Journal of Quantitative Spectroscopy & Radiative Transfer 203 (2017) 557–564

Contents lists available at ScienceDirect

Journal of Quantitative Spectroscopy & Radiative Transfer

journal homepage: www.elsevier.com/locate/jqsrt

Cycle bases to the rescue

Roland Tóbiás, Tibor Furtenbacher, Attila G. Császár*

MTA-ELTE Complex Chemical Systems Research Group, H-1117 Budapest, Pázmány Péter sétány 1/A, Hungary

ARTICLE INFO

Article history:
Received 3 January 2017
Received in revised form
20 March 2017
Accepted 21 March 2017
Available online 22 March 2017

Keywords:
Cycle basis
Information system
ECART
Rovibronic transitions
Energy levels
MARVEL
Database

ABSTRACT

Cycle bases of graph theory are introduced for the analysis of transition data deposited in line-by-line rovibronic spectroscopic databases. The principal advantage of using cycle bases is that outlier transitions –almost always present in spectroscopic databases built from experimental data originating from many different sources– can be detected and identified straightforwardly and automatically. The data available for six water isotopologues, $H_2{}^{16}O$, $H_2{}^{17}O$, $H_2{}^{18}O$, $HD^{16}O$, $HD^{17}O$, and $HD^{18}O$, in the HITRAN2012 and GEISA2015 databases are used to demonstrate the utility of cycle-basis-based outlier-detection approaches. The spectroscopic databases appear to be sufficiently complete so that the great majority of the entries of the minimum cycle basis have the minimum possible length of four. More than 2000 transition conflicts have been identified for the isotopologue $H_2{}^{16}O$ in the HITRAN2012 database, the seven common conflict types are discussed. It is recommended to employ cycle bases, and especially a minimum cycle basis, for the analysis of transitions deposited in high-resolution spectroscopic databases.

1. Introduction

Advances in the instrumentation of high-resolution molecular spectroscopy and the analysis of the spectra [1] have led to a considerable increase in the experimental rovibronic information available for even rare isotopologues of molecules containing just a few atoms. Knowledge of line-by-line (LBL) spectroscopic data is required in a number of scientific and engineering applications. The large amount of LBL data obtained experimentally is deposited in scientific information systems [2–17], in the best cases in a critically evaluated and annotated form. Identifying and correcting assignment conflicts and experimental errors in measured LBL data of the spectroscopic databases requires sophisticated tools. These tools are not readily available, as handling an extreme amount of sometimes overlapping and sometimes conflicting high-resolution spectroscopic data –obtained independently by a large number of laboratories– has only a relatively brief history. In this paper we concentrate on the identification and handling of transition conflicts in LBL databases.

The traditional tools employed by high-resolution spectroscopists [1] to assign lines in observed spectra, as well as checking transition entries in LBL databases, include the use of approximate model Hamiltonians and combination difference (CD) relations. Spectroscopists have been accustomed to fitting a small set of spectroscopic constants, present in the model Hamiltonian chosen, to their observed data. These Hamiltonians interpolate well the experimental data obtained but their extrapolation capabilities are limited. A problem with CD relations is that usually only a few of them are available during a given measurement.

It must also be pointed out that the number of “experimentally measured” energy levels seems to grow, over time, much less rapidly than the number of measured and assigned transitions [18]. This is a significant hindrance when the full set of energy levels is required [19,20]. Due to the Rydberg–Ritz combination principle [21,22], if energy levels are known accurately, the transitions they are involved in are also known accurately; therefore, it appears that it is more advantageous to concentrate on measuring transitions that define new energy levels accurately, rather than measuring more and more unknown transitions connecting accurately known energy levels.

Conflicts in energy levels and transitions deposited in spectroscopic information systems are realized regularly by the scientific community developing and using LBL data. These conflicts are reported to the developers of the databases and are taken into consideration when the regular incremental upgrades, for example that of the HITRAN [2,7,11,17,23] or the GEISA [8,14] databases, takes place, utilizing also the latest set of experimental and quantum chemical information. When corrections are introduced, they improve the overall quality of the database but at the same time new conflicts may be introduced, some of which get rectified only at the next stage of incremental corrections. It is highly desirable to have a facility which identifies more or less routinely and automatically existing conflicts among the transitions deposited in

* Corresponding author.
E-mail address: csaszar@chem.elte.hu (A.G. Császár).

http://dx.doi.org/10.1016/j.jqsrt.2017.03.031
0022-4073/© 2017 Elsevier Ltd. All rights reserved.

a spectroscopic database, points out the source of the conflict, and suggests ways to avoid the conflict(s). As shown below, a tool capable of achieving all this becomes available if rovibronic energy levels and transitions, as well as the LBL databases, are viewed in the context of spectroscopic networks (SN) [18,24–27].

The concept of networks is used ubiquitously when studying complex phenomena related to nature, society, and communication [28]. A network-theoretical view of high-resolution spectroscopy has also been developed during the last decade [24–27,29,30]. SNs are obtained by forming a set of vertices and edges from rovibronic energy levels and transitions, respectively. As shown in a number of publications [18,31–40], SNs provide a highly useful framework to improve the LBL spectroscopic knowledge available, mix experimental and first-principles data, suggest new and useful measurements, etc. Energy levels forming the vertices of experimental SNs exhibit a heavy-tailed degree distribution and there exists a huge number of cycles of widely varying size within experimental SNs, irrespective of the molecule studied. The enormous number of cycles in experimental SNs of molecules was used previously to explain why SNs are extremely robust in the event of the random removal of vertices [24].

Now let us return to the problem that the transition wavenumbers and the underlying energy levels in a database storing primarily experimental data coming from various measurements may contradict each other. Inconsistencies among the spectroscopic data can be recognized in a number of ways. Deviations in the data set can be uncovered, for example, by analyzing the repeated measurements or taking advantage of outlier tests [41–43]. We are taking another direction, based on the observation that SNs have the somewhat unusual network characteristic that the vertices of the SNs (*i.e.*, the energy levels) can be ordered in a systematic and unique way based on their term values. An intrinsic property of SNs is thus the law of energy conservation (LEC): transitions forming a cycle must satisfy an ordered sum law. This in turn provides a nice opportunity to check the correctness of transitions that form cycles of the SN, circumventing explicit consideration of energy levels. This task is further helped by the fact that there are only a very limited number of energy levels which are not part of any cycle.

To demonstrate how to use the LEC in the analysis of observed transitions, a simple subset of a $H_2^{17}O$ line list, a 4-membered cycle, is taken from Ref. [29] (see Fig. 1). In this example, the LEC requires that the properly signed sum corresponding to the exact wavenumbers of the transitions $(000)1_{1,1} \leftarrow (000)0_{0,0}$, $(000)2_{0,2} \leftarrow (000)1_{1,1}$, $(011)1_{0,1} \leftarrow (000)2_{0,2}$, and $(011)1_{0,1} \leftarrow (000)0_{0,0}$ should be zero. Under experimental conditions [29], a “discrepancy” computed from the transitions,

$$\left|36.9312 + 33.0747 + 5273.8278 - 5343.8340\right| \text{ cm}^{-1} = 0.0003 \text{ cm}^{-1}$$

suggests that the LEC is satisfied well within a “threshold” provided by the measured uncertainties of the wavenumbers of transitions,

$$(0.0001 + 0.013 + 0.0010 + 0.0021) \text{ cm}^{-1} = 0.0162 \text{ cm}^{-1}.$$

If a certain measured transition wavenumber has lower accuracy, the above “discrepancy” may become larger than the “threshold”, indicating the incompatibility of the observed data. This example illustrates that cycles are suitable for the detection of measurement errors and inconsistencies in LBL data deposited in high-resolution spectroscopic databases. Detecting and treating cycles when there is a large number of them requires a sophisticated approach. Here we explore such an approach, the use of cycle bases (CB) of graph (network) theory. It should be mentioned that cycle bases, especially minimum cycle bases (MCB), have found application in another field of chemistry, in chemical graph theory [44].

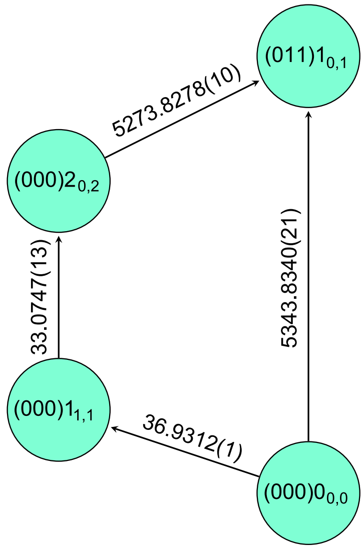

Fig. 1. Four experimentally measured transitions, represented by arrows, connecting labeled energy levels of the $H_2^{17}O$ database of Ref. [29]. The transitions obey, within experimental uncertainty, the law of energy conservation. Nodes denote the energy levels with standard labeling of rovibrational levels, $(v_1v_2v_3)\ J_{K_a,K_c}$ (see text). Numbers on the arrows represent wavenumbers in cm^{-1}, with uncertainties given in parentheses. Levels are placed along a hypothetical horizontal approximate energy axis.

2. Methodological details

Parts of the basics of network theory [28] useful to understand SNs have recently been reviewed by two of the present authors [18]. Nevertheless, to help to follow the upcoming discussion, it is worth repeating some of the definitions and extending some others with certain mathematical rigor. Many of the terms and principles characterizing CBs are found in Ref. [45].

2.1. Preliminaries

Let L and T be nonempty sets whose entries are called *vertices* and *edges*, and let $I: T \to L \times L$ be the (not necessarily injective) *incidence* function, where $L \times L = \{(\lambda_1, \lambda_2): \lambda_1, \lambda_2 \in L\}$ is the *set of ordered pairs* of L, providing the *source* and *target* vertex of the edges, respectively. Then, the quadruple $N_S = \langle L, T, I, \vartheta \rangle$ is a *vertex-weighted, directed multigraph*, where $\vartheta: L \to \mathbb{R}$ is a weight function defined on the vertices. Such multigraphs will serve as a model of *spectroscopic networks*, with the following interpretation: (a) L is the set of (energy) *levels*, (b) T is the set of *transitions*, and (c) $\vartheta(\lambda)$ is the exact (spectroscopic) *term* corresponding to $\lambda \in L$. Furthermore, $\vartheta(\lambda_1) \leq \vartheta(\lambda_2)$ for each $\tau \in T$ with $I(\tau) = (\lambda_1, \lambda_2)$, where λ_1 and λ_2 are referred to as the *lower and upper levels* of the transition τ, respectively. The transitions $\tau, \tau' \in T$ are *coincident* if $I(\tau) = I(\tau')$. The set of all the transitions coincident to $\tau \in T$ is the *coincidence class* of τ.

In certain cases it is advantageous to replace the multigraph specified above by less complicated structures. Let (a) $L{\bullet}L = \{\Lambda \subseteq L: 1 \leq |\Lambda| \leq 2\}$ be the *set of unordered pairs* of L, (b) $\mathcal{I}: T \to L{\bullet}L$ be the *undirected incidence function* of N_S defined as $\mathcal{I}(\tau) = \{\lambda_1, \lambda_2\}$ for any $\tau \in T$ with $I(\tau) = (\lambda_1, \lambda_2)$, (c) $\mathcal{I}_{\tilde{T}}: \tilde{T} \to L{\bullet}L$ be the *restriction* of the function $\mathcal{I}$ to $\tilde{T} \subseteq T$ constructed as $\mathcal{I}_{\tilde{T}}(\tau) = \mathcal{I}(\tau)$ for all $\tau \in \tilde{T}$, and (d) T^* be the *set of unique transitions*, namely, one of the largest subset of T for which $\mathcal{I}|_{T^*}$ is injective. With the notation introduced, $\mathcal{N}_S = \langle L, T, \mathcal{I}, \vartheta \rangle$ and $\mathcal{G} = \langle L, T^*, \mathcal{I}|_{T^*} \rangle$ are the *underlaying network* and the *contraction graph* (CG) of N_S, respectively.

To reflect the *Rydberg–Ritz combination principle* [21,22], the function $\varsigma: T \to \mathbb{R}$ should also be introduced,

$$\varsigma(\tau) = \vartheta(\lambda_2) - \vartheta(\lambda_1), \tag{1}$$

where $\tau \in T$ with $I(\tau) = (\lambda_1, \lambda_2)$. An essential property of the function ς is that coincident transitions have the same wavenumbers. Since addition of a real number to the term values does not change the definition of ς, we restrict ourselves to considering $\min_{\lambda \in L}\vartheta(\lambda)$=0.

Note that ς is subject to the LEC, *i.e.*, for any $\tau_1, \tau_2 \in T$ with $I(\tau_1) = (\lambda_1, \lambda_2)$ and $I(\tau_2) = (\lambda_2, \lambda_3)$, the expression

$$\varsigma(\tau_1) + \varsigma(\tau_2) = \vartheta(\lambda_2) - \vartheta(\lambda_1) + \vartheta(\lambda_3) - \vartheta(\lambda_2) = \vartheta(\lambda_3) - \vartheta(\lambda_1) \tag{2}$$

does not depend on $\vartheta(\lambda_2)$. To extend Eq. (2), consider a subset of T, denoted by $\Theta = \{\tau_i: 1 \le i \le K\}$, such that $\mathcal{I}(\tau_i) = \{\lambda_i, \lambda_{i+1}\}$ for all $1 \le i \le K$ with $\lambda_{K+1} = \lambda_1$, and let $\Phi_\Theta: \Theta \to \{-1, 1\}$ be the so-called *spectral sign function*,

$$\Phi_\Theta(\tau_i) = \begin{cases} 1, & \text{if } I(\tau_i) = (\lambda_i, \lambda_{i+1}) \\ -1, & \text{otherwise.} \end{cases} \tag{3}$$

For the transitions of Θ the LEC can be formulated as

$$\sum_{i=1}^{K} \Phi_\Theta(\tau_i)\varsigma(\tau_i) = 0. \tag{4}$$

As all measured wavenumbers suffer from measurement uncertainties, the term *spectroscopic realization* should be defined. A spectroscopic realization of N_S is the quintuple $R_S = \langle L, T, I, \sigma, \delta\rangle$, whereby $\sigma: T \to \mathbb{R}$ and $\delta: T \to \mathbb{R}^+$ are the *experimental wavenumber* and *uncertainty* functions, respectively. Using the contraction graph $\mathcal{G}$, the *contracted realization* of R_S, $R_S^* = \left\langle L, T^*, I|_{T^*}, \sigma^*, \delta^*\right\rangle$, is obtained, whereby $\sigma^*: T^* \to \mathbb{R}$ and $\delta^*: T^* \to \mathbb{R}$. Based on the spectroscopic information related to the coincidence class of a $\tau^* \in T^*$, denoted by s^*, δ^* and σ^* can be chosen as

$$\delta^*(\tau^*) = \sqrt{\frac{|s^*|}{\sum_{\tau \in s^*} \frac{1}{\delta^2(\tau)}}}, \tag{5}$$

$$\sigma^*(\tau^*) = \frac{\delta^{*2}(\tau^*)}{|s^*|} \sum_{\tau \in s^*} \frac{\sigma(\tau)}{\delta^2(\tau)}, \tag{6}$$

where $\delta^*(\tau^*)$ and $\sigma^*(\tau^*)$ correspond to an estimate of the expectation value of $\varsigma(\tau^*)$ and its uncertainty, respectively. We call $\sigma^*(\tau^*)$ and $\delta^*(\tau^*)$ the *contracted wavenumber* and the *contracted uncertainty*, respectively.

To provide a graph-theoretical interpretation of the LEC, it is necessary to recall certain (*edge-induced*) *subgraphs* of $\mathcal{G}$, *i.e.*, graphs of the form $\mathcal{G}_{\text{sub}} = \langle L_{\text{sub}}, T_{\text{sub}}, \mathcal{I}|_{T_{\text{sub}}}\rangle$ with $L_{\text{sub}} = \cup_{\tau \in T_{\text{sub}}} \mathcal{I}(\tau)$ and $T_{\text{sub}} \subseteq T^*$. $\mathcal{G}_{\text{sub}}$ is a *path* of length $\mathcal{L}$ between λ_1 and $\lambda_{\mathcal{L}+1}$ if

$$L_{\text{sub}} = \{\lambda_k \in L: 1 \le k \le \mathcal{L} + 1\}, \tag{7}$$

$$T_{\text{sub}} = \{\tau_l \in T^*: 1 \le l \le \mathcal{L}\}, \tag{8}$$

and

$$\mathcal{I}(\tau_l) = \{\lambda_l, \lambda_{l+1}\}(1 \le l \le \mathcal{L}). \tag{9}$$

The graph $\mathcal{G}$ is *connected* if there exists a path between any pair of levels, else it is *disconnected*. $\mathcal{G}_{\text{sub}}$ is a *component* of $\mathcal{G}$ if it is connected and for any $\tau \in T^*\backslash T_{\text{sub}}$ the graph

$$\mathcal{G}_{\text{sub}}[\tau] = \left\langle L_{\text{sub}} \cup \mathcal{I}(\tau), T_{\text{sub}} \cup \{\tau\}, \mathcal{I}|_{T_{\text{sub}} \cup \{\tau\}}\right\rangle \tag{10}$$

is disconnected. With the help of Eqs. (7)–(10), $\mathcal{G}_{\text{sub}}[\tau]$ is a *cycle* of length $\mathcal{L} + 1$ in $\mathcal{G}$ if $\tau \in T^*$ with $\mathcal{I}(\tau) = \{\lambda_1, \lambda_{\mathcal{L}+1}\}$. Each cycle $C = \left\langle L_C, T_C, \mathcal{I}|_{T_C}\right\rangle$ in $\mathcal{G}$ can be associated with a spectral sign function Φ_{T_C}, satisfying the LEC (see Eq. (4)):

$$\sum_{\tau \in T_C} \Phi_{T_C}(\tau)\varsigma(\tau) = 0. \tag{11}$$

The set of all equations of the form of Eq. (11) may show linear dependency in any graph $\mathcal{G}$; therefore, one needs to find a subset of the set of all the cycles in $\mathcal{G}$ whose LEC equations are maximally linearly independent. However, to form such a subset, further information is needed. The cycles $C' = \left\langle L_{C'}, T_{C'}, \mathcal{I}|_{T_{C'}}\right\rangle$ and $C'' = \left\langle L_{C''}, T_{C''}, \mathcal{I}|_{T_{C''}}\right\rangle$ are said to be *disjoint* if $T_{C'} \cap T_{C''} = \emptyset$. Clearly, if C' and C'' are non-disjoint, another cycle of $\mathcal{G}$, called the *composition* (or *sum*) of C' and C'', can be expressed as

$$C' \oplus C'' = \left\langle L_{C' \oplus C''}, T_{C'}\Delta T_{C''}, \mathcal{I}|_{T_{C'}\Delta T_{C''}}\right\rangle, \tag{12}$$

where

$$T_{C'}\Delta T_{C''} = (T_{C'}\backslash T_{C''}) \cup (T_{C''}\backslash T_{C'}) \tag{13}$$

is the *symmetric difference* of $T_{C'}$ and $T_{C''}$, *i.e.*, those parts of the two sets which are not common, and

$$L_{C' \oplus C''} = \cup_{\tau \in T_{C'}\Delta T_{C''}} \mathcal{I}(\tau). \tag{14}$$

C' and C'' are *independent* if $T_{C'}\Delta T_{C''} \ne \emptyset$.

2.2. Cycle bases

One of the largest sets of independent cycles, $\mathcal{C}$, contains exactly $\nu = |T^*| - |L| + c$ cycles [46,47], where c is the number of components in $\mathcal{G}$, and ν is the *cyclomatic number* [46,47] of $\mathcal{G}$. Furthermore, the elements of $\mathcal{C}$ span all cycles in $\mathcal{G}$, *i.e.*, there is a subset $\bar{\mathcal{C}} \subseteq \mathcal{C}$ of non-disjoint cycles for each cycle C in $\mathcal{G}$ such that the composition of the cycles in $\bar{\mathcal{C}}$ equals to C. The set $\mathcal{C}$, whose entries are the *basic cycles* (BCs) of $\mathcal{G}$, is called a *cycle basis* (CB) of $\mathcal{G}$. Since $\cup_{i=1}^{c}\mathcal{C}_i$ spans all the cycles of $\mathcal{G}$, where $\mathcal{C}_i$ is a CB of the ith component in $\mathcal{G}$, all $\mathcal{C}_i$s can be studied separately. Without loss of generality, in what follows only connected CGs are investigated.

To understand the construction of a CB in $\mathcal{G}$, additional definitions are needed. The connected graph $\mathcal{G}$ is a *tree* if it is *acyclic*, *i.e.*, it has no cycles. Using the notation introduced in the previous subsection, $\mathcal{G}_{\text{sub}}$ is a *spanning tree* (ST) in $\mathcal{G}$ if it is a tree and $L_{\text{sub}} = L$. It can be shown that $|T_{\text{sub}}| = |L| - 1$. If $\mathcal{G}_{\text{sub}}$ is a ST in $\mathcal{G}$ and $\tau \in T^*\backslash T_{\text{sub}}$ with $\mathcal{I}(\tau) = \{\lambda', \lambda''\}$, then there is a path between λ' and λ'', denoted by $P_\tau = \langle L_\tau, T_\tau, \mathcal{I}|_{T_\tau}\rangle$. Seemingly, the graph $C_\tau = P_\tau[\tau]$ is a cycle in $\mathcal{G}$, where [ ] has the same meaning as in Eq. (10).

It is apparent that $C_{\tau'}$ and $C_{\tau''}$ are independent for any τ', $\tau'' \in T^*\backslash T_{\text{sub}}$ because $\tau', \tau'' \in (T_{\tau'} \cup \{\tau'\})\Delta(T_{\tau''} \cup \{\tau''\})$. Thus, $|T^* \backslash T_{\text{sub}}| = |T^*| - |L| + 1$ cycles, called *fundamental cycles* (FCs), can be obtained, which span any cycle in $\mathcal{G}$ giving a *fundamental cycle basis* (FCB) induced by the spanning tree $\mathcal{G}_{\text{sub}}$ to $\mathcal{G}$.

The following must be noted about FCBs and their use in spectroscopy: (a) different FCs belong to different spanning trees; (b) the number of possible spanning trees for the contraction graph of a usual spectroscopic realization is enormous; and (c) the topology of a spanning tree influences the lengths of the FCs in a cycle basis.

It is useful to adopt a cycle basis $\mathcal{C}$ such that the lengths of BCs should be reduced by decreasing $\sum_{C \in \mathcal{C}} |T_C|$, called the *total cost*, with the cycles $C = \langle L_C, T_C, \mathcal{I}|_{T_C}\rangle$. Despite the fact that building a ST whose FCB has minimal total cost is an NP-incomplete problem [48], so-called *minimum cycle bases* (MCB), having minimum total cost, can be constructed in polynomial time [49]. FCBs are suitable for handling extremely large graphs because STs can be built with efficient spanning-tree-search algorithms (*e.g.*, BFS (breadth-first search) [50], DFS (depth-first search) [51], as well as Kruskal's [52] and Prim's [53] algorithms) well suited to optimize certain parameters of the ST. In this paper MCBs are utilized in the analysis of spectroscopic data, the implementation of FCBs based on Prim's algorithm is mentioned, as well.

2.3. The ECART code

When a cycle basis is available, the basic cycles can be characterized with suitable measures. To quantify how badly the experimental lines break the conservation equations of the form of Eq. (11), we introduce a few measures.

Using the notations in Eq. (11), the (*contracted circular*) *discrepancy* of cycle $C \in \mathcal{C}$ is

$$\mathcal{D}(C) = \left| \sum_{\tau\in T_C} \Phi_{T_C}(\tau)\sigma^*(\tau) \right|, \tag{15}$$

providing a suitable measure of the violation of the LEC in *C*. With correct experimental uncertainties, $\mathcal{D}(C)$ should be close to

$$\mathcal{T}(C) = \sum_{\tau\in T_C} \delta^*(\tau), \tag{16}$$

where $\mathcal{T}(C)$ is the (*contracted circular*) *threshold* of *C*. The basic cycle *C* is claimed to be *good* if the LEC holds within experimental accuracy, that is

$$\mathcal{E}(C) \le \kappa, \tag{17}$$

where $\mathcal{E}(C) = \mathcal{D}(C) - \mathcal{T}(C)$ is the (*contracted circular*) *elongation* of *C*, and κ is a *cut-off parameter*. If *C* is not good, it is labeled as *bad*.

Next, we explore how many cycles a bad transition is linked to. $\tau \in T$ is *linked* to a cycle $C \in \mathcal{C}$ if τ is coincident with a $\tau^* \in T_C$. A bad basic cycle *C* has at least one $\tau \in T$ linked to *C* such that $\delta(\tau)$ is *optimistic*, *i.e.*, it is smaller than the proper uncertainty. In what follows a transition with an optimistic uncertainty is called an *outlier*. Bad basic cycles imply the presence of outliers, hence transitions linked to at least one bad cycle are suspected to have optimistic uncertainties. Transitions not linked to bad cycles are called *harmless*, *i.e.*, they do not cause conflicts if there are no suspicious transitions in the given SN (which is not necessarily true if further lines are added to the network).

The labeled BCs help to make several practical recommendations for the outlier analysis: (a) careful inspection of suspicious transitions is strongly recommended; (b) it is worth starting the revision of the dataset with transitions linked only to bad BCs; (c) the worst cycles, those with the largest elongations, should be treated first, sorting BCs in descending order of their elongation appears to be a meaningful approach; (d) after solving some of the problems, it is recommended to recalculate the discrepancies; and (f) if there are too many suspicious transitions in a CG, it is advised to first cleanse the limited network induced by them.

A code, named ECART, based on BCs, to detect and identify outliers in a LBL spectroscopic database was written in the C++ language, where ECART stands for “Energy Conservation Analysis of Rovibronic Transitions”. Our implementation utilizes the parallel, open-source MCB code of Dutta [54], which was converted into a function and interfaced with a module based on the Hopcroft–Tarjan algorithm [55]. When the MCB is generated, the labels of the BCs are determined by means of Eq. (3) and (15)–(17). As to the cut-off parameter, κ = 0.04 cm^{-1} is selected. In the output of the ECART code basic cycles are qualified as “good” or “bad”, and transitions are listed as “harmless” or “suspicious”.

In ECART a FCB-based approach is also accessible, building a spanning tree with Prim's algorithm [53]. To eliminate the problem of disconnectedness, the *c* components of the CG are linked with $c-1$ “artificial” transitions, which do not have measured wavenumbers, so that the CG should become connected. (It can readily be seen that these artificial transitions do not change the cyclomatic number of a CG.) To obtain the paths designated by the non-ST transitions, the *lowest common ancestor* search [56] is utilized, providing a FCB for the CG.

3. Data sources

To assess the utility of building cycle bases for a LBL database, we used the datasets of six water isotopologues available in HITRAN2012 [11] and three corresponding datasets of GEISA2015 [14]. The HITRAN2012 and GEISA2015 data were downloaded from Refs. [57] and [58], respectively.

For the uncertainties of the HITRAN2012 lines 10^{1-n} cm^{-1} was used, where *n* is the “error code” obtained from the HITRAN files, except for lines with *n*=0, which were ignored during our analysis. Since the line positions of the GEISA2015 water isotopologues generally do not carry uncertainties, a fixed value of 0.01 cm^{-1} is used, suitable for identifying rough outliers among the data. (Of course, the line positions may have much higher accuracy. In particular, the most intense CRDS lines have an uncertainty much less than 10^{-5} cm^{-1}.) In addition to transitions with *n*=0, lines with no assignment and some

Table 1
Selected characteristics of the datasets of water isotopologues employed during this study, taken from the HITRAN2012 [11] and GEISA2015 [14] databases with CN=cyclomatic number of the contraction graph, FBCs=number of four-membered basic cycles in the MCB and SBCs=number of six-membered basic cycles in the MCB.

Database	Species	Lines	Sources	Components	Lines ignored	CN	FBCs	SBCs
HITRAN	$H_2{}^{16}O$	142,045	13	1	7304	117,052	117,050	2
	$H_2{}^{17}O$	27,544	5	2	10,473	14,456	14,456	–
	$H_2{}^{18}O$	39,903	4	4	6198	28,456	28,456	–
	$HD^{16}O$	13,327	5	1	467	9421	9421	–
	$HD^{17}O$	175	1	4	–	81	78	3
	$HD^{18}O$	1611	2	1	–	1252	1252	–
GEISA	$H_2{}^{16}O$	119,885	15	2	–	107,921	107,921	–
	$H_2{}^{17}O$	26,215	10	2	8338	15,049	15,049	–
	$H_2{}^{18}O$	39,613	11	4	6547	28,060	28,060	–
	$D_2{}^{16}O$	5971	2	2	–	4079	4079	–
	$D_2{}^{18}O$	162	1	34	–	14	12	2

which were declared to be of dubious quality are also neglected.

Table 1 contains selected characteristics of the HITRAN2012 [11] and GEISA2015 [14] datasets employed during this study. Due to the robustness of the SNs examined, there are very few six-membered basic cycles and there exist no cycles of length greater than 6 in the CB. Inspecting the number of components, it should also be noted that the $H_2^{16}O$ network from HITRAN2012 contains only one component instead of two (see the later discussion), while the $D_2^{18}O$ set from GEISA2015 is quite fragmented.

In what follows we will work with the standard labeling $(v_1v_2v_3)J_{K_a,K_c}$, where $(v_1v_2v_3)$ and J_{K_a,K_c} denote the standard normal mode and asymmetric-top quantum numbers, respectively [59].

4. Outlier identification using ECART

This section considers the transition conflicts detected in the datasets of Table 1 by the ECART code. As Table 2 indicates, there are several bad basic cycles in the data sets analyzed. Nevertheless, none of the high-intensity lines are problematic, an important conclusion for modelers using HITRAN2012 and GEISA2015 data. In what follows only the seven most typical conflict types are discussed (see Table 3 to 5). Upon request, the complete set of the basic cycles of the water isotopologues considered are available from the authors.

4.1. Violation of selection rules

Building a spectroscopic network of a molecule from measured transitions should mean that the well-established selection rules of high-resolution molecular spectroscopy are satisfied by the deposited transitions. Nevertheless, it happens that some of the transitions present in information systems violate some of the selection rules [34].

In the present study, the $H_2^{16}O$ and $H_2^{18}O$ sets of the HITRAN2012 database contain 29 and 8 forbidden transitions, respectively. Two of the problematic transitions, denoted by t_1 and t_2 in Table 3, are included in the basic cycle drawn in Fig. 2. In HITRAN2012, t_1 is claimed to be taken from source nu-44 (Ref. [34]), but in fact there is no such energy level with the $(511)8_{1,8}$ label in the source indicated (see Table 4). Thus, this transition should be deleted from the dataset for at least two reasons. It should be noted that there is a link, represented by T_1 in Table 5 (*vide infra*), which has the same wavenumber and it is part of the HITRAN2012 database, as well. Likewise, we could not find t_2 in the publication encoded in HITRAN2012 as nu-24. However, T_2 in Table 5 has an only slightly different wavenumber with an assignment drastically different from that of t_2. Predicting the wavenumber of the $(032)6_{0,6} \leftarrow (010)5_{0,5}$ transition with the difference between the terms of L_1 and L_2, 10,522.847487 cm^{-1} is obtained, whose deviation from the original value is comparable to the observed discrepancy of Fig. 2.

Table 2
Characteristics of basic cycles (BCs) in the water datasets of the HITRAN2012 [11] and GEISA2015 [14] information systems, with LD=largest discrepancy in cm^{-1}, LE=largest elongation in cm^{-1}, and LIST=largest intensity among the suspicious transitions (STRs) in cm·molecule^{-1}.

Database	Species	Bad BCs	STRs	LD	LE	LIST
HITRAN	$H_2^{16}O$	2 415	5 188	547.015	545.805	6.12×10^{-24}
	$H_2^{17}O$	6	21	21.907	21.905	5.95×10^{-25}
	$H_2^{18}O$	96	266	95.026	95.004	2.29×10^{-21}
	$HD^{16}O$	22	72	0.667	0.645	1.08×10^{-23}
	$HD^{17}O$	–	–	0.000	–0.040	–
	$HD^{18}O$	–	–	0.000	0.000	–
GEISA	$H_2^{16}O$	–	–	0.021	–0.059	–
	$H_2^{17}O$	1	4	2.502	2.462	5.31×10^{-25}
	$H_2^{18}O$	73	159	95.026	94.986	2.25×10^{-22}
	$D_2^{16}O$	–	–	0.062	0.022	–
	$D_2^{18}O$	–	–	0.065	0.005	–

4.2. Deviant coincidence classes

Coincident transitions may form part of information systems, they are there due to repeated measurements, occasionally with drastically different wavenumbers. Then, the need arises to eliminate those coincident transitions which are in conflict with each other, making their coincidence classes deviant.

Such problems exist for the $H_2^{16}O$ and $H_2^{18}O$ sets of the HITRAN2012 database, yielding 344 and 11 deviant coincidence classes, respectively. One example of a deviant coincidence class contains the transitions $t_3 - t_5$, given in Table 3. The discrepancy of the basic cycle containing them is similar to the difference between the wavenumbers of t_3 and t_5. Although the line position of t_4 can be reproduced from the supplementary material of Ref. [60], the computed term value of L_4 is in disagreement with L_3, based on the experimental wavenumber of T_3. As far as t_5 is concerned, similarly to t_2, it does not form part of the source nu-24. Since the $(300)8_{7,1} \leftarrow (000)7_{6,2}$ transition is part of a conflict-free BC of the $H_2^{16}O$ dataset of the GEISA2015 database, taken from an independent source with a wavenumber of 10,826.299909 cm^{-1} [61], t_3 can be accepted as a validated transition.

4.3. Inconsistent labeling of energy levels

If LBL data are collected from several sources, the consistency of the energy levels must be ensured. For example, three and one lines of the four-membered 1862nd basic cycle of the $H_2^{16}O$ dataset of the HITRAN2012 database are taken from the nu-44 and nu-24 sources, respectively. In the source nu-44, the transition $(300)2_{1,2} \leftarrow (000)3_{1,2}$ coming from nu-24 was relabeled according to T_7, providing the line T_6. Consequently, adopting t_6 and the lines consistent with T_6 at the same time leads to a conflict in the SN. As an alternative to t_6, the line T_8 of Table 5 is suggested.

4.4. Nearly identical discrepancies

Certain outliers can induce almost identical discrepancies if they are the transitions responsible for conflicts in several basic cycles. For example, let us consider t_7 of the $H_2^{17}O$ dataset of HITRAN2012. Comparing this line against T_9 leads to the conclusion that the difference between the positions of T_9 and t_7, 21.9072 cm^{-1}, fully explains the discrepancy of 21.907 cm^{-1} of the corresponding BC.

4.5. Constructive and destructive interferences

Since the deviation of the measured wavenumbers from their exact values may be either negative or positive, they can increase or decrease the discrepancies, giving rise to constructive or destructive interferences.

During the process of revising t_8, t_9, and t_{10} we observed that t_8 and t_9 interfere constructively in basic cycle 2. Between t_8 and t_{10} a destructive interference is found as their deviations cancel out in the 1681st basic cycle of the $H_2^{18}O$ GEISA2015 dataset. The t_{10} to T_{12} replacement eliminates the discrepancy of the 55th basic cycle.

4.6. Conflicts within the same source

Needless to say, not only the complete database should be consistent, but also the individual sources.

Table 3
Characteristics of certain outliers in the selected datasets with ID=line identifier, UN=uncertainty in cm^{-1}, Source code=reference code used in the databases, BC#=line number of the basic cycle including the transition under consideration (with its discrepancy in parentheses), and Match=matching the transitions with the rows of Tables 4 and Table 5.

Species	ID	Wavenumber/$cm^{\{-1\}}$	UN	Assignment	Source code	BC#	Match
$H_2{}^{16}O$ (HITRAN)	t_1	21,295.39010	10^{-1}	$(511)8_{1,8} \leftarrow (000)7_{0,7}$	nu-44	1 (545.805)	T_1
	t_2	11,069.73629	10^{-1}	$(032)6_{0,6} \leftarrow (010)5_{0,5}$	nu-24	1 (545.805)	L_1, L_2; T_2
	t_3	10,826.30108	10^{-2}	$(300)8_{7,1} \leftarrow (000)7_{6,2}$	nu-44	2 (301.648)	L_3, L_5; T_3
	t_4	10,881.31000	10^{0}	$(300)8_{7,1} \leftarrow (000)7_{6,2}$	nu-42	2 (301.648)	L_4, L_6
	t_5	11,130.97150	10^{-3}	$(300)8_{7,1} \leftarrow (000)7_{6,2}$	nu-24	2 (301.648)	T_4; T_5
	t_6	10,461.71537	10^{-3}	$(041)8_{6,3} \leftarrow (000)7_{6,2}$	nu-24	1862 (0.125)	T_8
$H_2{}^{17}O$ (HITRAN)	t_7	7,419.90980	10^{-3}	$(101)9_{3,6} \leftarrow (000)8_{3,5}$	nu-30	1 (21.907) 2 (21.907)	T_9
$H_2{}^{18}O$ (GEISA)	t_8	6,763.71810	10^{-2}	$(021)3_{1,3} \leftarrow (000)4_{1,4}$	T04	2 (5.434)	T_{10}
	t_9	6,766.43620	10^{-2}	$(120)3_{2,1} \leftarrow (000)4_{1,4}$	T04	2 (5.434)	T_{11}
	t_{10}	6,814.66380	10^{-2}	$(021)3_{1,3} \leftarrow (000)3_{1,2}$	T04	55 (2.716) 1681 (0.001)	T_{12}
	t_{11}	5.6731090	10^{-2}	$(010)15_{9,6} \leftarrow (010)14_{10,5}$	TE3	1 (95.026)	L_7; L_8

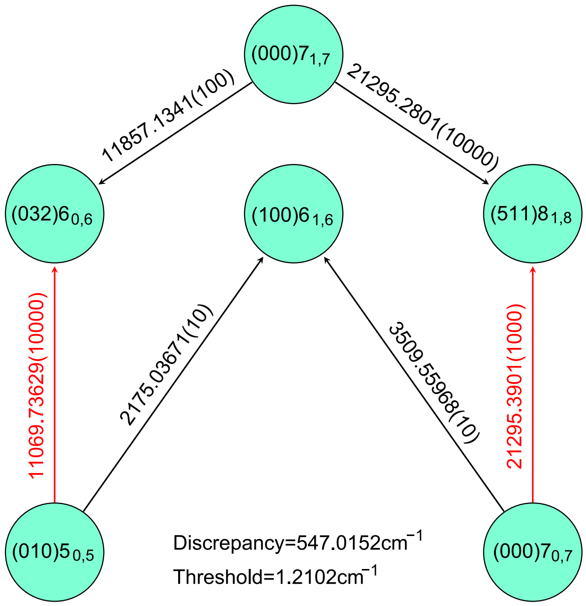

Fig. 2. The worst $H_2{}^{16}O$ basic cycle in the HITRAN2012 water database. Red transitions linking orto and para isomers are forbidden [34]. The positions of the nodes do not reflect their energy ordering.

The worst basic cycle of the $H_2{}^{18}O$ network in GEISA2015 is also the worst basic cycle in the $H_2^{18}O$ set of the HITRAN2012 database, and the transitions involved originate, according to GEISA2015, from a single source (see also Fig. 3). According to the bibliography of HITRAN2012, the lines, which are exactly the same as in GEISA2015, are taken, in fact, from nu-0 [62] and nu-48 [32]. Making a prediction for the wavenumber of t_{11}, 100.699331 cm^{-1} is obtained, correcting the worst discrepancy from 95.026 to 0.0002 cm^{-1}.

4.7. Directed basic cycles

Directed cycles have the same spectral sign for each transition linked to them. Obviously, there may be no directed basic cycle in the dataset due to the definition of spectroscopic networks. In the water datasets directed cycles could not be found, but it is important to bear in mind this possibility when performing the ECART analysis of further networks.

Table 4
Reference MARVEL ([32,34]) and ab initio [60] energy levels of two water isotopologues, with ID=line identifier, Term=term value in cm^{-1}, and UN=uncertainty in cm^{-1}.

Species	ID	Term	UN	Assignment	Source
$H_2{}^{16}O$	L_1	12,443.61328	2.32×10^{-4}	$(032)6_{0,6}$	Ref. [34]
	L_2	1,920.765793	2.92×10^{-6}	$(010)5_{0,5}$	Ref. [34]
	L_3	12,042.49077	1.20×10^{-4}	$(300)8_{7,1}$	Ref. [34]
	L_4	12,097.47106	–	$(300)8_{7,1}$	Ref. [60]
	L_5	1,216.189692	4.00×10^{-6}	$(000)7_{6,2}$	Ref. [34]
	L_6	1,216.161059	–	$(000)7_{6,2}$	Ref. [60]
$H_2{}^{18}O$	L_7	5,781.621682	6.35×10^{-4}	$(010)15_{9,6}$	Ref. [32]
	L_8	5,680.922351	4.04×10^{-4}	$(010)14_{10,5}$	Ref. [32]

5. Conclusions

Many information systems containing experimental rovibronic high-resolution spectroscopic data list transitions selected by experts on the basis that they are the most accurate ones among those available [23]. This widely employed policy in choosing the transitions to be deposited in spectroscopic information systems means that the transitions selected may not be fully compatible with each other. This incompatibility problem may be due to several factors, namely: wrong decision(s) made during the selection procedure; incorrect transfer of the correct experimental data to the database; incompatibility of the labels of the transitions used by the different groups measuring the conflicting transitions; wrong assignment of the measured line during an experimental study. These problems show up and can be identified if all transitions of a molecule are subjected to a joint validation and refinement procedure, suggested and used successfully in the MARVEL (Measured Active Rotational-Vibrational Energy Levels) procedure [25,29,30]. However, even then it has not been straightforward to identify those transitions which are seemingly incorrect or deviate substantially from other measured and assigned transitions.

Incompatibilities have been checked and identified in this work for six water isotopologues, $H_2{}^{16}O$, $H_2{}^{17}O$, $H_2{}^{18}O$, $HD^{16}O$, $HD^{17}O$, and

Table 5
Reference transitions of three species [32,34], with ID=line identifier, and Source code=numbered IUPAC tag taken from the sources.

Species	ID	Wavenumber/cm$^{\{-1\}}$	Uncertainty/cm$^{\{-1\}}$	Assignment	Source code
$H_2{}^{16}O$	T_1	21,295.39010	1.00×10^{-3}	$(610)8_{1,8} \leftarrow (000)7_{0,7}$	05ToNaZoSh.16370
	T_2	11,069.73509	2.00×10^{-3}	$(013)6_{0,6} \leftarrow (010)5_{0,5}$	02BrToDu_C.2379
	T_3	10,826.31590	1.78×10^{-2}	$(300)8_{7,1} \leftarrow (000)7_{6,2}$	05ToNaZoSh.04458
	T_4	11,130.97039	1.16×10^{-3}	$(102)8_{7,1} \leftarrow (000)7_{6,2}$	02BrToDu_C.2455
	T_5	11,130.96398	7.15×10^{-3}	$(102)8_{7,1} \leftarrow (000)7_{6,2}$	08ToTe_C.5036
	T_6	10,461.71425	3.09×10^{-3}	$(300)2_{1,2} \leftarrow (000)3_{2,1}$	02BrToDu_C.950
	T_7	10,461.71935	1.00×10^{-3}	$(300)2_{1,2} \leftarrow (000)3_{2,1}$	08ToTe_C.2376
	T_8	10,461.59240	5.00×10^{-3}	$(041)8_{6,3} \leftarrow (000)7_{6,2}$	D_08ToTe.02375
$H_2{}^{17}O$	T_9	7,441.81700	1.00×10^{-3}	$(101)9_{3,6} \leftarrow (000)8_{3,5}$	94Tothb.819
$H_2{}^{18}O$	T_{10}	6,766.43610	1.00×10^{-3}	$(021)3_{1,3} \leftarrow (000)4_{1,4}$	06LiNaSoVo.720
	T_{11}	6,763.71875	1.00×10^{-3}	$(120)3_{2,1} \leftarrow (000)4_{1,4}$	07MiWaKaCa.83
	T_{12}	6,817.38150	1.00×10^{-3}	$(021)3_{1,3} \leftarrow (000)3_{1,2}$	94Tothb.116
	T_{13}	6,814.66350	1.00×10^{-3}	$(120)3_{2,1} \leftarrow (000)3_{1,2}$	06LiNaSoVo.838

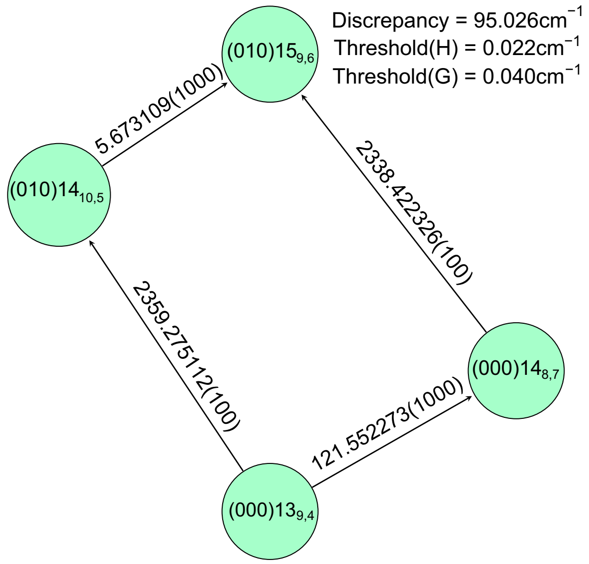

Fig. 3. The worst basic cycle of the $H_2{}^{18}O$ dataset of the GEISA2015 database, which is the worst basic cycle of the $H_2{}^{18}O$ network in HITRAN2012, as well. Threshold(H) and Threshold(G) denote the thresholds of this cycle in the HITRAN2012 and GEISA2015 databases, respectively.

$HD^{18}O$, part of two information systems, HITRAN2012 [11] and GEISA2015 [14]. Identification of incompatible transitions is done here without any reference to the underlying energy levels, utilizing only the labels of the transitions. The process, following a graph-theoretical view of high-resolution spectra and line-by-line spectroscopic databases, is based on the cycle bases of graph (network) theory and the following characteristics of spectroscopic networks: (a) experimental SNs form an enormous number of cycles, and it is straightforward to identify a fundamental cycle basis spanning all the possible cycles of arbitrary size; (b) the cyclomatic number of the largest experimental SNs studied is still relatively small; and (c) the law of energy conservation must be satisfied by all cycles containing correct transitions with correct uncertainties. For smaller experimental SNs a minimum cycle basis (MCB) can also be formed, whereby almost all cycles are as small as possible: since all experimental SNs studied are bipartite [18], all the cycles have to be of even size and the smallest ones are of length 4. For larger experimental SNs finding the MCB may be too expensive but even if the cycle basis is of cost, checking the fundamental cycles as suggested here allows the straightforward identification of problematic, outlying transitions.

While our protocol, named ECART for Energy Conservation Analysis of Rovibronic Transitions, identified no transition conflicts for $HD^{17}O$ and $HD^{18}O$, $D_2{}^{16}O$, and $D_2{}^{18}O$ in the case of HITRAN2012 [11] and for the large dataset of $H_2{}^{16}O$ in the case of GEISA2015 [14], it yielded a considerable number of conflicting transitions for other water isotopologues in these information systems. For example, for $H_2{}^{16}O$ in the case of HITRAN2012 there are 5188 conflicting transitions identified. Following an in-depth analysis, some of the most typical transition conflicts were successfully remedied.

Acknowledgements

The authors thank the NKFIH (grant number K119658) for financial support. Discussions with Professor Jonathan Tennyson, Professor János Tóth, and Professor Bill Poirier on some of the topics of this paper are gratefully acknowledged. Professor Nicole Jacquinet-Husson is thanked for providing useful pieces of information on the GEISA2015 database. The authors wish to thank Mr. Debarshi Dutta, as well, for providing technical assistance in using his parallel MCB code.

References

[1] M. Quack, F. Merkt (eds.), Handbook of high-resolution spectroscopy. Wiley; 2014.

[2] McClatchey RA, Benedict WS, Clough SA, Burch DE, Calfee RF, Fox K, Rothman LS, Garing JS, AFCRL atmospheric absorption line parameters compilation. Tech. rep., AFCRL-TR-73-0096. Air Force Cambridge Research Laboratories; 1973.

[3] Pickett HM, Poynter RL, Cohen EA, Delitsky ML, Pearson JC, Müller HSP. Submillimeter, millimeter, and microwave spectral line catalog. J Quant Spectrosc Rad Transf 1998;60:883–890.

[4] Müller HSP, Schlöder F, Stutzki J, Winnewisser G. The Cologne database for molecular spectroscopy, CDMS: A useful tool for astronomers and spectroscopists. J Mol Struct 2005;742:215–227.

[5] Császár AG, Fazliev AZ, Tennyson J, W@DIS - Prototype information system for systematization of spectral data of water. Abstracts of the twentieth colloquium on high resolution molecular spectroscopy; 2007. URL ⟨http://vesta.u-bourgogne.fr/hrms/⟩.

[6] Rothman LS, Jacquemart D, Barbe A, Benner DC, Birk M, Brown LR, et al. The HITRAN 2004 molecular spectroscopic database. J Quant Spectrosc Radiat Transf 2005;96:139–204.

[7] Rothman LS, Gordon IE, Barbe A, Benner DC, Bernath PF, Birk M, et al. The HITRAN 2008 molecular spectroscopic database. J Quant Spectrosc Radiat Transf 2009;110:533–572.

[8] Jacquinet-Husson N, Crepeau L, Armante R, Boutammine C, Chédin A, Scott NA, et al. The 2009 edition of the GEISA spectroscopic database. J Quant Spectrosc Radiat Transf 2011;112.

[9] Rothman LS, Gordon IE, Barber RJ, Dothe H, Gamache RR, Goldman A, et al. HITEMP, the high-temperature molecular spectroscopic database. J Quant Spectrosc Radiat Transf 2010;111:2139–2150.

[10] Tashkun S, Perevalov V. CDSD-4000: high-resolution, high-temperature carbon dioxide spectroscopic databank. J Quant Spectrosc Rad Transf 2011;112:1403–1410.

[11] Rothman LS, Gordon IE, Babikov Y, Barbe A, Chris Benner D, Bernath PF, et al. The HITRAN2012 molecular spectroscopic database. J Quant Spectrosc Rad Transf 2013;130:4–50.

[12] Ba YA, Wenger C, Surleau R, Boudon V, Rotger M, Daumont L, et al. MeCaSDa and ECaSDa: methane and ethene calculated spectroscopic databases for the virtual atomic and molecular data centrez. J Quant Spectrosc Rad Transf 2013;130:62–68.

[13] Babikov Y, Mikhailenko S, Barbe A, Tyuterev V. S&MPO-An information system for ozone spectroscopy on the webz. J Quant Spectrosc Rad Transf 2014;145:169–196.

[14] Jacquinet-Husson N, Armante R, Scott NA, Chedin A, Crepeau L, Boutammine C, et al. The 2015 edition of the GEISA spectroscopic database. J Mol Spectrosc 2016;327:31–72.

[15] Tennyson J, Yurchenko SN, Al-Refaie AF, Barton EJ, Chubb KL, Coles PA, et al. The ExoMol database: Molecular line lists for exoplanet and other hot atmospheres. J Mol Spectrosc 2016;327:73–94.

[16] ⟨http://www.respecth.hu/⟩; 2016.

[17] Gordon IE, Rothman LS, Hill C, Kochanov RV, Tan Y, Bernath PF, et al., The HITRAN2016 molecular spectroscopic database. J. Quant. Spectrosc. Rad. Transfer, in press.

[18] Császár AG, Furtenbacher T, Árendás P. Small molecules-big data. J Phys Chem A 2016;120:8949–8969.

[19] Furtenbacher T, Szidarovszky T, Hrubý J, Kyuberis AA, Zobov NF, Polyansky OL, et al. Definitive ideal-gas thermochemical functions of the $H_2^{16}O$ molecule. J Phys Chem Ref Data 2016;45:043104.

[20] Simkó I, Furtenbacher T, Hrubý J, Zobov NF, Polyansky OL, Tennyson J, et al. Recommended ideal-gas thermochemical functions for heavy water and its substituent isotopologues. J Phys Chem Ref Data 2017 (submitted for publication).

[21] Ritz W. On a new law of series spectra. Astrophys J 1908;28:237.

[22] Perkampus H-H. Encyclopedia of spectroscopy. VCH; 1995.

[23] Rothman LS. The evolution and impact of the HITRAN molecular spectroscopic database. J Quant Spectrosc Rad Transf, Transf 2010;111:1565–1567.

[24] Császár AG, Furtenbacher T. Spectroscopic networks. J Mol Spectrosc 2011;266:99–103.

[25] Császár AG, Czakó G, Furtenbacher T, Mátyus E. An active database approach to complete spectra of small molecules. Ann Rep Comp Chem 2007;3:155–176.

[26] Furtenbacher T, Arendás P, Mellau G, Császár AG. Simple molecules as complex systems. Sci Rep 2014;4:4654.

[27] Arendás P, Furtenbacher T, Császár AG. On spectra of spectra. J Math Chem 2016;54:806–822.

[28] Newman MEJ. Networks.Oxford: Oxford University Press; 2010.

[29] Furtenbacher T, Császár AG, Tennyson J. MARVEL: measured active rotational-vibrational energy levels. J Mol Spectrosc 2007;245:115–125.

[30] Furtenbacher T, Császár AG. MARVEL: measured active rotational-vibrational energy levels. II Algorithmic Improv, J Quant Spectrosc Radiat Transf 2012;113:929–935.

[31] Furtenbacher T, Császár AG. On employing $H_2^{16}O$, $H_2^{17}O$, $H_2^{18}O$, and $D_2^{16}O$ lines as frequency standards in the 15 - 170 cm^{-1} window. J Quant Spectrosc Radiat Transf 2008;109:1234–1251.

[32] Tennyson J, Bernath PF, Brown LR, Campargue A, Carleer MR, Császár AG, et al. Critical evaluation of the rotational-vibrational spectra of water vapor. Part I. Energy levels and transition wavenumbers for $H_2^{17}O$ and $H_2^{18}O$. J Quant Spectrosc Radiat Transf 2009;110:573–596.

[33] Tennyson J, Bernath PF, Brown LR, Campargue A, Carleer MR, Császár AG, et al. Critical evaluation of the rotational-vibrational spectra of water vapor. Part II. Energy levels and transition wavenumbers for $HD^{16}O$, $HD^{17}O$, and $HD^{18}O$. J Quant Spectrosc Radiat Transf 2010;110:2160–2184.

[34] Tennyson J, Bernath PF, Brown LR, Campargue A, Császár AG, Daumont L, et al. IUPAC critical evaluation of the rotational-vibrational spectra of water vapor. Part III: energy levels and transition wavenumbers for $H_2^{16}O$. J Quant Spectrosc Rad Transf 2013;117:29–58.

[35] Tennyson J, Bernath PF, Brown LR, Campargue A, Császár AG, Daumont L, et al. IUPAC critical evaluation of the rotational-vibrational spectra of water vapor. Part IV. Energy levels and transition wavenumbers for $D_2^{16}O$, $D_2^{17}O$ and $D_2^{18}O$. J Quant Spectrosc Radiat Transf 2014;142:93–108.

[36] Fábri C, Mátyus E, Furtenbacher T, Mihály B, Zoltáni T, Nemes L, et al. Variational quantum mechanical and active database approaches to the rotational-vibrational spectroscopy of ketene. J Chem Phys 2011;135:094307.

[37] Al Derzi AR, Furtenbacher T, Yurchenko SN, Tennyson J, Császár AG. MARVEL analysis of the measured high-resolution spectra of $^{14}NH_3$. J Quant Spectrosc Radiat Transf 2015;161:117–130.

[38] Császár AG, Furtenbacher T. Promoting and inhibiting tunneling via nuclear motions. Phys Chem Chem Phys 2016;18:1092–1104.

[39] Furtenbacher T, Szabó I, Császár AG, Bernath PF, Yurchenko SN, Tennyson J. Experimental energy levels and partition function of the $^{12}C_2$ molecule. Astrophys J Suppl 2016;224:44.

[40] McKemmish LK, Masseron T, Sheppard S, Sandeman E, Schofield Z, Furtenbacher T, et al. MARVEL analysis of the measured high-resolution rovibronic spectra of $^{48}Ti^{16}O$. Astrophys. J. Suppl. 2017;228:15.

[41] Rousseeuw PJ, Leroy AM. Robust regression and outlier detection, 589. John Wiley & Sons; 2005.

[42] Esbensen KH, Guyot D, Westad F, Houmoller LP. Multivariate data analysis-in practice: an introduction to multivariate data analysis and experimental design. Multivar Data Anal 2002.

[43] Watson JKG. The use of term-value fits in testing spectroscopic assignments. J Mol Spectrosc 1994;165:283–290.

[44] Kirby EC, Klein DJ, Mallion RB, Pollak P, Sachs H. A theorem for counting spanning trees in general chemical graphs and its particular application to toroidal fullerenes. Croat Chem Acta 2004;77:263–278.

[45] Diestel R, Král D, Seymour P. Graph theory, Oberwolfach Rep 2016;13(1):51–86.

[46] Euler L. Elementa doctrinae solidorumz. Novi Comm Acad Sci Imp Petropol 1752;4:109–140.

[47] Cauchy AL. Recherche sur les polyedres. J Éc Polytech 1813;9:68–86.

[48] Deo N, Prabhu G, Krishnamoorthy MS. Algorithms for generating fundamental cycles in a graph. ACM Trans Math Softw (TOMS) 1982;8(1):26–42.

[49] Mehlhorn K, Michail D. Implementing minimum cycle basis algorithms. J Exp Algorithmics 2007;11:2–5.

[50] Lee CY. An algorithm for path connections and its applications. IRE Trans Electron Comput 1961:346–365.

[51] Harary F. Graph theory.Reading: Addison-Wesley; 1969.

[52] Kruskal JB. On the shortest spanning sub tree of a graph and the traveling salesman problem. Proc Am Math Soc 1956;7:58.

[53] Prim RC. Shortest connection networks and some generalizations. Bell Syst Tech J 1957;36:1389–1401.

[54] URL ⟨https://github.com/deepai/Parallel-MinimumCycleBasis⟩. [Accessed 26 February. 2017].

[55] Hopcroft J, Tarjan R. Algorithm 447: efficient algorithms for graph manipulation. Comm ACM 1973;16:372–378.

[56] Harel D, Tarjan RE. Fast algorithms for finding nearest common ancestors. SIAM J Comp 1984;13:338–355.

[57] HITRANonline; 2016. URL ⟨https://http://www.hitran.org/⟩.

[58] GEISAonline; 2016. URL ⟨http://www.pole-ether.fr/etherTypo/?Id=950⟩.

[59] Kroto HW. Molecular rotation spectra.New York: Dover; 1992.

[60] Barber R, Tennyson J, Harris G, Tolchenov R. A high-accuracy computed water line list. Mon Not Roy Astron Soc 2006;368:1087–1094.

[61] Tolchenov RN, Tennyson J. Water Line Parameters from Refitted Spectra constrained by empirical upper state levels: study of the 9500-14500 cm^{-1} region. J Quant Spectrosc Rad Transf 2008;109:559–568.

[62] Rothman L, Gamache R, Goldman A, Brown L, Toth R, Pickett H, et al. The HITRAN database: 1986 edition. Appl Opt 1987;26:058–4097.