Revisiting van der Waals Radii: From Comprehensive Structural Analysis to Knowledge-Based Classification of Interatomic Contacts

Ivan Yu. Chernyshov,* Ivan V. Ananyev,[b] and Evgeny A. Pidko*[a, c]

This paper is dedicated to W.D.S. Motherwell who showed the power of structural chemistry to the first author.

Weak noncovalent interactions are responsible for structure and properties of almost all supramolecular systems, such as nucleic acids, enzymes, and pharmaceutical crystals. However, the analysis of their significance and structural role is not straightforward and commonly requires model studies. Herein, we describe an efficient and universal approach for the analysis of noncovalent interactions and determination of van der Waals radii using the line-of-sight (LoS) concept. The LoS allows to unambiguously identify and classify the “direct” interatomic contacts in complex molecular systems. This approach not only provides an improved theoretical base to molecular “sizes” but also enables the quantitative analysis of specificity, anisotropy, and steric effects of intermolecular interactions.

1. Introduction

Van der Waals (vdW) radii ($R_{vdW}$) are one of the cornerstones of contemporary chemistry. The visual perception and basic analysis of chemical structures and architectures are strongly related to the concept of vdW radii, which provides the basic definition of the atomic “sizes”. Even though this concept has little physical grounding, it is easily interpretable and therefore widely used in the theoretical chemistry, particularly in design of force fields (in the Lennard-Jones parameter form),[1] dispersion corrections,[2] and COSMO-RS-type solvate models.[3] These models operate – directly or indirectly – by the intermolecular interatomic distances, which are regarded as corresponding to the minimum energy of atom-atom interactions, and are usually defined as a function of the sum of vdW radii. However, it has been noted[4, 5] that the sum of $R_{vdW}$ for available vdW sets including the most widely used one tabulated by Bondi[6] consistently underestimate the position of the energy minimum by as much as 0.3–0.4 Å. An illustrative example is a C6Cl6 crystal with the shortest r(Cl–Cl) = 3.65 Å that exceeds significantly the $2R_{vdW}^{Bondi}$(Cl) = 3.50 Å. The latter can only be achieved at a very high pressure of ca. 1.2 GPa.[7] Moreover, about 2.0% (>4000) of unique organic molecular crystals in the Cambridge Structural Database (CSD)[8] do not contain intermolecular contacts shorter than the sum of $R_{vdW}^{Bondi}$ (see the first section of Supporting Information for details). The question arises: why are all bonds in these crystals longer than the distance, which is usually considered as the most probable, in other words, closest to the energy minimum of the respective interatomic potential?

This inconsistency is rooted in the indirect nature of the approaches used to obtain the vdW parameters from experimental structural datasets. Conventionally, $R_{vdW}$ are derived from the position of the vdW peak in the distributions of contact distances between nonbonded atoms (Figure 1a,b).[6,9,10] Such distributions represent a superposition of a gaussian curve due to the vdW peak, and a rapidly growing function due to the randomly distributed contacts (Figure 1a). Because the position of the vdW peak maximum ($D_{max}$) is often hidden within the dataset for other contacts (Figure 1b), the sum of vdW radii is normally taken as the half-height of the vdW peak distance, $D_{half}$ instead of the $D_{max}$ (Figure 1a). However, $D_{max}$ has a clearer physical meaning as corresponding to the most probable distance for the considered interaction.[6] On the contrary, $D_{half}$-based vdW radii are qualitative, as the interpretation of their sum is not straightforward. It should be mentioned here that the relation between statistical distributions and energy surfaces is not straightforward and cannot be quantitatively determined without introducing arbitrary and unwarranted assumptions.[11] However, low-energy regions of molecular potential energy surfaces can be recognized and mapped from distributions of observed structures, thus $D_{max}$ can be...
considered an approximation of the “minimum energy position”.\(^{[4,11]}\)

Although alternative knowledge-based\(^{[12]}\) approaches to derive \(R_{vdW}\) were proposed, similar to the original Bondi’s strategy they fail to extract the most probable interatomic distances from the structural data. In addition to statistical approaches, several \textit{ab initio} computational ways to derive atomic sizes and \(R_{vdW}\) have been proposed.\(^{[13,14]}\) However, the real systems are much more complex than the computational models. As a result, such computational approaches provide only qualitative radii values and their use for the analysis of real extended systems therefore necessitates further investigations to ensure an adequate connection with the experimental results.

In this work, we introduce a new method to define the \(D_{\text{max}}\) and, accordingly, to compute \(D_{\text{max}}\)-based vdW radii (\(R_{\text{max}}\)) for almost any intermolecular interaction directly from diverse structural datasets. The key idea is to eliminate the background from randomly distributed contacts using the Line-of-Sight\(^{[15]}\) concept. It considers a pair of atoms as interacting only when they “see” each other (Figure 1d) because no other atom in the structure intrudes between them or, in other words, shields them from one another (for the exact definition see the SI). We identify three main types of contact shielding, namely, the covalent, intra- and intermolecular shielding depending on the type of shielding atom (Figure 1d), with only the covalent shielding having the characteristic distances comparable to those of the \(D_{\text{max}}\). A representative example of the contributions from the LoS and shielded contacts for the complete dataset of N···O contacts is illustrated in Figure 1c.

**2. Results and Discussion**

We propose to implement the Line-of-Sight\(^{[10]}\) concept (LoS) to eliminate the background from randomly distributed contacts, which is equivalent to classification of contacts into corresponding and noncorresponding to interatomic interactions. Thus, we start with understanding the physical meaning of this procedure. LoS concept considers the atoms as interacting only when they “see” each other (Figure 1d) because no other atom in the structure intrudes between them or, in other words, shields them from one another (for the exact definition see the SI). We identify three main types of contact shielding, namely, the covalent, intra- and intermolecular shielding depending on the type of shielding atom (Figure 1d), with only the covalent shielding having the characteristic distances comparable to those of the \(D_{\text{max}}\). A representative example of the contributions from the LoS and shielded contacts for the complete dataset of N···O contacts is illustrated in Figure 1c.
Figure 2. Dependence of NCI surface (isosurface value is set to 0.5) in the H$_2$N···H$_2$CF system on N···C angle and N···(H)C shielding.
The results in Table 1 reveal that the LoS model yields vdW radii consistently exceeding the values obtained by D$_{\text{eff}}$-based approaches by as much as 0.10–0.20 Å resulting in the increase of the sum of $R_{\text{vdW}}$ by 0.2–0.4 Å. This is attributed to the transition of the analysis from the rather arbitrary $D_{\text{eff}}$ parameters to more specific and rigorously defined $D_{\text{max}}$ values. We argue that $R_{\text{max}}$ are more physically sound as they directly reflect the most probable contact distance, whereas $R_{\text{eff}}$ appear to correspond to strongly shortened interactions, which are usually denoted as “specific” in the chemical literature.

Interestingly, the vdW radii estimates from quantum chemical calculations for isolated atoms ($R_{\text{QM}}$) by Rahm et al.$^{[14]}$ defined as the average distance from the nucleus to a point where the electron density falls to 0.001 a.u., also exceed $R_{\text{eff}}$ (Table 1). The comparison with the data obtained in this work shows that these theoretical $R_{\text{QM}}$ values cohere well with $R_{\text{max}}$ for C, N, O, and F, whereas they quite uniformly exceed $R_{\text{max}}$ for S, Se, Cl, Br and I by 0.15–0.20 Å. This implies that the electron density parameters for weak interactions at the most probable distance depend only weakly on the interacting elements.

Furthermore, our data analysis allows identifying the contact-angle dependencies in vdW radii. A representative example is the interhalogen interactions C–Hal$_1$–Hal$_2$–C, Hal–Cl, Br, I (X-bonds, XBs), which are usually classified as types I and II.$^{[19]}$ Type I XBs are characterized by /C–Hal$_1$–Hal$_2$–C$^\perp$, and are usually nonspecific and weak, whereas type II XBs are characterized by /C–Hal$_1$–Hal$_2$–C$^\perp$ and /C–Hal$_1$–Hal$_2$–C$^\perp$ and are usually strong and shorter than type I XBs due to the $\sigma$-hole interaction.$^{[20]}$ Therefore, our approach can specifically be used to discriminate the respective noncovalent interactions and obtain different $R_{\text{max}}$(Hal) values for 90° and 180° C–Hal–X angles (see for example Figures S3–27–29, 49–51, 64–66, 73–75). It should also be noted, that different $R_{\text{max}}$ values were obtained for H atoms from H–H and H–X, X ≠ H contacts (1.21 Å and 1.29 Å, respectively). This should be considered when analyzing interactions between aliphatic tails.

It should be noted, that $R_{\text{max}}$ are in fact a condensed representation of the analyzed contacts and therefore can be instrumental for the more detailed analysis of noncovalent interactions and statistical definition of qualitative structural features of chemical systems. For example, specific interactions and sterically hindered contacts can be automatically identified by $D_{\text{max}}$, substantially deviating from the sum of the respective vdW radii. Let us consider two representative examples of CH···O and C···O contacts, for which the $D_{\text{max}}$ are, respectively, shorter or longer than the sum of $R_{\text{max}}$.

CH···O contacts show $D_{\text{max}}$ values shorter by 0.14 Å than the sum of $R_{\text{max}}$. This deviation is much larger than that (0.02 Å) detected for the related CH–N and CH–F interactions (Table S1). The CH···O contacts are shorter by 0.18 Å than CH–N and very similar to CH–F (Figure 3a), although an opposite trend can be seen in the C$_{\text{sp}2}$···O/N/F contacts (Figure 3b). Such deviation of $D_{\text{max}}$ from the sum of $R_{\text{max}}$ implies that CH···O contacts are highly specific and their structure-forming role is more significant than that of the other CH···X contacts, including CH–F, despite the similar electrostatic nature of these contact types.$^{[22]}$

The C$_{\text{sp}3}$···O contacts formed by quaternary carbon exceed by 0.30 Å the corresponding $R_{\text{max}}$ sum, which, in turn, cohere well with the $D_{\text{max}}$ values for the C···O contacts formed by primary (C$_{\text{sp}3}$) and tertiary (C$_{\text{sp}3}$) carbon atoms (Figure 4). Such a shift of the C$_{\text{sp}3}$···O vdW peak is in line with the expected significant sterical repulsions between the neighbors of the interacting carbon atom and the oxygen atom and indicates significant steric hindrance of C$_{\text{sp}3}$ atoms. This means that the concept of vdW radii should be applied with a special care when dealing with the atoms in confined tetrahedral or octahedral environments (e.g. Si, Bi, most of d, f-elements, and other elements considered by Alvarez in Ref. [10]), which would show an increased effective size due to the unaccounted steric

### Table 1. Van der Waals radii of elements typical for organic compounds.

<table>
<thead>
<tr>
<th>Atom</th>
<th>Van der Waals radii [Å]</th>
<th>Atom or contact type$^{[16]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_{\text{eff}}$</td>
<td>$R_{\text{QM}}$</td>
</tr>
<tr>
<td></td>
<td>Bondi$^{[1]}$</td>
<td>R&amp;J$^{[2]}$</td>
</tr>
<tr>
<td>H</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>C</td>
<td>1.20</td>
<td>–</td>
</tr>
<tr>
<td>N</td>
<td>1.55</td>
<td>1.56</td>
</tr>
<tr>
<td>O</td>
<td>1.52</td>
<td>1.56</td>
</tr>
<tr>
<td>F</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>S</td>
<td>1.47</td>
<td>1.44</td>
</tr>
<tr>
<td>Cl</td>
<td>1.75</td>
<td>1.74</td>
</tr>
<tr>
<td>Se</td>
<td>1.90</td>
<td>–</td>
</tr>
<tr>
<td>Br</td>
<td>1.85</td>
<td>1.85</td>
</tr>
<tr>
<td>I</td>
<td>1.98</td>
<td>2.00</td>
</tr>
</tbody>
</table>

[a] Atom types used for determination of van der Waals radii in this work. If there are several radii for one element, the first row lists ‘default’ $R_{\text{eff}}$ for the current element that can be compared with the values from other vdW radii sets. R and Z in formulae stands for C-bonded and any monovalent substituents, respectively. [b] Water molecules were excluded as $D_{\text{max}}$ of the respective interactions were systematically different from those for ROH/R$_2$O. [c] Carboxylates and charged atoms were excluded from the datasets as the $D_{\text{max}}$ of the corresponding interactions were systematically different from those for uncharged atoms, which is apparently due to the increased electrostatic contribution.
effects. The corresponding $D_{\text{max}}$ values will merely indicate the position of the first maximum of the radial distribution function and make little sense in terms of noncovalent interactions. Only the atoms in an “open” environment such as trigonal, square planar or square pyramidal configurations can be used to determine $R_{vdW}$ from contact distance distributions without shielding effects to be crucial. Even for the extended dataset considered herein these conditions were fully satisfied only for the 10 elements, which data are summarized in Table 1.

These examples illustrate that despite all advantages, the element-defined $R_{\text{max}}$ radii still fail to generally define the most probable distances of intermolecular interactions. The preferred and more accurate approach is to directly utilize $D_{\text{max}}$ values determined for the given atom pair contact rather than the sum of $vdW$ radii. In this way, the structural analysis will automatically account for the specificity of the analyzed interactions, their anisotropy, and the impact of steric effects. We have developed an algorithm and implemented it in a script (deposited at the GitHub\textsuperscript{23}) that allows searching the CSD for intermolecular contacts with certain geometry (ConQuest output) with their subsequent classification as line-of-sight or shielded contacts (Figure 5). The produced datasets can readily be used for the determination of the $D_{\text{max}}$ using any standard table processing or data analysis software. Such $D_{\text{max}}$ values obtained for specific contacts with a certain geometry can be used to create high-throughput geometry-based descriptors of
3. Conclusions

In summary, a new approach for automated analysis and classification of noncovalent interactions has been developed. The key feature of the reported methodology is the filtration of the direct interatomic contacts from the other background structural features based on the LoS concept. The validity of this approach has been confirmed by density functional theory calculations. The LoS concept was used to recalculate $R_{max}$ for light elements, which turned out to be substantially larger than those currently accepted by the chemistry community. Our analysis reveals that the underestimations of the “atomic sizes” noted earlier for most tabulated vdW radii ($R_{max}$) is solely related to the deficiencies of the datasets used for their determination. The $R_{max}$ introduced here are free from statistical bias and are based on clear physical grounds. We anticipate the high practical utility of $R_{max}$ and particularly, their atom type-specific variations for improving of molecular sizes in different chemical methods and approaches, and for the analysis of intermolecular interactions on a wide variety of systems including the experimental structural databases and the results of theoretical calculations on molecular and condensed systems relevant to various fields of chemistry and material sciences. The current LoS approach implemented now in the relevant software enables the quantitative analysis of specificity, anisotropy and steric effects of intermolecular interactions while benchmarking databases as well as studying specific systems.

Computational Methods

Version 5.39 with 4 updates (up to August 2018) of CSD\textsuperscript{[40]} was used for selection of organic crystals containing H(D), B, C, N, O, F, Si, P, S, Cl, As, Se, Br and I atoms. Disordered, erroneous, polymeric, pressurized, powder structures and experiments with R-factor $>0.075$ were removed from consideration. 224001 selected crystals were used for search of unique intermolecular contacts A–B with distance $D(A\ldots B)$ up to 7.0 Å. C–H, N–H and O–H bond lengths were normalized to CCDC/ConQuest defaults: C–H: 1.089 Å, N–H: 1.015 Å, O–H: 0.993 Å. Acetylenic $C_\equiv H$ bond lengths were normalized to 1.06 Å (neutron diffraction: ACETYL05, RALDEN01, XEHLEB, ZULDEP01), and S–H bond lengths were normalized to 1.34 Å (microwave data: H$_2$S\textsuperscript{[28]} neutron diffraction: NALCSY02). For each contact the following information were collected: (1) contact geometry, (2) chemical nature of contact atoms, (3) shielding atom, and (4) contact shielding value. A total of c.a. 640 000 000 contacts were found, more than 40 000 000 of which was LoS depending on used vdW radii (41 346 551 for the final version from Table 1). These data were used further to build and analyze the distribution of distances of various contacts and is available for download\textsuperscript{[28]}. Obtained line-of-sight A–B contacts were used to plot histograms of contact distance distributions for all possible A, B atom type pairs (Table S1). Primary analysis of these distributions combined with chemical common sense allowed to select atom types and, therefore, contacts used for the van der Waals radii determination. The main requirements were: (1) there are enough contacts to determine $D_{max}$ so that line-of-sight contacts distribution contains a line-of-sight peak that is similar to the Gaussian function; (2) contact corresponds to non-specific interaction; (3) influence of steric effects on A–B distance can be excluded; (4) addition of the contact to the list of contacts used for van der Waals radii
determination does not change any radius significantly (> 0.05 Å). This selection is subjective to a certain degree, however, it’s well suited for our final goal which was to obtain \( R_{\text{emp}} \) corresponding to weak nonspecific interactions that are not affected by any effects. Selected contact types were used for \( R_{\text{max}} \) determination and are marked with ‘+’ sign in the second column of Table S1. Using the data of Table S1 together with the data on intermolecular contacts,\(^{30} \) allows one to reproduce this work, or to choose another set of contact types to determine another version of \( R_{\text{max}} \).

Quantum chemical calculations were performed with the Gaussian 16 rev. B.01\(^{29} \) program at the B3LYP/6-311 + + G(d,p) level of theory. The NCIS analysis was performed with the Multiwfn package.\(^{30} \) Geometry of the \( \text{H}_2\text{N} \cdots \text{CH}_2\text{F} \) complex was obtained by merging optimized \( \text{NH}_3 \) and \( \text{CH}_2\text{F} \) molecules into the staggered \( C_2v \) structure with \( D(\text{N} \cdots \text{C}) = 3.0 \) Å. The \( \text{N} \cdots \text{F} \) angle was changed so that the \( \text{H}_2\text{N} \cdots \text{CH}_2\text{F} \) complex retains the \( \text{N} \cdots (\text{H})\text{C} \) symmetry plane. NCI surface (0.5 isosurface value) dependence on \( \text{N} \cdots \text{F} \) angle remains unchanged with \( D(\text{N} \cdots \text{C}) \) increasing.

Acknowledgements

This project has received no external funding and was driven by pure love to science and curiosity. SurfSARA and NWO are acknowledged for providing access to supercomputer resources.

Keywords: CSD analysis • intermolecular contacts • intermolecular interactions • molecular crystals • Van der Waals radii

[28] https://doi.org/10.4121/uuid:33c65813-e9ef-41f1-910c-660d18ecf097
Seeing trees behind the forest! A method based on the line-of-sight concept to efficiently eliminate the background from randomly distributed contacts and analyze intermolecular interactions within diverse extended structural datasets is presented. This approach allowed to more specifically define the vdW parameters from the bulk structural data and to identify atom-type specific vdW radii with clear physical meaning.