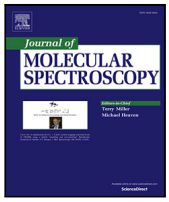
Journal of Molecular Spectroscopy 401 (2024) 111901

Contents lists available at ScienceDirect

Journal of Molecular Spectroscopy

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



Predicting the rotational dependence of line broadening using machine learning

Elizabeth R. Guest, Jonathan Tennyson *, Sergei N. Yurchenko

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

ARTICLE INFO

Dataset link: 10.5281/zenodo.10631728

Keywords:
Machine learning
Line broadening

ABSTRACT

Correct pressure broadening is essential for modelling radiative transfer in atmospheres, however data are lacking for the many exotic molecules expected in exoplanetary atmospheres. Here we explore modern machine learning methods to mass produce pressure broadening parameters for a large number of molecules in the ExoMol data base. To this end, state-of-the-art machine learning models are used to fit to existing, empirical air-broadening data from the HITRAN database. A computationally cheap method for large-scale production of pressure broadening parameters is developed, which is shown to be reasonably (69%) accurate for unseen active molecules. This method has been used to augment the previously insufficient ExoMol line broadening diet, providing air-broadening data for all ExoMol molecules, so that the ExoMol database has a full and more accurate treatment of line broadening. Suggestions are made for improved air-broadening parameters for species present in atmospheric databases.

1. Introduction

The characterisation and modelling of exoplanetary atmospheres require large volumes of laboratory spectroscopic data [1,2]. Simulations have demonstrated the need to deal correctly with line broadening in the atmospheres of exoplanets [3,4]. In general, exoplanetary atmospheric models are limited by insufficient data; particular areas where more information is needed include collisional broadening and line mixing parameters. Indeed, the lack of suitable collision broadening parameters is given as the number one requirement in a recent review of laboratory data needs to aid understanding exoplanetary atmospheres by Fortney et al. [5]. Current studies of hot atmospheres use at best qualitative estimates of pressure-broadening parameters for many molecules and molecular ions. Exoplanets have many potential compositions [6] and have been observed at wide ranges of temperatures [7] with hot planets orbiting close to their host stars providing the most reliable observations. Uncommon molecules on Earth are expected to be highly important for exoplanetary atmospheric processes, such as metal hydrides and oxides. In order to observe the huge expected variety of molecular species, a huge amount of spectroscopic data must be produced [8,9] to match the spectral features observed. This paper represents a first step towards meeting the pressure broadening portion of this need.

The necessity for this work is due of the sparsity of data for pressure broadening parameters as pressure broadening is unknown for the majority of collisional pairs. Many papers have laid out the need for more data for all molecule broadener pairs [3,5,10,11]. The importance of pressure broadening increases with more modern telescopes, as the impact of broadening scales with resolution. The differences in spectral cross sections due to pressure broadening have been shown to be significant [12–14]. For instance the James Webb Space Telescope (JWST), with resolution $R \sim 1000 - 3000$, will have errors in cross sections of up to 40% when pressure broadening data is missing [7]. Pressure broadening is also important because exoplanet spectra are optically thick. When spectral lines are fully saturated, line intensity alone is no longer sufficient to determine opacity. Pressure broadening is therefore an important component of the opacity in optically thick conditions. The importance of pressure broadening as a spectral parameter is therefore clear. As we show below, the need for improved treatment of line broadening is not restricted to exoplanets with air-broadening parameters; line broadening is poorly determined for some key molecules in our own atmosphere.

HITRAN [15] is a database of spectroscopic parameters, used for simulation of the transmission and emission of light in the Earth's atmosphere [16]. It is the major supplier of spectroscopy data for Earth-based studies. The high temperature extension HITEMP [17,18] is also aimed at terrestrial applications but serves a similar need as the ExoMol database [8]. Both HITRAN and HITEMP currently have well populated air broadening values for their databases. HITRAN has recently been expanding its list of perturbers to also include H_2, He, H_2O and CO_2

* Corresponding author.
E-mail address: j.tennyson@ucl.ac.uk (J. Tennyson).

https://doi.org/10.1016/j.jms.2024.111901
Received 31 October 2023; Received in revised form 10 February 2024; Accepted 16 March 2024
Available online 18 March 2024

broadening [19]. The extensive set of HITRAN air broadening data is used as the training data in our modelling.

The line profile used in this work is the Voigt profile. This is a profile that takes into account the two major sources of line broadening, pressure broadening and thermal Doppler broadening. While it is simple to provide Doppler-broadening parameters for a given molecule and temperature, the same is not true for collisional, or pressure, broadening, which are the focus of the current paper. Within the assumption of a Lorentz profile, collisional broadening is parameterised by γ, the half width half maximum (HWHM) of the Lorentzian line profile. γ is a function of temperature and the active and perturbing molecules; for the rotation–vibration spectra considered here, γ, in principle, also depends on the initial and final rotational and vibrational quantum numbers of each transition. Full dependence is difficult to properly characterise as there are relatively few data with full quantum number dependence provided.

Machine learning (ML) has been chosen to fill the gap in broadening data due to its relative simplicity. The goal of this work is to establish an efficient procedure for production of line broadening parameters with a special emphasis on completeness. Here we use ML techniques in conjunction with the line broadening data from HITRAN to predict pressure broadening by air parameters for arbitrary active molecules. The ExoMol database [8] is a large resource of spectroscopic data for atmospheric and astrophysical purposes. It contains line lists for almost 100 molecules (about 280 if isotopologues are counted) with a special emphasis on completeness at high temperatures. A lack of broadening data is a major problem for the many species in the ExoMol database. Laboratory measurements or theoretical estimates of γ for most of these species are unlikely to be produced soon. The accuracy of our ML predictions will be limited by the accuracy of training data, as well as by the assumptions made when choosing variables (features) to train on. Our work is the first application of ML for the production of the line shape parameters and one of the first for high-resolution spectroscopy applications in general. The ML methodology developed has been used to fill pressure broadening parameters by air in the ExoMol data base.

2. Methodology

In general, ML applications require using a dataset for training on, and an independent test set typically comprising 10% to 20% of the data. Here, since we only have 48 molecules to train on (see below), instead of rigidly splitting between a training and a test set cycle, we conducted a series of ML runs where training on 43 molecules is used to make predictions for the other 5. The series of runs was designed so that our model can be validated for each of the 48 molecules, using a training model that did not contain that molecule. We also investigated a similar procedure whereby 47 molecules were used for training, and 1 used for testing at a time. This did not noticeably improve results.

Our final operational model is trained on all 48 molecules which can be used to make predictions for molecules not included in the original dataset. The steps followed in our study are summarised in Fig. 1 and are described in turn below.

2.1. Input data

For our training data we use the air broadening data provided by the HITRAN database [15]. The HITRAN data comes from a collection of sources; empirical, *ab initio* and semi-empirical fits. Their data is designed primarily for Earth-based observations, with data for room temperature and pressure conditions. Every line in the HITRAN database has an assigned value for $\gamma \equiv \gamma_{\rm air}$, for air as the perturbing species, as well as $\gamma_{\rm self}$ for self-broadening. Air is taken as 80% N_2 and 20% O_2 in theoretical calculations. The predictions in this paper only concern the air broadening γ and will therefore be applicable also for room temperature and pressure gases, where air is the primary source of broadening. The temperature dependence of the (air) line broadening is described in HITRAN by the temperature exponent $n_{\rm air}$. This is not considered in the present study which is confined the HITRAN reference temperature of 296 K.

There are 55 species included in the HITRAN2020 database. We collected HITRAN's air-broadening data with the final operational model presented here using the data provided as of August 2023. In this work, 48 molecules were used for training and are listed in Table 1. O, SO_3, NO^+, NF_3, CF_4, $ClONO_2$ and SF_6 were not used for training. Here, O is not a molecule, while the other six species had very large and costly lists of transitions, and no accurate air-broadening data.

The availability of line broadening data in HITRAN2020 is summarised in Table 1. As shown in Table 1, HITRAN has air broadening data for all of its molecules, and only a few molecules have other broadening data. We have used only air broadening data, because there is a good quantity of data available for this perturber.

For almost all molecules in HITRAN's database, the air broadening values are the same for each isotopologue. There are two Cl containing molecules which have minor differences in broadening coefficient. For species with an overwhelmingly dominant isotopologue, we only used data from the main, dominant isotopologue. For molecules containing atoms with multiple common isotopes, notably Cl and Br, we took the broadening data for both isotopes. The only difference in treating these isotopologues was in the species mass.

2.1.1. Cleaning data

It was necessary to perform analysis of the data prior to training our models. Apart from the six molecules discussed above, only minimal amounts of data were excluded. Most non-parent isotopologues were excluded. Incomplete data (such as transitions where J was missing) was thrown out. Data where ΔJ was more than 2 were excluded, as it was assumed these largely correspond to data errors in the HITRAN database. This was a few hundred transitions total, a small minority of the data points, most of which were unphysical. We have queried the validity of these lines with HITRAN. The data for C_2H_6 with a reported error code of 4 ($\geq$10% and <20%) were also excluded due to the data being incorrect. This issue has since been reported to and remedied by HITRAN.

Data was replicated for molecules with fewer transitions, so that all molecules used had the same number of data points. For each molecule, the transitions were copied an integer number of times, until all have close to the same number of lines as HNO_3, the molecule with the most transitions. This was done as a way of weighting the data so that all molecules have equal impact. If this was not done, then the training data would be dominated by larger molecules with greater numbers of spectral lines in the HITRAN database, such as HNO_3 and SO_2. Our final set of HITRAN data used for training and testing is provided on zenodo.

2.1.2. Features of the data

In machine learning, the variables used to describe the data are known as features. The features of our data are the measurable properties of a molecule and its transitions. A subset of the properties given in the HITRAN database become features of our data. Based on standard statistical measures, those properties which are not useful for predicting line broadening are excluded. Other features from other sources are introduced, see below. As we only study air-broadening in this work, all features used are associated with the active molecule and not the perturber.

Many features were trialled for training. Using many features at once is detrimental to final model performance, so only the statistically most important features are retained in our final model.

In the course of this work we tested the following features (all quantum numbers refer to both upper and lower states):

- Features taken from the HITRAN database:
  - J, the total rotational quantum number;

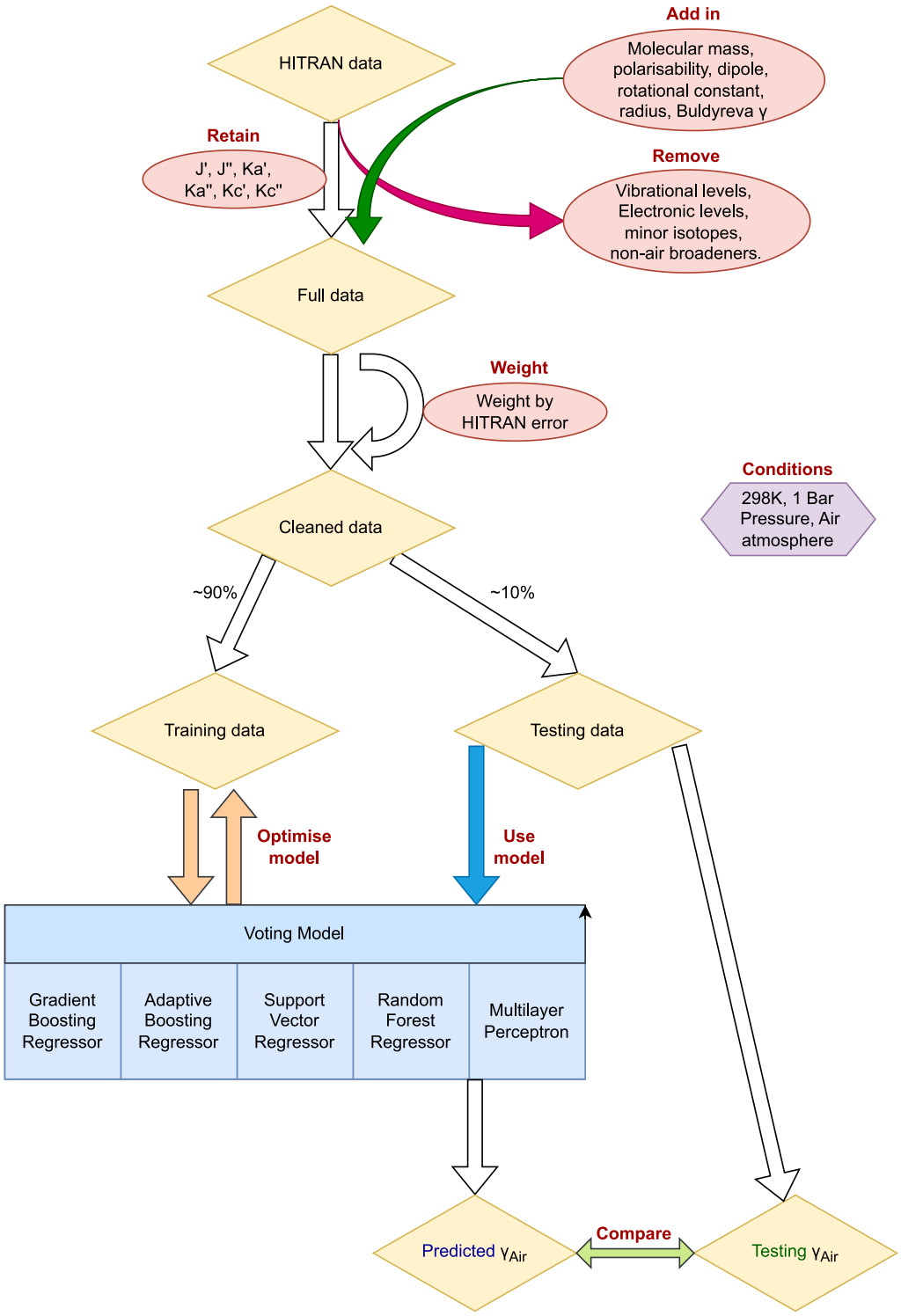


Fig. 1. Schematic representation of the pipeline used to produce our results.

- K, the projection of J on the z axis of spherical and symmetric top molecules;
- K_a and K_c, the projections of J along the axes of lowest and highest inertia, respectively. These quantum numbers are only valid for asymmetric top type molecules;
- $\tilde{\nu}$, transition wavenumber;
- S_w, line intensity;
- A, Einstein A coefficient;
- v_1, first vibrational quantum number. Tried only for diatomics;
- Λ, the projection of orbital angular momentum onto the principal axis. Tried only for open shell diatomics;
- S, the total spin quantum number. Tried only for open shell diatomics;
- Ω, the projection of total angular momentum onto the principal axis. Tried only for open shell diatomics;

Table 1
Proportion of lines in the HITRAN database with broadening parameters for the given perturbers.

Molecule	Total number of lines	Air	He	H_2	CO_2	H_2O	Self
H_2O	319 886	1	0	0	0	0	1
CH_3F	1 499	1	0	0	0	0	1
$COCl_2$	309 914	1	0	0	0	0	1
C_2N_2	71 775	1	0	0	0	0	1
H_2S	36 556	1	1	1	0	1	1
C_4H_2	251 245	1	0	0	0	0	1
HOBr	4 358	1	0	0	0	0	1
C_2H_6	63 516	1	0	0	0	0	1
CH_3I	178 247	1	0	0	0	0	1
H_2	3 480	1	0	0	0	0	1
CH_3Br	36 911	1	0	0	0	0	1
HO_2	38 804	1	0	0	0	0	1
HNO_3	950 864	1	0	0	0	0	1
NO	251 898	1	0	0	0	0	1
HCN	128 239	1	1	1	0	0	1
CH_3Cl	219 575	1	0	0	0	0	1
O_2	15 505	1	0	0	0	1	1
SO_2	549 425	1	1	1	1	0	1
CO	1 344	1	1	1	1	1	1
CS	1 088	1	0	0	0	0	1
CH_3OH	19 897	1	0	0	0	0	1
O_3	314 183	1	0	0	0	0	1
HCOOH	187 596	1	0	0	0	0	1
ClO	11 501	1	0	0	0	0	1
PH_3	104 759	1	1	1	0	0	1
HOCl	16 276	1	0	0	0	0	1
OCS	21 776	1	1	1	1	0	1
CH_4	309 863	1	0	0	0	1	1
C_2H_4	59 536	1	0	0	0	0	1
SO	42 916	1	0	0	0	0	1
H_2O_2	126 983	1	0	0	0	0	1
COF_2	168 793	1	0	0	0	0	1
N_2O	33 265	1	1	0	1	1	1
H_2CO	40 670	1	1	1	1	0	1
C_2H_2	74 335	1	1	1	1	0	1
HCl	17 800	1	1	1	1	0	1
HC_3N	248 273	1	0	0	0	0	1
NO_2	171 058	1	0	0	0	0	1
HF	8 090	1	1	1	1	0	1
CS_2	45 758	1	0	0	0	0	1
HI	3 161	1	0	0	0	0	1
N_2	1 107	1	0	0	0	0	1
HBr	6 070	1	0	0	0	0	1
OH	55 698	1	1	1	0	0	1
NH_3	76 605	1	1	1	1	1	1
CH_3CN	3 572	1	0	0	0	0	1
CO_2	174 446	1	1	1	0	1	1
GeH_4	60 878	1	0	0	0	0	1

- Calculated features from other sources:
  - Approximate K_a and K_c which treat all molecules as asymmetric rotors with both K_a and K_c used as rotational quantum numbers. K values for different rotor types were merged. To make a consistent 'K' feature across all rotor types, some approximations were made as described in Table 2;
  - m, rotational quantum number, which distinguishes the branches, calculated by:
    - O-branch: $m = -J''$
    - P-branch: $m = -J''$
    - Q-branch: $m = J''$
    - R-branch: $m = J'' + 1$
    - S-branch: $m = J'' + 1$
  - Molecular mass;
  - Molecular polarisability, with the data sources described in Table 3;

Table 2
Mapping of K values to give consistent features across rotor types.

Rotor type	$K_a^{(\text{approx})}$	$K_c^{(\text{approx})}$
Asymmetric	K_a	K_c
Prolate symmetric	K	$J - K$
Oblate symmetric	$J - K$	K
Spherical	$J/2$	$J/2$
Linear	0	J

  - Permanent molecular dipole moment, data sources described in Table 3;
  - Rotor type: symmetric linear, asymmetric linear, symmetric top, asymmetric top, and spherical top;
  - A, B and C rotational constants for asymmetric molecules, data sources described in Table 3. For symmetric systems, values are copied across all constants. For linear systems, B and C are the same and A is assumed to be ∞, approximated at 100 000. This is done to have consistent features across rotor types;
  - R_{VdW}, kinetic (Van der Waals) diameter of active molecule, data sources described in Table 3;
  - m_m, order of the molecules' leading multipole moment, described in [20];
  - γ_B predicted by the formula of Buldyreva et al. [20], see below;
  - O, P, Q, R and S branch numbers. The branch of the transition of each data point would have its number recorded, and the other 4 features would have NaN values;
  - Open/not-open shell molecular type, labelled 1 or 0.
  - ω_e, the harmonic wavenumber. Tried only for diatomics;
  - $\omega_e\chi_e$, the first anharmonic vibrational constant. Tried only for diatomics;
  - mass ratio of atoms, $\frac{M_L}{M_h}$, the mass of the lighter atom over the mass of the heavier atom. Tried only for diatomics;
  - Mean bond length, tried only for diatomics.

The final list of 16 features used to describe the data is given by: J', J'', m, $K_a''^{(\text{approx})}$, $K_c''^{(\text{approx})}$, $K_a'^{(\text{approx})}$, $K_c'^{(\text{approx})}$, molecular mass, molecular dipole, molecular polarisability, rotational constants A, B, and C; multiple moment order m_m, Van der Waals radius R_{VdW}, and finally the approximate line broadening parameter γ_B.

m was chosen as the principal quantum number to use and plot; in practice these plots are almost entirely symmetric about $m = 0$ which means that effectively the key parameter is J''; γ is well-known to be highly J-dependent. Broadening is high for low rotational speeds, and tends to a constant for high rotations. Various K values also have a large effect on γ [21,22], so these quantum numbers were used in the final model. It was assumed that γ was minimally dependant on vibrational state, and for diatomics it was seen that this feature made little difference. Line characteristics, such as $\tilde{\nu}$ and S_w, were also seen to have little importance.

Mass, polarisibility, and rotational constant were seen to be important for describing the differences been molecules. Labelling transitions types or rotor types was seen to make little difference to predictions, perhaps because this information is already contained in the rotational constants.

γ_B has been calculated using a formula derived by Tsao et al. [23], expanded on in [20]. The papers give a prediction for γ for a pair of molecules, which depends on a small number of input variables. Their equation provides insight into some of the descriptors which are important for determining γ, such as m_m and the Van der Waals radius. This value has no J-dependence. γ_B improved our model's predictions, but it was not one of the most important features. γ_B is compared against in our results, demonstrating our improvement for predicting air-broadening

Table 3
Sources of data for features. NIST refers to their CCCBDB (Computational Chemistry Comparison and Benchmark DataBase) [24]; experimental data were chosen where possible. NIST - e is experimental data, NIST - se is semi-empirical calculated data, NIST - HF is data calculated using the Hartree–Fock method, and NIST - ga is data calculated using a group additivity method. JPL [25] and CDMS [26] are complimentary databases, both providing information on molecular transitions. QDB is the Quantemol DataBase [27]. Unknown Kinetic Diameters were estimated by comparison to the size of similar molecules. A full list of diameters used is included in the supplementary information.

Molecule	Dipole	Polarisibility	Rotational constants	Kinetic diameter
$COCl_2$	NIST - e	NIST - e	NIST - e	Estimated
C_2N_2	NIST - e	NIST - e	NIST - e	Estimated
H_2S	NIST - e	NIST - e	NIST - e	[28]
C_4H_2	NIST - e	NIST - e	NIST - e	Estimated
HOBr	NIST - e	NIST - HF	NIST - e	Estimated
CH_3I	NIST - e	NIST - e	NIST - e	Estimated
CH_3Br	NIST - e	NIST - e	NIST - e	Estimated
HNO_3	NIST - e	NIST - se	NIST - e	Estimated
HCN	NIST - e	NIST - e	NIST - e	[29]
CH_3Cl	NIST - e	NIST - e	NIST - e	Estimated
HCOOH	NIST - e	NIST - e	NIST - e	[30]
PH_3	NIST - e	NIST - e	NIST - e	Estimated
HOCl	NIST - e	NIST - ga	NIST - e	Estimated
SO	NIST - e	NIST - ga	NIST - e	Estimated
HC_3N	NIST - e	NIST - se	NIST - e	Estimated
CS_2	NIST - e	NIST - e	NIST - e	Estimated
HI	NIST - e	NIST - e	NIST - e	Estimated
CH_3CN	NIST - e	NIST - e	NIST - e	Estimated
H_2O	NIST - e	NIST - e	JPL	[31]
CH_3F	NIST - e	NIST - e	NIST - e	QDB
C_2H_6	NIST - e	NIST - e	NIST - e	[31]
H_2	NIST - e	NIST - e	NIST - e	[32]
HO_2	JPL	[33]	JPL	[31]
NO	NIST - e	NIST - e	JPL	[32]
O_2	NIST - e	NIST - e	CDMS	[32]
SO_2	NIST - e	CDMS	NIST - e	QDB
CO	JPL	NIST - e	JPL	[32]
CS	JPL	NIST - se	JPL	QDB
CH_3OH	JPL	NIST - e	JPL	[31]
O_3	JPL	NIST - e	JPL	QDB
ClO	JPL	[34]	JPL	QDB
OCS	JPL	NIST - e	JPL	QDB
CH_4	NIST - e	NIST - e	NIST - e	[32]
C_2H_4	NIST - e	NIST - e	NIST - e	[32]
H_2O_2	JPL	NIST - se	JPL	[31]
COF_2	JPL	[35]	JPL	QDB
N_2O	JPL	NIST - e	JPL	QDB
H_2CO	JPL	NIST - e	JPL	[31]
C_2H_2	NIST - e	NIST - e	JPL	[32]
HCl	JPL	NIST - e	JPL	QDB
NO_2	JPL	NIST - e	JPL	QDB
HF	JPL	NIST - e	JPL	QDB
N_2	NIST - e	NIST - e	JPL	[32]
HBr	NIST - e	NIST - e	JPL	QDB
OH	JPL	NIST - HF	JPL	[36]
NH_3	JPL	NIST - e	JPL	QDB
GeH_4	NIST - e	NIST - e	NIST - se	Estimated
CO_2	NIST - e	NIST - e	NIST - e	[32]

2.1.3. Weighting data

The stated accuracy of line broadening in the HITRAN database is highly variable. Approximately half of the data has a known uncertainty, often based on experimental determinations. For the rest of the data, HITRAN provides estimates both for the data and associated uncertainty. For example, for some molecules there are no known air-broadening parameters and HITRAN simply provides very rough estimates or a single γ value as a placeholder. All line broadening data in HITRAN are provided with uncertainty estimates using an error code from 0 to 8 as summarised in Table 4. This allows the data to be filtered by the accuracy of broadening parameters. We used the weighting described in this table to weight each of our data points. This was so that the poor data is disregarded, but without throwing it out entirely. The uncertainties of the HITRAN data were used as error bars to validate the fitting of our data. Not all models had an inbuilt weighting function, in this case data was replicated proportionally to reproduce the weighting assigned in Table 4.

Table 4
Weight assigned to our training data based on HITRAN error code.

HITRAN error code	Uncertainty	Weighting
0	Unreported or unavailable	$(1/500\,000)^2$
1	Default or constant	$(1/20\,000)^2$
2	Average or estimate	$(1/1000)^2$
3	≥20%	$(1/50)^2$
4	≥10% and <20%	$(1/15)^2$
5	≥5% and <10%	$(1/10)^2$
6	≥2% and <5%	$(1/10)^2$
7	≥1% and <2%	$(1/10)^2$
8	<1%	$(1/10)^2$

2.2. Machine learning process

An overview of ML principles used here is given by Sarker [37]. Various machine learning tools have been compared in this work to find optimum results. The models used here are provided by the scikit-learn python package [38].

Feature scaling is an important method used in machine learning, to normalise the features of the data. This allows them to be treated equally when the input data is on different scales. Pre-built models such as those given by scikit-learn often assume unit variance for gradient descent algorithms, and so using this scale on the data improves the speed of training. For this reason, the input data was scaled to zero mean and unit variance using the scikit-learn StandardScaler tool. This makes different features consistent for inclusion into models.

As discussed above, the data was spit into a testing and a training set. Data points (transitions) common to each molecule were kept together. The molecules in our data were randomly split into two sets, 43 used for training and 5 retained for testing the model. The data was split along these lines, so that every time a model was trained, there would be data for 5 molecules that could be used to score the prediction.

This training procedure was run independently ten times on a different set of testing and training molecules. This led to there being ten almost identical trained models to analyse every time a model architecture was created. The difference between these ten was the molecules used in its training and testing set. As a limited set of molecules was present, this procedure meant that a machine learning model architecture could be judged based off how it made predictions for all molecules in the HITRAN database. For all plots in this paper, the molecule of interest was in the testing set when the model was being trained. This means that all results for molecules consider the said molecule as unseen. Careful checks were done to ensure that no data from the molecule being tested was included in the training set.

Once split, the data was shuffled to remove any bias in the ML runs on the order in which the data was provided.

The 'final operational model' available on zenodo is one last training run performed on all HITRAN data, with the optimised model parameters and hyperparameters described here. This can be used to make predictions for molecules currently outside of the HITRAN database.

2.2.1. Machine learning models

Ensemble methods have been shown to be effective tools in machine learning [39]. These form the basis of much of the trialled models. All models trialled were regressors. The models trialled were Gradient Boosting [40], Adaptive Boosting [41], Random Forest Regression [42], Decision Tree Regression [43], Support Vector Regression with an RBF (radial basis function) kernel [44] (SVR), a Stochastic Gradient Descent linear model [45] (SGD), Multi-Layer Perceptron [46] (MLP), Voting Regressor [47,48], and a Dummy Regressor as a baseline.

Gradient boosting, adaptive boosting, random forest and voting models are all examples of ensemble methods. They combine predictions of weaker models, giving strong results. A decision tree is a tree shaped model, separating data into categories using if-else type rules. Support vector machines optimise boundaries between data types. In regression, the boundary is the prediction. Using a kernel allows non-linear boundaries. Stochastic gradient descent is a method for finding an optimal linear fit to data. We do not expect a linear model to fit our data well. An MLP is a basic neural network, made up of multiple layers of fully connected neurons, each having a nonlinear activation function. A dummy regressor is a baseline model, which simply returns the mean of the training data.

The most effective model was made using a Voting Regressor. This is a meta-estimator which averages a set of base models. The base models used are gradient boosting (scikit-learn's HistGradientBoostingRegressor), adaptive boosting (scikit-learn's AdaBoostRegressor), support vector regression (scikit-learn's SVR), random forest (scikit learn's RandomForestRegressor) and a multi-layer perceptron (scikit learn's MLPRegressor). These models were all used equally as estimators in scikit-learn's VotingRegressor, referred to our voting model below.

2.2.2. Hyperparameters of the model

The fixed properties of a machine learning model are known as model hyperparameters. These are things like the learning rate, or the number of neurons used in an MLP. The variable properties of a model are known as model parameters. These are things like the coefficients in linear regression, or the weighting of neurons in an MLP. The ML algorithm alters its parameters until an optimally trained line broadening model is found, within the defined model hyperparameters. This line broadening model is a function of the features that characterise molecules and their transitions.

Various hyperparameters of each ML model were trialled to find the best description of line broadening. Scikit-learn's GridSearchCV and RandomisedSearchCV were used to systematically search the parameter space for each type of model, and score them by the accuracy of their predicted γ. The different model hyperparameters were ranked based on their scores, and the best ones retained in the final voting model.

The default model parameters given by scikit-learn were determined to be optimal in most cases. The random forest had a few altered model hyperparameters; we used 'criterion' = mse, 'minimum weight fraction of leaf' = 0.001 and 'number of estimators' = 10. The MLP also had a few specific model hyperparameters; 'α' = 0.01, 'hidden layer sizes' = (30, 30), 'learning rate' = adaptive, 'number of iterations with no change before stopping' = 1, 'tolerance' = $1e^{-5}$. The voting regressor used 'number of jobs' = −1 to optimise parallelisation of training.

2.2.3. Scoring data

We compare the scores using two metrics, root mean square error (RMSE), and R^2. RMSE is calculated the standard way, and presented as a percentage error, to match the errors in the training data provided by HITRAN. The RMSEs were calculated compared to HITRAN's more accurate data, datapoints with error codes of 3 and above. The R^2 score, referred to as 'score' in our results, shows the success of the model. It is defined as $(1 - \frac{u}{v})$, where u is the residual sum of squares, given by $\sum(y_{\text{true}} - y_{\text{pred}})^2$, and v is the total sum of squares, given by $\sum(y_{\text{true}} - \text{mean}(y_{\text{true}}))^2$. The optimal score allowed is 1, and bad scores can be any negative number. This definition is all as is provided by [38].

3. Results

Various baseline ML models were trialled. Those using random forest type models were observed to be good, both because they produced stable curves, and for speed of computation. This supported our extended use of them in the voter model finally chosen. The MLP, which is a neural network type model, often scored the best, however it was more prone to overfitting. This was seen with high J predictions of γ, which did not tend to a constant as indicated by theory. Optimisation of the α and $tolerance$ hyperparameters of the MLP was done to minimise this. Using a final model that combines the predictions of several sub-models was optimal for making accurate, but not overfitted, predictions.

Even though not all components of the final model chosen are effective predictors on their own, it was observed that the 5 final components chosen, when combined, gave the best final score, with the best looking m-dependence.

In our figures we compare our predicted m-dependence and hence given their symmetry J-dependence, with that in HITRAN for each molecule. It should be noted that for molecules where γ depends on the vibrational or K quantum numbers, there will be multiple values of γ for each m. Fig. 2 shows how well various models replicate the m-dependence of line broadening, justifying the final model chosen.

Both Table 5 and the plots demonstrate that the voter model is the best of these considered. This is why the voter model was chosen as our final model type; all results presented below use this model.

We compare our results to those molecules in HITRAN with accurate line broadening data in Figs. 3 and 4. This is a validation of the accuracy of machine learning predictions. For molecules where HITRAN is the most accurate we do not always replicate their accuracy, shown in Fig. 3. For other key HITRAN molecules we match γ well, shown in Fig. 4.

In Fig. 5 we compare our results to the line broadening data from HITRAN for molecules with only a small amount of accurate line broadening data. In these cases, our results lie within the HITRAN error bars. This shows that machine learning predictions are state-of-the-art for the majority of molecules for which there is little experimental line broadening.

Table 6 compares the accuracy of predictions made for different types of molecule.

The average RMSE of the final tested voter model was 25.8%. The average uncertainty in the HITRAN data was at 12%, when only their accurate data is included. The average uncertainty of all HITRAN data can be estimated at 33%. This shows that, while we have not matched HITRAN's uncertainties on their most accurate data, we provide reasonable estimates for unseen molecules.

Table 5
Comparison of models created using a reduced dataset, for ease of computation. Models were trained iteratively, so that all molecules were included in the testing dataset once. RMSE's were calculated for all molecules individually, and the final RMSE given was calculated by averaging the RMSE of all molecules. The semi empirical model of Buldyreva *et al.* is included for comparison.

Model	RMSE/%
Decision Tree	35.9
Random Forest	32.3
Gradient Boosting	38.0
Adaptive Boosting	32.9
SVR	40.5
SGD	36.2
Dummy Regressor	42.4
MLP	27.9
Voting Regressor	29.3
Buldyreva Model	49.0

Table 6
Comparison of the number of HITRAN molecules for each type of rotor, and the mean percentage RMSE for each.

Rotor type	Number of molecules	RMSE/%
Asymmetric	17	15.4
Prolate symmetric	6	19.2
Oblate symmetric	2	23.6
Spherical	1	43.0
Linear	22	30.3

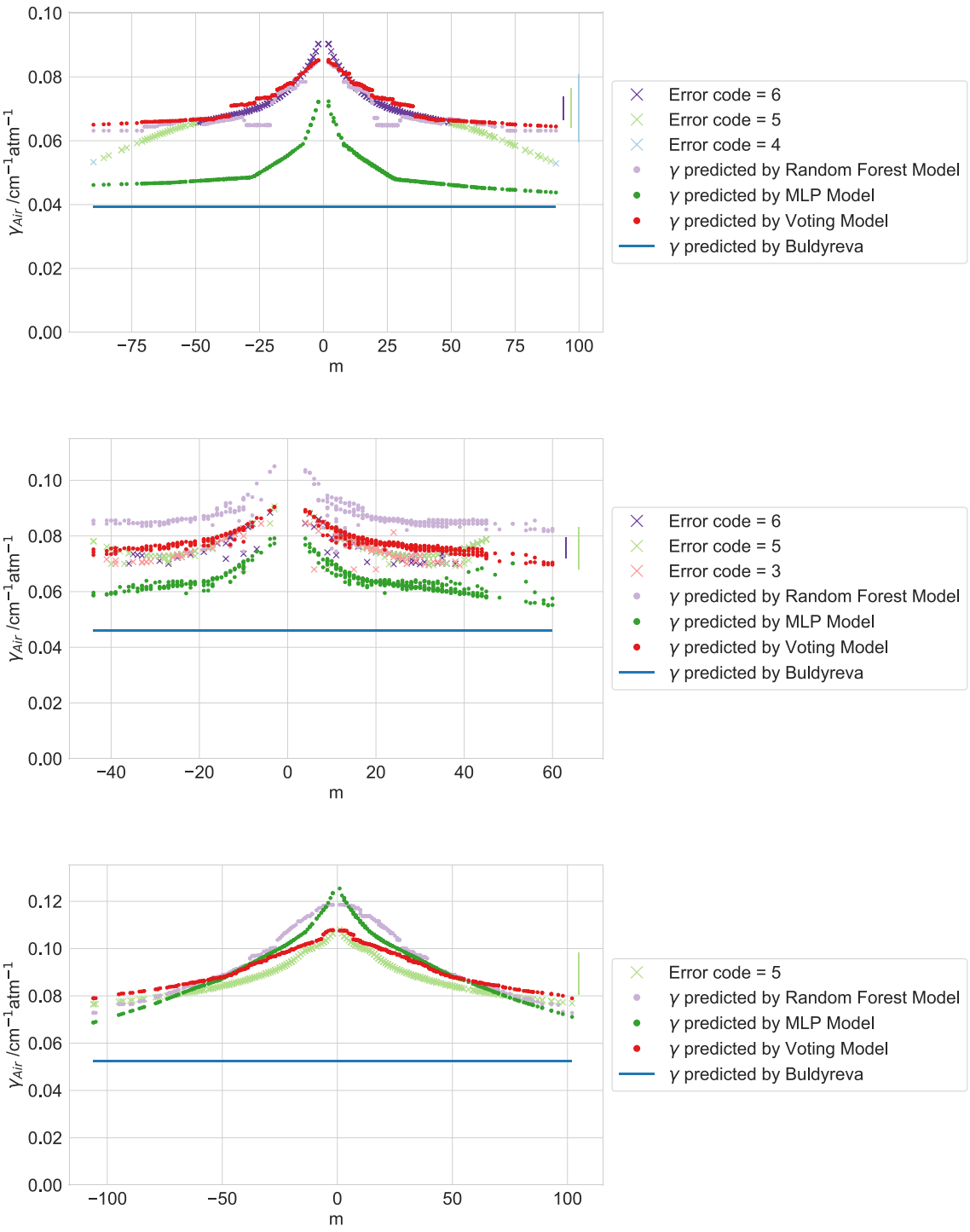


Fig. 2. Comparison of γ_{air} predictions from various models to known HITRAN data for CO_2 (top), C_2H_6 (middle) and OCS (bottom). The HITRAN data is given by crosses labelled by their error codes; the uncertainty of each error code is given in Table 4. The semi empirical model of Buldyreva et al. is included for comparison.

Calculating these scores on a line-by-line basis, it is seen that 69% of line broadening predicted falls within HITRAN's error bars. This is deemed good. Too much matching to imperfect data would imply overfitting.

4. Discussion

In the following, results for a few key molecules are discussed; a full set of comparative figures are given in the supplementary data. In general, the results obtained are quite good. 69% of the predicted data lies within the uncertainty ranges of the training data. This is validation that unknown molecules can have air broadening parameters predicted.

It is difficult to score our results consistently. One reason is that a good model for some molecules may make others worse. Good RMSE or scores may mean overfitting, which can sometimes be seen when bumpy curves are predicted. As the ExoMol project looks at high temperature atmospheres, predictions at high J values are the priority. The stable results at high J demonstrated by our model are therefore necessary, even at the expense of poorer fitting at low J.

Some data in the HITRAN database is estimated, as very little experimental data is available. For molecules like H_2O_2, HITRAN has

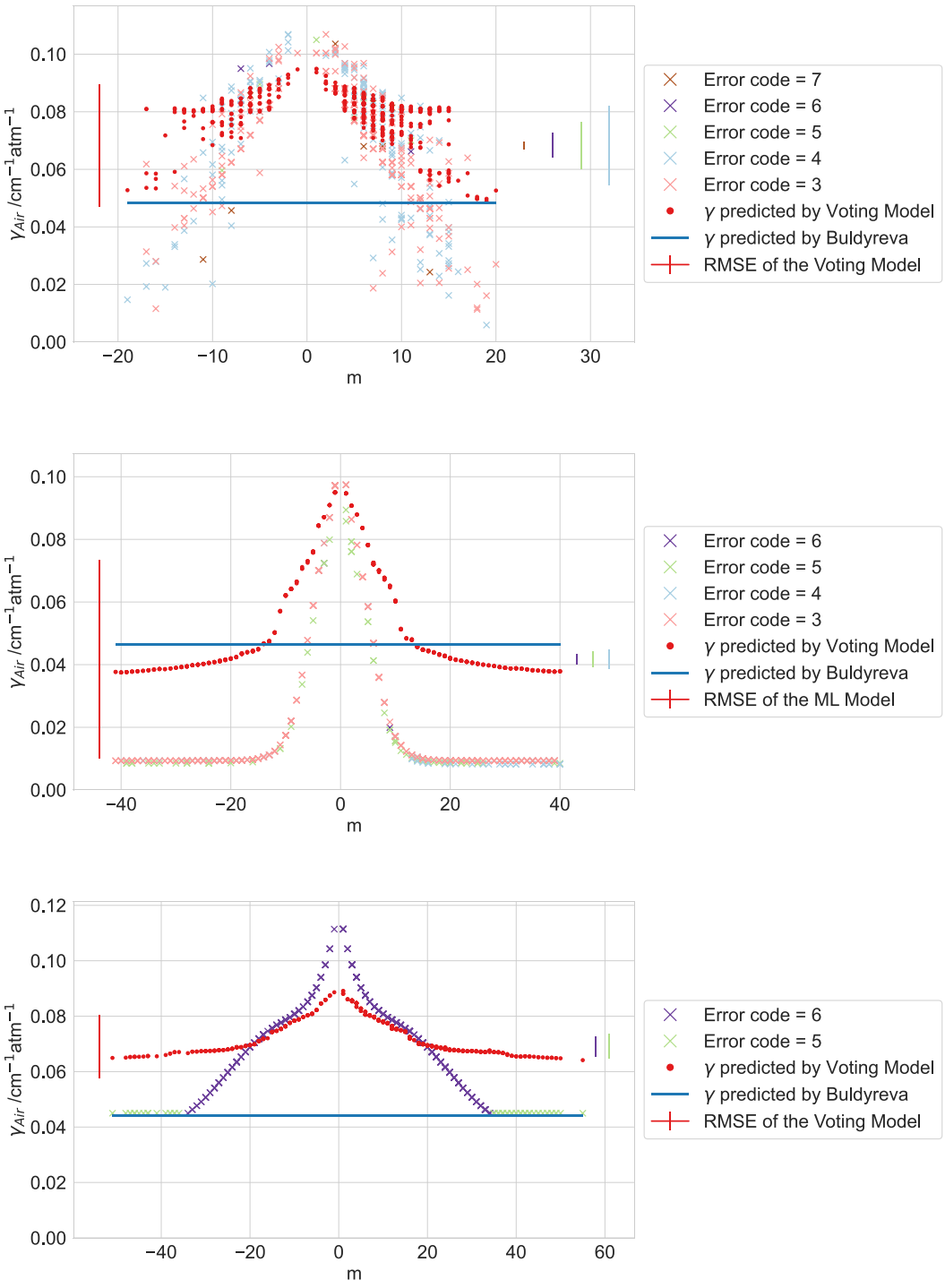


Fig. 3. Comparison of γ_{air} predictions from the final tested model, to some of the best HITRAN data, for H_2O (top), HCl (middle) and C_2H_2 (bottom). The HITRAN data is given by crosses labelled by their error codes; the uncertainty of each error code is given in Table 4. The semi empirical model of Buldyreva et al. is included for comparison.

provided a single number guess for line broadening, 0.1 cm^{-1} atm^{-1}. In Fig. 6 we show our predictions for these molecules, as an example of the use of our model. Our results show a good shape for the J dependence with reasonable overall estimates for the magnitude of γ. HONO is a future molecule of interest for HITRAN, for which our model makes reasonable estimates, shown in Fig. 7; we are not aware of any air-broadening data for this molecule. This is an example of predictions made by our model that could be used to update databases such as HITRAN and ExoMol.

As a use case for our work, we show some typical ExoMol molecules here, SO, HCN and SO_2, see Fig. 8. There is little accurate data to compare our predictions to, however, we consider our predictions to be reasonable. Particularly interesting is the case of SO; the HITRAN air-broadening data for SO is a copy of that for O_2. While these species have similar open-shell electronic structures, something we found not important when testing features, SO has a dipole while O_2 does not. One would expect the SO's dipole to lead to increased broadening and as can be seen in Fig. 8, our predicted γ's are significantly larger than

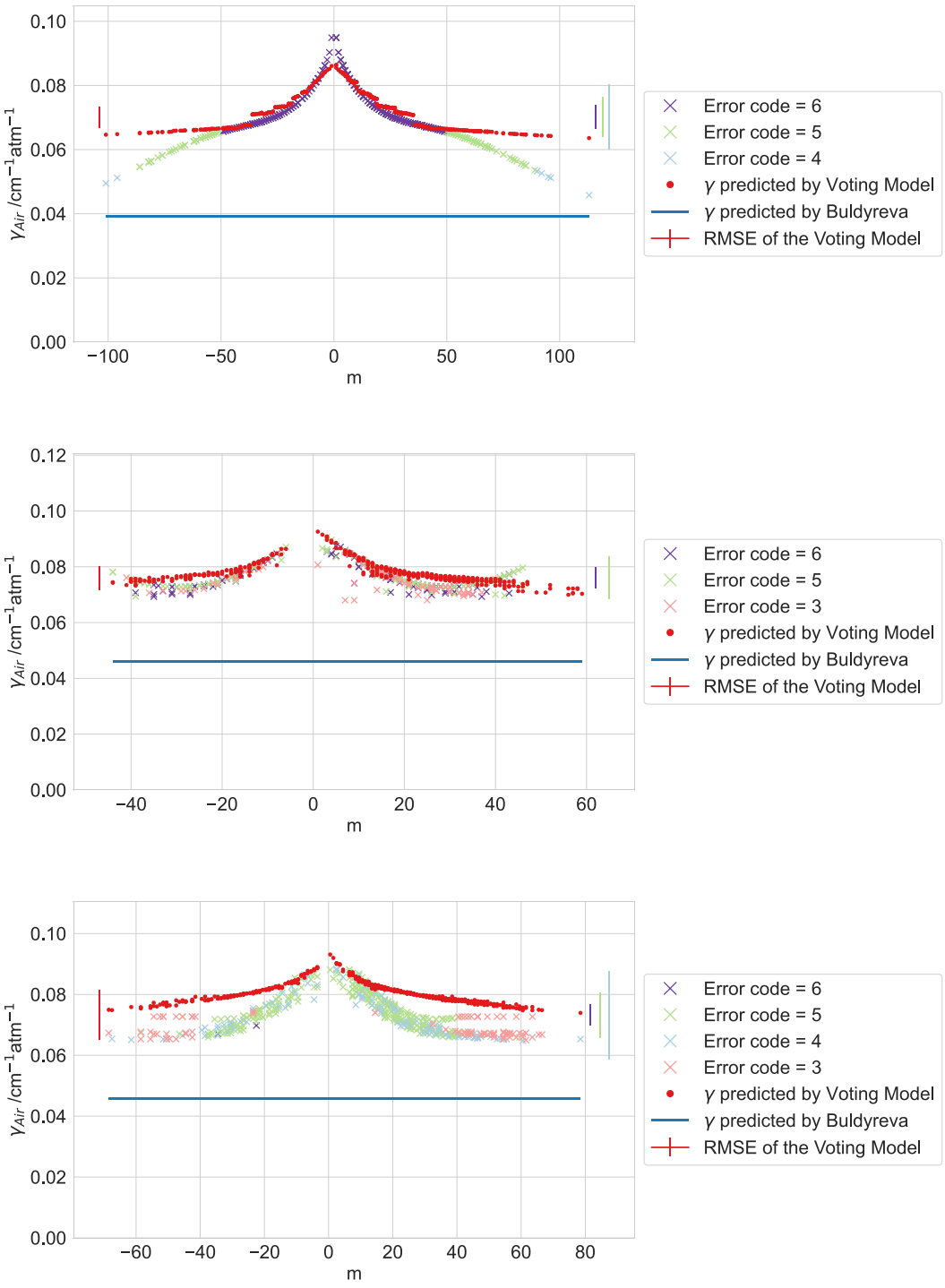


Fig. 4. Comparison of γ_{air} predictions from the final tested model, where our results match HITRAN's data very well, for CO_2 (top), C_2H_6 (middle) and NO_2 (bottom). The HITRAN data is given by crosses labelled by their error codes; the uncertainty of each error code is given in Table 4. The semi empirical model of Buldyreva et al. is included for comparison.

those currently used by HITRAN. We suggest that our predictions are likely to be more accurate.

5. Conclusions and outlook

Here we use machine learning to train a model to reproduce the effects of air-broadening on molecular transitions at 296 K and atmospheric pressure using data from the HITRAN data base. The model successfully predicts about 69% of the data provided by HITRAN within HITRAN uncertainties. For a significant number of molecules the HITRAN data is actually highly uncertain: γ being unavailable, fixed as a constant or estimated. Our model can be used to predict broadening parameters for these species and we suggest in most cases our predictions are likely to be more reliable than the data currently estimated, in the absence of any specific measurements or theoretical calculations. Our model is able to predict pressure broadening parameters within

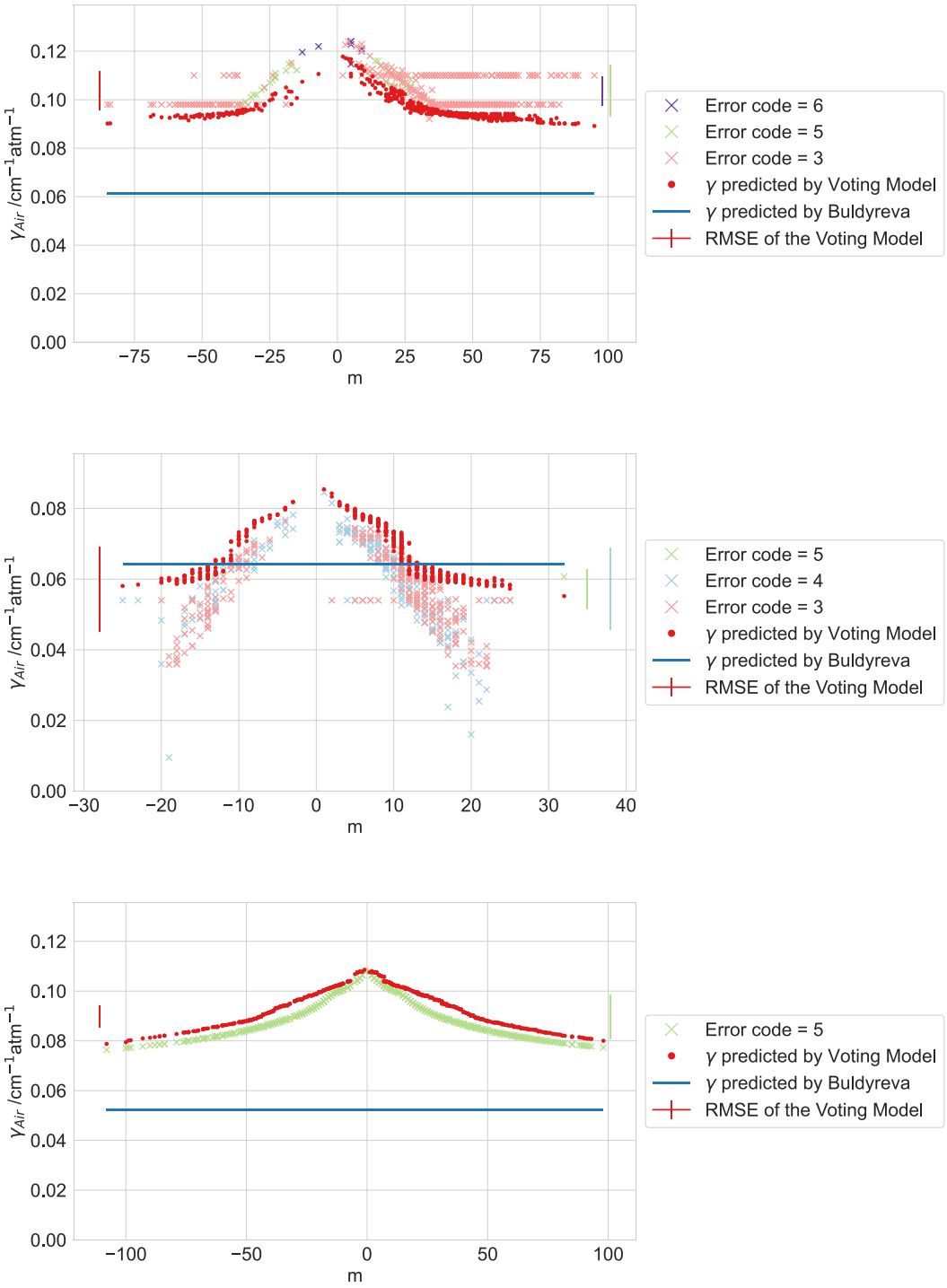


Fig. 5. Comparison of γ_{air} predictions from the final tested model for more unusual HITRAN molecules, for HNO_3 (top), PH_3 (middle) and OCS (bottom). The HITRAN data is given by crosses labelled by their error codes; the uncertainty of each error code is given in Table 4. The semi empirical model of Buldyreva et al. is included for comparison.

error bars 69% of the time. This is in line with our expectation of the accuracy of such predictions. Matching the training data with 100% accuracy would imply an overfitting of the data which are not perfect. Broadening parameters generated with our model will be used to update the rather crude ExoMol pressure-broadening diet [13] and thus populate the database with these important data.

This work represents a first step into field of machine learning of broadening parameters. There are clearly a number of future directions all of which we plan to explore. Using our model to predict air-broadening γ's for molecules not contained in the HITRAN database is the most straightforward extension and the results can compared with predictions made using semi-empirical approximations [20,49]. For studies of exoplanet atmosphere broadening parameters are required over very extended temperature ranges. The data on high temperature broadening is limited but there are theoretical constructs, eg. [50,51], which should facilitate the extension of the data to higher temperatures. More challenging is the need to include broadeners other than air.

Being able to estimate pressure broadening for any molecule would be of great use to exoplanet modellers. There is a fair amount of broadening data in the literature where air is not the broadener. HITRAN has

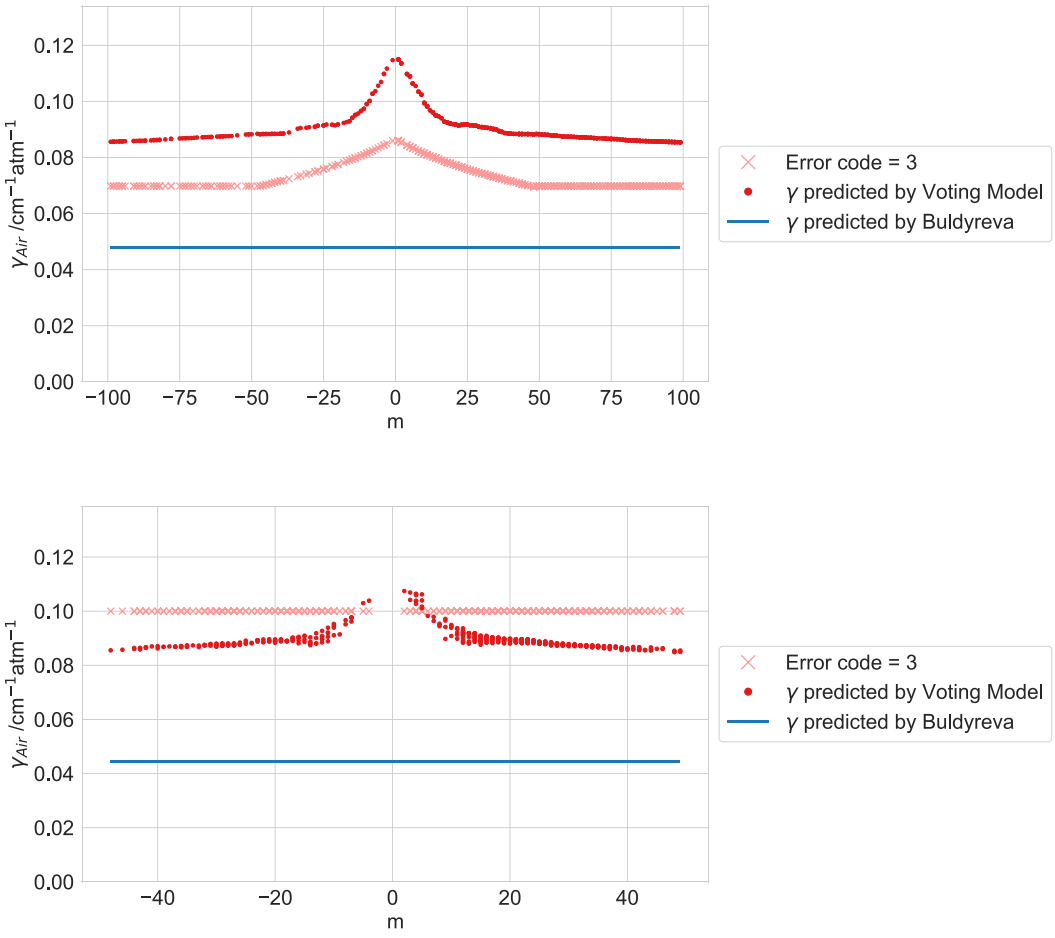


Fig. 6. Comparison of γ_{Air} predictions from the final tested model, to HITRAN estimates based on no empirical data, for CS (top) and H_2O_2 (bottom). The HITRAN data is given by crosses labelled by their error codes; the uncertainty of each error code is given in Table 4. The semi empirical model of Buldyreva et al. is included for comparison.

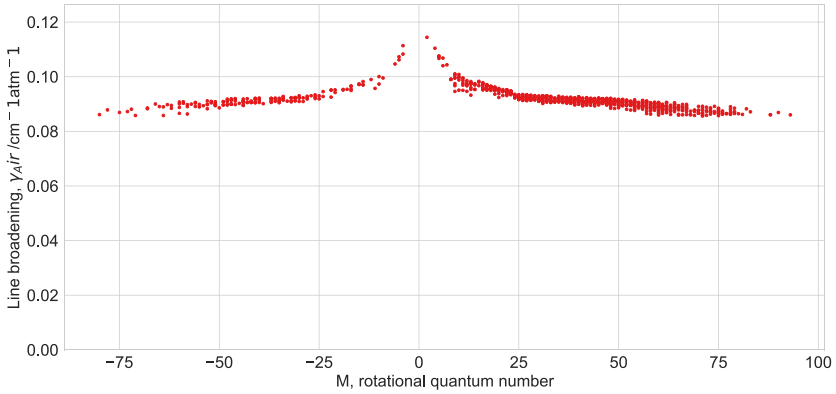


Fig. 7. Predictions of γ_{Air} for HONO, a molecule without much empirical data.

pressure broadening parameters for some active molecules perturbed by H_2, He, H_2O and CO_2. Other sources of data include other databases which have some broadening data, such as GEISA [52], HITEMP [18] and ExoMol [8]. There are also a lot of experimental measurements in the literature, largely giving accurate data for a few lines or a particular spectral region. Theoretically, Gamache and co-workers have computed comprehensive, temperature dependent list of broadening parameters for systems such as CO_2 broadened by water [53] and water broadened by H_2 [54]. These various studies could be combined to form a suitable training set to look at the effect of different broadening species.

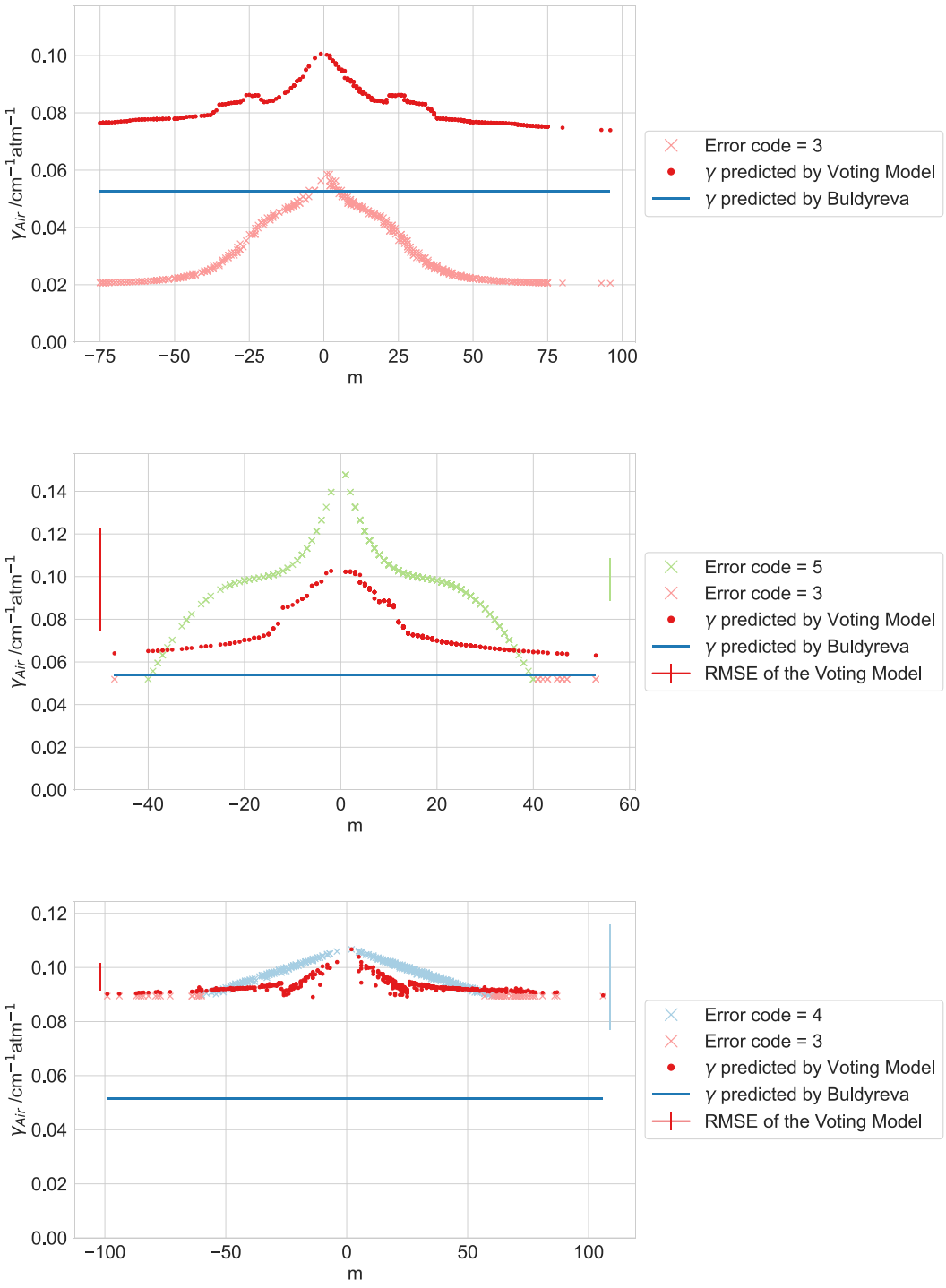


Fig. 8. Comparison of γ_{Air} predictions from the final tested model, to the HITRAN data for SO (top), HCN (middle) and SO_2 (bottom). The HITRAN data is given by crosses labelled by their error codes; the uncertainty of each error code is given in Table 4. The semi empirical model of Buldyreva et al. is included for comparison.

Considering the intrinsic difficulty in obtaining line broadening parameters experimentally, we believe that ML has the potential to be a powerful complementary tool.

The air broadening data described in this paper are available on the ExoMol website, for all exotic species in the ExoMol database.

CRediT authorship contribution statement

Elizabeth R. Guest: Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Jonathan Tennyson:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Sergei N. Yurchenko:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jonathan Tennyson reports financial support was provided by European Research Council. Elizabeth Guest reports administrative support, article publishing charges, and travel were provided by Science and

Technology Facilities Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The training data, available for the model trained on all inputs for predicting air-broadening for new molecules are available on the ExoMol zenodo area via doi 10.5281/zenodo.10631728.

Acknowledgements

This work was supported by ERC Advanced Investigator Project 883830 (ExoMolHD) in collaboration with the UCL Centre for Doctoral Training in Data Intensive Science, funded by the STFC training grant reference ST/P006736/1.

Appendix A. Supplementary data

Plots of γ versus J comparing our predictions with HITRAN for all 48 molecules are given as supplementary material. A full list of kinetic diameters used in this work is also included.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jms.2024.111901.

References

[1] J. Tennyson, S.N. Yurchenko, The ExoMol atlas of molecular opacities, Atoms 6 (2018) 26, http://dx.doi.org/10.3390/atoms6020026.

[2] K.L. Chubb, M. Rocchetto, S.N. Yurchenko, M. Min, I. Waldmann, J.K. Barstow, P. Molliere, A.F. Al-Refaie, M. Phillips, J. Tennyson, The ExoMolOP database: Cross-sections and K-tables for molecules of interest in high-temperature exoplanet atmospheres, Astron. Astrophys. 646 (2021) A21, http://dx.doi.org/10.1051/0004-6361/202038350.

[3] G. Tinetti, J. Tennyson, C.A. Griffiths, I. Waldmann, Water in exoplanets, Philos. Trans. R. Soc. Lond. Ser. A 370 (2012) 2749–2764, http://dx.doi.org/10.1098/rsta.2011.0338.

[4] L. Anisman, K.L. Chubb, Q. Changeat, B. Edwards, S.N. Yurchenko, J. Tennyson, G. Tinetti, Cross-sections for heavy atmospheres: H_2O self-broadening, J. Quant. Spectrosc. Radiat. Transfer 283 (2022) 108146, http://dx.doi.org/10.1016/j.jqsrt.2022.108146.

[5] J. Fortney, T.D. Robinson, S. Domagal-Goldman, A.D.D. Genio, I.E. Gordon, E. Gharib-Nezhad, N. Lewis, C. Sousa-Silva, V. Airapetian, B. Drouin, R.J. Hargreaves, X. Huang, T. Karman, R.M. Ramirez, G.B. Rieker, J. Tennyson, R. Wordsworth, S.N. Yurchenko, A.V. Johnson, T.J. Lee, M.S. Marley, C. Dong, S. Kane, M. López-Morales, T. Fauchez, T. Lee, K. Sung, N. Haghighipour, S. Horst, P. Gao, D.-y. Kao, C. Dressing, R. Lupu, D.W. Savin, B. Fleury, O. Venot, D. Ascenzi, S. Milam, H. Linnartz, M. Gudipati, G. Gronoff, F. Salama, L. Gavilan, J. Bouwman, M. Turbet, Y. Benilan, B. Henderson, N. Batalha, R. Jensen-Clem, T. Lyons, R. Freedman, E. Schwieterman, J. Goyal, L. Mancini, P. Irwin, J.-M. Desert, K. Molaverdikhani, J. Gizis, J. Taylor, J. Lothringer, R. Pierrehumbert, R. Zellem, N. Batalha, S. Rugheimer, J. Lustig-Yaeger, R. Hu, E. Kempton, G. Arney, M. Line, M. Alam, J. Moses, N. Iro, L. Kreidberg, J. Blecic, T. Louden, P. Mollière, K. Stevenson, M. Swain, K. Bott, N. Madhusudhan, J. Krissansen-Totton, D. Deming, I. Kitiashvili, E. Shkolnik, Z. Rustamkulov, L. Rogers, L. Close, The need for laboratory measurements and ab initio studies to aid understanding of exoplanetary atmospheres, in: Astro2020: Decadal Survey on Astronomy and Astrophysics, 2019, http://dx.doi.org/10.48550/arXiv.1905.07064, arXiv:1905.07064.

[6] N.M. Batalha, Exploring exoplanet populations with NASA's kepler mission, Proc. Natl. Acad. Sci. 111 (2014) 12647–12654, http://dx.doi.org/10.1073/pnas.1304196111.

[7] C. Hedges, N. Madhusudhan, Effect of pressure broadening on molecular absorption cross sections in exoplanetary atmospheres, Mon. Not. R. Astron. Soc. 458 (2016) 1427–1449, http://dx.doi.org/10.1093/mnras/stw278.

[8] J. Tennyson, S.N. Yurchenko, A.F. Al-Refaie, V.H.J. Clark, K.L. Chubb, E.K. Conway, A. Dewan, M.N. Gorman, C. Hill, A.E. Lynas-Gray, T. Mellor, L.K. McKemmish, A. Owens, O.L. Polyansky, M. Semenov, W. Somogyi, G. Tinetti, A. Upadhyay, I. Waldmann, Y. Wang, S. Wright, O.P. Yurchenko, The 2020 release of the ExoMol database: Molecular line lists for exoplanet and other hot atmospheres, J. Quant. Spectrosc. Radiat. Transfer 255 (2020) 107228, http://dx.doi.org/10.1016/j.jqsrt.2020.107228.

[9] J.C. Zapata Trujillo, M.M. Pettyjohn, L.K. McKemmish, High-throughput quantum chemistry: empowering the search for molecular candidates behind unknown spectral signatures in exoplanetary atmospheres, Mon. Not. R. Astron. Soc. 524 (2023) 361–376, http://dx.doi.org/10.1093/mnras/stad1717.

[10] J.J. Fortney, T.D. Robinson, S. Domagal-Goldman, D.S. Amundsen, M. Brogi, M. Claire, D. Crisp, E. Hebrard, H. Imanaka, R. de Kok, M.S. Marley, D. Teal, T. Barman, P. Bernath, A. Burrows, D. Charbonneau, R.S. Freedman, D. Gelino, C. Helling, K. Heng, A.G. Jensen, S. Kane, E.M.R. Kempton, R.K. Kopparapu, N.K. Lewis, M. Lopez-Morales, J. Lyons, W. Lyra, V. Meadows, J. Moses, R. Pierrehumbert, O. Venot, S.X. Wang, J.T. Wright, The need for laboratory work to aid in the understanding of exoplanetary atmospheres, 2016, http://dx.doi.org/10.48550/ARXIV.1602.06305.

[11] R.S. Freedman, M.S. Marley, K. Lodders, Line and mean opacities for ultracool dwarfs and extrasolar planets, Astrophys. J. Suppl. 174 (2008) 504–513, http://dx.doi.org/10.1086/521793.

[12] E. Gharib-Nezhad, M.R. Line, The influence of H_2O pressure broadening in high-metallicity exoplanet atmospheres, Astrophys. J. 872 (2019) 27, http://dx.doi.org/10.3847/1538-4357/aafb7b.

[13] E.J. Barton, C. Hill, M. Czurylo, H.-Y. Li, A. Hyslop, S.N. Yurchenko, J. Tennyson, The ExoMol diet: H_2 and He line-broadening parameters, J. Quant. Spectrosc. Radiat. Transfer 203 (2017) 490–495, http://dx.doi.org/10.1016/j.jqsrt.2017.01.028.

[14] P. Niraula, J. de Wit, I.E. Gordon, R.J. Hargreaves, C. Sousa-Silva, R.V. Kochanov, The impending opacity challenge in exoplanet atmospheric characterization, Nat. Astron. 6 (2022) 1287–1295, http://dx.doi.org/10.1038/s41550-022-01773-1.

[15] I.E. Gordon, L.S. Rothman, R.J. Hargreaves, R. Hashemi, E.V. Karlovets, F.M. Skinner, E.K. Conway, C. Hill, R.V. Kochanov, Y. Tan, P. Wcisło, A.A. Finenko, K. Nelson, P.F. Bernath, M. Birk, V. Boudon, A. Campargue, K.V. Chance, A. Coustenis, B.J. Drouin, J. Flaud, R.R. Gamache, J.T. Hodges, D. Jacquemart, E.J. Mlawer, A.V. Nikitin, V.I. Perevalov, M. Rotger, J. Tennyson, G.C. Toon, H. Tran, V.G. Tyuterev, E.M. Adkins, A. Baker, A. Barbe, E. Canè, A.G. Császár, A. Dudaryonok, O. Egorov, A.J. Fleisher, H. Fleurbaey, A. Foltynowicz, T. Furtenbacher, J.J. Harrison, J. Hartmann, V. Horneman, X. Huang, T. Karman, J. Karns, S. Kassi, I. Kleiner, V. Kofman, F. Kwabia-Tchana, N.N. Lavrentieva, T.J. Lee, D.A. Long, A.A. Lukashevskaya, O.M. Lyulin, V.Y. Makhnev, W. Matt, S.T. Massie, M. Melosso, S.N. Mikhailenko, D. Mondelain, H.S.P. Müller, O.V. Naumenko, A. Perrin, O.L. Polyansky, E. Raddaoui, P.L. Raston, Z.D. Reed, M. Rey, C. Richard, R. Tóbiás, I. Sadiek, D.W. Schwenke, E. Starikova, K. Sung, F. Tamassia, S.A. Tashkun, J. Vander Auwera, I.A. Vasilenko, A.A. Vigasin, G.L. Villanueva, B. Vispoel, G. Wagner, A. Yachmenev, S.N. Yurchenko, The hitran2020 molecular spectroscopic database, J. Quant. Spectrosc. Radiat. Transfer 277 (2022) 107949, http://dx.doi.org/10.1016/j.jqsrt.2021.107949.

[16] I.E. Gordon, L.S. Rothman, Hitran, 2015, URL https://hitran.org/.

[17] L.S. Rothman, I.E. Gordon, R.J. Barber, H. Dothe, R.R. Gamache, A. Goldman, V.I. Perevalov, S.A. Tashkun, J. Tennyson, Hitemp, the high-temperature molecular spectroscopic database, J. Quant. Spectrosc. Radiat. Transfer 111 (2010) 2139–2150, http://dx.doi.org/10.1016/j.jqsrt.2010.05.001.

[18] R.J. Hargreaves, I.E. Gordon, L.S. Rothman, S.A. Tashkun, V.I. Perevalov, A.A. Lukashevskaya, S.N. Yurchenko, J. Tennyson, H.S.P. Müller, Spectroscopic line parameters of NO, NO_2, and N_2O for the HITEMP database, J. Quant. Spectrosc. Radiat. Transfer 232 (2019) 35–53, http://dx.doi.org/10.1016/j.jqsrt.2019.04.040.

[19] Y. Tan, F.M. Skinner, S. Samuels, R.J. Hargreaves, R. Hashemi, I.E. Gordon, H_2, he, and CO_2 pressure-induced parameters for the HITRAN database. II. Line lists of CO_2, N_2O, CO, SO_2, OH, OCS, H_2CO, HCN, PH_3, H_2S, and GeH_4, Astrophys. J. Suppl. 262 (2022) 40, http://dx.doi.org/10.3847/1538-4365/ac83a6.

[20] J. Buldyreva, S.N. Yurchenko, J. Tennyson, Simple semi-classical model of pressure-broadened infrared/microwave linewidths in the temperature range 200–3000 K, RAS Tech. Instrum. 1 (2022) 43–47, http://dx.doi.org/10.1093/rasti/rzac004.

[21] B.A. Voronin, N.N. Lavrentieva, T.P. Mishina, T. Y.Chesnokova, M.J. Barber, J. Tennyson, Estimate of the $J'J$ dependence of water vapor line broadening parameters, J. Quant. Spectrosc. Radiat. Transfer 111 (2010) 2308–2314.

[22] Q. Ma, R.H. Tipping, N.N. Lavrentieva, Pair identity and smooth variation rules applicable for the spectroscopic parameters of H_2O transitions involving high-J states, Mol. Phys. 109 (2011) 1925–1941, http://dx.doi.org/10.1080/00268976.2011.599343.

[23] C.J. Tsao, B. Curnutte, Line-widths of pressure-broadened spectral lines, J. Quant. Spectrosc. Radiat. Transfer 2 (1962) 41–91, http://dx.doi.org/10.1016/0022-4073(62)90013-4.

[24] R. Johnson, Nist 101. Computational chemistry comparison and benchmark database (1999-11-01 1999).

[25] H.M. Pickett, R.L. Poynter, E.A. Cohen, M.L. Delitsky, J.C. Pearson, H.S.P. Müller, Submillimeter, millimeter, and microwave spectral line catalog, J. Quant. Spectrosc. Radiat. Transfer 60 (1998) 883–890.

[26] C.P. Endres, S. Schlemmer, P. Schilke, J. Stutzki, H.S.P. Müller, The cologne database for molecular spectroscopy, CDMS, in the virtual atomic and molecular data centre, VAMDC, J. Mol. Spectrosc. 327 (2016) 95–104, http://dx.doi.org/10.1016/j.jms.2016.03.005.

[27] J. Tennyson, S. Mohr, M. Hanicinec, A. Dzarasova, C. Smith, S. Waddington, B. Liu, L.L. Alves, K. Bartschat, A. Bogaerts, S.U. Engelmann, T. Gans, A.R. Gibson, S. Hamaguchi, K.R. Hamilton, C. Hill, D. O'Connell, S. Rauf, K. van 't Veer, O. Zatsarinny, The 2021 release of the Quantemol database (QDB) of plasma chemistries and reactions, PSST 31 (2022).

[28] G. Liu, A. Cadiau, Y. Liu, K. Adil, V. Chernikova, I.-D. Carja, Y. Belmabkhout, M. Karunakaran, O. Shekhah, C. Zhang, A.K. Itta, S. Yi, M. Eddaoudi, W.J. Koros, Enabling fluorinated MOF-based membranes for simultaneous removal of H_2S and CO_2 from natural gas, Angew. Chem. Int. Ed. 57 (2018) 14811–14816, http://dx.doi.org/10.1002/anie.201808991.

[29] C.A. Scholes, Hydrogen cyanide recovery by membrane gas separation, Chem. Eng. J. 386 (2020) 124049, http://dx.doi.org/10.1016/j.cej.2020.124049.

[30] J. Jae, G.A. Tompsett, A.J. Foster, K.D. Hammond, S.M. Auerbach, R.F. Lobo, G.W. Huber, Investigation into the shape selectivity of zeolite catalysts for biomass conversion, J. Catal. 279 (2011) 257–268, http://dx.doi.org/10.1016/j.jcat.2011.01.019.

[31] A.W. Jasper, J.A. Miller, Lennard–jones parameters for combustion and chemical kinetics modeling from full-dimensional intermolecular potentials, Combust. Flame 161 (2014) 101–110, http://dx.doi.org/10.1016/j.combustflame.2013.08.004.

[32] R.W. Baker, Other Membrane Processes, John Wiley & Sons, Ltd, 2012, p. 557, http://dx.doi.org/10.1002/9781118359686.ch13, (Chapter 13).

[33] A. Sharipov, B. Loukhovitski, A.V. Pelevkin, Diffusion coefficients of electronically excited molecules, Phys. Chem. Kinet. Gas Dyn. 22 (2021) 913, http://dx.doi.org/10.33257/PhChGD.22.1.913.

[34] S. Du, J.S. Francisco, G.K. Schenter, B.C. Garrett, Ab initio and analytical intermolecular potential for ClO–H_2O, J. Chem. Phys. 126 (2007) 114304, http://dx.doi.org/10.1063/1.2566537.

[35] Csid:9246, chemspider, 2007, URL http://www.chemspider.com/Chemical-Structure.9246.html.

[36] A. Khayar, J. Bonamy, Calculation of mean collision cross sections of free radical oh with foreign gases, J. Quant. Spectrosc. Radiat. Transfer 28 (1982) 199–212, http://dx.doi.org/10.1016/0022-4073(82)90023-1.

[37] I. Sarker, Machine learning: Algorithms, real-world applications and research directions, SN Comput. Sci. 2 (2021) 160, http://dx.doi.org/10.1007/s42979-021-00592-x.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[39] O. Sagi, L. Rokach, Ensemble learning: A survey, WIREs Data Min. Knowl. Discov. 8 (2018) e1249, http://dx.doi.org/10.1002/widm.1249.

[40] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Statist. 29 (2001) 1189–1232, http://dx.doi.org/10.1214/aos/1013203451.

[41] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. System Sci. 55 (1997) 119–139, http://dx.doi.org/10.1006/jcss.1997.1504.

[42] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, http://dx.doi.org/10.1023/A:1010950718922.

[43] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, Taylor & Francis, 1984.

[44] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27, http://dx.doi.org/10.1145/1961189.1961199.

[45] J. Kiefer, J. Wolfowitz, Stochastic estimation of the maximum of a regression function, Ann. Math. Stat. 23 (1952) 462–466, http://dx.doi.org/10.1214/aoms/1177729392.

[46] G.E. Hinton, Connectionist learning procedures, Artificial Intelligence 40 (1989) 185–234.

[47] K. An, J. Meng, Voting-averaged combination method for regressor ensemble, in: D.-S. Huang, Z. Zhao, V. Bevilacqua, J.C. Figueroa (Eds.), Advanced Intelligent Computing Theories and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 540–546.

[48] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (1992) 241–259, http://dx.doi.org/10.1016/S0893-6080(05)80023-1.

[49] J. Buldyreva, R.P. Brady, S.N. Yurchenko, J. Tennyson, Collisional broadening of molecular rovibronic lines, J. Quant. Spectrosc. Radiat. Transfer 313 (2024) 108843, http://dx.doi.org/10.1016/j.jqsrt.2023.108843.

[50] R.R. Gamache, B. Vispoel, On the temperature dependence of half-widths and line shifts for molecular transitions in the microwave and infrared regions, J. Quant. Spectrosc. Radiat. Transfer 217 (2018) 440–452, http://dx.doi.org/10.1016/j.jqsrt.2018.05.019.

[51] N. Stolarczyk, F. Thibault, H. Cybulski, H. Jozwiak, G. Kowzan, B. Vispoel, I.E. Gordon, L.S. Rothman, R.R. Gamache, P. Wcislo, Evaluation of different parameterizations of temperature dependences of the line-shape parameters based on *ab initio* calculations: Case study for the HITRAN database, J. Quant. Spectrosc. Radiat. Transfer 240 (2020) 106676, http://dx.doi.org/10.1016/j.jqsrt.2019.106676.

[52] T. Delahaye, R. Armante, N. Scott, N. Jacquinet-Husson, A. Chédin, L. Crépeau, C. Crevoisier, V. Douet, A. Perrin, A. Barbe, V. Boudon, A. Campargue, L. Coudert, V. Ebert, J.-M. Flaud, R. Gamache, D. Jacquemart, A. Jolly, F. Kwabia Tchana, A. Kyuberis, G. Li, O. Lyulin, L. Manceron, S. Mikhailenko, N. Moazzen-Ahmadi, H. Müller, O. Naumenko, A. Nikitin, V. Perevalov, C. Richard, E. Starikova, S. Tashkun, V. Tyuterev, J. Vander Auwera, B. Vispoel, A. Yachmenev, S. Yurchenko, The 2020 edition of the GEISA spectroscopic database, J. Mol. Spectrosc. (2021) 111510, http://dx.doi.org/10.1016/j.jms.2021.111510.

[53] L. Regalia, E. Cousin, R.R. Gamache, B. Vispoel, S. Robert, X. Thomas, Laboratory measurements and calculations of line shape parameters of the H_2O-CO_2 collision system, J. Quant. Spectrosc. Radiat. Transfer 231 (2019) 126–135, http://dx.doi.org/10.1016/j.jqsrt.2019.04.012.

[54] R.R. Gamache, B. Vispoel, C.L. Renaud, K. Cleghorn, L. Hartmann, Vibrational dependence, temperature dependence, and prediction of line shape parameters for the H_2O-H_2 collision system, Icarus 326 (2019) 186–196, http://dx.doi.org/10.1016/j.icarus.2019.02.011.