Machine learning prediction of state-to-state rate constants for astrochemistry

Duncan Bossion, Gunnar Nyman, Yohann Scribano

Artificial Intelligence Chemistry 2 (2024) 100052

Contents lists available at ScienceDirect

Artificial Intelligence Chemistry

journal homepage: www.journals.elsevier.com/artificial-intelligence-chemistry

Machine learning prediction of state-to-state rate constants for astrochemistry

Duncan Bossion [a,b], Gunnar Nyman [a], Yohann Scribano [c,*]

[a] Department of Chemistry and Molecular Biology, University of Gothenburg, SE-405 30, Gothenburg, Sweden
[b] Institut de Physique de Rennes, UMR-CNRS 6251, Université de Rennes, F-35000 Rennes, France
[c] Laboratoire Univers et Particules de Montpellier, UMR-CNRS 5299, Université de Montpellier, Place Eugène Bataillon, 34095 Montpellier, France

ARTICLE INFO

Keywords:
Astrochemistry
Collisional molecular processes
Methods
Machine learning
Quantum and quasi-classical molecular reaction dynamics

ABSTRACT

In this work, we investigate the possibility to use an artificial neural network to predict a large number of accurate state-to-state rate constants for atom-diatom collisions, from available rates obtained at two different accuracy levels, using a few accurate rates and many low-accuracy rates. The $H + H_2 \rightarrow H_2 + H$ chemical reaction is used to benchmark our neural network, as both low and high accuracy state-to-state rates are available in the literature. Our artificial neural network is a multilayer perceptron, using 8 input neurons including the low-accuracy rate constants, with the high accuracy rate constants as the output neuron. The use of machine learning to predict rate constants is very encouraged, as the rates obtained are accurate, even using as low as 1% of the full dataset to train the neural network, and improve greatly the low accuracy rates previously available. This approach can be used to generate full rate constant datasets with a consistent accuracy, from sparse rates obtained with various methods of different accuracies.

1. Introduction

The interpretation of astronomical molecular spectra requires the knowledge of the population of the molecules in their rovibrational states. Moreover, in the interstellar medium, Local Thermal Equilibrium (LTE) is not always established and an accurate knowledge of the reaction rate constants, at the state-to-state level, is required to solve the radiative transfer equations [1,2]. Experimental determination of those state-to-state reaction rates is very challenging, and they can fortunately be determined thanks to the state-of-the-art of computational chemistry [3,4,5]. Those calculations are generally conducted under the Born-Oppenheimer approximation[6], in which ab initio quantum chemistry calculations are done at fixed nuclear geometries in order to produce a global potential energy surface (PES) able to describe all possible chemical rearrangement involved in the collisional process. For an atom-diatom collision, the state-to-state processes involving diatomic species as reactant/product can be described as:

$$\mathrm{A} + \mathrm{BC}(v,j) \rightarrow \begin{cases} \mathrm{A} + \mathrm{BC}(v',j') & \text{inelastic arrangement}, \\ \mathrm{B} + \mathrm{AC}(v',j') & 1^{st} \text{ reactive arrangement}, \\ \mathrm{C} + \mathrm{AB}(v',j') & 2^{nd} \text{ reactive arrangement} \end{cases} \quad (1)$$

where v and j are the vibrational and rotational quantum numbers of a diatomic molecule, respectively. The kinetic efficiency of those elementary processes is specified through the state-to-state reaction rate defined as:

$$k_{v'j'\leftarrow vj}(T) = \left(\frac{8}{\pi\mu k_B^3 T^3}\right)^{1/2} \int_0^\infty \sigma_{v'j'\leftarrow vj}(E_c) E_c e^{-\frac{E_c}{k_B T}} dE_c \quad (2)$$

where $\sigma_{v'j'\leftarrow vj}(E_c)$ is a state-to-state resolved collisional cross-section, E_c is the collisional energy, k_B is Boltzmann's constant, μ is the reduced mass for the A–BC motion, and T is the temperature. Collisional cross-sections and reaction rate constants are computed using either the Quasi-Classical Trajectory (QCT) [see [3,7,8, 9,10,11], and references therein] or the Time Independent Quantum Mechanical (TIQM) [see [4, 5,12,13,14], and references therein] approaches and have been provided to the astrophysical/astrochemistry community during the last three decades. The adopted methods intrinsically have different ranges of applicability in terms of temperatures and internal quantum states of the reactants. Exact TIQM calculations are obviously the most accurate, especially at low temperature, where quantum effects such as tunnelling, zero point energy or resonances have a high impact. However, they are difficult to implement for calculations at high temperatures in terms

* Corresponding author.
E-mail addresses: duncan.bossion@univ-rennes.fr (D. Bossion), nyman@chem.gu.se (G. Nyman), yohann.scribano@umontpellier.fr (Y. Scribano).

https://doi.org/10.1016/j.aichem.2024.100052
Received 13 July 2023; Received in revised form 17 January 2024; Accepted 30 January 2024
Available online 3 February 2024

of computing capacities since many rovibrational states/channels have to be included in the partial wave expansion of the scattering wavefunction. Indeed, memory and CPU costs scale as the square and the cube of the total number of rovibrational channels, respectively. The close-coupling approach becomes prohibitive when the number of coupled-channels exceed $\sim$ 10000, and thus limits in practice its application to low temperature (up to a few thousand kelvins) or collisional energy regime. Moreover, this prevents the use of such sophisticated and accurate methods for heavy triatomic systems, or more generally for polyatomic systems. On the contrary, although QCT calculations fail at low and very low temperatures, they become more and more accurate when the temperature increases. Moreover, QCT simulations are not limited by the computational time and memory cost specific to TIQM calculations. They can thus provide state-to-state data for all possible ro-vibrational states of the reactant (up to the dissociation limit) and for high temperatures.

In recent years, the use of statistical learning techniques applied to chemical problems has gained considerable interest [15,16]. In particular, supervised machine learning (for which the features related to the data aimed to be predicted are known) has seen a growing interest for chemical reactivity (see the review by [17] and references therein). It was, for example, used to obtain thermal reaction rate constants (see the review by [18] and references therein) of a large number of organic chemical reactions. Machine learning technics were also successfully used as a tool to obtain accurate PES for a fraction of the cost of direct ab initio calculations [19,20], allowing for dynamics studies of reactions involving larger molecules than otherwise accessible. Recently, machine learning was also used in non-adiabatic ab initio chemistry and applied to photochemistry [21,22]. It has also been used for direct applications to astrochemistry to predict binding energies between molecules and surfaces [23], valuable information to study adsorption and desorption of molecules on interstellar ice or dust, or to predict chemical abundances and make chemical inventories of astrophysical media [24]. Some studies focused on machine learning to predict rate constants directly. Thermal rates, having a relatively smooth behaviour with no sharp changes related to the temperature, are ideal candidates for neural network models. Those can be predicted based solely on information concerning the PES, and are typically not of high accuracy, but can complete datasets for reactions with missing data as was done by Komp and Valleau [25] to predict 1.5 million rates, or using exact calculations for the training, leading to more accurate results [26,27,28]. For Non Local Thermal Equilibrium (NLTE) conditions, inelastic state-to-state rate constants have been predicted using Artifical Neural Network (ANN). Indeed, ANN was used to obtain rates outside the accessible range of exact quantum methods [29], leading to rates of limited accuracy, as neural network models are usually more efficient at predicting data lying in the range of the training. The latter was done to complete state-to-state rates for the $\mathrm{N}(^4\mathrm{S}) + \mathrm{NO}(^2\Pi) \rightarrow \mathrm{O}(^3\mathrm{P}) + \mathrm{N}_2(\mathrm{X}^1\Sigma_g^+)$ reaction, using for the training QCT rates widespread over the whole range of ro-vibrational states considered for the predictions [30].

In this work, we present a different approach based on machine learning, using an artificial neural network (ANN), to predict state-to-state rate constants $k_{v'j' \leftarrow vj}(T)$ with high accuracy over a large domain of temperatures of interest. We benchmark our approach on the $\mathrm{H} + \mathrm{H}_2 \rightarrow \mathrm{H}_2 + \mathrm{H}$ chemical reaction for which we have a large set of data (computed with both TIQM and QCT methods). Our approach uses a limited number of high-accuracy rate constants, and can include low-accuracy rates, in order to complete large ensembles of state-to-state rate constants required in particular for applications in astrophysical media under NLTE conditions like photon-dominated regions. In Sec. 2 we briefly describe the basic concept of artificial neural network and present our architecture. In Sec. 2.2 the data manipulations are extensively discussed, including the choice of input parameters and the transformations that we apply in order to generate data efficiently usable by an artificial neural network. Sec. 3 explains the protocol, specifying the metrics we use to define the quality of our models. We present and discuss the results in Sec. 4, followed by conclusions in Sec. 5.

2. Artificial neural network for state-to-state rate constants

The artificial neurons of neural networks are inspired by the biological neurons [31]. Even if the neuron model used in the algorithms is highly simplified, it retains the same core principle. Each neuron is connected to other neurons that pass it weighted values, similarly to biological neurons for which each dendrite, connected to a different neuron, has its own sensibility to a received signal. After receiving and summing the weighted values obtained from the previous artificial neurons, each neuron applies an activation function and transmits the results to the next connected neurons. This activation function is used to introduce non-linearity in the network and allows building accurate models when the relation between the input data (the features) and the output data (the labels) is highly non-linear.

In supervised machine learning, the algorithm is trained at a specific task on a set of data called the training database, for which input and output data are known. During this stage, the weights of each neuron are updated by a back-propagation technique, minimizing the error computed by comparing the output value of the network to the true value (given by the labels). In order to converge the weights, the network is looped over many times until the error reaches a threshold value. The network configuration and the weights of each neuron build the model. The ANN makes predictions by applying this model to a set of features different from the ones provided during the training stage and for which no label is available.

2.1. Multilayer perceptron architectures

The multilayer perceptron (MLP) architecture is one of the most typical neural networks [32]. It consists of three types of layers of nodes: an input layer, at least one hidden layer, and an output layer. Each node of the hidden layers is a neuron that uses a non-linear activation function $f(x)$, with x an input of this neuron. The two most commonly used functions are the hyperbolic tangent and the Rectify Linear Unit [ReLU; 33]. In this work, we use the latter, defined by

$$f(x) = \max(x, 0) = \frac{x + |x|}{2} = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

also known as ramp function. All MLPs are fully connected, with each node in a given layer connected to all the neurons of the following layer (with a neuron-specific weight). A bias node is usually added to each layer, excluding the input layer, and is only connected to the neurons of this layer. It has no activation function (equivalent to a linear activation function) but has a neuron specific weight that is also optimized with the back-propagation technique. The bias allows a shift of the activation function of each neuron. Within this work, we consider two different MLP architectures:

a) The first neural network, MLP1, takes as input a set of 8 features, which are described in Sec. 2.2. The network consists of 4 hidden layers composed of 64, 64, 64 and 48 neurons, respectively, as well as bias neurons (a scheme of MLP1 is presented in Fig. 1). Each neuron of the hidden layers applies the ReLU activation function. In total the number of adjustable parameters is 12 065.
b) The second neural network, MLP2, takes as input a set of 7 features (which does not include the low-accuracy state-to-state rate constants as explained bellow). It is composed of 4 hidden layers of 32, 48, 64 and 64 neurons, respectively, as well as bias neurons, accounting for 9201 adjustable parameters.

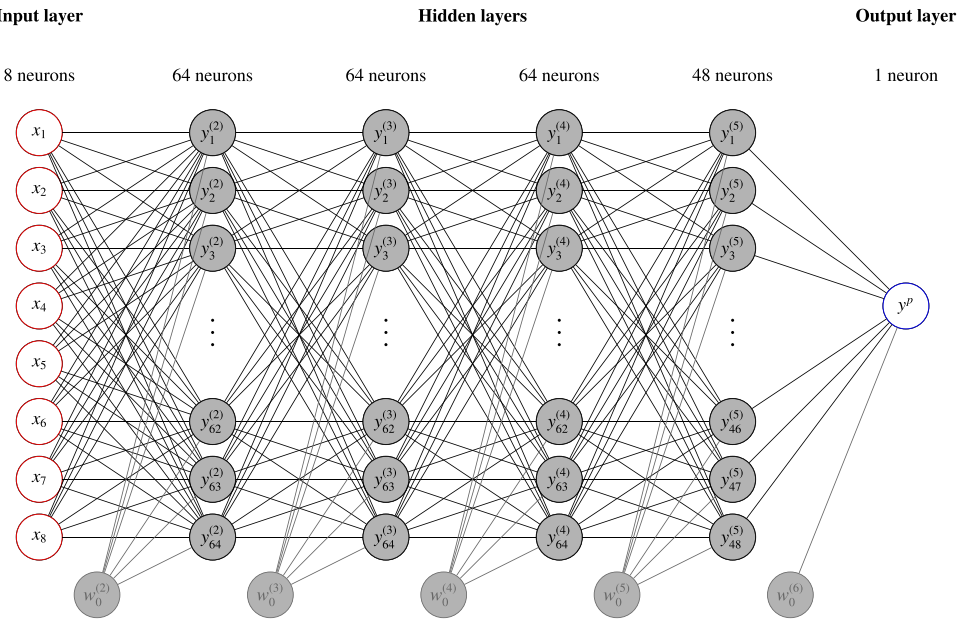

Fig. 1. Our multilayer perceptron architecture MLP1, composed of 4 hidden layers of 64, 64, 64 and 48 neurons each and 8 features, $x_1 \ldots x_8$, for the input neurons and one output neuron, y^p. Bias neurons, $w_0^{(L)}$ connected to the hidden layers, and to the output layer, $L = 6$, are represented. We also use in this work another multilayer perceptron, MLP2, of 4 hidden layers of 32, 48, 64 and 64 neurons each, with only 7 features.

The algebraic expression of the output $y_j^{(L)}$ of the j-th neuron in the L-th layer in our MLP is recursively defined as:

$$y_j^{(L)}(\mathbf{X}) = f\left(w_{0,j}^{(L)} + \sum_{i=1}^{N_L} w_{i,j}^{(L)} y_i^{(L-1)}(\mathbf{X})\right), \tag{4}$$

with $\mathbf{X} = \{x_1, x_2, \ldots, x_8\}$, N_L the number of neurons in the L-th layer, $w_{0,j}^{(L)}$ the bias of the j-th neuron of this layer, $w_{i,j}^{(L)}$ the weight applied by the j-th neuron of the L-th layer to the output of the i-th neuron of the $(L-1)$-th layer named $y_i^{(L-1)}$, and finally $f(x)$ is the activation function defined in Eq. (3). For the first layer (the input layer), we have $y_j^{(1)}(\mathbf{X}) = x_j$. For the output of the single neuron of the last layer (the output layer), we define $y_1^{(6)}(\mathbf{X}) \equiv y^p$. To control the overfitting of the data, we monitor the loss on both the training and a validation set (defined in Sec. 2.2) during the training stage. This loss, which is standard for this type of MLP, is the mean square error regression, that is the square of the difference between the predicted and the true value, completed by a L2-regularization term that penalizes complex models:

$$\text{Loss} = \frac{1}{2n}\sum_i^n |y_i - y_i^p|^2 + \frac{\alpha}{2n}\sum_i^n |w_{i,j}|^2 \tag{5}$$

where α is a parameter that controls the magnitude of the penalization, y_i is a label and y_i^p its corresponding predicted value, and the index i loops over all the n elements of the training dataset. This work uses the Scikit-learn package in Python [34] to create the MLPs.

2.2. Datasets of the neural network

We aim to use machine learning, in particular an ANN, to predict accurate state-to-state rate constants based on a limited number of accurate but numerically expensive rates, and possibly using a high number of numerically cheap but less accurate rates. Two sets of data have to be considered:

- the training dataset, for which the accurate rates are known and will be used as labels,
- the dataset, for which no accurate data is known (we may have low accuracy data available).

In order to assess the accuracy and efficiency of our ANN model to predict state-to-state rate constants, we use it on the H_3 system, in particular on the $H + H_2 \rightarrow H_2 + H$ reaction. This system presents the advantage to have been thoroughly studied in the past with various methods [3,9,35,36]. We will consider as high accuracy rates, the state-to-state rates obtained on this system by TIQM calculations [4,5,14]. All the state-to-state rate constants (considering transitions between all the possible rovibrational states) have also been obtained by the QCT method by some of the authors [10] and will be used as the low accuracy rates.

Predicting data of interest using machine learning requires first building the predictive model. This model is built during the learning stage, for which two subsets of the training dataset are required: a list of inputs called features, and the expected outputs called labels. The features have to be carefully chosen in order to optimize the learning of the algorithm. Features that are directly correlated have to be avoided to not bias the algorithm and overfit the data. On the other hand, any feature of importance, independent of the other features, has to be included to avoid underfitting the data, and hence missing important behaviour. The features that we define for the state-to-state rate constants are i) the initial and final rotational and vibrational quantum numbers $\{j, v\}$ and $\{j', v'\}$, respectively (four input neurons), ii) the initial and final internal energies $E_{v,j}$ and $E_{v',j'}$ respectively for the diatomic fragments (2 input neurons), iii) the temperature T (1 input neuron). The low accuracy data, here the state-to-state rate constants obtained with QCT calculations ($k^c_{v',j' \leftarrow v,j}(T) \equiv k^c_i$, with i an index that runs over all possible rovibrational transitions $v, j \rightarrow v', j'$ and for some selected temperatures) are also possible input data (1 further input neuron). The labels, representing the high accuracy data, are the state-to-state rate constants ($k^q_{v',j' \leftarrow v,j}(T) \equiv k^q_i$) obtained by TIQM calculations for a limited number of rovibrational transitions. We should also emphasize that the efficiency of the training is ensured by the consideration of both the rovibrational quantum numbers and their associated rovibrational energies. Indeed, some rovibrational energy levels can be quasi-degenerate and the use of both features lead to an easier characterization of the transition in the

training step.

3. Methodology and metrics of the algorithm

3.1. Rescaling the datasets

Once the determination of the features is done, they have to be scaled in the most pleasing way for the algorithm, to obtain the best predictions. This includes having features that have zero-mean and unit-variance. Outliers can be problematic as they will concentrate all the remaining inliers to very close values. Our features are all (except the feature k_i^c) bounded values that do not contain outliers, hence the easiest and most appropriate operation is a standard scaling, consisting in removing the mean, and scaling to unit-variance. This has to be performed for each feature x_r independently, $x_r^{\text{scal}} \equiv \widetilde{x}_r = \frac{x_r - \overline{x_r}}{\sigma}$. This scaling cannot be applied to the k_i^c features, as the rate constants contain possible outliers and have values too small for a neural network to perform well (values ranging from 10^{-29} to 10^{-8}). We hence first take the negative logarithm (with $\log \equiv \log_{10}$) in order to obtain a distribution of values between 8 and 29. The distribution is still very uneven and not properly scaled, with outliers at the largest values (corresponding to the smallest rates). To solve this issue, we apply a power transformation, that will lead to a more gaussian-like distribution with a zero-mean and unit-variance, in particular we apply the Yeo-Johnson transform [37], defined by

$$\widetilde{x}_r^{(\lambda)} = \begin{cases} \Big((x_r+1)^{\lambda}-1\Big)/\lambda & \text{if } \lambda \neq 0, \quad x \geq 0, \\ \log(x_r+1) & \text{if } \lambda = 0, \quad x \geq 0, \\ -\Big((-x_r+1)^{2-\lambda}-1\Big)/(2-\lambda) & \text{if } \lambda \neq 2, \quad x < 0, \\ -\log(-x_r+1) & \text{if } \lambda = 2, \quad x < 0. \end{cases} \tag{6}$$

It is a parametric function of λ, which is optimised through maximum likelihood.

The labels k_i^q also need to be transformed. Usually, the transformation or scaling of the label is not necessarily similar to the transformation applied to the features, and is done separately. Here, the nature of k_i^q and k_i^c is the same. We hence chose to use the same procedure, and in particular the same transformation, for those two variables. We define the transformation on the labels, as it is the quantity of interest that we want the most adapted for the learning algorithm. We then apply this transformation with its parameters to k_i^c. The opposite was tested and leads to less accurate models, as the transformation is then adapted to the feature k_i^c, and applying it to k_i^q leads to labels with non zero-mean and unit-variance. Using the same procedure but executing the transformation independently for the labels and the input rates leads to less accurate models, as the correspondence between these two variables is then lost. Fig. 2 shows the distribution of the k_i^q feature before and after the described transformation is applied. The data are predicted with the same rescaling, hence to obtain the final value (in our case the rate constants), the inverse transformations have to be applied.

3.2. Performance metrics

We consider a full dataset composed only of data for which we have both the features and the labels. Furthermore, to test both the MLP1 and the MLP2, we select a dataset for which we have both TIQM and QCT rate constants available. This makes it possible to compute the error on the predicted data and assess the usability of an ANN to obtain accurate state-to-state rate constants. A metric used to estimate the accuracy of our model is the mean absolute error (MAE),

$$\text{MAE} \equiv \frac{\sum_i^n |\log k_i^p - \log k_i^q|}{n}, \tag{7}$$

with n the size of the training dataset and i is a running index over all the elements of this dataset (all the $v,j \to v',j'$ transitions for all the temperatures available), k_i^p represents the predicted state-to-state rate constants. This error gives an estimation of the average error. The other metric used is the root mean square error (RMSE),

$$\text{RMSE} \equiv \sqrt{\frac{\sum_i^n (\log k_i^p - \log k_i^q)^2}{n}}, \tag{8}$$

which gives a strong penalty for data that strongly deviate from the exact values. These two metrics are complementary, as together they give an estimation of how generally close to the correct values the predictions are, but also of the number of outliers that are far away from the correct value. It is important to note that those errors are not computed on the rate constants directly as they vary by many orders of magnitudes and the low values would be negligible. The errors are computed on the logarithm with base 10 of the rate constants, so a difference of several orders of magnitude will be as penalized for small rates as for large rates.

The full dataset is divided into three subsets: a training set, a validation set, and a test set. The training set is used to optimize the weights of the neural network, which are initially randomly chosen, and build the model. We use the validation set to avoid underfitting or overfitting the training set. This is done by checking the performance of the model on the validation set, that is not used to train the model, after each iteration over the network (during the weight optimization). When the loss on the validation set is not improving by at least 10^{-4} for 10 iterations, the model is considered converged and the training ends. The performance of a given model is obtained by evaluating the loss of the trained model. The test set is used to test the model on a set that was not involved in any way in the training process.

In this work, the full dataset is first divided into two subsets of different proportions (specified later) for which the elements are split after being randomized, one part being the test set, and the remaining

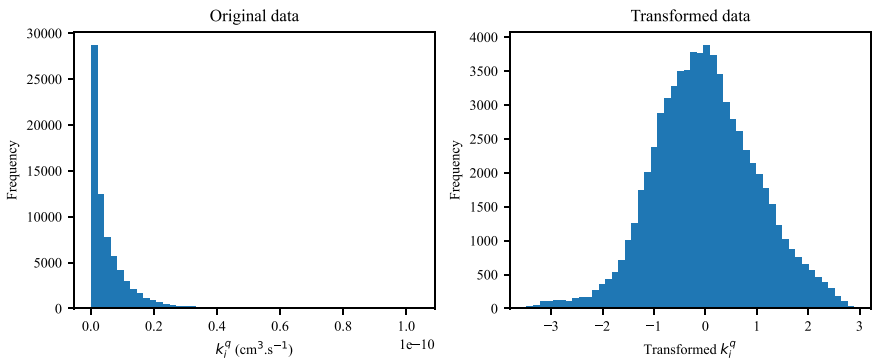

Fig. 2. Histogram of the labels k_i^q, the accurate rate constant with i representing a given rovibrational transition, before and after transformation. We apply a Yeo-Johnson transform [37] the negative of the logarithm with base 10 of the original k_i^q labels. The transformed data have zero-mean and unit-variance.

being the ensemble of data used during the training. Of the latter, 10% are selected at random to create the validation set, and the 90% remaining form the training set.

To choose the best architecture for our MLP once the different subsets are created, and evaluate the stability of the network, we perform five-fold cross-validation. It consists in splitting the training set into five random subsets of identical size and using four of those grouped as a training set and the last one as a validation set (note that this validation set is used to obtain the accuracy of the model when it is converged, as opposed to the validation set mentioned previously which is used to check the convergence of the model and stop the iterations on the network). This procedure is repeated a total of five times, each time defining another of the five subset as the validation set. This presents the advantage of having each data of the training set used both as training (in four of the five training sets) and as validation (in one of the validation sets). The loss used to choose the best architecture is taken as the best average loss over the five splits. In order to choose among different architectures with a similar average loss, the smallest standard deviation is favoured.

4. Results and discussion

We present in Fig. 3 the MAE (left panel) and the RMSE (right panel) of the QCT results (blue dotted line) and of the predicted results, including (green solid line) or not (red dashed lines) the k_i^c feature as input neuron hence using respectively our MLP1 and MLP2. We compute these errors with respect to the size of the training dataset considered, in proportion to the full dataset, starting with as low as 1% of the full dataset used for the training, to 50%. Table 1 presents the number of data in the training and test datasets, depending on the fraction of the full dataset considered for the training. The maximum efficiency of the MLP is clearly lying where the size of the training dataset is very small compared to the size of the full dataset, which is a behaviour of interest for our application. Above a training dataset size of around 20% of the full dataset, the improvement on the MAE and RMSE is very small, contrarily to small training datasets. In particular, between 1% and 10%, the accuracy of the predictions improves a lot by small increases of the size of the training dataset. It is interesting to note that when exact values of at least 5% of the total amount of rates are available, widespread over the entire energy range, the predicted data using only those exact rates to train the model will generate results more accurate than the ones obtained by direct QCT calculations. On the other hand, if this training dataset is a small fraction of the full dataset, adding the QCT value as an input neuron is highly improving the accuracy of the predictions, with an MAE of the predictions about two times smaller than predictions using only 7 input neurons (excluding k_i^c), and an RMSE about 1.5 times smaller. Overall, adding the 8-th input neuron is always leading to more accurate predictions, hence it is advised to use those low accuracy data when available. The biggest improvement over the QCT rates is observed on the RMSE value, showing that the number of outliers is dramatically decreased in the predictions, with a value of the RMSE at least twice smaller even using only 1% of the full dataset for the training. It is important to note that we add a layer of complexity in this choice of data, as all the channels lead to the same product (here the two reactive channels and the inelastic channel lead to H_2). This is impacting the performances of the neural network, in particular because the reactions conserving the parity of the rotational quantum number will include both contributions from the reactive and the inelastic channels, whereas the reactions not conserving this parity will only include the reactive channels (due to ortho-/para-H_2 conservation). The behaviour of the inelastic and reactive channels is different, with tunnelling impacting the reactive channel. The neural network performs even more efficiently for reactions in which the state-to-state rate constants of the different channels are treated separately, as it means a more direct correlation between the inputs and outputs. This is illustrated in Fig. 4 for the reactive channel.

Table 1
Number of state-to-state rate constants in the training and test datasets, depending on the fraction of the full dataset used for the training.

Fraction used for training	0%	1%	10%	50%
Training dataset size	0	715	7155	35775
Test dataset size	71550	70835	64395	35775

Fig. 4 shows the MAE and RMSE, considering only the state-to-state rates of reactive collisions. As expected, the predictions are better than considering a mixture of reactive and inelastic channels. The QCT rates are less accurate as tunnelling plays an important role in reactive collisions. Similarly to Fig. 3, for very small fractions of the full dataset used for the training, adding the QCT feature improves the predictions. On the other hand, contrarily to Fig. 3, the predictions using only 7 features are at least as accurate as the direct QCT calculations even with very small datasets. In addition, for training datasets above around 10% of the full dataset, adding the QCT rate as a feature does not lead to more accurate predictions. This can be explained by a stronger correlation between input and output when only reactive collisions are considered, making the predictions with only 7 features more accurate, while QCT rates are less accurate and may deviate the predictions of the model from the exact values.

To have a better insight on the accuracy of the predictions, we represent in Fig. 5 the MAE for the MLP using 8 features and with a training dataset representing 1% (green solid line) and 10% (red dashed line) of the full dataset. It is clear that the biggest improvement over the QCT rates is situated at low temperatures, where the QCT calculations

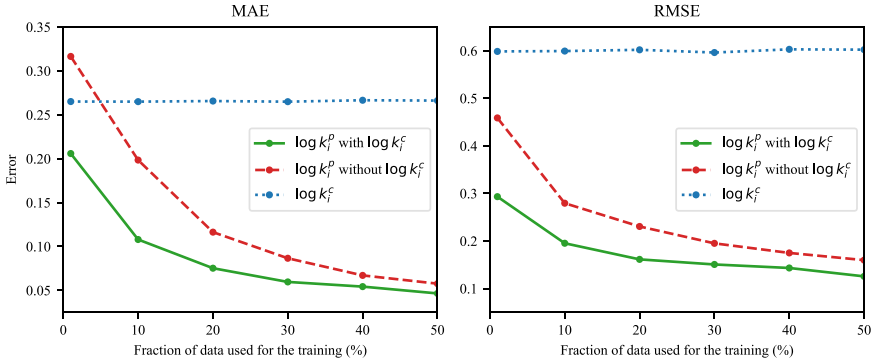

Fig. 3. Mean absolute error (left panel) and root mean squared error (right panel) of the predicted logarithm of the rate constants, $\log k_i^p$, using our artificial neural network, depending on the fraction of the full dataset considered for the training (from 1% to 50%), to predict the remaining fraction. Green solid lines: the MLP1 architecture has 8 features, including the QCT rate constants, k_i^c; red dashed lines: the MLP2 architecture, only considering 7 features, excluding the QCT rate constants; blue dotted lines: the errors obtained directly from the QCT calculations.

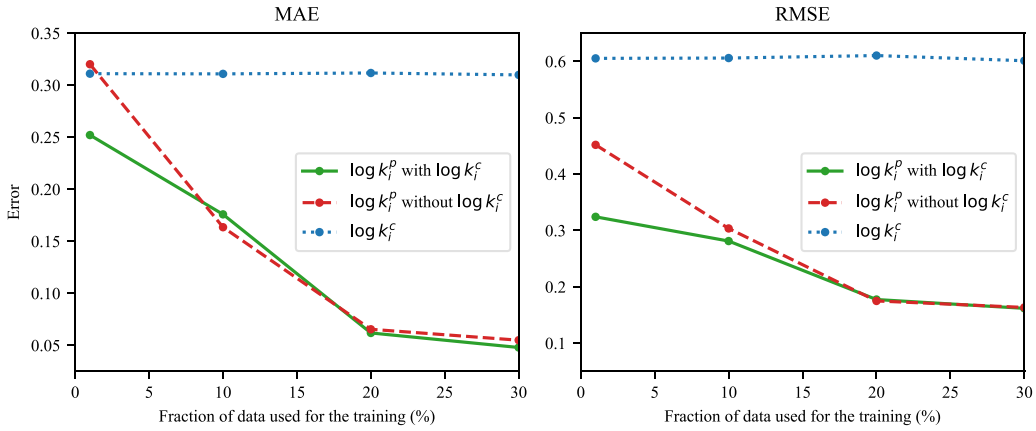

Fig. 4. Mean absolute error (left panel) and root mean squared error (right panel) of the predicted logarithm of the rate constants, $\log k_i^p$, considering only the purely reactive rates, depending on the fraction of the full dataset considered for the training (from 1% to 30%) to predict the remaining fraction. Green solid lines: the MLP1 architecture has 8 features, including the QCT rate constant, k_i^c; red dashed lines: the MLP2 architecture, only considering 7 features; blue dotted lines: the errors obtained directly from the QCT calculations.

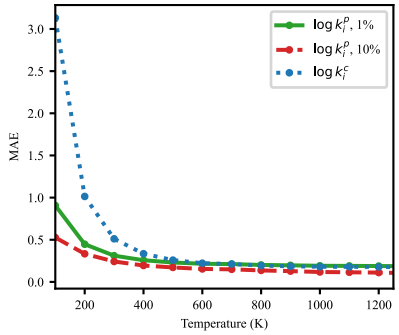

Fig. 5. Mean absolute error of the predictions (using the 8 features) for 1% (green solid line) and 10% (red dashed line) of the full dataset used for the training, compared to the QCT results (blue dotted line), as functions of the temperature.

fail to capture the quantum effects. Over 600 K, using 1% of the full dataset for the training leads to predictions of the same accuracy as the QCT calculations. The big advantage of using machine learning as shown here is that the predictions have a relatively consistent accuracy over the whole energy and temperature range considered. The RMSE is not presented here but has the same behaviour.

To illustrate those averaged errors, we present in Fig. 6 scatter plots of all the rates of the test dataset at 100 K and at 1000 K. We represent the exact quantum rates on the x-axis and the QCT (blue dots) and predicted rates (orange and red dots) in the y-axis. Predictions using 1% of the full dataset for the training are represented in the top panels (orange dots), while predictions using 10% of the full dataset for the training are represented in the bottom panels (red dots). The predicted data are always more accurate than the QCT results, as expected from the errors represented in Figs. 3 and 5, and we clearly see the improvement of the predictions over the QCT calculations at low temperatures. In particular, at 100 K there is a large difference between the distribution of the QCT data and that of the predicted data, including the predictions with a training dataset of 1% of the full dataset. While the predictions are not exact, they are distributed almost symmetrically around the exact values (black solid line). A large correction in particular of the missing tunnelling effect is applied by our MLP to the QCT values given as input to generate the predictions. At high temperature (here 1000 K) the QCT and predicted rates using 1% of the full dataset for the training are of similar accuracy, showing that the neural network is particularly efficient with small training dataset when the difference between the exact and the QCT rates is large. Otherwise, it requires a larger training dataset of at least 10% to show a valuable improvement.

5. Conclusions

We present an artificial neural network able to accurately predict state-to-state rate constants, taking as input the initial and final rovibrational states, the internal energy, the temperature, and possibly a low accuracy state-to-state rate constant. This network is trained and tested on the $H + H_2$ chemical reaction, for which a large amount of data is already available from both high accuracy calculations (using TIQM method) and low accuracy calculations (using the QCT method), making it an ideal candidate for benchmarking the efficiency of this machine learning approach. After training, we show that the predictions are a major improvement compared to the low accuracy rate constants, in particular in the tunnelling regime, usually very challenging for low accuracy methods like the QCT method.

We show that by computing only about 5% of the total number of rates wanted with a high accuracy method and using those results for the training, an artificial neural network can already predict rates at a higher accuracy than the ones obtained using the QCT method, in only a few minutes on a desktop computer. In addition, when higher accuracy predictions are desired, adding the results obtained with a low accuracy method as input will highly improve the predictions when the number of exact rates available accounts for only a few percent of the total amount of rates to predict.

The present work contains both the $H + H_2$ reactive and inelastic channels in the same dataset. But since those two collisional arrangements (inelastic and reactive arrangements) exhibit different behaviour, we plan to incorporate them as two distinct features in our neural network, this work is ongoing. Thus, the good performance of our model on this benchmark chemical reaction allows us to be very confident in it and its ability to become a standard tool for the astrochemistry community in the production of large number of state-to-state rate constants. This is a major step towards generating complete datasets with consistent accuracy, usable in astrochemical models, with the aim to motivate the use of machine learning, in particular to predict reaction rate constants and molecular cooling functions [2].

Moreover, we should highlight that within our training, we do not explicitly specify the characteristics and size of the molecular reactant/product system. The efficiency of our scheme is based on the use of quantum number levels and internal energy states of the involved molecular species in the chemical reaction. Our methodology can thus be easily extended to more complex systems, such as atom-polyatomic collisional processes in gas phase. In the present study, we have

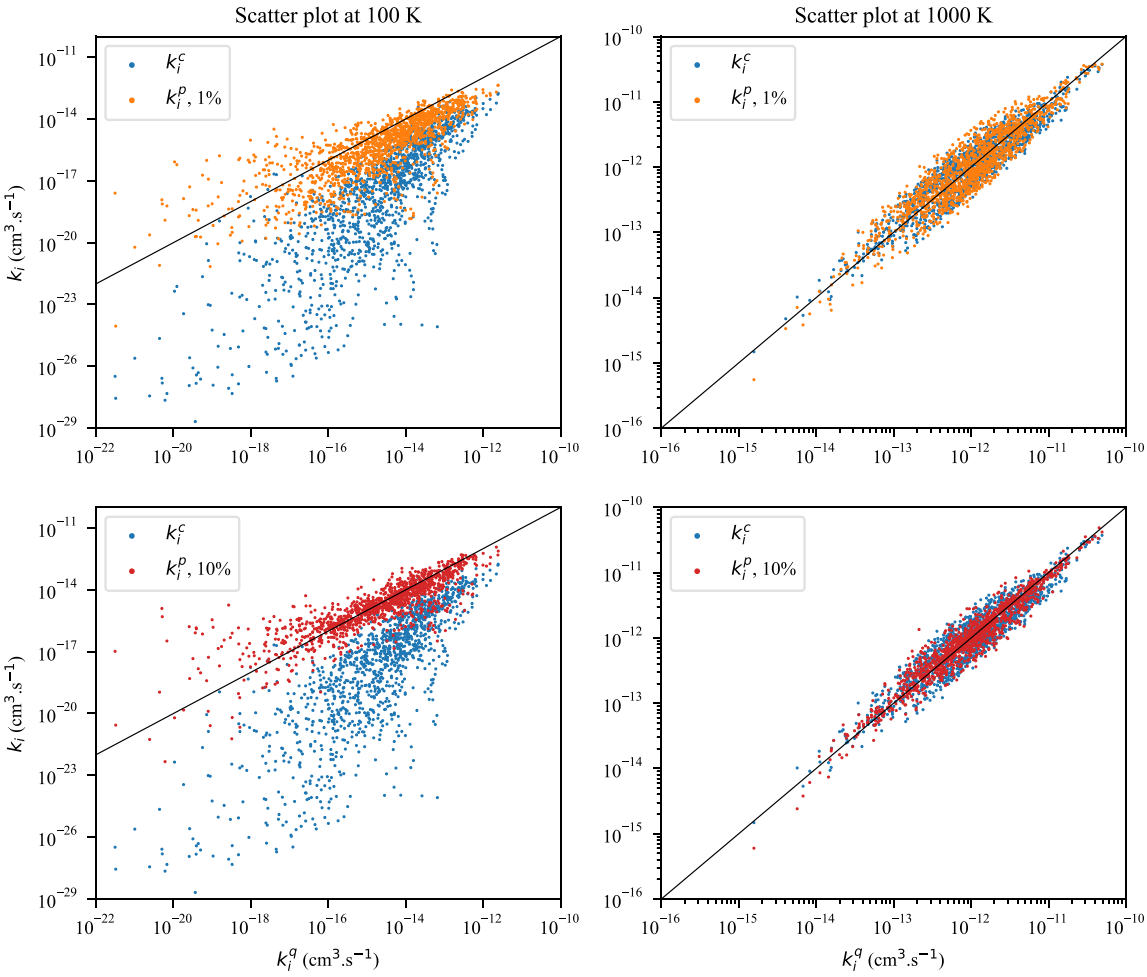

Fig. 6. Rate constants obtained from QCT calculations (k_i^c; blue dots), and predicted with our MLP including k_i^c, k_i^p, using 1% (orange dots) and 10% (red dots) of the full dataset for the training, compared to the exact values of the test dataset (99% and 90% of the full dataset, respectively) for two different temperatures: 100 K (two panels on the left) and 1000 K (two panels on the right) versus exact quantum rates, k_i^q. Perfect agreement is modelled by the black solid line.

incorporated some quantum effects arising at low temperature such as the tunnelling effect thanks to the use of low temperature reaction rate constants computed with the TIQM method. We plan to incorporate other quantum effects (spin-orbit coupling, non-adiabatic effects, …) which can also have an impact on chemical reactivity over a large domain of collisional energy. For high collisional energies and chemical reactions involving several electronic excited states (beyond the Born Oppenheimer picture), we may investigate the use of the Time Dependent Quantum Mechanical method (TDQM), which can be computationally advantageous compared to the TIQM method. We emphasize that our approach, based on data sets computed at two accuracy levels, could also be used in other contexts than chemical reactivity. For example, it could be possible to use a similar approach in the context of rovibrational molecular spectroscopy, for which the production of spectra with several millions of lines is crucial for the characterization of the chemical composition of exoplanetary atmospheres.

CRediT authorship contribution statement

Scribano yohann: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Conceptualization. **Nyman Gunnar:** Writing – review & editing, Writing – original draft, Funding acquisition. **Bossion Duncan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Gunnar Nyman reports financial support was provided by Alice Wallenberg Foundation.

Data availability

The training/validation/test dataset are available upon request.

Acknowledgements

D. B. and Y. S. acknowledge discussions with J. Itam-Pasquet at the early stage of this work. D. B. and G. N. acknowledge support from the Knut and Alice Wallenberg Foundation through grant nr. KAW 2020.0081.

References

[1] C.M. Coppola, S. Longo, M. Capitelli, F. Palla, D. Galli, Vibrational level population of H_2 and H_2^+ in the early universe, Astrophys. J. Suppl. Ser. 193 (7) (2011) 1–11, https://doi.org/10.1088/0067-0049/193/1/7/meta.

[2] M.M. Coppola, F. Lique, F. Mazzia, F. Esposito, M.V. Kazandjian, Temperature and density dependent cooling function for H_2 with updated H_2/H collisional rates, Mon. Not. R. Astron. Soc. 486 (2) (2019) 1590–1593, https://doi.org/10.1093/mnras/stz927.

[3] M.E. Mandy, P.G. Martin, Collisional excitation of H_2 molecules by H atoms, Astrophys. J. Suppl. Ser. 86 (1993) 199. ⟨https://ui.adsabs.harvard.edu/abs/1993ApJS.86.199M⟩.

[4] F. Lique, P. Honvault, A. Faure, Ortho?para-H_2 conversion processes in astrophysical media, Int. Rev. Phys. Chem. 33 (1) (2014) 125–149, https://doi.org/10.1080/0144235X.2014.897443.

[5] F. Lique, Revisited study of the ro-vibrational excitation of H_2 by H: towards a revision of the cooling of astrophysical media, Mon. Not. R. Astron. Soc. 453 (1) (2015) 810–818, https://doi.org/10.1093/mnras/stv1683.

[6] M. Born, R. Oppenheimer, Zur quantentheorie der molekeln, Ann. der Phys. 389 (20) (1927) 457–484, https://doi.org/10.1002/andp.19273892002.

[7] D.G. Truhlar, J.T. Muckerman, Atom-molecule collision theory: A guide for the experimentalist, R. B. Bernstein, 1979.

[8] F.J. Aoiz, L. Bañares, V.J. Herrero, Recent results from quasiclassical trajectory computations of elementary chemical reactions, J. Chem. Soc. Fraday Trans. 94 (1998) 2483–2500. ⟨https://pubs.rsc.org/en/content/articlelanding/1998/ft/a803469i⟩.

[9] F.J. Aoiz, L. Bañares, V.J. Herrero, The H + H_2 reactive system. Progress in the study of the dynamics of the simplest reaction, Int. Rev. Phys. Chem. 24 (1) (2005) 119–190, https://doi.org/10.1080/01442350500195659.

[10] D. Bossion, Y. Scribano, F. Lique, G. Parlant, Ro-vibrational excitation of H_2 by H extended to high temperatures, Mon. Not. R. Astron. Soc. 480 (3) (2018) 3718–3724, https://doi.org/10.1093/mnras/sty2089.

[11] D. Bossion, Y. Scribano, G. Parlant, State-to-state quasi-classical trajectory study of the D + H_2 collision for high temperature astrophysical applications, J. Chem. Phys. 150 (8) (2019) 084301, https://doi.org/10.1063/1.5082158.

[12] P. Honvault, M. Jorfi, T. González-Lezana, A. Faure, L. Pagani, Ortho-Para H_2 conversion by proton exchange at low temperature: an accurate quantum mechanical study, Phys. Rev. Lett. 107 (2) (2011) 023201, https://doi.org/10.1103/PhysRevLett.107.023201.

[13] P. Honvault, Y. Scribano, State-to-State quantum mechanical calculations of rate coefficients for the $D^+ + H_2 \rightarrow HD + H^+$ reaction at low temperature, J. Phys. Chem. A 117 (39) (2013) 9778–9784, https://doi.org/10.1021/jp3124549.

[14] F. Lique, P. Honvault, A. Faure, Ortho?para-H_2 conversion by hydrogen exchange: comparison of theory and experiment, J. Chem. Phys. 137 (15) (2012) 154303, https://doi.org/10.1063/1.4758791.

[15] M. Ceriotti, C. Clementi, O. Anatole von Lilienfeld, Machine learning meets chemical physics, J. Chem. Phys. 154 (16) (2021) 160401, https://doi.org/10.1063/5.0051418.

[16] J.A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, A. Tkatchenko, Combining machine learning and computational chemistry for predictive insights into chemical systems, Chem. Rev. 121 (16) (2021) 9816–9872, https://doi.org/10.1021/acs.chemrev.1c00107.

[17] M. Meuwly, Machine learning for chemical reactions, Chem. Rev. 121 (16) (2021) 10218–10239, https://doi.org/10.1021/acs.chemrev.1c00033.

[18] E. Komp, N. Janulaitis, S. Valleau, Progress towards machine learning reaction rate constants, Phys. Chem. Chem. Phys. 24 (2022) 2692–2705, https://doi.org/10.1039/D1CP04422B.

[19] S. Käser, M. Meuwly, Transfer learned potential energy surfaces: accurate anharmonic vibrational dynamics and dissociation energies for the formic acid monomer and dimer, Phys. Chem. Chem. Phys. 24 (2022) 5269–5281, https://doi.org/10.1039/D1CP04393E.

[20] Y. Hashimoto, T. Takayanagi, T. Murakami, Theoretical calculations of the thermal rate coefficients for the interstellar $NH_3^+ + H_2 \rightarrow NH_4+ +$ H reaction on a new Δ -machine learning potential energy surface, ACS Earth Space Chem. 7 (3) (2023) 623–631, https://doi.org/10.1021/acsearthspacechem.2c00384.

[21] P. Dral, M. Barbatti, Molecular excited states through a machine learning len, Nat. Rev. Chem. 5 (2021) 388–405, https://doi.org/10.1038/s41570-021-00278-1.

[22] B.-X. Xue, M. Barbatti, P. Dral, Machine learning for absorption cross sections, J. Phys. Chem. A 124 (35) (2020) 7199–7210, https://doi.org/10.1021/acs.jpca.0c05310.

[23] T. Villadsen, N.F.W. Ligterink, M. Andersen, Predicting binding energies of astrochemically relevant molecules via machine learning, Astron. Astrophys 666 (2022) A45, https://doi.org/10.1051/0004-6361/202244091.

[24] K.L.K. Lee, J. Patterson, A.M. Burkhardt, V. Vankayalapati, M.C. McCarthy, B. A. McGuire, Machine learning of interstellar chemical inventories, Astrophys. J. Lett. 917 (1) (2021) L6, https://doi.org/10.3847/2041-8213/ac194b.

[25] E. Komp, S. Valleau, Machine learning quantum reaction rate constants, J. Phys. Chem. A 124 (41) (2020) 8607–8613, https://doi.org/10.1021/acs.jpca.0c05992.

[26] P.L. Houston, A. Nandi, J.M. Bowman, A machine learning approach for prediction of rate constants, J. Phys. Chem. Lett. 10 (17) (2019) 5250–5258, https://doi.org/10.1021/acs.jpclett.9b01810.

[27] A. Nandi, J.M. Bowman, P. Houston, A machine learning approach for rate constants. II. Clustering, training, and predictions for the $O(^3P) + HCl \rightarrow OH + Cl$ reaction, J. Phys. Chem. A 124 (28) (2020) 5746–5755, https://doi.org/10.1021/acs.jpca.0c04348.

[28] P.L. Houston, A. Nandi, J.M. Bowman, A machine learning approach for rate constants. III. Application to the $Cl(^2P) + CH_4 \rightarrow CH_3 + HCl$ reaction, J. Phys. Chem. A 126 (33) (2022) 5672–5679, https://doi.org/10.1021/acs.jpca.2c04376.

[29] D.A. Neufeld, Rate coefficients for the collisional excitation of molecules: estimates from an artificial neural network, Astrophys. J. 708 (1) (2009) 635, https://doi.org/10.1088/0004-637X/708/1/635.

[30] D. Koner, O.T. Unke, K. Boe, R.J. Bemish, M. Meuwly, Exhaustive state-to-state cross sections for reactive molecular collisions from importance sampling simulation and a neural network representation, J. Chem. Phys. 150 (21) (2019) 211101, https://doi.org/10.1063/1.5097385.

[31] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (4) (1943) 115–133, https://doi.org/10.1007/BF02478259.

[32] T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical learning: data mining, inference and prediction, 2nd edition.,, Springer, 2009. ⟨http://www-stat.stanford.edu/~tibs/ElemStatLearn/⟩.

[33] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, In: Proceedings of the 27th International Conference on Machine Learning, ICML'10, Omnipress, USA, 2010, 807–814.⟨http://dl.acm.org/citation.cfm?id=3104322.3104425⟩.

[34] 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. ⟨https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf⟩.

[35] S.A. Wrathmall, D.R. Flower, The rovibrational excitation of H_2 induced by H, J. Phys. B: At. Mol. Opt. Phys. 40 (16) (2007) 3221, https://doi.org/10.1088/0953-4075/40/16/003.

[36] S.A. Wrathmall, A. Gusdorf, D.R. Flower, The excitation of molecular hydrogen by atomic hydrogen in astrophysical media, Mon. Not. R. Astron. Soc. 382 (1) (2007) 133–138, https://doi.org/10.1111/j.1365-2966.2007.12420.x.

[37] I.-K. Yeo, R.A. Johnson, A new family of power transformations to improve normality or symmetry, Biometrika 87 (4) (2000) 954–959. ⟨http://www.jstor.org/stable/2673623⟩.