Journal of Quantitative Spectroscopy & Radiative Transfer 272 (2021) 107756

Contents lists available at ScienceDirect

Journal of Quantitative Spectroscopy & Radiative Transfer

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

autoECART: Automatic energy conservation analysis of rovibronic transitions

Roland Tóbiás[a,*], Kristóf Bérczi[b], Csaba Szabó[c], Attila G. Császár[d]

[a] Laboratory of Molecular Structure and Dynamics, Institute of Chemistry, ELTE Eötvös Loránd University and ELKH-ELTE Complex Chemical Systems Research Group, Pázmány Péter sétány 1/A, Budapest H-1117 Hungary
[b] MTA-ELTE Egerváry Research Group, Department of Operations Research, ELTE Eötvös Loránd University, Pázmány Péter sétány 1/C, Budapest H-1117, Hungary
[c] Department of Algebra and Number Theory, ELTE Eötvös Loránd University, Pázmány Péter sétány 1/C, Budapest H-1117, Hungary
[d] Laboratory of Molecular Structure and Dynamics, Institute of Chemistry, ELTE Eötvös Loránd University and MTA-ELTE Complex Chemical Systems Research Group, Pázmány Péter sétány 1/A, Budapest H-1117, Hungary

ARTICLE INFO

Article history:
Received 16 April 2021
Revised 16 May 2021
Accepted 17 May 2021
Available online 24 May 2021

Keywords:
Network theory
Constrained empirical energies
autoECART
GEISA

ABSTRACT

Despite decades of diligent work on their development, line-by-line (LBL) spectroscopic information systems, widely employed by scientists and engineers, may still contain a number of incorrect rovibronic lines. A novel heuristic protocol, relying on cycle bases of spectroscopic networks (SN) and a system of linear constraints, is proposed to unravel incorrect transitions present in the database. The algorithm is named autoECART, standing for automatic Energy Conservation Analysis of Rovibronic Transitions. The autoECART method is tested on synthetic SNs constructed from spectroscopic databases of nine water isotopologues. The systematic numerical tests demonstrate the outstanding capability of the autoECART procedure to identify almost all of the outliers generated randomly in these synthetic SNs. As a 'real-life' example, the GEISA-2019-$H_2{}^{18}O$ database is scrutinized, revealing 17 rough outlier lines. Developers of spectroscopic information systems are encouraged to utilize autoECART for cleansing the huge number of transitions deposited in LBL databases.

© 2021 The Author(s). Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

Accurate and detailed knowledge of rovibronic spectra of small molecules is crucial to numerous scientific and engineering applications. Description and understanding of combustion [1,2], radiative transfer [3,4], chemical vapor deposition [5], (exo)planetary atmospheres [6], *etc.*, require dependable line-by-line (LBL) spectroscopic information for a variety of chemical species. Extensive spectral knowledge is needed to find traces of molecules in the interstellar medium and in atmospheres of (exo)planets. Experimental lines, augmented with first-principles transitions, are often employed to deduce the complete, highly accurate energy-level structure of species [7–9], which in turn provides their partition functions [10–17] for the determination of accurate ideal-gas temperature-dependent thermochemical quantities.

These often interdisciplinary applications acted as drivers of several significant advancements in the field of high-resolution molecular spectroscopy [18]. As a result, millions of observed and billions of first-principles transitions have been determined, some of which have been deposited in spectroscopic information systems, such as HITRAN [19], GEISA [20], and CDMS [21]. Administration, annotation, and maintenance of these popular spectroscopic databases call for sophisticated theoretical and computational procedures during the construction, accumulation, confirmation, treatment, visualization, and distribution of LBL data. Efficient algorithms for these tasks are not readily available, partly because active management of an excessive number of occasionally conflicting rovibronic lines and states deduced by various laboratories does not have a long history [22].

Incompatibilities (*e.g.*, misassignments, calibration errors, misprints, or underestimated uncertainties) in rovibronic LBL datasets

[*] Corresponding author.
E-mail addresses: roland.tobias@ttk.elte.hu (R. Tóbiás), attila.csaszar@ttk.elte.hu (A.G. Császár).

https://doi.org/10.1016/j.jqsrt.2021.107756
0022-4073/© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

are repeatedly revealed both by the developers and the users of spectroscopic information systems. Despite the fact that flawed entries are corrected during the periodic updates of spectroscopic databases, contradictory transitions might appear upon the inclusion of new data sources. Consequently, it is a highly important question whether a universal and almost automatic protocol can be devised that is suitable for identifying and treating the conflicts among the processed lines right after the extension of the LBL catalogues with new data.

Needless to say, there are traditional methods for the detection of incompatible transitions, including effective Hamiltonian fitting, combination difference analysis, and different multivariate outlier tests [23]. These approaches are generally too cumbersome and ineffective in localizing all the erroneous lines among hundreds of thousands of LBL entries. As advocated here, the concept of spectroscopic networks (SN, see Ref. [24]) offers an elegant way to resolve the outlier problem *via* the exploitation of all the interdependencies among the elementary data types (wavenumbers, uncertainties, and rovibronic assignments) of the LBL records. In this study, internal conflicts are sought exclusively for these elementary parameters.

The mathematical theory of SNs is well founded and by now amply discussed in the literature [22,24–28]. As shown repeatedly, viewing rovibronic transitions and energy levels as edges and nodes, respectively, of a huge, directed, loop-free multigraph may give a better insight into the characteristics of and interconnections among rovibronic states.

The principle of SNs and the active database [29] approach provided the basis for the development of the MARVEL (Measured Active Rotational-Vibrational Energy Levels) procedure [30–33]. MARVEL has been deployed for the critical evaluation of experimental high-resolution spectra of 24 small molecules [14,31,34–51].

A highly useful property of SNs is that they possess a very large number of cycles, all of which should satisfy the law of energy conservation (LEC, see Ref. [28]). LEC prescribes that the discrepancy (absolute signed sum of the transition wavenumbers) must not be greater than the threshold (sum of the wavenumber uncertainties) for any cycle. Undoubtedly, violation of LEC within a cycle indicates that at least one of the underlying transitions is incorrect. Thus, LEC gives an excellent opportunity to evaluate the compatibility of spectral lines participating in cycles, without an explicit reference to the energy values of the rovibronic states.

In Ref. [28], a cycle-basis-based strategy, called Energy Conservation Analysis of Rovibronic Transitions (ECART), was proposed to facilitate the exploration of incorrect LBL entries in a semi-automated way. The principal objective of the present paper is to formulate a significantly improved and fully automatic version of the ECART algorithm, named autoECART, which is able to unveil all of the contaminating lines causing serious conflicts in spectroscopic databases. As shown below for synthetic SNs and for the GEISA-2019-$H_2{}^{18}O$ [52] database, autoECART works remarkably well and enables the complete utilization of the available spectral information that can be extracted from the rovibronic transitions.

2. Theoretical background

2.1. Spectroscopic networks (SN)

SNs are weighted, directed, loopless multigraphs, where (a) the *nodes* denote energy levels (states), (b) the *edges* are rovibronic transitions (lines), oriented from their lower-energy states to the upper ones, and (c) (task-dependent) *edge weights* can be assigned to the lines. Since LBL datasets are often composed of experimental, empirical, and theoretical transitions in a heavily mixed form, the term ‘spectroscopic network’ is used here in its fully general meaning.

To extract the full experimental and theoretical information accessible, one should investigate certain network elements (for instance, components, paths, cycles, and bridges) of the SN. Unlike in traditional network theory, these elements are defined here without considering the edge directions.

A *component* is a maximal connected collection of energy levels. A *path* stands for a series of linked, unrepeated transitions with distinct states. A *cycle* consists of a path and a line connecting the first and last energy levels of this path. As multiple edges are also permitted, cycles of length two may also occur in SNs. If a transition is not part of any cycle, then it is named a *bridge*.

A specific component of a SN is a *principal component* (PC) if it holds the lowest-energy state of a molecular nuclear-spin isomer; otherwise, it is called a *floating component* (FC). It should be emphasized that SNs often contain two or more PCs and several FCs. For every component, there is a (not necessarily unique) minimal subset of transitions, a *spanning tree* (ST), making the energy levels connected within that component. A set comprising STs for each component forms a *spanning forest* (SF) of the SN. All the lines outside a SF specify *basic cycles* with some transitions of this SF, whose collection is a *cycle basis* (CB). Each cycle can be written as a symmetric difference of the basic cycles [53].

SFs can be restricted such that various conditions hold for the edge weights. SFs can be obtained, *e.g.*, by the depth-first search (DFS) and breadth-first search (BFS) techniques, as well as by the Kruskal and Dijkstra methods [53]. Hereafter, BFS will be applied, with unit weights, to build a SF of the SN.

Subnetworks are important graph structures for network-based analyses. A subnetwork $\mathcal{N}$ is (a) built from certain transitions and *all* the energy levels of the SN and (b) represented by a *participation matrix*, $\mathbf{P} = \mathrm{diag}(P_1, P_2, \ldots, P_{N_\mathrm{T}})$, where N_T is the number of transitions in the SN, and P_i denotes the *participation coefficient* ($P_i = 1$ if the ith line of the SN is ‘inserted’ into $\mathcal{N}$, otherwise $P_i = 0$). Since $\mathcal{N}$ embraces all the energy levels and only a few transitions of the SN, some states may be *isolated points* (energy levels without incident lines) within $\mathcal{N}$. Additionally, as each subnetwork of the SN contains the same nodes, the set operations (union, intersection, subtraction, symmetric difference, and so forth) and relations (*e.g.*, membership and inclusion) specified over the transition sets of subnetworks can be transferred to the subnetworks themselves. For example, $\mathcal{N}_1 \cup \mathcal{N}_2$ is also a subnetwork, comprising all the lines of the subnetworks $\mathcal{N}_1$ and $\mathcal{N}_2$. Moreover, $\mathcal{N}_1 \subseteq \mathcal{N}_2$ means that $\mathcal{N}_2$ covers all the transitions of $\mathcal{N}_1$.

For the purposes of the present study, two special subnetworks should be defined. The *leading subnetwork*, $\mathcal{N}_\mathrm{LS}$, is a subnetwork of transitions not rejected by the user nor discarded algorithmically during the outlier analysis of the SN. The *atomic subnetwork*, $\diamondsuit$, is a subnetwork without lines.

2.2. Empirical energy values

Relying on the *Ritz principle* [54], the exact transition wavenumbers (ς_i; $1 \le i \le N_\mathrm{T}$) are given as

$$\varsigma_i = E_{\mathrm{up}(i)} - E_{\mathrm{low}(i)}, \tag{1}$$

where $\mathrm{up}(i)$ and $\mathrm{low}(i)$ are the indices of the upper and lower energy levels of the ith line, respectively, E_j is the exact energy of the jth state ($1 \le j \le N_\mathrm{L}$), while N_T and N_L are the number of transitions and energy levels in the SN, respectively. Since Eq. (1) is invariant under a shift of $E_{\mathrm{up}(i)}$ and $E_{\mathrm{low}(i)}$ with the same constant for all the components of the SN, it is standard practice to replace E_j with

$$e_j = E_j - E_{\mathrm{core}(\mathrm{comp}(j))}, \tag{2}$$

where e_j and $\mathrm{comp}(j)$ are the *relative energy* and the *component index* of the jth state, respectively, and $\mathrm{core}(k)$ is the index related to the lowest-energy level (*core*) of the kth component within $\mathcal{N}$. Trivially,

$$e_{\mathrm{core}(k)} = 0 \tag{3}$$

is valid for all $1 \le k \le N_\mathrm{c}$, where N_c is the number of components in $\mathcal{N}$. Thus, ς_i can be rewritten as

$$\varsigma_i = e_{\mathrm{up}(i)} - e_{\mathrm{low}(i)}. \tag{4}$$

To simplify the characterization of SNs, we decided to slightly abuse our notation and language by replacing 'e_j' with 'E_j' and using the term 'energy' rather than 'relative energy'. Since in SNs the energies of the rovibronic states are always related to the core energies of the components in a subnetwork $\mathcal{N}$, this practice should not cause any distortion in conveying the scientific message.

Instead of ς_i, only its measured or computed estimate, σ_i, can be derived, which is augmented, in an ideal case, with its δ_i uncertainty. Assuming uncorrelated errors with zero expected values for the σ_is, the energies can be obtained by minimizing the following objective function:

$$\boldsymbol{\Omega}(\boldsymbol{\epsilon}) = (\boldsymbol{\sigma} - \mathbf{R}\boldsymbol{\epsilon})^\mathrm{T}\mathbf{PW}(\boldsymbol{\sigma} - \mathbf{R}\boldsymbol{\epsilon}), \tag{5}$$

where (a) $\boldsymbol{\epsilon} = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_{N_\mathrm{L}}\}^\mathrm{T}$ symbolizes the variable vector, (b) $\boldsymbol{\sigma} = \left\{\sigma_1, \sigma_2, \ldots, \sigma_{N_T}\right\}^\mathrm{T}$ is the wavenumber vector, (c) $\mathbf{P}$ symbolizes the participation matrix of subnetwork $\mathcal{N}$, (d) $\mathbf{W} = \mathrm{diag}(w_1, w_2, \ldots, w_{N_\mathrm{T}})$ is a diagonal (*statistical*) *weight matrix* with elements $w_i = \delta_i^{-2}$, and (e) $\mathbf{R} = \{r_{ij}\}$ is the *Ritz matrix* [27]. The parameter r_{ij} is $+1/-1$ if the jth energy level is the upper/lower state of the ith line, otherwise $r_{ij} = 0$. Taking Eq. (3) into consideration, a minimum of $\boldsymbol{\Omega}(\boldsymbol{\epsilon})$, denoted with $\bar{\mathbf{E}} = \{\bar{E}_1, \bar{E}_2, \ldots, \bar{E}_{N_\mathrm{L}}\}^\mathrm{T}$, can be obtained by solving an overdetermined system of inhomogeneous linear equations:

$$\bar{E}_{\mathrm{core}(1)} = \bar{E}_{\mathrm{core}(2)} = \ldots = \bar{E}_{\mathrm{core}(N_\mathrm{c})} = 0, \tag{6}$$

$$\mathbf{G}\bar{\mathbf{E}} = \mathbf{F}, \tag{7}$$

where $\mathbf{G} = \mathbf{R}^\mathrm{T}\mathbf{PWR}$ is the *weighted Gram–Schmidt matrix* of $\mathbf{R}$ and $\mathbf{F} = \mathbf{R}^\mathrm{T}\mathbf{PW}\boldsymbol{\sigma}$ is the *vector of free terms*. Note that Eqs. (6) and (7), where the $\bar{E}_j$ entries are known as *empirical energies*, are instrumental in the *conventional MARVEL procedure* [30,32]. Employing the $\bar{E}_j$ values, an *empirical wavenumber*, $\bar{\sigma}_i = \bar{E}_{\mathrm{up}(i)} - \bar{E}_{\mathrm{low}(i)}$, can be assigned to the ith line.

2.3. Constrained empirical energies

Unfortunately, unique wavenumber uncertainties are not published in most spectroscopic data sources, only typical values for *segments* (sets of lines of the same data source with similar uncertainties) are provided. For this reason, where unavoidable, these average-case uncertainties, quoted as *estimated segment uncertainties* (ESU, see Ref. [33]), are adopted as approximations for the real δ_i values in $\mathcal{N}$.

It was shown in Ref. [33] that this approximation may lead to distortions in the empirical energies. To mitigate these harmful distortions, a restrictive procedure (*constrained MARVEL algorithm*) was proposed in Ref. [33] for experimental SNs. Within this scheme, (a) a *kernel subnetwork* ($\mathcal{N}_\mathrm{k}$) is formed from the accurate (trustworthy) lines of $\mathcal{N}$, (b) empirical wavenumbers are determined for these reliable transitions by evaluating Eqs. (6) and (7) with $\mathcal{N} = \mathcal{N}_\mathrm{k}$, and (c) the empirical energies are calculated for $\mathcal{N}$ under the restriction that the empirical wavenumbers derived in (b) are unchanged. This method facilitates the treatment of arbitrarily long, embedded $\mathcal{N}^{(1)} \subseteq \mathcal{N}^{(2)} \subseteq \ldots \subseteq \mathcal{N}^{(q)}$ sequences by gradually setting $\mathcal{N}_\mathrm{k} = \mathcal{N}^{(p-1)}$ and $\mathcal{N} = \mathcal{N}^{(p)}$ $(p = 2, 3, \ldots, q)$. In the rest of this section, a brief description is given for the calculation of constrained empirical energies.

Designate the core indices of subnetwork $\mathcal{N}_\mathrm{k}$ with $\mathrm{core_k}(1)$, $\mathrm{core_k}(2)$, ..., $\mathrm{core_k}(N_\mathrm{k,c})$, where $N_\mathrm{k,c} \ge N_\mathrm{c}$ means the number of components in $\mathcal{N}_\mathrm{k}$. Considering that the cores of $\mathcal{N}$ are also the cores of $\mathcal{N}_\mathrm{k}$, one can rearrange the components of $\mathcal{N}_\mathrm{k}$ such that $\mathrm{core_k}(l) = \mathrm{core}(l)$ is met for all $1 \le l \le N_\mathrm{c}$. Based on this rearrangement, the following linear constraints can be imposed upon the $\bar{E}_j$ values for all $1 \le j \le N_\mathrm{L}$:

$$\bar{E}_j - \bar{E}_{\mathrm{core_k}(\mathrm{comp_k}(j))} = \beta_j, \tag{8}$$

where the jth state is associated with (a) its component index inside $\mathcal{N}_\mathrm{k}$, $\mathrm{comp_k}(j)$, and (b) the β_j *bound*, which is the jth entry of the $\boldsymbol{\beta}$ vector representing the solution of Eqs. (6) and (7) under $\mathcal{N} = \mathcal{N}_\mathrm{k}$. With these constraints, only the core energies of $\mathcal{N}_\mathrm{k}$ are linearly independent variables, while all the other energies can be expressed from Eq. (8). These constraints can also be written as

$$\bar{\mathbf{E}} = \mathbf{C}\bar{\mathbf{H}} + \boldsymbol{\beta}, \tag{9}$$

where $\bar{\mathbf{H}} = \{\bar{H}_1, \bar{H}_2, \ldots, \bar{H}_{N_\mathrm{k,c}}\}^\mathrm{T}$ including $\bar{H}_m = \bar{E}_{\mathrm{core_k}(m)}$, and $\mathbf{C} = \{c_{jm}\}$ together with

$$c_{jm} = \begin{cases} 1, & \text{if } m = \mathrm{comp_k}(j), \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Substituting Eq. (9) into Eqs. (6) and (7) and performing some algebraic manipulations, the following system of linear equations is obtained:

$$\bar{H}_1 = \bar{H}_2 = \ldots = \bar{H}_{N_\mathrm{c}} = 0, \tag{11}$$

$$\mathbf{G_C}\bar{\mathbf{H}} = \mathbf{F_C}, \tag{12}$$

where

$$\begin{aligned} \mathbf{G_C} &= \mathbf{C}^\mathrm{T}\mathbf{GC}, \\ \mathbf{F_C} &= \mathbf{C}^\mathrm{T}\left(\mathbf{F} - \mathbf{G}\boldsymbol{\beta}\right). \end{aligned} \tag{13}$$

After solving Eqs. (11) and (12) for $\bar{\mathbf{H}}$, the *constrained empirical energy values* can be deduced from Eq. (9).

If $\mathcal{N}_\mathrm{k} = \lozenge$, each state forms a distinct component within $\mathcal{N}_\mathrm{k}$, implying that none of the energy values are constrained by Eq. (9). In this case, $\mathbf{C} = \mathbf{I}_{N_\mathrm{L} \times N_\mathrm{L}}$ and $\boldsymbol{\beta} = \mathbf{0}_{N_\mathrm{L}}$, where $\mathbf{I}_{N_\mathrm{L} \times N_\mathrm{L}}$ is the N_L $\times N_\mathrm{L}$ identity matrix, and $\mathbf{0}_{N_\mathrm{L}}$ is the $N_\mathrm{L} \times 1$ zero vector, reducing Eqs. (11) and (12) to Eqs. (6) and (7), respectively.

2.4. Consistency types

The consistency of SNs is a crucial notion for the network-theoretical analysis of *outliers* (transitions with incorrect elementary parameters). A subnetwork $\mathcal{N}$ of the SN is called *consistent* if there is an $\boldsymbol{\eta} = \{\eta_1, \eta_2, \ldots, \eta_{N_\mathrm{L}}\}^\mathrm{T}$, *potential vector* obeying

$$\left|\sigma_i - \eta_{\mathrm{up}(i)} - \eta_{\mathrm{low}(i)}\right| \le \delta_i \tag{14}$$

for every $1 \le i \le N_\mathrm{T}$ (with $P_i = 1$). It is proved in Appendix A that $\mathcal{N}$ is consistent if and only if all the cycles are *regular* within subnetwork $\mathcal{N}$. A regular cycle, comprising transitions with indices $\mathcal{I}_1, \mathcal{I}_2, \ldots,$ $\mathcal{I}_\mathcal{L}$ and signs $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_\mathcal{L}$, should be subject to

$$\mathcal{D} \le \mathcal{T}, \tag{15}$$

where

$$\mathcal{D} = \left|\sum_{k=1}^{\mathcal{L}} \mathcal{S}_k \sigma_{\mathcal{I}_k}\right| \text{ and } \mathcal{T} = \sum_{k=1}^{\mathcal{L}} \delta_{\mathcal{I}_k} \tag{16}$$

are the *discrepancy* and the *threshold* of the investigated cycle, respectively, and $\mathcal{L}$ is the *length* of this cycle. Here the signs $\mathcal{S}_k$ need to satisfy

$$\sum_{k=1}^{\mathcal{L}} \mathcal{S}_k \mathrm{row}_{\mathcal{I}_k}(\mathbf{R}) = \mathbf{0}_{N_\mathrm{L}}, \tag{17}$$

where $\mathrm{row}_q(\mathbf{R})$ is the transpose of the qth row in $\mathbf{R}$. This equation carries the most transparent algebraic formulation of the *law of energy conservation* (LEC, see Ref. [28]) for a cycle.

A fundamental feature of consistency is that it is *balanced:* if $\mathcal{N}$ is consistent, then any $\mathcal{N}^* \subseteq \mathcal{N}$ is also consistent [due to the exclusive presence of cycles adhering to Eq. (15)]. This balanced characteristics connotes that the same $\boldsymbol{\eta}$ potential vector can be applied in Eq. (14) to verify the consistency of $\mathcal{N}^*$ and $\mathcal{N}$.

By restricting Eq. (14), two special forms of consistency can be defined. A consistent $\mathcal{N}$ subnetwork is *strongly consistent* if the entries of $\boldsymbol{\eta}$ are empirical energy values [*i.e.*, $\boldsymbol{\eta}$ satisfies Eqs. (6) and (7)]. If $\mathcal{N}$ is a consistent subnetwork, and the entries of $\boldsymbol{\eta}$ are constrained empirical energies [namely, $\boldsymbol{\eta}$ follows the structure of Eq. (9)], $\mathcal{N}$ is *conditionally consistent*. In these cases, Eq. (14) takes the form of

$$d_i \le 0, \tag{18}$$

where $d_i = |\Delta_i| - \delta_i$ and $\Delta_i = \sigma_i - \overline{\sigma_i}$ are the *defect* and the *residual* of the ith transition, respectively. In Appendix B, five potential misconceptions about the three versions of consistency are discussed.

Due to the use of approximate δ_i values, the consistency of SNs might be compromised. To suppress these mild deformity effects, Eq. (14) should be relaxed to

$$\left|\sigma_i - \eta_{\mathrm{up}(i)} - \eta_{\mathrm{low}(i)}\right| \le \delta_{\mathrm{r},i}, \tag{19}$$

where $\delta_{\mathrm{r},i} \ge \delta_i$ is called the *relaxed uncertainty*. To keep the ancillary input data to a minimum, it is advisable to set the same $\delta_{\mathrm{r},i}$ value for the transitions of each segment within a given database. If each transition of $\mathcal{N}$ follows Eq. (19), then $\mathcal{N}$ is termed *quasi-consistent*. Evidently, quasi-consistency of subnetwork $\mathcal{N}$ holds exclusively in the case that each cycle of $\mathcal{N}$ complies with

$$\mathcal{D} \le \mathcal{T}_{\mathrm{r}}, \tag{20}$$

where the $\mathcal{T}$ threshold of Eq. (16) is substituted with its relaxed form,

$$\mathcal{T}_{\mathrm{r}} = \sum_{k=1}^{\mathcal{L}} \delta_{\mathrm{r},\mathcal{I}_k}. \tag{21}$$

A cycle is labelled as *bad* if it contradicts Eq. (20); otherwise, it is labelled as *good*.

Similarly to ‘simple’ consistency, quasi-consistency can also be restrained in two distinct ways. If $\mathcal{N}$ is quasi-consistent, and $\boldsymbol{\eta}$ is comprised of empirical energies matching Eqs. (6) and (7), then $\mathcal{N}$ is *strongly quasi-consistent*. Analogously, a quasi-consistent subnetwork $\mathcal{N}$ is *conditionally quasi-consistent* if $\boldsymbol{\eta}$ is a vector of constrained empirical energy values [derived from Eqs. (11) to (12)]. In these cases, Eq. (19) can be written in the form

$$d_{\mathrm{r},i} \le 0, \tag{22}$$

where $d_{\mathrm{r},i} = |\Delta_i| - \delta_{\mathrm{r},i}$ designates the *relaxed defect* of the ith transition. It should be stressed that the replacement $\delta_i \to \delta_{\mathrm{r},i}$ is not applied during the evaluation of Eqs. (6) and (7) or (11) and (12). Since Eqs. (14) and (19) differ only in their right-hand side, the three quasi-consistency types exhibit the same properties as the three variants of simple consistency.

2.5. Difficulties of outlier detection in SNs

Regrettably, the identification of outliers in SNs is far from being trivial. The most severe problem is that, due to the ill-defined characteristics of outliers from a mathematical viewpoint, no exact condition can be provided for their presence, which is further aggravated by the diversity of the underlying SNs.

There is no doubt that if quasi-consistency is infringed in a SN, then this SN should contain at least one outlier. Nevertheless, as attested in Appendix B (see misconceptions M4 and M5), quasi-consistency is not sufficient for the absence of flawed transitions. Thus, so-called *latent outliers*, that is incorrect lines not producing positive relaxed defects, cannot be eliminated merely *via* network-theoretical tools. For this reason, following the detection of non-latent outliers, the transitions of the SN must be subjected to external justification, like selection-rule analysis or EH modeling, in order to catch latent outliers.

Thanks to the robustness of large SNs [22], they include a huge number of short (2- and 4-membered) cycles and only few bridges. This robustness induces a notable decrease in the number of latent outliers, increasing the utility of network-based analyses.

To ensure the automatic detection of non-latent outliers, one should look for an effective heuristic algorithm, which is able to decontaminate the SN by assembling a *blacklist* of *hypothetical outliers* (BHO) from the faulty lines. In this BHO list, (a) most (if not all) of the non-latent outliers of the SN should appear, and (b) the number of *pseudo-outliers*, ascribed to deficiencies of the applied procedure, is as small as possible. Within this approximate protocol, quasi-consistency needs to be confirmed by checking conditional quasi-consistency, which provides an easy-to-assess (but, indeed, a slightly strict) criterion for the fulfillment of Eq. (19) for all $1 \le i \le N_{\mathrm{T}}$.

3. The autoECART algorithm

Following the introduction of the ECART (Energy Conservation Analysis of Rovibronic Transitions) procedure [28], efforts have been made to train this approach for the automatic generation of optimal BHOs. Our first attempt to determine a trustworthy BHO was based on the reduction of the outlier problem to a mixed integer linear programming (MILP) model (see Appendix C), assuming that the SN can be made free of outliers by neglecting a handful of lines with a minimum total weight. Nevertheless, during the numerical tests, the resulting system of equations turned out to be overly ill-conditioned and numerically too ineffective for SNs containing in excess of 200 000 transitions.

To produce a suitable BHO, we also tried to evaluate the number of good and bad basic cycles, n_{good} and n_{bad}, respectively, for each transition and perform a simple statistical analysis of the $n_{\mathrm{bad}}/(n_{\mathrm{bad}} + n_{\mathrm{good}})$ ratios. Although this method indicated properly several outliers in our test SNs, it was not able to minimize the number of pseudo-outliers in the BHO compiled.

Finally, we devised a successful algorithm, built upon the joint utilization of BFS-type CBs and constrained empirical energies. In the remainder of this section, our autoECART protocol, capable of yielding a reliable BHO, is described in detail.

As clear from the modular structure of the autoECART method (see Fig. 1), its scheme is founded on a purity check and the double execution of the *supervised outlier expulsion* (*SOE*). The purity check is employed to test the conditional quasi-consistency of $\mathcal{N}_{\mathrm{LS}}$ with respect to $\mathcal{N}_{\mathrm{anc}}$, where $\mathcal{N}_{\mathrm{anc}}$ represents the *anchor subnetwork* including those lines of $\mathcal{N}_{\mathrm{LS}}$ whose blacklisting is not allowed. If this type of consistency is satisfied, then the autoECART procedure is completed.

SOE (a) processes a *rigid* ($\mathcal{N}_{\mathrm{rig}}$) and an *elastic* ($\mathcal{N}_{\mathrm{ela}}$) *subnetwork* of the SN, (b) creates an $\mathcal{N}_{\mathrm{mer}} = \mathcal{N}_{\mathrm{rig}} \cup \mathcal{N}_{\mathrm{ela}}$ *merged subnetwork*, and (c) makes subnetwork $\mathcal{N}_{\mathrm{mer}}$ conditionally quasi-consistent, subject to $\mathcal{N}_{\mathrm{rig}}$, by assembling a BHO from the lines of $\mathcal{N}_{\mathrm{ela}}$, while keeping all the transitions of $\mathcal{N}_{\mathrm{rig}}$. For details on the stages of SOE, see the next two subsections.

The autoECART method is executed over $\mathcal{N}_{\mathrm{LS}} \setminus \mathcal{N}_{\mathrm{anc}}$. The transitions of the $\mathcal{N}_{\mathrm{anc}}$ subnetwork are specified by the user within the input (if no such lines are provided, then $\mathcal{N}_{\mathrm{anc}} = \Diamond$ will be set). During the first call of SOE (with $\mathcal{N}_{\mathrm{ela}} = \mathcal{N}_{\mathrm{anc}}$ and $\mathcal{N}_{\mathrm{rig}} = \Diamond$), a BHO is extracted from $\mathcal{N}_{\mathrm{anc}}$ to warrant that $\mathcal{N}_{\mathrm{anc}}$ is condition-

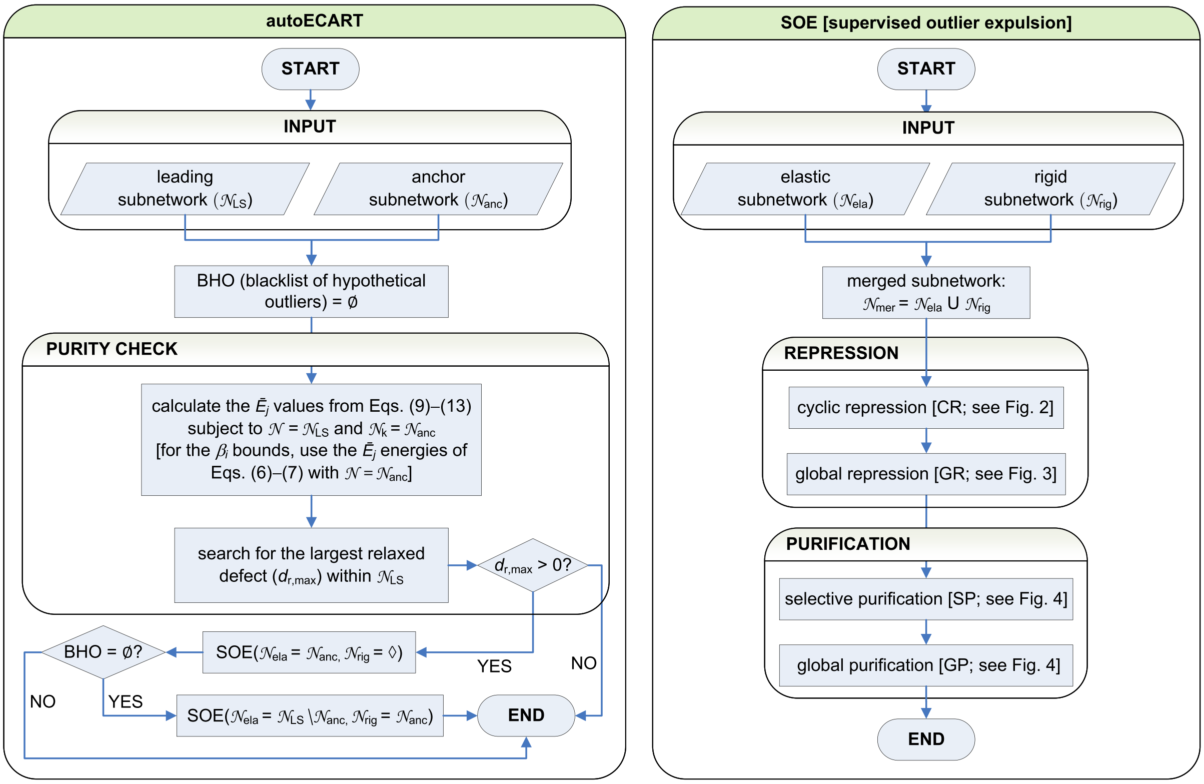

autoECART
START
INPUT
leading subnetwork ($\mathcal{N}_{\rm LS}$)
anchor subnetwork ($\mathcal{N}_{\rm anc}$)
BHO (blacklist of hypothetical outliers) = ∅
PURITY CHECK
calculate the $\bar{E}_j$ values from Eqs. (9)–(13) subject to $\mathcal{N} = \mathcal{N}_{\rm LS}$ and $\mathcal{N}_{\rm k} = \mathcal{N}_{\rm anc}$ [for the β_i bounds, use the $\bar{E}_j$ energies of Eqs. (6)–(7) with $\mathcal{N} = \mathcal{N}_{\rm anc}$]
search for the largest relaxed defect ($d_{\rm r,max}$) within $\mathcal{N}_{\rm LS}$
$d_{\rm r,max} > 0$?
YES
NO
SOE($\mathcal{N}_{\rm ela} = \mathcal{N}_{\rm anc}$, $\mathcal{N}_{\rm rig} = \Diamond$)
BHO = ∅?
NO
YES
SOE($\mathcal{N}_{\rm ela} = \mathcal{N}_{\rm LS} \setminus \mathcal{N}_{\rm anc}$, $\mathcal{N}_{\rm rig} = \mathcal{N}_{\rm anc}$)
END
SOE [supervised outlier expulsion]
START
INPUT
elastic subnetwork ($\mathcal{N}_{\rm ela}$)
rigid subnetwork ($\mathcal{N}_{\rm rig}$)
merged subnetwork: $\mathcal{N}_{\rm mer} = \mathcal{N}_{\rm ela} \cup \mathcal{N}_{\rm rig}$
REPRESSION
cyclic repression [CR; see Fig. 2]
global repression [GR; see Fig. 3]
PURIFICATION
selective purification [SP; see Fig. 4]
global purification [GP; see Fig. 4]
END

Fig. 1. Overview of the autoECART protocol and its key procedure, SOE (supervised outlier expulsion). Steps of SOE are defined in Figs. 2–4. For details, see text.

ally quasi-consistent with respect to $\Diamond$. If the BHO is not empty, then the execution of the autoECART procedure is interrupted, and the user is requested to remove the entries of BHO from $\mathcal{N}_{\rm anc}$. If there is no conflict within $\mathcal{N}_{\rm anc}$, the SOE module is recalled, with $\mathcal{N}_{\rm ela} = \mathcal{N}_{\rm LS} \setminus \mathcal{N}_{\rm anc}$ and $\mathcal{N}_{\rm rig} = \mathcal{N}_{\rm anc}$, to produce a BHO from the $\mathcal{N}_{\rm LS} \setminus \mathcal{N}_{\rm anc}$ subnetwork, and then the autoECART procedure is terminated.

Apparently, the BHO list constructed by autoECART needs to be thoroughly checked, and its transitions should be either corrected or excluded from $\mathcal{N}_{\rm LS}$ by the database builder. If certain BHO entries turn out to be pseudo-outliers, they should be placed into the $\mathcal{N}_{\rm anc}$ subnetwork. After these modifications, the autoECART protocol is to be repeated, and the BHO list should be reanalyzed until no outliers are found. When an empty BHO is received, it is worth setting $\mathcal{N}_{\rm k} = \Diamond$ and iterating the autoECART procedure until conditional quasi-consistency is reached for $\mathcal{N}_{\rm k} = \Diamond$, as well.

3.1. Repression procedure (RP)

During the repression procedure, a *preserved subnetwork* ($\mathcal{N}_{\rm pres}$) is defined, by setting $\mathcal{N}_{\rm pres} = \mathcal{N}_{\rm mer}$, and then successively transporting its putative outliers into the *greylist of hypothetical outliers* (*GHO*) until the conditional consistency of $\mathcal{N}_{\rm pres}$ is reached. $\mathcal{N}_{\rm pres}$ will aid, as a constraint in Eq. (22), the assessment of the GHO lines (see also Section 3.2).

After the initialization of $\mathcal{N}_{\rm pres}$, an iterative algorithm (*cyclic repression, CR*), invoking randomly chosen BFS-type CBs, is executed in an attempt to make $\mathcal{N}_{\rm pres}$ quasi-consistent. The CR protocol, whose flowchart is shown in Fig. 2, (a) removes all the bad, shortest, edge-disjoint basic cycles of the CBs from $\mathcal{N}_{\rm pres}$, and (b) places the lines of the eliminated cycles (outside $\mathcal{N}_{\rm rig}$) in the GHO. When no bad cycles can be identified, $\mu_{\rm retr}$ retrials are permitted to build further CBs and reveal their bad cycles 'hidden' so far, where $\mu_{\rm retr}$ is the (user-specified) *retrial margin*.

As the CR procedure allows only a non-exhaustive enumeration and repression of bad cycles, there is no guarantee that $\mathcal{N}_{\rm pres}$ has no incompatible lines. Thus, a *global repression* (*GR*) should be carried out to quench all the problematic transitions disrupting the quasi-consistency of $\mathcal{N}_{\rm pres}$. The GR method accomplishes three important tasks. First, the lines of $\mathcal{N}_{\rm pres} \setminus \mathcal{N}_{\rm rig}$ disobeying Eq. (22) are moved into the GHO. Second, GHO transitions with $d_{{\rm r},i} \le 0$ are reinstated into $\mathcal{N}_{\rm pres}$. One should be cautious with the latter step because the $d_{{\rm r},i}$ values of the reinstated lines may sometimes become positive after the recalculation of the constrained empirical energy values, initiating an infinite loop within the GR procedure. This convergence issue is avoided by permitting the reinstatement only $\mu_{\rm recu}$ times, where $\mu_{\rm recu}$ is the so-called *recuperation margin*. Third, the bridges produced by the CR/GR procedures, among which there may be possible outliers for $\mathcal{N}_{\rm pres}$, are deactivated and placed into the GHO list. The determination of bridges could be implemented by assembling a BFS-type CB and exploiting that bridges cannot take part in the basic cycles of this CB.

3.2. Purification procedure (PP)

Upon completion of the RP procedure, a conditionally quasi-consistent subnetwork $\mathcal{N}_{\rm pres}$ of $\mathcal{N}_{\rm mer}$ is obtained. Therefore, the lines of $\mathcal{N}_{\rm pres}$ provide a suitable restriction for the empirical energy values to collect the outliers of $\mathcal{N}_{\rm ela}$. For this specific purpose, a purified subnetwork, set to $\mathcal{N}_{\rm pur} = \mathcal{N}_{\rm mer}$, is constructed, from which transitions are progressively transferred into BHO to gain conditional quasi-consistency for $\mathcal{N}_{\rm pur}$ in a process, called *purification procedure* (*PP*), highly similar to the RP protocol. The PP method returns $\mathcal{N}_{\rm pur}$ as a near-largest, hopefully outlier-free subnetwork of $\mathcal{N}_{\rm mer}$ and yields the final BHO list, the main result of the autoECART procedure.

The *selective purification* (*SP*) technique, shown schematically in Fig. 4, is an elaborate protocol to cleanse the subnetwork $\mathcal{N}_{\rm pur} \setminus$

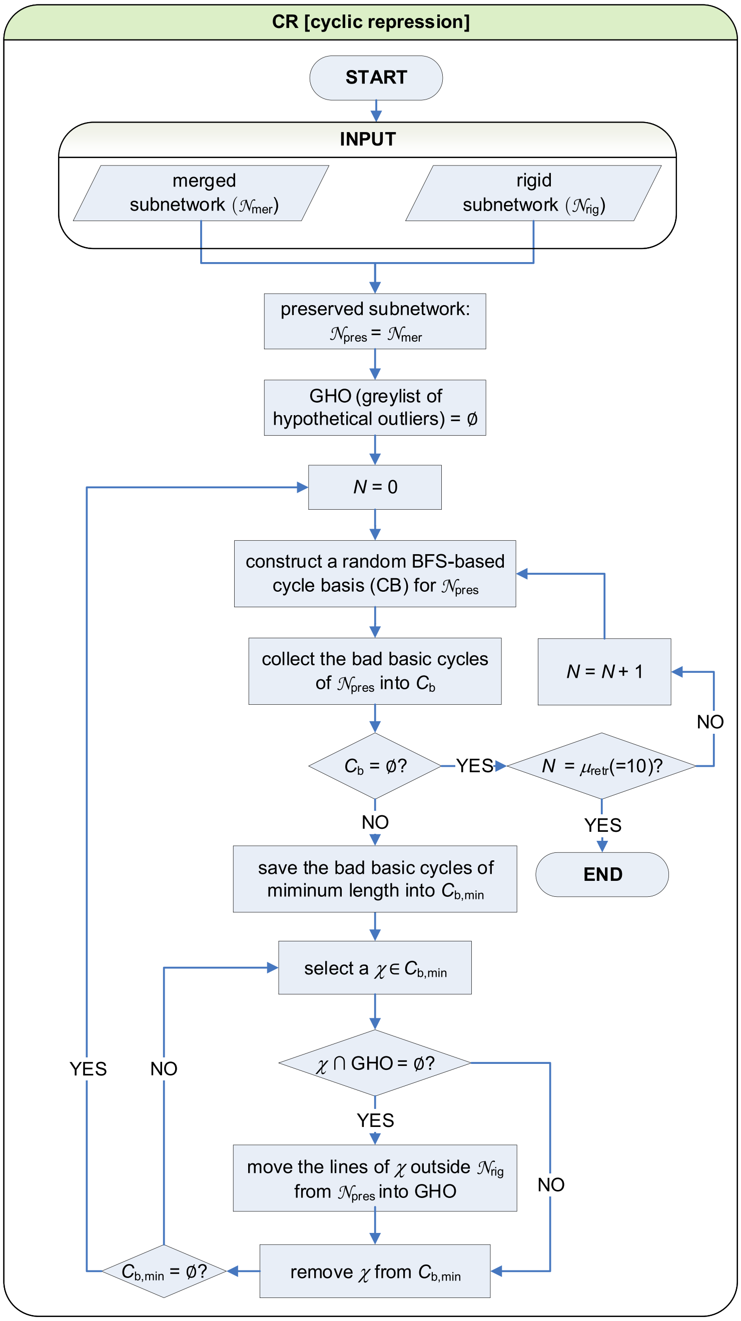

CR [cyclic repression]
START
INPUT
merged subnetwork ($\mathcal{N}_{\rm mer}$)
rigid subnetwork ($\mathcal{N}_{\rm rig}$)
preserved subnetwork: $\mathcal{N}_{\rm pres} = \mathcal{N}_{\rm mer}$
GHO (greylist of hypothetical outliers) = ∅
$N = 0$
construct a random BFS-based cycle basis (CB) for $\mathcal{N}_{\rm pres}$
collect the bad basic cycles of $\mathcal{N}_{\rm pres}$ into $C_{\rm b}$
$N = N + 1$
$C_{\rm b} = \emptyset$?
YES
$N = \mu_{\rm retr}(=10)$?
NO
YES
END
NO
save the bad basic cycles of miminum length into $C_{\rm b,min}$
select a $\chi \in C_{\rm b,min}$
YES
NO
$\chi \cap$ GHO = ∅?
YES
move the lines of χ outside $\mathcal{N}_{\rm rig}$ from $\mathcal{N}_{\rm pres}$ into GHO
NO
$C_{\rm b,min} = \emptyset$?
remove χ from $C_{\rm b,min}$

Fig. 2. Cyclic repression. The symbol '(=10)' denotes the suggested value of $\mu_{\rm retr}$. For details, see text.

$\mathcal{N}_{\rm pres}$ polluted with outliers. Within the SP method, the transitions of $\mathcal{N}_{\rm pur} \setminus \mathcal{N}_{\rm pres}$ having $d_{{\rm r},i} > 0$ are transposed into BHO. These transpositions are implemented in small fractions, censoring those transitions first for which $d_{{\rm r},i} > \phi_{\rm sp} d_{\rm r,max}$, where $d_{\rm r,max}$ is the maximum defect in $\mathcal{N}_{\rm pur} \setminus \mathcal{N}_{\rm pres}$, and $\phi_{\rm sp}$ means the *selective purification factor*. Notice that the application of $\mathcal{N}_{\rm pres}$ as a constraint not only makes the autoECART approach more sensitive to outliers, but it also leads to a significant speed-up for the subsequent calculation of the empirical energies. This acceleration is attributed to the fact that $\mathbf{G_C}$ is considerably smaller than $\mathbf{G}$ [see Eq. (13)].

After the completion of SP, it is advisable to test whether the transitions of $\mathcal{N}_{\rm pur}$ meet Eq. (22), even if $\mathcal{N}_{\rm k} = \mathcal{N}_{\rm rig}$ is utilized in Eqs. (9)–(13) rather than $\mathcal{N}_{\rm k} = \mathcal{N}_{\rm pres}$. In this spirit, a *global purification* (*GP*) is executed (see Fig. 4), which in fact corresponds to the call of the GR procedure with the (BHO, $\mathcal{N}_{\rm pur}$, $\mathcal{N}_{\rm rig}$, $\mathcal{N}_{\rm mer}$) argument list (see Fig. 3). At the termination of the GP step, a conditionally consistent $\mathcal{N}_{\rm pur}$ subnetwork is obtained, which is expected to have no outliers within its cycles.

It must be emphasized that the autoECART approach, despite its effectiveness (Section 4, *vide infra*), is not an 'exact' method (exact outlier detection protocols cannot be devised due to the ill-defined nature of the outliers from a mathematical point of view). In other words, it may occur that another heuristic algorithm could find some outliers remaining imperceptible for the autoECART procedure. For lack of a better option, one must accept this kind of deficiency of approximate methods, maintaining the possibility to constantly improve them, where feasible.

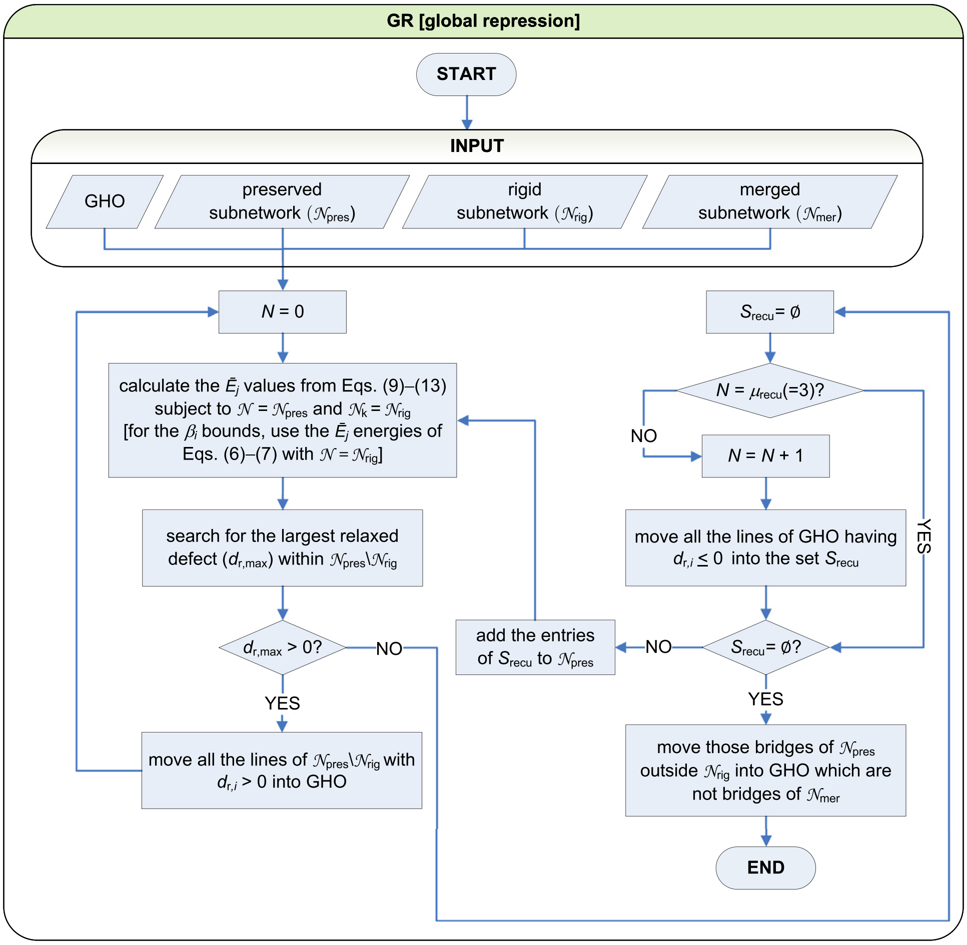

GR [global repression]
START
INPUT
GHO
preserved subnetwork ($\mathcal{N}_{\rm pres}$)
rigid subnetwork ($\mathcal{N}_{\rm rig}$)
merged subnetwork ($\mathcal{N}_{\rm mer}$)
$N = 0$
calculate the $\bar{E}_j$ values from Eqs. (9)–(13) subject to $\mathcal{N} = \mathcal{N}_{\rm pres}$ and $\mathcal{N}_{\rm k} = \mathcal{N}_{\rm rig}$ [for the β_i bounds, use the $\bar{E}_j$ energies of Eqs. (6)–(7) with $\mathcal{N} = \mathcal{N}_{\rm rig}$]
search for the largest relaxed defect ($d_{\rm r,max}$) within $\mathcal{N}_{\rm pres}\backslash\mathcal{N}_{\rm rig}$
$d_{\rm r,max} > 0$?
NO
YES
move all the lines of $\mathcal{N}_{\rm pres}\backslash\mathcal{N}_{\rm rig}$ with $d_{{\rm r},i} > 0$ into GHO
$S_{\rm recu} = \emptyset$
$N = \mu_{\rm recu}(=3)$?
NO
$N = N + 1$
YES
move all the lines of GHO having $d_{{\rm r},i} \le 0$ into the set $S_{\rm recu}$
$S_{\rm recu} = \emptyset$?
NO
add the entries of $S_{\rm recu}$ to $\mathcal{N}_{\rm pres}$
YES
move those bridges of $\mathcal{N}_{\rm pres}$ outside $\mathcal{N}_{\rm rig}$ into GHO which are not bridges of $\mathcal{N}_{\rm mer}$
END

Fig. 3. Global repression. The symbol '(=3)' denotes the suggested value of $\mu_{\rm recu}$. For details, see text.

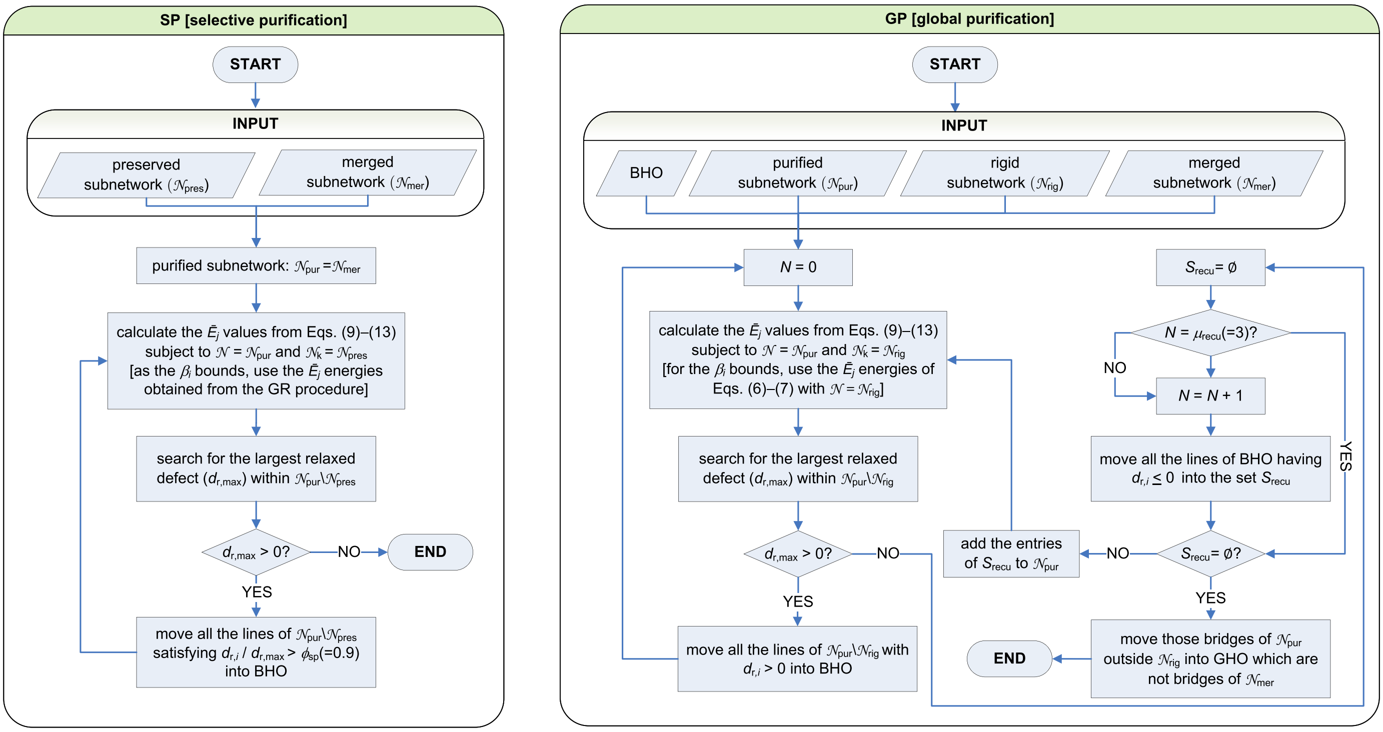

SP [selective purification]
START
INPUT
preserved subnetwork ($\mathcal{N}_{\rm pres}$)
merged subnetwork ($\mathcal{N}_{\rm mer}$)
purified subnetwork: $\mathcal{N}_{\rm pur} = \mathcal{N}_{\rm mer}$
calculate the $\bar{E}_j$ values from Eqs. (9)–(13) subject to $\mathcal{N} = \mathcal{N}_{\rm pur}$ and $\mathcal{N}_{\rm k} = \mathcal{N}_{\rm pres}$ [as the β_i bounds, use the $\bar{E}_j$ energies obtained from the GR procedure]
search for the largest relaxed defect ($d_{\rm r,max}$) within $\mathcal{N}_{\rm pur}\backslash\mathcal{N}_{\rm pres}$
$d_{\rm r,max} > 0$?
NO
END
YES
move all the lines of $\mathcal{N}_{\rm pur}\backslash\mathcal{N}_{\rm pres}$ satisfying $d_{{\rm r},i} / d_{\rm r,max} > \phi_{\rm sp}(=0.9)$ into BHO
GP [global purification]
START
INPUT
BHO
purified subnetwork ($\mathcal{N}_{\rm pur}$)
rigid subnetwork ($\mathcal{N}_{\rm rig}$)
merged subnetwork ($\mathcal{N}_{\rm mer}$)
$N = 0$
calculate the $\bar{E}_j$ values from Eqs. (9)–(13) subject to $\mathcal{N} = \mathcal{N}_{\rm pur}$ and $\mathcal{N}_{\rm k} = \mathcal{N}_{\rm rig}$ [for the β_i bounds, use the $\bar{E}_j$ energies of Eqs. (6)–(7) with $\mathcal{N} = \mathcal{N}_{\rm rig}$]
search for the largest relaxed defect ($d_{\rm r,max}$) within $\mathcal{N}_{\rm pur}\backslash\mathcal{N}_{\rm rig}$
$d_{\rm r,max} > 0$?
NO
YES
move all the lines of $\mathcal{N}_{\rm pur}\backslash\mathcal{N}_{\rm rig}$ with $d_{{\rm r},i} > 0$ into BHO
$S_{\rm recu} = \emptyset$
$N = \mu_{\rm recu}(=3)$?
NO
$N = N + 1$
YES
move all the lines of BHO having $d_{{\rm r},i} \le 0$ into the set $S_{\rm recu}$
$S_{\rm recu} = \emptyset$?
NO
add the entries of $S_{\rm recu}$ to $\mathcal{N}_{\rm pur}$
YES
move those bridges of $\mathcal{N}_{\rm pur}$ outside $\mathcal{N}_{\rm rig}$ into GHO which are not bridges of $\mathcal{N}_{\rm mer}$
END

Fig. 4. Selective (left panel) and global (right panel) purification. The symbols '(=0.9)' and '(=3)' denote the recommended values of parameters $\phi_{\rm sp}$ and $\mu_{\rm recu}$, respectively (see also text).

4. Performance of the autoECART code

The novel approach detailed in Section 3 was programmed, in the C++ language, into a code called autoECART. Our implementation uses an in-house version of the BFS and lowest-common-ancestor [55] searches, together with the sparse Cholesky factorization of the Eigen software package [56].

In this section, the utility of the autoECART program is analyzed for a few synthetic networks. Investigation of *synthetic spectroscopic networks* (SSN) is important to gain full control over the outlier analysis. For reasons of simplicity, each SSN will be identical to its $\mathcal{N}_{\rm LS}$ subnetwork, and $\mathcal{N}_{\rm anc}$ will be set to $\diamond$.

During the utility analysis, a specific SSN is constructed as follows: (a) a *contour SN* (*CSN*) is assembled from a subset of the non-excluded transitions contained in the IUPAC [35,40] and W2020 [51] databases of six ($HD^{16}O$, $HD^{17}O$, $HD^{18}O$, $D_2{}^{16}O$, $D_2{}^{17}O$, and $D_2{}^{18}O$) and three ($H_2{}^{16}O$, $H_2{}^{17}O$, and $H_2{}^{18}O$) molecules, respectively, (b) $\delta_i = \min(\delta_{0,i},\ 0.05\ \mathrm{cm}^{-1})$ is set for the line uncertainties, where $\delta_{0,i}$ is the uncertainty of the ith transition specified in Refs. [35,40,51], (c) the empirical energies are calculated by solving Eqs. (6) and (7), and (d) the synthetic analogue of the CSN, simply termed synthetic SN, is derived by performing the substitution

$$\sigma_i \leftarrow |\bar{E}_{\mathrm{up}(i)} - \bar{E}_{\mathrm{low}(i)} + o_i + a_i| \tag{23}$$

for every $1 \le i \le N_{\rm T}$. In Eq. (23), the $o_i \in [-\delta_i/2, \delta_i/2)$ uniform random variable imitates the *observational error* of the ith transition, and a_i denotes the *additional error* defined as

$$a_i = (1 - \mathcal{B}_i)\,\Sigma(c_i, \mathrm{CCR})\,\xi_i M_i. \tag{24}$$

Eq. (24) comprises the following parameters: (a) $\mathcal{B}_i$ is a binary variable designating whether the ith line corresponds to a bridge of the SN ($\mathcal{B}_i = 1$) or not ($\mathcal{B}_i = 0$), (b) $\xi_i \in \{-1, 1\}$ means a uniform random sign, (c) $M_i \in [\rho_1, \rho_2)$ is a uniform random number with its lower and upper limits $\rho_1 > 0$ and $\rho_2 \ge \rho_1$, respectively, (d) $c_i \in [0, 1)$ is another uniform random number, (e) $\mathrm{CCR} \in [0, 1]$ is the *critical contamination ratio*, and (f) $\Sigma(c_i, \mathrm{CCR})$ is a binary *switch function* with

$$\Sigma(c_i, \mathrm{CCR}) = \begin{cases} 1, & \text{if}\, c_i \le \mathrm{CCR}, \\ 0, & \text{otherwise}. \end{cases} \tag{25}$$

Eq. (25) suggests that the outliers of SSNs are those non-bridge lines which have nonzero additional errors. One can also realize that the CCR factor limits the $N_{\rm O}/N_{\rm T}$ fraction, where $N_{\rm O}$ is the number of outliers generated.

Evidently, a CSN can be associated with several randomly chosen SSNs. A way to provide high-diversity SSNs for the same CSN is to vary ρ_1, ρ_2, and CCR during the application of the construction scheme. The free choice of ρ_1 and ρ_2 allows the introduction of a *draft* and a *fine construction* for all CSNs in the following fashion:

$$\begin{aligned} \text{draft}: &\ \rho_1 = 0.1\ \mathrm{cm}^{-1} \text{ and } \rho_2 = 1000\ \mathrm{cm}^{-1}, \\ \text{fine}: &\ \rho_1 = 0.1\ \mathrm{cm}^{-1} \text{ and } \rho_2 = 1\ \mathrm{cm}^{-1}. \end{aligned} \tag{26}$$

In a similar way, one can form a number of quite dissimilar SSNs from a CSN by adjusting the CCR parameter in Eq. (24).

It is also important to scrutinize how the appearance of multiple edges influences the effectiveness of autoECART. This examination may be conducted by establishing two CSNs, a simple and a multiple one, for a given $H_2{}^XO$/HD^XO/$D_2{}^XO$ isotopologue ($X = 16, 17, 18$) from the non-excluded lines of Refs. [35,40,51]. A multiple CSN of an isotopologue is composed of all the non-excluded lines of the related W2020/IUPAC dataset; a simple CSN is created likewise, but only one line is retained from transitions with the same assignment.

Following these considerations, we set up 18 (nine simple and nine multiple) CSNs and derived 56 SSNs (28-28 in draft and fine modes) by adopting multiple CCR values (0.01, 0.02, 0.03, 0.05, 0.07, and 0.1) for $H_2{}^{16}O$ and a single value of CCR = 0.01 for the other eight molecules. Then, the autoECART code was deployed to unveil the outliers of these SSNs [by setting $\delta_{\mathrm{r},i} = \delta_i$ in Eq. (22)]. The numerical results are collected in Tables 1 and 2.

As displayed in Tables 1 and 2, the CSNs forming the bases of the SSNs embrace only a few network bridges (as compared to the total number of lines) and sufficient variety, both in size and topology [22], to critically evaluate the efficiency of the autoECART program. The draft-mode SSNs serve as to simulate the impact of possible line misassignments, while their fine-mode siblings mimic the potential mistakes in the transition wavenumbers. One can also observe in the penultimate columns of Tables 1 and 2 that the outliers generated increase the defects for an excessive number of correct lines, extremely aggravating the identification of the true outliers. In defiance of this serious complication, the autoECART code copes intelligently with discovering these outliers, yielding only a small number of pseudo-outliers (typically 10-15 % of the BHO). This ratio is slightly larger (20-25 %) for the W2020-based SSNs of H_2^XO ($X = 16, 17, 18$), which is ascribed to the virtual and complementary lines [51] causing poorer connections within the underlying CSNs.

It is also worth noting that most outliers are successfully debunked: the exceptional cases appearing in fine mode at higher CCRs are most certainly due to cancellation of almost identical additional errors. As to the two CSN types, no differences can be seen in the results achieved for SSNs of simple and multiple CSNs. The CPU time, being minuscule even for the largest SSNs, varies nearly quadratically with $N_{\rm T}$ at a fixed CCR value, while the growth does not follow a rigorous trend with respect to CCR, but also depends on the number of BFS forests utilized in the CR procedure.

Briefly, the tests performed with SSNs prove the general utility of the autoECART program to itemize the outliers caused by misassignments and flawed wavenumbers in a completely automated way. However, as observed in Table 2, there may be some outliers unidentified if all their bad cycles are destroyed during the CR process by repressing a couple of correct transitions of these cycles. Similarly, those outliers whose additional errors and the stated uncertainties are close to each other can also be identified only with difficulty or their exploration may involve considerably more pseudo-outliers. Note that the assignments of the BHO lines must be thoroughly analyzed, allowing the recognition of systematic misassignments in a specific source segment (these systematic errors may also induce some latent outliers).

5. Outliers in the GEISA-2019-$H_2{}^{18}O$ line list

As an illustrative and didactic example, the GEISA-2019-$H_2{}^{18}O$ dataset [52] is chosen to characterize the effectiveness of the autoECART program to detect outliers in popular spectroscopic information systems. This dataset comprises 41 214 rovibrational transitions, of which 6547 are only partially assigned. These incompletely assigned spectral lines had to be neglected during the present analysis.

Since most of the transition wavenumbers do not have individual uncertainties in the dataset, a conservative estimate of $\delta_i = 0.01\ \mathrm{cm}^{-1}$ is applied here for each transition, allowing the recognition of rough outliers. Moreover, $\delta_{\mathrm{r},i} = 0.05\ \mathrm{cm}^{-1}$ is used for each line as a relaxed uncertainty, required by the autoECART protocol. While one can adopt tighter uncertainty intervals to identify more subtle outliers, such a detailed analysis is beyond the scope of this proof-of-concept study.

As a result of its first execution, the autoECART program blacklisted 20 transitions, reported in Table 3. This table also shows the W2020-HotWat78 [51,59] hybrid linelist counterparts of the

Table 1
Outlier detection in draft-mode synthetic spectroscopic networks (SSN) of nine water isotopologues.

species[a]	CSN type[b]	CCR[c]	N_T[d]	N_B[e]	N_O[f]	N_IO[g]	N_PO[h]	N_BHO[i]	N_pd[j]	t_run[k]/s
$\mathrm{H_2{}^{16}O}$	simple	0.01	114 033	2413	1114	1114	351	1465	110 205	36.7
$\mathrm{H_2{}^{16}O}$	simple	0.02	114 033	2413	2277	2277	581	2858	110 439	38.6
$\mathrm{H_2{}^{16}O}$	simple	0.03	114 033	2413	3352	3352	717	4069	110 533	37.6
$\mathrm{H_2{}^{16}O}$	simple	0.05	114 033	2413	5481	5481	1145	6626	110 762	40.9
$\mathrm{H_2{}^{16}O}$	simple	0.07	114 033	2413	7804	7804	1604	9408	110 858	42.1
$\mathrm{H_2{}^{16}O}$	simple	0.10	114 033	2413	11 019	11 019	2255	13 274	111 018	46.3
$\mathrm{H_2{}^{17}O}$	simple	0.01	17 830	1265	151	151	83	234	16 490	1.3
$\mathrm{H_2{}^{18}O}$	simple	0.01	26 697	1466	230	230	78	308	25 147	2.3
$\mathrm{HD^{16}O}$	simple	0.01	36 057	1912	368	368	68	436	34 123	3.7
$\mathrm{HD^{17}O}$	simple	0.01	443	13	5	5	2	7	430	0.0
$\mathrm{HD^{18}O}$	simple	0.01	7186	445	69	69	15	84	6736	0.3
$\mathrm{D_2{}^{16}O}$	simple	0.01	43 045	2287	354	354	61	415	40 625	4.9
$\mathrm{D_2{}^{17}O}$	simple	0.01	547	97	5	5	2	7	436	0.1
$\mathrm{D_2{}^{18}O}$	simple	0.01	9748	672	83	83	17	100	9055	0.5
$\mathrm{H_2{}^{16}O}$	multiple	0.01	287 659	1894	2775	2775	176	2951	279 770	98.0
$\mathrm{H_2{}^{16}O}$	multiple	0.02	287 659	1894	5793	5793	243	6036	282 200	117.5
$\mathrm{H_2{}^{16}O}$	multiple	0.03	287 659	1894	8500	8500	398	8898	282 847	128.0
$\mathrm{H_2{}^{16}O}$	multiple	0.05	287 659	1894	13 999	13 999	596	14 595	283 386	148.6
$\mathrm{H_2{}^{16}O}$	multiple	0.07	287 659	1894	20 266	20 266	742	21 008	283 637	163.5
$\mathrm{H_2{}^{16}O}$	multiple	0.10	287 659	1894	28 561	28 561	1022	29 583	283 941	173.1
$\mathrm{H_2{}^{17}O}$	multiple	0.01	26 894	1168	235	235	40	275	25 288	2.3
$\mathrm{H_2{}^{18}O}$	multiple	0.01	65 619	1132	665	665	42	707	62 560	7.0
$\mathrm{HD^{16}O}$	multiple	0.01	53 369	1736	507	507	40	547	51 205	6.3
$\mathrm{HD^{17}O}$	multiple	0.01	483	13	5	5	1	6	470	0.2
$\mathrm{HD^{18}O}$	multiple	0.01	8729	431	74	74	7	81	8264	0.4
$\mathrm{D_2{}^{16}O}$	multiple	0.01	52 842	2236	486	486	50	536	50 109	7.5
$\mathrm{D_2{}^{17}O}$	multiple	0.01	583	93	6	6	3	9	463	0.2
$\mathrm{D_2{}^{18}O}$	multiple	0.01	12 001	656	111	111	9	120	11 272	0.6

[a] Chemical formula of a water isotopologue. [b] Type of the underlying countour spectroscopic network (CSN). [c] Critical contamination ratio (CCR) of a SSN, defined in Eq. (25). [d] Number of transitions in a SSN. [e] Number of bridges in a SSN. [f] Number generated outliers in a SSN. [g] Number of identified outliers in a SSN. [h] Number of pseudo outliers in a SSN. [i] Number of transitions in the blacklist of hypothetical outliers (BHO) of a SSN. [j] Number of transitions with positive relaxed defects in a SSN. [k] Running time, related to a Lenovo Legion Y530 (81FV00T4HV) Notebook, for a SSN, in seconds.

Table 2
Outlier detection in fine-mode synthetic spectroscopic networks (SSN) of nine water isotopologues[a].

species	CSN type	CCR	N_T	N_B	N_O	N_IO	N_PO	N_BHO	N_pd	t_run/s
$\mathrm{H_2{}^{16}O}$	simple	0.01	114 033	2413	1072	1072	318	1390	68 022	29.2
$\mathrm{H_2{}^{16}O}$	simple	0.02	114 033	2413	2251	2251	613	2864	81 850	33.3
$\mathrm{H_2{}^{16}O}$	simple	0.03	114 033	2413	3413	3413	793	4206	91 182	35.6
$\mathrm{H_2{}^{16}O}$	simple	0.05	114 033	2413	5405	**5403**	1160	6563	98 431	38.1
$\mathrm{H_2{}^{16}O}$	simple	0.07	114 033	2413	7841	**7839**	1633	9472	101 522	43.3
$\mathrm{H_2{}^{16}O}$	simple	0.10	114 033	2413	11 101	**11 097**	2053	13 150	103 568	44.6
$\mathrm{H_2{}^{17}O}$	simple	0.01	17 830	1265	158	158	56	214	12 008	1.4
$\mathrm{H_2{}^{18}O}$	simple	0.01	26 697	1466	268	268	95	363	17 269	2.3
$\mathrm{HD^{16}O}$	simple	0.01	36 057	1912	349	349	59	408	25 314	3.4
$\mathrm{HD^{17}O}$	simple	0.01	443	13	3	3	1	4	395	0.1
$\mathrm{HD^{18}O}$	simple	0.01	7186	445	57	57	8	65	4896	0.3
$\mathrm{D_2{}^{16}O}$	simple	0.01	43 045	2287	445	445	77	522	26 197	5.1
$\mathrm{D_2{}^{17}O}$	simple	0.01	547	97	5	5	3	8	229	0.1
$\mathrm{D_2{}^{18}O}$	simple	0.01	9748	672	78	78	13	91	5465	0.5
$\mathrm{H_2{}^{16}O}$	multiple	0.01	287 659	1894	2951	2951	226	3177	173 312	89.4
$\mathrm{H_2{}^{16}O}$	multiple	0.02	287 659	1894	5769	5769	286	6055	216 798	107.6
$\mathrm{H_2{}^{16}O}$	multiple	0.03	287 659	1894	8628	8628	325	8953	220 356	120.8
$\mathrm{H_2{}^{16}O}$	multiple	0.05	287 659	1894	14 506	14 506	534	15 040	256 112	145.4
$\mathrm{H_2{}^{16}O}$	multiple	0.07	287 659	1894	19 893	19 893	832	20 725	269 809	169.4
$\mathrm{H_2{}^{16}O}$	multiple	0.10	287 659	1894	28 629	**28 623**	1020	29 643	271 404	177.9
$\mathrm{H_2{}^{17}O}$	multiple	0.01	26 894	1168	252	252	38	290	13 651	2.2
$\mathrm{H_2{}^{18}O}$	multiple	0.01	65 619	1132	639	639	47	686	34 504	6.5
$\mathrm{HD^{16}O}$	multiple	0.01	53 369	1736	515	515	35	550	39 915	6.1
$\mathrm{HD^{17}O}$	multiple	0.01	483	13	6	6	2	8	348	0.1
$\mathrm{HD^{18}O}$	multiple	0.01	8729	431	77	77	9	86	5842	0.4
$\mathrm{D_2{}^{16}O}$	multiple	0.01	52 842	2236	514	514	75	589	34 386	7.4
$\mathrm{D_2{}^{17}O}$	multiple	0.01	583	93	8	8	2	10	330	0.2
$\mathrm{D_2{}^{18}O}$	multiple	0.01	12 001	656	113	113	27	140	7340	0.7

[a] Headings of the columns are defined in the footnote to Table 1. Where N_O and N_IO are different, the corresponding numbers are indicated in boldface.

suspicious GEISA lines, facilitating the discovery of possible mistakes. Following a careful analysis of the GEISA data, the suspicious lines could be divided into four groups (I-IV, see Table 3).

Group I contains three transitions, #1, #4, and #5, for which no rigorous matches could be found in the W2020-HotWat78 catalogue. Line #1 taks part in a bad cycle of the GEISA-2015-$H_2^{18}O$ database (see Fig. 3 of Ref. [28]). This line is possibly taken from HITRAN [19]. As this transition will be deleted from HITRAN 2020 [60] due to its unsuccessful validation, we recommend to exclude line #1 from the GEISA-2019 database, as well. As to transitions #4 and #5, their counterparts could be found only in the W2020-$H_2^{17}O$ and W2020-$H_2^{16}O$ compilations, respectively. As long as the managers of the GEISA-2019 database find these matches convincing, they should move lines #4 and #5 into the GEISA-2019-$H_2^{17}O$ and GEISA-2019-$H_2^{16}O$ line collections, respectively (relying on the related W2020 line assignments). An alternative option is to simply remove these faulty transitions from GEISA-2019.

Group II is made up of five spectral lines, #2, #3, #7, #18, and #20. Since these transitions are completely reproduced by their W2020/HotWat78 siblings, they are pseudo-outliers, which do not require corrections.

Table 3
Analysis of the blacklist of hypothetic outilers obtained for the GEISA-2019-$H_2{}^{18}O$ database [52][a].

#	σ/cm^{-1}	S/cm molecule^{-1}	Assignment	Source	Group
1	5.673 109	3.511×10^{-35}	$(010)15_{9,6} \leftarrow (010)14_{10,5}$	TE3	I
	–	–	–	–	
2	2 338.422 326	1.273×10^{-29}	$(010)15_{9,6} \leftarrow (000)14_{8,7}$	TE3	II
	2 338.35(20)	1.276×10^{-29}	$(010)15_{9,6} \leftarrow (000)14_{8,7}$	HotWat78	
3	2 359.275 112	1.698×10^{-29}	$(010)14_{10,5} \leftarrow (000)13_{9,4}$	TE3	II
	2 359.26(20)	1.702×10^{-29}	$(010)14_{10,5} \leftarrow (000)13_{9,4}$	HotWat78	
4	4 283.968 400	2.140×10^{-27}	$(001)9_{9,1} \leftarrow (000)8_{7,2}$	S13	I
	[4 283.908 30(50)	1.141×10^{-27}	$(001)11_{7,5} \leftarrow (000)10_{5,6}$	W2020-$H_2^{17}O$]	
5	5 488.928 220	1.720×10^{-27}	$(110)12_{3,10} \leftarrow (000)11_{2,9}$	G96	I
	[5 488.928 4(10)	5.688×10^{-29}	$(031)9_{5,5} \leftarrow (020)8_{5,4}$	W2020-$H_2^{16}O$]	
6	**5 737.821 200**	7.839×10^{-30}	$(030)8_{7,2} \leftarrow (000)7_{4,3}$	CA9	III
	5 737.720 37(50)	7.336×10^{-30}	$(030)8_{7,2} \leftarrow (000)7_{4,3}$	W2020	
7	5 845.752 700	6.545×10^{-29}	$(110)12_{3,10} \leftarrow (000)11_{0,11}$	CA9	II
	5 845.752 70(50)	7.366×10^{-29}	$(110)12_{3,10} \leftarrow (000)11_{0,11}$	W2020	
8	6 608.254 670	4.264×10^{-26}	$(\mathbf{021})\mathbf{3}_{\mathbf{1,3}} \leftarrow (000)4_{3,2}$	CA9	IV
	6 608.254 99(10)	4.295×10^{-26}	$(120)3_{2,1} \leftarrow (000)4_{3,2}$	W2020	
9	6 610.972 890	5.510×10^{-27}	$(\mathbf{120})\mathbf{3}_{\mathbf{2,1}} \leftarrow (000)4_{3,2}$	CA9	IV
	6 610.973 30(10)	5.524×10^{-27}	$(021)3_{1,3} \leftarrow (000)4_{3,2}$	W2020	
10	6 705.239 100	1.610×10^{-26}	$(\mathbf{021})\mathbf{3}_{\mathbf{1,3}} \leftarrow (000)3_{3,0}$	CA9	IV
	6 705.239 53(10)	1.548×10^{-26}	$(120)3_{2,1} \leftarrow (000)3_{3,0}$	W2020	
11	6 707.957 320	2.810×10^{-27}	$(\mathbf{120})\mathbf{3}_{\mathbf{2,1}} \leftarrow (000)3_{3,0}$	CA9	IV
	6 707.957 84(10)	2.860×10^{-27}	$(021)3_{1,3} \leftarrow (000)3_{3,0}$	W2020	
12	6 763.717 770	1.030×10^{-24}	$(\mathbf{021})\mathbf{3}_{\mathbf{1,3}} \leftarrow (000)4_{1,4}$	CA9	IV
	6 763.718 04(10)	9.141×10^{-25}	$(120)3_{2,1} \leftarrow (000)4_{1,4}$	W2020	
13	6 766.435 990	9.160×10^{-25}	$(\mathbf{120})\mathbf{3}_{\mathbf{2,1}} \leftarrow (000)4_{1,4}$	CA9	IV
	6 766.436 34(10)	8.318×10^{-25}	$(021)3_{1,3} \leftarrow (000)4_{1,4}$	W2020	
14	6 814.663 380	1.060×10^{-25}	$(\mathbf{021})\mathbf{3}_{\mathbf{1,3}} \leftarrow (000)3_{1,2}$	CA9	IV
	6 814.663 64(10)	1.010×10^{-25}	$(120)3_{2,1} \leftarrow (000)3_{1,2}$	W2020	
15	6 817.381 600	4.030×10^{-25}	$(\mathbf{120})\mathbf{3}_{\mathbf{2,1}} \leftarrow (000)3_{1,2}$	CA9	IV
	6 817.381 95(10)	3.595×10^{-25}	$(021)3_{1,3} \leftarrow (000)3_{1,2}$	W2020	
16	6 908.557 590	1.940×10^{-24}	$(\mathbf{021})\mathbf{3}_{\mathbf{1,3}} \leftarrow (000)2_{1,2}$	CA9	IV
	6 908.557 88(10)	1.717×10^{-24}	$(120)3_{2,1} \leftarrow (000)2_{1,2}$	W2020	
17	6 911.275 810	8.510×10^{-25}	$(\mathbf{120})\mathbf{3}_{\mathbf{2,1}} \leftarrow (000)2_{1,2}$	CA9	IV
	6 911.276 19(10)	7.695×10^{-25}	$(021)3_{1,3} \leftarrow (000)2_{1,2}$	W2020	
18	7 396.534 000	7.643×10^{-30}	$(002)13_{7,7} \leftarrow (000)12_{8,4}$	CA9	II
	7 396.70(20)	7.341×10^{-30}	$(002)13_{7,7} \leftarrow (000)12_{8,4}$	HotWat78	
19	8 141.363 000	6.673×10^{-30}	$(\mathbf{050})\mathbf{11}_{\mathbf{5,7}} \leftarrow (000)12_{2,10}$	CA9	IV
	8 141.50(20)	7.457×10^{-30}	$(031)11_{2,9} \leftarrow (000)12_{2,10}$	HotWat78	
20	8 151.504 000	9.509×10^{-30}	$(050)11_{5,7} \leftarrow (000)12_{2,10}$	CA9	II
	8 151.64(20)	1.026×10^{-29}	$(050)11_{5,7} \leftarrow (000)12_{2,10}$	HotWat78	
21	7 426.475 000	1.566×10^{-29}	$(\mathbf{002})\mathbf{13}_{\mathbf{7,7}} \leftarrow (000)12_{8,4}$	CA9	IV
	7 426.49(20)	1.547×10^{-29}	$(101)13_{8,5} \leftarrow (000)12_{8,4}$	HotWat78	
22	7 793.709 000	9.543×10^{-30}	$(\mathbf{002})\mathbf{13}_{\mathbf{7,7}} \leftarrow (000)12_{6,6}$	CA9	IV
	7 793.72(20)	9.334×10^{-30}	$(101)13_{8,5} \leftarrow (000)12_{6,6}$	HotWat78	

[a] The first column contains the serial numbers of the BHO transitions provided by the double execution of the autoECART program. Where possible, the BHO lines are paired with their W2020 [51] counterparts to reveal the origin of the conflicts detected. For lines #4 and #5, only poor matches are found with a $H_2^{17}O$ and a $H_2^{16}O$ transition, respectively; thus, these W2020 lines are placed in square brackets. The second and third columns contain the wavenumbers (σ) and the intensities (S) of the individual transitions, respectively. The intensities are related to room temperature and corrected for the natural $H_2^{18}O$ abundance of 0.2 %. The fourth column specifies an assignment for each line as follows: $(v_1' \, v_2' \, v_3')J'_{K_a',K_c'} \leftarrow (v_1'' \, v_2'' \, v_3'')J''_{K_a'',K_c''}$, where (a) $'$ and $''$ refer to the upper and lower states, respectively, (b) $(v_1 \, v_2 \, v_3)J_{K_a,K_c}$ designates a rovibrational state, (c) $(v_1 \, v_2 \, v_3)$ is composed of the vibrational normal-mode quantum numbers reflecting the Mulliken convention [57], and (d) J_{K_a,K_c} stands for the standard asymmetric-top rotational quantum numbers [58]. The fifth column enumerates the GEISA sources (TE3, S13, G96, and CA9), as well as the sources of the reference lines (W2020 [51] and HotWat78 [59]). The meaning of the four source tags is not clarified on the GEISA website [52]. The last column tabulates four types of group indices: (I) unmatchable or poorly matchable transition, (II) confirmed line (pseudo-outlier), (III) possible transcription error, and (IV) misassignment. The possibly erroneous spectroscopic data are highlighted in boldface. The last two transitions isolated by the solid line were blacklisted during the re-execution of the autoECART code after moving the lines of group II into $\mathcal{N}_{\mathrm{anc}}$.

Group III is formed by only line #6, for which a residual of $\sim$ 0.1 cm^{-1} is obtained at the termination of the autoECART program. This large residual agrees perfectly with the deviation between the GEISA-2019 and W2020 wavenumbers, 0.100 83 cm^{-1}, implying that a transcription error was made in the first decimal place of the GEISA wavenumber value by the builders of the GEISA-2019 catalogue.

Group IV encompasses transitions #8–#17 and #19 with incorrect upper-state quantum numbers. As the W2020/HotWat78 counterparts indicate, these assignment problems are due to the interchanges $(0\,2\,1)3_{1,3} \leftrightarrow (1\,2\,0)3_{2,1}$ and $(0\,5\,0)11_{5,7} \leftrightarrow (0\,3\,1)\ 11_{2,9}$.

After (a) ignoring transitions #1, #4, and #5, (b) rectifying the putative typographical error in line #6, (c) reassigning transitions #8–#17 and #19, as well as (d) transporting the adjusted lines into $\mathcal{N}_{\mathrm{anc}}$, the autoECART code was re-executed, detecting further two BHO transitions (see lines # 21 and #22 in Table 3). These two transitions pertain to group IV. Following the reassignment of lines #21 and #22, as well as the relocation of the transitions from $\mathcal{N}_{\mathrm{anc}}$ to $\mathcal{N}_{\mathrm{LS}}$, the third execution of autoECART found no further BHO transitions in the repaired database.

6. Conclusions

It is next to impossible to identify all the incorrect transitions in huge line-by-line (LBL) rovibronic databases without heavy reliance on state-of-the-art theoretical and computational methods. This paper addresses the theoretical and practical aspects of outlier detection in LBL databases *via* network theory. As a result, a robust and efficient heuristic algorithm is provided, that yields, in an automated fashion, a comprehensive list of those outliers which can be detected *via* network-theoretical tools.

Our newly developed approach is named autoECART, standing for automatic Energy Conservation Analysis of Rovibronic Transitions. autoECART exploits the facts that (a) LBL databases can be represented with spectroscopic networks (SN), (b) these SNs have a large number of cycles, and (c) all these cycles need to satisfy the law of energy conservation (LEC). Although the technicalities related to the autoECART protocol are slightly involved, as shown in Figs. 1–4, one should keep in mind that the two basic operations are as follows: (i) the bad cycles, including at least one outlier, are judiciously selected and disconnected from the SN, and (ii) a blacklist of hypothetical outliers is built by constraining the empirical energies to the remaining (most likely outlier-free) subnetwork during a least-squares iteration.

The efficacy of the autoECART protocol has been evaluated on synthetic SNs of various size, where the outliers are known *a priori*. The ensuing systematic tests corroborate that (a) autoECART is capable of detecting nearly all the outliers with just a few exceptions, (b) the number of pseudo-outliers, corresponding to valid lines found to be suspicious by the autoECART algorithm, is fairly small, and (c) the running time is insignificant even for SNs composed of hundreds of thousands of transitions.

Since autoECART is an approximate procedure, its sensitivity could be enhanced in the future with the inclusion of other powerful heuristics. Nevertheless, none of these improvements could lift the theoretical limitations concerning latent outliers (that is, those outlying transitions whose detection is impossible based upon the available spectral information). These special outliers can be diagnosed only by augmenting the SN with additional transitions deduced, *e.g.*, from new, appropriately designed spectroscopic measurements [61,62].

The autoECART protocol has been employed to discover rough outliers in one real LBL dataset. For this purpose, the GEISA-2019-H$_2{}^{18}$O line collection [52] was selected, in which 22 doubtful lines could be recognized by our autoECART code. Of these potential outliers, a comparison with the hybrid W2020-HotWat78 [51,59] database revealed 17 truly anomalous lines. It should also be noted that there are $\sim$6500 transitions with partial assignments in the GEISA-2019-H$_2{}^{18}$O dataset: they could not be checked during this study, because autoECART demands unique labels for the investigated spectral lines.

As a long-term initiative, we plan to produce a web-based implementation of the autoECART approach, able to process different input formats used in prominent spectroscopic information systems, such as HITRAN [19] and GEISA [52]. This online form of the autoECART code could greatly alleviate the laborious work of the managers of spectroscopic databases to assemble consistent sets of rovibronic transitions. (As to the extremely large high-temperature line lists, like HITEMP [63], autoECART must be accelerated further by technical or coding improvements.)

Such a web application would also aid the decontamination of newly constructed or updated MARVEL [30,32,33] datasets, where the lines collected from dozens of data sources must be corrected and synchronized. Finally, note that a user-friendly autoECART software could considerably help the efforts of high-school students involved in academic research as active participants of MARVEL projects [42,43,46,47,64].

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgement

RT and AGC received support from the ELTE Institutional Excellence Program (TKP2020-IKA-05), from the grant VEKOP-2.3.2-16-2017-000, and from NKFIH (no. K119658). CS is thankful for financial support from NKFP (no. FK-124814). KB acknowledges support from János Bolyai Research Fellowship of the Hungarian Academy of Sciences, as well as from projects no. NKFI-128673, ÚNKP-20-5, and TKP2020-NKA-06 provided by the NRDI found of Hungary and financed under the FK_18, the ÚNKP-20-5, and the Thematic Excellence Programme (National Challenges Subprogramme) funding schemes, respectively.

Appendix A. A necessary and sufficient condition for consistency

This section provides a rigorous proof for the proposition that a subnetwork $\mathcal{N}$ of the SN is consistent if and only if all of its cycles are regular. Before the proof, some notions and symbols should be defined.

Consider the following connected sequence of energy levels, signs, and transitions:

$$\omega = \lambda_{J_1}[S_1, \tau_{I_1}]\lambda_{J_2}[S_2, \tau_{I_2}]\ldots\lambda_{J_L}[S_L, \tau_{I_L}]\lambda_{J_{L+1}}, \tag{A.1}$$

where (a) L is the *length* of ω, (b) τ_i and λ_j indicate the ith transition and the jth energy level of the SN, respectively, while (c)

$$\mathbb{J}[\omega] = \langle J_p : 1 \le p \le L+1 \rangle, \tag{A.2}$$

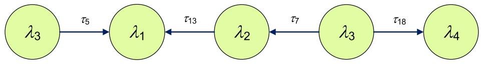
$$\mathbb{I}[\omega] = \langle I_q : 1 \le q \le L \rangle, \tag{A.3}$$

and

$$\mathbb{S}[\omega] = \langle S_q : 1 \le q \le L \rangle \tag{A.4}$$

are freely chosen lists of energy-level indices, transition indices, and signs, respectively. The ω sequence is known as a *walk* (see Fig. A.1) if I_q and S_q meet the following connections for $1 \le q \le L$:

$$\mathrm{up}(I_q) \in \{J_q, J_{q+1}\}, \tag{A.5}$$



State indices	Transition indices	Signs	Lower-state indices	Upper-state indices
$J_1 = 3$	$I_1 = 5$	$S_1 = +1$	low(I_1) = 3	up(I_1) = 1
$J_2 = 1$	$I_2 = 13$	$S_2 = -1$	low(I_2) = 2	up(I_2) = 1
$J_3 = 2$	$I_3 = 7$	$S_3 = -1$	low(I_3) = 3	up(I_3) = 2
$J_4 = 3$	$I_4 = 18$	$S_4 = +1$	low(I_4) = 3	up(I_4) = 4
$J_5 = 4$				

Fig. A.1. Exemplary walk of length 4 with its parameters. Utilizing the notation of Eq. (A.1), this walk can be given as $\omega = \lambda_{J_1}[S_1, \tau_{I_1}]\lambda_{J_2}[S_2, \tau_{I_2}]\lambda_{J_3}[S_3, \tau_{I_3}]\lambda_{J_4}[S_4, \tau_{I_4}]\lambda_{J_5}$.

$$\text{low}(I_q) = \begin{cases} J_q, & \text{if } \text{up}(I_q) = J_{q+1}, \\ J_{q+1}, & \text{if } \text{up}(I_q) = J_q, \end{cases} \tag{A.6}$$

$$S_q = \begin{cases} +1, & \text{if } \text{up}(I_q) = J_{q+1}, \\ -1, & \text{if } \text{up}(I_q) = J_q. \end{cases} \tag{A.7}$$

For each walk ω and any $\boldsymbol{\eta} = \{\eta_1, \eta_2, \ldots, \eta_{N_L}\}^{\mathrm{T}}$, an extremely useful relation can be derived:

$$\begin{aligned} &\sum_{q=1}^{L} S_q[\eta_{\text{up}(I_q)} - \eta_{\text{low}(I_q)}] = \sum_{q=1}^{L}[\eta_{J_{q+1}} - \eta_{J_q}] = \\ &\eta_{J_2} - \eta_{J_1} + \eta_{J_3} - \eta_{J_2} + \ldots + \eta_{J_L} - \eta_{J_{L-1}} + \\ &\eta_{J_{L+1}} - \eta_{J_L} = \eta_{J_{L+1}} - \eta_{J_1}. \end{aligned} \tag{A.8}$$

Note that Eq. (A.8) becomes zero if $J_{L+1} = J_1$.

The shorthand notation that walk ω passes from λ_{J_1} to $\lambda_{J_{L+1}}$ is $J_1 \rightsquigarrow J_{L+1}$. Furthermore, if i is an entry of $\mathbb{U}[\omega]$, then this statement is denoted as $i \in \mathbb{U}[\omega]$, where $\mathbb{U} \in \{\mathbb{J}, \mathbb{I}\}$. In what follows, walks are designated by extending the symbol ω with extra marks (like prime, tilde, bar, superscripts, *etc.*), while their parameters (L, J_p, I_q, and S_q) are augmented with the same marks. For example, the parameters of ω' are indicated as L', J'_p, I'_q, and S'_q.

For walks, two operations can be defined. If $J_L \in \{\text{up}(i), \text{low}(i)\}$ in Eq. (A.2), then $\omega' = \omega \oplus \tau_i$ is another walk of length $L' = L + 1$, where

$$J'_p = \begin{cases} J_p, & \text{if } 1 \le p \le L', \\ \text{up}(i), & \text{if } p = L' + 1 \text{ and } J_{L'} = \text{low}(i), \\ \text{low}(i), & \text{if } p = L' + 1 \text{ and } J_{L'} = \text{up}(i), \end{cases} \tag{A.9}$$

$$I'_q = \begin{cases} I_q, & \text{if } 1 \le q \le L, \\ i, & \text{if } q = L', \end{cases} \tag{A.10}$$

$$S'_q = \begin{cases} S_q, & \text{if } 1 \le q \le L, \\ +1, & \text{if } q = L' \text{ and } J_{L'} = \text{low}(i), \\ -1, & \text{if } q = L' \text{ and } J_{L'} = \text{up}(i), \end{cases} \tag{A.11}$$

while $\oplus$ is named *suffixation*. Provided that ω' and ω'' are walks with $J''_1 = J'_{L'}$, $\omega = \omega' \odot \omega''$ is also a walk with $L = L' + L''$, where $\odot$ is an operation called *concatenation*,

$$U_q = \begin{cases} U'_q, & \text{if } 1 \le q \le L' + t_U, \\ U''_{q-L'}, & \text{if } L' + t_U + 1 \le q \le L + t_U, \end{cases} \tag{A.12}$$

for any $U \in \{J, I, S\}$, while $t_J = 1$ and $t_I = t_S = 0$.

In this section, certain special walks will be given particular attention. A walk ω of length zero, which includes only λ_{J_1} with no transitions, is called a *trivial walk* and denoted with $\omega^{\triangledown}$. As long as both $\mathbb{I}[\omega]$ and $\mathbb{J}[\omega]$ embrace distinct indices, ω becomes a *path*. If $i \not\in \mathbb{I}[\omega]$, and $\text{up}(i), \text{low}(i) \in \{J_1, J_{L+1}\}$ for a path ω, then $\omega \oplus \tau_i$ is a *cycle*. It can be demonstrated straightforwardly that the signs of $\mathbb{S}(\omega)$ form a solution of Eq. (17) for a cycle ω of $\mathcal{N}$.

Adapting Eq. (15) to a cycle ω and making a few trivial rearrangements of Eq. (15) results in

$$\sum_{q=1}^{L}\left[\delta_{I_q} + S_q\sigma_{I_q}\right] \ge 0 \tag{A.13}$$

and

$$\sum_{q=1}^{L}\left[\delta_{I_q} - S_q\sigma_{I_q}\right] \ge 0. \tag{A.14}$$

By choosing

$$\mathcal{W}_i^{\langle +1\rangle} = \delta_i + \sigma_i \tag{A.15}$$

and

$$\mathcal{W}_i^{\langle -1\rangle} = \delta_i - \sigma_i \tag{A.16}$$

for all $1 \le i \le N_{\mathrm{T}}$, one can introduce the following parameters for the left-hand sides of Eqs. (A.13)–(A.14):

$$\mathcal{K}^{+}[\omega] = \sum_{q=1}^{L}\mathcal{W}_{I_q}^{\langle S_q\rangle} \tag{A.17}$$

and

$$\mathcal{K}^{-}[\omega] = \sum_{q=1}^{L}\mathcal{W}_{I_q}^{\langle -S_q\rangle}, \tag{A.18}$$

which are referred to as the *cost* and the *anti-cost* of walk ω, respectively. By means of the cost $\mathcal{K}^{\pm}[\omega]$, one can say that ω is regular if and only if

$$\mathcal{K}^{+}(\omega), \mathcal{K}^{-}(\omega) \ge 0. \tag{A.19}$$

Clearly, $\mathcal{K}^{+}[\omega]$ and $\mathcal{K}^{-}[\omega]$ can be assigned to an arbitrary walk ω, but regularity continues to be interpreted only for cycles.

Taking advantage of $\mathcal{W}_i^{\langle +1\rangle}$ and $\mathcal{W}_i^{\langle -1\rangle}$, Eq. (14) can be reformulated as two separate inequalities,

$$\mathcal{W}_i^{\langle +1\rangle} - \left[\eta_{\text{up}(i)} - \eta_{\text{low}(i)}\right] \ge 0 \tag{A.20}$$

and

$$\mathcal{W}_i^{\langle -1\rangle} + \eta_{\text{up}(i)} - \eta_{\text{low}(i)} \ge 0. \tag{A.21}$$

These relations can be unified *via* $s \in \{-1, 1\}$:

$$\mathcal{W}_i^{\langle s\rangle} - s\left[\eta_{\text{up}(i)} - \eta_{\text{low}(i)}\right] \ge 0. \tag{A.22}$$

Armed with the notions, symbols, and formulas introduced in Eqs. (A.1)–(A.22), one can prove the equivalence of the following two predicates: *A*) $\mathcal{N}$ is a consistent subnetwork of the SN, and *B*) all the cycles of $\mathcal{N}$ are regular.

I. Proof of $A \Rightarrow B$:

Consider a vector $\boldsymbol{\eta} = \{\eta_1, \eta_2, \ldots, \eta_{N_L}\}^{\mathrm{T}}$ satisfying Eq. (A.22) for all $1 \le i \le N_{\mathrm{T}}$ with $P_i = 1$ and pick a cycle ω of $\mathcal{N}$.

Case (a):

Due to the arbitrariness of the sign s, one can write

$$\sum_{q=1}^{L}\left\{\mathcal{W}_{I_q}^{\langle S_q\rangle} - S_q\left[\eta_{\text{up}(I_q)} - \eta_{\text{low}(I_q)}\right]\right\} \ge 0. \tag{A.23}$$

Expanding the left side of Eq. (A.23) results in

$$\sum_{q=1}^{L}\mathcal{W}_{I_q}^{\langle S_q\rangle} - \sum_{q=1}^{L}S_q\left[\eta_{\text{up}(I_q)} - \eta_{\text{low}(I_q)}\right] \ge 0. \tag{A.24}$$

Since the second term should be zero due to Eq. (A.8) and $J_{L+1} = J_1$, Eq. (A.24) can be simplified to

$$0 \le \sum_{q=1}^{L}\mathcal{W}_{I_q}^{\langle S_q\rangle} \equiv \mathcal{K}^{+}[\omega]. \tag{A.25}$$

Case (b): Replacing S_q with $-S_q$ in Eq. (A.23) also yields a nonnegative expression:

$$\sum_{q=1}^{L}\left\{\mathcal{W}_{I_q}^{\langle -S_q\rangle}+S_q\left[\eta_{\mathrm{up}(I_q)}-\eta_{\mathrm{low}(I_q)}\right]\right\}\geq 0, \tag{A.26}$$

from which

$$0\leq\sum_{q=1}^{L}\mathcal{W}_{I_q}^{\langle -S_q\rangle}\equiv\mathcal{K}^{-}[\omega] \tag{A.27}$$

can be deduced, in analogy with Eq. (A.24). Combination of Eqs. (A.25) and (A.27) leads to the consequence that ω should be a regular cycle; in other words, $A\Rightarrow B$.

II. Proof of $B\Rightarrow A$:

To show the converse statement, one needs to construct $K_j\rightsquigarrow j$ paths for all $1\leq j\leq N_{\mathrm{L}}$, where $K_j=\mathrm{core}(\mathrm{comp}(j))$ is the index of the core within the component of λ_j. Let $\omega^{\langle\langle j\rangle\rangle}$ illustrate a *cheapest* $K_j\rightsquigarrow j$ path, *i.e.*, for which $\mathcal{K}^{+}[\omega^{\langle\langle j\rangle\rangle}]$ is minimal. For the λ_{K_j} state, the cheapest path is ω^{∇}, which is of zero cost by definition. Now, take τ_i and examine the costs of $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i$ and $\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\oplus\tau_i$.

1. *Analysis of* $\mathcal{K}^{+}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i]$:

This problem can be split up into two cases, depending on whether $\lambda_{\mathrm{up}(i)}$ takes part in $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}$ ($\mathrm{up}(i)\in\mathbb{J}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}]$) or not ($\mathrm{up}(i)\not\in\mathbb{J}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}]$).

Case (a): $\mathrm{up}(i)\not\in\mathbb{J}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}]$

If state $\lambda_{\mathrm{up}(i)}$ is not included in $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}$, then $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i$ is another $K_{\mathrm{up}(i)}\rightsquigarrow\mathrm{up}(i)$ path, which must not be cheaper than $\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}$. Under these circumstances,

$$\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\right]\quad\leq\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i\right]=\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\right]+\mathcal{W}_i^{\langle +1\rangle}, \tag{A.28}$$

in which $\mathcal{W}_i^{\langle +1\rangle}$ is employed because the last state of $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}$ is $\lambda_{\mathrm{low}(i)}$, providing a +1 sign for τ_i within $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i$ and its cost. After a simplification,

$$\mathcal{W}_i^{\langle +1\rangle}-\mathcal{K}^{+}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}]+\mathcal{K}^{+}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}]\geq 0, \tag{A.29}$$

a relation quite similar to Eq. (A.20).

Case (b): $\mathrm{up}(i)\in\mathbb{J}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}]$

Assuming that $\lambda_{\mathrm{up}(i)}$ participates in $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}$, this path can be decomposed as follows:

$$\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}=\omega^{K_{\mathrm{up}(i)}\rightsquigarrow \mathrm{up}(i)}\odot\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}, \tag{A.30}$$

where $\omega^{K_{\mathrm{up}(i)}\rightsquigarrow \mathrm{up}(i)}$ and $\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}$ are $K_{\mathrm{up}(i)}\rightsquigarrow\mathrm{up}(i)$ and $\mathrm{up}(i)\rightsquigarrow\mathrm{low}(i)$ paths, respectively, and the trivial connection $K_{\mathrm{low}(i)}=K_{\mathrm{up}(i)}$ is taken into account. Applying Eq. (A.30), the cost of $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i$ can be formulated as

$$\begin{aligned}&\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i\right]=\\&\mathcal{K}^{+}\left[\omega^{K_{\mathrm{up}(i)}\rightsquigarrow \mathrm{up}(i)}\odot\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}\oplus\tau_i\right]=\\&\mathcal{K}^{+}\left[\omega^{K_{\mathrm{up}(i)}\rightsquigarrow \mathrm{up}(i)}\right]+\mathcal{K}^{+}\left[\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}\oplus\tau_i\right],\end{aligned} \tag{A.31}$$

where the additivity of the cost terms is taken into account together with the fact that the signs of the individual walks remain the same after their concatenation. The path $\omega^{K_{\mathrm{up}(i)}\rightsquigarrow \mathrm{up}(i)}$ in Eq. (A.31) is certainly not cheaper than $\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}$, yielding

$$\mathcal{K}^{+}[\omega^{K_{\mathrm{up}(i)}\rightsquigarrow \mathrm{up}(i)}]\geq\mathcal{K}^{+}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}]. \tag{A.32}$$

Dissecting the structure of the walk $\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}\oplus\tau_i$, two subcases can be envisioned.

Subcase #1:

If $i\in\mathbb{I}[\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}]$, the path $\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}$ includes solely the τ_i line (with a sign −1); otherwise, this path possesses repetitive states, which is not allowed. Suffixing $\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}$ with τ_i (*via* a sign +1), the cost of $\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}\oplus\tau_i$ turns into

$$\begin{aligned}\mathcal{K}^{+}\left[\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}\oplus\tau_i\right]&=\mathcal{W}_i^{\langle -1\rangle}+\mathcal{W}_i^{\langle +1\rangle}=\delta_i-\sigma_i+\delta_i+\sigma_i\\&=2\delta_i\geq 0.\end{aligned} \tag{A.33}$$

Subcase #2:

If $i\not\in\mathbb{I}[\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}]$, the walk $\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}\oplus\tau$ must be a cycle. Relying on the premise that all the cycles are regular within $\mathcal{N}$ [see predicate B)],

$$\mathcal{K}^{+}[\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}\oplus\tau_i]\geq 0 \tag{A.34}$$

is fulfilled, due to Eq. (A.19). Thus, $\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}\oplus\tau_i$ exhibits a nonnegative cost, regardless of whether τ_i takes part in $\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}$. Utilizing Eqs. (A.31)–(A.34), the cost of $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i$ can be estimated from below as

$$\begin{aligned}\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i\right]&=\mathcal{K}^{+}\left[\omega^{K_{\mathrm{up}(i)}\rightsquigarrow \mathrm{up}(i)}\right]+\mathcal{K}^{+}\left[\omega^{\mathrm{up}(i)\rightsquigarrow \mathrm{low}(i)}\oplus\tau_i\right]\\&\geq\mathcal{K}^{+}\left[\omega^{K_{\mathrm{up}(i)}\rightsquigarrow \mathrm{up}(i)}\right]\geq\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\right].\end{aligned} \tag{A.35}$$

Since the sign of τ_i is +1 in $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i$, Eq. (A.35) becomes

$$\begin{aligned}\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i\right]&=\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\right]+\mathcal{W}_i^{\langle +1\rangle}\\&\geq\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\right],\end{aligned} \tag{A.36}$$

from which the following inequality is deduced:

$$\mathcal{W}_i^{\langle +1\rangle}-\mathcal{K}^{+}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}]+\mathcal{K}^{+}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}]\geq 0, \tag{A.37}$$

which is identical to the relation given in Eq. (A.29).

In short, Eq. (A.37) should hold for τ_i, irrespective of whether $\lambda_{\mathrm{up}(i)}$ lies on the $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}$ path or not.

2. *Analysis of* $\mathcal{K}^{+}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\oplus\tau_i]$:

The treatment applied here is analogous to that used for the investigation of $\mathcal{K}^{+}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\oplus\tau_i]$.

Case (a): $\mathrm{low}(i)\not\in\mathbb{J}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}]$

Under this condition, $\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\oplus\tau_i$ is an alternative $K_{\mathrm{low}(i)}\rightsquigarrow\mathrm{low}(i)$ path, which is by no means cheaper than $\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}$:

$$\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\right]\leq\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\oplus\tau_i\right]=\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\right]+\mathcal{W}_i^{\langle -1\rangle}, \tag{A.38}$$

where $\mathcal{W}_i^{\langle -1\rangle}$ is entered as τ_i should be attached to $\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}$ *via* $\lambda_{\mathrm{up}(i)}$, requiring a sign of −1 for τ_i. After a trivial rearrangement,

$$\mathcal{W}_i^{\langle -1\rangle}+\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\right]-\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\right]\geq 0, \tag{A.39}$$

showing high resemblance to Eq. (A.21).

Case (b): $\mathrm{low}(i)\in\mathbb{J}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}]$

With this assumption, $\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}$ can be partitioned as follows:

$$\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}=\omega^{K_{\mathrm{low}(i)}\rightsquigarrow \mathrm{low}(i)}\odot\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)}, \tag{A.40}$$

where $\omega^{K_{\mathrm{low}(i)}\rightsquigarrow \mathrm{low}(i)}$ and $\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)}$ denote two paths of types $K_{\mathrm{low}(i)}\rightsquigarrow\mathrm{low}(i)$ and $\mathrm{low}(i)\rightsquigarrow\mathrm{up}(i)$, respectively, and $K_{\mathrm{up}(i)}=K_{\mathrm{low}(i)}$ is considered. In view of Eq. (A.40), the expression of $\mathcal{K}^{+}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\oplus\tau_i]$ can be fragmented as follows:

$$\begin{aligned}&\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\oplus\tau_i\right]=\mathcal{K}^{+}\left[\omega^{K_{\mathrm{low}(i)}\rightsquigarrow \mathrm{low}(i)}\odot\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)}\oplus\tau_i\right]=\\&\mathcal{K}^{+}\left[\omega^{K_{\mathrm{low}(i)}\rightsquigarrow \mathrm{low}(i)}\right]+\mathcal{K}^{+}\left[\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)}\oplus\tau_i\right].\end{aligned} \tag{A.41}$$

where

$$\mathcal{K}^{+}[\omega^{K_{\mathrm{low}(i)}\rightsquigarrow \mathrm{low}(i)}] \geq \mathcal{K}^{+}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}], \tag{A.42}$$

and $\mathcal{K}^{+}[\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)} \oplus \tau_i] \geq 0$ is verified below.

Subcase #1:

Presume that $i \in \mathbb{I}[\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)}]$, $\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)}$ is a path, whose single line, τ_i, is provided with a sign of +1. Then, τ_i can be appended to $\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)}$ with a sign of –1, producing the following cost:

$$\mathcal{K}^{+}[\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)} \oplus \tau_i] = \mathcal{W}_i^{\langle +1\rangle} + \mathcal{W}_i^{\langle -1\rangle} \geq 0. \tag{A.43}$$

Subcase #2:

The opposite condition, the relation $i \not\in \mathbb{I}[\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)}]$, implies that $\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)} \oplus \tau$ is a cycle. Demanding the regularity of the cycles within $\mathcal{N}$ [see predicate B)],

$$\mathcal{K}^{+}[\omega^{\mathrm{low}(i)\rightsquigarrow \mathrm{up}(i)} \oplus \tau_i] \geq 0 \tag{A.44}$$

is obtained, after applying Eq. (A.19). Converting Eqs. (A.41)–(A.44) into a single relation, Eq. (A.41) can be rewritten as

$$\mathcal{K}^{+}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle} \oplus \tau_i] \geq \mathcal{K}^{+}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}]. \tag{A.45}$$

Owing to the fact that τ_i is coupled to $\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}$ with a sign of –1 in $\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle} \oplus \tau_i$, Eq. (A.45) becomes

$$\begin{aligned}\mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle} \oplus \tau_i\right] &= \mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}\right] + \mathcal{W}_i^{\langle -1\rangle} \\ &\geq \mathcal{K}^{+}\left[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}\right],\end{aligned} \tag{A.46}$$

which can be rearranged to

$$\mathcal{W}_i^{\langle -1\rangle} + \mathcal{K}^{+}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}] - \mathcal{K}^{+}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}] \geq 0. \tag{A.47}$$

Hence, $\mathrm{low}(i) \not\in \mathbb{J}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}]$ and $\mathrm{low}(i) \in \mathbb{J}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}]$ lead to the same inequality [see Eqs. (A.37) and (A.47)].

In conclusion, it is ascertained that the regularity of the cycles within $\mathcal{N}$ involves

$$\mathcal{W}_i^{\langle s\rangle} - s\left\{\mathcal{K}^{+}[\omega^{\langle\langle \mathrm{up}(i)\rangle\rangle}] - \mathcal{K}^{+}[\omega^{\langle\langle \mathrm{low}(i)\rangle\rangle}]\right\} \geq 0 \tag{A.48}$$

for any $s = \pm 1$ and each $1 \leq i \leq N_{\mathrm{T}}$ with $P_i = 1$. Therefore, by selecting $\eta_j = \mathcal{K}^{+}[\omega^{\langle\langle j\rangle\rangle}]$ for all $1 \leq j \leq N_{\mathrm{L}}$, Eq. (A.48) translates to Eq. (A.22), manifesting the consistency of $\mathcal{N}$ and thereby attesting $B \Rightarrow A$.

Appendix B. Five feasible misconceptions (M1–M5) about consistency types

Fallacies hidden in the network-theory-based definitions of consistency types are collected here and refuted with counterexamples. These exemplary SNs of minimal size are depicted in Fig. A.2.

M1. Consistency implies strong consistency (false).

If this claim was true, then a SN comprising a single regular cycle should be strongly consistent, like that shown in Fig. A.2(a). One can observe that the cycle of Fig. A.2(a) is indeed regular, as its discrepancy,

$$\begin{aligned}\mathcal{D} &= |\sigma_1 + \sigma_2 - \sigma_3 - \sigma_4| \\ &= |1.0000 + 1.0000 - 1.0000 - 1.0100| \\ &= 0.0100\ \mathrm{cm}^{-1}\end{aligned} \tag{B.1}$$

is smaller than its threshold,

$$\begin{aligned}\mathcal{T} &= \delta_1 + \delta_2 + \delta_3 + \delta_4 \\ &= 0.0001 + 0.0025 + 0.0025 + 0.0050 = 0.0101\ \mathrm{cm}^{-1}.\end{aligned} \tag{B.2}$$

Now, let $\mathcal{N}$ denote the example SN itself and calculate the empirical energies *via* Eqs. (6)–(7). To this end, one needs to set up **P**, **W**, $\boldsymbol{\sigma}$, and **R** (by neglecting trailing zeros) as follows:

$$\mathbf{P} = \mathrm{diag}(1, 1, 1, 1), \tag{B.3}$$

$$\begin{aligned}\mathbf{W} &= \mathrm{diag}\left(\delta_1^{-2}, \delta_2^{-2}, \delta_3^{-2}, \delta_4^{-2}\right) \\ &= \mathrm{diag}\left(10^8, 1.6 \times 10^5, 1.6 \times 10^5, 4 \times 10^4\right)\ \mathrm{cm}^2,\end{aligned} \tag{B.4}$$

$$\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}^{\mathrm{T}} = \{1, 1, 1, 1.01\}^{\mathrm{T}}, \tag{B.5}$$

$$\mathrm{R} = \begin{array}{c|cccc} & \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \\ \hline \tau_1 & -1 & 1 & 0 & 0 \\ \tau_2 & 0 & -1 & 1 & 0 \\ \tau_3 & 0 & 0 & 1 & -1 \\ \tau_4 & -1 & 0 & 0 & 1 \end{array} \tag{B.6}$$

From these arrays, one can build the **G** matrix and the **F** vector of Eq. (7) in the following way:

$$\begin{aligned}\mathbf{G} &= \mathbf{R}^{\mathrm{T}}\mathbf{P}\mathbf{W}\mathbf{R} \\ &= \begin{pmatrix} 1.0004 \times 10^8 & -1 \times 10^8 & 0 & -4 \times 10^4 \\ -1 \times 10^8 & 1.0016 \times 10^8 & -1.6 \times 10^5 & 0 \\ 0 & -1.6 \times 10^5 & 3.2 \times 10^5 & -1.6 \times 10^5 \\ -4 \times 10^4 & 0 & -1.6 \times 10^5 & 2 \times 10^5 \end{pmatrix},\end{aligned} \tag{B.7}$$

and

$$\mathbf{F} = \begin{pmatrix} -1.0004 \times 10^8 \\ 9.984 \times 10^7 \\ 3.2 \times 10^5 \\ -1.196 \times 10^5 \end{pmatrix}. \tag{B.8}$$

As the sum of the rows in **G** is a zero vector, **G** is a singular matrix. This means that Eq. (7) is underdetermined in itself. By replacing the first row/column pair of the **G** matrix with that of the 4 × 4 identity matrix, a nonsingular $\mathbf{G}_1$ matrix is obtained:

$$\mathbf{G}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1.0016 \times 10^8 & -1.6 \times 10^5 & 0 \\ 0 & -1.6 \times 10^5 & 3.2 \times 10^5 & -1.6 \times 10^5 \\ 0 & 0 & -1.6 \times 10^5 & 2 \times 10^5 \end{pmatrix}. \tag{B.9}$$

Additionally, by substituting the first entry of **F** with zero, another auxiliary array can be introduced:

$$\mathbf{F}_1 = \begin{pmatrix} 0 \\ 9.984 \times 10^7 \\ 3.2 \times 10^5 \\ -1.196 \times 10^5 \end{pmatrix}. \tag{B.10}$$

With the aid of $\mathbf{G}_1$ and $\mathbf{F}_1$, the vector of empirical energies can be expressed as

$$\bar{\mathbf{E}} = \mathbf{G}_1^{-1}\mathbf{F}_1 \approx \begin{pmatrix} 0 \\ 1.000\,003 \\ 2.001\,669 \\ 1.003\,335 \end{pmatrix}, \tag{B.11}$$

where $\mathbf{G}_1^{-1}$ is the inverse $\mathbf{G}_1$ matrix, and the entries of $\bar{\mathbf{E}}$ are given in cm^{-1}. the $d_i = |\Delta_i| - \delta_i = |\sigma_i - \bar{E}_{\mathrm{up}(i)} + \bar{E}_{\mathrm{low}(i)}| - \delta_i$ relation ($1 \leq i \leq 4$) leads to the following defects:

$$\begin{aligned} d_1 &= |1 - 1.000\,003 + 0| - 1 \times 10^{-4} \\ &\approx -9.7 \times 10^{-5}\ \mathrm{cm}^{-1}, \\ d_2 &= |1 - 2.001\,669 + 1.000\,003| - 2.5 \times 10^{-3} \\ &\approx -8.34 \times 10^{-4}\ \mathrm{cm}^{-1}, \\ d_3 &= |1 - 2.001\,669 + 1.003\,335| - 2.5 \times 10^{-3} \\ &\approx -8.34 \times 10^{-4}\ \mathrm{cm}^{-1}, \\ d_4 &= |1.01 - 1.003\,335 + 0| - 5 \times 10^{-3} \\ &\approx +\mathbf{1.665} \times \mathbf{10^{-3}}\ \mathrm{cm}^{\mathbf{-1}}. \end{aligned} \tag{B.12}$$

As the d_4 value is positive, Eq. (18) is violated, suggesting that $\mathcal{N}$ is not strongly consistent. Therefore, consistency does not imply strong consistency.

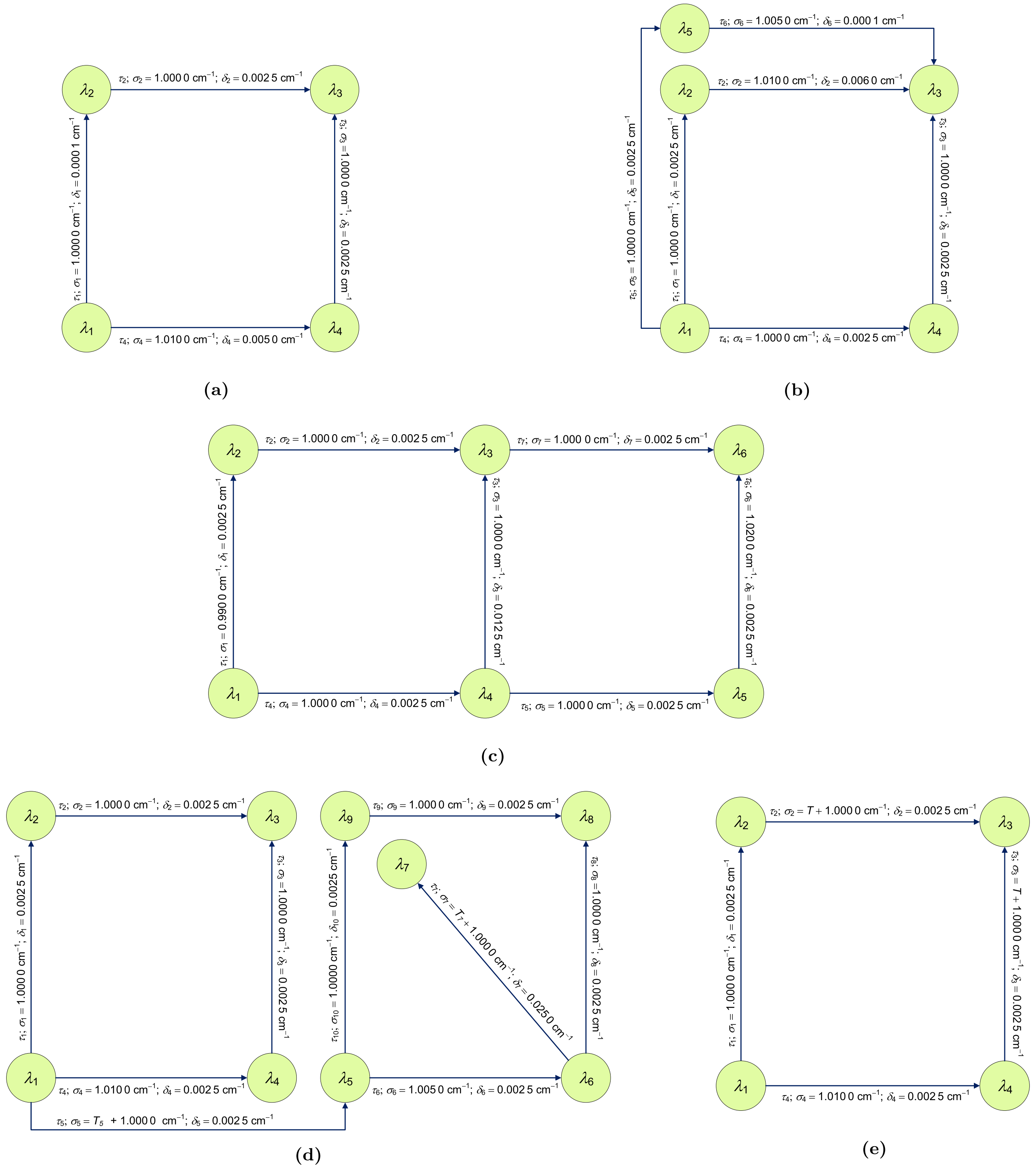

λ_2
τ_2; $\sigma_2 = 1.0000$ cm^{-1}; $\delta_2 = 0.0025$ cm^{-1}
λ_3
τ_1; $\sigma_1 = 1.0000$ cm^{-1}; $\delta_1 = 0.0001$ cm^{-1}
τ_3; $\sigma_3 = 1.0000$ cm^{-1}; $\delta_3 = 0.0025$ cm^{-1}
λ_1
τ_4; $\sigma_4 = 1.0100$ cm^{-1}; $\delta_4 = 0.0050$ cm^{-1}
λ_4
(a)
λ_5
τ_6; $\sigma_6 = 1.0050$ cm^{-1}; $\delta_6 = 0.0001$ cm^{-1}
λ_2
τ_2; $\sigma_2 = 1.0100$ cm^{-1}; $\delta_2 = 0.0060$ cm^{-1}
λ_3
τ_5; $\sigma_5 = 1.0000$ cm^{-1}; $\delta_5 = 0.0025$ cm^{-1}
τ_1; $\sigma_1 = 1.0000$ cm^{-1}; $\delta_1 = 0.0025$ cm^{-1}
τ_3; $\sigma_3 = 1.0000$ cm^{-1}; $\delta_3 = 0.0025$ cm^{-1}
λ_1
τ_4; $\sigma_4 = 1.0000$ cm^{-1}; $\delta_4 = 0.0025$ cm^{-1}
λ_4
(b)
λ_2
τ_2; $\sigma_2 = 1.0000$ cm^{-1}; $\delta_2 = 0.0025$ cm^{-1}
λ_3
τ_7; $\sigma_7 = 1.0000$ cm^{-1}; $\delta_7 = 0.0025$ cm^{-1}
λ_6
τ_1; $\sigma_1 = 0.9900$ cm^{-1}; $\delta_1 = 0.0025$ cm^{-1}
τ_3; $\sigma_3 = 1.0000$ cm^{-1}; $\delta_3 = 0.0125$ cm^{-1}
τ_6; $\sigma_6 = 1.0200$ cm^{-1}; $\delta_6 = 0.0025$ cm^{-1}
λ_1
τ_4; $\sigma_4 = 1.0000$ cm^{-1}; $\delta_4 = 0.0025$ cm^{-1}
λ_4
τ_5; $\sigma_5 = 1.0000$ cm^{-1}; $\delta_5 = 0.0025$ cm^{-1}
λ_5
(c)
λ_2
τ_2; $\sigma_2 = 1.0000$ cm^{-1}; $\delta_2 = 0.0025$ cm^{-1}
λ_3
λ_9
τ_9; $\sigma_9 = 1.0000$ cm^{-1}; $\delta_9 = 0.0025$ cm^{-1}
λ_8
λ_7
τ_1; $\sigma_1 = 1.0000$ cm^{-1}; $\delta_1 = 0.0025$ cm^{-1}
τ_3; $\sigma_3 = 1.0000$ cm^{-1}; $\delta_3 = 0.0025$ cm^{-1}
τ_{10}; $\sigma_{10} = 1.0000$ cm^{-1}; $\delta_{10} = 0.0025$ cm^{-1}
τ_7; $\sigma_7 = T_7 + 1.0000$ cm^{-1}; $\delta_7 = 0.0250$ cm^{-1}
τ_8; $\sigma_8 = 1.0000$ cm^{-1}; $\delta_8 = 0.0025$ cm^{-1}
λ_1
τ_4; $\sigma_4 = 1.0100$ cm^{-1}; $\delta_4 = 0.0025$ cm^{-1}
λ_4
λ_5
τ_6; $\sigma_6 = 1.0050$ cm^{-1}; $\delta_6 = 0.0025$ cm^{-1}
λ_6
τ_5; $\sigma_5 = T_5 + 1.0000$ cm^{-1}; $\delta_5 = 0.0025$ cm^{-1}
(d)
λ_2
τ_2; $\sigma_2 = T + 1.0000$ cm^{-1}; $\delta_2 = 0.0025$ cm^{-1}
λ_3
τ_1; $\sigma_1 = 1.0000$ cm^{-1}; $\delta_1 = 0.0025$ cm^{-1}
τ_3; $\sigma_3 = T + 1.0000$ cm^{-1}; $\delta_3 = 0.0025$ cm^{-1}
λ_1
τ_4; $\sigma_4 = 1.0100$ cm^{-1}; $\delta_4 = 0.0025$ cm^{-1}
λ_4
(e)

Fig. A.2. Counterexamples for possible misconceptions detailed in Appendix B.

M2. Strong consistency is balanced (false).

To disprove this misconception, take Fig. A.2(b) as an example. Initializing $\mathcal{N}$ as the entire SN and applying the strategy shown in Eqs. (B.3)–(B.12), one can derive the following empirical energies:

$$\begin{aligned}
\bar{E}_1 &\approx 0\ \text{cm}^{-1}, \\
\bar{E}_2 &\approx 0.999\,102\ \text{cm}^{-1}, \\
\bar{E}_3 &\approx 2.003\,931\ \text{cm}^{-1}, \\
\bar{E}_4 &\approx 1.001\,965\ \text{cm}^{-1}, \\
\bar{E}_5 &\approx 0.998\,932\ \text{cm}^{-1},
\end{aligned} \tag{B.13}$$

and defects:

$$\begin{aligned}
d_1 &\approx -1.602 \times 10^{-3}\ \text{cm}^{-1}, \\
d_2 &\approx -8.29 \times 10^{-4}\ \text{cm}^{-1}, \\
d_3 &\approx -5.35 \times 10^{-4}\ \text{cm}^{-1}, \\
d_4 &\approx -5.35 \times 10^{-4}\ \text{cm}^{-1}, \\
d_5 &\approx -1.432 \times 10^{-3}\ \text{cm}^{-1}, \\
d_6 &\approx -9.8 \times 10^{-5}\ \text{cm}^{-1},
\end{aligned} \tag{B.14}$$

signaling that $\mathcal{N}$ is strongly consistent. If τ_6 is left out from $\mathcal{N}$ by resetting $P_6 = 0$ in $\mathbf{P}$, the quantities of Eqs. (B.13) and (B.14) are mutated into

$$\begin{aligned}&\bar{E}_1 \approx 0\ \mathrm{cm}^{-1},\\&\bar{E}_2 \approx 0.998\,858\ \mathrm{cm}^{-1},\\&\bar{E}_3 \approx 2.002\,283\ \mathrm{cm}^{-1},\\&\bar{E}_4 \approx 1.001\,142\ \mathrm{cm}^{-1},\\&\bar{E}_5 \approx 1.000\,000\ \mathrm{cm}^{-1},\end{aligned} \tag{B.15}$$

and

$$\begin{aligned}&d_1 \approx -1.358 \times 10^{-3}\ \mathrm{cm}^{-1},\\&d_2 \approx +\mathbf{5.75} \times \mathbf{10^{-4}}\ \mathrm{cm}^{-\mathbf{1}},\\&d_3 \approx -1.358 \times 10^{-3}\ \mathrm{cm}^{-1},\\&d_4 \approx -1.358 \times 10^{-3}\ \mathrm{cm}^{-1},\\&d_5 \approx -2.500 \times 10^{-3}\ \mathrm{cm}^{-1}.\end{aligned} \tag{B.16}$$

Hence, a subnetwork of this strongly consistent SN is not strongly consistent, sharply contradicting the assumed balanced nature of strong consistency.

M3. If a cycle basis of a SN is consistent, then the SN itself should be also consistent (false).

Unfortunately, this is also a claim which is 'too good to be true'. Upon the analysis of Fig. A.2(c), one can realize that the two 4-membered cycles $\chi_1 = \tau_1 - \tau_2 - \tau_3 - \tau_4$ and $\chi_2 = \tau_3 - \tau_5 - \tau_6 - \tau_7$ are regular, while the cycle of length 6, $\chi_3 = \tau_1 - \tau_2 - \tau_7 - \tau_6 - \tau_5 - \tau_4$, is not regular:

$$\begin{aligned}&\mathcal{D}(\chi_1) = 0.01\ \mathrm{cm}^{-1} \le \mathcal{T}(\chi_1) = 0.02\ \mathrm{cm}^{-1},\\&\mathcal{D}(\chi_2) = 0.02\ \mathrm{cm}^{-1} \le \mathcal{T}(\chi_2) = 0.02\ \mathrm{cm}^{-1},\\&\mathcal{D}(\chi_3) = \mathbf{0.03}\ \mathrm{cm}^{-1} > \mathcal{T}(\chi_3) = 0.015\ \mathrm{cm}^{-1}.\end{aligned} \tag{B.17}$$

As long as $\{\chi_1, \chi_2\}$ is selected as a cycle basis of the SN, this cycle basis, made up of only regular cycles, is consistent, while the SN itself is not consistent due to the non-regular cycle χ_3, contradicting claim M3.

M4. A consistent SN has no outliers (false).

Pick a SN with bridges, like that of Fig. A.2(d). This SN holds two 4-cycles ($\chi_1 = \tau_1 - \tau_2 - \tau_3 - \tau_4$ and $\chi_2 = \tau_6 - \tau_8 - \tau_9 - \tau_{10}$) with the following discrepancies and thresholds:

$$\begin{aligned}&\mathcal{D}(\chi_1) = 0.01\ \mathrm{cm}^{-1} \le \mathcal{T}(\chi_1) = 0.01\ \mathrm{cm}^{-1},\\&\mathcal{D}(\chi_2) = 0.005\ \mathrm{cm}^{-1} \le \mathcal{T}(\chi_2) = 0.01\ \mathrm{cm}^{-1}.\end{aligned} \tag{B.18}$$

These inequalities suggest that all the cycles within this SN are regular, certifying its consistency. The consistency of this SN can be destructed only by inflating $\mathcal{D}(\chi_1)$ or $\mathcal{D}(\chi_2)$ over $\mathcal{T}(\chi_1)$ or $\mathcal{T}(\chi_2)$, respectively. As the discrepancies are not influenced by σ_5 or σ_7, these bridge wavenumbers can be shifted with the two (arbitrary) transcription errors, T_5 and T_7, without affecting the consistency of the SN. Hence, a consistent SN may carry bridge outliers.

M5. A bridgeless consistent SN cannot contain out liers (false).

It is easy to show that the absence of bridges in consistent SNs is not a warranty for the lack of outliers. Fig. A.2(e) exhibits an example SN, corresponding to a 4-membered cycle, where σ_2 and σ_3 suffer from the same transcription error T. The discrepancy and the threshold of this χ cycle are as follows:

$$\mathcal{D}(\chi) = 0.01\ \mathrm{cm}^{-1} \le \mathcal{T}(\chi) = 0.01\ \mathrm{cm}^{-1}, \tag{B.19}$$

independently of the actual value of T. It seems that the presence of outliers τ_2 and τ_3 in this SN does not collide with its consistency, disproving that bridgeless SNs are free of outliers. Of course, one can select different transcription errors for τ_2 and τ_3 such that the regularity of χ is retained: the single criterion is that these errors should be close to each other to an extent not permitting $\mathcal{D}(\chi)$ to grow above $\mathcal{T}(\chi)$.

Appendix C. The MILP formalism

As mentioned in Sec. 3, one of our attempts has been to build a mixed integer linear programming (MILP) model for the automated detection of outliers in SNs. Here a concise summary is given on how to treat the outlier problem within the MILP approach.

The idea behind a MILP-based approximation of outlier detection is that one can find an

$$\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_{N_\mathrm{L}}\}^\mathrm{T} \tag{C.1}$$

potential and a

$$\mathcal{P} = \mathrm{diag}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{N_\mathrm{T}}) \tag{C.2}$$

participation matrix for the SN which satisfy

$$\sigma_i - \mathcal{E}_{\mathrm{up}(i)} + \mathcal{E}_{\mathrm{low}(i)} \le \delta_{\mathrm{r},i} + (1 - \mathcal{P}_i)\phi_\mathrm{pen}, \tag{C.3}$$

$$-\sigma_i + \mathcal{E}_{\mathrm{up}(i)} - \mathcal{E}_{\mathrm{low}(i)} \le \delta_{\mathrm{r},i} + (1 - \mathcal{P}_i)\phi_\mathrm{pen}, \tag{C.4}$$

and

$$\mathcal{P}_i \le P_{\mathrm{LS},i} \tag{C.5}$$

for each $1 \le i \le N_\mathrm{T}$, where ϕ_pen denotes a sufficiently huge *penalty factor* and $P_{\mathrm{LS},i}$ is the ith diagonal entry in the participation matrix of $\mathcal{N}_\mathrm{LS}$. In the case of $\mathcal{P}_i = 1$, Eqs. (C.3) and (C.4) turn into

$$|\sigma_i - \mathcal{E}_{\mathrm{up}(i)} + \mathcal{E}_{\mathrm{low}(i)}| \le \delta_{\mathrm{r},i}, \tag{C.6}$$

implying that the subnetwork determined by $\mathcal{P}$ must be quasi-consistent. For $\mathcal{P}_i = 0$, $(1 - \mathcal{P}_i)\phi_\mathrm{pen}$ warrants that the left sides of Eqs. (C.3) and (C.4) can be confined below an upper limit. Furthermore, Eq. (C.5) prescribes that solely those quasi-consistent subnetworks should be considered which do not contain the excluded lines of $\mathcal{N}_\mathrm{LS}$.

Since a $(\mathcal{E}, \mathcal{P})$ pair adhering to Eqs. (C.3)–(C.5) always exists for a suitably large ϕ_pen, one can choose a $(\mathcal{E}, \mathcal{P}) = (\tilde{\mathcal{E}}, \tilde{\mathcal{P}})$ pair as to maximize the function

$$O(\mathcal{P}) = \sum_{i=1}^{N_\mathrm{T}} \varpi_i \mathcal{P}_i, \tag{C.7}$$

subjected to Eqs. (C.3)–(C.5), where ϖ_i is a weight factor for line τ_i (selected as $\varpi_i = 1$ or $\varpi_i = \delta_i^{-2}$). This maximizing condition forces most of the transitions with the largest possible weights to be inserted in the subnetwork $\tilde{\mathcal{N}}$ specified by $\tilde{\mathcal{P}}$. Within the scheme established by Eqs (C3)–(C5) and (C7), $\tilde{\mathcal{N}}$ can be regarded as the optimal quasi-consistent subnetwork of $\mathcal{N}_\mathrm{LS}$, while $\mathcal{N}_\mathrm{LS} \setminus \tilde{\mathcal{N}}$ gives the BHO for the SN.

Despite the simplicity of the MILP model in Eqs. (C.3)–(C.5) and (C.7), which can be solved, *e.g., via* the *branch and cut* protocol [65] incorporated into the GLPK package [66], the solution of this model suffers from numerical issues. The reason behind the observed instability is that a ϕ_pen value too large, meeting the criterion

$$\phi_\mathrm{pen} > \sum_{i=1}^{N_\mathrm{T}} P_{\mathrm{LS},i}\sigma_i, \tag{C.8}$$

should be adopted to obtain a generally solvable model, making the resultant inequalities rather ill-conditioned. This technical difficulty indicates that the $\tilde{\mathcal{P}}_i$ parameters, handled as floating-point numbers which are rounded to integers, might become inaccurate, failing to provide a correct and optimal quasi-consistent subnetwork $\tilde{\mathcal{N}}$ for the SN. Consequently, the MILP-based outlier scheme of Eqs. (C.3)–(C.5) and (C.7), in its current form, is not capable of yielding a trustworthy BHO for practical SNs containing hundreds of thousands of transitions.

References

[1] Allen MG. Diode laser absorption sensors for gas-dynamic and combustion flows. Meas Sci Technol 1998;9:545–62.

[2] Webber ME, Kim S, Sanders ST, Baer DS, Hanson RK, Ikeda Y. In situ combustion measurements of CO_2 by use of a distributed-feedback diode-laser sensor near 2.0 μm. Appl Opt 2001;40:821–8.

[3] Ramanathan V, Vogelmann aM. Greenhouse effect, atmospheric solar absorption and the Earth's radiation budget: from the Arrhenius-Langley era to the 1990s. Ambio 1997;26:38–46.

[4] Bernath PF. The spectroscopy of water vapour: experiment, theory and applications. Phys Chem Chem Phys 2002;4:1501–9.

[5] Ashfold MN, May PW, Petherbridge JR, Rosser KN, Smith JA, Mankelevich YA, et al. Unravelling aspects of the gas phase chemistry involved in diamond chemical vapour deposition. Phys Chem Chem Phys 2001;3:3471–85.

[6] Pollack JB, Dalton JB, Grinspoon D, Wattson RB, Freedman R, Crisp D, et al. Near-infrared light from Venus' nightside: a spectroscopic analysis. Icarus 1993;103:1–42.

[7] Polyansky OL, Császár AG, Shirin SV, Zobov NF, Barletta P, Tennyson J, et al. High-accuracy *ab initio* rotation-vibration transitions for water. Science 2003;299:539–42.

[8] Voronin B, Tennyson J, Tolchenov R, Lugovskoy A, Yurchenko S. A high accuracy computed line list for the HDO molecule. Mon Not R Astron Soc 2010;402:492–6.

[9] Tennyson J, Yurchenko SN. ExoMol: molecular line lists for exoplanet and other atmospheres. Mon Not R Astron Soc 2012;425:21–33.

[10] Martin JML, François J-P, Gijbels R. First principles computation of thermochemical properties beyond the harmonic approximation. I. Method and application to the water molecule and its isotopomers. J Chem Phys 1992;96:7633–45.

[11] Vidler M, Tennyson J. Accurate partition function and thermodynamic data for water. J Chem Phys 2000;113:9766–71.

[12] Aarset K, Császár AG, Sibert E, Allen WD, Schaefer III HF, Klopper W, et al. Anharmonic force field, vibrational energy levels, and barrier to inversion of SiH_3^-. J Chem Phys 2000;112:4053–63.

[13] Wenger C, Champion J-P, Boudon V. The partition sum of methane at high temperature. J Quant Spectrosc Rad Transf 2008;109:2697–706.

[14] 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 S 2016;224:44.

[15] Furtenbacher T, Szidarovszky T, Hruby 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.

[16] Simkó I, Furtenbacher T, Dénes N, Szidarovszky T, Hrubý J, Zobov NF, et al. Recommended ideal-gas thermochemical functions for heavy water and its substituent isotopologues. J Phys Chem Ref Data 2017;46:023104.

[17] Gamache RR, Roller C, Lopes E, Gordon IE, Rothman LS, Polyansky OL, et al. Total internal partition sums for 166 isotopologues of 51 molecules important in planetary atmospheres: Application to HITRAN 2016 and beyond. J Quant Spectrosc Rad Transf 2017;203:70–87.

[18] Merkt F, Quack M. Molecular quantum mechanics and molecular spectra, molecular symmetry, and interaction of matter with radiation. In: Quack M, Merkt F, editors. Handbook of high-resolution spectroscopy. Wiley, Chichester; 2011. p. 1–55.

[19] Gordon IE, Rothman LS, Hill C, Kochanov RV, Tan Y, Bernath PF, et al. The HITRAN 2016 molecular spectroscopic database. J Quant Spectrosc Rad Transfer 2017;203:3–69.

[20] 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.

[21] Endres CP, Schlemmer S, Schilke P, Stutzki J, Müller HS. The Cologne database for molecular spectroscopy, CDMS, in the virtual atomic and molecular data centre, VAMDC. J Mol Spectrosc 2016;327:95–104.

[22] Császár AG, Furtenbacher T, Árendás P. Small molecules – big data. J Phys Chem A 2016;120:8949–69.

[23] Watson JKG. The use of term-value fits in testing spectroscopic assignments. J Mol Spectrosc 1994;165:283–90.

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

[25] Furtenbacher T, Császár AG. The role of intensities in determining characteristics of spectroscopic networks. J Mol Spectrosc 2012;1009:123–9.

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

[27] Árendás P, Furtenbacher T, Császár AG. On spectra of spectra. J Math Chem 2016;54:806–22.

[28] Tóbiás R, Furtenbacher T, Császár AG. Cycle bases to the rescue. J Quant Spectrosc Radiat Transf 2017;203:557–64.

[29] Császár AG, Czakó G, Furtenbacher T, Mátyus E. An active database approach to complete spectra of small molecules. Annu Rep Comput Chem 2007;3:155–76.

[30] Furtenbacher T, Császár AG, Tennyson J. MARVEL: measured active rotational-vibrational energy levels. J Mol Spectrosc 2007;245:115–25.

[31] Tennyson J, Bernath PF, Brown LR, Campargue A, Carleer MR, Császár AG, et al. IUPAC 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;110:573–96.

[32] Furtenbacher T, Császár AG. MARVEL: measured active rotational-vibrational energy levels. II. Algorithmic improvements. J Quant Spectrosc Radiat Transf 2012;113:929–35.

[33] Tóbiás R, Furtenbacher T, Tennyson J, Császár AG. Accurate empirical rovibrational energies and transitions of $H_2{}^{16}O$. Phys Chem Chem Phys 2019;21:3473–95.

[34] 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;109:1234–51.

[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 II. Energy levels and transition wavenumbers for $HD^{16}O$, $HD^{17}O$, and $HD^{18}O$. J Quant Spectrosc Radiat Transf 2010;110:2160–84.

[36] Fábri C, Mátyus E, Furtenbacher T, Nemes L, Mihály B, Zoltáni T, et al. Variational quantum mechanical and active database approaches to the rotational-vibrational spectroscopy of ketene, H_2CCO. J Chem Phys 2011;135:094307.

[37] 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 Radiat Transf 2013;117:29–80.

[38] Furtenbacher T, Szidarovszky T, Fábri C, Császár AG. MARVEL analysis of the rotational-vibrational states of the molecular Ions H_2D^+ and D_2H^+. Phys Chem Chem Phys 2013;15:10181–93.

[39] Furtenbacher T, Szidarovszky T, Mátyus E, Fábri C, Császár AG. Analysis of the rotational-vibrational states of the molecular ion H_3^+. J Chem Theor Comput 2013;9:5471–8.

[40] 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;117:93–108.

[41] 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;161:117–30.

[42] 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 S 2017;228:15.

[43] Chubb KL, Joseph M, Franklin J, Choudhury N, Furtenbacher T, Császár AG, et al. MARVEL analysis of the measured high-resolution spectra of C_2H_2. J Quant Spectrosc Radiat Transf 2018;204:42–55.

[44] Tóbiás R, Furtenbacher T, Császár AG, Naumenko OV, Tennyson J, Flaud J-M, et al. Critical evaluation of measured rotational-vibrational transitions of four sulphur isotopologues of $S^{16}O_2$. J Quant Spectrosc Radiat Transf 2018;208:152–63.

[45] Chubb KL, Naumenko O, Keely S, Bartolotto S, MacDonald S, Mukhtar M, et al. MARVEL analysis of the measured high-resolution rovibrational spectra of $H_2{}^{32}S$. J Quant Spectrosc Radiat Transf 2018;218:178–86.

[46] McKemmish LK, Borsovszky J, Goodhew KL, Sheppard S, Bennett AFV, Martin ADJ, et al. MARVEL analysis of the measured high-resolution rovibronic spectra of $^{90}Zr^{16}O$. Astrophys J 2018;867:33.

[47] Furtenbacher T, Horváth M, Koller D, Sólyom P, Balogh A, Balogh I, et al. MARVEL analysis of the measured high-resolution rovibronic spectra and definitive ideal-gas thermochemistry of the $^{16}O_2$ molecule. J Phys Chem Ref Data 2019;48:023101.

[48] Darby-Lewis D, Shah H, Joshi D, Khan F, Kauwo M, Sethi N, et al. MARVEL analysis of the measured high-resolution spectra of ^{14}NH. J Mol Spectrosc 2019;362:69–76.

[49] McKemmish LK, Syme A-M, Borsovszky J, Yurchenko SN, Tennyson J, Furtenbacher T, et al. An update to the MARVEL data set and ExoMol line list for $^{12}C_2$. Mon Not R Astron Soc 2020;497:1081–97.

[50] Furtenbacher T, Tóbiás R, Tennyson J, Polyansky OL, Császár AG. W2020: A database of validated rovibrational experimental transitions and empirical energy levels of $H_2{}^{16}O$. J Phys Chem Ref Data 2020;49:033101.

[51] Furtenbacher T, Tóbiás R, Tennyson J, Polyansky OL, Kyuberis AA, Ovsyannikov RI, et al. The W2020 database of validated rovibrational experimental transitions and empirical energy levels of water isotopologues. Part II. $H_2{}^{17}O$ and $H_2{}^{18}O$ with an update to $H_2{}^{16}O$. J Phys Chem Ref Data 2020;49:043103.

[52] GEISA website. https://geisa.aeris-data.fr/.

[53] Diestel R. Graph theory. Berlin: Springer; 2006.

[54] Ritz W. On a new law of series spectra. Astrophys J 1908;28:237–43.

[55] Harel D, Tarjan RE. Fast algorithms for finding nearest common ancestors. SIAM J Comput 1984;13:338–55.

[56] Eigen v3 website, last accessed on 2021/05/27 12:52:46. http://eigen.tuxfamily.org.

[57] Mulliken RS. Report on notation for the spectra of polyatomic molecules. J Chem Phys 1955;23:1997–2011.

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

[59] Polyansky OL, Kyuberis AA, Lodi L, Tennyson J, Ovsyannikov RI, Zobov N. ExoMol molecular line lists XIX: high accuracy computed line lists for $H_2{}^{17}O$ and $H_2{}^{18}O$. Mon Not R Astron Soc 2017;466:1363–71.

[60] Gordon IE, Rothman LS, Hargreaves RJ, Hashemi R, Karlovets EV, Skinner FM, et al. The hitran2020 molecular spectroscopic database. J Quant Spectrosc Radiat Transf 2021. In press

[61] Árendás P, Furtenbacher T, Császár AG. From bridges to cycles in spectroscopic networks. Sci Rep 2020;10:1–13.

[62] Tóbiás R, Furtenbacher T, Simkó I, Császár AG, Diouf ML, Cozijn FMJ, et al. Spectroscopic-network-assisted precision spectroscopy and its application to water. Nat Commun 2020;11:1708.

[63] Rothman LS, Wattson RB, Gamache RR, Schroeder J, McCann A. HITRAN, HAWKS and HITEMP high-temperature molecular database. Proc Soc Photo Opt Inst 1995;2471:105–11.

[64] Sousa-Silva C, McKemmish LK, Chubb KL, Baker J, Barton EJ, Gorman MN, et al. Original Research By Young Twinkle Students (ORBYTS): when can students start performing original research? Phys Educ 2018;53:015020.

[65] Padberg M, Rinaldi G. A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems. SIAM Rev Soc Ind Appl Math 1991;33:60–100.

[66] GLPK (GNU Linear Programming Kit) library, last accessed on 2021/05/27 12:52:46. https://www.gnu.org/software/glpk/.