Supporting Information for

Vibrational Spectroscopic Map, Vibrational Spectroscopy, and Intermolecular Interaction

Carlos R. Baiz[1], Bartosz Błasiak[2], Jens Bredenbeck[3], Minhaeng Cho[*,4,5], Jun-Ho Choi[6], Steven A. Corcelli[7], Arend G. Dijkstra[8], Chi-Jui Feng[9], Sean Garrett-Roe[10], Nien-Hui Ge[11], Magnus W. D. Hanson-Heine[12], Jonathan D. Hirst[12], Thomas L. C. Jansen[13], Kijeong Kwac[4], Kevin J. Kubarych[14], Casey H. Londergan[15], Hiroaki Maekawa[11], Mike Reppert[16], Shinji Saito[17], Santanu Roy[18], James L. Skinner[19], Gerhard Stock[20], John E. Straub[21], Megan C. Thielges[22], Keisuke Tominaga[23], Andrei Tokmakoff[9], Hajime Torii[24], Lu Wang[25], Lauren J. Webb[26], and Martin T. Zanni[27]

[1]Department of Chemistry, University of Texas at Austin, Austin, TX 78712, U.S.A. [2]Department of Physical Chemistry and Quantum Chemistry, Wroclaw University of Science and Technology, Wybrzeże, Wyspiańskiego 27, 50-370 Wrocław, Poland. [3]Johann Wolfgang Goethe-University, Institute of Biophysics, Max-von-Laue-Str. 1, 60438, Frankfurt am Main, Germany.[4]Center for Molecular Spectroscopy and Dynamics, Seoul 02841, Republic of Korea. [5]Department of Chemistry, Korea University, Seoul 02841, Republic of Korea. [6]Department of Chemistry, Gwangju Institute of Science and Technology, Gwangju 61005, Republic of Korea. [7]Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN 46556, U.S.A. [8]School of Chemistry and School of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, U.K. [9]Department of Chemistry, James Franck Institute and Institute for Biophysical Dynamics, University of Chicago, Chicago, IL 60637, U.S.A. [10]Department of Chemistry, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A. [11]Department of Chemistry, University of California at Irvine, Irvine, CA 92697-2025, U.S.A. [12]School of Chemistry, University of Nottingham, Nottingham, University Park, Nottingham, NG7 2RD, U.K. [13]University of Groningen, Zernike Institute for Advanced Materials, Nijenborgh 4, 9747 AG Groningen, The Netherlands. [14]Department of Chemistry, University of Michigan, 930 N. University Ave., Ann Arbor, MI 48109, U.S.A. [15]Department of Chemistry, Haverford College, Haverford, Pennsylvania 19041, U.S.A. [16]Chemical Physics Theory Group, Department of Chemistry, University of Toronto, Toronto, Ontario M5S 3H6, Canada. [17]Department of Theoretical and Computational Molecular Science, Institute for Molecular Science, Myodaiji, Okazaki, 444-8585, Japan. [18]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6110, U.S.A. [19]Institute for Molecular Engineering, University of Chicago, Chicago, IL 60637, U.S.A. [20]Biomolecular Dynamics, Institute of Physics, Albert Ludwigs University, 79104 Freiburg, Germany. [21]Department of Chemistry, Boston University, Boston, MA 02215, U.S.A. [22]Department of Chemistry, Indiana University, 800 East Kirkwood, Bloomington, Indiana 47405, U.S.A. [23]Molecular Photoscience Research Center, Kobe University, Nada, Kobe 657-0013, Japan [24]Department of Applied Chemistry and Biochemical Engineering, Faculty of Engineering, and Department of Optoelectronics and Nanostructure Science, Graduate School of Science and Technology, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu 432-8561, Japan.[25]Department of Chemistry and Chemical Biology, Institute for Quantitative Biomedicine Rutgers University, 174 Frelinghuysen Road,

Piscataway, NJ 08854, U.S.A. [26]Department of Chemistry, The University of Texas at Austin, 105 E. 24th Street, STOP A5300, Austin, Texas 78712, U.S.A. [27]Department of Chemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706-1396, U.S.A.

*Email: mcho@korea.ac.kr (MC)

Each map file is a plain text file with the extension '.vbm' denoting 'vibration map'. It consists of sections whose titles start with the '%' character. The map file essentially consists of the following three parts: (i) the identity of the molecule, the vibrational mode, and its property, e.g., frequency shift, for which the map is developed, (ii) the structural information on the concerned molecule and the interaction sites (the number of those sites and their locations) at which electrostatic perturbations should be calculated, and (iii) the parameter values for mapping those perturbations to the properties. None of the sections defined below are mandatory. The author of vbm files can select any sections necessary for his own vbm file and the sections with the same title can occur multiple times.

1. Specification of the vibrational probe molecule, the vibrational mode, and its property

The first part of each vbm file consists of the following sections for providing the name of the map, the authors, the date when this file was created, the reference (published paper) reporting the map, as well as the information on the vibrational probe molecule, the vibrational mode, and its property:

%name
%authors
%date
%references
%description

2. Structural information of the molecule and the interaction sites

The main information in this part of the vbm file is the specifications of the interaction sites at which electrostatic potentials and/or electric fields originating from the environment should be calculated. To specify a given interaction site, it is necessary to provide structural information of the molecule. The format for specifying those structural parameters for general molecules is presented below in subsection 2.a.

The cases of flexible polymeric systems need special attention because intrachain electrostatic interactions should be carefully taken into account. For each vibrational chromophore, the range that is treated as "external" in the calculations of electrostatic perturbations should be explicitly specified. In the cases of peptides and proteins, because the major vibrational chromophores are located on the peptide bonds, which are rather rigid and common to most of the combinations of connected amino acid residues, a specially-defined concise way of representing the interaction sites is often regarded as useful. The relevant file format will be explained in subsection 2.b.

2.a. Cases of general molecules
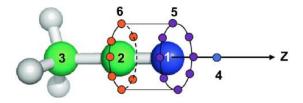The interaction sites on the vibrational chromophore are defined with the following four sections:

%numbers
        %structure
        %sites on
        %sites off

The section %numbers consists of three numbers on a single line: the number of atoms in the molecule, the number of interaction sites that are located on the atomic sites ("sites on"), and the number of interaction sites that do not coincide with the atomic sites ("sites off"). For example,

        %numbers
        6   3   17

This example represents the case of the acetonitrile molecule in Figure S1, in which there are 6 atoms in the molecule, three of which are used as interaction sites, e.g., C, C, and N atoms, and 17 interaction sites that do not coincide with the atomic sites.

Figure S1. Distribution of interaction sites for MeCN. Eight sites are distributed on the circle around N and C atoms (taken from Fig. 6 of Ref. [1]).

The section %structure contains the Cartesian coordinates of the atoms in the vibrational chromophore. The format of this section is

        %structure
        n   *atom*   *x*   *y*   *z*

where *n* is the sequential integer, *atom* is the name of the atom, and *x, y, z* are the Cartesian coordinates in Å. For example, in the case of acetonitrile:

        %structure
        1   N   0.000000   0.000000   1.424943
        2   C   0.000000   0.000000   0.289409
        3   C   0.000000   0.000000   -1.179370
        4   H   0.000000   1.018383   -1.544945
        5   H   0.881946   -0.509192   -1.544945
        6   H   -0.881946   -0.509192   -1.544945

The actual values of the atomic coordinates of the molecule may change not only by the choice of the origin and the orientations of the axes of the coordinate system but also by the methods used to obtain these coordinates, such as the type of molecular dynamics force fields or the theoretical level and the basis set in quantum chemical calculations. Therefore, the definition of the interaction sites in the "%sites on" and "%sites off" sections should not depend on the

detailed values of the coordinates in the %structure section. Instead, they are defined by the use of only the indices of the atoms in the "%structure" section.

The section "%sites on" is to specify the interaction sites that coincide with the atoms of the chromophore molecule. This section consists of lines having the sequential index of each interaction site followed by the number denoting the atomic site. For example,

```
%sites on
1    2
2    3
3    5
```

This example means that three interaction sites numbered as 1, 2, and 3 are located on the second, third, and fifth atoms of the chromophore molecule, respectively.

The section "%sites off" contains information on the interaction sites that do not coincide with any atom of the chromophore molecule. The format of this section is explained in the following sequence: (i) defining a local coordinate system, (ii) interaction sites located on the bonds, (iii) general cases of off-atom interaction sites, and (iv) defining dummy sites.

(i) Defining a local coordinate system: The four lines starting with d0, d1, d2, d3 are used to define a local molecule-fixed coordinate system. The line starting with d0 specifies the position of the local origin, the format of which is

```
d0    i
```

which denotes a selection of atom I (the ith atom in the %structure section) as the local origin. The other three lines specify mutually orthogonal unit vectors d_1, d_2, and d_3. The format of the line starting with d1 is

```
d1    j
```

which means that the unit vector d_1 is taken in the direction from atom I (taken as the origin) to atom j. The format of the line starting with dn ($n = 2, 3$) is

```
dn    j    k
```

where j and k are atom indices defined in the "%structure" section or one of the dm ($m < n$) unit vectors. Then dn specifies the unit vector that is in the direction of the cross product (vector product) of the two vectors defined by j and k. Let us consider the following example:

```
d0    2
d1    1
d2    1    4
d3    1    d2
```

In this example, the 'd0' line denotes a selection of atom 2 as the local origin. The 'd1' line defines the unit vector d_1 in the direction of r_{21}, which is the vector from atom 2 (the local origin) to atom 1. The 'd2' line defines the unit vector d_2 in the direction of the cross product $r_{21} \times r_{24}$. The 'd3' line defines the unit vector d_3 in the direction of the cross product $r_{21} \times d_2$. In Figure S2, this local coordinate system is represented for an acetonitrile molecule as an example. We emphasize that the introduction of 'd0' is

just to define the three directions specified by the d_1, d_2, and d_3 unit vectors. After the definition of these three unit vectors, the position of 'd0' will not be used any more. We will define each off-atom interaction site with its own local origin. Thus the selection of 'd0' can be done in a way that is most convenient to define the d_1, d_2, and d_3 unit vectors.
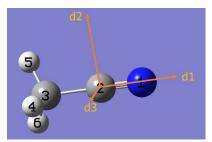

Figure S2. The local coordinate system defined by the three vectors **d_1, d_2, and d_3.**

(ii) Interaction sites located on the bonds: The format is as follows:

n b j k [r]

Here, n denotes that this line is for the nth interaction site. The letter 'b' is used to denote that this line is to define an interaction site located on the bond connecting atoms j and k. When the optional parameter 'r' is absent, the interaction site is located on the exact midpoint of the bond. When 'r' is given, $r = R_{jn} / R_{jk}$ where R_{jn} is the distance between atom j and the interaction site, and R_{jk} is the distance between atoms j and k.

(iii) General cases of off-atom interaction sites:

n I r s t

Here, n denotes that this line is for the nth interaction site. The number I is used to denoting the index of the atom to be used as the origin for defining this interaction site. The three real numbers r, s, and t are the coefficients used to define the relative position of this interaction site from atom I as $r * d_1 + s * d_2 + t * d_3$. Atom I should be nearest to the interaction site n so that the definition is least dependent on the detailed coordinate values in the section '%structure'.

(iv) Defining dummy sites: There is a special case of $n = 0$. In this case, the line begins with 0 and defines a site in the format of (ii) or (iii). And this site is not counted as an interaction site but can be used in the definition of other interaction sites. The line beginning with 0 can occur any number of times in the 'site off' section. For example,

          0   b   1   2
          1   0   0.0   0.7   0.7
          0   b   2   3
          2   0   0.0   0.7   0.7

The second line defines interaction site 1 as the vector $0.0 * d_1 + 0.7 * d_2 + 0.7 * d_3$ from site 0 which is defined by the first line as the midpoint of the bond connecting

atoms 1 and 2. Then the third line redefines site 0 as the midpoint of the bond connecting atoms 2 and 3. Then the fourth line uses this site 0 to define another interaction site.

Sites with negative n is also possible and treated in the same way as site 0. One can define a site with a negative number n to avoid overwriting by another site 0. The sites with negative numbers can be defined in any order and will not be counted as the actual interaction sites.

The numbering of the off-atom interaction sites in the format (ii) and (iii) starts from $m+1$ if there are m interaction sites defined in the '%sites on' section.

2.b. Case of peptides and proteins
For peptides and proteins, the following sections may be used to define the interaction sites:

%structure residues
%sites type
%dihedral

The section "%structure residues" consists of lines having a sequential integer followed by the name of the residue. For example,

%structure residues
1   ALA
2   GLY
3   ALA

The section "%sites type" is used to specify which atoms are to be considered as interaction sites in the amide bond connecting two residues. For example:

%sites type
1   2   4
1   O
2   C
3   N
4   H

The first line consists of three numbers among which the first two numbers (n and $n+1$) or ($n+1$ and n) denote two residues defined in the "%structure residues" section and the third number is the number of atoms to be listed in the next lines with one atom per line. The above example represents that the four atoms (O, C, N, H) in each amide bond connecting residues 1 and 2 are regarded as interaction sites. There is a special case that the first two integers in the first line are both zero. This case represents all the amide bonds that are not explicitly defined in this section. In the following example, we define eight different kinds of interaction sites:

%sites type
1   2   4
3   O

S7

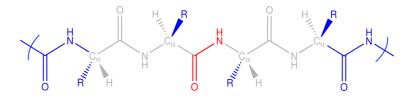|	
2	C
3	N
3	H
0 0	4
3	O
6	C
7	N
8	H

This example shows that the first five lines denote the O, C, N, H atoms of the amide bond between residues 1 and 2, and the second five lines denote the O, C, N, H atoms of all the remaining amide bonds.

In the cases of the peptides and proteins, it is necessary for the frequency map to assign the atoms (called "excluded atoms") that should be excluded in the calculations of the electrostatic potentials and/or electric fields operating on respective vibrational chromophores. Usually, these are the atoms that are sufficiently close to the interaction sites of each peptide bond. The list of excluded atoms can be given in this section using the index 'e'. Note that all the atoms constituting each vibrational chromophore (which are explicitly defined with a number index in this section) will also be excluded in the calculations of the electrostatic potentials and/or electric fields operating on it.

We use 'CA', 'CB', 'CC', … to denote the alpha carbon, beta carbon, gamma carbon, and so on. We use 'HA', 'HB', … to represent the hydrogen atoms bonded to CA, CB, …., respectively. We use 'CA-N' to denote the alpha carbon connected to the N atom of the peptide bond and 'CA-C' the alpha carbon connected to the C atom of the peptide bond. We use 'CA-N1' to denote the alpha carbon of the residue next to the 'CA-N' alpha carbon, 'CA-C1' the alpha carbon of the residue next to the 'CA-C' alpha carbon. As an example, consider the peptide A shown in Figure S3 (taken from the supplementary material of ref [2]). When the electrostatic perturbation on the peptide bond shown in red is calculated, the grey atoms, as well as the red atoms themselves, are to be excluded. If we suppose that the peptide bond shown in red connects residues 4 and 5 and only the O atom of it is the interaction site for mapping, the content of '%site type' related to this part of peptide A will be as follows:

%sites	type		
4	5	20	
1	O		# interaction site and also excluded atom
e	C		# excluded atoms denoted by the 'e' index
e	N		
e	H		
e	CA-N		
e	HA-N		
e	O-N		
e	C-N		
e	N-N		

e	H-N
e	CA-N1
e	HA-N1
e	CA-C
e	HA-C
e	O-C
e	C-C
e	N-C
e	H-C
e	CA-C1
e	HA-C1

Figure S3. An example of peptides. Atoms displayed in blue are included in the calculation of the electrostatic perturbation on the amide bond displayed in red.

To provide parameters for coupling maps between neighboring amide bonds, it is necessary to specify dihedral angles to give conformation of the neighboring amide groups. For this purpose, we use the section "%dihedral". For the most general case, a dihedral angle can be specified by a list of four atoms. For example,

1	2	3	4
5	6	7	8
9	10	11	12

These three lines define three dihedral angles and will be called dihedral angles 1, 2, and 3 when the coupling map is defined.

Alternatively, each line has only one integer followed by a letter 'N' or 'C' when the structure is defined in the '%structure residues' section with a list of residue names. Then, this number denotes the residue having the alpha-carbon to define the phi and psi angles in peptides or proteins. For example:

3	N
3	C

The letter 'N' is used to define the phi angle which is the dihedral angle of C-N-C(alpha)-C. The letter 'C' is used to define the psi angle which is the dihedral angle of N-C(alpha)-C-N. It should be emphasized that %structure residue and %sites type are specially made only for the case of atomic interaction sites of peptides and proteins. If one wants off-interaction sites

for a peptide, he should use %structure, %sites on, and %sites off sections and give full information of the atomic coordinates.

3. Mapping parameters

After structural information is presented as described above, this part of the vbm file is to provide the actual values of mapping parameters and consists of the following sections:

%map interaction
%map param
%map dihedral
%map coupling

The two sections "%map interaction" and "%map param" give the actual values of the parameters. These two sections are always given together in a pair in the order of "%map interaction" and then "%map param". The basic format of "%map interaction" consists of five lines as follows:

%map interaction
Target property
Unperturbed value of the target property
Perturbation
Unit of map parameters
Information on the data type of map parameters [Full|Reduced]

The first two lines are the name and the unperturbed value of the target property. The third line describes the type of perturbation, e.g., electrostatic potential, electric field, etc. The fourth line specifies the unit of the map parameters. The "information on the data type of map parameters" on the fifth line concerns the mathematical form of the quantity such as scalar, vector, and tensor expressed in numbers. For example, 1 for a scalar, 3 for a three-dimensional vector, "3 3" for a 3-by-3 matrix, "3 3 3" for a tensor with triple indices. One may also specify the "[Full|Reduced]" option on this line, which will be explained below.

In general, one or more types of perturbations on each interaction site may be effective in calculating the value V of the target property, so that it is represented (in the case of a single interaction site)

$$V = a_1 P_1 + a_2 P_2 + a_3 P_3 + \cdots$$

where a_i are parameters in scalar, vector, or tensor form corresponding to the perturbation P_i. The perturbation P_i can be a single quantity such as the electrostatic potential at the interaction site or can be a product of quantities such as a product of electric fields in two different directions.

When the right-hand side of Eq. (1) is a sum of N terms, then the last three lines of "%map interaction" should be repeated N times. For example, when the target property is the vibrational frequency shift $\Delta\omega$ and the perturbations are the electrostatic potential ϕ and the electric field E so that

S10

$$\Delta\omega = a_1\phi + \boldsymbol{a}_2 \cdot \boldsymbol{E}$$

where a_1 is a scalar parameter and $\boldsymbol{a}_2$ is a vector, the "%map interaction" and "%map param" sections are given as follows:

        %map interaction
        Frequency   # target property
        1707   # unperturbed value
        Electrostatic potential
        cm^-1/(atomic unit of electric potential)
        1
        Electrostatic field
        cm^-1/(atomic unit of electric field)
        3

        %map param
        0.0012   # parameter values for electrostatic potentials
        0.0003   # repeated *n* times if there are *n* interaction sites
        -0.0054
        0.0064
        -0.0014
        -0.0011
        0.40   -1.21   0.79   # parameter values for electric fields
        13.54   22.54   28.55   # repeated *n* times if there are
        50.55   28.55   22.54   # *n* interaction sites
        13.54   0.79   -1.21
        -1.89   -1.14   5.76
        14.64   24.00   34.82

When the fifth line of "%map interaction" contains two or more numbers, such as "3 3" (for a 3-by-3 matrix) or "3 3 3" (for a tensor with triple indices), it can have additionally "Full" or "Reduced" to denote whether the matrix given in "%map param" is in the "full" form or "reduced" form. For example, in the case of a second rank tensor,

        reduced – for a second rank tensor we need only six components:

$$A_{11}, A_{12}, A_{13}, A_{22}, A_{23}, A_{33}$$

        full – for a second rank tensor we need nine components:

$$A_{11}, A_{12}, A_{13}, A_{21}, A_{22}, A_{23}, A_{31}, A_{32}, A_{33}$$

It is always assumed that the rightmost index changes first when reading the data. In the "%map param" section, the number of rows and columns to represent the parameter values can be freely chosen with one rule that the rightmost index changes first. For example, the full second rank tensor can be represented as

$$A_{11} \; A_{12} \; A_{13} \; A_{21} \; A_{22} \; A_{23} \; A_{31} \; A_{32} \; A_{33}$$

or

$$A_{11} \; A_{12} \; A_{13}$$

$$A_{21} \; A_{22} \; A_{23}$$

$$A_{31} \; A_{32} \; A_{33}$$

The section "%map dihedral" describes frequency shifts due to nearest neighbor interactions of local amide oscillators in peptides and proteins. The format is such that the first three lines are as follows:

> N string
>
> unit
>
> θ_{min} θ_{max} $\theta_{interval}$ [θ_{min} θ_{max} $\theta_{interval}$]

Here N denotes an integer referring to the Nth residue defined in the "%structure reside" section. The 'string' is either 'c' or 'n' to represent the N-terminus side (-CONH-Cα-) or C-terminus side (-Ca-CONH-) amide group, respectively, where Cα represents the alpha carbon of the Nth residue. For example,

> 1   c

means that the map parameters are for the frequency shift of the amide group on the C-terminus side due to the changes in the dihedral angles of the first residue. If N is zero, the map applies to all the residues that are not explicitly specified in this section. In the second line the unit for the parameters is given. In the third line, θ_{min}, θ_{max}, $\theta_{interval}$ are the minimum and maximum values and the grid size of the phi and psi dihedral angles with respect to the alpha-carbon of the Nth residue. If these values are different between phi and psi dihedral angles, it is possible to give all six values: θ_{min}, θ_{max}, $\theta_{interval}$ for the phi dihedral angle and θ_{min}, θ_{max}, $\theta_{interval}$ for the psi dihedral angle. From the third line, a matrix of parameter values is given. For example:

> 1   c
>
> cm^-1
>
> -180   180   90
>
> 12 23 34 45 12   # values for (-180,-180), (-180,-90),(-180,0),(-180,90),(-180,180)
>
> 56 67 78 89 56   # values for (-90,-180), (-90,-90),(-90,0),(-90,90),(-90,180)
>
> 91 12 23 34 91   # values for (0,-180), (0,-90),(0,0),(0,90),(0,180)
>
> 45 56 67 78 45   # values for (90,-180), (90,-90),(90,0),(90,90),(90,180)
>
> 33 44 55 66 33   # values for (180,-180), (180,-90),(180,0),(180,90),(180,180)

The section "%map coupling" provides parameters for the coupling term between two vibration modes such as the $J_{ij}(t)$ term in ref. [3]. The format is the same as "%map dihedral" with one

exception that the 'string' is not necessary on the first line, so that the format of the first three lines are:

N

unit

θ_{min} θ_{max} $\theta_{interval}$ [θ_{min} θ_{max} $\theta_{interval}$]

Here N is an integer denoting the Nth residue defined in the "%structure reside" section. If N is zero, the map applies to all the residues that are not explicitly specified in this section. 'unit' is the unit of the parameters. θ_{min}, θ_{max}, $\theta_{interval}$ are the minimum and maximum values and the grid size of the phi and psi dihedral angles with respect to the alpha-carbon of the Nth residue. From the fourth line, a matrix of parameter values is given.

Supplementary references

(1)   Choi, J.-H.; Oh, K.-I.; Lee, H.; Lee, C.; Cho, M. Nitrile and Thiocyanate IR Probes: Quantum Chemistry Calculation Studies and Multivariate Least-Square Fitting Analysis. *J. Chem. Phys.* **2008,** *128*, 134506.
(2)   Reppert, M.; Roy, A. R.; Tokmakoff, A. Isotope-Enriched Protein Standards for Computational Amide I Spectroscopy. *J. Chem. Phys.* **2015,** *142*, 125104.
(3)   Roy, S.; Lessing, J.; Meisl, G.; Ganim, Z.; Tokmakoff, A.; Knoester, J.; Jansen, T. L. C. Solvent and Conformation Dependence of Amide I Vibrations in Peptides and Proteins Containing Proline. *J. Chem. Phys.* **2011,** *135*, 234507.