PCCP

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Phys. Chem. Chem. Phys.*, 2025, **27**, 23000

Neural network ensemble for computing cross sections of rotational transitions in $H_2O + H_2O$ collisions

Bikramaditya Mandal, [a] Dmitri Babikov, [b] Phillip C. Stancil, [c] Robert C. Forrey, [d] Roman V. Krems [e] and Naduvalath Balakrishnan *[a]

Received 23rd July 2025,
Accepted 25th September 2025

DOI: 10.1039/d5cp02812d

rsc.li/pccp

Water (H_2O) is one of the most abundant molecules in the universe and is found in a wide variety of astrophysical environments. Rotational transitions in $H_2O + H_2O$ collisions are important for modeling environments rich in water molecules but they are computationally intractable using quantum mechanical methods. Here, we present a machine learning (ML) tool using an ensemble of neural networks (NNs) to predict cross sections to construct a database of rate coefficients for rotationally inelastic transitions in collisions of complex molecules such as water. The proposed methodology utilizes data computed with a mixed quantum-classical theory (MQCT). We illustrate that efficient ML models using NNs can be built to accurately interpolate in the space of 12 quantum numbers for rotational transitions in two asymmetric top molecules, spanning both initial and final states. We examine various architectures of data corresponding to each collision energy, symmetry of water molecules, and excitation/de-excitation rotational transitions, and optimize the training/validation data sets. Using only about 10% of the computed data for training, the NNs predict cross sections of state-to-state rotational transitions in $H_2O + H_2O$ collisions with an average relative root mean squared error of 0.409. Thermally averaged cross sections, computed using the predicted state-to-state cross sections (~90%) and the data used for training and validation (~10%), were compared against those obtained entirely from MQCT calculations. The agreement is found to be excellent with an average percent deviation of about ~13.5%. The methodology is robust, and thus applicable to other complex molecular systems.

1 Introduction

Several state-of-the-art observatories, such as the Atacama Large Millimeter Array (ALMA) radio telescope, and other space-exploration telescopes, including the Spitzer space telescope and, most recently, the James Webb space telescope (JWST), have been deployed to collect spectroscopic data. The spectra obtained from these telescopes show signatures of water molecules with isotopic constitutions, namely H_2O and HDO. Observation and analysis of water isotopologues in cometary comae, in relation to their abundances on Earth and other solar system bodies, can yield valuable insights into the early history of Earth and, by extension, the solar system. Water is also found in a large variety of astrophysical environments. For example, water is detected in cometary comae,[1,2] cold interstellar molecular clouds,[3] stellar photospheres and circumstellar envelopes,[3] and atmospheres of icy planets.[4] Water represents the major reservoir of oxygen and, thus, controls the chemistry of many species in the gas phase and also on grain surfaces.[3,5] In star-forming regions, water emission dominates the process of gas cooling.[6] This process creates the brightest line in the radio frequency or maser-like radiation in the GHz range,[7] which carries information about physical conditions in these environments. The low temperature under prestellar core conditions leads to freezing of most volatile compounds onto surfaces of grains.[8] The variability of the D/H ratio in different molecules, particularly in water, can yield essential details on its formation conditions in our solar system.

[a] Department of Chemistry and Biochemistry, University of Nevada, Las Vegas, Nevada 89154, USA. E-mail: naduvala@unlv.nevada.edu
[b] Department of Chemistry, Marquette University, Milwaukee, WI 53233, USA
[c] Department of Physics and Astronomy and the Center for Simulational Physics, University of Georgia, Athens, GA 30602, USA
[d] Department of Physics, Penn State University, Berks Campus, Reading, PA 19610, USA
[e] Department of Chemistry, University of British Columbia, Vancouver, BC V6T 1Z1, Canada

Atmospheres of icy planets and their moons, such as Jovian moons, are known to have an anisotropic distribution of water vapor, affecting the properties of the observed water line. While most of such atmosphere is collision-less, the sub-solar point supports intense sublimation and photoinduced desorption, which results in a distribution that is not in local thermodynamic equilibrium (non-LTE) and driven by molecular collisions, such as excitation and de-excitation of H_2O. Interpreting these and many other observations requires numerical modeling and relies on the knowledge of precise excitation and quenching schemes for *ortho*- and *para*-H_2O. Large uncertainties of rate coefficients for these transitions can affect the predictions of astrophysical models by orders of magnitude.[9,10] To characterize emission as a function of coma radius, modeling with radiation transfer codes, such as RADEX,[11] LIME,[12] or MOLPOP,[13] is necessary, which in turn requires collision rate coefficients as an input. To understand and correlate with the observed rotational spectra from ALMA or JWST missions, one requires state-to-state collisional rate coefficients for the rotational excitation and quenching processes. These rate coefficients are difficult to calculate for complex collision systems such as $H_2O + H_2O$ and $HDO + H_2O$. Astrophysical models also require inelastic scattering rate coefficients for a range of other complex collision systems, including $CH_3OH + CO$, $H_2O + HCN$, $H_2CO + CO$, and $H_2O + CH_3OH$.[14]

Databases such as BASECOL[15] and LAMDA[16] have been developed to simplify the process of obtaining rate coefficients for different molecular systems. The rate coefficients can be computed by a quantum mechanical treatment of the collision problem as implemented in a few codes available to the scientific community, such as MOLSCAT,[17,18] HIBRIDON,[19] and TwoBC[20] or using a mixed quantum/classical theory (MQCT),[21] when quantum calculations are not practical.

The inelastic collisions of H_2O with H_2O are almost impossible to study using fully quantum methods because the water molecule has a dense spectrum of quantum states.[22–24] However, significant progress has been made recently by Mandal *et al.* to study rotationally inelastic collisions of two water molecules using MQCT.[25–27] MQCT has proven its ability to produce accurate results with computational efficiency for inelastic collisions of several molecular systems.[21,23,25–39] In this approach, the relative translational motion of collision partners is treated classically using the mean-field trajectories method, while rotations and vibrations (*i.e.*, internal degrees of freedom of the colliding molecules) are treated quantum mechanically.[21] In the current implementation of the MQCT method for $H_2O + H_2O$ collisions, the H_2O molecules are treated as rigid rotors.[21]

Using the MQCT methodology, a database of thermally averaged cross sections (TACSs) (averaged over a thermal population of rotational levels of the partner H_2O molecule) was first published by Mandal and Babikov[26] followed by a database of thermal rate coefficients.[27] This database of TACSs and the rate coefficients contained 231 transitions in *para*-H_2O and 210 transitions in *ortho*-H_2O (both treated as the target molecule) in the temperature range of $5 \leq T \leq 1000$ K. In a subsequent study, a database of both rotational temperature ($T_{\rm rot}$) and kinetic temperature ($T_{\rm kin}$) dependent rate coefficients was built to model non-LTE environments using RADEX for $H_2O + H_2O$ collisions.[25]

While MQCT can be applied to collisions of two water molecules, the computational complexity remains challenging. As elaborated by Mandal and Babikov,[26] the computation involves evaluation of matrix elements of the interaction potential in a basis of rotational wave functions of the water molecules that are required for the simulation of mixed quantum/classical trajectories. In the prior work, the computation of these matrices alone required about $\sim$2.7 M CPU hours in the HPC facility Raj at Marquette University (AMD Rome 2 GHz processors, memory 512 GB). Additionally, the total cost of the scattering calculations (trajectory simulations) for six collision energies was about 5.25 M CPU hours using the same HPC facility. Altogether, the cost of the MQCT calculations of rate coefficients for $H_2O + H_2O$ collisions was nearly 8 million CPU hours. More importantly, several months of human work were needed to manage the ongoing simulations and carry out numerous post-processing analyses to convert state-to-state cross sections to the rate coefficients to be deposited into the databases. While significant speedup in the computation of the relevant coupling matrices has been achieved recently, the trajectory simulations remain computationally demanding.

The challenge of such massive computational tasks is twofold. First, computing the TACSs for 231 transitions in *para*- and 210 transitions in *ortho*-H_2O required generating over a million cross sections for individual state-to-state transitions at each collision energy. Independent calculations for a total of 3268 initial states combining both the target and quencher molecule needed to be completed to build the database. Secondly, each of these simulations for individual initial states and collision energies required computation of four large interaction potential matrices, each containing about 1.35 million coupling elements. With these two factors, building an extended database for such a complex system using direct calculations is often not practical. Machine learning may instead prove useful for constructing such a complex database.

Machine learning (ML) has found widespread applications in recent years in many areas of physics and chemistry, including condensed matter physics,[40–42] nuclear physics,[43–45] astronomy,[46–48] particle physics,[49–51] quantum many-body physics,[52–54] cosmology[55–57] and fitting of multi-dimensional potential energy surfaces (PESs) from electronic structure calculations. Permutationally invariant polynomials (PIPs) combined with neural networks (NNs) by Bowman, Guo, and others[58–69] and Gaussian process regression (GPR) by Krems and coworkers[70–75] have been widely adopted for building PESs of complex molecular systems. ML has also been used to predict rate coefficients for inelastic collisions of diatomic molecules for astrophysical modeling from a smaller set of available data and improve the accuracy of approximate quantum scattering calculations.[47,75–80] Quantum machine learning is also being actively explored.[81–87]

To our knowledge, ML has so far not been applied to develop a complex database of rate coefficients for collisions involving two triatomic molecules. In this work, our goal is to reduce the computational effort needed to build databases for complex colliding partners, such as H_2O + H_2O, by implementing and incorporating ML algorithms into this process so that more such databases can be produced, and made available to the modeling community. For this purpose, we make use of previously computed cross sections for individual state-to-state transitions for collisions of two water molecules as a benchmark, and for training machine learning models. The goal is to use the available data to explore if it is feasible to construct a reliable model for efficient interpolation in the large space of quantum states of two triatomic molecules.

This paper is organized as follows: Section 2 briefly discusses the theory to compute thermally averaged rate coefficients for H_2O + H_2O collisions, data pre-processing and architecture of the machine learning models employed here. In Section 3, we discuss results obtained from the ML models using NNs and compare them against the MQCT data not used for training the ML models. A summary of our findings is given in Section 4.

2 Methods

2.1 Thermally averaged cross sections (TACSs)

The process of building a database of rotationally inelastic rate coefficients for H_2O + H_2O collisions by computing thermally averaged rate coefficients, $k_{n_1\to n_1'}(T_{\text{rot}}, T_{\text{kin}})$, is explained in detail by Mandal *et al.*[25] Only relevant equations to compute the TACSs are provided below.

Thermal rate coefficients for state-to-state transitions at a given kinetic temperature T_{kin} are computed by averaging the corresponding cross sections over a Maxwell–Boltzmann distribution of relative velocities for all relevant collision energies, E_c, as follows:

$$k_{n_1 n_2\to n_1' n_2'}(T_{\text{kin}}) = \frac{v_{\text{ave}}(T_{\text{kin}})}{(k_B T_{\text{kin}})^2} \times \int_{E_c=0}^{\infty} E_c \sigma_{n_1 n_2\to n_1' n_2'}(E_c) e^{-\frac{E_c}{k_B T_{\text{kin}}}} dE_c, \tag{1}$$

where k_B is the Boltzmann constant, $v_{\text{ave}}(T_{\text{kin}}) = \sqrt{8k_B T_{\text{kin}}/\pi\mu}$ is the average collision velocity, μ is the reduced mass of the collision complex, and the subscripts n_1n_2 and $n_1'n_2'$ indicate the initial and final states, respectively. Each n is a composite index that represents a full set of quantum numbers for one molecule. For example, for water molecules, n denotes $j_{k_A k_C}$, where j is the rotational quantum number and k_A and k_C are the projections of j along the axis of the largest and smallest moment of inertia, I_A and I_C, respectively. Furthermore, for *para*-H_2O (nuclear spins of two H atoms are anti-parallel), k_A + k_C is even and for *ortho*-H_2O (nuclear spins of two H atoms are parallel), k_A + k_C is odd. Since our focus is on the target H_2O molecule, we compute the rate coefficients for water molecules by summing over all final states and averaging over all initial states of its collision partner (quencher):

$$k_{n_1\to n_1'}(T_{\text{rot}}, T_{\text{kin}}) = \sum_{n_2} w_{n_2}(T_{\text{rot}}) \sum_{n_2'} k_{n_1 n_2\to n_1' n_2'}(T_{\text{kin}}). \tag{2}$$

In eqn (2), the thermal populations or weights $w_{n_2}(T_{\text{rot}})$ of the initial states of the quencher are defined as follows:

$$w_{n_2}(T_{\text{rot}}) = \frac{(2j_2+1)e^{-\frac{E_2}{k_B T_{\text{rot}}}}}{Q_2(T_{\text{rot}})}, \tag{3}$$

where E_2 represents the energies of the rotational states n_2 of the quencher. The denominator $Q_2(T_{\text{rot}})$ in eqn (3) is the rotational partition function of the quencher given by

$$Q_2(T_{\text{rot}}) = \sum_{n_2} (2j_2+1)e^{-\frac{E_2}{k_B T_{\text{rot}}}}. \tag{4}$$

For more details of the computation of Q_2, see ref. 25.

The computation of the state-to-state rate coefficients in eqn (1) reaches practical limits for complex systems, like H_2O + H_2O, due to the enormous numbers of individual state-to-state transitions $n_1n_2 \to n_1'n_2'$. In the work reported by Mandal and Babikov,[26] 231 *para–para* and 210 *ortho–ortho* transitions were computed for the target H_2O, considering a maximum value of j_1 = 7. This required a rotational basis set with 38 states each for the *para*- and *ortho*-isomers of the quencher H_2O molecule for which a maximum value of j_2 = 10 was adopted. This led to over a million individual state-to-state transitions $n_1n_2 \to n_1'n_2'$, considering all the initial states of the molecular system as elaborated in the Introduction section. Ideally, these calculations need to be done on a grid of collision energy dense enough to perform the integral over the collision energy E_c, as shown in eqn (1). The human effort needed to manually check millions of individual transitions and implement them in eqn (1) is very labor intensive.

To tackle this challenge, an alternative approach was introduced by exchanging the order of integration in eqn (1) with the summation in eqn (2) as follows:

$$k_{n_1\to n_1'}(T_{\text{rot}}, T_{\text{kin}}) = \frac{v_{\text{ave}}(T_{\text{kin}})}{(k_B T_{\text{kin}})^2} \times \int_{E_c=0}^{\infty} \sigma_{n_1\to n_1'}(E_c, T_{\text{rot}}) e^{-\frac{E_c}{k_B T_{\text{kin}}}} E_c dE_c. \tag{5}$$

In eqn (5), since the summation over the states of the quencher H_2O molecule is now carried out before the integration over collision energy, a thermally averaged cross section for the transition $n_1 \to n_1'$ of the target H_2O molecule is introduced:

$$\sigma_{n_1\to n_1'}(E_c, T_{\text{rot}}) = \sum_{n_2} w_{n_2}(T_{\text{rot}}) \sum_{n_2'} \sigma_{n_1 n_2\to n_1' n_2'}(E_c). \tag{6}$$

The TACSs are computed as follows: first, all the individual state-to-state transition cross sections are summed over the final states of the quencher H_2O molecule, and then the resulting sums are averaged over the initial states of the quencher H_2O molecule for a given value of rotational temperature T_{rot} and collision energy E_c. Since the number of rotational transitions between the states of a target molecule is relatively small, it is much easier to check the behavior of all TACSs before they are integrated over the collision energy in eqn (5). As previously mentioned, in the work of Mandal and Babikov,[26] the number of rotational transitions in *para*- and *ortho*-water considering only de-excitation processes was 231 and 210, respectively, and the number of collision energies was six, making them easier to manually check and ensure proper behavior.

The computed TACSs for these six collision energies can then be used for analytical fits to compute kinetic temperature dependent rate coefficients as described in detail by Mandal *et al.*[25] However, in this work, the computed TACSs are the main goal; therefore, the details of computing rate coefficients using analytical expressions are not discussed here. These 231 transitions in *para*- H_2O and 210 transitions in *ortho*-H_2O for the target molecule are used in this work as a benchmark of the ML predictions.

2.2 Details of the machine-learning method

The TACSs described in the previous section require the individual state-to-state transition cross sections considering initial and final states of the target as well as the quencher H_2O molecules. As stated in the Introduction section, the goal of the present study is to effectively reduce the computational cost of the MQCT calculations to evaluate these state-to-state transition cross sections. This is achieved by the methodology described in the ensuing sections.

2.2.1 Data analysis and pre-processing for machine learning. We begin by analyzing the available data for both excitation and de-excitation of the target H_2O molecule. Previous studies by Mandal and Babikov[26] showed that the dependencies of cross sections on the energy difference between the initial and final states of the colliding partners, given by $\Delta E = E_{\text{initial}} - E_{\text{final}}$, exhibit a single-exponential decay near the $\Delta E = 0$ regime, and a double-exponential decay over the entire range of ΔE. This is illustrated in Fig. 1. The exponential decay is displayed for large ΔE on both excitation ($\Delta E < 0$) and quenching ($\Delta E > 0$) wings. In this work, our focus is to exploit this exponential decay of the state-to-state cross sections with ΔE and use that for our advantage as selection criteria for preparing the training data set for the NNs. A recent study by Joy *et al.* found a similar trend for $H_2O + H_2$ collisions.[36] The authors of ref. 36 fitted their data analytically to compute the coefficients using exponential functions.

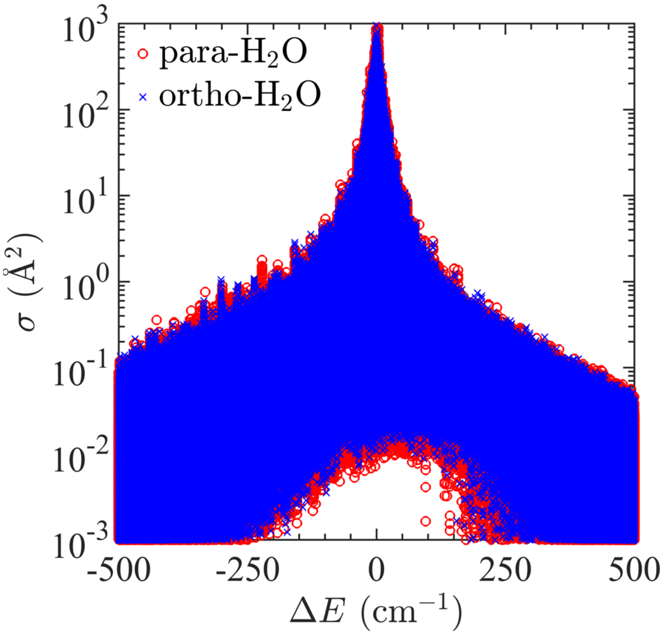

Fig. 1 State-to-state cross sections for rotational transitions in H_2O + H_2O collisions as functions of the energy difference between initial and final rotational levels (ΔE). Results for *para*-H_2O and *ortho*-H_2O targets are shown by open circles (red) and crosses (blue), respectively.

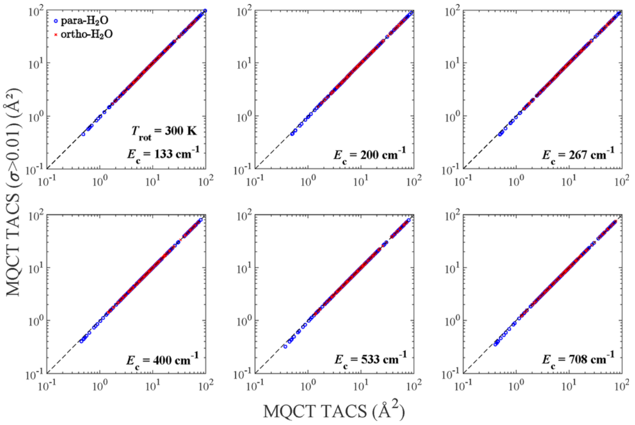

Fig. 2 Comparison of TACSs evaluated with all individual state-to-state cross sections from MQCT calculations with those computed by eliminating $\sigma < 0.01$ Å^2. The dashed black line is the perfect agreement, while blue circles and red crosses correspond to the TACSs for *para* and *ortho*-H_2O targets, respectively.

As a result of the exponential decay, the cross sections vary by several orders of magnitude as the energy difference ΔE increases. Therefore, we start by testing whether the very small cross sections are necessary when the individual cross sections are converted to the TACS. The small-magnitude cross sections are expected to add unnecessary noise to the data, leading to increased complexity of the NNs. Fig. 2 displays a comparison between TACSs for state-to-state transitions with all individual cross sections included compared to the case in which cross sections smaller than 0.01 Å^2 were omitted. The TACSs including all individual cross sections are plotted along the horizontal axis, while the TACSs computed without $\sigma < 0.01$ Å^2 are plotted along the vertical axis. The black dashed line is the perfect agreement while blue circles and red crosses represent the TACSs for *para* and *ortho*-H_2O targets, respectively. It is clear that omitting $\sigma < 0.01$ Å^2 in eqn (5) has no appreciable effect on the accuracy of the TACSs. Therefore, we eliminate individual state-to-state cross sections with a magnitude of <0.01 Å^2 to build our machine learning models.

For training and validation, we take slices from different ranges of ΔE to make a subset of the entire data as follows:

$$\begin{aligned}\text{Data}_{\text{train-validation}} = \Big\{&\sigma_{n_1n_2\to n_1'n_2'}(E_\text{c})|\\ &\left(0\leq\left|\Delta E_{n_1n_2\to n_1'n_2'}\right|\leq 10\ \text{cm}^{-1}\right)\wedge\\ &\left(45\leq\left|\Delta E_{n_1n_2\to n_1'n_2'}\right|\leq 50\ \text{cm}^{-1}\right)\wedge\\ &\left(95\leq\left|\Delta E_{n_1n_2\to n_1'n_2'}\right|\leq 100\ \text{cm}^{-1}\right)\wedge\\ &\left(145\leq\left|\Delta E_{n_1n_2\to n_1'n_2'}\right|\leq 150\ \text{cm}^{-1}\right)\wedge\\ &\left(195\leq\left|\Delta E_{n_1n_2\to n_1'n_2'}\right|\leq 200\ \text{cm}^{-1}\right)\wedge\\ &\left(245\leq\left|\Delta E_{n_1n_2\to n_1'n_2'}\right|\leq 250\ \text{cm}^{-1}\right)\wedge\\ &\left(295\leq\left|\Delta E_{n_1n_2\to n_1'n_2'}\right|\leq 300\ \text{cm}^{-1}\right)\wedge\\ &.\\ &.\\ &.\\ &\Big\}.\end{aligned}\tag{7}$$

The datasets used for training, validation, and testing are displayed in Fig. 3 as a function of ΔE. The red circles represent the subset of data used for training and validation, while the blue crosses indicate the remaining data used for testing. We note that the different energy slices chosen for training and validation may not always include the smallest and the largest ΔE for a given initial state. To avoid extrapolation in ΔE, these lowest or highest energy transitions are also added to the training data for these initial states. Thus, there is a subset of points in the training dataset that lies outside the region specified by eqn (7), as shown in Fig. 3. About 80% of the data points selected in this way are used for training, while the remaining 20% are used for validation.

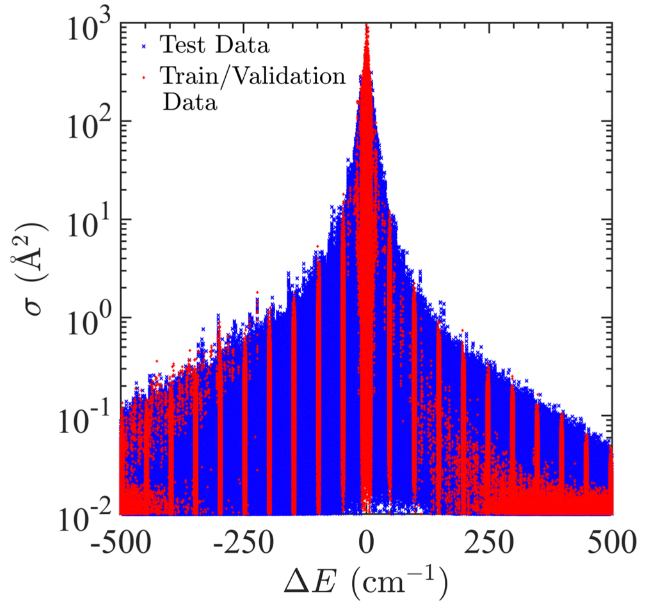

Fig. 3 A visual representation of the data used for training and validation as well as testing. The red circles denote the data for training and validation and the blue crosses denote the data for testing as given by eqn (7).

In general, we aim to design the training and validation dataset that consists of cross section values for which the energy difference of the transition lies within the entire range so that there are no predictions to be made outside the range of the training data. All the cross sections for the remaining transitions are used as the test data set. In this work, multiple slices are made through the whole data set for ML models to optimize the computational efficiency and accuracy of NNs, as discussed later.

Due to the differing slopes of the excitation ($\Delta E < 0$) and quenching ($\Delta E > 0$) wings, a single neural network was unable to adequately capture the behavior of both. Therefore, separate NNs were constructed for these two regimes. Also, each collision energy and *para*- and *ortho*-H_2O symmetries were treated separately to build distinct NNs. For each of the six collision energies, we trained four NNs based on combinations of excitation and quenching transitions for both *para* and *ortho* symmetries, resulting in a total of 24 ML models trained and validated on their respective datasets.

Each of our dataset has thirteen features as input parameters for the NNs: rotational quantum numbers of the initial and final states of the first water molecule $\left(j_1k_{a_1}k_{c_1},j_1'k_{a_1}'k_{c_1}'\right)$, the second water molecule $\left(j_2k_{a_2}k_{c_2},j_2'k_{a_2}'k_{c_2}'\right)$, and the energy difference between the initial and final states of the molecular system, ΔE. The NNs are designed to interpolate over all these input features. Since the input features are composed of different data types (integers for rotational quantum numbers of the initial and final states while float for the energy difference) with a large variation in the magnitude of the input data, they needed to be scaled for the NNs to work optimally. This is done by using the "StandardScaler" function from the "sklearn" package to have zero-mean and unit-variance as $\tilde{x} = \dfrac{x-u}{s}$, where u and s are the mean and the standard deviation of the features, respectively. This standardization of the input data is done so that none of the features get higher weights just because of their magnitude being larger than the values of other features. Note that this scaling is applied only to the input features as listed previously, and not the cross section, *i.e.*, output. The same transformation is applied uniformly to all three data sets: training, validation and testing.

In our data analysis, we found that the dependent feature, *i.e.*, cross sections for individual state-to-state rotational transitions, vary by several orders of magnitude. We found that the NNs do not perform well for data that vary over several orders of magnitude. To resolve this issue, we used the logarithm of the cross section in our ML modeling [$y = \log_{10}(\sigma)$] as the target. The predictions are then converted back to cross sections as $\sigma = 10^y$.

2.2.2 NN details. The NNs all have the same architecture, each characterized by one input layer with thirteen neurons

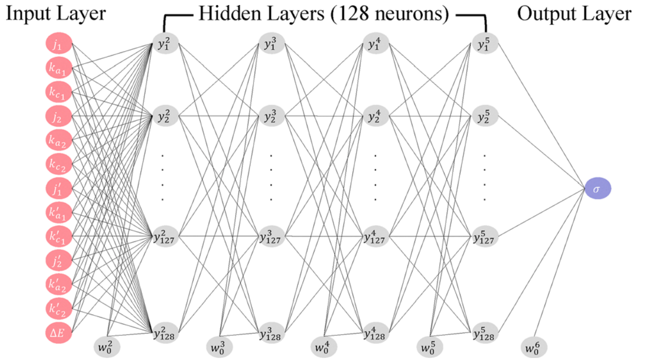

Fig. 4 Architecture of the ML models composed of four hidden layers with each having 128 neurons, thirteen features for the input layer, and one output neuron for the logarithm of the cross sections.

corresponding to the specific features of our dataset. The optimal number of hidden layers following the input layer was determined through an exploratory search to optimize the NN performance. The root mean squared error or RMSE for the test data corresponding to two, three and four hidden layers, each with 128 neurons, were, respectively, 0.81, 0.74 and 0.73 Å^2. Therefore, we decided to build our ML models with four hidden layers, each with 128 neurons. There is one output layer with a single neuron corresponding to the logarithm of the cross sections. A schematic diagram of our NNs is shown in Fig. 4.

The rectified linear unit (ReLU) [$f(x) = \max(0,x)$] is used as the activation function in all hidden layers. We used the Adam optimizer to train the NNs with a learning rate of 0.0001.[88] The details of the NNs including the number of parameters for each layer are provided in Table 1.

Our ML models were built using a TensorFlow with a batch size of 32 and a maximum of 300 epochs.[89] An early stopping mechanism with a patience of 30 using the validation dataset was adopted to prevent the NNs from overfitting. Additionally, we used ridge regularization (*i.e.*, L2 regularization) with a penalty of 0.01 to the weights of the kernels for each hidden layer. This discourages large weights and reduces the complexity of our NNs. A summary of the NN hyperparameters is provided in Table 2.

The mean-squared error function, $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_{\text{MQCT}}^{i} - y_{\text{predicted}}^{i}\right)^2$, was used as the loss function to quantitatively analyze the performance of our ML models applied to the training data. Here, n is the total number of samples used, and the variables y_{MQCT}^{i} and $y_{\text{predicted}}^{i}$ represent the logarithm of the actual cross sections from MQCT calculations and the cross sections predicted by the NNs, respectively. We also monitored the $\text{RMSE} = \sqrt{\text{MSE}}$ for the interpretation of the predicted data.

Table 1 Summary of the architecture and technical details of the NNs

Layer name	No. of neurons	No. of parameters	Activation
Input	13	13	
Hidden layer 1: dense	128	1792	ReLU
Hidden layer 2: dense	128	16 512	ReLU
Hidden layer 3: dense	128	16 512	ReLU
Hidden layer 4: dense	128	16 512	ReLU
Output	1	129	

Table 2 A summary of the hyper-parameters and their values used for training the ML models

Hyperparameters	Type/value	Additional details
Batch size	32	
Optimizer	Adam[88]	Learning rate = 0.0001
Kernel regularization	Ridge (L2)	Regularization weight = 0.01
Maximum no. of epochs	300	
Early stopping	Implemented	Patience = 30

3 Results

3.1 Optimization of training, validation and test datasets

First, we analyzed the whole dataset of individual state-to-state cross sections for rotational transitions in the collision of two H_2O molecules. This dataset included all combinations of *ortho*- and *para*-H_2O, *i.e.*, *ortho–ortho*, *ortho–para*, *para–ortho*, and *para–para* combinations considering both the target and quencher H_2O molecules and for all six collision energies. Fig. 1 displays this cross section data as a function of the energy difference (ΔE) between the initial and final rotational states. Cross sections for *para*-H_2O as the target molecule are shown by red open circles, while blue crosses represent the same for *ortho*-H_2O as the target. They are characterized by a single exponential decay for an $|\Delta E|$ of $\leq$ 50 cm^{-1} followed by a second exponential decay as $|\Delta E|$ increases. To adequately capture the exponential decay of the cross sections with ΔE, we adopted different slices of the entire data set for training and validation and built several NNs as part of the optimization process. The accuracy of the ML models was determined at two different levels. First, all the individual state-to-state transitions were tested against the whole test dataset. Second, thermally averaged cross sections were computed using the NN predictions and compared against the actual MQCT TACS.

3.1.1 Dataset 1. We started to build our NNs using the data shown in Fig. 3 and specified by eqn (7). The predicted cross sections for state-to-state transitions were compared against the actual MQCT cross sections from the test data for both *para* and *ortho*-H_2O targets as shown in Fig. S1 and S2, respectively, of the Supplementary Information (SI). The resulting TACSs using these predicted data are compared against the actual MQCT TACSs, as shown in Fig. S3 of the SI. Note that the computation of TACSs requires both the NN predictions and the data used for training and validation. The agreement at the level of individual cross sections is found to be reasonable for both *para* and *ortho*-H_2O targets, while the agreement at the level of TACS is found to be excellent.

3.1.2 Dataset 2. We explored if we can reduce the size of the training dataset, while preserving prediction accuracy, to improve efficiency. Therefore, instead of composing the subset of data for training and validation at every ΔE = 50 cm^{-1}, we built a subset with a step of 100 cm^{-1} in ΔE:

$$\begin{aligned}\text{Data}_{\text{train-validation}} = \Big\{&\sigma_{n_1 n_2 \to n_1' n_2'}(E_\text{c}) \big| \\ &\left(0 \le \left|\Delta E_{n_1 n_2 \to n_1' n_2'}\right| \le 10\ \text{cm}^{-1}\right) \wedge \\ &\left(95 \le \left|\Delta E_{n_1 n_2 \to n_1' n_2'}\right| \le 100\ \text{cm}^{-1}\right) \wedge \\ &\left(195 \le \left|\Delta E_{n_1 n_2 \to n_1' n_2'}\right| \le 200\ \text{cm}^{-1}\right) \wedge \\ &\left(295 \le \left|\Delta E_{n_1 n_2 \to n_1' n_2'}\right| \le 300\ \text{cm}^{-1}\right) \wedge \\ &. \\ &. \\ &\Big\}.\end{aligned} \tag{8}$$

The sets of training/validation and test data from Dataset 2 are displayed in Fig. S4. The comparison with the MQCT data for individual state-to-state cross sections became slightly worse for both *para* and *ortho*-H_2O targets as shown in Fig. S5 and S6, respectively. The predicted TACS also displays larger discrepancies with the MQCT TACS as shown in Fig. S7 of the SI.

3.1.3 Dataset 3.

The training data need to reflect dataset 1 due to its double exponential feature. The ML models need to capture the change in slope across different ranges of ΔE, and the training data should reflect this behavior. As explained earlier, the slope of the exponential decay changes rapidly near ΔE = 50 cm^{-1}, and so these data should be a part of the training set as shown in eqn (9) below. Moreover, we explored reducing the range of the maximum magnitude of ΔE to check if we really need to sample the entire data. We systematically reduced the value of ΔE and found that eliminating data above $|\Delta E| \geq 300$ cm^{-1} has a minimal effect on the overall TACS. We thus arrived at the following dataset:

$$\begin{aligned}\text{Data}_{\text{train-validation}} = \Big\{&\sigma_{n_1 n_2 \to n_1' n_2'}(E_\text{c}) \big| \\ &\left(0 \le \left|\Delta E_{n_1 n_2 \to n_1' n_2'}\right| \le 10\ \text{cm}^{-1}\right) \wedge \\ &\left(45 \le \left|\Delta E_{n_1 n_2 \to n_1' n_2'}\right| \le 50\ \text{cm}^{-1}\right) \wedge \\ &\left(95 \le \left|\Delta E_{n_1 n_2 \to n_1' n_2'}\right| \le 100\ \text{cm}^{-1}\right) \wedge \\ &\left(145 \le \left|\Delta E_{n_1 n_2 \to n_1' n_2'}\right| \le 150\ \text{cm}^{-1}\right) \wedge \\ &\left(195 \le \left|\Delta E_{n_1 n_2 \to n_1' n_2'}\right| \le 200\ \text{cm}^{-1}\right) \wedge \\ &\left(245 \le \left|\Delta E_{n_1 n_2 \to n_1' n_2'}\right| \le 250\ \text{cm}^{-1}\right) \wedge \\ &\left(295 \le \left|\Delta E_{n_1 n_2 \to n_1' n_2'}\right| \le 300\ \text{cm}^{-1}\right)\Big\}.\end{aligned} \tag{9}$$

This is the training/validation set that we label as the best performing and the most optimized for reliable predictions. We adopt the data structure shown by eqn (9) for collision energies of 133, 200, and 267 cm^{-1}. Because the density of individual cross sections near the elastic peak decreases rapidly with collision energy, for E_c = 400 cm^{-1} and higher, we replace ΔE of $<$ 10 cm^{-1} by ΔE of $<$ 15 cm^{-1} in eqn (9) to have adequate sampling near ΔE = 0 cm^{-1}. The resulting training/validation and test data corresponding to individual state-to-state rotational transitions are displayed in Fig. 5 for the *para*-H_2O molecule at the highest and lowest collision energies. For other collision energies and *ortho*-H_2O molecules, the dataset is very similar.

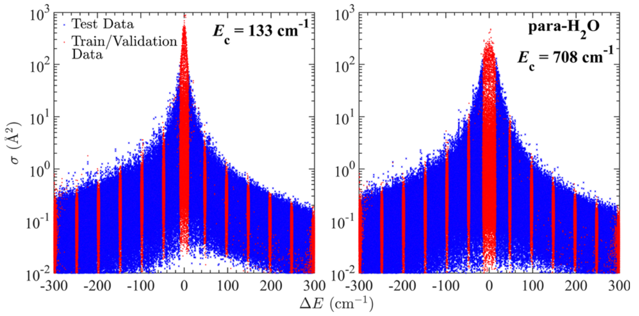

Fig. 5 The training/validation and test data for all state-to-state transitions at the highest and lowest collision energies, based on dataset 3, are presented as a function of ΔE for the *para*-H_2O molecule. The dataset for other collision energies and *ortho*-H_2O is very similar.

A summary of the sizes of the training, validation and test datasets from eqn (9) is given in Table 3 that includes both exchange symmetries of *para* and *ortho*-H_2O molecules and both excitation and quenching transitions. The size of the training data is about ~10% of the entire dataset after removing the small-magnitude cross sections and limiting the range of energy gaps to ΔE of $\leq$ 300 cm^{-1}. The size of the training, validation and test datasets remains almost the same for the lower three collision energies for both *para* and *ortho* symmetries and for both excitation and quenching transitions. The same applies to the higher three collision energies, but the size of the training, validation and test datasets is slightly different as explained before.

The predicted cross sections for the individual state-to-state rotational transitions of both *para* and *ortho*-H_2O molecules are shown in Fig. 6 and 7, respectively. The horizontal axis represents the MQCT cross sections while the NN predictions are plotted along the vertical axis. The black dashed line would be the perfect agreement while the red dots represent the comparison. It is seen that the predicted cross sections agree reasonably well with the MQCT cross sections and are within the range of acceptable accuracy for both *para* and *ortho* symmetries of the H_2O molecule. While the smaller cross sections (corresponding to larger ΔE) exhibit higher discrepancies (relative to their magnitudes), their overall contribution to TACSs is less significant.

To further quantify the errors in the NN predictions for each collision energy, specific symmetry of the H_2O molecule (*para* or *ortho*), and excitation/quenching transitions (ΔE $<$ 0 or

Table 3 A summary of the sizes of training, validation and test datasets for both *para*- and *ortho*-H_2O molecules and for both excitation ($\Delta E < 0$) and quenching ($\Delta E > 0$) transitions is listed. The corresponding relative RMSE is also listed in the last column

E_c (cm^{-1})	H_2O symmetry	ΔE sign	Training data size	Validation data size	Test data size	Relative RMSE
133	*para*	(+ve)	34 422	7772	273 223	0.430
200	*para*	(+ve)	34 424	7772	273 232	0.379
267	*para*	(+ve)	34 445	7778	273 374	0.421
400	*para*	(+ve)	39 165	8957	267 713	0.402
533	*para*	(+ve)	39 162	8957	267 783	0.404
708	*para*	(+ve)	39 199	8966	267 967	0.397
133	*ortho*	(+ve)	33 161	7494	259 189	0.429
200	*ortho*	(+ve)	33 164	7495	259 241	0.384
267	*ortho*	(+ve)	33 186	7501	259 329	0.389
400	*ortho*	(+ve)	37 526	8586	254 107	0.387
533	*ortho*	(+ve)	37 541	8590	254 221	0.380
708	*ortho*	(+ve)	37 552	8592	254 307	0.367
133	*para*	(−ve)	36 196	8213	291 562	0.438
200	*para*	(−ve)	36 137	8199	291 201	0.442
267	*para*	(−ve)	36 161	8205	291 181	0.462
400	*para*	(−ve)	40 837	9374	285 384	0.414
533	*para*	(−ve)	40 864	9380	285 566	0.431
708	*para*	(−ve)	40 891	9387	285 943	0.404
133	*ortho*	(−ve)	35 544	8089	280 565	0.444
200	*ortho*	(−ve)	35 525	8084	280 295	0.430
267	*ortho*	(−ve)	35 528	8085	280 334	0.416
400	*ortho*	(−ve)	39 844	9163	275 045	0.424
533	*ortho*	(−ve)	39 859	9167	275 191	0.409
708	*ortho*	(−ve)	39 880	9172	275 370	0.381

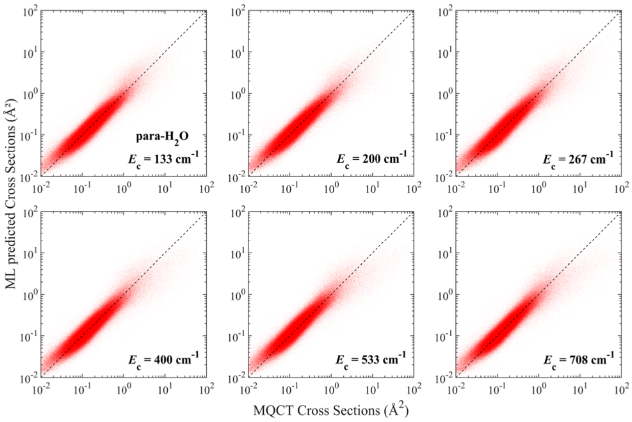

Fig. 6 A comparison of ML predicted state-to-state cross sections against the actual MQCT results for the *para*-H_2O molecule.

$\Delta E > 0$), we report in the last column of Table 3 the relative RMSE or RRMSE, defined as:

$$\text{RRMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sigma^i_{\text{MQCT}} - \sigma^i_{\text{predicted}}}{\sigma^i_{\text{MQCT}}}\right)^2}, \tag{10}$$

where σ^i_{MQCT} refers to the MQCT cross sections not used for training or validation. The RRMSE values range from ∼37% to ∼46% with an average value of ∼41%.

We have also examined whether a similar level of accuracy can be reached with a fewer number of hidden layers and number of neurons in each layer to reduce the complexity of the NNs since the training dataset in this case is relatively small compared to dataset 1. We found that reducing the number of hidden layers does not significantly make the prediction worse with respect to the RMSE values nor meaningfully improve the computational efficiency. Therefore, we retained four hidden layers with 128 neurons each.

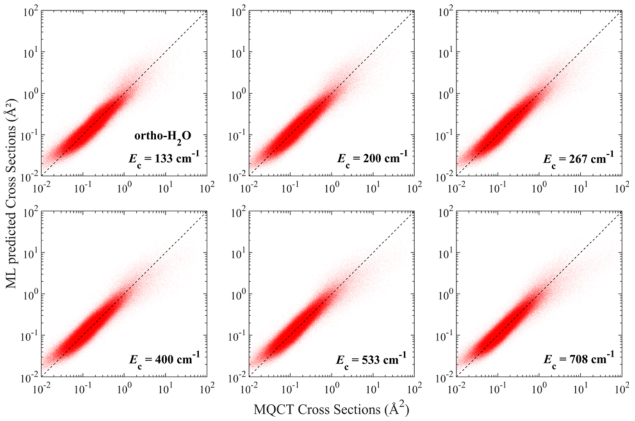

Fig. 7 A comparison of ML predicted state-to-state cross sections against the actual MQCT results for the *ortho*-H_2O molecule.

Finally, we composed the TACSs following eqn (6) using the predicted individual state-to-state transitions for the test data combined with original training and validation data and compared against the actual MQCT TACS. Fig. 8 displays this comparison. The blue circles and red crosses correspond to the *para* and *ortho* symmetries of the H_2O molecule, respectively. These TACSs are computed for a rotational temperature T_{rot} = 300 K. The agreement is excellent between the actual MQCT TACS and the NN predictions. The agreement does not improve significantly when more data are included by extending the range of the ΔE as shown in Fig. S3 of the SI using dataset 1.

It should be noted that the original MQCT state-to-state cross sections were divided into a training and validation set (about 10%) and a test set (about 90%). In the comparison

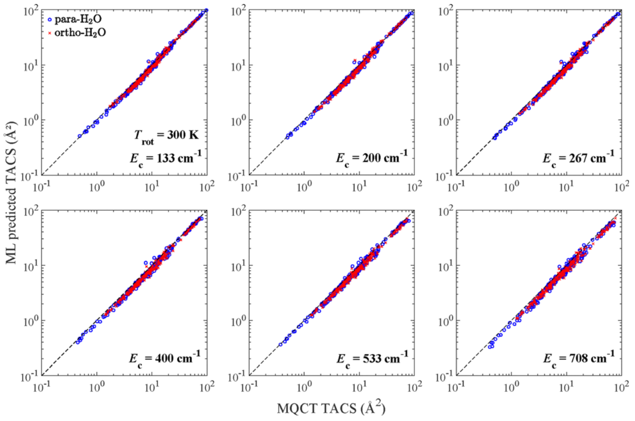

Fig. 8 A comparison of thermally averaged cross sections (TACSs) computed using state-to-state cross sections from MQCT calculations and our NN predicted cross sections is presented in different panels for six different collision energies. Blue empty circles and red crosses denote, respectively, transitions in target *para*-H_2O and *ortho*-H_2O molecules.

provided in Fig. 8, the test set was replaced by the cross sections produced by NN predictions. To demonstrate the important contributions of the NN predicted data toward the overall TACSs, we computed the TACSs by using only the training and validation data. The resulting TACSs are compared against those obtained by including the NN predictions and the actual MQCT TACS in Fig. 9 for *ortho*-H_2O. Similar results for *para*-H_2O are provided in Fig. S8 of the SI. The TACSs using all of the original MQCT cross sections are plotted along the horizontal axis, while the red crosses and blue circles correspond to the TACS computed with and without incorporating the NN predictions into the training and validation data. The dashed black curve would represent the perfect agreement. These red crosses are the same as shown in Fig. 8. It can be seen that the blue circles deviate significantly from the perfect agreement represented by the black dashed diagonal line. Thus, the contribution of the NN-predicted cross sections is significant accounting for nearly 90% of the original MQCT cross sections not used for training and validation. Therefore, the approach presented here can be applied to other complex molecular systems to substantially reduce the complexity involved in rate coefficient calculations.

The TACSs are one of the main ingredients for astrophysical models and serve as an important input to the numerical codes of radiative transfer modeling, such as RADEX, MOLPOP or LIME. To confirm that the accuracy of the NNs implemented here is sufficient for these models, we computed the percentage deviation of the TACSs as predicted by our NNs and the original MQCT TACSs. The computed percentage deviation is then averaged over all transitions for target H_2O molecules over 231 *para*-H_2O and 210 *ortho*-H_2O transitions. The resulting data, presented in Table 4, illustrate that the agreement is excellent at the level of TACS despite larger deviations at the state-to-state level. Because TACSs are the main quantity that is relevant for modeling energy transfer in astrophysical environments, the level of accuracy attained in our ML models is adequate for astrophysical models. The average percentage deviation is about ~13% and ~14%, while the largest error is about ~15% and ~17% for *para* and *ortho*-H_2O targets, respectively. This is very encouraging since our goal is to reduce the requirements of the computational resources to explicitly compute the relevant TACSs. This is achieved in our proposed workflow without losing significant accuracy but at a very low computational cost.

Table 4 Average percent deviation between the ML predicted TACS and the MQCT TACS for both *ortho*- and *para*-H_2O targets at different collision energies

	Average % difference	
E_c (cm^{-1})	*para*-H_2O	*ortho*-H_2O
133	11.87	10.00
200	13.80	13.73
267	12.50	14.18
400	13.06	14.72
533	13.30	14.28
708	15.00	16.99

To illustrate the accuracy of TACSs derived from the NNs, a plot of percent deviation for all 441 transitions of the target H_2O molecule considering both *ortho*- and *para*-H_2O symmetries for all six collision energies is displayed in Fig. 10 as a function of the magnitude of the actual MQCT TACSs. It is seen that for larger values of TACSs, the percentage deviation remains in the range of ~10–20% considering all collision energies and both symmetries of the target H_2O molecule. The agreement is better for lower values of collision energies and decreases slightly at higher collision energies. We also notice an increase in the percentage deviation for a lower magnitude of the TACSs for both *para* and *ortho*-H_2O transitions, which is expected.

3.1.4 Dataset 4. Because the raw data for constructing and testing the NNs come from MQCT calculations, which are computationally expensive, ideally one would like to have the smallest set of MQCT data for training and validation. With this in mind, we tested a further reduction of the size of the training and validation data by limiting the state-to-state transitions to an $|\Delta E|$ of $\leq$ 200 cm^{-1}. The resulting NN predictions are

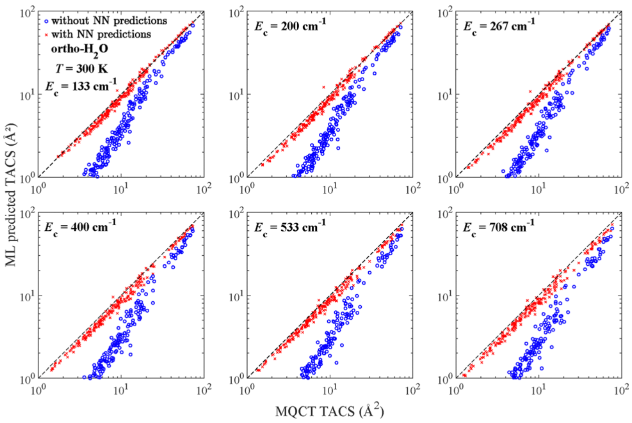

Fig. 9 Thermally averaged cross sections computed with (red crosses) and without (blue circles) incorporating the ML predicted cross sections into the training and validation data for *ortho*-H_2O molecules plotted against the original MQCT TACS. The red crosses are the same as in Fig. 8.

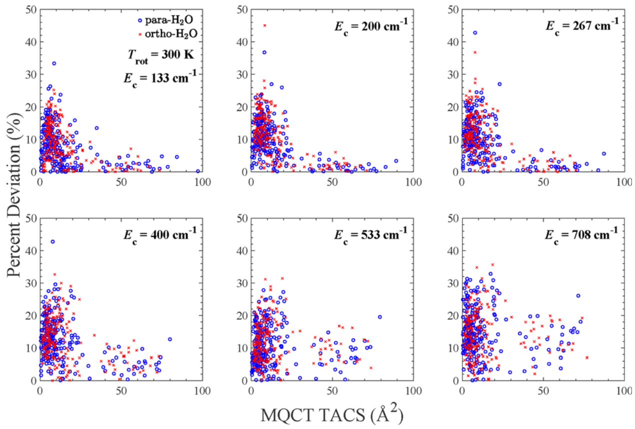

Fig. 10 Percentage deviation of the ML predicted TACS from the actual MQCT TACS as a function of the magnitude of the MQCT TACS.

shown in Fig. S9 and S10 of the SI for *para* and *ortho* H_2O, respectively. The TACSs are shown in Fig. S11 of the SI. Clearly, the agreement is not at the same level as with dataset 3. Higher discrepancies are visible at the level of individual cross sections as well as TACSs but they remain reasonable. We conclude that dataset 3 provides the best balance between performance and accuracy.

3.2 Computational efficiency of the ML models

The efficiency of the ML approach is dependent on the size of the training and validation dataset since that is the data one needs to compute using MQCT. Therefore, dataset 4 is the most efficient while dataset 1 is the least efficient. Due to the higher error of the predicted results from dataset 4, we deem it unacceptable. This leads us to conclude that dataset 3 provides an optimal choice to achieve acceptable accuracy while maintaining the cost of computing the required MQCT data for training and validation reasonably small. Thus, the following analysis pertains to dataset 3.

Besides the size of the training and validation data, there are several other approximations that one can make to reduce the computational cost of the proposed ML approach. One can eliminate small-magnitude cross sections to reduce noise, but these cross sections are hard to identify without scattering calculations. Therefore, we do not consider this in our efficiency estimation. Another approximation is made based on the magnitude of the range of $|\Delta E|$. This can be explored prior to large scale MQCT calculations in order to achieve the most efficient workflow proposed in our methodology.

On average, the size of the test data is ~7.5 times larger than the sum of the sizes of the training and validation data. It was found from previous studies by Mandal *et al.*[32] that the cost of MQCT calculations, within the adiabatic trajectory approximation (AT-MQCT) methodology, increases as N^3 (typical of coupled-channel calculations), where N is the number of rotational states in the basis. Note that a larger rotational basis is needed for higher $|\Delta E|$ transitions. Thus, excluding higher $|\Delta E|$ transitions, as discussed in the various architectures of datasets, can drastically reduce the computation cost.

The efficiency of the ML approach should consider both the cost of constructing the relevant transition matrices and the actual cost of propagating the trajectories. However, these matrices now also contain significantly reduced numbers of transitions compared to the original MQCT calculations. Computation of each of these matrices in the original MQCT calculations required about 676 000 CPU hours and a total of four matrices were computed. However, it is difficult to estimate the cost savings in computing these matrices with a smaller number of transitions. Additionally, a newer version of MQCT is expected to significantly speedup these calculations making them computationally less demanding compared to the MQCT trajectory simulations. Thus, we do not include it in our estimation of computational efficiency.

As reported earlier,[26] the cost of the MQCT trajectory calculations was about 5 250 000 million CPU hours. Considering that the test data for dataset 3 are about 7 times larger than the training and the validation data and that computationally demanding high $|\Delta E|$ transitions ($|\Delta E| > 300\ \text{cm}^{-1}$) were excluded from the dataset, we estimate the cost for the trajectory calculations to be around 100 000 CPU hours for producing the relevant data for training and validation. Overall, we expect about a factor of 50 savings in the computational cost from our approach. The actual CPU time for training the NNs is insignificant (about ~5.5 CPU hours) once the training and validation data are in place. This is a remarkable gain in efficiency. Using the methodology presented here, an expanded database of much needed rotational transitions in water molecules can be computed at a reasonable computational cost.

4 Conclusions

In this work, we implemented a machine-learning method to reduce the cost of computing state-to-state and thermally averaged cross sections for collisions of complex molecular systems, such as $H_2O + H_2O$. Due to the methodological as well as computational limitations, inelastic collisions of these systems remain largely unexplored while the rate coefficients are much needed for the astrophysics community. Computational costs of building a database of rate coefficients for rotational transitions in $H_2O + H_2O$ collisions were estimated to be ~5–8 million CPU hours when only first 231 and 210 transitions of *para*- and *ortho*-H_2O molecules, respectively, were explored. However, this is not sufficient and in this work, we present a methodology to expand the database further to include more rotational transitions at a significantly reduced computational cost.

Prior applications of GP and NNs for predicting state-to-state cross sections/rate coefficients for atom–diatom scattering involved interpolation in spaces of quantum states defined by only four quantum numbers (v, j and v', j') of the diatomic molecules.[77] Similarly, in the application of GP models for diatom–diatom scattering, at most 8 quantum numbers (v_i, j_i, v_i',j_i' of two diatomic molecules) are considered in the work of Jasinski *et al.*[47] though recent studies of Mihalik *et al.*[79] and Wang *et al.*[80] considered only changes in ro-vibrational levels of SiO in collisions of SiO and the ground state *para*-H_2. Here, the ML approach is shown to yield reliable results for interpolation in the space of 12 quantum numbers for $H_2O + H_2O$ collisions. Our approach of using an ensemble of NNs for a range of collision energies paves a pathway for efficient computation of rate coefficients for state-to-state rotational transitions in collisions of two asymmetric top molecules.

For practical purposes, an estimated speed up factor of ~50 is expected using our proposed pipeline exploiting the physics behind the energy transfer process (exponential decay of rate coefficients with energy gap) and utilizing deep learning algorithms. The cross section data for rotationally inelastic scattering of two H_2O molecules computed using MQCT shows a double exponential feature as a function of the energy difference between the initial and final states. NNs constructed and demonstrated in this work appear to capture this feature during training and yield models that successfully predict cross sections for state-to-state rotational transitions in H_2O + H_2O collisions from which accurate thermally averaged cross sections are derived. Our tested NNs achieved an excellent accuracy level for a higher magnitude of thermally averaged cross sections. In the future, we hope to use this methodology to compute additional rotational transitions in water to extend the existing database.

The methodology presented here is robust and general and can be implemented for other systems of complex colliding partners, such as HCN + H_2O and HDO + H_2O, and collisions of atoms and diatoms with other polycyclic aromatic hydrocarbons. By utilizing the machine learning models using neural networks as proposed in this work, these computationally demanding scattering calculations are expected to become significantly more affordable. This proposed workflow is expected to open a new avenue in the near future to populate databases such as BASECOL for astrophysical modeling.

Author contributions

N. B. and B. M. conceived the project. B. M. and D. B. generated the original MQCT data. B. M. constructed and tested the ML models with assistance from all co-authors. All co-authors contributed to data analysis, validation, and manuscript preparation.

Conflicts of interest

There are no conflicts to declare.

Data availability

The data supporting this article have been included in the main article and the supplementary information (SI). Supplementary information: neural network ensemble for computing cross sections of rotational transitions in H_2O + H_2O collisions. Additional supporting material on ML models for different datasets of training, validation and testing and comparison of NN predictions against original MQCT data. See DOI: https://doi.org/10.1039/d5cp02812d.

Acknowledgements

This work is supported in part by NSF grant no. PHY-2409497 (N. B.) and NASA grant 80NSSC22K1167 (P. C. S.). D. B. was supported by NSF grant CHE-2102465. R. V. K. acknowledges support from the Natural Sciences and Engineering Research Council (NSERC) of Canada. This research also used HPC resources at Marquette University funded in part by NSF award CNS-1828649 and the RebelX cluster at UNLV.

References

1 D. Bockelee-Morvan, U. Calmonte, S. Charnley, J. Duprat, C. Engrand, A. Gicquel, M. Hassig, E. Jehin, H. Kawakita and B. Marty, *et al.*, *Space Sci. Rev.*, 2015, **197**, 47.
2 L. Dones, R. Brasser, N. Kaib and H. Rickman, *Space Sci. Rev.*, 2015, **197**, 191.
3 E. F. van Dishoeck, L. E. Kristensen, J. C. Mottram, A. O. Benz, E. A. Bergin, P. Caselli, F. Herpin, M. R. Hogerheijde, D. Johnstone and R. Liseau, *et al.*, *Astron. Astrophys.*, 2021, **648**, A24.
4 P. Hartogh, E. Lellouch, R. Moreno, D. Bockelee-Morvan, N. Biver, T. Cassidy, M. Rengel, C. Jarchow, T. Cavalie and J. Crovisier, *et al.*, *Astron. Astrophys.*, 2011, **532**, L2.
5 E. Roueff and F. Lique, *Chem. Rev.*, 2013, **113**, 8906.
6 E. F. van Dishoeck, L. E. Kristensen, A. O. Benz, E. A. Bergin, P. Caselli, J. Cernicharo, F. Herpin, M. R. Hogerheijde, D. Johnstone and R. Liseau, *et al.*, *Publ. Astron. Soc. Pac.*, 2011, **123**, 138.
7 E. M. L. Humphreys, *Proc. Int. Astron. Union*, 2007, **3**, 471.
8 K. E. Mandt, O. Mousis, B. Marty, T. Cavalie, W. Harris, P. Hartogh and K. Willacy, *Space Sci. Rev.*, 2015, **197**, 297.
9 J. Loreau, A. Faure and F. Lique, *J. Chem. Phys.*, 2018, **148**, 244308.
10 M. B. Khalifa, E. Quintas-Sanchez, R. Dawes, K. Hammami and L. Wiesenfeld, *Phys. Chem. Chem. Phys.*, 2020, **22**, 17494.
11 F. F. S. Van der Tak, J. H. Black, F. L. Schoier, D. J. Jansen and E. F. van Dishoeck, *Astron. Astrophys.*, 2007, **468**, 627.
12 C. Brinch and M. R. Hogerheijde, *Astron. Astrophys.*, 2010, **523**, A25.
13 A. A. Ramos and M. Elitzur, *Astron. Astrophys.*, 2018, **616**, A131.
14 N. Biver, J. Boissier, D. Bockelée-Morvan, J. Crovisier, H. Cottin, M. Cordiner, N. Roth and R. Moreno, *Astron. Astrophys.*, 2022, **668**, A171.
15 M. L. Dubernet, C. Boursier, O. Denis-Alpizar, Y. A. Ba, N. Moreau, C. M. Zwolf, M. A. Amor, D. Babikov, N. Balakrishnan and C. Balança, *et al.*, *Astron. Astrophys.*, 2024, **683**, A40.
16 F. F. S. van der Tak, F. Lique, A. Faure, J. H. Black and E. F. van Dishoeck, *Atoms*, 2020, **8**, 15.
17 J. M. Hutson and C. R. Le Sueur, molscat, bound and field, version 2020.0, 2020.
18 J. M. Hutson and C. R. Le Sueur, *Comput. Phys. Commun.*, 2019, **241**, 9.
19 M. H. Alexander, P. J. Dagdigian, H. J. Werner, J. Kłos, B. Desrousseaux, G. Raffy and F. Lique, *Comput. Phys. Commun.*, 2023, **289**, 108761.
20 R. V. Krems, *TwoBC-quantum scattering program*, University of British Columbia, Vancouver, Canada, 2006.
21 B. Mandal, D. Bostan, C. Joy and D. Babikov, *Comput. Phys. Commun.*, 2024, **294**, 108938.
22 P. J. Agg and D. C. Clary, *J. Chem. Phys.*, 1991, **95**, 1037.
23 C. Boursier, B. Mandal, D. Babikov and M. L. Dubernet, *Mon. Not. R. Astron. Soc.*, 2020, **498**, 5489.
24 G. Buffa, O. Tarrini, F. Scappini and C. Cecchi-Pestellini, *Astrophys. J., Suppl. Ser.*, 2000, **128**, 597.
25 B. Mandal, M. Zoltowski, M. Cordiner, F. Lique and D. Babikov, *Astron. Astrophys.*, 2024, **688**, A208.
26 B. Mandal and D. Babikov, *Astron. Astrophys.*, 2023, **671**, A51.

27 B. Mandal and D. Babikov, *Astron. Astrophys.*, 2023, **678**, A51.

28 B. Mandal, A. Semenov and D. Babikov, *J. Phys. Chem. A*, 2018, **122**, 6157.

29 B. Mandal, A. Semenov and D. Babikov, *J. Phys. Chem. A*, 2020, **124**, 9877.

30 B. Mandal, PhD thesis, Marquette University, 2021.

31 B. Mandal, C. Joy, A. Semenov and D. Babikov, *ACS Earth Space Chem.*, 2022, **6**, 521.

32 B. Mandal, C. Joy, A. Bostan, D. Eng and D. Babikov, *J. Phys. Chem. Lett.*, 2023, **14**, 817.

33 D. Bostan, B. Mandal, C. Joy and D. Babikov, *Phys. Chem. Chem. Phys.*, 2023, **25**, 15683.

34 D. Bostan, B. Mandal and D. Babikov, *Phys. Chem. Chem. Phys.*, 2024, **26**, 27567.

35 C. Joy, B. Mandal, D. Bostan and D. Babikov, *Phys. Chem. Chem. Phys.*, 2023, **25**, 17287.

36 C. Joy, B. Mandal, D. Bostan, M. L. Dubernet and D. Babikov, *Faraday Discuss.*, 2024, **251**, 225.

37 C. Joy, D. Bostan, B. Mandal and D. Babikov, *Astron. Astrophys.*, 2024, **692**, A229.

38 D. Bostan, B. Mandal, C. Joy, M. Zoltowski, F. Lique, J. Loreau, E. Quintas-Sanchez, A. Batista-Planas, R. Dawes and D. Babikov, *Phys. Chem. Chem. Phys.*, 2024, **26**, 6627.

39 A. Semenov, B. Mandal and D. Babikov, *Comput. Phys. Commun.*, 2020, **252**, 107155.

40 H. Moustafa, P. M. Lyngby, J. J. Mortensen, K. S. Thygesen and K. W. Jacobsen, *Phys. Rev. Mater.*, 2023, **7**, 014007.

41 S. S. Rahaman, S. Haldar and M. Kumar, *J. Phys.: Condens. Matter*, 2023, **35**, 115603.

42 S. Ciarella, M. Chiappini, E. Boattini, M. Dijkstra and L. M. C. Janssen, *Mach. Learn.: Sci. Technol.*, 2023, **4**, 025010.

43 N. N. Ma, T. L. Zhao, W. X. Wang and H. F. Zhang, *Phys. Rev. C*, 2023, **107**, 014310.

44 G. P. A. Nobre, D. A. Brown, S. J. Hollick, S. Scoville and P. Rodrguez, *Phys. Rev. C*, 2023, **107**, 034612.

45 N. Mallick, S. Prasad, A. N. Mishra, R. Sahoo and G. G. Barnaföldi, *Phys. Rev. D*, 2023, **107**, 094001.

46 G. Fang, S. Ba, Y. Gu, Z. Lin, Y. Hou, C. Qin, C. Zhou, J. Xu, Y. Dai and J. Song, *et al.*, *Astron. J.*, 2023, **165**, 35.

47 A. Jasinski, J. Montaner, R. C. Forrey, B. H. Yang, P. C. Stancil, N. Balakrishnan, J. Dai, R. A. Vargas-Hernández and R. V. Krems, *Phys. Rev. Res.*, 2020, **2**, 032051.

48 M. Lochner, L. Rudnick, I. Heywood, K. Knowles and S. S. Shabala, *Mon. Not. R. Astron. Soc.*, 2023, **520**, 1439.

49 P. Ilten, T. Menzo, A. Youssef and J. Zupan, *SciPost Phys.*, 2023, **14**, 027.

50 A. Butter, T. Plehn, S. Schumann, S. Badger, S. Caron, K. Cranmer, F. A. Di Bello, E. Dreyer, S. Forte and S. Ganguly, *et al.*, *SciPost Phys.*, 2023, **14**, 079.

51 P. Eller, A. T. Fienberg, J. Weldert, G. Wendel, S. Böser and D. F. Cowen, *Nucl. Instrum. Methods Phys. Res., Sect. A*, 2023, **1048**, 168011.

52 D. M. Anstine and O. Isayev, *J. Phys. Chem. A*, 2023, **127**, 2417.

53 C. Miles, R. Samajdar, S. Ebadi, T. T. Wang, H. Pichler, S. Sachdev, M. D. Lukin, M. Greiner, K. Q. Weinberger and E. A. Kim, *Phys. Rev. Res.*, 2023, **5**, 013026.

54 Y. H. Zhang and M. Di Ventra, *Phys. Rev. B*, 2023, **107**, 075147.

55 P. Lemos, M. Cranmer, M. Abidi, C. Hahn, M. Eickenberg, E. Massara, D. Yallup and S. Ho, *Mach. Learn.: Sci. Technol.*, 2023, **4**, 01LT01.

56 D. de Andres, G. Yepes, F. Sembolini, G. Martnez-Muñoz, W. Cui, F. Robledo, C. H. Chuang and E. Rasia, *Mon. Not. R. Astron. Soc.*, 2023, **518**, 111.

57 I. Gómez-Vargas, J. Briones Andrade and J. A. Vázquez, *Phys. Rev. D*, 2023, **107**, 043509.

58 B. J. Braams and J. M. Bowman, *Int. Rev. Phys. Chem.*, 2009, **28**, 577.

59 C. Qu, Q. Yu and J. M. Bowman, *Annu. Rev. Phys. Chem.*, 2018, **69**, 151.

60 P. L. Houston, C. Qu, Q. Yu, P. Pandey, R. Conte, A. Nandi and J. M. Bowman, *J. Chem. Theory Comput.*, 2024, **20**, 3008.

61 B. Jiang and H. Guo, *J. Chem. Phys.*, 2013, **139**, 054112.

62 J. Li, B. Jiang and H. Guo, *J. Chem. Phys.*, 2013, **139**, 204103.

63 B. Jiang and H. Guo, *J. Chem. Phys.*, 2014, **141**, 034109.

64 C. Xie, X. Zhu, D. R. Yarkony and H. Guo, *J. Chem. Phys.*, 2018, **149**, 144107.

65 B. Jiang, J. Li and H. Guo, *Int. Rev. Phys. Chem.*, 2016, **35**, 479.

66 R. Biswas, R. Rashmi and U. Lourderaj, *Resonance*, 2020, **25**, 59.

67 R. Biswas, U. Lourderaj and N. Sathyamurthy, *J. Chem. Sci.*, 2023, **135**, 22.

68 R. Biswas, F. A. Gianturco, K. Giri, L. González-Sánchez, U. Lourderaj, N. Sathyamurthy and E. Yurtsever, *Artif. Intell. Chem.*, 2023, **1**, 100017.

69 A. Kushwaha and T. J. Dhilip Kumar, *Int. J. Quantum Chem.*, 2023, **123**, e27007.

70 J. Dai and R. V. Krems, *Mach. Learn.: Sci. Technol.*, 2023, **4**, 045027.

71 J. Dai and R. V. Krems, *J. Chem. Phys.*, 2022, **156**, 184802.

72 K. Asnaashari and R. V. Krems, *Mach. Learn.: Sci. Technol.*, 2021, **3**, 015005.

73 H. Sugisawa, T. Ida and R. V. Krems, *J. Chem. Phys.*, 2020, **153**, 114101.

74 J. Dai and R. V. Krems, *J. Chem. Theory Comput.*, 2020, **16**, 1386.

75 R. V. Krems, *Phys. Chem. Chem. Phys.*, 2019, **21**, 13392.

76 M. Meuwly, *Chem. Rev.*, 2021, **121**, 10218.

77 D. Bossion, G. Nyman and Y. Scribano, *Artif. Intell. Chem.*, 2024, **2**, 100052.

78 J. Arnold, D. Koner, S. Kaser, N. Singh, R. J. Bemish and M. Meuwly, *J. Phys. Chem. A*, 2020, **124**, 7177.

79 D. E. Mihalik, R. Wang, B. H. Yang, P. C. Stancil, T. J. Price, R. C. Forrey, N. Balakrishnan and R. V. Krems, *J. Chem. Phys.*, 2025, **162**, 024116.

80 R. Wang, D. E. Mihalik, B. H. Yang, P. C. Stancil, T. J. Price, R. C. Forrey, N. Balakrishnan and R. V. Krems, *Astrophys. J.*, 2025, **991**, 140.

81 E. Torabian and R. V. Krems, *Phys. Rev. Res.*, 2023, **5**, 013211.

82 A. Dawid, J. Arnold, B. Requena, A. Gresch, M. Płodzień, K. Donatella, K. A. Nicoli, P. Stornati, R. Koch, M. Büttner, *et al.*, *arXiv*, 2022, preprint, arXiv:2204.04198, DOI: **10.48550/arXiv.2204.04198**.

83 P. Kairon, J. Jäger and R. V. Krems, *arXiv*, 2025, preprint, arXiv:2501.07433, DOI: **10.48550/arXiv.2501.07433**.

84 Y. Zhang and Q. Ni, *Quantum Eng.*, 2020, **2**, e34.

85 M. Cerezo, G. Verdon, H. Y. Huang, L. Cincio and P. J. Coles, *Nat. Comput. Sci.*, 2022, **2**, 567.

86 C. Ciliberto, M. Herbster, A. D. Ialongo, M. Pontil, A. Rocchetto, S. Severini and L. Wossnig, *Proc. R. Soc. A*, 2018, **474**, 20170551.

87 M. Schuld, I. Sinayskiy and F. Petruccione, *Contemp. Phys.*, 2015, **56**, 172.

88 D. P. Kingma, *arXiv*, 2014, preprint, arXiv:1412.6980, DOI: **10.48550/arXiv.1412.6980**.

89 M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015.