ExoMatch: automated matching and assignment of experimental spectra using ExoMol line lists

Phillip A. Coles

February 13, 2020

Department of Physics and Astronomy, University College London, London WC1E 6BT, UK

1 Introduction

Assignment of experimentally measured rotational-vibrational spectra is important as it allows the spectra to be of practical use outside the immediate conditions under which they were recorded. Full assignment of an experimentally measured spectrum involves characterising the upper and lower state rotational-vibrational quantum numbers and energies of the individual transitions. The lower state energies are necessary to determine the temperature dependence of the line intensities, and the quantum numbers are necessary to accurately approximate the effects of line broadening.

A common and versatile method of assigning an experimentally measured spectrum is by direct comparison with an equivalent theoretically calculated spectrum which, by nature of the calculations, will contain information about the upper and lower state energies and quantum numbers. Several databases exist that contain theoretically calculated rotational-vibrational spectra (ref.) that can be used for such a purpose, a notable example is the ExoMol project (???) which has generated theoretical line lists for over 30 (???) molecules of key importance to modelling exoplanet and terrestrial atmospheres.

An important precursor to the assignment procedure is the derivation of line positions and line intensities from the raw experimental absorption or emission spectra. This is necessary to help mitigate incorrect assignments due to line blending. Once the experimental data has been converted to this line list format the assignment procedure is as follows. Firstly, tentative assignments are performed by matching the experimentally determined lines to their theoretical counterparts, either by eye or by developing a script that compares the observed and calculated line positions and line intensities between the two line lists. Each tentative assignment is then either validated, or not, using a method based on Combination Differences. Whereas the latter step may be easily automated, matching of the experimental lines to their theoretical counterparts is not so straightforward, as there are often systematic inaccuracies in the theoretical line list, the exact dependencies (regarding the rotational and vibrational quantum numbers) of which are not known until after they have been used to successfully assign spectra. This uncertainty is exacerbated in molecules that have dense, complex spectra such as ammonia, making the procedure difficult to automate and extremely time-consuming to perform manually.

In this paper we report a code to automate the optimal matching of experimental and theoretical line lists, and then assign the experimental spectra using the matches as a starting point. Our matching procedure uses an implementation of the Hungarian Algorithm, a well known combinatorial optimization algorithm used to solve the general 'assignment problem'[1] in mathematics. In the following sections we first...

2 Optimal matching of spectra

The problem of matching lines between two line lists can be formulated using graph theory by representing each line list as one set of non-adjacent vertices in a complete weighted bipartite graph. Using this representation each line corresponds to a vertex, each edge that connects two vertices corresponds to a possible match, and the weight of that edge represents the cost of that match. The aim is to find the maximum matching, that is, a matching where every experimental line is matched to a theoretical line, where the sum of edge weights included in the matching is minimum.

Converting this problem to a formal mathematical definition we denote the two disjoint sets of vertices as U and V composed of indivual elements u_i and v_j where $i = 1, ..., N$ and $j = 1, ..., M$. If we define set U to correspond to the theoretical line list and V to correspond to the experimental line list then $N \geq M$. The objective is then to find

$$\min \sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij} x_{ij} \tag{1}$$

[1] Note that in this context 'assignment' does not refer to the attachment of rotational-vibrational quantum numbers to measured spectra as it does in the spectroscopic usage. In the remainder of the text we refer only to the spectroscopic sense of the word.

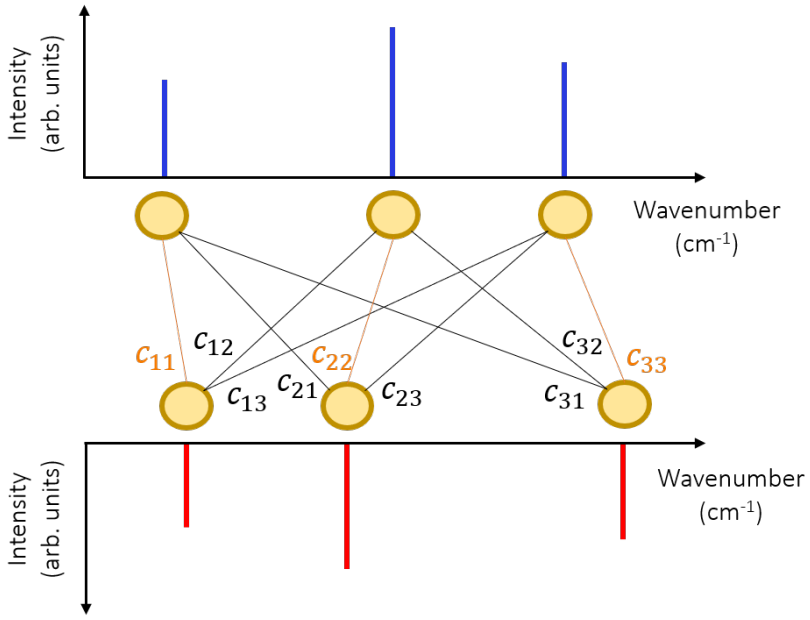

Figure 1: A caption

subject to

$$\begin{aligned} \sum_{i=1}^{N} x_{ij} &= 1 \qquad \forall\, j \\ \sum_{j=1}^{M} x_{ij} &\leq 1 \qquad \forall\, i \\ x_{ij} &\in \{1, 0\} \qquad \forall\, i, j \end{aligned} \tag{2}$$

where $x_{ij} = 1$ if u_i is matched to v_j and 0 otherwise, and c_{ij} is the weight of the edge connecting u_i to v_j which is henceforth refer to as the cost of the match. An example situation for two line lists, each containing three lines, is illustrated in Figure 1. The orange vertices show a possible maximum matching between the two line lists, in this case the total cost of the matching is given by $c_{11} + c_{22} + c_{33}$.

2.1 Cost function

The objective function c_{ij} measures the cost of matching u_i to v_j. In the absence of any prior assignments, the two fundamental quantities that are always provided in an experimental line list are transition wavenumber ν and line intensity I. We denote ν_j^{o} and ν_i^{c} as the experimentally observed and theoretically calculated line positions respectively, I_j^{o} and I_i^{c} as the corresponding line intensities, $\tilde{\nu}_{ij} = \nu_j^{\mathrm{o}} - \nu_i^{\mathrm{c}}$ and $\tilde{I}_{ij} = I_j^{\mathrm{o}}/I_i^{\mathrm{c}}$. Assuming only these quantities are available we propose a cost function of the form

$$c_{ij}(\tilde{\nu}, \tilde{I}) = \sqrt{|\tilde{\nu}_{ij}|^2 + k \cdot \left|\log(\tilde{I}_{ij})\right|^2} \tag{3}$$

where k is a constant used to control the relative weight of $\tilde{\nu}_{ij}$ and $\tilde{I}_{ij}$. The magnitude of k will predominantly depend on the relative accuracy of the line positions compared to line intensities in the theoretical calculations, which in turn depend on the accuracy of the underlying potential energy surface (PES) and dipole moment surface (DMS), as well as other thresholds employed in the calculation of the rotational-vibrational energy levels.

Of course we are free to choose the form of c_{ij} and several other forms were tested, including those with higher order polynomial terms and asymptotic terms. However, the relatively simple form given in Eq. 3 seemed to work best. In principle even quantum numbers could be incorporated into the matching by including terms that reduce to zero if the labelling of u_i agrees with that of v_j, and asymptote to infinity otherwise.

2.2 Matching Algorithm

Sets of equations Eqs. 1 and 2 describe a non-square generalisation of the square ($N = M$) Linear Assignment Problem (LAP). Several algorithms have been developed to solve this generalised case, referred to as the Rectangular Linear Assignment Problem (RLAP), such as those detailed in Refs. cite. Alternatively the problem can be solved using standard LAP algorithms cite by adding an additional $N - M$ vertices to set V and allocating zero weight to all edges that connect to them. This is equivalent to adding an additional $N - M$ lines to the experimental line list that can be matched to any theoretical line with zero additional cost. We have chosen this approach rather than solving the

RLAP, as it has allowed us to use a particularly efficient implementation of the Hungarian Algorithm that involves $\mathcal{O}(N^3)$ operations rather than $\mathcal{O}(N^4)$, as was the case for the original algorithm.

Timing data for the algorithm, measured when applied to real spectra, is presented in Figure 2. In practical use one would rarely be required to match more than a few thousand lines in one sitting, instead a large spectrum would be split into regions naturally based on the vibrational band structure, then only the strongest fraction of lines would be matched and used as a starting point for assignment. This means that execution times would be restricted to seconds rather than minutes.

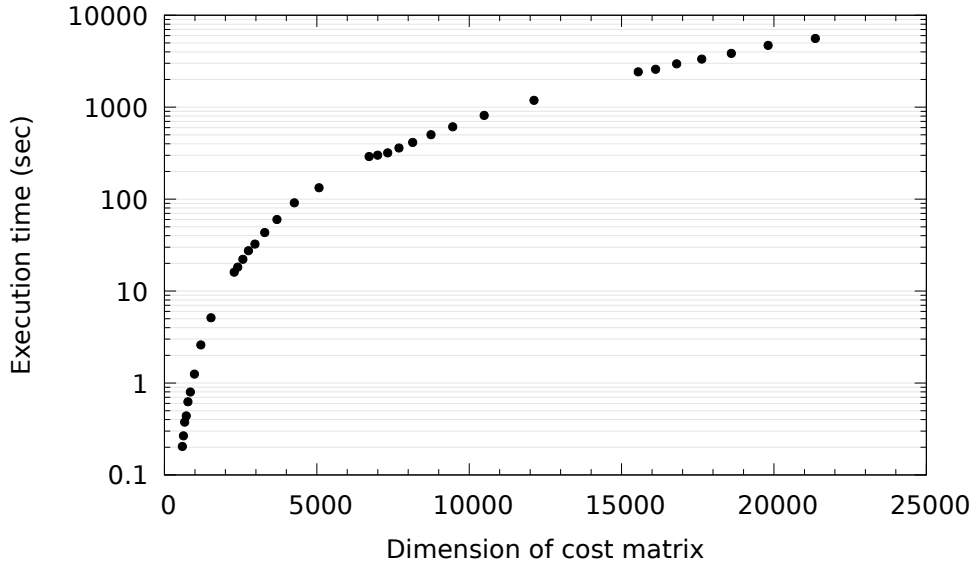

Figure 2: Two lines sharing the same numerical label indicates that they are a matched pair.

An excerpt of the matchings generated by solving the LAP (cost matrix coefficient $k = 0.5$) in two particularly difficult spectral regions of $^{12}CH_4$ (see Section 5.2) is presented in Fig. 3. For a human performing a by-eye comparison, generating a matching would involve careful consideration of the relative line positions and intensities of the theoretical and experimental data sets in both regions, and would likely yield a similar or identical matching (depending on the individual's interpretation) to those presented in Fig. 3 requiring a timescale orders of magnitude longer than ExoMatch. Alternatively, a comprehensive algorithmic approach would involve recursively searching for a best match for each line (i.e. where each matched pair agrees closely in both line position and line intensity) that simultaneously allows its neighbours to be partnered in an equal fashion. As each line can only have one partner, a reasonable approach would be to base the matching on some quantity related to the line position residuals and intensity ratios of all the matched pairs simultaneously, wherein we have arrived at a situation analogous to the Linear Assignment Problem implemented in ExoMatch.

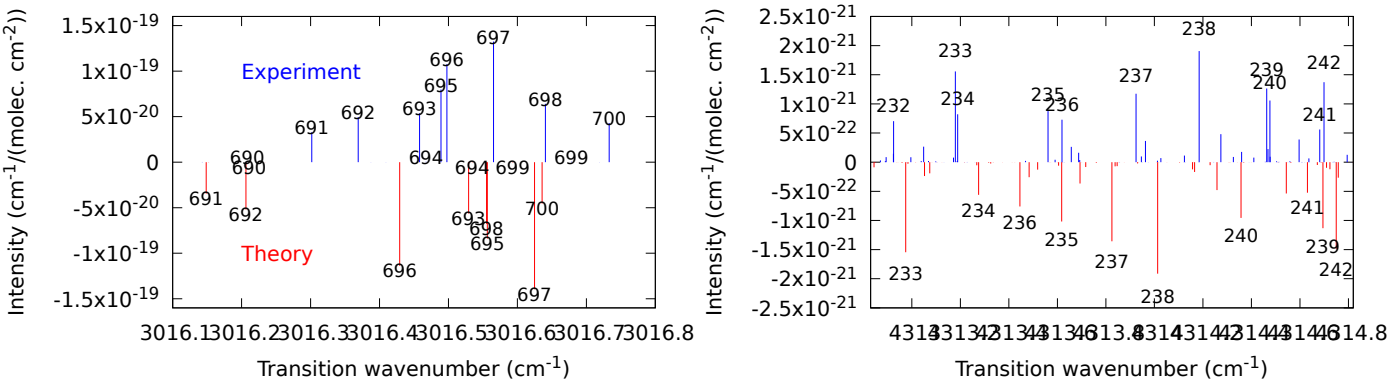

Figure 3: Two lines sharing the same numerical label indicates that they are a matched pair.

Pseudocode???

3 Validation using Combination Differences

Once the minimum-cost matching has been generated, each match must undergo a validation procedure to determine whether the experimental line indeed corresponds to its theoretically calculated partner. If successful, the experimentally

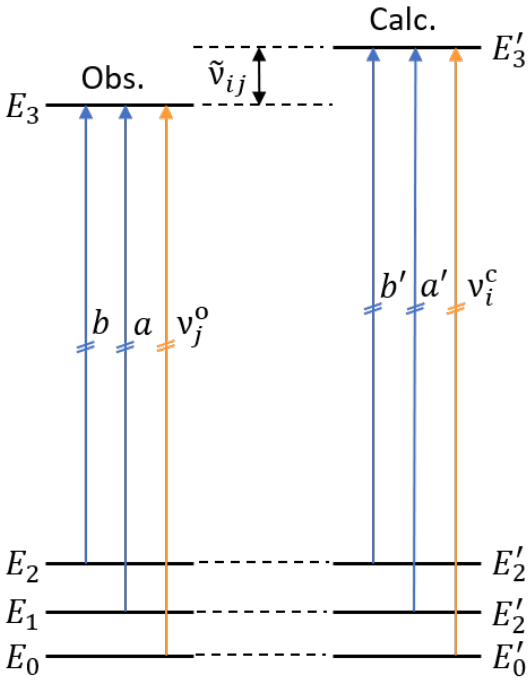

Figure 4: A caption

determined line can be characterised by the quantum numbers of its partner, and the line is said to be assigned. The validation procedure utilised in this work is based on the well known Ground State Combination Differences procedure, and has been used to successfully assign multiple infrared spectra in the past cite.

Figure 4 shows two energy level diagrams, each illustrating the same set of three allowed transitions (coloured vertical arrows) as given in the observed (left diagram) and calculated (right diagram) line lists. Let v_j, A and B denote experimentally observed transitions with wavenumbers ν_j^o, a and b respectively, that occur between lower states E_0, E_1, E_2 and the upper state E_3. The same transitions represented in the theoretical line list are then u_i, A' and B', occuring at wavenumbers ν_i^c, a' and b' respectively, between lower states E'_0, E'_1, E'_2 and the upper state E'_3. Providing that v_j (left orange arrow) has been matched to u_i (right orange arrow), and assuming the difference between the observed and calculated line positions can be attributed solely to the error in the theoretical upper state energy (i.e. that of E'_3), the exact values of a and b can be estimated as $a \approx a' + \tilde{\nu}_{ij}$ and $b \approx b' + \tilde{\nu}_{ij}$. This is a reasonable assumption to make if E'_0, E'_1 and E'_2 belong to the ground vibrational state, as purely rotational states are usually predicted with < 0.01 cm^{-1} accuracy in modern nuclear motion calculations cite. Even if they are not, this level of accuracy can be achieved simply by replacing the energetic values of E'_0, E'_1 and E'_2 by their previous experimentally derived values where available cite.

Once the values of a and b have been predicted using the above method, the existence of A and B in the experimental line list must be confirmed. If A and B are indeed found at wavenumbers $a-\Delta < a < a+\Delta$ and $b-\Delta < b < b+\Delta$ where Δ is the experimental uncertainty, and the observed line intensities of A and B agree satisfactorily with the theoretical intensities of A' and B', then the upper and lower state quantum numbers of u_i, A' and B' can be assigned to lines v_j, A and B respectively. In this case the original match (u_i,v_j) is considered validated, and the lines A and B that have consequently been mapped onto A' and B' are referred to as Combination Difference (CD) partners. The number of CD partners required to confidently validate an initial match will generally depend on how strong and isolated the lines involved are. For strong, isolated lines only one partner may be enough, whereas for weaker lines inhabiting dense spectral regions three or more may be necessary to ensure confidence in the assignments.

4 Program overview

ExoMatch was designed to take two line lists, one experimentally derived and one theoretically calculated containing unique upper and lower state quantum labels and energies, and return a comprehensive list of assignments using the procedures outlined in Sections 2 and 3. A flow chart of the code is presented in Fig. 5.

The execution of the program involves the following steps:

1. Read the input instruction;
2. Read the experimental line positions and line intensities, and the theoretical line positions, line intensities, upper and lower state quantum numbers and energies;
3. Prepare the sets of theoretical lines U and experimental lines V to be matched. Write U and V to files (if required);
4. If user supplies matching and $N_{\text{iter}} = 1$, go to Step 7;
5. Prepare cost matrix c_{ij};
6. Determine minimum cost matching $O \subseteq U \times V$. Output matches;

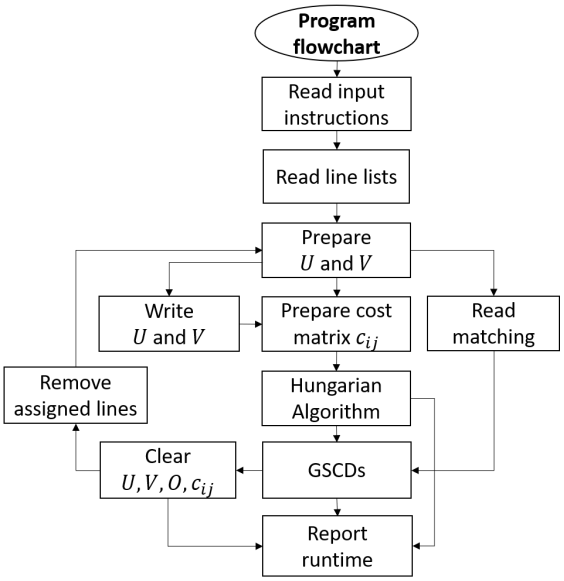

Figure 5: Structure of program

7. Search for GSCD partners for each match, starting with strongest lines. Output assignments;
8. If user specifies $N_{\mathrm{itmax}} > 1$, clear U, V and O, then remove assigned lines from lines lists and go to Step 3;
9. Report elapsed time. Terminate;

The ExoMatch input file structure is displayed in Table 1. The (complete set of) keywords listed correspond to parameters to be set by the user and are immediately followed by their parameter values. The ordering of keywords in the input file does not matter.

Table 1: Caption.

Keyword	Value	Type	Description
`obsfile`	`explinelist.inp`	String	File name of experimental line list
`calcfile`	`calclinelist.inp`	String	File name of theoretical line list
`obsrange`	`0 10000`	Integer	Wavenumber range of experimental lines to be matched
`calcrange`	`0 10001`	Integer	Wavenumber range of calculated lines to be matched
`calcIthresh`	`0.5e-24`	Real	Lower intensity threshold for calculated lines to be matched
`obsIthresh`	`1.0e-24`	Real	Lower intensity threshold for experimental lines to be matched
`readmatches`	`matching.inp`	String	Name of file containing matches. Can be omitted.
`matching`	`print`	String	Write lines be matched to separate files
`GSCD`		N/A	Perform GSCDs
`CDthresh`	`read/0.005`	String/Real	Experimental line position uncertainty for GSCDs
`Iratio`	`0.7`	Real	Defines acceptable limit of $I_{\mathrm{obs}}/I_{\mathrm{calc}}$ and $I_{\mathrm{calc}}/I_{\mathrm{obs}}$ for GSCD partners
`numGSCDs`	`2`	Integer	Number of GSCD partners required for assignment
`Nquanta`	`3`	Integer	Number of quanta specifying upper and lower states
`maxiter`	`2`	Integer	Number of iterations of matching + GSCDs to perform

The matching is performed on a subset of experimental and theoretical lines controlled by the parameters `obsrange`, `calcrange`, `obsIthresh` and `calcIthresh`. These subsets are referred to as sets V and U respectively. In the example provided in Table 1 the experimental line list is to be read from the file named `explinelist.inp`, and only experimental lines in the range $0 - 10\ 000$ cm^{-1} with intensities greater than 1.0×10^{-24} cm^{-1}/(molec. cm^{-2}) are to be included in the matching, i.e. in set V. The inclusion of the `readmatches` keyword indicates the initial matching is to be read from a file, in this case named `matching.inp`. However, this is generally not recommended unless the matching was previously generated using the current input parameters, as the line indexing supplied in `matching.inp` must agree with the indexing determined by `obsrange`, `calcrange`, `obsIthresh` and `calcIthresh`. To circumvent this, the `GSCDs` keyword can be omitted if the user only wishes to generate an optimal matching, which can then be copied to `matching.inp`, modified, and read back into the program.

Unlike the matching, the GSCDs procedure utilises the full spectrum contained in `obsfile` and `calcfile`. In the case that GSCDs are to be performed, the experimental uncertainty Δ can either be read from the input file or line-by-line from the third column of `obsfile`, the latter option allows each line to possess its own uncertainty. The number of GSCD partners, excluding the original match, required for the set to be considered assigned is controlled by `numGSCDs`, and the tolerance for differences between the observed and calculated line intensities is controlled by `Iratio`. Specifically, the GSCD partners of the matched lines u_i and v_j with intensities I_i and I_j, must have intensities that fall within $I_{\mathrm{ratio}} \times I_i/I_j$ and the reciprocal of this value.

4.1 Input line lists

Example experimental and theoretical input line list files are shown in Tables 2 and 3. The experimental line list should contain at least two fields; line position and line intensity, with an optional third field containing the experimental uncertainty of each line (see Table 2). Any additional information, e.g. previous assignments, that follows the uncertainty field will not be used but will be printed in the GSCD section output, and optionally in the matching output by including the keywords `matchinfo all` or `matchinfo obs` in the input file. The theoretical line list file should also contain line positions and line intensities, followed by `Nquanta` upper state quantum numbers, then the upper state energy, `Nquanta` lower state quantum numbers, and finally the lower state energy of each line. Again, any additional information placed after this will not be used in the assignments, but will be printed in the GSCD output, and optionally in the matching output using the keywords `matchinfo all` or `matchinfo calc`. It is important that the `Nquanta` quantum numbers provided are sufficient to uniquely define each upper and lower state for the GSCDs procedure to work. A simple example of such a scheme would employ J, Γ and $\mathrm{N_b}$, where J is the total angular momentum quantum number, Γ is the total state symmetry, and $\mathrm{N_b}$ is the index of the state in the J, Γ block.

Table 2: Excerpt from example experimental line list input file. Line positions and line intensities given in arbitrary units, uncertainty corresponds only to the line positions.

Line position	Line Intensity	Uncertainty
9120.85110	4.050e-24	0.005
9121.14780	3.880e-24	0.005
9121.20860	2.240e-24	0.001
9121.28060	1.030e-24	0.001
9122.36189	1.970e-24	0.005

Table 3: Excerpt from example theoretical line list input file. Line positions and line intensities given in arbitrary units. (n'_1, n'_2, n'_3) correspond to upper state quantum numbers, E' to upper state energy, (n''_1, n''_2, n''_3) to lower state quantum numbers, and E'' to lower state energy. Energies must be given in the same units as the line positions.

Line position	Line Intensity	n'_1	n'_2	n'_3	E'	n''_1	n''_2	n''_3	E''
9120.65979	1.461953e-25	7	5	579	9800.4950	8	2	2	679.8352
9121.03436	7.516349e-25	4	5	330	9207.6914	3	2	1	86.6570
9121.23723	4.517523e-25	10	6	1472	9982.2336	11	3	1	860.9963
9121.35818	1.942792e-25	3	5	243	9240.5966	3	2	2	119.2384
9121.42686	3.451439e-25	10	6	1473	9982.4232	11	3	1	860.9963

4.2 Output

As standard, ExoMatch will output the results of each matching and each GSCDs search it performs. Figure 6 shows the format of the output from the matching routine. The first column contains to the index of the (experimental) line in set V and the second column contains the index of its matched (theoretical) partner in set U. The columns that follow are the corresponding experimental and theoretical line positions and intensities, and lastly the entire row from the theoretical line list file is printed.

Figure 6 shows an example output from the GSCDs routine...

```
...

383   968   9672.8801 5.402e-24   9672.8107 3.94853e-24   9.67281066e+03 3.94853e-24 7 3...
384   967   9672.8854 4.920e-23   9672.6596 5.27165e-23   9.67265964e+03 5.27165e-23 5 5...
385   971   9673.1054 9.640e-24   9672.9975 9.39824e-24   9.67299754e+03 9.39823e-24 5 3...
386   969   9673.1297 4.566e-23   9672.8957 1.46177e-23   9.67289576e+03 1.46177e-23 2 5...
...
```

Figure 6: A figure.

5 Test cases

Several databases exist that aggregate lists of experimentally derived line-by-line parameters such as line positions and line intensities for molecules of terrestrial and astronomical importance. Most notable of these is the HITRAN (HIgh

```
...

[25] 9672.88543 4.920e-23 9672.65964 5.27165385e-23 0.22579
9554.13437 1.82e-23 9553.90880 2.07229894e-23 0.22557(0.22579) 0.8783 || Obs line:...  || Calc line:...
9640.46875 1.98e-24 9640.24325 2.27016158e-24 0.22550(0.22579) 0.8722 || Obs line:...  || Calc line:...
9672.88543 4.92e-23 9672.65964 5.27165385e-23 0.22579(0.22579) 0.9333 || Obs line:...  || Calc line:...
Derived experimental energy:  9938.11162

[26] 9669.13624 4.859e-23 9669.11741 5.39931096e-23 0.01883
Derived experimental energy:  0
...
```

Figure 7: A figure.

Resolution TRANsmission) database cite, that currently contains line-by-line data for 49 molecules and their isotopologues, out of which a significant proportion are partially or fully assigned upper and lower state quantum numbers and energies. The line lists contained in HITRAN represent the most comprehensive source of assigned experimental data available, and therefore provide an invaluable source with which to test the reliability of ExoMatch.

We selected two molecules that have a large number of assigned lines present in HITRAN 2016 cite, namely ammonia (NH_3) and methane (CH_4), and reassigned their respective HITRAN 2016 line lists using ExoMatch in conjunction with the theoretically calculated line lists CoYuTe cite and YT10to10 cite from the ExoMol project cite. Our assignments are then compared to those present in HITRAN to validate our method. The ExoMol line lists are available online in standard ExoMol format cite, and for the purpose of converting to ExoMatch input format (see Table 3) the ExoCross cite program was used.

5.1 Ammonia (NH_3)

Ammonia is a pyramidal tetratomic molecule exhibiting six vibrational degrees of freedom. The labelling of rovibrational states of NH_3 has been discussed extensively by Down et al. cite who suggests the following 13 useful quantum numbers to uniquely define each state:

$$\Gamma_{\rm tot}, \nu_1, \nu_2, \nu_3, \nu_4, L_3, L_4, L, i, \Gamma_{\rm vib}, J, K, \Gamma_{\rm rot}$$

where ν_i, L_i and L are vibrational normal mode quantum numbers, J is the total rotational angular momentum and K is its projection onto the molecule-fixed $z-$axis, i is the inversion parity, and $\Gamma_{\rm rot}$, $\Gamma_{\rm vib}$ and $\Gamma_{\rm tot}$ are the rotational, vibrational and total symmetries respectively. The good quantum labels J and $\Gamma_{\rm tot}$ correspond to conserved quantities so should be consistent between sources. The remaining labels are spoiled by Coriolis coupling and centrifugal distortion, but are still commonly used in various combinations for comparative purposes.

ExoMatch was used to match and assign HITRAN spectra in four strongly absorbing spectral regions of $^{14}NH_3$. Input parameters, output statistics and timing data for the analysis are shown in Table 4. Only transitions from the vibrational ground and ν_2 bands with $J \leq 20$ are included in the assignment statistics, as only these states in the theoretical line list were replaced with the corresponding experimental values citeMARVEL. The number of incorrect matches provided by ExoMatch was estimated by counting the number of cases where the $(J', \Gamma'_{\rm tot}) \leftarrow (J'', \Gamma''_{\rm tot})$ labelling in HITRAN 2016 disagreed with that of the CoYuTe match. For the entire set of 4932 matches this occurred in less than 4% of cases. Performing the same comparison using the more complete labelling scheme $(J', K', v'_{\rm str}, v'_{\rm bnd}, i', \Gamma'_{\rm tot}) \leftarrow (J'', K'', v''_{\rm str}, v''_{\rm bnd}, i'', \Gamma''_{\rm tot})$, where $v_{\rm str} = v_1 + v_3$ and $v_{\rm bnd} = 2v_2 + v_4$, discrepancies were found in less than 6% of the overall matches. Of course the success of the matching will predominantly depend on the accuracy of the theoretical calculations, and for poorly predicted, complex, or incomplete spectra the likelihood of ExoMatch providing correct matches decreases substantially.

In each region, two iterations of the matching + GSCDs procedures were performed, and the number of discrepancies between the ExoMatch assignment of $(J', \Gamma'_{\rm tot}) \leftarrow (J'', \Gamma''_{\rm tot})$ and that of HITRAN 2016 was counted. Only theoretical lines with intensities greater than 1×10^{-28} cm/molecule (Regions 1 and 2) or 1×10^{-25} cm/molecule (Regions 3 and 4) were supplied to ExoMatch for performing GSCDs. Weaker lines are far denser, and less accurately predicted than strong lines, so truncating the theoretical line list in this manner reduces the probability of ExoMatch finding false GSCD partners. The experimental wavenumber uncertainty was set to a constant value of 0.0005 or 0.001 $\rm cm^{-1}$ rather than using the line-by-line values specified by the HITRAN 2016 uncertainty indices, which were often as large as 0.1 or 0.01. Using a smaller estimated uncertainty was found to significantly reduce the number of discrepancies between the ExoMatch assignments and those in HITRAN 2016, at the cost of fewer overall assignments. In this respect it is worth noting that almost all assignment discrepancies listed in Table 4 are associated with HITRAN lines with transition wavenumber uncertainty codes of 2, corresponding to large experimental uncertainties between 0.01 and 0.1 $\rm cm^{-1}$. Such lines are particularly vulnerable to incorrect assignments using methods such as GSCDs or Effective-Hamiltonian fits, and until they are remeasured with higher accuracy it is difficult to say with certainty the cause of these conflicts.

Table 4: caption.

	Region 1	Region 2	Region 3	Region 4
`obsrange`	600−1300	1400−1900	3100−3600	4250−4600
`calcrange`	600−1300	1400−1900	3100−3600	4250−4600
`obsIthresh`	5×10^{-23}	5×10^{-23}	1×10^{-22}	1×10^{-23}
`calcIthresh`	2×10^{-23}	2×10^{-23}	0.7×10^{-22}	0.9×10^{-23}
# obs. lines in matching	1205	1627	1317	783
# calc. lines in matching	1599	2168	1683	1476
Matching exec. time (s)	5.2	14.4	5.3	11.6
# incorrect matches[a]	3	11	137	34
`CDthresh`	0.0005	0.0005	0.0005	0.001
`Iratio`	0.8	0.8	0.8	0.8
`numGSCDs`	2	2	2	1
# total assignments	2883	1518	816	804
# assignment discrepancies[a]	1	10	10	9

[a] applies if any of the following HITRAN labelling disagrees with that of ExoMatch: $(J', \Gamma'_{\rm tot}) \leftarrow (J'', \Gamma''_{\rm tot})$

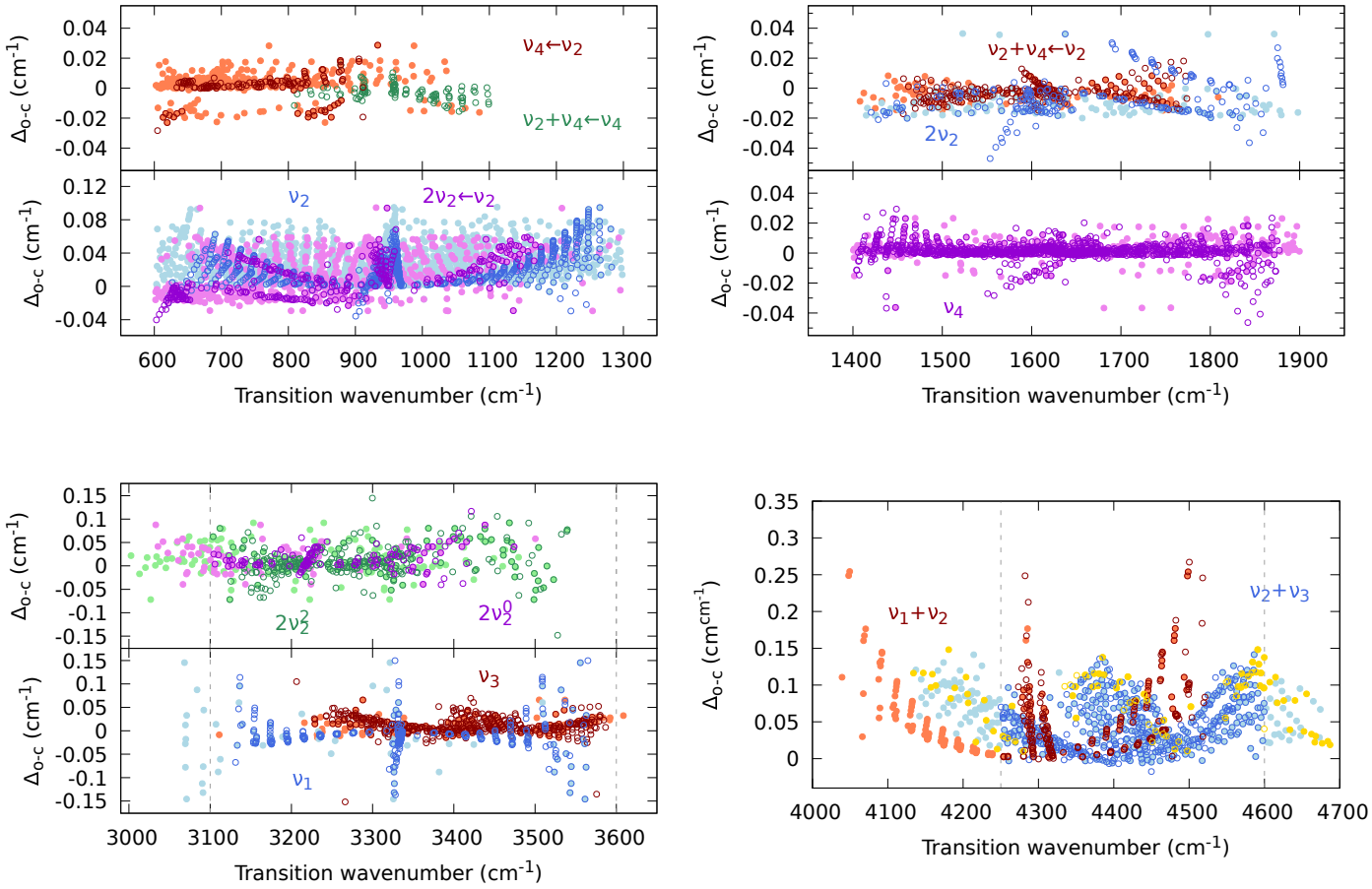

Figure 8: Two lines sharing the same numerical label indicates that they are a matched pair.

Figure 8 displays the transition wavenumber residuals of the matchings (hollow circles) and resultant GSCD assignments (solid circles), split by their CoYuTe vibrational band labels, for each of the four regions investigated. A number of HITRAN lines assigned by ExoMatch contained only partial quantum number assignments (78 in total), these are shown in yellow. Not shown are the matches or assignments whose HITRAN J and $\Gamma_{\rm tot}$ quantum number labels conflict with those found by ExoMatch. In all regions the energy residuals are predominantly structured in a manner that is characteristic of the systematic deficiencies inherent in variational calculations. This systematic structure is a strong indicator of a correct match or assignment.

5.2 Methane (CH_4)

Methane is a five atomic spherical top molecule exhibiting nine vibrational degrees of freedom. The quantum state labelling scheme employed in YT10to10 is discussed in Refcite and we adopt the same descrption, utilising the following 15 labels:

$$\Gamma_{\rm tot}, \nu_1, \nu_2, L_2, \nu_3, L_3, M_3, \nu_4, L_4, M_4, \Gamma_{\rm vib}, J, K, \tau_{\rm rot}, \Gamma_{\rm rot}$$

where ν_i, L_i and M_i are vibrational normal mode quantum numbers, J is the total rotational angular momentum and K is its projection onto the molecule-fixed $z-$axis, $\tau_{\rm rot}$ is the rotational parity, and $\Gamma_{\rm rot}$, $\Gamma_{\rm vib}$ and $\Gamma_{\rm tot}$ are the rotational, vibrational and total symmetries respectively.

ExoMatch was used to match and assign HITRAN 2016 $^{12}CH_4$ spectra in the pentad (2700−3300 cm^{-1}) and octad (3700−4600 cm^{-1}) regions. Input parameters used for the analysis, output statistics and timing data are shown in Table 5. Only transitions from the ground vibrational state with $J \leq 10$ are included in the assignment statistics, as these were the only states for which we replaced the YT10to10 predictions with their experimentally derived values citeNikitin. The HITRAN 2016 parametrisation of CH_4 contains no label for total symmetry, therefore, in the absence of a second good quantum label, an incorrect match or assignment discrepancy was counted as one in which the labelling $(J', v'_{\rm str}, v'_{\rm bnd}) \leftarrow (J'', \Gamma''_{\rm rot}, v''_{\rm str}, v''_{\rm bnd})$, where $v_{\rm str} = v_1 + v_3$ and $v_{\rm bnd} = v_2 + v_4$, in HITRAN 2016 disagreed with that of the YT10to10 match. Although $\Gamma_{\rm rot}$ is not a good quantum number, for transitions to the ground vibrational state, which comprised all but a few matches, the $\Gamma_{\rm rot}$ label acts in the place of total symmetry $\Gamma_{\rm tot}$. For the inputs listed in Table 5, ExoMatch provided the correct match at least 80% of the time, and did so for hundreds of lines in a matter of seconds.

Table 5: caption.

	Region 1	Region 2a	Region 2b
`obsrange`	2400−3300	4100−4500	4500−4700
`calcrange`	2400−3300	4100−4500	4500−4700
`obsIthresh`	2×10^{-22}	5×10^{-22}	5×10^{-23}
`calcIthresh`	1×10^{-22}	3×10^{-22}	3×10^{-23}
# obs. lines in matching	1057	395	188
# calc. lines in matching	1538	571	294
Matching exec. time (s)	6.6	0.4	0.1
# incorrect matches[a]	210	66	47
`CDthresh`	0.001	0.0005	0.0005
`Iratio`	0.8	0.8	0.8
`numGSCDs`	2	1	1
# total assignments	1197	306	170
# assignment discrepancies[a]	3	2	5

[a] applies if any of the following HITRAN labelling disagrees with that of ExoMatch: $(J', v'_{\rm str}, v'_{\rm bnd}) \leftarrow (J'', \Gamma''_{\rm rot}, v''_{\rm str}, v''_{\rm bnd})$.

ExoMatch was run for two iterations of the matching + GSCDs procedure in each region. As with the ammonia analysis, only YT10to10 lines with intensities greater than 1×10^{-25} cm^{-1}/(molecules cm^{-2}) were supplied as an input, and the experimental wavenumber uncertainty was set to a constant value of 0.001 or 0.0005 cm^{-1}. From comparing the number of assignments found by ExoMatch to the size of each matching, it is clear that roughly one-quarter to one-half of matches in each region resulted in a complete GSCD set. For the GSCD sets that were found, a small number of discrepancies were noted between the ExoMatch assignments and those present in HITRAN 2016 (see Table 5). Again these are associated solely with lines that posses high transition wavenumber uncertainties in HITRAN (i.e. uncertainty codes 2 or 3). For this reason neither an incorrect ExoMatch or HITRAN assignment can be ruled out with certainty. Nevertheless, the overall agreement is very good and illustrates the reliability of the ExoMatch procedure.

Figure 9 displays the transition wavenumber residuals of the matchings (hollow circles) and GSCD assignments (solid circles) in the pentad (upper) and octad (lower) regions. Likely incorrect matches (black crosses), as defined above, are also shown. In both regions the energy residuals are structured in the same systematic way that was observed for ammonia. Finally,

6 Conclusion

A Appendix

A.1 Basic handling

The appendix

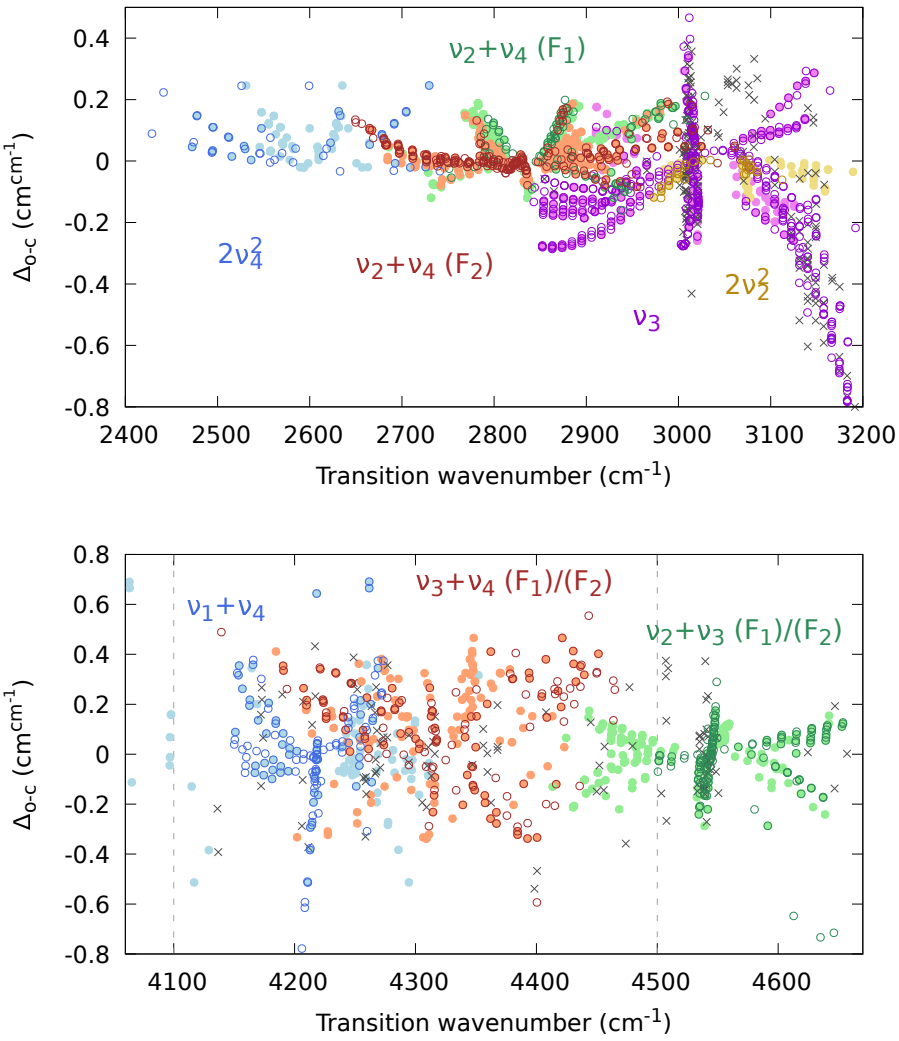

Figure 9: Two lines sharing the same numerical label indicates that they are a matched pair.