for one particular family of compounds only.[3] Therefore, if one wants to embrace a broader chemical space with broader molecular diversity, an efficient strategy would be to define highly transferable descriptors that would ideally be associated with the most frequent chemical functional groups.

The electron density distribution function $\rho(\vec{r})$ is a property of key importance in molecular recognition phenomena. Accordingly, the work of Popelier, based on Bader's *Atoms in Molecules* theory,[4] proves that it is relevant to exploit 3D $\rho(\vec{r})$ data by taking advantage of its topology.[5−8] It has also been shown that topological analyses of $\rho(\vec{r})$ allow simplification of the 3D distribution of $\rho(\vec{r})$ in reduced representations made of critical points (CPs) without losing significant information. Bond CPs are, for example, very helpful to discard covalent from ionic bonds[9] as well as to detect hydrogen bonds from high resolution Electron Density Maps (EDMs) and further infer QSAR models.[10] Such reduced molecular representations have previously been developed and utilized in our Computational Chemistry Group (PCI) to study interactions between small organic compounds and macromolecules[11,12] as well as between macromolecules,[13,14] to superimpose pharmacological ligands,[15−18] and to model inorganic zeolite systems.[19]

In this context, our study focuses on the association between functional group and CPs of the electron density, particularly the peaks (*i.e.*, local maxima), and their properties. More specifically, the purpose of the study was to implement a totally automated process to generate a massive amount of functional group descriptors based on $\rho(\vec{r})$ peaks. Hence, the following strategy was developed. First, several procedures were implemented to automate the calculation of EDMs, to locate the peaks, and to associate them with functional groups. The entire process was then applied to a large set of organic small molecules from the Cambridge Structural Database (CSD). To evaluate the discriminating power of the proposed descriptors toward the functional groups, the results were analyzed with statistical tools such as principal component analysis (PCA) and hierarchical clustering (HC).

In the sequel, we first expose the methodology designed to calculate automatically the EDM, retrieve their peaks, and associate them with functional groups. Then, the automated procedures are applied to a large set of molecules, and the resulting data are analyzed and discussed. Discussions first focus on general qualitative trends and then are oriented toward quantitative comparison of various chemical groups. Finally, we describe how to classify those functional groups based on their $\rho(\vec{r})$ similarity.

* Corresponding author phone: +32 (0)81 72 54 62; fax: +32 (0)81 72 54 66; e-mail: Julien.burton@fundp.ac.be.

[†] Present address: Chemogenomics Laboratory, Translational Genomics Research Institute (TGen), 13208 E. Shea Blvd, Suite 110, Scottsdale, AZ 85259.

10.1021/ci800148z CCC: $40.75
Published on Web 10/03/2008

METHODOLOGY

Data Selection. Peak descriptors were calculated for a set of 62,936 molecules, considered as large enough to perform reliable statistics. Their 3D coordinates were extracted from CSD, September 2004 release.[20,21] To restrict the study to a pharmacological-relevant chemical space, the study was restricted to molecules containing C, H, O, N, P, S, F, Cl, Br, and I atoms only. Polymers and cocrystallized compounds were excluded. This large organic-like compound collection was built using the Conquest interface[22] from CSD.

Automation of Electron Density Map Calculation and Topological Analysis. In X-ray diffraction (XRD) structure solving, the electron density map (EDM) reconstructed through Fourier transform of measured diffraction intensities is compared to the one calculated on the basis of the hypothetic 3D coordinates of atoms and tabulated parameters such as the atomic scattering factors. The 3D coordinates are then affined to minimize the difference between the experimental EDM and the calculated EDM. Generally, a promolecular model is used for comparison with experiment, which means that $\rho(\vec{r})$ is built from scattering factors tabulated for independent atoms. Consequently, it is only necessary to consider the atomic coordinates $\vec{r}$ in the calculations without taking explicitly into account the nature of the chemical bonds.

The calculation of an EDM simulating the XRD structure solving process can be handled with a program like XTAL.[23] XTAL utilizes the following input parameters: the unit cell (UC) parameters, the nature of the constituting atoms, the space group, a selected crystallographic resolution, and, if needed, the anisotropic displacement parameter. With these parameters, XTAL simulates the reflections of the X-ray beams on the lattice planes at a selected level of resolution and computes the structure factors $F(\vec{h})$ of the molecule

$$F(\vec{h}) = \sum_{j=1}^{n_{\mathrm{at}}} f_j e^{-B_j\left(\frac{\sin\theta}{\lambda}\right)^2} e^{2\pi i \vec{h}.\vec{r}_j} \tag{1}$$

where f_j is the atomic scattering factor of atom j (tabulated), B_j is the anisotropic displacement parameter, $\vec{r}_j$ is the position of the atom, 2θ is the angle between the primary and the diffracted X-ray beam of wavelength λ (set to the copper Kα ray = 1.5418 Å), and $\vec{h}$ is a reciprocal space vector. The EDM can then be computed by the Fourier transform of the structure factors

$$\rho(\vec{r}) = \frac{1}{\mathrm{V}} \sum_{\vec{h}=-\infty}^{+\infty} F(\vec{h}) e^{-2\pi i \vec{h}.\vec{r}} \tag{2}$$

where V is the volume of the crystallographic UC. The crystallographic resolution of the EDM, related to θ and λ, can be defined by the minimum distance d_{min} between two lattice planes in the reflection process:

$$\left(\frac{\sin\theta}{\lambda}\right)_{\mathrm{max}} = \frac{1}{2d_{\mathrm{min}}} \tag{3}$$

The intentional choice of a resolution value permits to smooth the EDM at a particular level of detail. A high resolution (d_{min} ~1 Å) allows the description of a molecule at atomic level, whereas a lower resolution (d_{min} ~3 Å) focuses on groups of atoms. In XRD experiments, the resolution is purposely high to obtain very detailed EDM, which will facilitate the structure solving. In our case, the resolution is deliberately lowered to depict functional groups instead of individual atoms. An important point toward that goal is to identify the crystallographic resolution that associates one and only one peak of $\rho(\vec{r})$ with one functional group, which was achieved by Binamé et al.[24] Their strategy was based on the methodology detailed above. In this last work, the authors identified the conditions of calculation for the EDM, especially discovering the crystallographic resolution to associate one peak per chemical substructure moiety. Practically, they defined a set of common functional groups and calculated EDM and peaks of $\rho(\vec{r})$ at several crystallographic resolutions between 2.0 and 5.0 Å for a set of well-known pharmacological compounds. They concluded that the best resolution (d_{min}) value was 2.6 Å as approximately 91% of the substructures are associated with a single peak. We thus based our description of functional groups on this assessment and chose the same resolution to build our strategy.

A quantum mechanical (QM) method coupled to a resolution reduction of the EDM, e.g., using a wavelet smoothing method, would require heavy computational resources. In comparison, the strategy described above calculates simple summations of Fourier transforms of structure factors; the smoothing is ensured by setting the desired crystallographic resolution (d_{min}). Moreover, the undeniable precision of QM methods over our method vanishes when working at a high degree of smoothing as we do, hence the computational expense for using such methods would be unjustified.

In practice, the 3D coordinates and the nature of the atoms were read from *mol2* files extracted from CSD. Since the goal was to calculate descriptors for isolated molecules rather than for crystal structures, the EDM reconstruction procedure was adapted similarly as described in ref 13. $\rho(\vec{r})$ was calculated for a single molecule centered in an orthorombic cell whose lengths were augmented by adding an extra 5 Å on each side of the molecular structure. This resulted in a vanishing $\rho(\vec{r})$ function at the edges of the cell. Notice that the output files from CSD were treated by in-house procedures to remove small molecules (generally cocrystallized solvent) and to separate the structures in the case where several molecules were enclosed in single files. The grid step of the EDM was set to 0.5 Å, a compromise between the numerical precision for detecting the peaks and the calculation time. Indeed, a grid step of 0.75 Å does not permit locating all peaks on the grid, while a step of 0.25 Å requires longer calculation times. Binamé et al. demonstrated that the number of retrieved peaks with a 0.5 Å grid interval closely corresponds to the number of functional groups.

The ORCRIT program[25] was used to locate and identify the CPs, based on the local second derivatives (curvatures) of the EDM corresponding to the three eigenvalues of the diagonalized Hessian matrix. Our study focused on CPs for which the three curvatures of $\rho(\vec{r})$ are negative, *i.e.*, the peaks which correspond to $\rho(\vec{r})$ maxima. Note that other types of CPs include saddle points with 1 or 2 negative curvatures or pits with only positive curvatures. ORCRIT requires, as an input parameter, a cutoff $\rho(\vec{r})$ value to retrieve only the relevant CPs from the $\rho(\vec{r})$ distribution. The Fourier transform applied by XTAL to calculate $\rho(\vec{r})$ can, indeed, lead to ripples that generate nonrelevant CPs. The $\rho(\vec{r})$ cutoff value was

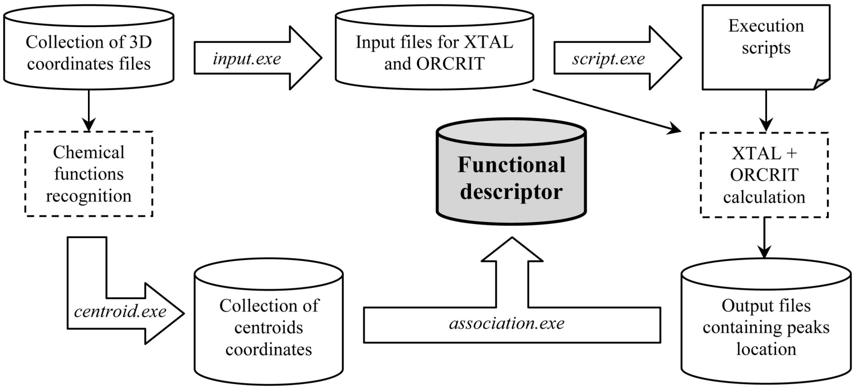
Collection of 3D coordinates files
input.exe
Input files for XTAL and ORCRIT
script.exe
Execution scripts
Chemical functions recognition
Functional descriptor
XTAL + ORCRIT calculation
centroid.exe
Collection of centroids coordinates
association.exe
Output files containing peaks location

Figure 1. General workflow of the calculation of the electron density peak descriptors, the functional group location, and the peak-functional group association.

thus fixed to 1.0 $e^-/Å^3$, a value commonly used in studies at that range of crystallographic resolution.[13,24]

The run of the two programs, XTAL and ORCRIT, as well as the management of the resulting files for the entire set of 62,936 molecules was implemented in a totally automated workflow. The final result after the XTAL and ORCRIT calculations is a list of CPs coordinates with four descriptors: the ED value at the peak position $\rho(0)$ and the three eigenvalues (EVs) of the Hessian matrix. Let us note that $\rho(0)$ corresponding to a maximum of $\rho(\vec{r})$ cannot be related to the one calculated at a QM level. Our method rebuilds the $\rho(\vec{r})$ distribution on the basis of a promolecular model where atomic $\rho(\vec{r})$ are centered on atomic nuclei. The smoothing brought by the medium resolution also modifies the shape of the electronic distribution comparing to QM computations.

Our procedures also involved the calculation of two additional peak descriptors, a fictitious ellipsoid volume,[11] calculated as

$$V = \frac{\pi^{3/2}\rho_{(0)}^{3/2}}{2^{3/2}|EV_x|^{1/2}|EV_y|^{1/2}|EV_z|^{1/2}} \tag{4}$$

and the maximal ellipsoid ellipticity

$$Ell = \ln\left(\frac{EV_{\max}}{EV_{\min}}\right) \tag{5}$$

where $EV_{\max}$ is the curvature with the highest absolute value, $EV_{\min}$ being the one with the lowest value. Approximately 180,000 peaks were eventually located, each being described by six descriptors. It can be surprizing to retrieve only ~3 functional groups per molecules, but one has to keep in mind that we were working with a set of 77 well chosen common substructures. A lot of complex functional groups arising from contiguous substructures were, for example, not taken into account.

Functional Groups Recognition. The next step consisted in selecting a set of the most relevant substructures usually present in pharmacological compounds with the help of fragment tables found in literature.[26−30] Their definition was also based on the analysis of 1200 pharmacological-like molecules provided by Aureus Pharma (Paris, France).[31] 77 functional substructures were finally retained on the basis of their occurrence frequency (See the detailed list of substructures in the Supporting Information.).

Each type of functional group was built and queried in the studied compound collection using the Conquest interface embedded in CSD. The process gave rise to a traditional problem in function recognition: some functions can be recognized in other functional moieties but may have lower chemical significance. For example, an *ester* contains both the *carbonyl* and the *ether* moieties; the central C atom of the *ester* could then be labeled according to each of the three substructures, whereas it should only be given the *ester* label as suggested by chemical common sense. To overcome the problem, a single rule was applied after the functional group recognition: "*If all the atoms of a substructure A are included in a substructure B, then A is not a functional group to be considered*". By doing so, all atoms were labeled only once by the most global and chemically relevant substructure, *i.e.*, corresponding to the best vision of the chemist.

Peak-Functional Group Association. For the purpose of assigning each peak to a functional group, centroid positions were calculated for each of the chemical moieties appearing in the 62,396 molecules, according to

$$\vec{r}_{centroid} = \frac{\sum_{i=1}^{m} \vec{r}_i \,.\, n_i}{\sum_{i=1}^{m} n_i} \tag{6}$$

where $\vec{r}_i$ and n_i are the position vector and the number of electrons for each atom i, respectively, and m is the number of atoms in the functional group. Such an expression locates each centroid toward atoms rich in electrons, just like a $\rho(\vec{r})$ peak is supposed to be. The role of the last procedure was to assign each peak of $\rho(\vec{r})$ to the closest centroid and thus to the associated functional group. A distance cutoff was arbitrarily applied to avoid irrelevant functional assignments: if the closest centroid to a peak is distant from more than 3.0 Å, the peak remains unassigned and is not taken into account. This prevents irrelevant peak-centroid association, the mean peak-functional group distance at 2.6 Å resolution being about 0.6 Å.[24] Once the assignment process was complete, a seventh peak descriptor was computed: the distance between the peak and the centroid of the assigned chemical substructure. A workflow of the calculation of the peak descriptors and association process with the functional groups is presented in Figure 1.

RESULTS AND DISCUSSION

XTAL and ORCRIT Computations. The computation of the $\rho(\vec{r})$ properties for 62,936 molecules by XTAL and

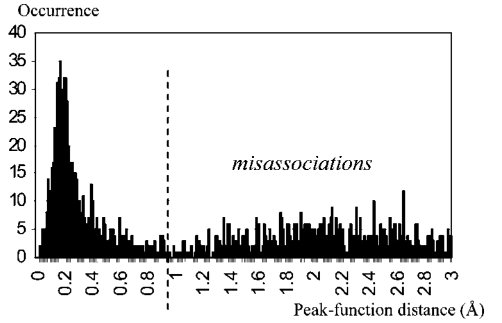

Figure 2. Distribution of the peak-functional group distance for 1373 carboxylic acids, illustrating the refinement process for the peak-function association.

ORCRIT was performed on Sun 900 MHz 64 bit CPUs in 119 h, the average calculation time for one molecule being 6.7 s. With such an amount of data to treat, the simulated XRD experiment (*cf.* Methodology) to reconstitute the EDM is advantageous to speed up the computation.

Refinement of the Peak-Functional Group Association. As mentioned earlier, a first general peak-centroid distance cutoff of 3.0 Å was applied to exclude undesired peak-functional group associations. This threshold was further refined for each particular functional substructure on the basis of the distribution of each peak-functional group distance. Indeed, most of the frequency histograms established for the distances (with bins of 0.01 Å) showed a dominant normal distribution followed by a minimum and, at larger distances, a noisy behavior corresponding mostly to undesired associations. This phenomenon is illustrated in Figure 2 presenting the distance frequency histogram computed from 1373 carboxylic acid functions in the data set of 62,936 molecules. To suppress the misassociations more efficiently, the peak-centroid distance cutoffs were adjusted in a way that eliminates peaks not belonging to the main distribution. This refinement step significantly reduced the number of retained peaks, for example to 657 for carboxylic acids. The threshold distance was defined as the global minimum of the distribution encountered after the highest maximum. The elimination of misassociated peaks successfully cleaned our data set and consequently tightened the distribution curves for the six other descriptors. Note however that the refinement could not be performed for all 77 functional groups as there are for example cases where the peak-centroid distance distribution did not present a clear maximum but a widespread curve. It is the case for functional groups solely composed of C and N atoms (alkenes, amines, highly substituted phenyls,...), i.e., chemical moieties with a low number of electrons. For these functional groups, the associated peak is considerably influenced by the neighborhood, leading to highly variable peak-chemical centroid distances and spread out distance distributions.

Analysis of the Distributions. Each descriptor for each functional group was plotted in histograms to observe its distribution. The trends followed by the descriptor values were investigated. An example of the normal distributions for three descriptors for the collection of 657 carboxylic acids is illustrated in Figure 3. In most cases, each descriptor follows a normal statistical distribution (Gaussian-like curve). In rare cases however, a normal distribution was not observed. For example, only 19 isothiocyanate functions (N=C=S) were retrieved in the whole set of molecules, which distorts the analysis.

Interestingly, several distinct or interpenetrated curves were observed for one given descriptor for a particular functional group. As we ensured the "traceability" of the functions during the calculation process, the meaning of these remarkable curves could be retrieved from the corresponding molecules. In several instances, some curves revealed that the descriptors could discriminate several subtypes of functional group. For example, the analysis of the peak-centroid distance for the fluoride function presented a second clearly separated cluster at a large distance (Figure 4). It corresponds to a subgroup exclusively composed of the trifluoro moiety (CF_3) which was not *a priori* defined as a functional group. The three C−F peaks actually merged, leading to a unique peak at a mean distance of 0.648 Å from the supposed C−F centroids as opposed to a 0.252 Å average distance for "classical" fluoride functions (Figure 5). The mean $\rho(0)$ for the trifluoro is also higher (3.084 $e^{-}/Å^{3}$) than the C−F one (2.255 $e^{-}/Å^{3}$). Peaks related to CHF_2 moieties are not merged and are thus included in the C−F distribution. The particular example of the trifluoro moiety indicates a reasonable discriminating power of our descriptors. Location of peaks associated with a para-substituted phenyl and a nitrile are also illustrated in Figure 5.

The variability of each of the six different descriptors within a functional group family is also noticeable. The $\rho(0)$ value and the highest curvature value of each peak presented the tightest curves, with average standard deviations for the 77 types of substructure approaching ~10 and ~18% of the mean value, respectively. As a reference, the standard deviation of the five other descriptors was comprised between ~33 and ~44% of the mean value. It is worth mentioning the particularly high standard deviation of the volume descriptor, which is calculated on the basis of four other descriptors. The variability of the volume descriptor is thus more important than individual variabilities due to additivity. Notice that, in our case, standard deviations are more depicting of the variation all over the population of studied functional groups than an error due to the lack of measure accuracy. Most of the time, high standards deviations are the consequence of the variety of structures in the studied data sets.

Comparison between Functional Groups. The mean descriptor values for each of the 77 chemical substructures were calculated and compared (Table 1). For the $\rho(0)$ value, functional groups with one or more atoms rich in electrons, i.e., P, S, Cl, Br, and I atoms, are well discriminated as they have $\rho(0)$ values above 3.000 $e^{-}/Å^{3}$. Functional groups with only C, O, N, or F atoms have a $\rho(0)$ ranging between 1.786 $e^{-}/Å^{3}$ (alkyne groups) and 2.847 $e^{-}/Å^{3}$ (5-membered heterocycles with four N atoms). Structurally similar functional groups have really similar $\rho(0)$ values. For example, even for phenyl rings that are distinguishable by their number of substitutions, the $\rho(0)$ value remains close to ~2.0 $e^{-}/Å^{3}$. For the 77 substructures, no correlation was observed between the number of electrons and the $\rho(0)$ value of the associated peak (R^2=0.083). This is consistent with the fact that $\rho(0)$depends not only from the number of electrons but also from their 3D distribution.

A high peak volume usually indicates diffuse electronic distribution in the function. It is particularly the case in large

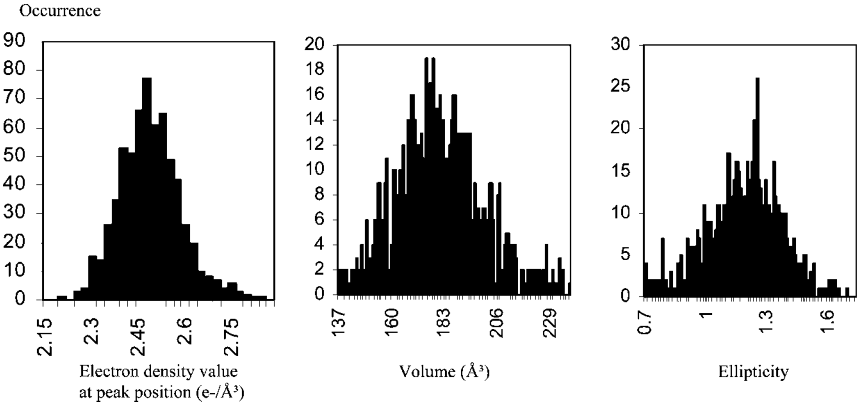

Figure 3. Distribution of three descriptors (electron density value at peak position, volume, and ellipticity of the peak) for the 657 carboxylic acids.

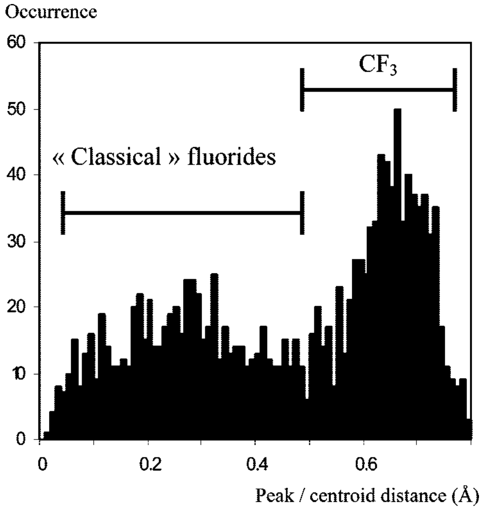

Figure 4. Distribution of the peak-centroid distance for 1494 fluorides showing that trifluoro moieties are discriminated.

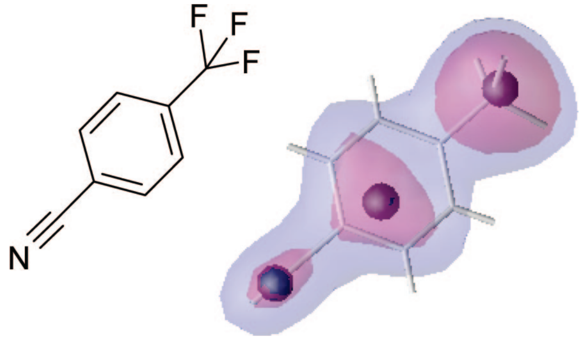

Figure 5. Example of the location of peaks associated with a nitrile, a para-substituted phenyl, and a single peak associated with the CF_3 moiety. Dark spheres represent the peaks; isocontours of electronic density are 1.0 $e^-/Å^3$ (light blue) and 1.5 $e^-/Å^3$ (purple).

substructures with few electrons such as phenyls, pyridine, ester, etc. Thus, the peak volume relates to the steric volume of the functional group itself.[11] The nearest neighborhood of a function can influence its peak volume. The amide moiety illustrates nicely this concept. Indeed, the more an amide is substituted (primary, secondary, and tertiary amine), the larger the volume. The neighboring entities smooth the peak curvatures and the volume increases subsequently. Halogens, that are characterized by a high $\rho(0)$ value, have a particularly low volume. In general a peak with high curvature will generally be associated with a low volume.

The conclusions about the ellipticity values mainly relate to ring functions. A high ellipticity is characteristic of an electron distribution accumulated preferentially in a given plane. Thus, higher ellipticities, *i.e.*, around ~ 2.0, tend to associate with cyclic functions, generally planar aromatic rings. Following this logic, lower ellipticities correspond to more spherical peaks (three curvatures nearly equal) indicating an isotropic distribution of the charge.

If considered globally, the three curvatures correlate well to $\rho(0)$ with R^2 values around 0.90. Nonetheless, each curvature individually brings different information. In-depth analysis reveals that the second curvature of the peak ellipsoid directly relates to the substitution of phenyl rings. Such information about substitution is not vehicled by any other descriptor. The plot of the second curvature versus the number of substituting groups of the phenyl moieties shows a linear dependence with R^2=0.987 when leaving out the 6-times-substituted phenyls (Figure 6). The third curvature is particularly low for all the cyclic moieties and corresponds to the smoother variation of $\rho(\vec{r})$ along the ring plane. As the third curvature is the denominator term in the ellipticity calculation, it explains why the ellipticity is high for peaks associated with ring-type structures.

The trends followed by the peak-centroid distance are more difficult to interpret due to the influence of neighboring functions, even for peaks among a given function type. In general the peaks associated with functions substituted multiple times are typically located at higher peak-centroid distances due to the influence of the neighboring groups. It is particularly true for all the ring systems that can be highly substituted, $d \sim 0.70$ Å. Terminal functional groups such as primary amide, carboxylic acid, aldehyde, nitrile, and nitro groups have a relatively small peak-centroid, $d \sim 0.20$ Å. It is also worth mentioning that two particular functional groups are characterized by an unusually high but well defined peak-functional group distance: carboximide C(=O)−NH−C=O with $d = 1.488 \pm 0.176$ Å and anhydride C(=O)−O−C=O with $d = 1.204 \pm 0.239$ Å. The apparent structural similarity between the two functional groups justifies a common explanation of the high peak-centroid distance. In most of the cases, two peaks are associated with these functions. Anhydride can be viewed as the merge of two ester moieties, each of them being characterized by its own peak (Figure 7). This is corroborated by the similarity between peaks of anhydrides and esters, the relative difference for $\rho(0)$, volume, ellipticity, and first and second curvatures not

Table 1. Mean and Standard Deviation of the 7 Descriptors for Each of the 77 Substructures out of the 62,936 Molecules

name	count[a]	$\rho(0)$ (e-/Å^3)	d (Å)	V (Å^3)	ellipticity	EV1	EV2	EV3
1-phenyl[b]	7435	1.965 ± 0.148	0.618 ± 0.203	404.853 ± 116.198	2.328 ± 0.750	−0.192 ± 0.026	−0.055 ± 0.028	−0.022 ± 0.014
2-phenyl(1,2)[b]	3154	1.946 ± 0.245	1.006 ± 0.516	293.682 ± 115.833	2.284 ± 1.625	−0.179 ± 0.032	−0.068 ± 0.041	−0.031 ± 0.025
2-phenyl(1,3)[b]	290	1.987 ± 0.151	0.685 ± 0.199	338.191 ± 95.991	2.074 ± 0.457	−0.188 ± 0.026	−0.068 ± 0.026	−0.026 ± 0.015
2-phenyl(1,4)[b]	3473	1.969 ± 0.149	0.694 ± 0.201	298.010 ± 79.092	1.741 ± 0.428	−0.185 ± 0.024	−0.060 ± 0.022	−0.035 ± 0.014
3-phenyl(1,2,3)[b]	452	1.931 ± 0.183	0.698 ± 0.203	336.667 ± 122.688	2.065 ± 0.588	−0.182 ± 0.025	−0.067 ± 0.034	−0.027 ± 0.017
3-phenyl(1,2,4)[b]	878	2.023 ± 0.158	0.700 ± 0.148	259.484 ± 53.999	1.623 ± 0.395	−0.188 ± 0.025	−0.074 ± 0.025	−0.040 ± 0.016
3-phenyl(1,3,5)[b]	72	1.932 ± 0.169	0.734 ± 0.198	305.360 ± 91.754	1.845 ± 0.543	−0.181 ± 0.026	−0.065 ± 0.029	−0.032 ± 0.017
4-phenyl(1,2,3,4)[b]	210	2.051 ± 0.552	1.263 ± 0.854	334.215 ± 85.185	1.872 ± 0.672	−0.194 ± 0.064	−0.074 ± 0.060	−0.040 ± 0.050
4-phenyl(1,2,3,5)[b]	1568	1.957 ± 0.159	0.773 ± 0.181	264.751 ± 71.425	1.627 ± 0.442	−0.182 ± 0.033	−0.073 ± 0.032	−0.039 ± 0.019
4-phenyl(1,2,4,5)[b]	262	2.009 ± 0.335	1.542 ± 0.799	302.341 ± 111.339	1.795 ± 0.702	−0.187 ± 0.056	−0.086 ± 0.053	−0.039 ± 0.036
5-phenyl[b]	373	1.997 ± 0.360	1.514 ± 0.793	344.108 ± 109.571	1.899 ± 0.689	−0.186 ± 0.050	−0.082 ± 0.052	−0.036 ± 0.032
6-phenyl[b]	451	2.020 ± 0.368	1.540 ± 0.761	273.510 ± 93.635	1.741 ± 0.638	−0.190 ± 0.064	−0.092 ± 0.056	−0.042 ± 0.034
5m-2N(1,2)[c]	288	2.436 ± 0.208	0.398 ± 0.197	244.735 ± 81.347	1.436 ± 0.454	−0.219 ± 0.035	−0.093 ± 0.028	−0.057 ± 0.023
5m-2N(1,3)[c]	409	2.429 ± 0.195	0.351 ± 0.175	244.702 ± 84.238	1.448 ± 0.430	−0.220 ± 0.031	−0.095 ± 0.028	−0.056 ± 0.021
5m-3N(1,2,3)[c]	66	2.663 ± 0.190	0.373 ± 0.118	187.827 ± 31.223	1.126 ± 0.267	−0.239 ± 0.026	−0.117 ± 0.025	−0.080 ± 0.020
5m-3N(1,2,4)[c]	129	2.595 ± 0.288	0.403 ± 0.255	237.478 ± 99.471	1.353 ± 0.486	−0.244 ± 0.067	−0.105 ± 0.043	−0.068 ± 0.029
5m-4N[c]	19	2.847 ± 0.108	0.224 ± 0.099	187.304 ± 23.892	1.039 ± 0.200	−0.250 ± 0.020	−0.119 ± 0.022	−0.090 ± 0.015
aldehyde	108	2.222 ± 0.103	0.162 ± 0.083	141.059 ± 15.761	0.852 ± 0.238	−0.188 ± 0.020	−0.144 ± 0.015	−0.082 ± 0.018
alkene	13669	1.906 ± 0.267	1.613 ± 0.769	277.442 ± 113.572	1.784 ± 1.197	−0.166 ± 0.055	−0.090 ± 0.046	−0.038 ± 0.027
alkyne	899	1.786 ± 0.254	0.266 ± 0.142	240.842 ± 97.052	1.610 ± 0.843	−0.163 ± 0.054	−0.134 ± 0.042	−0.043 ± 0.032
amide (primary)	131	2.335 ± 0.076	0.211 ± 0.078	189.686 ± 18.358	1.292 ± 0.186	−0.209 ± 0.021	−0.119 ± 0.019	−0.058 ± 0.009
amide (secondary)	1311	2.266 ± 0.142	0.340 ± 0.140	203.321 ± 48.931	1.289 ± 0.351	−0.204 ± 0.023	−0.104 ± 0.022	−0.059 ± 0.018
amide (tertiary)	1413	2.232 ± 0.128	0.397 ± 0.139	209.191 ± 66.074	1.336 ± 0.444	−0.202 ± 0.022	−0.103 ± 0.022	−0.057 ± 0.019
amidine	185	2.151 ± 0.301	1.155 ± 0.478	264.387 ± 97.973	1.736 ± 0.607	−0.203 ± 0.063	−0.105 ± 0.060	−0.043 ± 0.032
amine (primary)	489	2.060 ± 0.243	0.852 ± 0.215	227.733 ± 87.901	1.394 ± 0.385	−0.177 ± 0.029	−0.104 ± 0.028	−0.047 ± 0.020
amine (secondary)	1374	2.144 ± 0.487	1.125 ± 0.444	274.587 ± 90.389	1.575 ± 0.665	−0.173 ± 0.040	−0.106 ± 0.043	−0.046 ± 0.040
amine (tertiary)	1316	2.043 ± 0.170	0.504 ± 0.234	273.074 ± 93.931	1.508 ± 0.553	−0.167 ± 0.029	−0.078 ± 0.023	−0.041 ± 0.019
anhydride	176	2.360 ± 0.166	1.204 ± 0.239	221.643 ± 59.456	1.547 ± 0.462	−0.212 ± 0.025	−0.118 ± 0.024	−0.049 ± 0.020
azide	52	2.435 ± 0.236	0.390 ± 0.288	164.593 ± 24.488	1.402 ± 0.332	−0.222 ± 0.053	−0.176 ± 0.043	−0.056 ± 0.016
bromide	1099	7.340 ± 0.163	0.280 ± 0.070	101.014 ± 4.726	0.328 ± 0.074	−0.639 ± 0.043	−0.516 ± 0.026	−0.460 ± 0.026
carbamate	420	2.533 ± 0.124	0.230 ± 0.112	206.272 ± 37.473	1.335 ± 0.346	−0.237 ± 0.030	−0.107 ± 0.019	−0.065 ± 0.018
carboximide	741	2.233 ± 0.242	1.488 ± 0.176	204.628 ± 61.717	1.302 ± 0.394	−0.199 ± 0.031	−0.105 ± 0.027	−0.058 ± 0.024
carboxylic acid	657	2.477 ± 0.117	0.253 ± 0.142	180.894 ± 21.186	1.180 ± 0.192	−0.219 ± 0.024	−0.127 ± 0.022	−0.068 ± 0.013
chloride	2864	3.356 ± 0.151	0.870 ± 0.203	113.392 ± 8.798	0.524 ± 0.169	−0.288 ± 0.029	−0.235 ± 0.019	−0.171 ± 0.023
cyclopropane	296	2.071 ± 0.153	0.235 ± 0.118	226.291 ± 75.626	0.939 ± 0.374	−0.142 ± 0.027	−0.099 ± 0.023	−0.060 ± 0.024
diazo	175	2.283 ± 0.176	0.372 ± 0.206	202.647 ± 50.556	1.507 ± 0.453	−0.198 ± 0.023	−0.139 ± 0.027	−0.048 ± 0.019
disulfide	412	3.184 ± 0.253	0.629 ± 0.300	211.347 ± 58.259	1.769 ± 0.558	−0.286 ± 0.037	−0.209 ± 0.037	−0.056 ± 0.032
enamine	805	2.140 ± 0.383	1.577 ± 0.766	264.914 ± 93.390	1.584 ± 0.618	−0.187 ± 0.040	−0.091 ± 0.038	−0.046 ± 0.031
epoxyde	313	2.446 ± 0.220	0.206 ± 0.116	175.100 ± 30.463	0.711 ± 0.282	−0.179 ± 0.059	−0.133 ± 0.037	−0.089 ± 0.027
ester	6324	2.441 ± 0.155	0.292 ± 0.143	212.766 ± 50.641	1.469 ± 0.397	−0.226 ± 0.031	−0.118 ± 0.028	−0.056 ± 0.021
ether	5228	2.223 ± 0.143	0.510 ± 0.135	216.618 ± 69.031	1.466 ± 0.451	−0.186 ± 0.025	−0.127 ± 0.027	−0.047 ± 0.018
fluoride	640	2.255 ± 0.163	0.252 ± 0.125	203.686 ± 57.207	1.232 ± 0.405	−0.197 ± 0.025	−0.106 ± 0.028	−0.061 ± 0.021
furane	161	2.613 ± 0.146	0.286 ± 0.113	189.852 ± 25.835	1.106 ± 0.207	−0.235 ± 0.023	−0.109 ± 0.021	−0.079 ± 0.015
guanidine	75	2.338 ± 0.382	1.396 ± 0.860	244.944 ± 86.142	1.454 ± 0.539	−0.204 ± 0.043	−0.104 ± 0.045	−0.055 ± 0.031
halogen acid	22	2.811 ± 0.680	1.867 ± 0.840	123.232 ± 11.443	0.907 ± 0.455	−0.229 ± 0.070	−0.177 ± 0.073	−0.106 ± 0.054
hydrazine	215	2.232 ± 0.284	0.614 ± 0.251	280.771 ± 109.089	1.665 ± 0.606	−0.191 ± 0.038	−0.097 ± 0.040	−0.043 ± 0.027
hydrazone	260	2.220 ± 0.268	0.497 ± 0.261	220.358 ± 68.202	1.654 ± 0.520	−0.202 ± 0.061	−0.142 ± 0.054	−0.044 ± 0.025
hydroxilamine	35	2.315 ± 0.127	0.408 ± 0.166	208.534 ± 46.639	1.312 ± 0.369	−0.196 ± 0.024	−0.119 ± 0.027	−0.056 ± 0.019
hydroxyl	4373	2.070 ± 0.156	0.751 ± 0.194	214.039 ± 70.903	1.151 ± 0.400	−0.163 ± 0.028	−0.106 ± 0.027	−0.054 ± 0.019
imine	2908	2.254 ± 0.462	1.121 ± 0.474	258.885 ± 104.797	1.566 ± 0.614	−0.204 ± 0.065	−0.110 ± 0.059	−0.052 ± 0.040
iodide	581	12.036 ± 0.141	0.251 ± 0.102	80.621 ± 4.321	0.319 ± 0.081	−1.204 ± 0.047	−0.962 ± 0.024	−0.874 ± 0.022
isothiocyanate	19	2.978 ± 0.598	1.560 ± 0.778	137.179 ± 57.941	0.846 ± 0.639	−0.256 ± 0.075	−0.201 ± 0.073	−0.123 ± 0.051
ketone	4423	2.203 ± 0.291	0.375 ± 0.227	183.392 ± 48.769	1.126 ± 0.343	−0.201 ± 0.067	−0.118 ± 0.059	−0.068 ± 0.031
nitrile	1086	1.982 ± 0.112	0.228 ± 0.166	133.658 ± 25.500	0.919 ± 0.336	−0.175 ± 0.020	−0.142 ± 0.014	−0.073 ± 0.021
nitro	1393	2.803 ± 0.201	0.299 ± 0.175	159.714 ± 15.251	0.966 ± 0.248	−0.250 ± 0.058	−0.153 ± 0.047	−0.095 ± 0.017
nitroso	19	2.475 ± 0.305	0.249 ± 0.126	146.050 ± 15.714	1.123 ± 0.383	−0.224 ± 0.068	−0.151 ± 0.054	−0.076 ± 0.029
oxime	102	2.382 ± 0.135	0.403 ± 0.165	155.197 ± 19.863	1.062 ± 0.252	−0.205 ± 0.019	−0.150 ± 0.023	−0.073 ± 0.017
oximether	60	2.332 ± 0.279	1.378 ± 0.824	209.597 ± 52.863	1.535 ± 0.485	−0.198 ± 0.028	−0.142 ± 0.041	−0.048 ± 0.025
peroxyde	50	2.372 ± 0.165	0.339 ± 0.163	235.065 ± 62.504	1.519 ± 0.494	−0.191 ± 0.027	−0.120 ± 0.032	−0.047 ± 0.025
phosphine	254	3.031 ± 0.429	0.197 ± 0.105	151.613 ± 16.308	0.499 ± 0.215	−0.227 ± 0.094	−0.174 ± 0.067	−0.140 ± 0.065
phosphinoyl	98	3.326 ± 0.092	0.261 ± 0.083	150.623 ± 8.920	0.488 ± 0.209	−0.236 ± 0.027	−0.188 ± 0.018	−0.146 ± 0.019
phosphonate	158	3.811 ± 0.149	0.259 ± 0.111	151.836 ± 11.510	0.421 ± 0.142	−0.265 ± 0.033	−0.208 ± 0.020	−0.173 ± 0.017
pyrane	31	2.301 ± 0.906	0.814 ± 0.185	288.239 ± 74.903	1.950 ± 0.615	−0.216 ± 0.087	−0.095 ± 0.094	−0.044 ± 0.068
pyrazine	104	2.116 ± 0.147	0.587 ± 0.238	323.053 ± 95.645	2.018 ± 0.812	−0.200 ± 0.026	−0.067 ± 0.023	−0.031 ± 0.014
pyridine	896	2.103 ± 0.173	0.637 ± 0.198	304.245 ± 88.855	1.821 ± 0.466	−0.201 ± 0.028	−0.067 ± 0.030	−0.036 ± 0.020
pyrimidine	137	2.210 ± 0.162	0.738 ± 0.277	286.524 ± 85.953	1.865 ± 0.542	−0.208 ± 0.027	−0.082 ± 0.027	−0.036 ± 0.018
pyrrole	198	2.338 ± 0.174	0.286 ± 0.158	221.515 ± 42.371	1.365 ± 0.336	−0.216 ± 0.026	−0.091 ± 0.024	−0.058 ± 0.017
sulfate	10	3.639 ± 1.220	1.393 ± 1.177	142.05 ± 11.488	0.771 ± 0.607	−0.258 ± 0.082	−0.203 ± 0.086	−0.155 ± 0.091
sulfone	806	4.064 ± 0.130	0.285 ± 0.105	141.765 ± 8.533	0.417 ± 0.122	−0.290 ± 0.026	−0.236 ± 0.017	−0.192 ± 0.017
sulfoxyde	208	3.589 ± 0.149	0.280 ± 0.090	136.308 ± 12.062	0.555 ± 0.224	−0.281 ± 0.030	−0.221 ± 0.024	−0.162 ± 0.023
sultone	102	4.238 ± 0.125	0.212 ± 0.068	141.173 ± 7.477	0.391 ± 0.096	−0.303 ± 0.025	−0.241 ± 0.019	−0.205 ± 0.014
thioether	2401	3.140 ± 0.147	0.285 ± 0.172	132.643 ± 12.728	0.613 ± 0.181	−0.260 ± 0.026	−0.190 ± 0.023	−0.142 ± 0.021
thioketone	424	3.127 ± 0.150	0.412 ± 0.153	120.300 ± 10.436	0.617 ± 0.193	−0.265 ± 0.024	−0.217 ± 0.020	−0.145 ± 0.024
thiol	271	3.313 ± 0.248	0.670 ± 0.272	189.302 ± 49.520	1.408 ± 0.506	−0.278 ± 0.043	−0.228 ± 0.038	−0.074 ± 0.031
thiophene	194	3.225 ± 0.263	0.817 ± 0.145	140.341 ± 10.282	0.714 ± 0.198	−0.272 ± 0.033	−0.181 ± 0.030	−0.135 ± 0.021
trifluoro	469	3.084 ± 0.162	0.648 ± 0.061	166.501 ± 17.372	0.529 ± 0.149	−0.214 ± 0.031	−0.156 ± 0.021	−0.126 ± 0.016
urea	571	2.364 ± 0.128	0.340 ± 0.145	218.631 ± 54.284	1.392 ± 0.414	−0.218 ± 0.023	−0.098 ± 0.022	−0.057 ± 0.019

[a] Number of peaks used to calculate each mean value (after refinement). [b] The first number indicates the number of substitutions; numbers between brackets indicate the position of the substitutions. [c] Five membered heterocycles. The first number indicates the number of nitrogens (N) in the ring and numbers between brackets the position of the N atoms.

exceeding 6.5%. Only the third curvature differs more significantly by 13.3%. The same interpretation stands for carboximide which corresponds to a merge of two secondary amides. The analyses confirm that the crystallographic

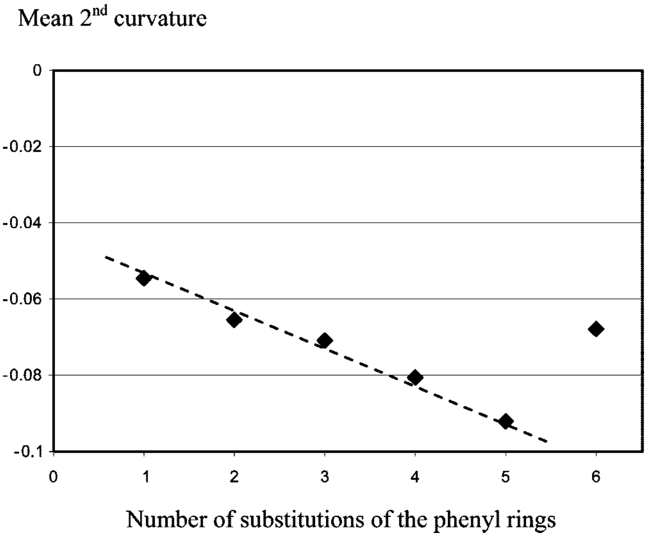

Figure 6. Plot of the second curvature versus the substitution of phenyl rings showing the linear dependence for the five first phenyl types.

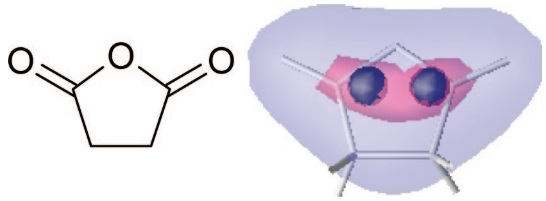

Figure 7. Example of two peaks associated with an anhydride function. Dark spheres represent the peaks; isocontours of electronic density are 1.0 e$^-$/Å^3 (light blue) and 2.0 e$^-$/Å^3 (purple).

resolution of 2.6 Å is the optimal value to associate one peak to one functional group. Of course larger functional groups associate with multiple peaks, as discussed above for both anhydride and carboximide functions. It is also valid to a lesser extent for cyclic structures. A qualitative summary of several function types with the various peak properties discussed here is presented in Table 2.

Principal Component Analysis (PCA). Extracting knowledge from large data tables is not a trivial task. Each descriptor captures important information about functional groups but is insufficient by itself to describe the substructures if taken away from its molecular context. Thus a PCA of the peak descriptors[32] was carried out for the purpose of regrouping all the characteristics within the descriptors while providing a more global view of the significance they vehicle. The peak-centroid distance descriptor was not included in the PCA as it was not well-defined for some functions as written above. Knowing that the peak-centroid distance can be large in some instances, considering it in the PCA would lead to significant delocalization of some datapoint clouds, otherwise well-defined in principle component (PC) space. The six considered descriptors were first normalized on the basis of a normal Gaussian distribution ($\mu = 0$, $\sigma^2 = 1$). Three PCs were calculated and reflected 96.67% of the variability contained in the six initial descriptors: PC1 = 73.95%, PC2 = 18.47%, PC3 = 4.25%. Such a conserved high variability denotes that reducing six dimensions to three is still reliable and that the representation is robust and easier to treat. The eigenvalues from the PCA indicate that none of the descriptors seems to be more influent than another.

Projecting the datapoints in the 3D space of the first three PCs led to a main, boomerang-shaped cloud of datapoints (Figure 8). The first important observation is that within this main cloud, the points are regrouped in smaller clouds, each corresponding to one functional group. Two clouds, outside the main one because of their extreme descriptor values, correspond to the iodide and bromide functions (not shown in Figure 8). As the sample of studied molecules is quite large, the data corresponding to a particular functional group are expected to have a broad dynamic range. But this does not affect the distribution in the 3D PC space, each functional group being well-defined. A second important observation deriving from the first one resides in the transferability of our descriptors. If a new datapoint corresponding to a given function type was calculated, it would likely project within the appropriate corresponding cloud because our sample collection derives from a high number of structurally diverse compounds. Therefore our descriptors are believed to provide a satisfactory description of the electronic properties of each functional group.

Some clouds can overlap to a certain extent, which indicates that certain particular functional groups are harder to discriminate from each other when utilizing our descriptor. Further analysis reveals that overlapping clouds correspond in most cases to chemically similar functional groups. Their overlap is thus not a weakness in our methodology, and the intersecting clouds could *a fortiori* be considered as indicators of similarity among functional groups. Totally overlapping clouds relate to very comparable functions, as ketone and aldehyde for example (Figure 9). Other related functional entities, taking different positional substitutions of phenyls, are partially overlapping, but associated clouds still reside in the same portion of the PC space.

Hierarchical Clustering (HC). Importantly, PCA concluded that clouds of similar functional groups are close in PC space, while dissimilar functions would be far from each other. This observation indicates that the Euclidian distance between cloud positions in PC space would be appropriate to measure functional similarity. Based on these concepts, HC was carried out to further quantify functional similarity and automate its calculation. HC arranges objects in a hierarchy (dendrogram) based on the similarity between them.[33] HC therefore utilizes a similarity metric; in our case the Euclidean distance between the cloud centroids in the 3D PC space. Note that the unweighted pair-group method with arithmetic mean (UPGMA)[34] was selected as cluster similarity method. The bromide and iodide moieties were not taken into account to calculate the dendrogram. These two functional groups are already well discriminated in the PC space and were ignored to avoid drawing the obvious conclusion that they are well discriminated, "polluting" the calculation with extreme numbers and reducing the dynamic range of other functional groups.

The resulting dendrogram is presented in Figure 10. At a coarse level (high Euclidian distance threshold), two main clusters are observed: one corresponding to electronically dense functional groups composed of halogens, sulfur, and phosphorus atoms (the 15 functions in top of the dendrogram) and the other one mainly composed of functional groups with atoms from the second period (the 60 functions in bottom part). Some of the trends discussed above are confirmed, as exemplified by the quick merge of the ketone and aldehyde

Table 2. Qualitative Summary of Peak Properties for Several Types of Functional Groups[a]

functional group type						
no. of e^- in atoms	no. of atoms	example	$\rho(0)$	volume	ellipticity	curvatures
high	high	sulfate	++	+	−	++
high	low	bromide	++	−	−	++
low	high	urea	+	++	+	+
low	low	nitrile	−	−	−	+
particular case: rings		pyridine	+/−	++	++	−

[a] "+" and "−" stand for high and low values, respectively.

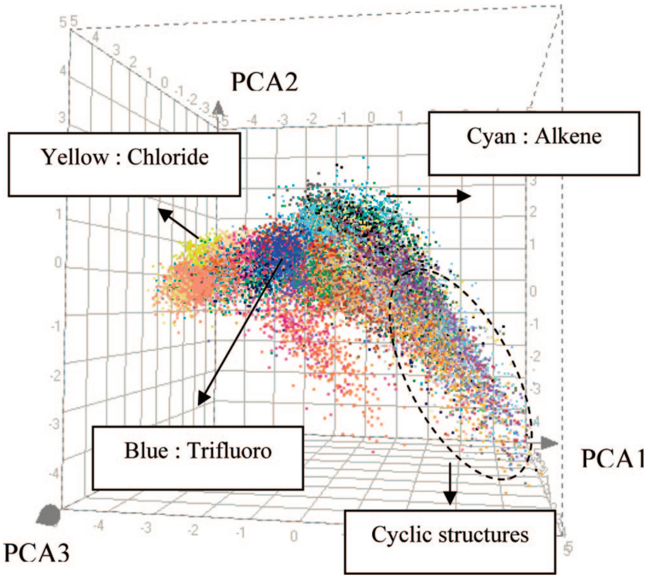

Figure 8. Projection of datapoints in principle component space corresponding to 75 substructures (bromide and iodide excluded) colored by types. Peaks of the same type are located in the same portion of space (75-color legend not shown for clarity). Several examples of functional group clouds are pinpointed.

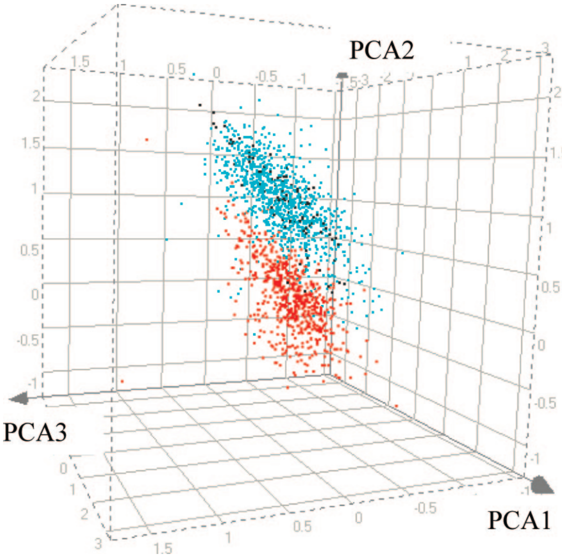

Figure 9. Zoom on the PC space containing the datapoint clouds corresponding to ketones (cyan dots), aldehydes (black dots), and carboxylic acids (red dots). Aldehydes totally overlap with ketones as they are really comparable electronically; carboxylic acids are well differentiated from the two other functions.

branches at the short 0.091 distance. It was also anticipated that phenyls would merge fast at low distance units. Generally speaking, structurally similar functional groups merge at lower distances.

Other interesting merges are observed, such as fluoride and carboximide. As odd as it can appear, peaks associated with these functions actually have similar properties. Another factor to keep in mind is that only the position of cloud centroids were utilized to measure Euclidian distances and not the shape or the orientation of each cloud.

Remarkably, HC permitted not only the highlighting of the relevance of the descriptors but also their limitation. Observing the dendrogram at the leaf-level of the hierarchical tree (Euclidian distance = 0.0), the functional groups were defined by us in a subjective way relying on chemical common sense. But, as similarity levels are increased, descriptors are progressively delivering their information, revealing which functional groups can be discriminated and which cannot. Thus looking at clusters at various similarity levels leads to functional equivalences at various levels of detail. That conclusion must be taken into account when the descriptors will be applied to study pharmacological ligands, for example.[35]

CONCLUSIONS AND OUTLOOK

The aim of our work was to explore the discriminating power of electron density $\rho_{(\vec{r})}$ (ED) peaks describing 77 types of functional groups extracted from 62,936 organic-like molecules from CSD. This was accomplished in two main steps: one being methodological and the other analytical.

The methodological part of the work was to implement several automated procedures to achieve a sequence of tasks. Electron density map (EDM) calculations were performed by XTAL on the basis of crystallographic parameters; the retrieval and identification of the peaks of the ED distribution was performed utilizing ORCRIT. EDMs were calculated at a medium crystallographic resolution of 2.6 Å. Consequently, one peak could be associated with one functional group. In parallel, we developed a workflow to automate the recognition of the functional groups encountered in the molecules. The final step consisted of assigning the peaks of $\rho_{(\vec{r})}$ to the corresponding chemical functions and calculating local descriptors.

In the analytical part, we first focused on the distributions and mean values of the descriptors, to underline some general, interesting trends that unveiled some of the information contained in the descriptors. Principal component analysis (PCA) was then used to view systematically how the peaks of the $\rho(\vec{r})$ were distributed in the space and demonstrated that similar functional groups are close in PC space while dissimilar functions are distant. Therefore, the Euclidian distance could be used as a similarity metric in a hierarchical clustering (HC) analysis, which further quantified functional similarity and organized functional moieties based on their electronic likeness.

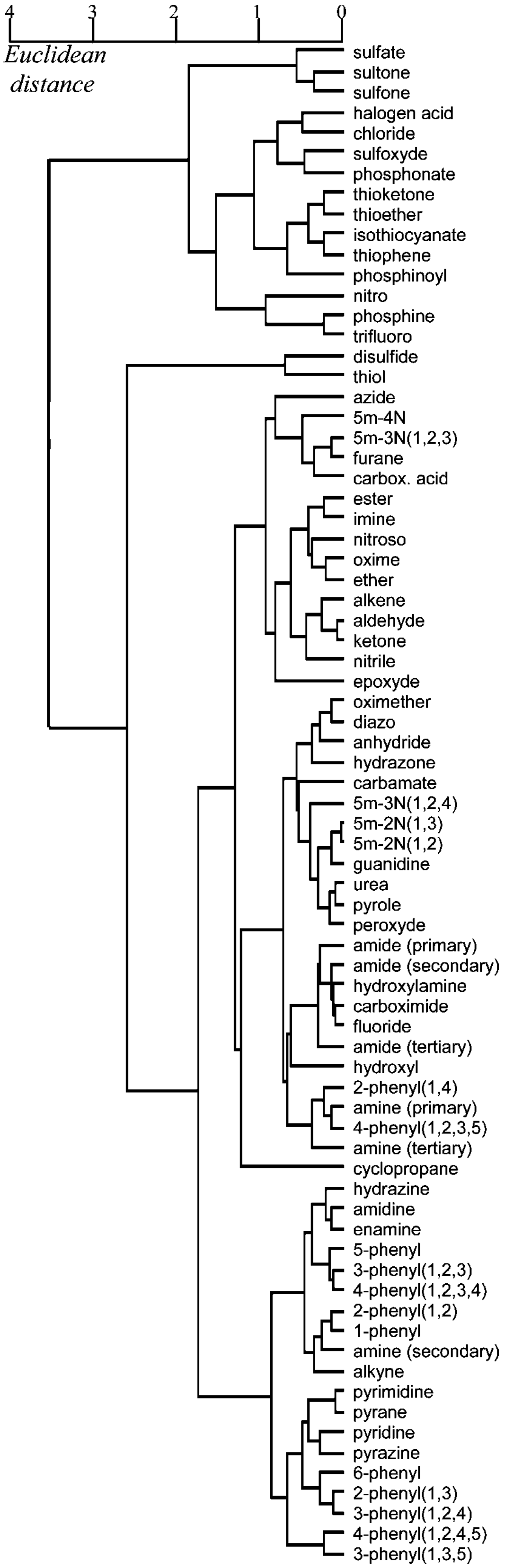

Figure 10. Dendrogram obtained by hierarchical clustering on the 75 cloud centroids from the PC space (iodide and bromide excluded).

On-going studies have leveraged this knowledge toward the implementation of molecular fingerprints derived from the hierarchical classification of the functional groups presented here. Indeed, at various functional similarity thresholds, bit strings indicating the presence/absence of functional moieties/clusters can be derived to build functional fingerprints at multiple levels of detail.

The ultimate validation of the descriptors is presently performed in studies aimed at classifying pharmacological ligands. Such studies relating to inhibition of cytochromes P450 2D6 and 1A2 inhibitors include a comparison to well-established molecular descriptors (MDL MACCS keys) and are in the process of being published.[36,35]

Abbreviations: CP, critical point; CSD, Cambridge Structural Database; EDM, electron density map; EV, eigenvalue; HC, hierarchical clustering; ORCRIT, Oak Ridge critical point; PCA, principal component analysis; PCI, Physico-Chimie Informatique; QM, quantum mechanics; QSAR, quantitative structure−activity relationships; UC, unit cell; UPGMA, unweighted pair-group method with arithmetic mean; XRD, X-ray diffraction.

ACKNOWLEDGMENT

J.B. wishes to thank the Belgian “Fonds pour la formation à la Recherche dans l’Industrie et dans l’Agriculture” (F.R.I.A.) for his Ph.D. fellowship. N.M. thanks the “Fonds National pour la Recherche Scientifique” (F.R.S.-F.N.R.S.) for her Scientific Research Worker support. Calculations were performed on the ISCF (Interuniversity Scientific Computing Facility) computers with the financial support of F.N.R.S. and the Loterie National under the convention number 2.4578.02. All authors would like to thank Prof. Gerald M. Maggiora for fruitful discussions.

Supporting Information Available: Representation of the 77 substructures treated in this study. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Oprea, T. I.; Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* **2004**, *8*, 349–358.

(2) Oprea, T. I. *Chemoinformatics in drug discovery;* Wiley-VCH: Weinheim, 2005.

(3) Doweyko, A. M. QSAR: dead or alive. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 81–89.

(4) Bader, R. F. W. *Atoms in molecules - A quantum theory;* Clarendon Press: London, 1995.

(5) Popelier, P. L. A. Molecular similarity and complementarity based on the theory of atoms in molecules. In *Molecular similarity in drug design*; Dean, P. M., Ed.; Cambridge, 1995; pp 215−240.

(6) Popelier, P. L. A. Integration of atoms in molecules: A critical examination. *Mol. Phys.* **1996**, *87*, 1169–1187.

(7) Popelier, P. L. A. MORPHY, a program for an automated “atoms in molecules” analysis. *Comput. Phys. Commun.* **1996**, *93*, 212–240.

(8) Popelier, P. L. A.; Smith, P. J. QSAR models based on quantum topological molecular similarity. *Eur. J. Med. Chem.* **2006**, *41*, 862–873.

(9) Popelier, P. L. A. Quantum molecular similarity. 1. BCP space. *J. Phys. Chem.* **1999**, *103*, 2883–2890.

(10) Koch, U.; Popelier, P. L. A. Characterization of C-H-O hydrogen-bonds on the basis of the charge-density. *J. Phys. Chem.* **1995**, *99*, 9747–9754.

(11) Leherte, L.; Allen, F. H. Shape information from critical point analysis of calculated electron density maps: Application to DNA-drug systems. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 257–272.

(12) Becue, A.; Burton, J.; Dury, L.; Hansenne, C.; Larin, A. V.; Latour, T.; Leherte, L.; Meurice, N.; Petit, J.; Vercauteren, D. P. In silico molecular similarity and complementarity based on the electron density. *Chim. Nouv.* **2007**, *94*, 14–21.

(13) Becue, A.; Meurice, N.; Leherte, L.; Vercauteren, D. P. Description of protein-DNA complexes in terms of electron-density topological features. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2003**, *59*, 2150–2162.

(14) Becue, A.; Meurice, N.; Leherte, L.; Vercauteren, D. P. Evaluation of the protein solvent-accessible surface using reduced representations in terms of critical points of the electron density. *J. Comput. Chem.* **2004**, *25*, 1117–1126.

(15) Meurice, N.; Leherte, L.; Vercauteren, D. P. Comparison of benzodiazepine-like compounds using topological analysis and genetic algorithms. *SAR QSAR Environ. Res.* **1998**, *8*, 195–232.

(16) Leherte, L.; Meurice, N.; Vercauteren, D. P. Critical point representations of electron density maps for the comparison of benzodiazepine-type ligands. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 816–832.

(17) Leherte, L. Application of multiresolution analyses to electron density maps of small molecules: Critical point representations for molecular superposition. *J. Math. Chem.* **2001**, *29*, 47–83.

(18) Leherte, L.; Meurice, N.; Vercauteren, D. P. Influence of conformation on the representation of small flexible molecules at low resolution: Alignment of endothiapepsin ligands. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 525–549.

(19) Leherte, L.; Vercauteren, D. P. Critical point analysis of calculated electron density maps at medium resolution: Application to shape analysis of zeolite-like systems. *J. Mol. Model.* **1997**, *3*, 156–171.

(20) Cambridge Strucural Database (CSD). http://www.ccdc.cam.ac.uk/products/csd/ (accessed July 1, 2008).

(21) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380–388.

(22) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; MacRae, C. F.; MacCabe, P.; Pearson, J.; Taylor, R. New software for searching the cambridge structural database and visualizing crystal structures. *Acta Cryst. B* **2002**, *58*, 389–397.

(23) Hall, S. R.; du Boulay, D. J.; Olthof-Hazekamp, R. *Xtal 3.7 System*;University of Western Australia: Crawley, 2000.

(24) Biname, J.; Meurice, N.; Leherte, L.; Glasgow, J.; Fortier, S.; Vercauteren, D. P. Use of electron density critical points as chemical function-based reduced representations of pharmacological ligands. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1394–1401.

(25) Johnson, C. K. *The Oak Ridge critical point network program;* Oak Ridge National Laboratory: Oak Ridge, 1977.

(26) Holliday, J. D.; Willett, P. Using a genetic algorithm to identify common structural features in sets of ligands. *J. Mol. Graphics Modell.* **1997**, *15*, 221–232.

(27) Cosgrove, D. A.; Willett, P. SLASH: A program for analysing the functional groups in molecules. *J. Mol. Graphics Modell.* **1998**, *16*, 19–32.

(28) Sheridan, R. P.; Miller, M. D. A method for visualizing recurrent topological substructures in sets of active molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 915–924.

(29) Gancia, E.; Montana, J. G.; Manallack, D. T. Theoretical hydrogen bonding parameters for drug design. *J. Mol. Graphics Modell.* **2001**, *19*, 349–362.

(30) Ertl, P. Cheminformatics analysis of organic substituents: Identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.

(31) Aureus Pharma homepage. www.aureus-pharma.com (accessed July 1, 2008).

(32) Jolliffe, I. T. *Principal component analysis*; Springer: Berlin, 2002.

(33) Kogan, J.; Nicholas, C.; Teboulle, M. *Grouping multidimensional data: Recent advances in clustering*; Springer: Berlin, 2006.

(34) Dawyndt, P.; De Meyer, H.; De Baets, B. UPGMA clustering revisited: A weight-driven approach to transitive approximation. *Int. J. Approx. Reason.* **2006**, *42*, 174–191.

(35) Burton, J. Danloy, E. Vercauteren, D. P. Fragment-based prediction of cytochrome P450 2D6 and 1A2 inhibition by recursive partitioning *submitted to SAR & QSAR Environ. Res.*

(36) Burton, J. Danloy, E. Petit, J. Maggiora, G. M. Vercauteren, D. P. Rough set theory for the prediction of inhibitors of cytochrome P450 1A2 and 2D6 *to be submitted to Chemom. Intell. Lab. Syst.*

CI800148Z