NEW UNIFORM AND REVERSIBLE REPRESENTATION OF 3D CHEMICAL STRUCTURES

Jure Zupan, Marjan Vračko, Marjana Novič
National Institute of Chemistry, 1000 Ljubljana, Hajdrihova 19, Slovenia

Received 22.8.1999

Abstract

New uniform and reversible *spectrum-like* representation of 3D chemical structures is explained. On a simple example of 3D structure of ethane, both the coding and decoding procedures are explained in detail. The *spectrum-like* representation is based on the projection of atoms specified by co-ordinate triplets $[x_i, y_i, z_i]$ on an arbitrarily large sphere using a Lorentzian shaped function dependent on atoms' position in the space. The new structure representation of a molecule with N atoms is defined as n-dimensional vector $\boldsymbol{S} = (s_1,s_2,..s_i,...s_n)$ with each component defined as a cumulative intensity s_i, at a given point i on the circle with and arbitrary radius. The cumulative intensity s_i (the i-th point on the circle at angle φ_i.) is a sum of N contributions s_{ji} of each atom j in the molecule.

$$s_i = \sum_{j=1}^{N} s_{ji} = \sum_{j=1}^{N} \frac{\rho_j}{(\varphi_i - \varphi_j)^2 + \sigma_j^2} \quad , \qquad i=1...number\ of\ divisions\ on\ a\ circle$$

The intensity function s_{ji} can be any bell shaped function. In our case the Lorentzian shape with maximum at angle φ_j, maximal intensity proportional to the ρ_j, and having the width, σ_j, related to the type of the new representation atom was chosen. The new representation is suitable for studying (modeling) various properties on a series of molecules having common skeletons or common structural parts. The changes of a new representation if a substituent on a given molecule is rotated is shown on an examples.

Introduction

Coding of chemical structures encompasses many different ways and different strategies for distinguishing different compounds, solutions, and materials. The computer era has initiated many different structure representations for efficient computer handling:

from fragment codes (Wiswesser-Line-Notation or WLN [1]), atom connectivity tables and topology indices [2] to sophisticated coding enabling 3-D similarity searches [3]. Unfortunately, many of such notations, notably chemical names, WLN, and the connection table representations are not uniform. This means that the representation cannot be expressed as a vector in the same *n*-dimensional space (fixed n for all structures in the study). Additionally, most of the nowadays used uniform molecular representations (topology indices, gnomic projections and Kohonen maps, for example) are not reversible.

An interesting aspect of structure coding history is that most of the attempts to find a new structure code were driven by a specific property elucidation goal in mind. With other words, the representations are mainly not proposed or invented just for its own sake, but with an aim to explain a property, or a set of properties, for which it is believed that is (are) related to the complete structure or at least to specific parts of it. For example, the choice of fragments is usually accomplished in such a way that they could be easily recognized by an appropriate analytical or spectroscopic method. Similarly, the hydrogen atoms in the connection tables are omitted because it is believed that in most applications hydrogen atoms are not needed or can be easily "attached" to the structure representation when such a need occurs. A good structure representation for modeling (be classical *via* analytical functions or by means of artificial neural networks) should fulfill the following four demands:

1. each compound should have one and only one code; with different codes for different structure (uniqueness),
2. each compound should be represented by the same number and type of variables (uniformness),
3. the structure should be possible to retrieve back from the code (reversibility), and
4. the rotated and/or translated structures should have the same code (translation and rotational invariance).

It is not difficult to fulfill each of the above four conditions separately, however, to fulfill all of them using the same representation is a very difficult task which has not yet being solved satisfactorily. The crucial property of any representation for modeling is its uniformness, i.e., all objects in the study (in our case the objects are molecular structures of compounds) must be represented with the same number of variables. For example, regardless of how many atoms the molecules consist off, all of them must be described by the same number of variables. This requirement is hard to fulfill, especially if the 3-D structural features of each molecule have to be incorporated into the representation.

In the present paper we shall discuss a method for representing 3-D chemical structures which is uniform, unique and reversible, but lacks the translation and rotational invariance. In general, the lack of the origin or rotation invariance may be regarded as a drawback when judging a representation. However, in some cases (one example will be shown later on) this property, namely the dependence of the representation on the choice of co-ordinate origin, offers quite a useful property which should be explored to the maximal benefit. If a structure representation depends on the choice of the co-ordinate origin it can be used only for comparative studies on **sets** of structures that can be somehow aligned to each other either by overlapping the skeletons or some other larger structural parts. The liberty to chose any co-ordinate origin offers a possibility to find **the** position of the origin from which the representation is most useful for a given task, i.e., the position from which the most relevant structural features could be characterized or "seen" best.

Uniformness as the most important aspect of any structure representation for handling a set of structures enables direct or inverse modeling of structure-property relationships. Without uniform representation no modeling (direct or inverse) is possible. There are immense variations of the problems that can be solved by modeling: from spectra-structure elucidation and structure-to-spectra simulation to the quantitative structure activity relationship (QSAR) problems (see for example ref.[4]). For solving this kind of problems the uniformity of the structure code is mandatory. The reason for this is that models (achieved either by the analytical functions or by artificial neural networks) require uniform code of objects. In formal words: mapping of chemical structures into the space of the sought property (or properties $Y(y_1,..y_n)$) can only be achieved from a measurement space having well defined metrics. Any model **A** can be built only if all objects in the study (chemical structures) are represented by vectors $\boldsymbol{X_S}$ having the same number of variables (axes) m, for each object:

$$\boldsymbol{Y_S}(y_{s1}, \ldots y_{sn}) = \mathbf{A}[\boldsymbol{X_S}(x_{s1}, x_{s2}, \ldots .x_{sm}), a_{11}, \ldots a_{ji} \ldots] \qquad /1/$$

The symbol **A** labels any modeling system (be a set of equations, the ensemble of neurons, or any other operator). The model **A** defined by the parameters a_{ji}, called coefficients, weights, pointers, or similar, performs the required mapping from the m-dimensional measurement space of structure representations $\boldsymbol{X_S}$ into the n-dimensional space of properties $\boldsymbol{Y_S}$.

Because the ultimate goal of modeling (for prediction) the properties of chemical compounds is to obtain knowledge how different structural parts (substituents, radicals, fragments, substructures, skeletons, etc.) influence the properties of interest, a 3D structure representation must be reversible. This means that it should be possible to decode the complete 3D structure from the representation. Unfortunately, any of the existing uniform structure representations that fulfills the condition of uniformness (gnomic projection [5], Gasteiger's approach [6] and several coding based on different topological indices [7] do not allow such an inversion. For example, from a set of either topological indices or any other set of descriptors (physico-chemical, biological or other) the structure cannot be reconstructed. The *spectrum-like* representation of 3D chemical structures discussed in this paper assures the uniformness and the reversibility.

Structure representation by projection

The main idea of the *spectrum-like* representation [8],[9] is to mimic a "light source" placed somewhere close to the molecule which casts "shadows" of atoms on the surface of an imaginary sphere drawn around the light source (Figure 1).

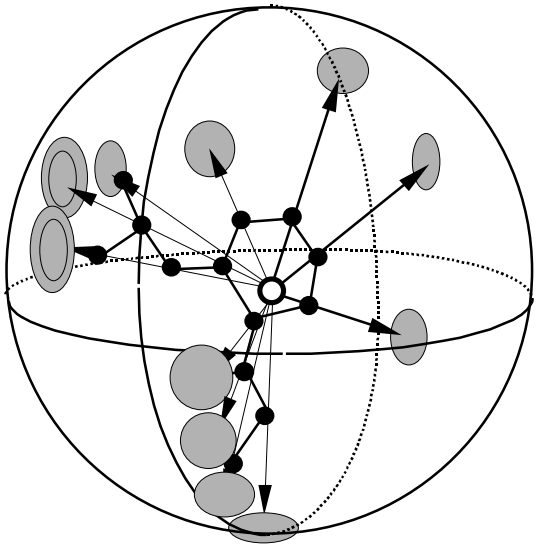

Figure 1 The molecule is placed into a sphere with an arbitrary radium R. At the centrum of the sphere is a light source which casts shadows of atoms on the spheres' surface. The intensity of the shadow of each atom is described by the 2D Lorentzian bell-shape function and depends on the distance of the atom from the light source and on the parameter σ as described by Equation /2/ For the actual calculation of the structure representation the shadows' projections into three perpendicular equatorial rings are calculated.

The positions and intensities of atoms' shadows on the surface of the sphere depend on the relative positions between the atoms and the light source. The complete shadow of

all atoms on an arbitrary equator of the imaginary sphere resembles a spectrum (Figure 2), hence, the name of the representation. The intensity s_{ji} of the shadow of an atom j (described by polar co-ordinates ρ_j and φ_j) on the equator at the point i (expressed by the polar angle φ_i) is evaluated by the Lorentzian function /2/:

$$s_{ji} = \frac{\rho_j}{(\varphi_i - \varphi_j)^2 + \sigma_j^2} \qquad /2/$$

The parameter σ_j enables each atom to be described by an additionally property, charge, for example. The Lorentzian function is chosen because of its simplicity. Any other appropriate function could be selected, if that would matter. In order to acquire the entire 3D information of the structure the shadows of atoms are projected onto three mutually perpendicular circles. For the complete representation of N triplets $[x_j,y_j,z_j]$ the co-ordinates z_j, y_j and x_j, of atoms are set to zero in sequence, hence, defining three "planar molecules" described by the $3N$ twoplets $[x_j,y_j]$, $[x_j,z_j]$, and $[y_j,z_j]$. The obtained three planar molecules are projected onto circles in the *(x,y)*, *(x,z)*, and *(y,z)*-planes, respectively, forming three equivalent sets of the new representation ***S.*** For the explanation of the Lorentzian projections of atoms on the circle the polar co-ordinates are more plausible than Cartesian ones. In the actual calculations when the atoms are described by twoplets of Cartesian co-ordinates the Equation /2/ is used in the rewritten form with Cartesian co-ordinates (see Equations /3/ and /4/. Figure 2 shows how for each atom j its shadow's intensity s_{ji} depends on the angle j_i on one circle.

The position of atom j is described by a polar co-ordinate pair (ρ_j,φ_j) in the internal coordinate system of the "light source". In all three notation of 2D-planes, i.e., in *(x,y)*, *(x,z)*, and *(y,z)*-planes, the first axis is abscissa while angle φ_j is between the abscissa and the vector ρ_j. The Lorentzian peak width parameter σ_j in Equations /1/ and /2/ is used to describe any individual property of the atom j (atomic or ionic radius, atomic number, ionization energy, electron affinity, charge, etc.). In many of our applications the parameters σ_j are set to σ_j = *1 + charge on atom j*. If the charge on atom j is negative σ_j is less than *1*, otherwise it is larger than *1*. Due to the fact that parameter σ_j can be specified for each atom the new proposed coding scheme is flexible enough to be adaptable to various types of problems for which "problem-specific" structure code is preferred. If only the space geometry and the shape of molecule has to be described, the parameter σ_j (Eq. /1/) is set equal to one for all atoms.

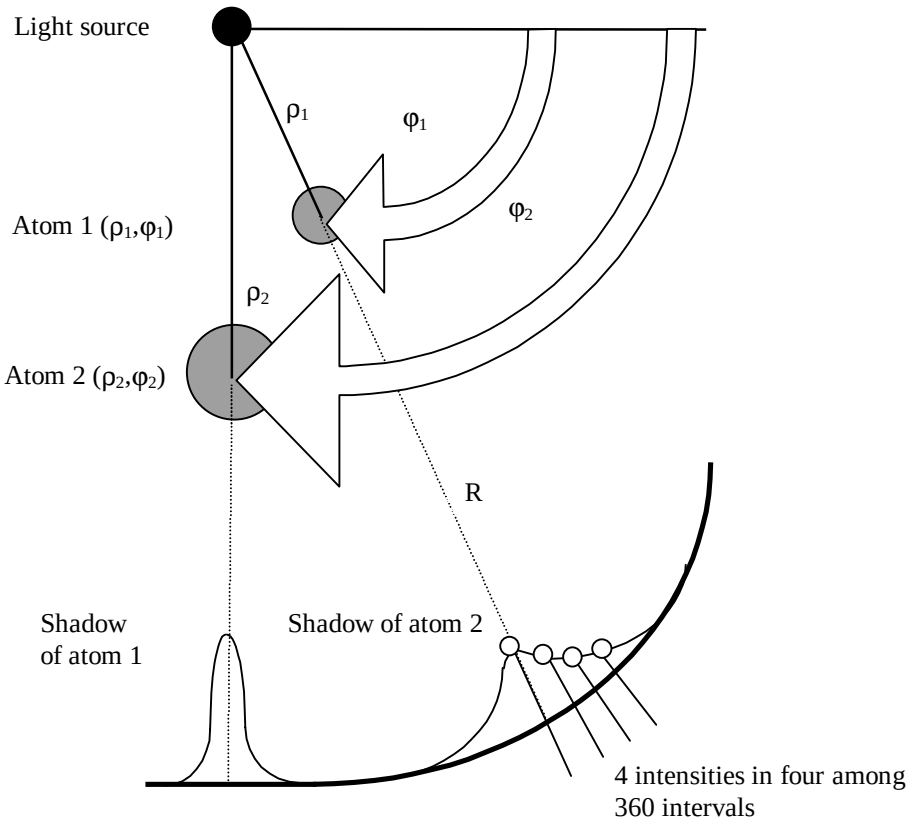

Figure 2 Contribution of atoms No. 1 and No. 2 (at the positions (r_1,φ_1) and (r_2,φ_2) to the intensity s_i at the interval (position) i on the circle with radium R, Each individual contribution has a Lorentzian bell-shape function. Circular trajectory with radium R to which the projections are made is divided into n intervals (360 in our example).

The intensities s_{ji} of *spectrum-like* representations at all positions φ_i are clearly additive. The cumulative formula for each variable s_i of the *spectrum-like* representation for the whole structure consisting of N atoms can be written as a sum:

$$s_i = \sum_{j=1}^{N} s_{ji} = \sum_{j=1}^{N} \frac{\rho_j}{(\varphi_i - \varphi_j)^2 + \sigma_j^2} \qquad \text{with } \varphi_i \text{ running from } \varphi_1 \text{ to } \varphi_{360} \qquad /3/$$

The number of variables s_i in each representation depends on the number of angles j_i which divides the equator around the molecule - the finer the division, the more precise the description. If the resolution of 1^o radial degree is chosen for the projection on each equator one spectrum has *360* intensities. Hence, a complete *spectrum-like* representation of each structure (projections of its structure into three perpendicular equators) has *1080* intensities. In general, the division of the circles should be adapted to the number of atoms N in the largest molecule of the study. After the division n is chosen, the new representation should be able to map each molecule, regardless of the number of its constituent atoms, into the same $3n$-dimensional space. In many studies where only

approximate positions of the substituents with the respect to the skeleton are sought the division *n* can be as low as *36* or even *n=18*. On the other hand, in cases when small differences in space positions of atoms are important, or for precise recoveries of structures back from the *spectrum-like* representations *n* can be as large as *720* (division to 0.5^{o}), thus making the full *spectrum-like* representation *2160* intensities (variables) long.

For actual evaluations of spectral intensities from Cartesian co-ordinate twoplets the following transformations are substituted into the Equation /2/:

$$\rho_j(x,y) = \sqrt{x_j^2 + y_j^2}\,, \quad \rho_j(y,z) = \sqrt{y_j^2 + z_j^2}\,, \quad \rho_j(x,z) = \sqrt{x_j^2 + z_j^2} \qquad /4/$$

$$\cos\varphi_j(x,y) = \frac{x_j}{\sqrt{x_j^2 + y_j^2}}, \qquad \cos\varphi_j(y,z) = \frac{y_j}{\sqrt{y_j^2 + z_j^2}}, \qquad \cos\varphi_j(x,z) = \frac{x_j}{\sqrt{x_j^2 + z_j^2}} \qquad /5/$$

hence:

$$s_i(x,y) = \sum_{j=1}^{N} \frac{\sqrt{x_j^2 + y_j^2}}{\left[\arccos\left(\frac{x_j}{\sqrt{x_j^2 + y_j^2}}\right) - \varphi_i(x,y)\right]^2 + \sigma_j^2} \qquad /6/$$

and equivalent forms for the intensities in *(x,z)* and *(y,z)* projections.

Because it is evident that the *spectrum-like* representation depends on the position of the light source there is of course a question how this position should be defined or found. For a modeling of properties applied in a connection with a class of compounds having identical skeletons or at least identical small substructure (a group of atoms), all molecules under study are first brought into such position that the common skeleton or substructure overlaps as much as possible. After this is achieved, the co-ordinate origin is either set at one of the atoms or at the center of gravity of the skeleton, or is found by any more sophisticated optimization method. The optimization criterion in a search for the optimal position of the co-ordinate origin is usually the broadest distribution of variances among all variables of the representation.

Coding of structure

As an example how the suggested representation $S = (s_1,s_2,..s_i,...s_n)$ can be obtained from the structure's *[x,y,z]* co-ordinates, the entire procedure for calculating the variables σ_i will be explained on a simple molecule of ethane (Figure 3).

The *[x,y,z]* co-ordinates of all eight atoms are given in Table I. Using Equation /2/ the Cartesian co-ordinates are transferred into the polar ones and the corresponding radii and angles of atoms in each of the three projections are calculated. For each atom the angular position (φ_i) at which the Lorentzian bell-shaped curve has its maximum and the maximal intensity ρ_i calculated from Equation /2/ are given for all three projections.

Table I. The co-ordinates of eight ethane atoms are in columns 3-5. In columns 6 to 11 are maximal intensities ρ_j and peak positions φ_i calculated according to Equations /5/ for all three projections.

j	Atom	x_j[Å]	y_j[Å]	z_j[Å]	$\rho_j(x,y)$	$\varphi_j(x,y)$	$\rho_j(x,z)$	$\varphi_j(x,z)$	$\rho_j(y,z)$	$\varphi_j(y,z)$
1	C	0.77	0.00	0.00	0.77	0.	0.77	0.	0.00	-
2	H	1.13	-0.74	0.74	1.35	327.	1.35	33.	1.05	135.
3	H	1.13	-0.27	-1.00	1.16	347.	1.51	318.	1.05	256.
4	H	1.13	1.00	0.27	1.51	42.	1.16	13.	1.05	15.
5	C	-0.77	0.00	0.00	0.77	180.	0.77	180.	0.00	-
6	H	-1.13	-1.00	-0.27	1.51	222.	1.16	193.	1.05	195.
7	H	-1.13	0.27	1.00	1.16	167.	1.51	138.	1.05	76.
8	H	-1.13	0.74	-0.74	1.35	146.	1.35	213.	1.05	316.

In order to obtain the complete *spectrum-like* representation as many Lorentzian curves as there are atoms in the molecule have to be added. All three resulting *spectrum-like* representations and the corresponding projections of the ethane molecule in the *(x,y)*, *(x,z)*, and *(y,z)*-planes are shown in Figures 3a, 3b, and 3c. By projecting any number of atoms onto a circular line of length 2π, the uniformity of the code is achieved. The actual resolution used for a specific application depends on the needs and computational capability of the user.

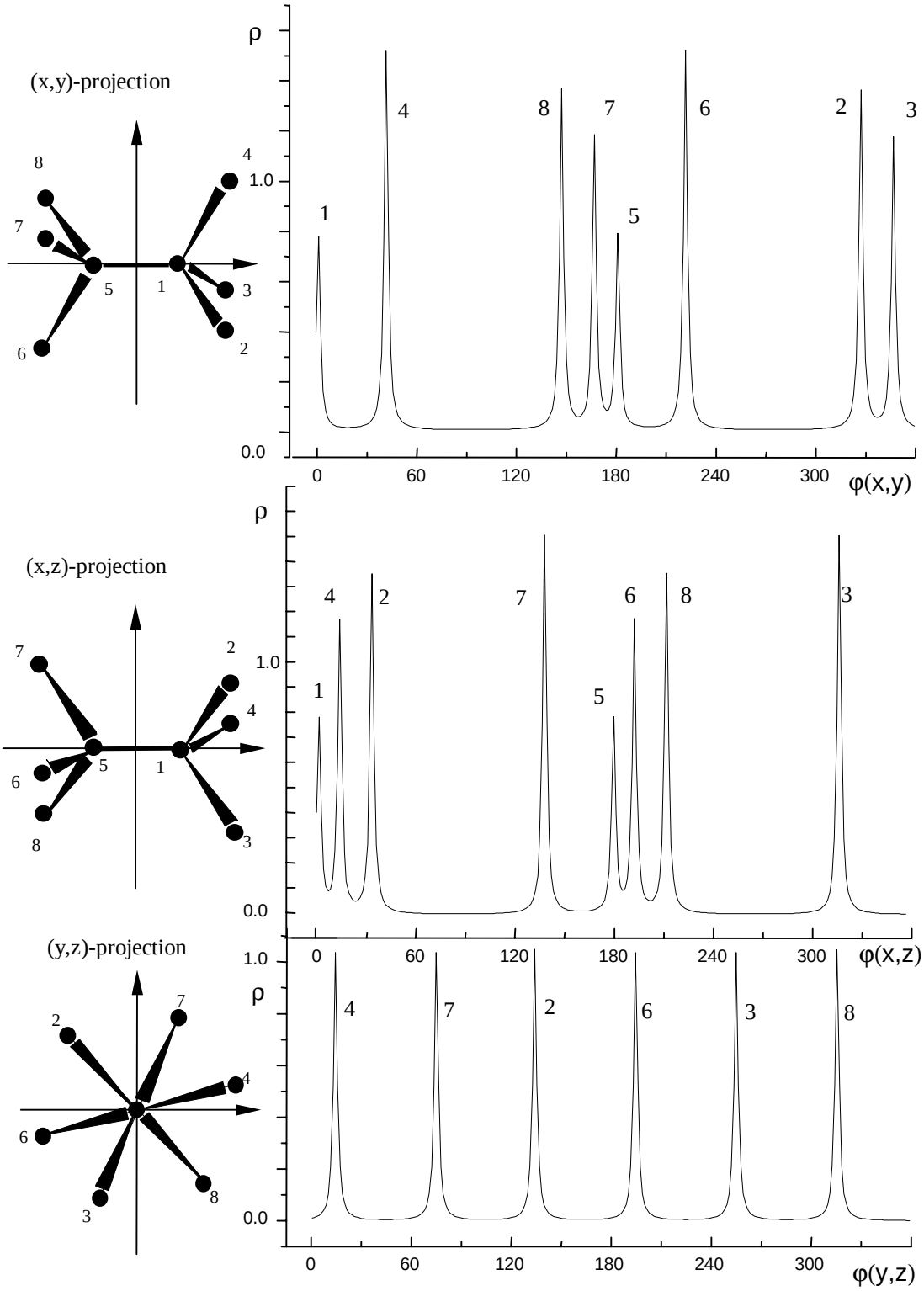
(x,y)-projection
ρ
1.0
0.0
0
60
120
180
240
300
φ(x,y)
(x,z)-projection
φ(x,z)
(y,z)-projection
φ(y,z)

Figure 3 *Spectrum-like* structure representations of ethane projected in *(x,y), (x,z),* and *(y,z)*-planes. The assignments of the peaks to eight ethane atoms are shown with figures standing next to the peaks. The intensities and positions are listed in Table I.

The smaller $2\pi/n$ ratio the larger the resolution of the representation. The more atoms has the largest molecule in the study, the finer is the resolution of the *spectrum-like* representation (division of the 2π circle into *n* intervals) is.

The translation invariance of the code is assured by setting the co-ordinate origin of the sphere within which the molecules are placed into the center of all *[x,y,z]* co-ordinates of the molecule. In many applications a certain atom (for example, atom common to all structures) is set into a center of co-ordinate origin. If any atom is in the central position it has no peak in the *spectrum-like* representation (see *(y,z)*-projection in the following example - Fig. 3c).

Decoding the spectrum-like *representation*

The basis for the reverse evaluation of *x, y,* and *z* co-ordinates of atoms are locations of peak positions in *spectrum-like* representations, i.e. the angles $\varphi_i = \varphi_j$ where maximal intensities ρ_j *(x,y),* ρ_j*(y,z),* and ρ_j*(x,z)* can be easily calculated using Equation /3/. In the next step the Cartesian coordinates are obtained:

$$x = \rho_j(x,y) \cos \varphi_j(x,y)$$
$$y = \rho_j(y,z) \cos \varphi_j(y,z)$$
$$z = \rho_j(x,z) \sin \varphi_j(x,z)$$

/7/

The last thing to do is to assign the obtained *x, y,* and *z* co-ordinates to the correct atoms. This assignment is not always straightforward because due to the symmetry of molecule some ambiguities can arise. Nevertheless, such cases can be easily handled by following a step-by step evaluation of the *x* and *y* co-ordinates and after that the eventual multiple choices for the assignment of the *z* co-ordinates can be resolved by logical elimination. An example of the decoding procedure is shown in Tables II and III. In this example the orientation of a molecule for which, due to its symmetric position, several possibilities for assignments of *[x, y]* pairs to the *z* co-ordinate arises, has been intentionally chosen. For the case of simplicity the σ_j values were taken to be 1.

Tables II-IV contain the information available for decoding, i.e., if all three *spectrum-like* representations of an unknown molecule are given. It should be noticed

that there is a small variation in coordinates' recovery due to the uncertainty of the resolution of 1 radial degree in all *spectrum-like* representations. Therefore, the average values for recovered co-ordinates with only two decimal places are given.

Table II. The information available from the *spectrum-like* representation in the (x,y)-plane (see Figure 3a).

Peak j	Peak position $\varphi_j(x,y)$ [deg]	Peak intensity $\rho_j(x,y)$ [Å]	$\cos \varphi_j(x,y)$	x $\rho_j \cos \varphi_j$ [Å]	$\sin \varphi_j(x,y)$	y $\rho_j \sin \varphi_j$ [Å]
1	0	0.77	1.000	0.77	0.000	0.00
2	42	1.51	0.743	1.12	0.669	1.01
3	146	1.35	-0.829	-1.12	0.559	0.75
4	167	1.16	-0.974	-1.13	0.225	0.26
5	180	0.77	-1.000	-0.77	0.000	0.00
6	222	1.51	-0.743	-1.12	-0.669	-1.01
7	327	1.35	0.839	1.13	-0.555	-0.75
8	347	1.16	0.974	1.13	-0.225	-0.26

Table III. The information available from the *spectrum-like* representation in the (x,z)-plane (see Figure 3b).

Peak j	Peak position $\varphi_j(x,z)$ [deg]	Peak intensity $\rho_j(x,z)$ [Å]	$\cos \varphi_j(x,z)$	x $\rho_j \cos \varphi_j$ [Å]	$\sin \varphi_j(x,z)$	z $\rho_j \sin \varphi_j$ [Å]
1	0	0.77	1.000	0.77	0.000	0.00
2	13	1.16	0.974	1.13	0.225	0.26
3	33	1.35	0.839	1.13	0.545	0.74
4	138	1.51	-0.743	-1.12	0.669	1.01
5	180	0.77	-1.000	-0.77	0.000	0.00
6	193	1.16	-0.974	-1.13	-0.225	-0.26
7	213	1.35	-0.839	-1.13	-0.545	-0.74
8	318	1.51	0.743	1.12	-0.669	-1.01

Table IV. The information available from the *spectrum-like* representation in the (y,z)-plane (see Figure 3c).

Peak j	Peak position $\varphi_j(y,z)$ [deg]	Peak intensity $\rho_j(y,z)$ [Å]	$\cos\varphi_j(y,z)$	y $\rho_j \cos \varphi_j$ [Å]	$\sin \varphi_j(y,z)$	z $\rho_j \sin \varphi_j$ [Å]
1	15	1.05	0.966	1.01	0.259	0.27
2	76	1.05	0.242	0.25	0.970	1.02
3	135	1.05	-0.707	-0.74	0.707	0.74

4	195	1.05	-0.966	-1.01	-0.259	-0.27
5	256	1.05	-0.242	-0.25	-0.970	-1.02
6	316	1.05	0.719	0.75	-0.695	-0.73

Two peaks (1 and 5) in Tables II and III are uniquely defined (having the same *x* coordinates in both tables) yielding the complete positions of two atoms: (0.77Å, 0.00Å, 0.00Å) and (-0.77Å, 0.00Å, 0.00Å). For the other six *x*-coordinates in Table II: 1.12Å, -1.12Å, -1.13Å, -1.12Å, 1.13Å, and 1.13Å, each *y*-coordinate is different: 1.01Å, 0.75Å, 0.26Å, -1.01Å, -0.75Å, and -0.26Å. The same is true for six *x*-coordinates in Table III: 1.13Å, 1.13Å, -1.12Å, -1.13Å –1.13Å, and 1.12Å; yileding *z*-coordinates of: 0.26Å, 0.74Å, 1.01Å, -0.26Å, -0.74Å, and -1.01Å.

Within the tolerance of ±0.01Å among various *x*-coordinates (1.12 and 1.13Å for example) the correct combination of *y* and *z* co-ordinates is easily resolved by the inspection of Table IV where *y* and *z* co-ordinate pairs uniquely determine the complete triplets *[x,y,z]* of the remaining six atoms (Table V). The obtained values are in fair agreement with the original co-ordinates of eight atoms shown in Table I. The maximal discrepancy of ±0.01 comes from the low resolution of the initial division. If the division was twice the original one (i.e., one intensity per 0.5^0 instead per 1.0^0), the error would be even lower.

Table V. The recovered coordinate triplets from the three *spectrum-like* representations as given on Figure 3a-3c.

Calculated			Actual		
x [Å]	y [Å]	z [Å]	x [Å]	y [Å]	z [Å]
0.77	0.00	0.00	0.77	0.00	0.00
-1.13	-0.75	-0.74	-1.13	-0.74	-0.74
1.12	-0.26	-1.01	1.12	-0.27	-1.00
1.13	1.01	0.26	1.13	1.01	0.26
-0.77	0.00	0.00	-0.77	0.00	0.00
-1.13	-1.01	-0.26	-1.13	-1.0o	-0.27
-1.12	0.26	1.01	-1.13	0.27	1.001
1.13	0.75	0.74	1.13	0.74	0.74

Applications

One of the advantages of the proposed *spectrum-like* 3D representation of structures is its additivity with the respect to the constituent atoms. The second advantage is a strict dependence on the actual position of each atom. If one combines these two properties it is easy to perceive that any conformation of a single compound can be coded differently. Nevertheless, the differences between the representation of two conformations are directly proportional to the distortion (rotation or stretching) of the substituent in question with the respect to the remaining skeleton.

Most conformers of a given molecule can be generated by rotating a part of a molecule (a substituent) around an axis (a bond defined by two fixed atoms) for a certain angle. For more complex conformers a combination of several rotations around different axes can be combined. In order to show how the method works and how the corresponding *spectrum-like* representations change from conformer to conformer, a simple case of rotation of a methyl group at one end of the ethane will be shown. See Appendix for the mathematics of the rotation (evaluation of the new coordinates) procedure.

From the new co-ordinates the new *spectrum-lik*e representation is calculated using Equation /6/. In order to show how the procedure works the methyl substituent (atoms No.: 2, 3, and 4 on the left-hand side of Figure 3) is rotated with the respect to the fixed methyl group (atoms No.: 6, 7, and 8). The rotation of the substituent is made around the bond between both carbon atoms No. 1 and No. 5 for 20^o and 40^o in the clockwise direction (Figure 4a). The change, or better the shifts of the corresponding peaks in the two *spectrum-like* representations (contributions in the (y,z)-plane only) of the two conformers compared to the representation of the original ethane molecule are shown in Figure 4b. Because the rotation around the bond between two carbons is in the clock-wise direction the peaks corresponding to hydrogen atoms No. 2, 3, and 4 are shifted towards smaller angle.

It is evident that the other two partial contributions (from the (x,y) and (x,z)-planes) to the new s*pectrum-like* representation of the conformers are changed correspondingly. On Figure 5a and 5b the (x,y)-plane projection of the same two ethane conformers and the accompanying parts of the *spectrum-like* representations are shown.

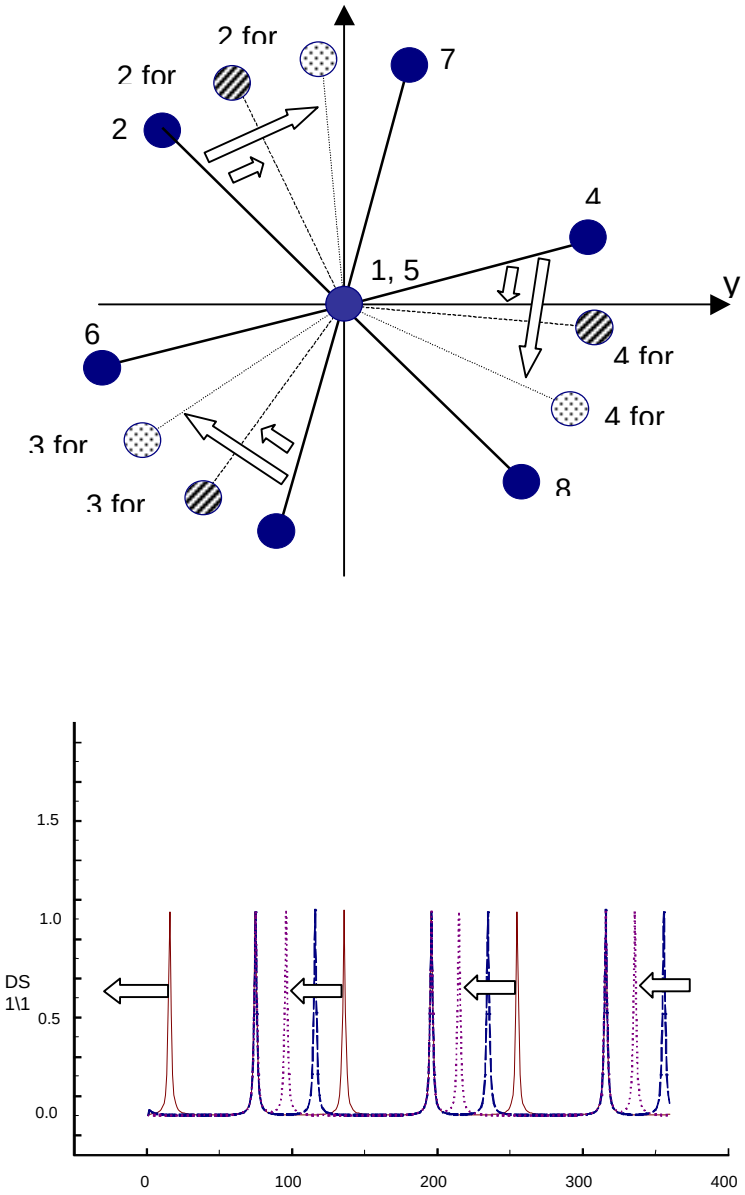

Figure 4. The (y,z)-projection of the three ethane conformers each with the rotated methyl group (hydrogen atoms No. 2, 3, and 4) for 20° in the clock-wise direction (above). The corresponding shifts in the *spectrum-like* representation are shown in (bottom). Dashed and dotted lines line represent "spectra" of two conformers rotated for 20 and 40°, respectively, from the full line *spectrum-like* representation of the original molecule (compare the representation on Figure 3 bottom).

Additionally, when coding the structures having the same skeleton or the same backbone structure (group of derivatives, analogues, etc.), all peaks in the new representation obtained for atoms of the common substructure can be simply subtracted from the representation. The subtraction of common peaks makes the representation

more sensitive to those parts of the structures that are actually relevant to the study. Due to the fact that most of the QSAR and other structure-property relationships, notably spectra-structure relationships, are made for the families of compounds, this representation can be an excellent tool for such purpose.

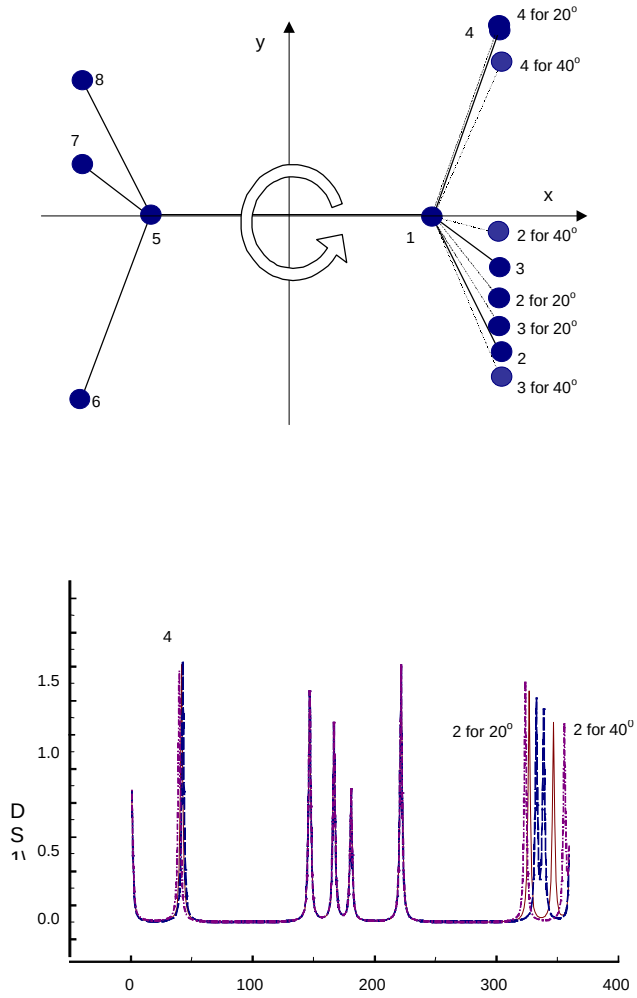

Figure 5 The (x-y)-plane projection of the same three ethane conformers (b). The arrow shows the direction of the coding of the *spectrum-like* representation. Bellow it can be seen that the small change of the atom No 4 position results in a small change of the corresponding peak positions in the second quadrant. On the other hand, relatively large changes of positions of atom No 2 result in larger changes of the peak shift.

Alternatively, a molecule can be represented by a single spectrum, which represents an ensemble of conformers. Such a spectrum can be easily calculated as an average over the spectra describing all different conformers. The average is the sum of all spectra divided by the number of conformers.

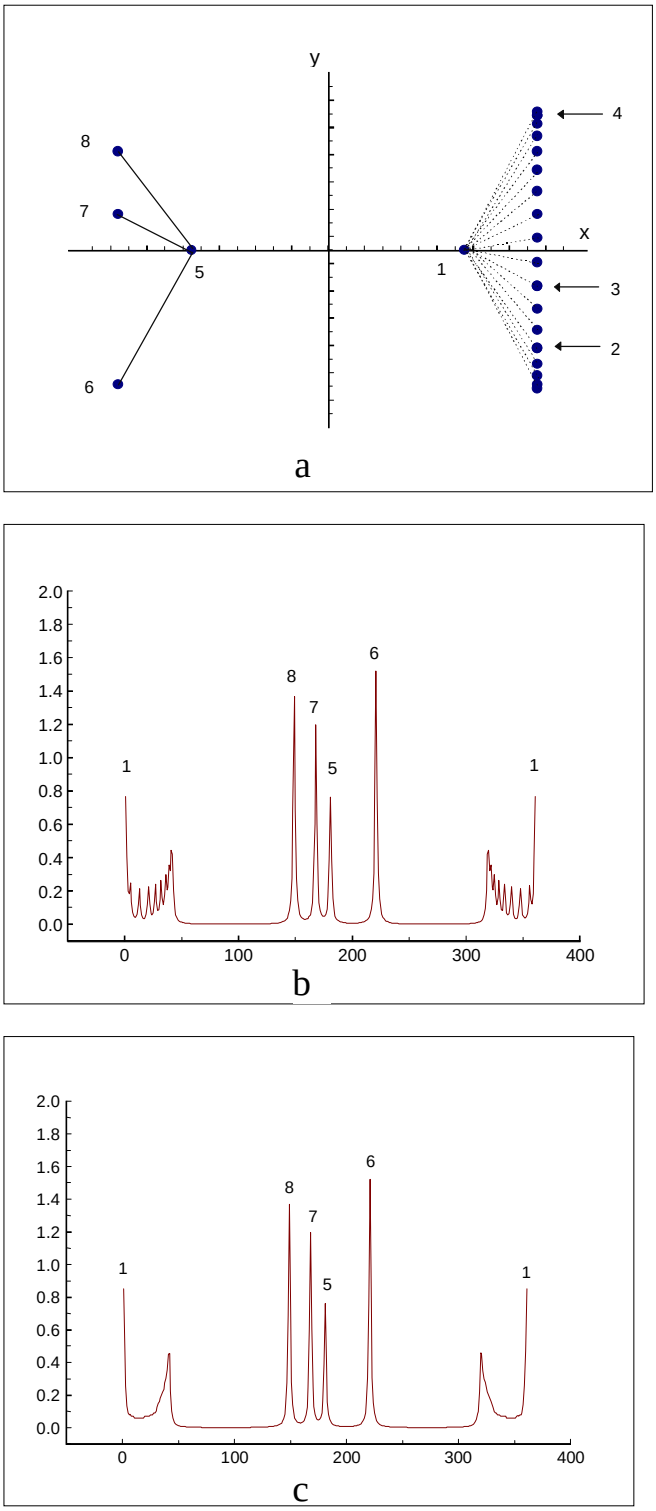

Figure 6. The (x-y) projection of six ethane rotamers (a). The atom Nos. 2, 3, and 4 are rotated each for 20° around the axis C1-C2. The average spectrum of six rotamers in the (x-y) plane exhibits clearly distinguished peaks for each rotated ataom (b). In the limit case the separate peak are merged into a continous concave shoulder (c). The peak in the middle of the shoulder corresponds to the atom No. 1.

As an example the three hydrogen atoms of the ethane (No. 2, 3, 4) are rotated around the C1-C2 bond for 20^o, 40^o, 60^o, 80^o, and 100^o, and projected into the (x-y) plane (Figure 6a). The average spectrum of the six rotamers is shown in Figure 6b. Comparing Figures 5 and 6b one can see that the peaks corresponding to the fixed atoms remain unchanged, while the peaks corresponding to the flexible atoms are spread over the space window. In the (x-y) projection, this window results in two intervals: one between 0 and 45^o and the second one between 315 and 360^o. If the number of rotamers is increased the resulting peaks merge into a broader parabolic shaped shoulder (see Figure 6c). Due to different density of atom projections in the (x-y) plane the shoulder features two separate maxima at each end. The only shortcoming of the average representation is the loss of the reversibility for the free rotated atoms of the structures.

Conclusion

The described *spectrum-like* representation of the structures is important for two reasons. First, it offers a uniform, i.e., a fixed-dimensional representation in which a wide variety of different structures can be coded and second, it offers the possibility for decoding of the structure in the backward direction from the representation. Most of the contemporary uniform structure representations are based on a group of several topological, shape/form, electronic, hydrophobic and other single variable properties from which decoding of the structures is practically impossible.

We are aware that with this representation the problems of rotational invariance still remain. However, if the proposed representation is used on a set of structures that are previously oriented or aligned in the same way, the code of the structures can easily be compared to each other. Even more, if structures in question are aligned to the same external co-ordinate system the feed-back information from the model about the parts of the structures and their spatial distributions responsible for the modeled property (for example a biological activity) can be deduced.

Acknowledgment

The financial support of the support of Ministry of Science and Technology of Slovenia within the Projects J1-8900 and J1-8901 is gratefully acknowledged.

References

[1] W. J. Wiswesser, *A Line-Formula Chemical Notation*; T.Y. Crowell, New York, 1954; and E. G. Smith, *The Wiswesser Line-Formula Chemical Notation*; McGraw Hill, New York, 1968.
[2] As review of different coding systems see for example: J. E. Ash, W. A. Warr, P. Willett, *Chemical Structure Systems*; Ellis Horwood, New York, 1991.
[3] See for example: *Concepts of Molecular Similarity*; M. A. Johnson and G. M. Maggiora, Eds., Wiley Interscience, New York, 1990.
[4] C. Hansh, A. Leo, *Exploring QSAR*; ACS Professional Reference Book, ACS, Washington, D.C., 1995, Chapter 3.
[5] P. L. Chau, P. M. Dean, Molecular recignition: 3D surface structure comparison by gnomonic projection, *J. Mol. Graphics* **1987**, 5, 97-100.
[6] J. H. Schuur, P. Selzer, J. Gasteiger, The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlation and Studies of Biological Activity, *J. Chem. Inf Comput. Sci*. **1996**, 36, 334-344,
[7] *From Chemical Topology to Three-Dimensional Geometry*; Ed. A. T. Balaban, Plenum Press, New York, 1997.
[8] M. Novič, J. Zupan, A New General and Uniform Structure Representation, in *Software Development in Chemistry 10, GDCh*; Editor: J. Gasteiger, Frankfurt a.M. 1996, p. 48-58.
[9] J. Zupan, M. Novič, General Type of a Uniform and Reversible Representation of Chemical Structures, *Anal. Chim. Acta* **1997**, 348, 409-418.

Povzetek

Predstavljena in razložena je nova enotna in reverzibilna spektralna predstavitev 3D kemijskih struktur. Celoten postopek kodiranja in dekodiranja nove predstavitve je razložen na primeru molekule etana. Nova spkralna predstavitev je zasnovana na projekciji vsakega posameznega atoma opisanega s trojico koordinat *[x,y,z]* na površino poljubno velike krogle. Projekcija je narejena z zvonasto krivuljo, katere intenziteta in širina sta odvisni od položaja atoma v prostoru in vrste atoma. Nova strukturna predstavtev molekule z *N* atomi je definirana kot n-dimenzionalni vektor $S=(s_1,s_2,\ldots s_n)$. Vsaka komponenta s_i predstavlja kumulativ-no intenziteto v točki i na ekvatorju omenjene poljubne krogle. Kumulativna intenziteta s_i na mestu *i* je vsota *N* prispevkov s_{ji} od vsakega atima *j* v molekuli :

$$s_i = \sum_{j=1}^{N} s_{ji} = \sum_{j=1}^{N} \frac{\rho_j}{(\varphi_i - \varphi_j)^2 + \sigma_j^2} \quad ,$$

i=1..število razdelkov na ekvatorju

Funkcija s katero računamo intenziteto, je načeloma lahko vsaka zvonasta krivulja. V našem primeru smo uporabili Lorentzovo funcijo. Nova predstavitev je posebej uporabna za modeliranje različnih lastnosti družin molekul z istim skeletom. Sprememba predstavitve molekule pri zasuku substituente je prikazana na primeru.

Appendix

The mathematics for obtaining the new co-ordinates if only a part of a structure is rotated around an arbitrary bond is the following (for numbering of atoms see Figure 3):

1. select the bond, i.e., two atoms A_5 and A_1 described by the coordinate triplets $[x_5,y_5,z_5]$ and $[x_1,y_1,z_1]$, around which the substituent will be rotated,

2. translate the coordinate system to atom A_5. This done by subtracting the triplet $[x_5,y_5,z_5]$ from coordinate triplets of all atoms. The new coordinates of atom A_1 $(x_1\text{-}x_5,\ y_1\text{-}y_5,\ z_1\text{-}z_5)$ is written as $A_1(x_0,\ y_0,\ z_0)$,

3. using matrix $\boldsymbol{T}$ rotatate the coordinate system with the origin in the atom A_5 in such a way that the new x-axis will point in the direction from A_5 to A_1 (X' and X stand for any rotated and old coordinate triplet of each atom in the molecule, respectively),

$$X' = \mathbf{T}X$$

$$T = \frac{1}{R}\begin{Vmatrix} x_0 & y_0 & z_0 \\ -\dfrac{x_0 y_0}{r} & -\dfrac{y_0 z_0}{r} & r \\ y_0 R/r & -x_0 R/r & 0 \end{Vmatrix}$$

4. rotate all atoms of the substituent on the bond A_1-A_2 with coordinates X' for angle φ:

where $r = \sqrt{x_0^2 + y_0^2}$ *and* $R = \sqrt{x_0^2 + y_0^2 + z_0^2}$

5. rotate the new coordinates of the entire molecule back to the original position:

$$X'' = \begin{Vmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & \sin\varphi \\ 0 & -\sin\varphi & \cos\varphi \end{Vmatrix} X'$$

$$X = \frac{1}{R}\begin{Vmatrix} x_0 & -\dfrac{x_0 z_0}{r} & \dfrac{y_0 R}{r} \\ y_0 & -\dfrac{y_0 z_0}{r} & -\dfrac{x_0 R}{r} \\ z_0 & r & 0 \end{Vmatrix} X''$$