DETERMINATION OF ATOMIC RADII FROM SMALL-MOLECULE CRYSTAL STRUCTURES

Master thesis

Systems Biology study programme

Vilnius University

STUDENT NAME:	Eglė Šidlauskaitė
STUDENT NUMBER:	1174420
SUPERVISOR:	Andrius Merkys
SUPERVISOR DECISION:	....
FINAL GRADE	....
DATE OF SUBMISSION:	15 May 2023

CONTENTS

1 LIST OF ABBREVIATIONS

BIC	Bayesian Information Criterion
CCDC	Cambridge Crystallographic Data Centre
CGI	Common Gateway Interface
CIF	Crystallographic Information File
CLPOs	Chemist’s Localized Property-Optimized Orbitals
CML	Chemical Markup Language
COD	Crystallography Open Database
CSD	Cambridge Structural Database
IUCr	International Union of Crystallography
SDF	Structural Data File
SMILES	Simplified Molecular Input Line Entry System
TSV	Tab-Separated Values
YAML	Yet Another Markup Language

2 INTRODUCTION

X-ray crystallography is an established method for determination of the exact atom positions in crystal structures, however, it is unable to capture chemical bonding. A heuristic widely used to determine if two atoms are connected by a chemical bond states that the distance between two chemically bonded atoms has to be smaller or equal to the sum of their atomic radii (Allen et al., 1979). Therefore, atomic radii tables have an important role in materials science, chemistry and similar fields of research.

A univocal method for atomic radii derivation is not yet determined which is evident from relatively large differences between results provided in different covalent radii tables published by Meng and Lewis (1991), Cordero et al. (2008) and Pyykkö and Atsumi (2009). Furthermore, the most commonly used atomic radii tables are derived using data from the Cambridge Structural Database (CSD) (Groom et al., 2016) which is distributed under a proprietary license and therefore restricts the usage and spread of such derivative datasets. This suggests a clear need to develop a completely independent workflow which would automatically calculate atomic radii tables using open crystallographic datasets such as the Crystallography Open Database (COD) (Gražulis et al., 2012).

This research aims to identify the van der Waals gap in the distributions of distances between atoms for each pair of elements. The van der Waals gap corresponds to the lowest density region between two density peaks, where the first one represents intramolecular bond lengths and the second represents distances between atoms affected by the van der Waals forces. This approach can be used to identify typical intramolecular bond lengths out of which the atomic radii table can be derived. The validity of the calculated atomic radii tables is tested against a dataset of small molecule chemical structures with known bonding.

All calculations are performed using an automated *Makefile* system. An automated workflow ensures that the provenance of the study is preserved. Furthermore, recomputations can be easily performed as more data becomes available in the future.

One of the goals of this study is to make the results of the calculations as well as the intermediate data freely accessible online. This goal is achieved in the form of an interactive website which can be used to review the final derived atomic radii table and compare it to other published radii tables.

3 AIM AND TASKS

The aim of this study is to derive a free and open atomic radii table using an automated and unsupervised methodology consisting of as few as possible assumptions. The main tasks of this project are:

1. Develop a system that can automatically identify crystal structures with features unsuitable for this study and filter these structures out.

2. Obtain interatomic distance data from the structures published in the COD.

3. Develop a system for automatic and unsupervised intramolecular bond length determination for each interatomic distance class.

4. Derive atomic radii from determined intramolecular bond length data.

5. Validate the derived table by using it to identify bonds in structures with known bonding data.

6. Create a website in which results of this study can be accessed and compared to previously published atomic radii tables.

4 LITERATURE REVIEW

4.1 Small molecule crystal structures

Small molecules are defined as inorganic, organometallic or non-polymeric organic molecules with low molecular weight (Li and Kang, 2020). Due to this characteristic, small molecules retain the ability to penetrate walls of cells and blood-organ barriers, thus, small organic molecules are used for novel drug discovery and development. Drug discovery requires thorough analyses of the structure of small molecules and their interactions with other molecules.

X-ray crystallography is a method universally used to determine the structure of small molecule crystals (Le Pevelen, 2010). Using this technique, information about the density of electrons in molecule crystals can be determined, which also reveals the three-dimensional structure of the molecules, including the exact positions of the atoms that constitute them. This experimental technique requires a beam of X-rays to be directed towards the crystal and resulting diffraction to be recorded.

The interaction between X-rays and the molecule results in X-ray diffraction (Stanjek and Äusler, 2004). X-ray diffraction occurs when an electron is located in the path of the X-ray beam. The diffracted beams are measured in terms of their angles and intensities. This data are later used to determine detailed information about the structure of the crystal.

The process of crystal structure determination has to be very precise and errors can occur in any step of this process (Kleywegt, 2000). Issues in this process may severely impact the results of further studies. The multipurpose tool `Platon` is widely used for crystal structure validation (Spek, 2003, 2009).

The unit cell is the smallest periodically repeating unit of a crystal (Massa, 2004). Unit cell is further reduced to an asymmetric unit by eliminating crystal symmetry. There are 230 distinct space groups that describe all possible combinations of symmetry operations in the three-dimensional space (Koster, 1957). Some structures are described using superspace groups which contain symmetry operations that are not eligible for transformations in a three-dimensional space (IUCr, 2006). Although such symmetry operations could be projected to the three-dimensional space group operations in order to restore the structure of the crystal, resulting positions of atoms and, therefore, distances between them are likely to be distorted.

In some cases the structure identified through X-ray crystallography may have missing atoms or parts of the structure. These issues may be caused by omission of profoundly disordered parts of the crystal (which may represent solvent molecules) from its model or errors such as incorrectly noted symmetry space group (Merkys, 2018). The fact that the structure has missing atoms can be identified by comparing the declared chemical formula to the chemical formula calculated from the unit cell of the crystal (Mounet et al., 2018). Additionally, unnatural voids in crystals can be determined by reconstructing the three-dimensional structure of the crystal and identifying the voids between molecules that are big enough to fit another molecule.

X-ray diffraction primarily measures the electron density and works well with electron-heavy atoms. The hydrogen atom has only one electron which scatters the radiation weakly and provides too little information for its position to be identified directly from experimental data. However, correct identification of hydrogen atom positions is vital to crystal structure determination as elimination of these atoms may cause shifts in the positions of the atoms that they are bonded (Harlow, 1996). Hence, hydrogen atom positions in a molecule crystal are usually interred from geometry and may not be exact. Although recent studies have stated that Hirshfeld atom refinement may be used to enable X-ray crystallography techniques to identify hydrogen atoms (Woińska et al., 2016), such approach is relatively new and not widely used yet. Therefore, it can be expected that in most of available crystallographic structures hydrogen atoms may not be located accurately.

In conclusion, X-ray crystallography is a method that allows to determine structural information of the small molecule crystal. However, there are certain limitations and error prone features of this method. These features need to be considered in order to avoid issues related to result accuracy in further studies.

4.2 Bonding in small molecule crystals

Although X-ray crystallography can reveal a lot of valuable information about the structure of small molecule crystals such as exact positions of the atoms that constitute them (Gražulis et al., 2015), it is not able to capture interactions between atoms. Interactions that occur in crystals can be differentiated into bonding (or intramolecular interactions) and non-bonding (or intermolecular interactions) (Williams, 2021).

Chemical bonding refers to strong attraction between atoms that allows molecules to retain their shape and structure (Atkins, 2023). These interactions result in relatively short interatomic distances. The main types of chemical bonding are ionic, metallic and covalent bonding (shown in Figure 1).

An ionic bond is formed when valence electrons are transferred from one atom to another. The atom which receives electrons gains a negative charge and becomes an anion. The atom which donates electrons becomes a positively charged cation. Met-

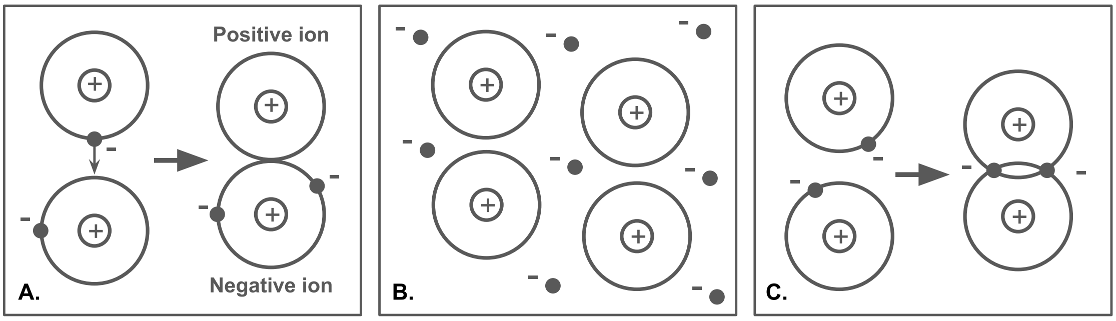


Figure 1. Visualization of chemical bonding types: A) ionic bonding; B) metallic bonding, C) covalent bonding.

als often become electron donors as they have few valence electrons. Consequently, nonmetals are usually the receptors of electrons because this way their valence orbital can be filled and, after the transfer, both ions obtain a stable electron configuration. The bond is formed due to the attraction between the oppositely charged ions.

In metals, atoms are closely packed together. Due to the close proximity between atoms, their valence shells overlap, enabling valence electrons to freely migrate between multiple atoms. As the electrons are not assigned to any specific atom, the atoms become cations which in turn are bound to the surrounding cloud of valence electrons. This type of bonding is called metallic bonding.

When two atoms in a stable molecule form an interatomic linkage by sharing an electron pair, the resulting bond is called a covalent bond (Bacskay et al., 1997). For a covalent bond to form between two atoms, their electronegativity has to be comparable, therefore covalent bonds are often formed between atoms of the same element. Furthermore, covalent bonds are rarely formed between metals as these atoms tend to have a low number of valence electrons and low ionization energy. Therefore, such elements are more prone to donating electrons than attracting them and the attraction is required for the covalent bond to form. Covalent bonds are induced by the electrostatic attraction between positively charged nuclei and negatively charged shared electrons.

In order to determine that there is a bond between two atoms, the distance between the atoms can be compared to the sum of their atomic radii. Each element in the periodic table has specific properties of the atom that represent the length of its radius in each type of bond that the element can form, such as covalent, ionic and metallic radii. When precise bonding data is unknown, a heuristic can be applied which states that two atoms are concluded to be connected by a chemical bond if the distance between them is smaller than or equal to the sum of the atomic radii of these atoms (Daintith, 2008).

Packing of molecules in a crystal is influenced by intermolecular forces. The van der Waals force is a weak intermolecular and distance-dependent interaction that can occur between atoms or molecules (Israelachvili, 1974). For any pair of elements, distances between bonded atoms are shorter than distances between atoms affected mainly by the van der Waals forces. A gap between these interactions results in an unpopulated range in the interatomic distance distribution which is called the van der Waals gap (Alvarez, 2013).

The paper published by Alvarez (2013) introduces a new method for calculation of the van der Waals radii and presents an interatomic distance distribution model. In this publication, interatomic distance distribution is described to consist of a chemical bond peak followed by a van der Waals gap, van der Waals interactions peak and random distribution (as shown in Figure 2). Random distribution is an exponential component that contains distances between atoms induced by weak intramolecular interactions and noise. This interatomic distance distribution model could be used to determine chemical bond lengths for interatomic distance classes based on the identification of the van der Waals gap.

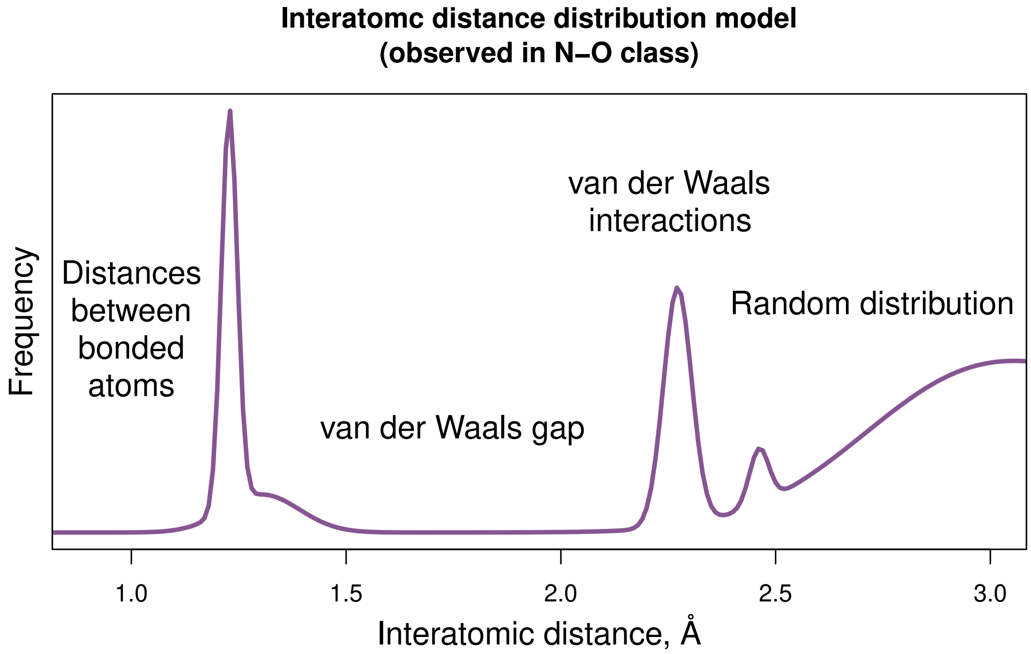


Figure 2. Interatomic distance distribution in N–O interatomic distance class generated using data from the COD. Four main features of the distribution model are shown – peak that represents distances between chemically bonded atoms, van der Waals gap, van der Waals interactions peak and random distribution. The random distribution in this case is not exponential due to the application of Voronoi cells.

Voronoi tessellation can be used to reduce the random component observed in the interatomic distance distribution by limiting iterations to direct neighbors only.

A Voronoi cell is a region of space that contains all space points closest to the atom (Olechnovič and Venclovas, 2014). This approach can help limit the number of identified interatomic interactions that belong to the random distribution by recording only the interactions between atoms in neighboring cells (as shown in Figure 3).

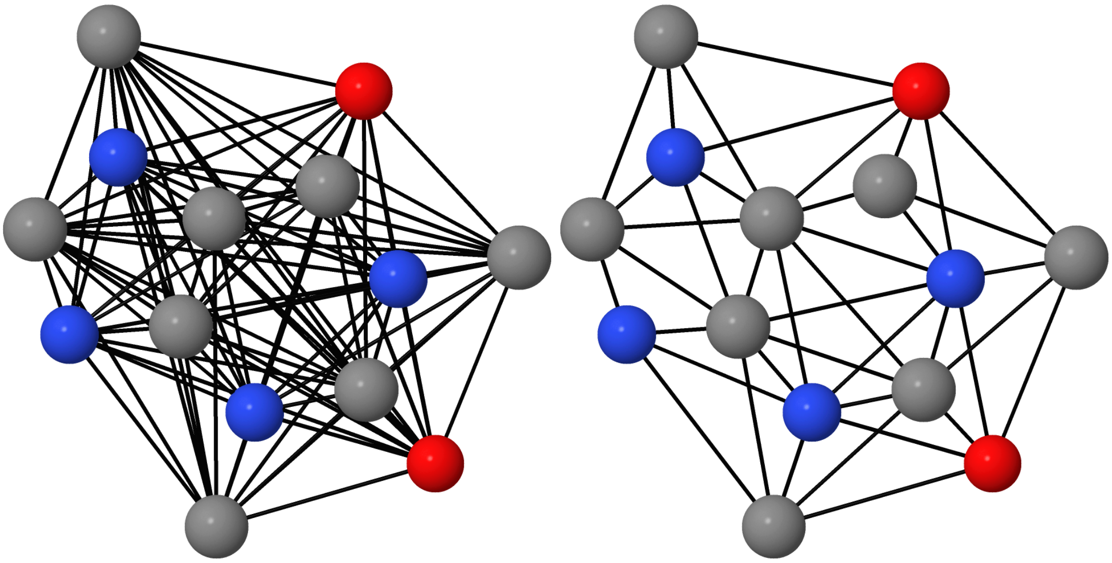


Figure 3. Visualization of Voronoi cell application to minimize the number of incorrectly identified interatomic distances. The graph on the left side shows all interatomic distances in a caffeine molecule. On the right side, interatomic distances are noted for atoms between neighboring Voronoi cells only.

The study by Alvarez (2013) also postulates that the interatomic distance distribution model can be approximated by a Gaussian mixture model. Different applications of this approach to covalent bond determination were tested and described in bachelor's thesis by Šidlauskaitė (2021), which discusses five different methodological approaches for the identification of the van der Waals gap. In each case, the Gaussian mixture model is fitted to the distribution of distances between atoms in each element pair class. The identification of the lowest density region which corresponds to the van der Waals gap is performed using the simplex method (Nelder and Mead, 1965). The results obtained from this study were tested by comparing connectivity of molecules using standardized simplified molecular-input line-entry system (SMILES) nomenclature (Weininger, 1988). Two sets of SMILES strings were generated: one using derived covalent radii and one using covalent radii published by Meng and Lewis (1991). The comparison revealed that in 8 out of 9 cases the generated SMILES strings matched which suggests that the approach used in this research is promising.

Atomic radii are useful in a lot of different fields of research related to materials science or chemistry (Zhang, 2018). Therefore, a number of covalent radii tables have been published. Although covalent radii determination is always based on the notion that two atoms are bonded if the distance between them is less than or equal to the sum of their covalent radii, a univocal method for covalent radii derivation is not yet determined. This is evident from different methodologies used in publications and significant differences between results provided in different covalent radii tables (as shown in Figure 4).

In conclusion, atomic radii are useful for determination of chemical bonding. Therefore, atomic radii tables are useful in a number of fields of scientific research and availability of precise atomic radii tables is of high importance. In organic chemistry, covalent bonds are formed substantially more often than ionic or metallic bonds. Although there is no univocal method to determine covalent radii, the interatomic distance distribution model published by Alvarez (2013) suggests an applicable approach for the determination of van der Waals gap and chemical bond length.

4.3 Published covalent radii tables

One of the first covalent radii tables was published by Meng and Lewis (1991). In this study, a geometry-based algorithm is used to analyze the three-dimensional structure of molecules in order to determine connectivity and the hybridization state of the atoms. The algorithm analyzes coordinates of heavy atoms and utilizes bond lengths in organic compounds published by Allen et al. (1987) that were derived using data from the September 1985 version of the CSD. The authors of this study include a covalent radii table which is used to determine atom connectivity in the molecules and is referenced to be sourced from Allen et al. (1987). However, the exact method used to derive these radii is not explained in either of the publications.

Although Meng and Lewis's publication demonstrates that this covalent radii table can successfully be used to determine atom connectivity in molecules, the ambiguity left behind the actual methodology used to derive this table suggests that these results are lacking a sufficiently detailed provenance record. Furthermore, it is important to note, that this covalent radii table depends on data supplied by CSD which is not an open-access database, therefore the results are not freely reusable. Based on the comparison with other published covalent radii tables, in some cases radii from this table differ from other tables quite significantly and do not follow the same trends based on atomic number variation (as shown in Figure 4).

The study published in 1995 by Batsanov (1995) is an example of a straightforward application of the covalent radii definition to the covalent radii determination methodology. In this paper, bond lengths in homonuclear diatomic molecules are measured

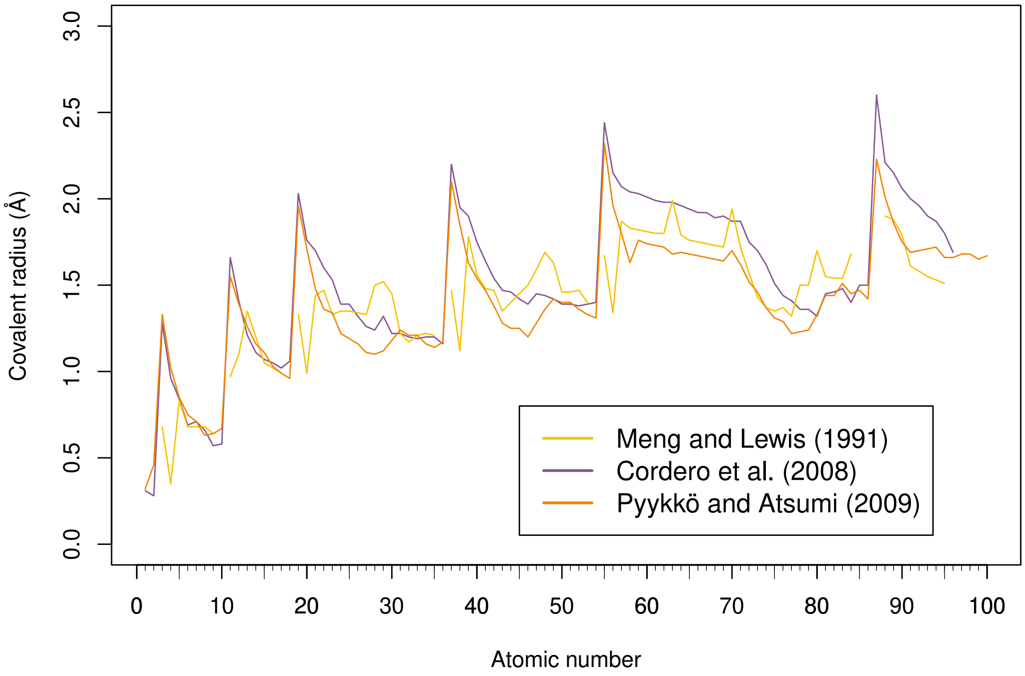


Figure 4. Comparison of covalent radii in published covalent radii tables.

experimentally and divided by two to determine the covalent radius of the atom. To determine bond lengths between atoms, simple gaseous molecules were analyzed using spectroscopy and simple solid substances were analyzed using X-ray studies. Additionally, homonuclear bond lengths were measured in compounds with organic or inorganic radicals, taking in consideration the order of the bond.

Batsanov’s study makes it very clear that covalent bond lengths may vary based on the type of measurement and state of the matter. Furthermore, the results consistently display that covalent bond is longer in single bonded atoms compared to double bonded atoms and double bonded atoms compared to triple bonded atoms. However, in this study only homonuclear bonds are analyzed and the sample size of the structures analyzed is not clear. The author also mentions that the system described in the study is intended for estimation of the interatomic distances which is sufficient to determine chemical bonding rather than determination of exact covalent radii values.

Publication by Cordero et al. (2008) describes a method of covalent radii determination that aims to derive radii for main organic elements that often form covalent bonds first and calculate covalent radii for other elements based on these values. First of all, covalent radii for nitrogen, carbon and oxygen are derived based on average bond lengths in N–N, C–C, C–N and C–O interatomic distance classes. The covalent radii of remaining elements are derived by calculating the average bond length of interatomic distance classes where one of the primary elements participates in the bond.

The results of this study show clear periodic trends of covalent radii (as shown in Figure 4). However, as the table was derived based on radii of only three selected organic elements, this study may be affected by selection bias. Furthermore, the data used in this study was taken from the CSD, resulting in the study not being freely reusable.

A different approach is taken in research for single-bond covalent radii published by Pyykkö and Atsumi (2008). In this case, all elements are treated as equal and the covalent radius of each element is derived independently using least-squares fit. Least-squares fit aims to minimize the squares of the differences between data points and predicted values (Björck, 1990). In order to refine experimental data, the study also used density functional theory (Orio et al., 2009) calculations. Furthermore, for elements for which no experimental data were available, authors predicted their covalent radii based on correlation between other determined radii and physical properties of the atoms.

The study used experimental and computed bond lengths, however, the authors do not specify the source of this data and the exact sample size. Additionally, in some cases, covalent radii had to be predicted instead of determined through calculations due to the lack of data. Furthermore, the calculated covalent radii for elements with atomic numbers between 55 and 80 differ quite significantly from values published in other covalent radii tables (as shown in Figure 4).

Another way to generate new covalent radii tables is by aggregating results from a few different sources. An example of such an approach is the covalent radii table published by Guha and coauthors in 2006 which is a part of the Blue Obelisk Data Repository (Guha et al., 2006). It is known that this table is heavily based on Pyykkö and Atsumi (2008), however it is not clear what methods were applied to make alterations to the data.

The Blue Obelisk Data Repository contains a wide assortment of chemical element features, covalent radii being one of them. Since the repository is open access, it is a great source of chemoinformatics data which can even be used by software developers to make their software interoperable with software developed by other authors. However, the covalent radii data is in principle taken from studies based on data from the CSD and there is a lack of details about the methodology used to derive or collect it.

Arguably the most popular covalent radii table is published by the CSD (CCDC, 2010). This table was not published in a separate paper – it is an assembly of data published by multiple authors and can be accessed as a component of the CSD itself. Thus, this data is distributed under a proprietary license which restricts the usage and spread of scientific information.

The main criticism for the CSD covalent radii is very well described in Cordero et al. (2008) – over the years these covalent radii values were manually amended to conform to connectivity of specific structures. Furthermore, elements for which the covalent radii could not be derived were assigned an arbitrary value of 1.50 Å. Therefore, the limitations of this covalent radii table are not only related to the usage of the proprietary license, but are also influenced by the limitations of the methodology used to define the values.

In conclusion, although there are a number of already published covalent radii tables, each of them has limitations. The most common limitation observed is the lack of availability for these results to be reproduced due to usage of proprietary data. Additionally, in some cases, the details behind the methodology used for covalent radii derivation are not clearly explained. Therefore, there is a clear need to develop an automated and unsupervised workflow that allows covalent radii tables to be automatically calculated using open crystallographic datasets such as the COD.

4.3.1 The most recent studies

One of the newest publications in the topic of covalent bond length determination was published by Nikolaienko et al. (2019). In this paper, density functional theory (Orio et al., 2009) is used for molecule geometry optimization. Information about molecule connectivity and covalent bond lengths is derived using chemist's localized property-optimized orbitals (CLPOs) analysis (Nikolaienko and Bulavin, 2019). CLPOs are obtained by identifying the precise localized representation of the first-order reduced density matrix (which describes one-electron features of the system) and introducing chemically motivated constraints in the process.

This methodology for covalent bond derivation relies on results of quantum-chemical computations instead of experimental data and involves only one adjustable parameter – threshold for bond ionicity. Therefore, the study presents a mostly automated approach to the task which allows to derive a non-empirical dataset of covalent bond lengths. Results of the study provide distances between atom pairs in molecules consisting of less than 12 atoms of elements that are metalloids and reactive nonmetals, excluding those that belong to group five of the periodic table.

A further study by Nikolaienko et al. (2020) has used this dataset for covalent radii derivation. In this study the authors state that in some structures the observed distribution of interatomic distances displays multimodal shapes that may not possess Gaussian distribution. Authors argue that in such cases least-squares approach (Björck, 1990) and maximum likelihood criteria (Leytham, 1984) which are commonly used in covalent radii derivation may not produce suitable results. Authors postulate that removal of conjugated bonds from the dataset creates a subset of data with distribution closer to normal distribution which is then used to derive first-principle covalent radii. Additionally, authors use an extension of the covalent radii definition by introducing the electronegativity difference into the equation of covalent bond length (O'Keefe and Brese, 1991).

The study states that derived covalent radii align well with corresponding values published by Pyykkö (2015). Furthermore, the approach is rather unique due to calculations being performed automatically and not requiring human-curated data processing. However, the study only considers 14 non-metallic elements. Therefore, a full covalent radii table is not available and it is not clear if the methodology presented would perform well with other periodic element families.

In the last decade, machine learning algorithms have been successfully applied to perform complicated computational tasks in a variety of fields of research, including chemistry (Goh et al., 2017). One of the most recent studies aiming to determine covalent radii was published by Nikolaienko and Bulavin (2021) and uses an approach based on machine-learning. In this study, a machine-learning binary classification model was used to identify bonded and non-bonded atom pairs. The model was trained using a support vector machine (Pisner and Schnyer, 2020) and two separate datasets with bond lengths of atom pairs belonging to each of the classes. The model was able to determine properties of individual chemical elements constituting chemical bonds, including covalent radii.

This study is a good demonstration of possible artificial intelligence applications in computational chemistry. However, it was only able to determine covalent radii for metalloids and reactive nonmetals, except for elements that belong to the periodic table group five. Additionally, stringent limitations were applied for the selection of molecules that can be analyzed, for example, only molecules containing up to 12 non-metallic chemical element atoms were considered; therefore, the results may be affected by selection bias.

In conclusion, it is evident that covalent radii determination is still a topical subject in the science community. The main goals of current studies are to achieve automation of covalent radii derivation process and to reduce the need of human-curation of the derived datasets. Additionally, an important target of these studies is to create a non-empirical approach of covalent radii derivation.

5 METHODS

5.1 Data

For this research 484 652 crystal structures described in crystallographic information file (CIF) format (Hall et al., 1991; Bernstein et al., 2016) and deposited in the COD[1] (revision 272900) were analyzed. The crystal structures described in these data files were restored to 3 x 3 x 3 supercells using tool `cif_fillcell` from *cod-tools*[2] (revision 9473) (Merkys et al., 2016) package. In order to determine distances between atoms in each crystal, tools `cif_contacts` from *crystal-contacts*[3] (revision 51) package and `cif_bonds_angles` from *atomclasses*[4] (revision 584) package were used. The tool `cif_contacts` is based on a software tool *voronota v1.18.1877*[5] (Olechnovič and Venclovas, 2014) and is used to enclose each atom in a Voronoi cell in order to take into account the contacts between atoms in neighboring cells only, reducing the number of long distance and obstructed interactions between atoms. The tool `cif_bonds_angles` is used to measure the distances between the selected atoms. Interatomic distances that involve hydrogen atoms were omitted from further analysis as the coordinates of these atoms usually are lower in quality (previously discussed in Section 4.1).

Interatomic distances were successfully measured in 358 626 CIF files. Further investigation revealed that for the remaining 26% of crystal structures distances between atoms could not be measured reliably due to marked and unmarked disorders. Distances between atoms were sorted into 3007 interatomic distance classes based on the chemical elements between which the distance was calculated. This data is stored in tab-separated values (TSV) files where each line describes an observed distance between atoms. The description constitutes of calculated distance in angstroms, atom labels (constructed from atom name, symmetry operator and translation in the crystal cell) (Gražulis et al., 2015), name of the original output file and COD ID of the structure in which the distance was identified (as displayed in the Figure 5).

5.2 Filtering

As previously discussed in Section 4.1, there are a number of features of experimentally measured crystallographic structures that may distort the results of this research. Additionally, based on the metadata provided in the CIF files, certain errors that could

[1] Located at `svn://www.crystallography.net/cod/cif`, revision 272900

[2] Located at `svn://www.crystallography.net/cod-tools/trunk`, revision 9473 (unreleased)

[3] Located at `svn://saulius-grazulis.lt/crystal-contacts/trunk`, revision 51 (unreleased)

[4] Located at `svn://saulius-grazulis.lt/atomclasses/trunk`, revision 584 (unreleased)

[5] Located at `https://github.com/kliment-olechnovic/voronota`

```
1.220843672   O2   N1         cif_contacts.out   1552948_cif_fillcell
3.401762186   O3   N2_3_565   cif_contacts.out   1552948_cif_fillcell
1.214520303   O3   N1_1_565   cif_contacts.out   1552948_cif_fillcell
3.059595731   O4   N1_3_555   cif_contacts.out   1552948_cif_fillcell
2.344228869   O4   N2         cif_contacts.out   1552948_cif_fillcell
```

Figure 5. Example of interatomic distance class data file, taken from N-O interatomic distance class.

affect this research can be identified. Therefore, input data needs to be analyzed and suspicious structures need to be removed prior to further calculations. The filtering was done on all 484 652 crystal structures from the COD.

Filtering is performed using a *Makefile* system. First, relevant information about each crystal structure is extracted from the corresponding CIF file using a *Perl* script `COD-features` and gathered in a TSV file. Then, based on the enabled filters, a COD ID list is generated which includes only structures without unsuitable features. Structures are filtered out only when information provided in the CIF file proves the structure to not fit the established requirements – lack of information for testing of a certain feature is not considered to be sufficient to remove the structure from further analysis. This list is further refined based on limitations on distances between atoms. The final filtered list of COD IDs is used to filter out rejected structures from each interatomic distance class.

5.2.1 Structural anomalies

The most straightforward way to ensure that a crystal structure model is complete is by comparing the chemical formula declared in the CIF file and chemical formula calculated based on the atomic coordinates. The declared chemical formula describes the chemical formula of a unit cell. Chemical formula is calculated using the tool `cif2cod` from *cod-tools* package and represents a formula unit. A number of formula units in the unit cell are defined under a CIF data name *_cell_formula_units_Z*.

Since in X-ray crystallography experiments hydrogen atom counts and coordinates usually are not determined precisely, hydrogen atoms and their counts are removed from both formulas before comparison. In order to compare declared and calculated chemical formulas, atom counts in the calculated chemical formula are multiplied by the number of formula units. In case the number of formula units is not declared in the CIF file, the comparison cannot be performed. Mismatches between declared and calculated chemical formulas were identified in 8.25% of all analyzed structures and such structures were removed from further calculations.

Some CIF data names are useful for quick identification of structural anomalies in crystal models. The *Perl* module `CODFlags.pm` from *cod-tools* package is useful for identification of such CIF file features. In order to identify some anomalies related to the method of site coordinate identification, superspace group usage and involvement of unmodeled solvent molecules, additional functions were added to this tool.

An important CIF data name for structure validation is *_atom_site_calc_flag* which indicates the method which was used to determine the coordinates of the site. The value *d* indicates that the coordinates were determined from diffraction measurements, value *c* or *calc* indicated that the coordinates were calculated from molecular geometry and value *dum* indicates that the site is a dummy site which does not represent a real atom, but rather an arbitrary point in space.

For the purpose of this research, experimentally determined atom coordinates are required as calculated coordinates may misrepresent distances between atoms. However, the exact position of hydrogen atoms can rarely be determined using X-ray crystallography. Therefore, data structures that have elements other than hydrogen, deuterium or tritium marked as sites with calculated coordinates are removed from further analysis. Additionally, if a dummy site is identified in the data file, the structure is omitted from further analysis. In total, about 0.11% of all analyzed crystal structures had to be rejected based on this requirement.

In order to restore the three-dimensional structure of a crystal which is described using superspace groups, the symmetry operations would need to be projected to the three-dimensional space group operations. This process may cause distorted positions of atoms which would significantly affect further calculations. For example, crystal structure with COD ID 2100486 is described using superspace groups and has an unusually short C–C aromatic bond (1.29 Å). Therefore, data structures with superspace group related CIF data names (*_space_group_ssg_**, *_space_group_symop_ssg_**, *_geom_angle_site_ssg_symmetry_** or *_geom_bond_site_ssg_symmetry_**) are omitted from further analysis. Only 0.06% of all analyzed structures had to be removed due to this requirement.

Some crystal structures may include unmodeled solvent molecules. In such cases, solvent molecules that are missing from the model create voids in the structure. These voids would have a negative impact on precision of identified distances between atoms by artificially increasing them. Such structures can be identified from CIF data names *_platon_squeeze_void_count_electrons* and *_smtbx_masks_void_count_electrons*. During filtering, 4.80% of all structures were identified to include unmodeled solvent molecules and were removed from further analysis.

Crystal density is an important characteristic that may signal possible inconsistencies in the crystal structure model. In a CIF file, density calculated from crystal cell and contents is noted under the *_exptl_crystal_density_diffrn* data name and experimentally measured density is noted under the *_exptl_crystal_density_meas* data name. In CIF files, density is measured in grams per cubic centimeter.

Crystal structures with very low density may indicate that the structure has missing atoms. Some structures may purposely have large voids resulting in low density, for example metal-organic frameworks. A study published by El-Kaderi et al. (2007) discusses three-dimensional covalent organic frameworks which are materials in which all interactions are covalent and are said to have extremely low densities of 0.17 g/cm^3, therefore, it was decided to use this value as the lower threshold for acceptable crystal densities.

On the contrary, extremely high crystal structure density may also indicate errors in the structure. For example, high density may be caused by unmarked partial occupancies or unmarked disorders. One of the densest materials in the world is osmium (Arblaster, 2014) which has a density of 22.59 g/cm^3, therefore this value was used as the upper threshold for acceptable crystal densities.

Crystal density can be calculated using tool `cif2cod` from *cod-tools* package. This tool calculates the volume and chemical formula of the unit cell. The chemical formula of the unit cell can be used to calculate the mass of the cell by multiplying the number of specific atoms in the cell by their corresponding atomic weight. By dividing the calculated mass by determined volume, crystal density can be determined.

A significant difference between density calculated from the cell chemical formula and calculated density declared in the CIF file could signify all previously discussed issues as well as other errors. Therefore, a ratio between declared calculated density (as per *_exptl_crystal_density_diffrn* CIF data name) and calculated density based on cell chemical formula and volume was determined for each molecule. Crystal structures with a calculated ratio lower than 0.75 or higher than 1.25 were removed from further analysis. Overall, about 2.53% of all analyzed structures had to be omitted from further analysis by applying this criterion.

Residual factor, or R-factor, is a measure of quality of a crystallographic model. R-factor evaluates how well the crystallographic model corresponds to the data gathered through the X-ray diffraction experiment. R-factor is calculated using formula:

$$R = \frac{\sum |F_{obs} - F_{calc}|}{\sum F_{obs}} \tag{1}$$

where R is the R-factor, F_{obs} is the observed structure-factor amplitudes, F_{calc} is the calculated structure-factor amplitudes and the sum includes all measured X-ray reflections.

In this study, structures with R-value higher than 0.1 are removed from the data pool. This limit was selected based on the IUCr data validation guidelines (IUCr, 2000). About 18.22% out of all analyzed data structures did not fit this requirement.

5.2.2 Measurements in unusual conditions

Temperature and pressure conditions applied on the crystal during an X-ray crystallography experiment are important features as abnormal conditions may affect the observed interaction between atoms and distort further calculations. CIF files have specific data names where such information can be noted. These measurements can be found under data names *_cell_measurement_pressure*, *_diffrn_ambient_pressure*, *_cell_measurement_temperature* and *_diffrn_ambient_temperature*, where *_cell_measurement_** data names refer to conditions at which the unit-cell parameters were measured and *_diffrn_ambient_** data names refer to the mean value of conditions at which the intensities were measured. Pressure is measured in kilopascals and temperature is measured in kelvins.

For this study, structures that were measured in temperature higher than 320 K or pressure outside the range of [80, 120] kPa were removed. Most such structures were observed in COD entries in ranges 1501500 – 1501600 and 9002000 – 9014000 and represent a number of measurements of the same compound in slightly different conditions performed in the same study, for example a sample of 43 entries from the former range with significantly higher temperature and pressure measurements is of tungsten carbide crystal published in a pressure-volume-temperature equation study by Litasov et al. (2010). About 1.72% of all analyzed data structures were filtered out based on this criterion.

5.2.3 Interatomic distance criteria

Distances between atoms were extracted from 358 626 CIF files. These structures were analyzed further by recording shortest and longest interatomic distances observed in each file. This analysis showed that some structures include particularly short or long atomic distances.

Observations of shorter distances may arise due to unmarked disorder, which may affect most of the distances in a crystal. For example, substitutional disorder is caused by different atoms occupying the same site in two unit cells (Neuburger, 2012). In order to define such disorder in a CIF file, different sites with partial occupancies are placed at identical coordinates. This disorder may cause distortion of the three dimensional crystal structure which can result in identification of abnormally short distances between atoms. In order to avoid distortion of the study results, all data structures with any instances of identified interatomic distances smaller than 1 Å were removed from further calculations.

Even after filtering the data, some random distribution (as shown in Figure 2) can be observed in most interatomic distance classes. This random distribution consists of distances between atoms that are interacting through forces other than intramolecular interactions or van der Waals forces. Additionally, some observations may be caused by noise induced by crystal structure disorders that may have been missed by prior filtering.

Partial removal of such noise can be achieved by applying an upper limit to interatomic distances. The largest covalent radius based on previously listed covalent radii tables is the radius of francium and is equal to 2.6 Å (Cordero et al., 2008). This would imply that in theory the largest distance between bonded atoms should be observed in the Fr–Fr interatomic distance class and be equal to 5.2 Å. Methodology used for intramolecular bond length determination in this research requires chemical bonds peak, van der Waals gap and the peak of van der Waals interactions to be included in the interatomic distance class. Therefore, in order to effectively remove structures with undetected disorders but keep a sufficient sample range to observe the interatomic distance distribution model described by Alvarez (2013), upper limit for distances between atoms is set to 6 Å. These interatomic distance limitations are responsible for removal of 7.59% out of 358 626 analyzed crystal structure models.

5.3 Determination of the intramolecular bond length

Chemical bond length has to be identified for each interatomic distance class. To complete this task, the van der Waals gap needs to be identified in the distribution of distances between atoms for each pair of elements. The van der Waals gap corresponds to the lowest density region between two density peaks, where the first one represents distances between atoms attracted by intramolecular forces and the second represents distances between atoms affected by the van der Waals forces.

5.3.1 Interatomic distance distribution approximation

In order to identify the lowest density region, interatomic distance distribution is approximated by a Gaussian mixture model. Program `fit-model` (Merkys, 2018) which uses functions from the *R* package *MixtureFitting v0.4.0*[6] was employed to determine parameters of mixture models by utilizing an expectation-maximization algorithm. This tool fits 10 mixture models to the data, each of which contains a different number of components (from 1 to 10). The program calculates parameters for each model: proportions, centers and standard deviations of its components, logarithm of the likelihood of the model and Bayesian Information Criterion (BIC) (Schwarz, 1978).

The number of observations in the class impacts the number of mixture models generated for it. Each component in the model is described using 3 parameters – proportion (proportions of all components sum up to 1), center and standard deviation. Therefore, in order to describe a model with n components $3n - 1$ parameters have to be determined. In this study, in order to ensure that a sufficient amount of data is used to approximate the models, at least 10 data points are required to determine each parameter. It means that in order to generate a Gaussian mixture model with n components interatomic distance class has to include at least $10 * (3n - 1)$ observations.

The BIC is used to determine which mixture model describes data the best while avoiding overfitting. If a Gaussian mixture model with only one component is identified as the best fitting, the class is omitted from further analysis, as in such case van der Waals gap cannot be identified. Furthermore, if the model involves more than one component but the components overlap in a way that does not allow to determine a local minimum between them, the class is omitted from further analysis as well.

5.3.2 Identification of the van der Waals gap

The location of the lowest density region is identified using the simplex method (Nelder and Mead, 1965). The simplex algorithm is an iterative process that consecutively checks if the marginal points of the set are optimal solutions to the function minimization problem before moving on to neighboring marginal points with lower values. Simplex function from the *R* package *MixtureFitting* is used to identify the lowest value of the function between two Gaussian mixture model peaks of interest.

In the case of 2 component Gaussian mixture model, determination of the van der Waals gap is quite straightforward – simplex method is used to determine the lowest density point between the two peaks and this area is considered to be the van der Waals gap. However, for most interatomic distance classes, Gaussian mixture models

[6] Located at `https://github.com/merkys/MixtureFitting`

containing more than 2 components were chosen as best fitting based on the BIC. A number of methods that could be used to determine which components in such a model represent different types of interactions were discussed in bachelor's thesis by Šidlauskaitė (2021). In the current study, two approaches for identification of mixture model components of interest were tested.

5.3.3 Method A: analysis of the first two components

Using the first approach, the first two components (components with centers closest to zero on the x-axis) are assumed to represent intramolecular bonds and van der Waals interactions respectively. Lowest density region between them is considered to represent the van der Waals gap. A quick review of results achieved using this method revealed that this method works very well for classes where the van der Waals gap is not very clearly defined. For example, in class C-Mg chemical bond length determined using this van der Waals gap identification method is equal to bond length calculated using covalent radii table published by Cordero et al. (2008) and very similar to bond length calculated using covalent radii table published by Pyykkö and Atsumi (2009) (as shown in Figure 6.A.).

However, this approach has difficulty analyzing classes in which single, double and triple covalent bonds have high numbers of observations which causes each of these interactions to form separate, although closely located peaks. An example of issues encountered using this method is shown in Figure 6.B. In C-N interatomic distance class, the first three components appear to represent different types of covalent interactions between atoms. These three components are followed by a clearly defined interval with no interatomic distance observations which represents the van der Waals gap. However, in this case the method locates the van der Waals gap prematurely.

5.3.4 Method B: analysis of neighboring component pairs

The second approach locates the minimal density using the simplex algorithm between each pair of neighboring model components. The interval between peaks in which the lowest function value was identified is assumed to be the van der Waals region. Consequently, the peak before this region is then considered to represent the distances between atoms forming intramolecular bonds and the peak after this region – distances between atoms affected by van der Waals interactions.

This method is able to deal with possible multi-component representation of the bond length distribution and takes into consideration all interatomic distance measurements hence reducing the probability of introduction of overgeneralization bias to the study.

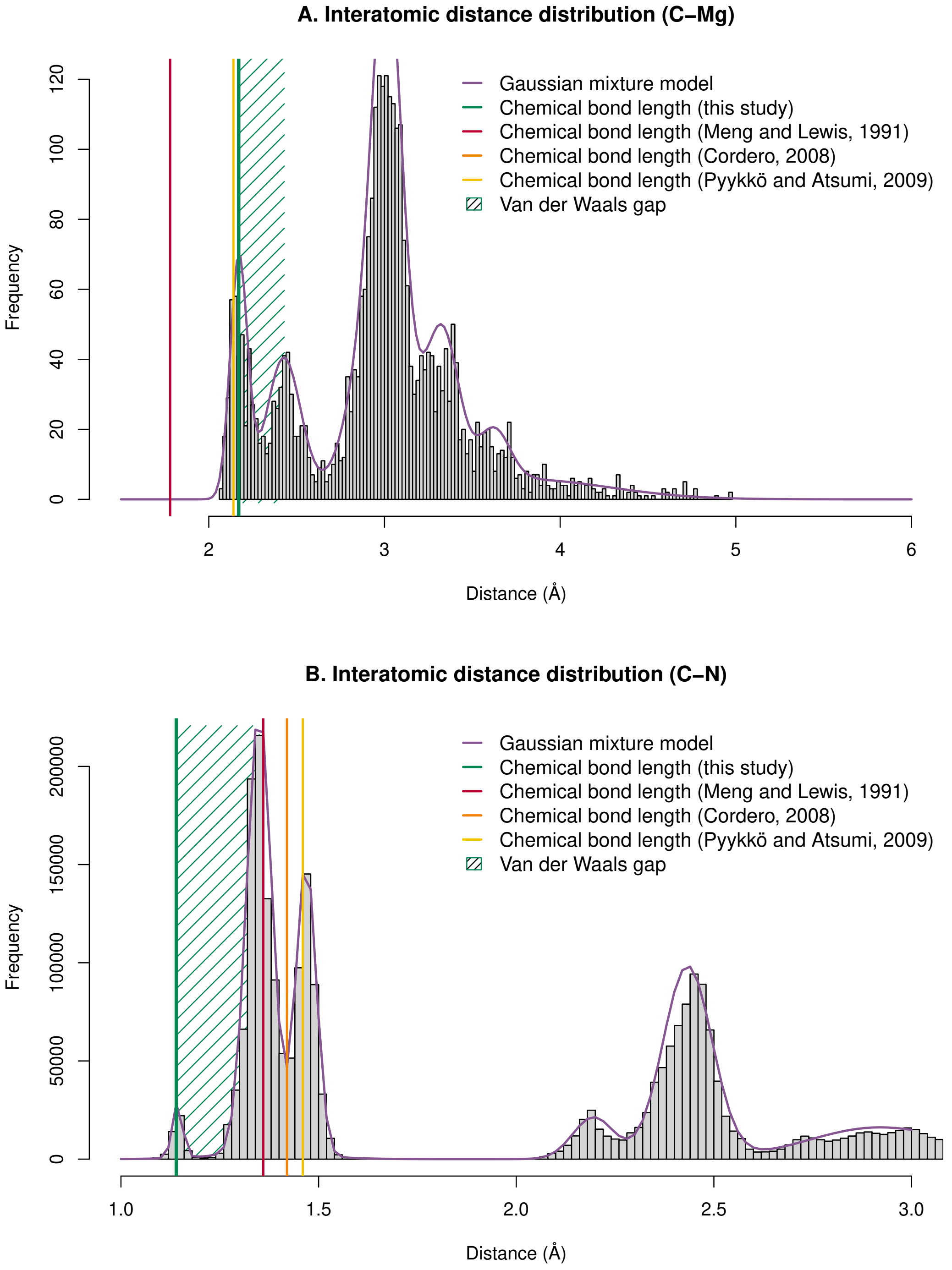


Figure 6. Visualization of the van der Waals gap identification method which considers the first two components of the distribution model to represent chemical bonds and van der Waals interactions respectively. Shown on C-Mg and C-N interatomic distance models.

Application on this method on C-N interatomic distance class is shown in Figure 7.A. In this case the algorithm compares all gaps between neighboring components and identifies the van der Waals gap location correctly which leads to a more precise determination of the chemical bond length.

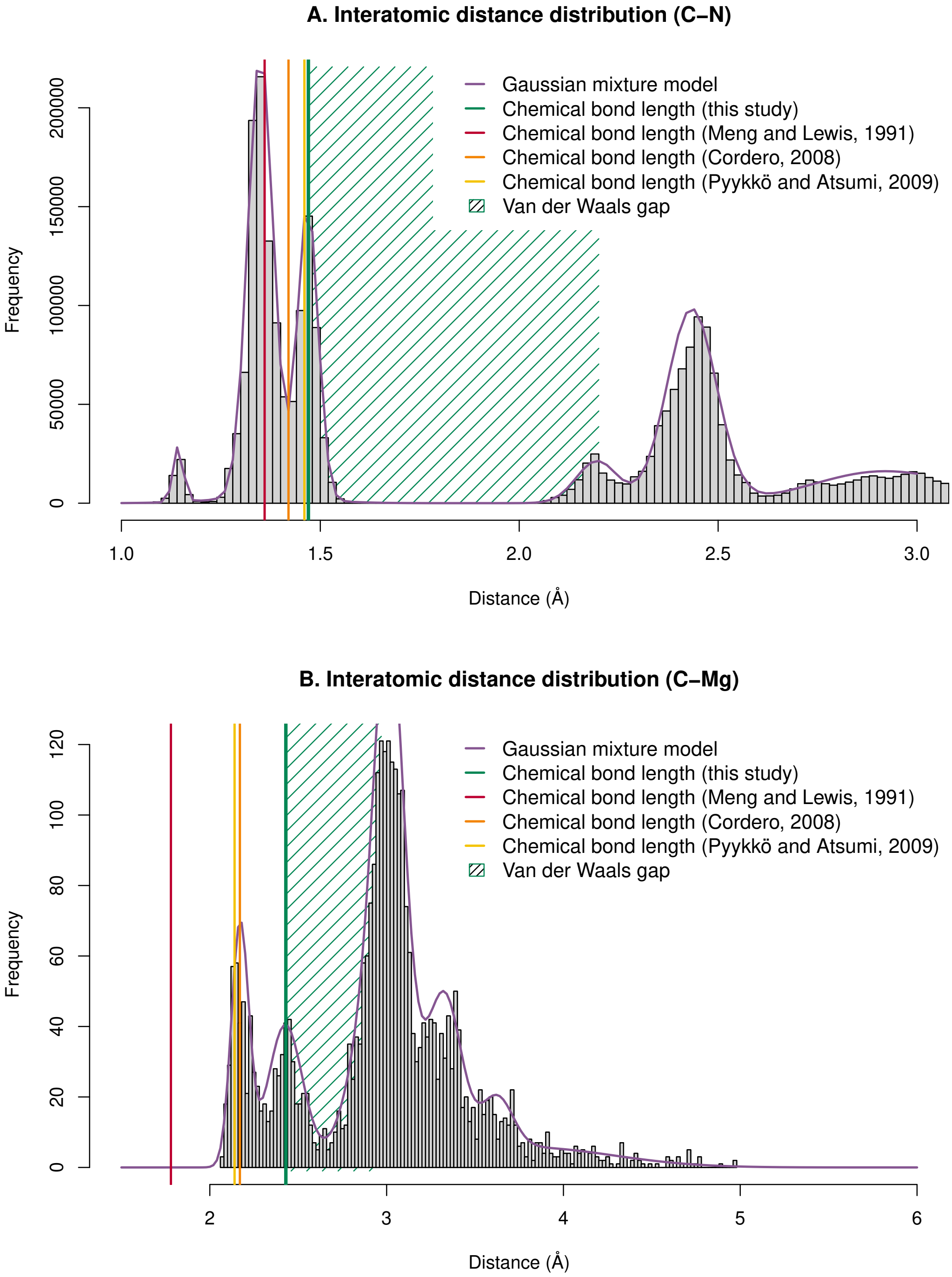


Figure 7. Visualization of the van der Waals gap identification method which considers gaps between each pair of neighboring components. Shown on C-N and C-Mg interatomic distance models.

However, the method performs less well for classes that do not display a clearly defined van der Waals gap. As each pair of components is evaluated, in such cases chemical bond length determined using this method is usually far greater compared to other sources. These issues are most often observed in classes that involve metals (example shown Figure 7.B.).

5.3.5 Additional considerations

In some classes small well-separated peaks of interatomic distance observations appear. These peaks can represent errors in the data or disorders in crystal structures that managed to pass through the filters. In other cases, the peaks are formed by unusual observations in classes that do not have a high number of total interatomic distance observations. For example, the interatomic distance class Cu-Sn has 439 total observations and a 7 component Gaussian mixture model was determined to describe data the best based on the BIC (as shown in Figure 8). The first component of this model describes 3 observations that are located around 1.42 Å. The next shortest observation in this class is located at 2.40 Å, resulting in an unpopulated range of over an angstrom in span.

Such distribution can lead both of the previously described methods to determine the intramolecular bond length to be substantially shorter or longer compared to distances calculated based on published atomic radii tables. An approach that could help solve this issue includes removal of components that have a significantly smaller proportion than other components of the model. However, after some testing, a threshold proportion separating significant components from components fitted to noise could not be determined. Therefore, this approach was not employed in this study.

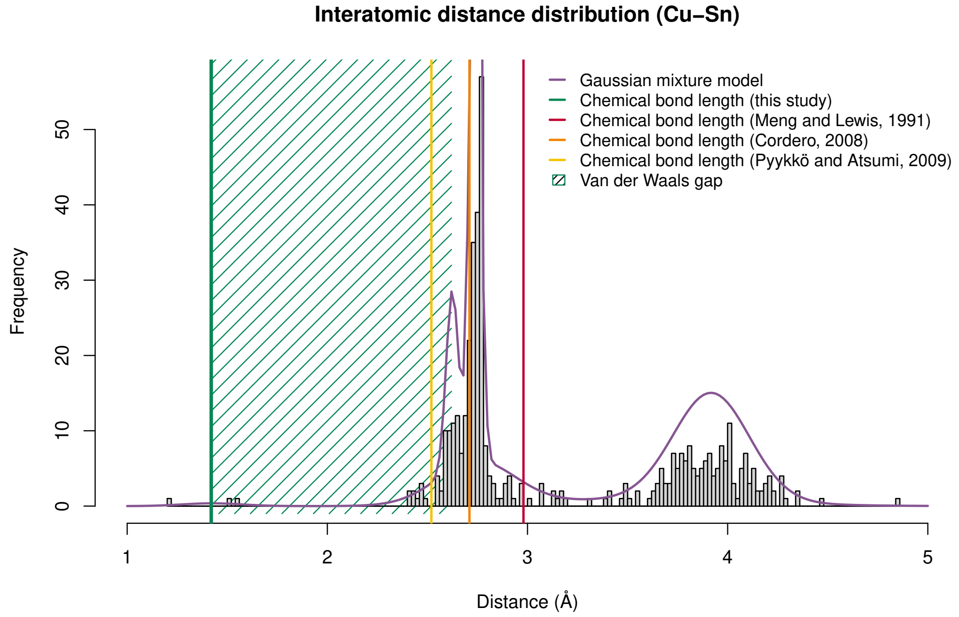


Figure 8. Visualization of a case where intramolecular bond length identified using both methods is substantially shorter compared to data from other sources. Shown on Cu-Sn interatomic distance model.

In conclusion, van der Waals gap identification allows one to determine which peak represents distances between atoms in the longest intramolecular bonds. The center of this peak is considered to be the maximum chemical bond length of that interatomic distance class. These calculations are automatically performed using a *Makefile* system with a number of *R* and *Perl* scripts involved. The results of these calculations are gathered in one TSV file.

5.4 Calculating atomic radii

The maximum chemical bond length is considered to be equal to the sum of atomic radii of two bonded atoms with a possibility to add an additional tolerance value. In order to resolve atomic radii of each element, an overdetermined system of equations is generated. This system of equations is stored as a matrix, where each column represents a different element and each row – an interatomic distance class. Numbers present in each row identify how many atoms of each element participate in the observed bond, hence the sum of values in each row is equal to 2. Additional column is added on the right side of the matrix which represents the result of each equation - intramolecular bond length determined for the specific class.

The overdetermined system of equations is solved using the weighted least squares method. The least squares method analyzes the difference between the observed value and the fitted value for each equation and aims to minimize the sum of the square values of these differences (Kiers, 1997). The weighted version of this method assumes that certain equations are more important than others and makes these equations more influential to the final result. Weights are added to the least squares method by solving for x:

$$A^T W A x = A^T W b \tag{2}$$

$$x = (A^T W A)^{-1} A^T W b \tag{3}$$

where A is the equation matrix, W is the diagonal weight matrix, x is the vector being solved for and b is the result vector.

Analyzed interatomic distance classes vary widely in the number of observations. For example, the largest class C–C involves more than 15 million observations. In comparison, the smallest class that can be fitted with a 2 component model, such as the Cr–K class, includes only 50 interatomic distance observations. Larger sample sizes suggest that interatomic distance distribution in these classes is reflected in more detail, can be approximated more precisely and provide more reliable intramolecular bond length estimations. Additionally, small classes are prone to specific issues related to the determination of the location of the van der Waals gap (as described in section 5.3.5). Therefore, the squared number of observations in each interatomic distance class was stored in a diagonal matrix and used as weights for the overdetermined system of equations. To solve the final system, the function `lsolve.ssor` from *R* package *Rlinsolve* (You, 2022) was used.

5.4.1 Variations of the atomic radii derivation method

Different variations of atomic radii determination method can be achieved by limiting which classes are included in the calculations based on the type of participating elements or by selectively combining the results from previously described van der Waals gap identification methods (discussed in Section 5.3.3 and Section 5.3.4). In this study, five variations of determined maximum intramolecular bond length datasets were tested and atomic radii were derived.

1. Gaps defined using Method A (described in Section 5.3.3). Calculated atomic radii table will be further referred to as table A.

2. Gaps defined using Method B (described in Section 5.3.4). Calculated atomic radii table will be further referred to as table B.

3. Gaps defined using Method B, using only the classes which involve at least one organic element. In this study, organic elements are defined as nonmetals (C, N, O, P, S, Se) and halogens (F, Cl, Br, I). This approach is influenced by the method used by Cordero et al. (2008) who prioritized defining covalent radii for the organic elements first. Furthermore, review of classes involving organic elements showed that van der Waals gap appears to be well-pronounced. Calculated atomic radii table will be further referred to as table B-ORG.

4. Gaps defined using Method A for classes that involve at least one element from the first two groups of the periodic table and using Method B for all remaining classes. In this study, this dataset will further be referred to as Method C. Although Method B appears to be a more scientifically sound solution that returns more cohesive results, it fails to correctly identify intramolecular bond length for classes that do not display a clearly defined van der Waals gap. The Method C is used to try and combat this issue as such cases usually involve elements from the first two groups of the periodic table. Calculated atomic radii table will be further referred to as table C.

5. Gaps defined using Method C, using only the classes which involve at least one organic element. Calculated atomic radii table will be further referred to as table C-ORG.

5.5 Validation of the results

To evaluate derived radii tables, connectivity between atoms in data from external sources can be tested. For this purpose, a *Perl* program `check_contacts`[7] (revision 2003) is used which reads the list of bonds provided in the input file, detects and reports both overlong and missing bonds in chemically annotated 3D structures. The bond lengths are evaluated based on the provided input atomic radii table in YAML format. This tool accepts structural data files (SDF) and chemical markup language (CML) files as input. Then, `AtomNeighbours.pm` package from *cod-tools* is used to identify neighboring atoms. An implementation of user interface for this program with selection of atomic radii tables generated in this study is available online[8].

Bonds between atoms listed in the input files are compared to expected chemical bond lengths based on the input radii table. Program informs about bonds that are longer than the maximum intramolecular bond length based on the provided radii table (overlong bonds) and distances between neighboring atoms that are not marked as bonded although should be based on the distance (missing bonds). Additionally, a tolerance value can be added to the maximum intramolecular bond length calculated using the provided atomic radii table in order to reduce the sensitivity of the validation process.

For testing, 236 352 SDF files were used. These SDF files are derived from the COD data using a chemical perception pipeline as described in Merkys et al. (2023). During validation, tolerance value of 0.35 Å is added.

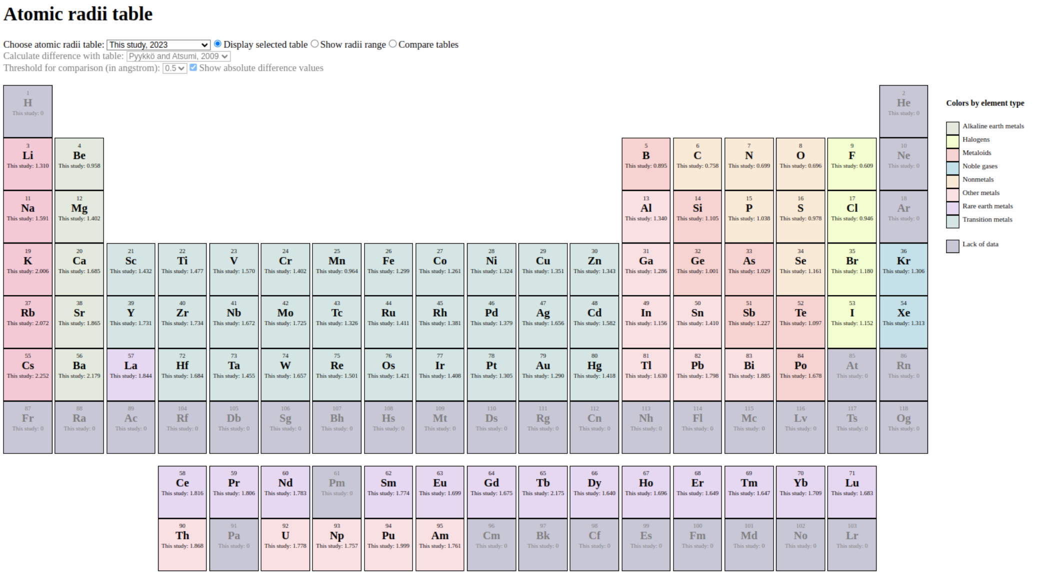
5.6 Access to results

An interactive website[9] for browsing calculated atomic radii, comparing different atomic radii tables and reviewing the details of the calculations has been developed. Atomic radii table is presented in the form of a periodic table (shown in Figure 9) where the derived radii values are shown next to each element. Selection menu at the top of the page provides options to visualize the range of atomic radii sizes or compare them

[7] Located at `svn://www.crystallography.net/contacts-in-COD/trunk/bin/check_contacts`, revision 2003

[8] Located at `http://databases.crystallography.lt:8080/contacts/website/cgi-bin/check_contacts.pl`

[9] Located at `http://databases.crystallography.lt:8080/contacts/website/cgi-bin/cov_radii_table.pl`



Atomic radii table

Choose atomic radii table: This study, 2023 ● Display selected table ○ Show radii range ○ Compare tables
Calculate difference with table: Pyykkö and Atsumi, 2009
Threshold for comparison (in angstrom): 0.5 ☑ Show absolute difference values

1 H This study: 0																	2 He This study: 0
3 Li This study: 1.310	4 Be This study: 0.958											5 B This study: 0.895	6 C This study: 0.758	7 N This study: 0.699	8 O This study: 0.696	9 F This study: 0.609	10 Ne This study: 0
11 Na This study: 1.591	12 Mg This study: 1.402											13 Al This study: 1.340	14 Si This study: 1.105	15 P This study: 1.038	16 S This study: 0.978	17 Cl This study: 0.946	18 Ar This study: 0
19 K This study: 2.006	20 Ca This study: 1.685	21 Sc This study: 1.432	22 Ti This study: 1.477	23 V This study: 1.570	24 Cr This study: 1.402	25 Mn This study: 0.964	26 Fe This study: 1.299	27 Co This study: 1.261	28 Ni This study: 1.324	29 Cu This study: 1.351	30 Zn This study: 1.343	31 Ga This study: 1.286	32 Ge This study: 1.001	33 As This study: 1.029	34 Se This study: 1.161	35 Br This study: 1.180	36 Kr This study: 1.306
37 Rb This study: 2.072	38 Sr This study: 1.865	39 Y This study: 1.731	40 Zr This study: 1.734	41 Nb This study: 1.672	42 Mo This study: 1.725	43 Tc This study: 1.326	44 Ru This study: 1.411	45 Rh This study: 1.381	46 Pd This study: 1.379	47 Ag This study: 1.656	48 Cd This study: 1.582	49 In This study: 1.156	50 Sn This study: 1.410	51 Sb This study: 1.227	52 Te This study: 1.097	53 I This study: 1.152	54 Xe This study: 1.313
55 Cs This study: 2.252	56 Ba This study: 2.179	57 La This study: 1.844	72 Hf This study: 1.684	73 Ta This study: 1.455	74 W This study: 1.657	75 Re This study: 1.501	76 Os This study: 1.421	77 Ir This study: 1.408	78 Pt This study: 1.305	79 Au This study: 1.290	80 Hg This study: 1.418	81 Tl This study: 1.630	82 Pb This study: 1.798	83 Bi This study: 1.885	84 Po This study: 1.678	85 At This study: 0	86 Rn This study: 0
87 Fr This study: 0	88 Ra This study: 0	89 Ac This study: 0	104 Rf This study: 0	105 Db This study: 0	106 Sg This study: 0	107 Bh This study: 0	108 Hs This study: 0	109 Mt This study: 0	110 Ds This study: 0	111 Rg This study: 0	112 Cn This study: 0	113 Nh This study: 0	114 Fl This study: 0	115 Mc This study: 0	116 Lv This study: 0	117 Ts This study: 0	118 Og This study: 0
			58 Ce This study: 1.816	59 Pr This study: 1.806	60 Nd This study: 1.783	61 Pm This study: 0	62 Sm This study: 1.774	63 Eu This study: 1.699	64 Gd This study: 1.675	65 Tb This study: 2.175	66 Dy This study: 1.640	67 Ho This study: 1.696	68 Er This study: 1.649	69 Tm This study: 1.647	70 Yb This study: 1.709	71 Lu This study: 1.683	
			90 Th This study: 1.868	91 Pa This study: 0	92 U This study: 1.778	93 Np This study: 1.757	94 Pu This study: 1.999	95 Am This study: 1.761	96 Cm This study: 0	97 Bk This study: 0	98 Cf This study: 0	99 Es This study: 0	100 Fm This study: 0	101 Md This study: 0	102 No This study: 0	103 Lr This study: 0	

Colors by element type

- Alkaline earth metals
- Halogens
- Metaloids
- Noble gases
- Nonmetals
- Other metals
- Rare earth metals
- Transition metals

- Lack of data

Figure 9. Radii browser website.

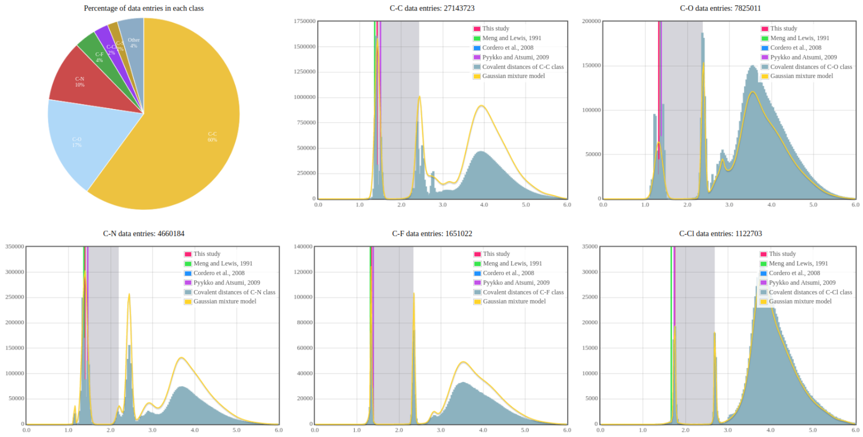


Figure 10. Details of the classes used in atomic radii derivation for carbon from the radii browser website. Not all histograms are displayed for brevity.

to values in other published radii tables (Meng and Lewis, 1991; Cordero et al., 2008; Pyykkö and Atsumi, 2009) and visualize the difference.

By clicking on each chemical element, users can review detailed information about classes that were used to derive the atomic radius (shown in Figure 10). For each element, a pie chart is shown, which displays the percentage of interatomic distance observations in each class that involves the selected element. Furthermore, histograms that display interatomic distance distribution in each of the involved interatomic distance classes are shown. In each histogram the best fitting Gaussian mixture model, identified intramolecular bond length and bond lengths calculated based on other published covalent radii tables are displayed. This functionality allows an easy way to review the analysis performed on each class and ensures transparency of the methodology applied.

The website was developed using *Perl common gateway interface (CGI)* and *JavaScript* plotting library *Flot*. The data used in the website is updated each time the results are recalculated.

6 RESULTS

6.1 Data preparation

After completing all the filtering steps, 297 973 crystal structures were accepted for covalent bond determination, which is approximately 61.5% of all structures published in the COD (as displayed in Figure 11). Interatomic distance observations in these structures populate 2897 interatomic distance classes. However, 896 of these classes have less than 20 observations, therefore Gaussian mixture models for them were not generated. Models were successfully generated for 2001 classes. Furthermore, 601 classes had to be removed from further study as a model with 1 component was identified as best describing the data.

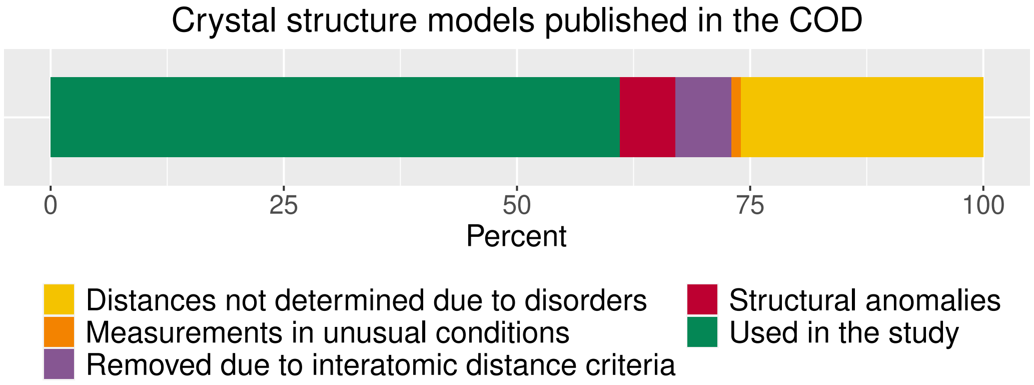


Figure 11. Results of data filtering (as described in Section 5.2).

In conclusion, suitable models were generated for 1400 interatomic distance classes. However, in some cases although the generated model had 2 components or more, the components were substantially overlapping. Therefore, some models may involve multiple peaks with no local minimums between them. For these classes the lowest density region cannot be identified and intramolecular bond length cannot be calculated. In total, maximum chemical bond length was determined for 1205 interatomic distance classes. After solving the overdetermined equation, covalent radii were successfully determined for 84 chemical elements.

6.2 Derived atomic radii tables

Full derived atomic radii tables can be found in Appendix 1. Figure 12 displays a comparison of atomic radii tables A, B and C. Graph shows that, as expected, atomic radii tables B and C are quite similar. Considering that 0.1 Å is a relatively substantial difference between radii, comparing the results from the two methods such difference is observed only for Li and Mg atoms. These atoms belong to periodic table groups 1 and

2, therefore, the results are directly affected by the usage of Method A for intramolecular bond length determination in related classes.

Values in atomic radii table A appear to be less consistent. The resulting radii are shorter compared to other published covalent radii tables. Furthermore, although the results follow the general atomic radii trend, a number of radius fluctuations appears to be higher.

In conclusion, this comparison revealed that methodology behind atomic radii table A is the worst performing approach for intramolecular bond length determination in the general sense. However, selective usage of data generated from Method A to construct the Method C dataset leads to improved atomic radii. Overall, the methodology behind atomic radii table C appears to be the best working approach out of the three tables that were compared.

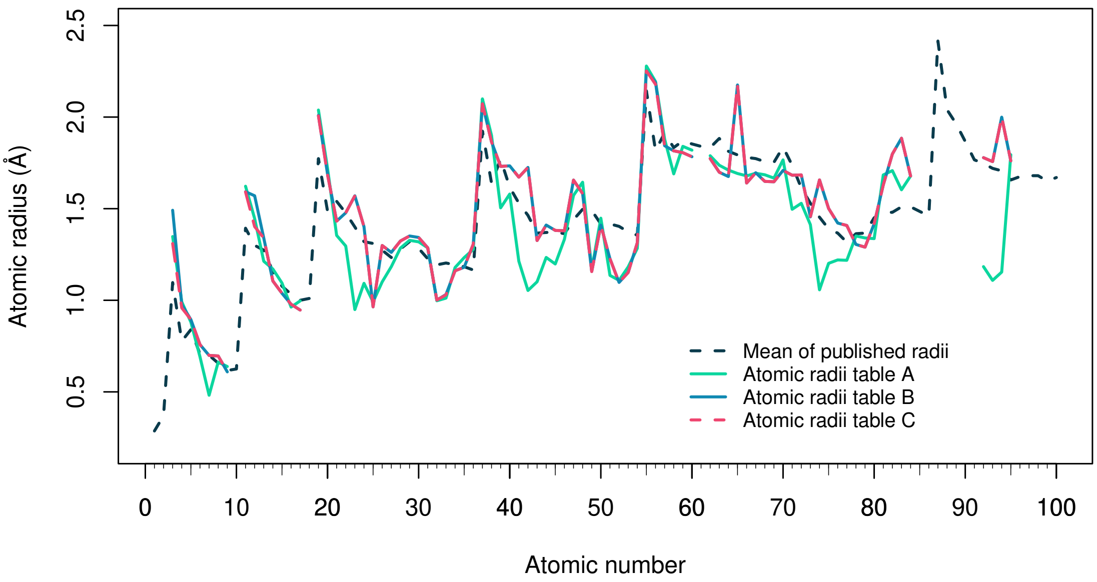


Figure 12. Comparison of atomic radii tables A, B and C (as described in Section 5.4.1). Published radii refer to tables of Meng and Lewis (1991), Cordero et al. (2008) and Pyykkö and Atsumi (2009).

6.2.1 Results using Method B

Comparison of radii in tables B and B-ORG is shown in Figure 13. In general, both atomic radii tables are quite similar, with radii differing in more than 0.1 Å for only 12 elements. Out of these elements in one case (Mg) radii are longer using the default dataset, and radii for the remaining cases (Na, K, Sc, Ge, Rb, In, Sb, Te, Cs, Hg and Pb) are longer using the organics dataset. Overall, radii from table B follow the general atomic radii trend more closely. Usage of the organics-focused dataset results in longer radii of Na, K, Rb and Cs which represent some of the peaks in the atomic radii distribution. In conclusion, omission of classes that are not related to organic elements did not lead to improvement of atomic radii.

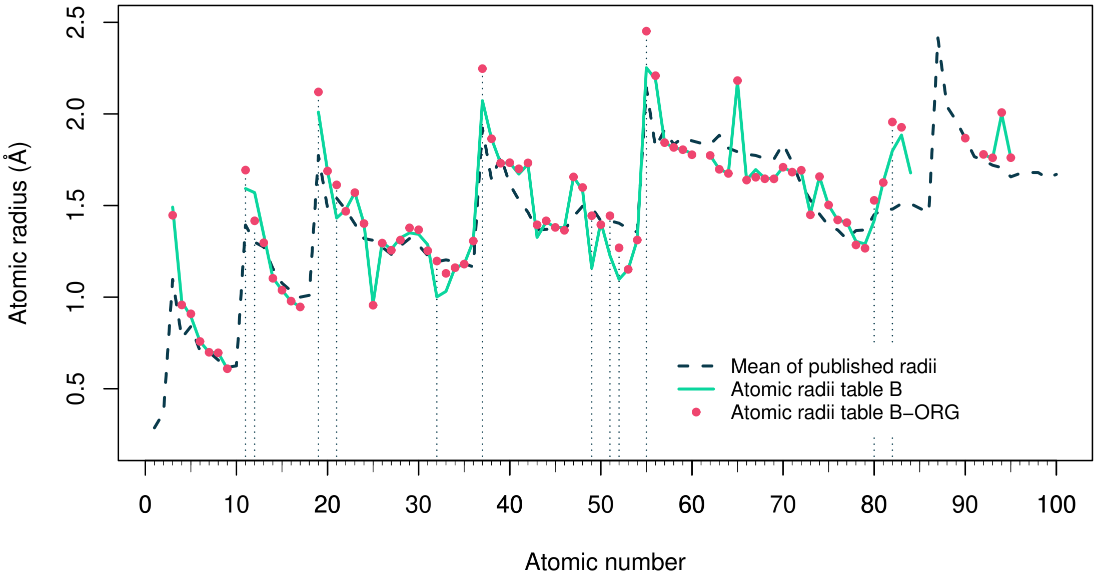


Figure 13. Comparison of atomic radii tables B and B-ORG (as described in Section 5.4.1). Published radii refer to tables of Meng and Lewis (1991), Cordero et al. (2008) and Pyykkö and Atsumi (2009).

6.2.2 Results using Method C

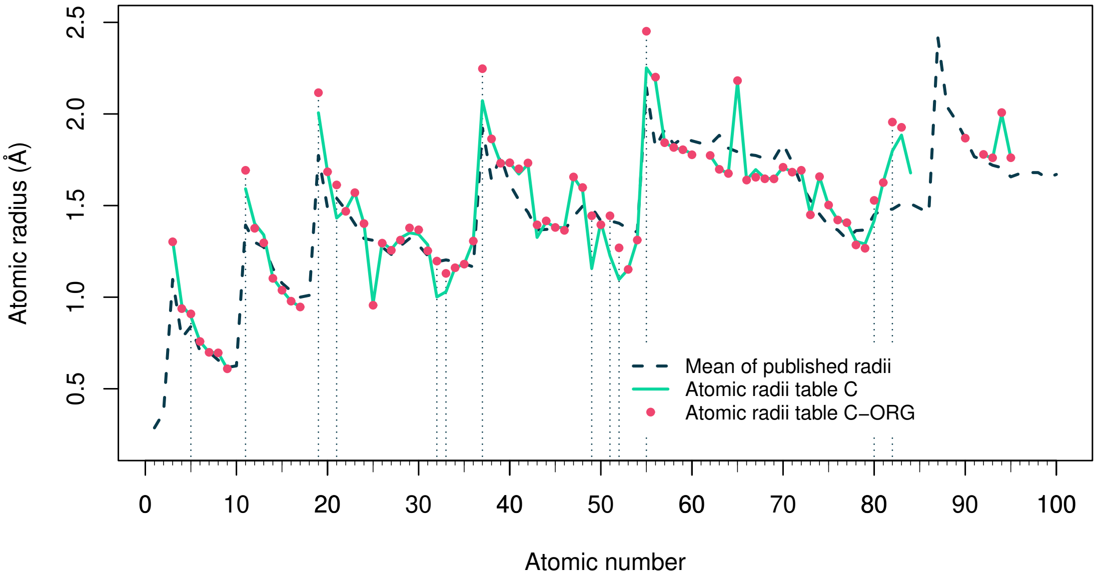


Figure 14. Comparison of atomic radii in tables C and C-ORG (as described in Section 5.4.1). Published radii refer to tables of Meng and Lewis (1991), Cordero et al. (2008) and Pyykkö and Atsumi (2009).

Atomic radii from tables C and C-ORG are shown in Figure 14. As seen from the graph, most of the radii are very similar. However, the atomic radii from the C-ORG table are longer than the radii from the C table for elements B, Na, K, Sc, Ge, As, Rb, In, Sb, Te, Cs, Hg, Pb. These differences result in stronger deviation of the atomic radii distribution from the usual trend. Therefore, it appears that the methodology behind atomic radii table C is the best working approach for atomic radii derivation.

6.3 Result validation

Atomic radii calculated using all datasets were tested by verifying bonding in structures as described in Section 5.5. Summary of verification results is provided in Table 1.

Table 1. Summary of validation results.

	SDF files with issues	Total number of errors	Out of which overlong
A	9.0%	319 023	88.7%
B	4.1%	108 329	32.7%
B-ORG	4.1%	104 960	24.9%
C	4.1%	106 911	34.2%
C-ORG	4.1%	104 277	26.0%
Meng and Lewis (1991)	4.5%	139 890	56.0%
Cordero et al. (2008)	4.2%	91 686	8.9%
Pyykkö and Atsumi (2009)	3.9%	82 187	63.6%

Note: Row names correspond to different intramolecular bond length data collection methods as described in the Section 5.4.1.

Based on the summary it is evident that atomic radii table A differs from the remaining tables the most. Using this table, errors were identified in more than twice as many SDF files compared to other radii tables. Furthermore, a high percentage of bonds that were identified as overlong, signals that radii in table A are comparably shorter than in other calculated tables. However, as some elements participate in a high number of interatomic contacts, the number of identified errors can be strongly affected by a change in the size of the radius of such an atom. For example, based on data shown in Appendix 1, the atomic radius of nitrogen atom in table A is 0.2 Å shorter than in other tables. As nitrogen atoms participate in a high number of interatomic contacts in the studied dataset, this difference alone highly inflates the number of identified errors. That can be confirmed based on the data in Appendix 2, as for table A the four out of five classes with most errors involve nitrogen. Results generated using the atomic radii tables B, B-ORG, C and C-ORG are very similar and both organics-focused radii tables identified a slightly lower percentage of overlong bonds.

More detailed comparison between all radii tables can be seen in Table 2. The highest number of identical errors were observed when using atomic radii table B-ORG and radii published by Cordero et al. (2008). However, there is also a close similarity between results achieved using the Cordero et al. (2008) table and radii tables B, C and C-ORG. It is important to note that these similarities are significantly higher than between results of any two of the previously published tables.

Table 2. Percentage of errors that are identical between two radii tables.

	Meng and Lewis (1991)	Cordero et al. (2008)	Pyykkö and Atsumi (2009)
A	5.5%	8.0%	5.9%
B	19.8%	38.4%	17.7%
B-ORG	20.7%	40.9%	18.2%
C	19.9%	38.6%	17.8%
C-ORG	20.8%	40.8%	18.1%
Meng and Lewis (1991)	X	22.3%	9.8%
Cordero et al. (2008)	22.3%	X	20.5%
Pyykkö and Atsumi (2009)	9.8%	20.5%	X

Note: Row names correspond to different intramolecular bond length data collection methods as described in the Section 5.4.1.

A table listing five classes with the most occurrences of each error type in each radii table can be found in Appendix 2. For all tables, one of the top 5 classes with most bonds identified as missing is Cu-Cu. Interatomic distance classes Fe-Fe and Mo-Mo are also frequently observed in this list. These observations are interesting as previously reviewed tables (Table 1 and Table 2) show that the amount of identical errors between tables and the proportion of overlong bonds to missing bonds varies quite widely.

7 DISCUSSION

Overall, all of the derived atomic radii tables follow the typical trend for atomic radii distribution. Out of four variations of calculated radii tables, table C was determined to be the most accurate and is considered to be the main result of this study. This decision is made based on the fact that radii from table C follow the atomic radii distribution of other tables very well (as shown in Figure 12) and this table perform similarly to the published atomic radii tables during the data validation process (as shown in Table 2). Although the C-ORG radii table appears to perform slightly better in some cases, methodology for derivation of the C-ORG table involves substantially more assumptions and limitations.

The main limitation of this study is the lack of the availability of open-access data on the interatomic distances classified by atom interaction type. In most of the previously published studies, analysis of strictly covalent bond length data is performed in order to derive covalent radii tables. In this study, the usage of Voronoi cells during the data preparation process is a valuable improvement that ensures that only relevant distances between atoms are considered. However, the final datasets still include a wide selection of interatomic interactions that affect the final interatomic distribution and impede the determination of the van der Waals gap. Therefore, the study results could be improved by further analyzing the interatomic distance distribution classes on a chemical level.

Comparable performance to other published covalent radii tables is an important achievement for the derived atomic radii dataset. Currently, *cod-tools* uses covalent radii from CCDC for small molecule crystal structure restoration. Further analysis of the validation results could lead to the use of the derived atomic radii in place of the current table.

The progress of this study has been presented as a poster presentation in the international conferences Open Readings 2022 and Open Readings 2023. In the year 2023 it received the best poster presentation award in its category.

8 CONCLUSIONS

All the tasks of this project were successfully achieved:

1. Interatomic distance data was collected from the Crystallography Open Database. The process involved calculation of the supercells, identification of the Voronoi neighbors and determination of the distances between neighboring atoms.

2. An automated system for data filtering was developed. Crystal structures with suspected disorders, structural anomalies, extreme measurement conditions or extreme interatomic distance observations were omitted from the study.

3. An automated system for intramolecular bond length determination was developed. Two main methods for van der Waals gap identification were explored and implemented.

4. Automated calculations for atomic radii derivation were performed. Six different approaches for initial dataset selection were reviewed.

5. Result validation was carried out to show close relation of the derived atomic radii to previously published results.

6. A website allowing to browse and compare atomic radii as well as review the data used to determine each value has been developed. Website can be accessed at `http://databases.crystallography.lt:8080/contacts/website/cgi-bin/cov_radii_table.pl`.

Completion of these tasks led to the achievement of the main aim of this study. During the study, an open atomic radii table has been derived and made freely accessible online. Methodology of this study is implemented in an automated and unsupervised manner.

9 ACKNOWLEDGEMENTS

This project has received funding from Vilnius University Science Support Fund (agreement No. MSF-JM-15/2021) and the Research Council of Lithuania (agreement No. S-MIP-23-87).

I would like to thank Andrius Merkys for valuable guidance during the project and Antanas Vaitkus for important insights. Furthermore, I would like to thank my classmates from VU MF Systems Biology 2023 course for continuous moral support.

References

1. Allen, F. H., Bellard, S.and Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. and Watson, D. G. (1979), ‘The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information’, *Acta Crystallographica Section B* 35(10), 2331–2339.

2. Allen, F. H., Kennard, O. and Watson, D. G. (1987), ‘Tables of Bond Lengths determined by X-Ray and Neutron Diffraction. Part I. Bond Lengths in Organic Compounds’, *Journal of the Chemical Society, Perkin Transactions 2* pp. S1–S19.

3. Alvarez, S. (2013), ‘A cartography of the van der Waals territories’, *Dalton Trans.* 42(24), 8617–8636.

4. Arblaster, J. W. (2014), ‘Osmium density challenge’, *Analytical and Bioanalytical Chemistry* 406(17), 4009–4011.

5. Atkins, P. W. (2023), ‘Encyclopedia Britannica’. Accessed: 2023-04-20.
   URL: *www.britannica.com/science/chemical-bonding*

6. Bacskay, G. B., Reimers, J. R. and Nordholm, S. (1997), ‘The Mechanism of Covalent Bonding’, *J. Chem. Educ.* 12, 1494.

7. Batsanov, S. S. (1995), ‘Experimental determination of covalent radii of elements’, *Russian Chemical Bulletin* 44, 2245–2250.

8. Bernstein, H. J., Bollinger, J. C., Brown, I. D., Gražulis, S., Hester, J. R., McMahon, B., Spadaccini, N., Westbrook, J. D. and Westrip, S. P. (2016), ‘Specification of the crystallographic information file format, version 2.0’, *Journal of Applied Crystallography* 49(1), 277–284.

9. Björck, A. (1990), *Handbook of Numerical Analysis*, Vol. 1, Elsevier, chapter Least squares methods, pp. 465–652.

10. CCDC (2010), ‘Elemental Radii’. Accessed: 2022-12-03.
    URL: *https://web.archive.org/web/20221023233244/https://www.ccdc.cam.ac.uk/support-and-resources/ccdcresources/Elemental_Radii.xlsx*

11. Cordero, B., Gómez, V., Platero-Prats, A. E., Revés, M., Echeverría, J., Cremades, E., Barragán, F. and Alvarez, S. (2008), ‘Covalent radii revisited’, *Dalton Transactions* (21), 2832–2838.

12. Daintith, J., ed. (2008), *A Dictionary of Chemistry*, 6 edn, Oxford University Press.

13. El-Kaderi, H. M., Hunt, J. R., Mendoza-Cortés, J. L., Côté, A. P., Taylor, R. E., O'Keeffe, M. and Yaghi, O. M. (2007), 'Designed synthesis of 3D covalent organic frameworks', *Science* 316(5822), 268–272.

14. Goh, G. B., Hodas, N. O. and Vishu, A. (2017), 'Deep learning for computational chemistry', *Journal of Computational Chemistry* 38, 1291–1307.

15. Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterottiand, L., Quirós, M., Serebryanaya, N. R., Moeck, P., Downs, R. T. and Le Bail, A. (2012), 'Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration', *Nucleic Acids Research* 40, D420–D427.

16. Gražulis, S., Merkys, A., Vaitkus, A. and Okulič-Kazarinas, M. (2015), 'Computing stoichiometric molecular composition from crystal structures', *Journal of Applied Crystallography* 48(1), 85–91.

17. Groom, C. R., Bruno, I. J., Lightfoot, M. P. and Ward, S. C. (2016), 'The Cambridge Structural Database', *Acta Cryst.* B72, 171–179.

18. Guha, R., Howard, M. T., Hutchison, G. R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J. and Willighagen, E. L. (2006), 'The Blue Obelisk-interoperability in chemical informatics', *Journal of Chemical Information and Modeling* 46(3), 991–998.

19. Hall, S. R., Allen, F. H. and Brown, I. D. (1991), 'The crystallographic information file (CIF): a new standard archive file for crystallography', *Acta Crystallographica Section A: Foundations of Crystallography* 47(6), 655–685.

20. Harlow, R. L. (1996), 'Troublesome Crystal Structures: Prevention, Detection, and Resolution', *J. Res. Natl. Inst. Stand. Technol.* 101(3), 327–339.

21. Israelachvili, J. N. (1974), 'The nature of van der Waals forces', *Contemporary Physics* 15, 159–178.

22. IUCr (2000), 'Data validation criteria, version 2000.06.09'. Accessed: 2022-05-01.
URL: *https://journals.iucr.org/services/cif/checking/RFACG_01.html*

23. IUCr (2006), 'Superspace group, From Online Dictionary of Crystallography'. Accessed: 2022-12-20.
URL: *https://dictionary.iucr.org/Superspace_group*

24. Kiers, H. A. (1997), ‘Weighted least squares fitting using ordinary least squares algorithms’, *Psychometrika* 62, 251–266.

25. Kleywegt, G. J. (2000), ‘Validation of protein crystal structures’, *Acta Crystallogr D Biol Crystallogr.* 56(3), 249–65.

26. Koster, G. F. (1957), *Solid state physics*, Vol. 5, Elsevier, chapter Space groups and their representations, pp. 173–256.

27. Le Pevelen, D. D. (2010), ‘Small Molecule X-Ray Crystallography, Theory and Workflow’, *Encyclopedia of Spectroscopy and Spectrometry (Second Edition)* pp. 2559–2576.

28. Leytham, K. M. (1984), ‘Maximum Likelihood Estimates for the Parameters of Mixture Distributions’, *Water Resources Research* 20, 896–902.

29. Li, Q. and Kang, C. (2020), ‘Mechanisms of Action for Small Molecules Revealed by Structural Biology in Drug Discovery’, *Int J Mol Sci.* 15, 5262.

30. Litasov, K. D., Shatskiy, A., Fei, Y., Suzuki, A., Ohtani, E. and Funakoshi, K. (2010), ‘Pressure-volume-temperature equation of state of tungsten carbide to 32 GPa and 1673 K’, *Journal of Applied Physics* 108.

31. Massa, W., ed. (2004), *Crystal Structure Determination*, Springer Berlin, Heidelberg.

32. Meng, E. C. and Lewis, R. A. (1991), ‘Determination of molecular topology and atomic hybridization states from heavy atom coordinates’, *Journal of Computational Chemistry* 12(7), 891–898.

33. Merkys, A. (2018), Extraction and Usage of Crystallographic Knowledge for Refinement and Validation of Molecular Models, PhD thesis, Vilnius University.
URL: *https://www.lvb.lt/permalink/f/16nmo04/ELABAETD31079741*

34. Merkys, A., Vaitkus, A., Butkus, J., Okulič-Kazarinas, M., Kairys, V. and Gražulis, S. (2016), ‘COD::CIF::Parser: an error-correcting CIF parser for the Perl language’, *Journal of Applied Crystallography* 49(1), 292–301.

35. Merkys, A., Vaitkus, A., Grybauskas, A., Konovalovas, A., Quirós, M. and Gražulis, S. (2023), ‘Graph isomorphism-based algorithm for cross-checking chemical and crystallographic descriptions’, *J Cheminform* 12.

36. Mounet, N., Gibertini, M., Schwaller, P., Campi, D., Merkys, A., Marrazzo, A., Sohier, T., Castelli, I. E., Cepellotti, A., Pizzi, G. and Marzari, N. (2018), ‘Two-dimensional

materials from high-throughput computational exfoliation of experimentally known compounds', *Nature Nanotechnology* 13, 246–252.

37. Nelder, J. A. and Mead, R. (1965), 'A Simplex Method for Function Minimization', *The Computer Journal* 7, 308–313.

38. Neuburger, M. (2012), Disorder in crystal structures: new approaches in finding the best model, PhD thesis, University of Basel.

39. Nikolaienko, T. Y. and Bulavin, L. A. (2019), 'Localized orbitals for optimal decomposition of molecular properties', *International Journal of Quantum Chemistry* 119(3), e25798.

40. Nikolaienko, T. Y. and Bulavin, L. A. (2021), 'The machine-learned radii of atoms', *Computational and Theoretical Chemistry* 1204, 113389.

41. Nikolaienko, T. Y., Chuiko, V. S. and Bulavin, L. A. (2019), 'The dataset of covalent bond lengths resulting from the first-principle calculations', *Computational and Theoretical Chemistry* 1163, 112508.

42. Nikolaienko, T. Y., Chuiko, V. S. and Bulavin, L. A. (2020), 'The covalent radii derived from the first-principle data', *Molecular Physics* 118(21-22), e1742937.

43. O'Keefe, M. and Brese, N. E. (1991), 'Atom sizes and bond lengths in molecules and crystals', *Journal of the American Chemical Society* 113(9), 3226–3229.

44. Olechnovič, K. and Venclovas, v. (2014), 'Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls', *Journal of Computational Chemistry* 35, 672–681.

45. Orio, M., Pantazis, D. A. and Neese, F. (2009), 'Density functional theory', *Photosynth Res* 102, 443–453.

46. Pisner, D. A. and Schnyer, D. M. (2020), *Machine Learning*, Academic Press, chapter 6 – Support vector machine, pp. 101–121.

47. Pyykkö, P. (2015), 'Additive Covalent Radii for Single-, Double-, and Triple-Bonded Molecules and Tetrahedrally Bonded Crystals: A Summary', *The Journal of Physical Chemistry A* 119(11), 2326–2337.

48. Pyykkö, P. and Atsumi, M. (2008), 'Molecular Single–Bond Covalent Radii for Elements 1–118', *Chemistry—A European Journal* 15(1), 186–197.

49. Pyykkö, P. and Atsumi, M. (2009), ‘Molecular Double–Bond Covalent Radii for Elements Li–E112’, *Chemistry—A European Journal* 15(46), 12770–12779.

50. Schwarz, G. (1978), ‘Estimating the Dimension of a Model’, *Annals of Statistics* 6, 461–464.

51. Spek, A. L. (2003), ‘Single-crystal structure validation with the program PLATON’, *J. Appl. Cryst.* 36, 7–13.

52. Spek, A. L. (2009), ‘Structure validation in chemical crystallography’, *Acta Cryst* D65, 148–155.

53. Stanjek, H. and Äusler, W. (2004), ‘Basics of X-ray Diffraction’, *Hyperfine Interactions* 154, 107–119.

54. Šidlauskaitė, E. (2021), ‘Determination of Covalent Bonds in Small Molecule Crystals’.
URL: *https://www.lvb.lt/permalink/f/8og6ah/ELABAETD107118730*

55. Weininger, D. (1988), ‘SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules’, *J. Chem. Inf. Comput. Sci.* 28(1).

56. Williams, L. D. (2021), ‘Molecular Interactions and the Behaviors of Biological Macromolecules’. Accessed: 2023-04-22.
URL: *https://williams.chemistry.gatech.edu/structure/molecular_interactions/mol_int.html#CVN*

57. Woińska, M., Grabowsky, S., Dominiak, P. M., Woźniak, K. and Jayatilaka, D. (2016), ‘Hydrogen atoms can be located accurately and precisely by x-ray crystallography’, *Sci Adv.* 27(5).

58. You, K. (2022), ‘Rlinsolve: Iterative solvers for (sparse) linear system of equations’.
URL: *https://cran.r-project.org/web/packages/Rlinsolve/index.html*

59. Zhang, Y. (2018), ‘Covalent Radii and New Applications’, *International Journal of Chemoinformatics and Chemical Engineering (IJCCE)* 7, 42–51.

10 SUMMARY

Vilnius University, Faculty of Medicine

Systems Biology study program

Eglė Šidlauskaitė

Determination of atomic radii from small-molecule crystal structures

Detailed analysis of small-molecule structures is a vital part of drug discovery. X-ray crystallography can be used to determine exact atom positions in crystal structures. Chemical bonding is usually inferred by comparing the distance between atoms to the sum of their atomic radii. However, there is no univocal method for the determination of atomic radii and commonly used radii tables are derived using data which is distributed under proprietary licenses.

In this study, a methodology and an independent workflow for automatic atomic radii derivation was developed. Crystal structures from the Crystallography Open Database were analyzed to filter out structures with unusual observations and used to obtain interatomic distance data. Typical maximum intramolecular bond length was determined for each pair of elements by fitting the Gaussian mixture model to the interatomic distance distribution and identifying the location of the van der Waals gap. The results were used to generate an overdetermined system of equations that was solved using the weighted least squares algorithm. As a result, atomic radii for 84 chemical elements were calculated.

Verification of the derived atomic radii table shows that it is able to detect connectivity in molecular entities comparably to other published atomic radii tables. The final atomic radii table, its comparison to other published covalent radii tables and intermediate data are freely accessible online.

11 SUMMARY IN LITHUANIAN

Vilniaus universitetas, Medicinos fakultetas

Sistemų biologijos studijų programa

Eglė Šidlauskaitė

Atomų spindulių nustatymas iš mažų molekulių kristalinių struktūrų

Išsami mažų molekulių struktūros analizė yra labai svarbi naujų vaistų paieškos dalis. Rentgeno spindulių kristalografija leidžia nustatyti tikslias atomų koordinates kristalinėse struktūrose. Cheminiai ryšiai tokiose struktūrose paprastai nustatomi lyginant atstumą tarp atomų su jų spindulių suma. Tačiau nėra vieningo metodo atomų spindulių nustatymui, o dažniausiai naudojamos spindulių lentelės yra sudarytos naudojant duomenis, kurie nėra atviros prieigos.

Šiame tyrime buvo sukurta automatinio atomų spindulių išvedimo metodika. Kristalų struktūros iš Atvirosios kristalografinės duomenų bazės (COD) buvo analizuojamos siekiant išfiltruoti struktūras su neįprastais stebėjimais ir panaudotos tarpatominių atstumų duomenims gauti. Tipinis didžiausias vidumolekulinio ryšio ilgis buvo nustatytas kiekvienos elementų poros tarpatominių atstumų pasiskirstymui pritaikant normaliųjų skirstinių mišinio modelį ir nustatant van der Valso tarpo vietą. Gauti rezultatai buvo panaudoti perteklinei lygčių sistemai sudaryti, kuri buvo išspręsta taikant svertinį mažiausių kvadratų metodą. Taip buvo apskaičiuoti 84 cheminių elementų atomų spinduliai.

Išvestos atomų spindulių lentelės validacija parodė, kad ji gali nustatyti jungtis molekulinėse esybėse panašiu tikslumu kaip ir kitos publikuotos atominių spindulių lentelės. Galutinė atominių spindulių lentelė, jos palyginimas su kitomis paskelbtomis kovalentinių spindulių lentelėmis ir tarpiniai duomenys yra patalpinti atvirai prieigai internete.

Appendix 1

Derrived atomic radii tables

Element	Atomic No.	A (1)	B (2)	B-ORG (3)	C (4)	C-ORG (5)
Li	3	1.348	1.491	1.447	1.310	1.302
Be	4	0.991	0.976	0.957	0.958	0.937
B	5	0.881	0.895	0.909	0.895	0.909
C	6	0.691	0.758	0.758	0.758	0.758
N	7	0.482	0.699	0.699	0.699	0.699
O	8	0.663	0.696	0.696	0.696	0.696
F	9	0.638	0.609	0.609	0.609	0.609
Na	11	1.622	1.593	1.693	1.591	1.692
Mg	12	1.446	1.570	1.417	1.402	1.376
Al	13	1.213	1.340	1.297	1.340	1.297
Si	14	1.170	1.105	1.103	1.105	1.103
P	15	1.094	1.038	1.038	1.038	1.038
S	16	0.963	0.978	0.978	0.978	0.978
Cl	17	0.997	0.947	0.946	0.946	0.946
K	19	2.038	2.010	2.120	2.006	2.116
Ca	20	1.720	1.689	1.688	1.685	1.684
Sc	21	1.354	1.432	1.613	1.432	1.613
Ti	22	1.296	1.477	1.469	1.477	1.469
V	23	0.949	1.570	1.570	1.570	1.570
Cr	24	1.093	1.402	1.402	1.402	1.402
Mn	25	0.991	0.964	0.956	0.964	0.956
Fe	26	1.100	1.299	1.295	1.299	1.295
Co	27	1.183	1.261	1.256	1.261	1.256
Ni	28	1.285	1.324	1.311	1.324	1.311
Cu	29	1.327	1.351	1.378	1.351	1.378
Zn	30	1.319	1.343	1.368	1.343	1.368
Ga	31	1.286	1.286	1.253	1.286	1.253
Ge	32	0.997	1.001	1.197	1.001	1.197
As	33	1.012	1.031	1.130	1.029	1.130
Se	34	1.179	1.161	1.161	1.161	1.160
Br	35	1.233	1.180	1.180	1.180	1.180
Kr	36	1.277	1.306	1.306	1.306	1.306
Rb	37	2.099	2.072	2.247	2.072	2.247

Note: Column names correspond to different intramolecular bond length data collection methods as described in the Section 5.4.1.

Element	Atomic No.	A (1)	B (2)	B-ORG (3)	C (4)	C-ORG (5)
Sr	38	1.900	1.865	1.865	1.865	1.864
Y	39	1.504	1.731	1.731	1.731	1.731
Zr	40	1.580	1.734	1.734	1.734	1.734
Nb	41	1.213	1.672	1.700	1.672	1.700
Mo	42	1.053	1.725	1.733	1.725	1.733
Tc	43	1.101	1.326	1.395	1.326	1.395
Ru	44	1.234	1.411	1.416	1.411	1.416
Rh	45	1.198	1.381	1.381	1.381	1.381
Pd	46	1.333	1.379	1.365	1.379	1.365
Ag	47	1.572	1.656	1.656	1.656	1.656
Cd	48	1.644	1.582	1.599	1.582	1.599
In	49	1.168	1.158	1.444	1.156	1.444
Sn	50	1.448	1.410	1.396	1.410	1.396
Sb	51	1.136	1.227	1.444	1.227	1.444
Te	52	1.109	1.097	1.269	1.097	1.269
I	53	1.182	1.152	1.152	1.152	1.152
Xe	54	1.283	1.313	1.312	1.313	1.312
Cs	55	2.279	2.252	2.452	2.252	2.452
Ba	56	2.194	2.189	2.209	2.179	2.201
La	57	1.880	1.844	1.843	1.844	1.843
Ce	58	1.690	1.816	1.817	1.816	1.817
Pr	59	1.840	1.806	1.805	1.806	1.805
Nd	60	1.819	1.783	1.777	1.783	1.777
Sm	62	1.789	1.774	1.774	1.774	1.774
Eu	63	1.737	1.699	1.698	1.699	1.698
Gd	64	1.711	1.675	1.675	1.675	1.675
Tb	65	1.692	2.175	2.182	2.175	2.182
Dy	66	1.679	1.640	1.639	1.640	1.639
Ho	67	1.693	1.696	1.654	1.696	1.654
Er	68	1.685	1.649	1.646	1.649	1.646
Tm	69	1.667	1.647	1.646	1.647	1.646
Yb	70	1.766	1.709	1.709	1.709	1.709
Lu	71	1.497	1.683	1.682	1.683	1.682
Hf	72	1.529	1.684	1.693	1.684	1.692
Ta	73	1.413	1.455	1.449	1.455	1.449
W	74	1.057	1.657	1.658	1.657	1.658
Re	75	1.202	1.501	1.503	1.501	1.503

Note: Column names correspond to different intramolecular bond length data collection methods as described in the Section 5.4.1.

Element	Atomic No.	A (1)	B (2)	B-ORG (3)	C (4)	C-ORG (5)
Os	76	1.220	1.421	1.421	1.421	1.421
Ir	77	1.218	1.408	1.407	1.408	1.407
Pt	78	1.351	1.305	1.285	1.305	1.285
Au	79	1.340	1.290	1.267	1.290	1.267
Hg	80	1.336	1.418	1.528	1.418	1.528
Tl	81	1.684	1.630	1.626	1.630	1.626
Pb	82	1.708	1.798	1.956	1.798	1.956
Bi	83	1.603	1.885	1.927	1.885	1.927
Po	84	1.678	1.678	-	1.678	-
Th	90	1.793	1.868	1.868	1.868	1.868
U	92	1.184	1.778	1.779	1.778	1.779
Np	93	1.108	1.757	1.761	1.757	1.761
Pu	94	1.154	1.999	2.008	1.999	2.008
Am	95	1.795	1.761	1.761	1.761	1.761

Note: Column names correspond to different intramolecular bond length data collection methods as described in the Section 5.4.1.

Appendix 2
Classes with most errors for each atomic radii table

Method	No.	Overlong		Missing	
		Class	Amount	Class	Amount
A (1)	1.	N-N	57317	Cu-Cu	2690
	2.	Fe-N	19736	Ag-Ag	2144
	3.	C-N	11247	Li-Li	1241
	4.	C-Ti	8839	Co-Co	1085
	5.	Co-N	8217	Ba-Ba	1069
B (2)	1.	Mn-N	7331	Mo-Mo	6591
	2.	Mn-O	6812	V-V	3120
	3.	C-Mn	2279	Cu-Cu	3048
	4.	I-I	1329	Fe-Fe	2724
	5.	Cl-Mn	946	Ag-Ag	2500
B - ORG (3)	1.	Mn-N	7440	Mo-Mo	6640
	2.	Mn-O	6854	V-V	3443
	3.	C-Mn	2309	Cu-Cu	3120
	4.	I-I	1329	Fe-Fe	2713
	5.	Cl-Mn	949	Ag-Ag	2500
C (4)	1.	Mn-N	7331	Mo-Mo	6591
	2.	Mn-O	6812	V-V	3120
	3.	C-Mn	2279	Cu-Cu	3048
	4.	I-I	1329	Fe-Fe	2724
	5.	Cl-Mn	946	Ag-Ag	2500
C-ORG (5)	1.	Mn-N	7440	Mo-Mo	6640
	2.	Mn-O	6855	Cu-Cu	3443
	3.	C-Mn	2309	V-V	3120
	4.	I-I	1329	Fe-Fe	2713
	5.	Cl-Mn	949	Ag-Ag	2500
Meng and Lewis (1991)	1.	K-O	9460	Cu-Cu	5358
	2.	Na-O	8259	Cu-O	3473
	3.	Li-O	7071	Fe-Fe	2868
	4.	Li-N	3442	Mo-Mo	2680
	5.	Ca-O	3148	Au-Au	2594

Method	No.	Overlong		Missing	
		Class	Amount	Class	Amount
Cordero et al. (2008)	1.	Hg-N	562	Mo-Mo	4358
	2.	N-Pb	313	Fe-Fe	3795
	3.	N-Zn	311	V-V	3053
	4.	Cd-O	295	Cu-Cu	2579
	5.	O-Pb	279	Co-Co	2127
Pyykkö and Atsumi (2009)	1.	Co-O	3527	Fe-Fe	1685
	2.	Ni-O	3371	Li-Li	1204
	3.	Mn-O	3190	Cu-Cu	1050
	4.	Mn-N	3054	Ru-Ru	925
	5.	Fe-N	2134	Ag-Ag	869