FOCUS ARTICLE

Molecular diatomic spectroscopy data

Laura K. McKemmish

School of Chemistry, University of New South Wales, Sydney, Australia

Correspondence
Laura K. McKemmish, School of Chemistry, University of New South Wales, 2052 Sydney, Australia.
Email: l.mckemmish@unsw.edu.au

Edited by: Peter Schreiner, Editor-in-Chief

Abstract
Accurate and comprehensive diatomic molecular spectroscopic data have long been vital in a wide variety of applications for measuring and monitoring astrophysical, industrial, and other gaseous environments. These data are also used extensively for benchmarking quantum chemistry and applications from quantum computers, ultracold chemistry and the search for physics beyond the standard model. Useful data can be highly detailed like line lists or summative like molecular constants, and obtained from theory, experiment, or a combination. There are plentiful (though not yet sufficient) data available, but these data are often scattered. For example, molecular constants have not been compiled since 1979 despite the existing compilation still being cited more than 200 times annually. Furthermore, the data are interconnected but updates in one type of data are not yet routinely applied to update interconnected data: in particular, new experimental and ab initio data are not routinely unified with other data on the molecule. This paper provides information and strategies to strengthen the connection between data producers (e.g., ab initio electronic structure theorists and experimental spectroscopists), data modelers (e.g., line list creators and others who connect data on one aspect of the molecule to the full energetic and spectroscopic description) and data users (astronomers, chemical physicists, etc.). All major data types are described including their source, use, compilation, and interconnectivity. Explicit advice is provided for theoretical and experimental data producers, data modelers, and data users to facilitate optimal use of new data with appropriate attribution.

This article is categorized under:
Theoretical and Physical Chemistry > Spectroscopy

KEYWORDS
data, diatomics, line lists, molecular constants, spectroscopy

1 | INTRODUCTION

Over the decades, there have been thousands of theoretical and computational studies of the high resolution spectroscopy of hundreds of diatomic molecules.¹⁻³ These data are used across the sciences including monitoring pollutants like CO, NO, SO in Earth's atmosphere,⁴ understanding astrophysical objects⁵ like planets⁶ and stars⁷ using molecules like TiO,⁸ assessing the accuracy of theoretical methodologies,⁹,¹⁰ creating and manipulating ultracold molecules, assessing the suitability of a molecule as a probe of variation of fundamental constants,¹¹,¹² for laser cooling¹³ and for quantum computing.¹⁴
The type of molecular spectroscopic data produced varies considerably but includes:

- assigned experimental transitions (spectral lines);
- energy levels;
- spectroscopic networks with self-consistent assigned transitions and energy levels with uncertainties;
- model Hamiltonian parameters, especially band and equilibrium constants;
- line lists, that is, sets of energy levels plus intensities of transitions between these levels;
- potential energy, spin–orbit and other energetic coupling curves;
- dipole and transition dipole moment curves;
- cross-sections of absorption versus frequency.

The data can be fully ab initio, experimental or modeled from a combination of both.

As a simplification, here researchers are categorized as data producers (notably ab initio electronic structure theorists and experimental spectroscopists), data modelers (e.g., line list and spectroscopic network creators) and data users (e.g., astronomers, chemical physicists). Data producers generally obtain data on one aspect of a molecule, for example, assigned transitions for a spectral band or dipole moment curves for the three lowest lying electronic states. Data modelers combine raw data into a format suitable for use in applications. For example, modelers may extrapolate from a small number of measured spectral lines to predict all transitions within frequency and intensity thresholds. Data modelers may also consolidate data from multiple individual experiments, for example, a harmonized set of molecular constants. The motivations and goals of data users are very diverse, with the group unified by the desire for easy and quick access to high-quality data with clearly communicated strengths and limitations.

Diatomic spectroscopic data are plentiful, yet contrary to expectations, existing data are not always sufficient for applications. For example, detection of molecules in exoplanets using cross-correlation of high-resolution ground-based measurements requires molecular line positions to be known to sub cm⁻¹ accuracy; for TiO this necessitated the creation of a new line list.¹⁵ As another example, europium diatomics are of potential interest in measuring variation in fundamental constants but the properties of their electronic states are not generally known.¹⁶

Problematically, available data on a single molecule are usually quite dispersed and it is an unexpectedly formidable task to collate and consolidate all available data. In the age of big data and machine learning, this is particularly concerning as the quality of any data science and/or artificial intelligence analysis depends first on the availability of well curated data sets. The difficulty of finding appropriate data means that many authors understandably often refer back to the last major collation of data on diatomic molecules; the famous Huber–Herzberg dataset of equilibrium molecular constants¹ (henceforth HH), published in 1979 after a decade of compilation. This data synthesis was hugely successful, with more than 9000 citations including at least 200 citations annually in the 2010s.

Although HH was an exemplary effort of its time, the data were provided as tables in a book. In the 21st century, researchers expect databases to be digitized, accessible, and queryable both graphically through a website and programmatically through application programming interface (API). Such databases are certainly available for some types of spectroscopic data and some molecules (see Table 1). However, this is not universal and some types of data, particularly ab initio data, do not appear to be compiled at all, an issue I highlight as a future opportunity.

A more challenging issue for our community is ensuring our databases is easily and regularly updated, with version control and thorough documentation of both initial data production and subsequent modifications. Often, data are interconnected; for examples, line lists rely on assigned experimental transitions and ab initio dipole moment curves. The technology and database systems are now in place for many (but not all) types of spectroscopic data to facilitate these updates and streamline dissemination and utilization of new data by users and in other types of data. However, these update mechanisms are rarely utilized by new data creators.

Some barriers restricting effective data production, consolidation, and frequent updating include: (1) the diversity of subdisciplines involved in producing, consolidating, and utilizing these data, (2) the variety of data types and complexity of their interconnections, (3) insufficient communication and understanding between subdisciplines on the most important data requirements, (4) technological issues with unfamiliar software and data formats, and (5) release of raw data and the inclusion of individual data into large databases, or individual databases into a centralized database, can (or at least is perceived to) reduce the citations to individual papers.

To address these barriers, this focus review has two main parts. In Section 2, I aim to enhance interdisciplinary understanding by concisely describing all the key types of spectroscopic data for diatomic molecules, identifying their interconnectivity, method of production, common use, and major compilations. Then in Section 3, I provide specific
### TABLE 1 Major data compilations of the spectroscopy of diatomic molecules

| Source | Data type | Strengths | # Mols | Molecules |
|--------|-----------|-----------|--------|-----------|
| **Spectroscopic networks** | | | | |
| MARVEL\(^{17-19}\) | SpecN, Lit | Centralized and self-consistent compilation of all experimental transitions data, uncertainties provided. | 10 | AlH, BeH, C\(_2\), CN, CP, NH, NO, O\(_2\), TiO, ZrO |
| **Molecular constants** | | | | |
| Huber–Herzberg 1979\(^1\) | EqConst | Extensiveness of molecular coverage, now available online via NIST Chem Webbook. | >100 | Many |
| Diatomic Molecular Spectroscopic Database (DMSD)\(^20\) | EqConst, FCF | Modern queryable database and user data contributions facilitated and encouraged; initially populated with some HH data | 173 | Many; 132, 35, and 6 molecules with Σ, Π, and Δ ground states respectively |
| **Rovibronic line lists** | | | | |
| HITRAN\(^4, 21\) | LL | Designed for Earth atmosphere applications with verified data of very high accuracy, with uncertainties and various line broadening parameters. | 13 | CO, CS, ClO, HBr, HF, HI, N\(_2\), NO, O\(_2\), OH, NO\(_+\) |
| HITEMP\(^22\) | LL | Similar to HITRAN but application to higher temperatures. | 3 | CO, NO, OH |
| ExoMol-main\(^3, 23–25\) | VLL, EL, SM | Completeness of each line list and validity at high temperature. Describes isotopologues easily and well. | 26 | AlH, AIO, BeH, CaH, CaO, C\(_2\), CS, HS, KCl, MgH, MgO, NaH, NaCl, NO, NS, PN, PH, PO, PS, ScH, SH, SiH, SiS, SiO, TiO, VO |
| MoLLIST\(^26\) | MHLL | Experimental basis, complementary coverage. | 20 | AlCl, AlF, CaH, CaF, C\(_2\), CH, CN, CP, CrH, FeH, KF, LiCl, LiF, MgF, MgH, NH, NaF, OH, OH\(^+\), TiH |
| ExoMol-collated\(^3, 27\) | LL | Central repository in single format for various rovibrational and rovibronic line lists, esp. ExoMol-main and MoLLIST | 52 | Non-MoLLIST/ExoMol-main: H\(_2\), HF, HCl, N\(_2\), SiO, SiO, ScH, LiH, LiH\(^+\), CO, HeH\(^+\), HD\(^+\), HD, YO |
| Kurucz\(^28, 29\) | LL | Breadth of included molecules and consistency of data compilation over time. | 45 | Notable: BH, BO, CH\(^+\), CN\(^+\), CO, CO\(^+\), CrH, CS, FeH, H\(_3\), KH, LaO, LiH, MgH, MgH\(^+\), MgO, NH, NH\(^+\), NO, NS, NaH, O\(_2\), ScO, SiC, SiH\(^+\), SiO\(^+\), SiS, TiH, YO, ZrO |
| GEISA 2015\(^30\) | LL |  | 11 | ClO, CO, HBr, HCl, HF, HI, N\(_2\), NO, NO\(^+\), O\(_2\), OH |
| Plez\(^31, 32\) | LL |  | 17 | AlH, C\(_2\), CaH, CH, CN, CO, CrH, FeH, MgH, NH, OH, SiH, SO, SIS, TiO, VO, ZrO |
| VALD\(^33, 34\) | LL | Vienna Atomic Line Data Base with some molecular line lists. | 8 | C\(_2\), CH, CN, CO, MgH, SiH, OH, TiO |
| **Microwave and millimeter-focused line lists** | | | | |
| Splatalogue (https://splatalogue.online/advanced1.php) | LL | Single portal for all LL data below | >100 | |
| CDMS\(^35–37\) | LL | Cologne Database for Molecular Spectroscopy | 84 | Many |

(Continues)
recommendations for four key groups—experimental spectroscopists, theoretical quantum chemists, data modelers, and data users—to ensure new data are effectively and easily utilized as well as appropriately attributed.

2 | DIATOMIC MOLECULAR SPECTROSCOPIC DATA

2.1 | Overview

This section describes major types of diatomic molecule data, their source, and major compilations.

Figure 1 provides a flowchart showing the interconnectivity of data types. Data are divided into five main categories: experimental spectroscopy, ab initio electronic structure theory, consolidated experimental spectroscopic network (usually from a MARVEL analysis), line list construction, and cross-sections.

Table 1 identifies major compilations of each data type; these are discussed further in the relevant subsection.

2.2 | Assigned experimental transitions

Definition: An assigned experimental transition consists of a transition frequency between an initial and final quantum state described with approximate quantum numbers and ideally an uncertainty.

Source: Experimental spectroscopy papers usually provide tabulations of assigned rovibronic spectral transitions for one or more spectral bands for one or more molecules.

Discussion: Assigned spectral transitions are the observable evidence of the molecule's energetic and spectroscopic properties.

Data compilation: Beyond individual papers (the main source of data), today these data are generally collated as consolidated experimental spectroscopic networks (see below). Previously, this type of data was usually collated as experimental line lists, in this format lacking self-consistency and often attribution to individual data sources.

2.3 | Rovibronic energy levels

Definition: Energy of rovibronic states (usually relative to the lowest energy quantum state in the molecule and ideally with uncertainties) with associated approximate quantum numbers describing the state.

Source: The most accurate set of energies is obtained by inverting the set of self-consistent MARVEL transition frequencies to obtain experimentally derived MARVEL energy levels.
Rovibronic energy levels can also be calculated from model Hamiltonian band or equilibrium constants, by solving the nuclear motion Schrödinger equation for a given energy spectroscopic model, or determined directly from experiment using combination differences if one observed transition involves the ground state.

**Common use:**

- to enable fitting of the energy spectroscopic model;
- to replace modeled energy levels in a line list and thus enable high precision frequency predictions;
- calculate molecular thermodynamics properties such as partition functions, specific heat, and cooling functions.

### 2.4 Consolidated experimental spectroscopic network

**Definition:** A large set of self-consistent assigned experimental transitions with uncertainties, and the associated inverted experimentally derived energy levels with uncertainties.

**Production:** Spectroscopic networks are usually produced through what has become known as a [Marvel](#) compilation and analysis (after the key enabling program). This involves: (1) a thorough literature review, (2) digitization and consolidation of all available assigned transitions into a single format, (3) using the [Marvel](#) software program, process data within first individual sources then multiple sources to produce a self-consistent spectroscopic network by increasing uncertainties, and removing transitions as required to ensure self-consistency of the whole spectroscopic network.

**Compilation:** Centralized to the Marvel online website with modern accessibility and database standards.
Updates: New experimental data should be immediately added to available MARVEL datasets, providing the dual benefits of cross-validation and ensuring the new data can be used easily to update line lists. The availability of MARVEL online substantially reduces the barrier to this task, which can be done through a graphical user interface. However, training with an experienced user can be useful.

Alternatively, well-formatted plain text data as supplementary information, ideally in the molecule's MARVEL format, and an email to the original compilation authors will generally enable new data to be added.

2.5 | Model Hamiltonian parameters (e.g., band and equilibrium constants)

Definition: The set of parameters that are used to approximate the energy of quantum states in a molecule based on each state's quantum numbers.

Discussion: Reasonable functional forms for the dependency of diatomic energy levels (and thus transition frequencies) on molecular quantum numbers can be obtained by approximating the full Hamiltonian of the system, that is, using a model Hamiltonian (e.g., a rigid rotor or Morse oscillator).43,44 These expressions have parameters that can be adjusted to maximize the fit between the predicted and experimental transition frequencies.

Components: The most common parameters are the band constants ($T_v$, $B_v$, $D_v$, etc.)—obtained when fitting rotational levels within a single vibrational state of a single electronic state—and equilibrium constants ($T_e$, $\omega_e$, $\omega_e\chi_e$, $B_e$, $D_e$, etc.)—obtained when fitting rotational levels from multiple vibrational levels of a single electronic state. For states with nonzero angular momentum and spin, the spin–orbit coupling, $A$, is also vital.

Higher-order couplings, such as the spin–spin and spin–rotation couplings, and off-diagonal couplings between electronic states are important for high accuracy applications.

Software: PGopher45 is the community standard program for fitting observed spectral transitions to model Hamiltonians.

Advantages: Good summary of the key features of a molecule's electronic structure, sufficient for many applications. Can be used to calculate partition functions.46

Limitations: Molecular constants are problematic when there is strong coupling between electronic states that lead to perturbations in quantum energy levels.47

Energy levels and spectra predicted from molecular constants do not extrapolate beyond experimental data as reliably as variational approaches based on fitted energy spectroscopic models.

Finally, it is important to note that combining molecular constants from multiple sources needs to be done carefully as molecular parameters can be strongly correlated, meaning two different fits can be inconsistent; for example, the value of rotational constant of an electronic state changes depending on whether the centrifugal distortion constants are included in the fit.

Sources: Many papers report equilibrium and band constants only for their data. It is less common to find a synthesis of all the molecular constants for a given molecule.

Major compilations: Currently, the best compilation is still HH from 1979. The easiest way to access most of this data is the Diatomic Molecular Spectroscopy Database (DMSD),20 which adheres to modern database standards. More importantly, DMSD is updateable by the community; inclusion of molecular constants post-1979 into DMSD is a key priority in diatomic data collation. Note that strongly correlated parameters—usually parameters of the same electronic state—should generally be obtained from a single source and jointly updated as required by a refit to all available experimental transition data.

Band constants are not centrally compiled, yet are preferred to equilibrium constants for high accuracy applications. The easiest pathway forward is probably the centralized collation of PGopher input files, with appropriate author attributions. These data could then be processed into a SQL database and made accessible online via, for example, the DMSD website.

2.6 | Line lists

Definition: A line list contains the most comprehensive model of a molecule's spectroscopy, containing energy levels, transition frequencies, and transition intensities.
Use: Line lists are most frequently used to predict absorption cross-sections and as input to sophisticated (e.g., radiative transfer) models of gaseous environments. Line lists can also be used to compute state lifetimes.

Types: Figure 1 identifies five produced types of line lists that are characterized by the

- the source of the interpolated and extrapolated energy levels and transitions—model Hamiltonians (MH) or energy spectroscopic model (ESM)—these are sometimes called perturbative and variational line lists respectively;
- whether experimentally derived MARVEL energy levels and/or experimental transitions are used to replace modeled energy levels.

A potential superior type of line list, not yet created, would use MARVEL energies where available, MH energies where reliable and ESM energies elsewhere.

Software: LEVEL has long been the standard program for computing diatomic line lists. However, its limited ability to describe multiple coupled electronic states means that it is being replaced by the new program DUO which treats this coupling systematically. Both programs solve the nuclear motion Schrödinger equation to convert an ESM to a set of energy levels, then combine the rovibronic wavefunction data with the intensity spectroscopy model to produce a final line list.

Recommendations: Line lists should be regularly updated as new sources of ab initio energetic data and/or assigned experimental transitions become available. To assist this, all input files and data to a produced line list should be included as supplementary information to published papers; for example, LEVEL or DUO input files, and the energy and intensity spectroscopic models. When experimental corrections are made to calculated data, the corrected and original data should be retained.

Major database compilations: Table 1 highlights that line lists are the most frequently compiled type of diatomic data. There is considerable diversity in data from different databases as shown for three representative diatomics in Table 2.

Two centralized database compilation websites have become prominent—the ExoMol website focusing on broad wavelength coverage and Splatalogue focusing on microwave lines—that have been very successful in enhanced data accessibility while retaining credit to the original line list creators.

2.7 Ab initio energetic data

Definition: Data produced entirely from ab initio quantum chemistry electronic structure calculations using a certain model chemistry (i.e., level of theory and basis set) that enable the calculation of the energies of rovibronic quantum states in a diatomic molecule.

Types: The most important types of energetic data are the potential energy curves and the diagonal and off-diagonal spin–orbit coupling curves (or at least their equilibrium values).

Production: The best data for multiple low-lying electronic states usually come from multi-reference configuration interaction calculations based on orbitals obtained from state-selective or state-averaged Complete Active Space Self-Consistent Field calculations. For the lowest lying electronic state of a given spin-symmetry with little multi-reference character, CCSD(T) with basis set extrapolation is usually the best practical choice.

Common use: Refined or raw ab initio energetic data are often a key component of an energy spectroscopic model, and thus a line list.

Beyond line list creation, potential energy curves and the ab initio coupling are essential for studies of dynamics, for example, through molecular dynamics.

Compilation: Currently very poor compilation with most potential energy curves and related energetics data in individual papers. Furthermore, many of these papers only contain figures or fitted constants, rather than the raw data.

Ab initio vibrational frequencies for diatomics for a range of model chemistries are available on the Computational Chemistry Comparison and Benchmark DataBase.

2.8 Energy spectroscopic model

Definition: The set of potential energy curves and coupling curves (e.g., spin–orbit, non-adiabatic) that together can be used with a nuclear motion program such as LEVEL or DUO to produce rovibronic energy levels for a molecule. Generally, energy spectroscopic models are fitted against available spectroscopic experimental data.
Production: Both LEVEL and DUO include functionality to produce a high accuracy ESM by fitting a parameterized set of potential energy curves and couplings (often obtained by ab initio energetics data) to assigned experimental transitions.

Use: The most common use of an ESM is to produce a line list. However, ESMs can be used instead of ab initio energetics data in a variety of applications, for example, to simulate reaction dynamics, and will reproduce experiment better than raw ab initio data.

Compilation notes: Centralized compilation of ESMs is scarce. ESMs are now routinely reported for new variationally produced ExoMol line lists, but are challenging to find for many older line lists or line lists, somewhat restricting their usefulness of their data for some applications and making future line list improvements more challenging.

2.9 | Intensity spectroscopic model (i.e., dipole and transition dipole moments)

Definition: The set of diagonal and off-diagonal dipole moment curves that are combined with an ESM to produce the intensities of spectral transitions within a molecular line list.

Production: Intensity information is almost always purely from ab initio calculations using the methods described above, rather than experimental measurements.

Verification: Accurate absolute intensities from experiment are challenging. Instead, relative intensities from lab measurements and real-world data (e.g., astronomical spectra) can be used to validate models. Furthermore, experimental lifetimes for excited rovibronic states can directly validate the absolute intensities in the models.
Compilation notes: Intensity spectroscopic models are now routinely included as part of ExoMol line lists and their online files.

Ab initio dipole and transition moments are available in literature, usually scattered. Unfortunately, data are often exclusively in figures and thus unusable, necessitating repeated calculations.

Updates: Line lists can be trivially updated with new dipole moment curves.

Corrections based on observational data on the relative intensity of bands are possible, but should be carefully considered and documented thoroughly.

2.10 | Cross-sections

Definition: The intensity of absorption as a function of frequency or wavelength in given environmental conditions (temperature, pressure, resolution).

Data sources: Although cross-sections are routinely catalogued for bigger molecules (e.g., by HITRAN\textsuperscript{21} and NIST\textsuperscript{52}), for diatomics the underlying data such as molecular constants are more commonly reported instead, though individual papers often have cross-sections as figures.

Modeled cross-sections can be obtained from line lists using the appropriate software (e.g., Exocross\textsuperscript{48} for ExoMol-formatted line lists) at a variety of temperatures and pressures. Limitations in experimental or observational resolution can be modeled using line broadening parameters.

Advantages: Cross-sections are the data type most closely linked with observable data, and can be a useful visualization aiding interpretation of the underlying spectroscopy of the model. For example, a cross-section plot is the easiest way to identify which spectral region has the strongest absorption for a given molecule and how that absorption changes with temperature and pressure.

2.11 | Compiled literature

Description: Molecule-sorted compiled lists of a significant number of references for diatomic spectroscopic data.

Motivation: There are two main motivations: (1) collate sources of spectroscopic data that have not yet been incorporated into major databases, (2) collate sources of diverse data and information beyond that categorized above that may be useful for data users—for example, predissociation and reaction rates.

Collation of all spectroscopic data on an individual diatomic can be quite time-consuming, and requires expert knowledge and skills to ensure comprehensive coverage of crucial data. Curated and updated list of literature for each molecule covering a wide variety of different data types is thus a worthwhile pursuit, admittedly one that is challenging to justify in isolation.

Production: The diatomic spectroscopic literature is quite scattered with significant amounts of data can only be found in the original manuscripts. Keyword based searches for molecular data on typical academic search engines are often cumbersome, especially for molecules where the diatomic form is not the most common occurrence of the element combination such as titanium oxide. Commonly, searching the citation web (i.e., references and citations) of initially identified papers has been the most effective way of collating relevant literature, but this relies on the connectivity of all relevant literature and is time consuming with existing tools.

Major compilations: There are two main sources of compiled literature data, the 2001 “Diref” database by Bernath and McLeod\textsuperscript{2}—an excellent resource but now out-dated—and the ExoMol bibliography (www.exomol.com).

The HH and MARVEL compilations both incorporate all relevant references in their data.

3 | ENSURING NEW DATA ARE EFFECTIVELY UTILIZED AND ATTRIBUTED

3.1 | Experimentalists

Data, most notably assigned transitions, provided by experimental spectroscopists are crucial to achieving spectroscopic accuracy for frequencies in models of diatomic spectroscopy. These data should always be provided in plain text format to facilitate use and avoid processing challenges and potential errors.
It is optimal to minimize the lag between production of new data and its utilization by data users; this goal can be achieved by incorporating the update of available databases alongside the publication of the spectroscopic analysis. Specifically, equilibrium constants should be added to the DMSD (noting band constants may be incorporated in future expansions). Where already compiled, new data should ideally be combined with the existing Marvel spectroscopic network, with existing line lists Marvelized to incorporate the improved experimental data. Inclusion of new data in the full spectroscopic network of the molecule is also beneficial to the verification of the experimental data assignments; previous consolidations have led to the identification and correction of calibration errors, and could also be useful in speeding up new assignments (noting that of course previous data can be removed if the initial spectral assignments are rejected).

Of course, it is disingenuous to ignore the barriers to this use of databases. Data producers may fear lack of attribution to their original research when their data are compiled into databases; three main remedies are recommended:

- Data modelers should ensure data structures incorporate proper attribution of data sources; see data modelers recommendations. The original source is definitely included in the DMSD and Marvel data. The situation is more complex with line lists as they are synthesis of all available data. The best way to ensure the data are attributed appropriately is probably to incorporate an update to the line list within an experimental paper so that users cite the update and the original paper.
- Increase understanding in the user community regarding the importance of data attribution; see the data user recommendations.
- Use of nontraditional metrics in quantifying the impact of the research, especially download counts.

The other barrier is the technical challenges and time associated with incorporating data into databases; recent progress has substantially reduced these barriers, specifically:

- Adding equilibrium constants to the DMSD website is trivial requiring only text entry. Complications could arise with complex Hamiltonians.
- To add new measured assigned transitions to a pre-existing spectroscopic network requires reformatting data, then running through the Marvel program using the online website first by itself then in combination with pre-existing Marvel data. If all experiments are error-free, this is a trivial step, but in practice there is normally a process of increasing uncertainties on some transitions and removing other transitions to obtain a self-consistent network. The tutorials and new features on the Marvel website are extremely helpful, but experienced guidance could be useful for new users and in difficult cases.
- Marvel energy levels can be readily added to an existing line list to create an updated Marvelized line list suitable for high-resolution purposes. This step is usually trivial involving converting between data representations. Because the energy levels (not transitions) are replaced, transitions not measured can be reproduced with experimental accuracies if upper and lower energy levels are in the experimental spectroscopic network. Isotopologue line lists can also be Marvelized by assuming that the error in the spectroscopic model is approximately isotopologue-independent, and thus the main isotopologue correction factor can be adopted.

As a final note to experimentalists, Marvel publications often explicitly identify where future experimental data would be most useful.

3.2 Theoreticians

Utilization of new data can be maximized by ensuring that

1. dipole moments curves are calculated alongside potential energy curves (note diagonal dipoles are usually calculated automatically but not extracted while transition moment need to be specifically requested but are vital for rovibronic spectroscopy);
2. the raw data for curves (potential energy, dipole moment) are made available in the supplementary information, ideally in plain text format;
Updated data on experimentally undetermined electronic states, and diagonal and off-diagonal spin–orbit coupling, either equilibrium values or curves, are also very valuable. Inclusion of sample input files is strongly encouraged to enhance data reproducibility.

Evaluating the effect of new dipole moment curves for molecules with existing line lists and spectroscopic models can be trivial, and can quickly lead to updates in the recommended line lists. With DUO line lists, one can simply replace existing curves in the input file, recalculate the line list and then use the ExoCross software package to produce and compare absorption cross-sections with the old and new dipole moment curves. Quantitative evidence of the effect of the changed dipole moment on the molecule's spectral properties could strongly enhance the theoretical publication. Theoreticians should discuss their results with the original line list creators; if the theoretical approach for dipole moments is clearly superior to the original model, then the newly generated line list can be adopted as the recommended version with associated citations. The inclusion of new potential energy curve data into line lists is more complicated and probably would need to be considered on a molecule-by-molecule basis in discussion with the line list creator.

Finally, broadly, there is a need for a centralized, format-consistent database of diatomic ab initio data, especially for machine learning applications.

3.3 Data Modelers

The ease with which data producers can incorporate new data into models, and data users can access the data including updates is central to the effectiveness of the research. To enable this, data modelers should:

- provide input files and all input data;
- provide well-documented software programs (e.g., using journals such as Computational Physics Communication and WIREs Computational Molecular Science), or utilize well-established programs;
- enable data access both graphically via websites and programmatically through, for example, APIs (e.g.,21,24);
- ensure references to original data (e.g., for ab initio calculations or experimental results) are retained and cited appropriately.

Mutually beneficial relationships between new model creators and data curators can be created that address the needs of both the model creators for attribution to enable career progression and funding, and users in terms of data accessibility; the incorporation of line list data from many authors into a central ExoMol website are an excellent example of this. These efforts are enhanced by explaining the importance of citing the original source to users (e.g., see below). Download counts can complement citations and demonstrate impact, and should thus be provided regularly to data contributors.

Regular small updates of individual molecule datasets, not substantial enough to warrant a new paper, should be encouraged. These intermittent updates could be documented through nontraditional publication such as arxiv or research notes, then compiled in regular database updates (already common with many major databases e.g., ExoMol and HITRAN). To support this, version control and clear data naming are essential. Unfortunately, the author knows of multiple instances where users found data inadequate for their research almost entirely because they were using out-of-date data.

3.4 Data users

The most important message to data users is the following: to enable modern data, all the researchers involved in producing this data must be supported from experimental spectroscopists, ab initio quantum chemists, line list creators, and so on. The easiest way to do this is through citations to individual data producers and modelers so they can quantitatively demonstrate the usefulness of their research. Although it is easier to just cite the database, without the citations to their individual data, data producers cannot justify the time required to put their data into an easily accessible form in a database—in fact, it might be counterproductive for them to do so!

Data users are encouraged to email their publications to data producers and/or modelers where the spectroscopy data proved particularly useful. Data are often used in more diverse applications than originally anticipated; letting data producers know about this diversity can strengthen the strength of their next proposal to collect the desired data.
Users are encouraged to reach out to modelers if they require data not currently available or if there are problems with the existing data.

4 | CONCLUSIONS

New high-resolution spectroscopic measurements and first-principles rovibronic calculations on small molecules are often motivated by the molecule’s importance in the accurate modeling and understanding of astrophysical, environmental, and atmospheric environments. To enable the new insights provided by these measurements and calculations to be utilized by applications experts, the data should be incorporated into centralized spectroscopic databases in a consistent, reliable, robust, and documented form.

The links between new spectroscopic data production and the use of spectral data in scientific and engineering applications can be greatly facilitated by new data-science approaches. This paper provides specific recommendations to experimental spectroscopists, ab initio quantum chemistry theorists, data modelers, and data users to take advantage of these new approaches and maximize the speed and impact of their research. The most important of these solutions are to prioritize data availability as plain text in supporting information and incorporated into databases, and for citations to primary data sources rather than just consolidated databases. The increased use of required Data Availability Statements in journals is strongly encouraged to support these aims.

High quality updateable and queryable databases are available for many diatomic species with line lists, assigned experimental transitions, and derived empirical energy levels data; expansion of these databases to other diatomics is ongoing research. The field would benefit from centralized repositories of ab initio electronic structure calculation results in a plain text format. The new DMDS20 incorporates many Huber–Herzberg equilibrium constants into a modern updateable database and should be embraced by the community as the path forward for this data type, and ideally expanded to the collation of molecular band constants.

ACKNOWLEDGMENTS

Thank you to all of the many people who have collaborated with me over the years on diatomic molecules to showcase the potential of these deceptively complex little systems, in particular Anna-Maree Syme, Jonathan Tennyson, Sergey Yurchenko, Maire Gorman and Lorenzo Lodi, and for comments on this manuscript draft from Juan Camilo Zapata, Anna-Maree Syme and Scott Kable.

CONFLICT OF INTEREST

The author has declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

No new data was created for this review article. Data discussed in this paper is available through the cited references and/or cited online databases.

ORCID

Laura K. McKemmish 🐐 https://orcid.org/0000-0003-1039-2143

RELATED WIREs ARTICLES

Computational challenges in Astrochemistry
VeloxChem: A Python-driven density-functional theory program for spectroscopy simulations in high-performance computing environments
Decoding chemical information from vibrational spectroscopy data: Local vibrational mode theory

REFERENCES

1. Huber K, Herzberg G. Molecular spectra and molecular structure: IV. Constants of diatomic molecules. New York: Springer Science & Business Media; 1979.
2. Bernath P, McLeod SD, DiRef, A Database of References Associated with the Spectra of Diatomic Molecules, J. Mol. Spectrosc., 2001; 207–287 2001.
3. Tennyson J, Yurchenko SN, Al-Refaei AF, Clark VHI, Chubb KL, Conway EK, et al. The 2020 release of the ExoMol database: molecular line lists for exoplanet and other hot atmospheres. *J Quant Spectrosc Radiat Transf.* 2020;255:107228.

4. Gordon IE, Rothman LS, Hill C, Kochanov RV, Tan Y, Bernath PF, et al. The HITRAN2016 molecular spectroscopic database. *J Quant Spectrosc Radiat Transf.* 2017;203:3–69.

5. Fortenberry RC. Quantum astrochemical spectroscopy. *Int J Quantum Chem.* 2017;117:81–91.

6. Konopacky QM, Barman TS, Macintosh BA, Marois C. Detection of carbon monoxide and water absorption lines in an exoplanet atmosphere. *Science.* 2013;339:1398–401.

7. Tsuji T. Molecules in stars. *Annu Rev Astron Astrophys.* 1986;24:89–125.

8. Piette AA, Madhusudhan N, McKemmish LK, Gandhi S, Masseron T, Welbanks L. Assessing spectra and thermal inversions due to TiO in hot Jupiter atmospheres. *Mon Not R Astron Soc.* 2020;496:3870–86.

9. O’Neill DP, Gill PM. Benchmark correlation energies for small molecules. *Mol Phys.* 2005;103:763–6.

10. Styszynski J. *Relativistic methods for chemists.* New York: Springer; 2010. p. 99–164.

11. Flambaum V, Kozlov M. Enhanced sensitivity to the time variation of the fine-structure constant and mp/mein diatomic molecules. *Phys Rev Lett.* 2007;99:150801.

12. Syme A-M, Mousley A, Cunningham M, McKemmish L. Diatomic rovibronic transitions as potential probes for proton-to-electron mass ratio across cosmological time. *Aust J Chem.* 2019;73(8):743–756.

13. Shuman ES, Barry JF, DeMille D. Laser cooling of a diatomic molecule. *Nature.* 2010;467:820–3.

14. DeMille D. Quantum computation with trapped polar molecules. *Phys Rev Lett.* 2002;88:067901.

15. McKemmish LK, Masseron T, Hоеijmakers HJ, Pérez-Mesa V, Grimm SL, Yurchenko SN, et al. ExoMol molecular line lists – XXXIII. The spectrum of titanium oxide. *Mon Not R Astron Soc.* 2019;488:2836–54.

16. Flambaum V. Personal communication.

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

18. Furtenbacher T, Csaszar AG. MARVEL: measured active rotational–vibrational energy levels. II. Algorithmic improvements. *J Quant Spectrosc Radiat Transf.* 2012;113:929–35.

19. Furtenbacher T, Császár AG. Marvel online 2.0. http://kkkr.chem.elte.hu/maravelonline/index.php.

20. Liu X, Truppe S, Meijer G, Pérez-Ríos J. The diatomic molecular spectroscopy database. *J Cheminform.* 2020;12. http://dx.doi.org/10.1186/s13321-020-00433-8.

21. Hill C, Gordon IE, Kochanov RV, Barrett L, Wilzewski JS, Rothman LS. HITRANonline: an online interface and the flexible representation of spectroscopic data in the HITRAN database. *J Quant Spectrosc Radiat Transf.* 2016;177:4–14.

22. Rothman L, Gordon I, Barber R, Dothe H, Gamache R, Goldman A, et al. HITEMP, the high-temperature molecular spectroscopic database. *J Quant Spectrosc Radiat Transf.* 2010;111:2139–50.

23. Tennyson J, Yurchenko SN. ExoMol: molecular line lists for exoplanet and other atmospheres. *Mon Not R Astron Soc.* 2012;425(1):21–33.

24. Tennyson J, Yurchenko SN, Al-Refaei AF, Barton EJ, Chubb KL, Coles PA, et al. The ExoMol database: molecular line lists for exoplanet and other hot atmospheres. *J Mol Spectrosc.* 2016;327:73.

25. Tennyson J, Lodi L, McKemmish LK, Yurchenko SN. The ab initio calculation of spectra of open shell diatomic molecules. *J Phys B.* 2016;49:102001.

26. Bernath PF. MolLISI: molecular line lists, intensities and spectra. *J Quant Spectrosc Radiat Transf.* 2020;240:106687.

27. Wang Y, Tennyson J, Yurchenko SN. empirical line lists in the ExoMol database. *Atoms.* 2020;8:7.

28. Kurucz RL. Including all the lines. *Can J Phys.* 2011;89:417–28.

29. Kurucz RL. Including all the lines: data releases for spectra and opacities. *Can J Phys.* 2017;95:825–7.

30. Jacquinet-Husson N, Armant R, Scott N, Chédin A, Crépeau L, Boutammine C, et al. The 2015 edition of the GEISA spectroscopic database. *J Mol Spectrosc.* 2016;327:31–72.

31. Plez B. Molecular data from plez. https://www.lupm.inp3.fr/users/plez/.

32. Plez B. A new TiO line list. *Astron Astrophys.* 1998;337:495.

33. Heiter U, Barklem P, Fossati L, Kildiyarova R, Kochukhov O, Kupka F, et al. VALD—an atomic and molecular database for astrophysics. *J Phys Conf Ser.* 2008;10:012011.

34. Kupka F, Dubernet M-L. VAMDC as a Resource for Atomic and Molecular Data and the New Release of VALD. *Open Astron.* 2011;20:503.

35. Müller HS, Thorwirth S, Roth D, Winnewisser G. The Cologne database for molecular spectroscopy, CDMS. *Astron Astrophys.* 2001;370:L49.

36. Müller HS, Schlöfer F, Stutzki J, Winnewisser G. The Cologne Database for Molecular Spectroscopy, CDMS: a useful tool for astronomers and spectroscopists. *J Mol Struct.* 2005;742:215.

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

38. Pickett H, Poynter R, Cohen E, Delitsky M, Pearson J, Müller H. Submillimeter, millimeter, and microwave spectral line catalog. *J Quant Spectrosc Radiat Transf.* 1998;60:883.

39. Jet Propulsion Laboratory (JPL). Millimeter and submillimeter spectral line catalog. https://spec.jpl.nasa.gov/.
40. Lovas FJ. NIST recommended rest frequencies for observed interstellar molecular microwave transitions—2002 revision. *J Phys Chem Ref Data*. 2004;33:177.
41. Kobayashi K, Tsunekawa S, Matsushima F, Yamamoto S, Ohishi M, et al. Toyama microwave atlas. http://moodle.sci.u-toyama.ac.jp/atlas/.
42. Phillips J, Davis S, Eakin D. Berkeley newsletter, analysis of molecular spectra (1974–2000).
43. Brown JM, Brown JM, Carrington A. *Rotational spectroscopy of diatomic molecules*. Cambridge: Cambridge University Press; 2003.
44. Lefebvre-Brion H, Field RW. *The spectra and dynamics of diatomic molecules: revised and enlarged edition*. Amsterdam: Elsevier; 2004.
45. Western CM. PGOPHER, A Program for Simulating Rotational, Vibrational and Electronic Spectra. *J Quant Spectrosc Radiat Transf*. 2017;186:221.
46. Barklem PS, Collet R. Partition functions and equilibrium constants for diatomic molecules and atoms of astrophysical interest. *Astron Astrophys*. 2016;588:A96.
47. Lefebvre-Brion H. *Perturbations in the spectra of diatomic molecules*. Amsterdam: Elsevier; 2012.
48. Yurchenko SN, Al-Refaie AF, Tennyson J. EXOCROSS: a general program for generating spectra from molecular line lists. *Astron Astrophys*. 2018;614:A131.
49. Le Roy RJ. LEVEL: a computer program for solving the radial Schrödinger equation for bound and quasibound levels. *J Quant Spectrosc Radiat Transf*. 2017;186:167–78.
50. Yurchenko SN, Lodi L, Tennyson J, Stolyarov AV. DUO: a general program for calculating spectra of diatomic molecules. *Comput Phys Commun*. 2016;202:262–75.
51. Johnson RD III, NIST E. NIST standard reference database 69. 2013.
52. Linstrom PJ, Mallard WG. The NIST chemistry webbook: a chemical data resource on the internet. *J Chem Eng Data*. 2001;46:1059–63.
53. McKemmish LK, Syme A-M, Borovszky J, Yurchenko SN, Tennyson J, Furtenbacher T, et al. An update to the MARVEL data set and ExoMol line list for 12C2. *Mon Not R Astron Soc*. 2020;497:1081–97.

**How to cite this article:** McKemmish LK. Molecular diatomic spectroscopy data. *WIREs Comput Mol Sci*. 2021;11:e1520. [https://doi.org/10.1002/wcms.1520](https://doi.org/10.1002/wcms.1520)