Potential Energy Surfaces of H_xO_y

Beryl Aribowo

July 20, 2022

Contents

1 Introduction

1.1 Atoms and molecules

Throughout the discussion of the functional forms of potential energy surfaces, several relevant concepts pertaining to the description of atoms are:

- The **nuclear geometry** defines that molecules consist of nuclei at the center of the atom, which the nuclei itself is surrounded by the **electron cloud**. The electron cloud describes the region which has a high probability of electrons to reside.
- The **atom type** a is an element from the periodic table of the elements, $a \in \{\text{H, He, Li, Be, ..., O, ...}\}$.
- The **nuclear charge** Z_i of the ith atom is equal to the number of the protons within atom i, which is the atomic number within the periodic table of the elements. The example of the nuclear charges is shown in Table 1.
- The **position** $\mathbf{x}_i$ of the atom ith is the location of the nucleus of atom i within the cartesian coordinates of three-dimensional space.
- The **interatomic distance** r_{ij} is the Euclidean distance between atom i and j,

$$r_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||_2 \tag{1}$$

where $||\mathbf{x}||_2 = \sqrt{x_1^2 + x_2^2 + x_3^2}$. Occasionally, the notation R_k ($k = 1, 2, 3...$) is preferred, it denotes the kth interatomic distances, this is relevant when the indices of the formula requires the enumeration of the distances in correlation to the data points, or when the indices are more convenient to enumerate in conjunction to other factors or terms which do not directly correlate to atomic combinations.

Table 1: Nuclear charges (or the atomic numbers) of nine atom types.

a	H	He	Li	Be	B	C	N	O	F	...
Z	1	2	3	4	5	6	7	8	9	...

1.2 Spectroscopic notation

The following description of the spectroscopic notation is obtained from Sharp [**?**]. The molecular spectroscopic notation $I(T)$ describes the state T of molecule I. T is composed of three parts: a letter, a spectroscopic symbol, and molecular orbital designation. The ground state is labeled X and excited states with the same multiplicity as the ground state are labelled in increasing order of energy with capital letters: $A, B, C, ...$; excited states with different multiplicity are labelled with lowercase letters: $a, b, c, ...$ The spectroscopic symbol which shows the spin multiplicity $2S + 1$ is represented by the superscripted number. The quantum number that represents angular momentum is Λ; the Greek letters $\Sigma, \Pi, \Delta...$ (follows $s, p, d...$) signify $\Lambda = 0, 1, 2, ...$

respectively. The subscript g or u, indicates even or odd symmetry respectively under inversion of wavefunction through the center of the moleculer. One denotes the even or odd symmetry under the reflection of wavefunction at any plane passing through both nuclei by the superscripted $+$ or $-$ respectively. For example, $I(T) = \mathrm{H}_2(X^1\Sigma_g^+)$ is a ground state hydrogen with $2S + 1 = 1$, $\Lambda = 1$, even symmetry under inversion of wavefunction, and even symmetry under the reflection of the wavefunction.

For convenience, consider the proposed rules by Wigner & Witmer [**?**] to determine the possible symbols for the diatomic molecular states given a pair of atomic states. For example, two atoms in identical ^3S states can form a diatomic molecule with $^1\Sigma_g^+$, $^3\Sigma_u^+$, or $^5\Sigma_g^+$ states; ^1S_g and ^1P_u possibly will result in $^1\Sigma_g^+$, $^1\Sigma_u^+$, $^1\Pi_g$, or $^1\Pi_u$. The symmetry is g or u if the sum of the angular momentum is even or odd respectively. The simplified correlation rules are shown in Table 2.

Table 2: Possible diatomic terms resulted from atomic terms.

Atomic terms	**Diatomic terms**
$S_g + S_g$ or $S_u + S_u$	Σ^+
$S_g + S_u$	Σ^-
$S_g + P_g$ or $S_u + P_u$	Σ^-, Π
$S_g + P_u$ or $S_u + P_g$	Σ^+, Π
$P_g + P_g$ or $P_u + P_u$	$\Sigma^+(2), \Sigma^-, \Pi(2), \Delta$
$P_g + P_u$	$\Sigma^+, \Sigma^-(2), \Pi(2), \Delta$

1.3 Potential energy surface

The **potential energy surface** (PES) describes the potential energy of a set of nuclear geometries of a particular molecule. By changing the nuclear geometry in a small amount in a desirable direction and evaluating it through the potential energy function, the PES is formed. In the simplest form, the PES is a system composed of a function f which computes the value of energy V relative to the intermolecular distances $\mathbf{R}$:

$$V := f(\mathbf{R}). \tag{2}$$

1.4 The Born-Oppenheimer approximation

The *ab initio* methods aim to find the solution of the non-relativistic solution of the electronic Schrdinger equation. Mainly this is done by the **Born-Oppenheimer** (BO) **approximation** [**?**]. The motivation of BO approximation is to alleviate the difficulty of computing the energy and wave function of molecules, for example the benzene molecule $\mathrm{C_6H_6}$ consists of 12 nuclei and 42 electrons, in a three-dimensional coordinates, results in a Schrodinger equation which is a partial differential eigenvalue equation with 162 variables ((3×12) nuclear + (3×42) electronic). The BO approximation is used to separate the (quantum mechanical) motion of the electrons from the motion of the nucleus, due to the fact that the ratio of the mass of electrons and nucleus are very large, which implies that given the same amount of kinetic energy, the nuclei move much more slowly than the electrons.

The main interest lies in the first step of the BO approximation, which is solving the electronic Schrdinger equation, yielding a wave function depending on electrons, formally

$$H_{\text{elec}}\chi(\mathbf{R}) = E_{\text{elec}}\chi(\mathbf{R}), \tag{3}$$

where H_{elec} is the electronic **Hamiltonian** which is the sum of electronic {kinetic energies, electronic repulsions, internuclear repulsions, electron–nuclear attractions}; $\chi(\mathbf{R})$ is the eigenfunction; $\mathbf{R}$ is the list of electronic coordinates; the eigenvalue E_{elec} is the PES. By varying $\mathbf{R}$, i.e., changing the nuclear geometry in small steps one can measure different values of the PES, this is referred as the **adiabatic approximation**.

On the other hand, **diabatic approximation** is used when the PES is multi-sheeted. The diabatic approach approximates the multi-valued energies through eigenvalues of a matrix (where each element is formed by polynomial functions), where higher eigenvalue represents higher level of energy (e.g., the energy of an excited state molecule), meanwhile the lowest eigenvalue represents the energy of the ground state; in the diabatic approximation, the **conical intersection** may occur, it happens when the curves representing the two eigenvalues intersect; another possible occurrence is when two eigenvalues are very close to each other, referred as the **avoided crossings**. [*** need to rewrite this subsec with more details - the BO eq ***]

1.5 Conical intersection

As mentioned previously, the conical intersection may happen when there exist multiple states of one molecule, the BO approximation breaks down and the PES of each state approaches each other, consequently nonadiabatic events take place.

1.5.1 Conical intersection of two electronic states

Matsika [**?**] summarized that the conical intersection can be described by the Hamiltonian

$$\mathbf{H} = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}, \tag{4}$$

where

$$H_{ij} = \langle \phi_i | H | \phi_j \rangle , \tag{5}$$

ϕ_k is the diabatic two-state basis used to expand the total wave function in place of the eigenfunctions of the electronic Hamiltonian. The eigenvalues of Eq.(4) are

$$E_{\pm} = \bar{H} \pm \sqrt{\Delta H^2 + H_{12}^2}, \tag{6}$$

where

$$\bar{H} = (H_{11} + H_{12})/2, \tag{7}$$

$$\Delta H = (H_{11} - H_{12})/2. \tag{8}$$

[*** introduce the rotation angle which defines diabatic -¿ adiabatic state transformation, rewquires complete BO eq ***] Two conditions must be satisfied for the matrix in Eq.(4) to be degenerate:

$$H_{11} - H_{22} = 0, \tag{9}$$

$$H_{12} = 0, \tag{10}$$

The above conditions describe two state conical intersection are satisfied in $N^{\text{int}} - 2$ subspace, which is referred as the seam or intersection space, where

$$N^{\text{int}} = 3N - 6 \tag{11}$$

denotes the degrees of freedom of a molecule with N-atoms. For example, diatomic molecules only have one degree of freedom, hence the degeneracy conditions will not be satisfied; this is the non-crossing rule (or the avoided crossing) described by NEUMANN & WIGNER [**?**].

1.5.2 Conical intersection of three electronic states

[*** complete the 3 states conical intersection ***]

The description of two-state conical intersection can be extended beyond two-state. MATSIKA & YARKONY [**?**] mentioned that according to the non-crossing rule of NEUMANN & WIGNER [**?**], in the absence of any symmetry, a seam of such conical intersections can exist in a subspace of dimension $N^{\text{int}} - 5$. The low dimensionality of the three-state seam makes the determination of its locus more complicated, hence efficient algorithm(s) for locating the conical intersection of three electronic states is required.

MATSIKA & YARKONY [**?**] described an algorithm to locate the three-state conical intersection.

1.5.3 N-fold degeneracy rule

KATRIEL & DAVIDSON [**?**] presented a general rule of M-fold degeneracy. Consider the BO electronic Hamiltonian for a polyatomic molecule, the matrix elements of this particular Hamiltonian is denoted as H_{ij}. For an $M \times M$ matrix, M-fold degeneracy requires to satisfy $M - 1$ diagonal conditions

$$H_{11} = H_{22} = ...H_{MM}, \tag{12}$$

and $M(M-1)/2$ off diagonal conditions

$$H_{12} = H_{13} = ...H_{1M} = H_{M-1,M} = 0, \tag{13}$$

which sums to $(M-1)(M+2)/2$ conditions; this also applies for higher-order matrix, however, with more complicated conditions for M-fold degeneracy. The maximum degeneracy of a molecule with N^{int} degrees of freedom is given by the largest M satisfying the inequality

$$(M-1)(M+2)/2 \leq 3N - 6. \tag{14}$$

1.6 Quality assessment

To measure the performance of the PES model, the root mean squared error (RMSE) is used. The formal definition of RMSE pertaining to the PES is

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(V_{i,\text{fit}} - V_{i,\text{ab initio}})^2}, \tag{15}$$

where n is the number of the data points (of the whole dataset),

$$V_{i,\text{fit}} := f(R_i) \tag{16}$$

is the energy value from the fitted function f, with the ith molecular distance from the dataset as the input parameter; $V_{i,\text{ab initio}}$ is the corresponding ith *ab initio* energy value. In a fitting scenario, the RMSE is used for both in training and validation scheme, where within the training scheme, RMSE acts as the objective function, i.e., it is used to tune the parameters of the function f, so that the RMSE of the whole dataset satisfies

$$\text{RMSE} \leq \delta, \quad 0 < \delta \ll 1, \tag{17}$$

on a scaled dataset where

$$R_i, V_i \in [0, 1], \text{ for } i = 1, 2, ..., n. \tag{18}$$

1.7 Unit conversion factors

The relevant information for PES particularly the atomic distances and energies are represented in various units throughout various sources of literature. The data available here are represented in $(R, V) =$ (Bohr, Hartree) units, where R is the atomic distance value and V is the energy value. The energy unit conversion factors are shown in Table 3, which is obtained from National Chiao Tung University's website [**?**], for example, 1 Hartree = 27.2107 eV; the atomic distance conversion factor is shown in Eq.(19).

Table 3: Atomic energy conversion factors.

	Hartree	**eV**	**kcal/mol**	**kJ/mol**	**cm^{-1}**
Hartree	1	27.2107	627.503	2625.5	219474.63
eV	0.0367502	1	23.0609	96.486 9	8065.73
kcal/mol	0.00159362	0.0433634	1	4.18400	349.757
kJ/mol	0.00038088	0.01036410	0.239001	1	83.593
cm^{-1}	4.55633×10^{-6}	1.23981×10^{-4}	0.00285911	0.0119627	1

$$1 \text{ atomic unit (a.u.)} = 1 \text{ Bohr} = 0.529177249 \text{ Angstrom (A)}. \tag{19}$$

2 Functional forms

2.1 Permutational-invariant PES (Braams & Bowman 2009)

Historically, the invariance property of the PES was first stated by Murrell et al [**?**], where it was noted that an invariant fitting basis with this can be represented by an integrity basis. In general, any method that incorporates the invariance property into the representation of the PES must express it in the terms of a set of variables that is closed under all permutations, e.g., Cartesian coordinates, internuclear distances, etc. The approach described by Braams & Bowman [**?**] uses monomial symmetrization alongside the invariant polynomial theory, where invariant bases are obtained in terms of **primary** and **secondary polynomials**.

The example of the monomial symmetrization application on acetylene (C_2H_2) by Zou et al [**?**] will lead into the general form of the monomial symmetrization later, the multinomials expansion for the potential is

$$V = \sum_{\substack{a,b,c,d,e,f \\ a+b+c+d+e+f \leq M}}^{M} C_{abcdef} \left[y_{12}^a y_{13}^b y_{14}^c y_{23}^d y_{24}^e y_{34}^f \right], \tag{20}$$

where y_{ij} is the **Morse variable** with the following form

$$y_{ij} = \exp(-r_{ij}/r_0), \tag{21}$$

where H atoms are labelled $1, 2$, and the C atoms are labelled $3, 4$; r_0 is a chosen parameter. The summation over all powers of each y is constrained by the total degree M. The Morse variable indices (consequently the internuclear distances r_{ij}) are written in lexical order $i < j$; N by convention is the number of atoms, which in this case is four. Another possible form of Eq.(21) is $y_{ij} = 1 - \exp(-r_{ij}/r_0)$.

As the number of variables in V increases, the number of terms grows nonlinearly for a fixed M. Monomials with identical coefficients could be grouped into a single polynomial given by the sum of monomials multiplied by one coefficient.

Braams & Bowman [**?**] proceed with the following equation to replace Eq(20):

$$V = \sum_{\substack{a,b,c,d,e,f \\ a+b+c+d+e+f \leq M}}^{M} D_{abcdef} S \left[y_{12}^a y_{13}^b y_{14}^c y_{23}^d y_{24}^e y_{34}^f \right], \tag{22}$$

where S denotes the symmetrization of the monomials. To symmetrize the monomials, the mapping of atom permutations to permutations of the internuclear distances is needed. For example, for A_3 molecules, let 1, 2, 3 denote the initial arrangement of the atoms; consider the permutation: 2, 3, 1. The initial internuclear distances r_{12}, r_{13}, r_{23} are mapped into r_{23}, r_{21}, r_{31}, which is then rewritten into the lexical order of r_{23}, r_{12}, r_{13}. Thus the monomial $r_{12}^a r_{13}^b r_{23}^c$ is mapped into $r_{23}^a r_{12}^b r_{13}^c$ or equivalent to the permuted powers $r_{12}^b r_{13}^c r_{23}^a$. The addition of four other monomials corresponding to the four additional permutations gives a fully symmetrized sum of monomials. Consequently y_{ij} is also permuted analogously, which yields the symmetrized basis. It is now clear that only one term and one unknown coefficient D_{abcdef} appears for every set of tuples (a, b, c, d, e, f) which are permutationally equivalent. The details of the symmetrization on larger and more complex molecule formations are such as A_2B, A_3, A_2B_2, etc, are provided in Table 1-5 of Braams & Bowman [**?**].

2.2 CHIPR (Rocha & Varandas 2021)

The **Combined-Hyperbolic-Inverse-Power-Representation (CHIPR)** was first developed by Varandas [**?**] and assessed for triatomic molecules H_3 and H_2O. It was recently updated by Rocha & Varandas [**?**] for the PES up to tetratomic molecule. In general, for simpler molecules, CHIPR is directly used as the potential energy value; for more complex molecules, for example molecules with multiple states (excited states and ground state), CHIPR acts as the components of each element of a matrix where each eigenvalue of the matrix represents one potential energy value on a particular state of the molecule.

2.2.1 General representation

In CHIPR, The molecular potential energy is represented as the summation of n energy terms:

$$V(\mathbf{r}) = \sum_{i=1}^{n} \mathcal{V}_S^{(i)}(\mathbf{r}), \tag{23}$$

where $\mathbf{r} = (r_{12}, r_{13}, ...)$ is the list of all interatomic distances, and the sub-index S denotes that the functional must reflect the appropriate permutational symmetry. $\mathcal{V}_S^{(1)}$ represents the one-body term (within this particular framework, it is set to 0),

$$\mathcal{V}_S^{(2)}(\mathbf{r}) = \sum_{i<j} V_{ij}^{(2)}(r_{ij}) \tag{24}$$

is the sum of two-body potentials, the sum of the three-body potentials is

$$\mathcal{V}_S^{(3)}(\mathbf{r}) = \sum_{i<j<k} V_{ijk}^{(3)}(r_{ij}, r_{ik}, r_{jk}), \tag{25}$$

$\mathcal{V}_S^{(4)}(\mathbf{r})$ represents the four-body interaction, etc. For example, in the case of tetratomic molecules, the assumed many-body expansion form of the CHIPR method is

$$V(\mathbf{r}) = \mathcal{V}_S^{(1)} + \mathcal{V}_S^{(2)}(\mathbf{r}) + \mathcal{V}_S^{(3)}(\mathbf{r}) + \mathcal{V}_S^{(4)}(\mathbf{r}), \tag{26}$$

The general form of the CHIPR n-body term is

$$V^{(n)}(\mathbf{r}) = \sum_{i_1=0,...,i_\tau=0}^{L} C_{i_1,...,i_\tau} \prod_{p=1}^{\tau} y_p^{i_p}, \tag{27}$$

where $C_{i_1,...,i_\tau}$ are the expansion coefficients (Table I by Varandas [?] provides the example of the index values of the expansion coefficients), y_p $(p = 1, 2, ...\tau)$ is the set of coordinates relative to some reference geometry, where

$$\tau = n(n-1)/2 \tag{28}$$

is the total number of internal degrees of freedom. Every y_p in Eq.(27) is expanded as

$$y_p = \sum_{\alpha=1}^{M} c_\alpha \phi_{p,\alpha}, \tag{29}$$

where c_α are contraction coefficients to be determined from fitting, with α as the index of each primitive function $\phi_{p,\alpha}$. The primitive function $\phi_{p,\alpha}$ has the form of

$$\phi_{p,\alpha} = \begin{cases} \mathrm{sech}^{\eta_\alpha}(\gamma_{p,\alpha}\rho_{p,\alpha}), & \text{if long-range terms are ignored,} \\ \left(\dfrac{\tanh(\beta_\alpha R_p)}{R_p}\right)^{\sigma_\alpha} \mathrm{sech}^{\eta_\alpha}(\gamma_{p,\alpha}\rho_{p,\alpha}), & \text{otherwise,} \end{cases} \tag{30}$$

where R_k, $k \in \mathbb{Z}^+$, $(R_1 := r_{12}, R_2 := r_{13}, R_3 := r_{23}, ...)$ is introduced as the atomic distance notation to simplify the indices' enumeration; the deviation coordinate is expressed as

$$\rho_{p,\alpha} = R_p - R_{p,\alpha}^{\mathrm{ref}}, \tag{31}$$

$\gamma_{p,\alpha}$ represents the nonlinear parameters, and the appearing constants are

$$\begin{aligned} \eta_\alpha &\equiv \eta = 1, \\ \sigma_\alpha &\equiv \sigma = 6, \\ \beta_\alpha &\equiv \beta = 1/5. \end{aligned} \tag{32}$$

To avoid strong linear dependencies, the reference geometries of the various basis functions are related by

$$R_{p,\alpha}^{\text{ref}} = \zeta (R_p^{\text{ref}})^{\alpha-1}, \tag{33}$$

where ζ and $R_{p,0}^{\text{ref}}$ are constants. Table II by VARANDAS [**?**] provides the example of the coefficients' values for ground state OH and O_2.

2.2.2 One-body term

The form of the one-body term is

$$\mathcal{V}_S^{(1)} = \sum_i V_i^{(1)}, \tag{34}$$

$\mathcal{V}_S^{(1)}$ is zero for the atomic ground state and positive for the atomic excited state. In the models for H_xO_y, the relevant states for the one-body term are

$$\begin{aligned} &\text{H}(^2S), \\ &\text{O}(^1D), \text{O}(^3P), \text{O}^+(^4S). \end{aligned} \tag{35}$$

2.2.3 Two-body terms

The functional form of two-body terms (diatomic molecule) is

$$V^{(2)}(r_{ij}) = \frac{Z_i Z_j}{r_{ij}} \sum_{l=1}^{L} C_l y^l, \tag{36}$$

where Z_i and Z_k are the nuclear charges of atoms i and k; the appearing y is the summation factor described in Eq.(29). The relevant states for H_xO_y are

$$\begin{aligned} &\text{H}_2(X^1\Sigma_g^+), \\ &\text{O}_2(^1\Delta_g), \text{O}_2(X^3\Sigma_g^-), \\ &\text{OH}(X^2\Pi), \text{OH}^+(X^3\Sigma^-). \end{aligned} \tag{37}$$

The pair potential energy curves are described in Section 4.

2.2.4 Three-body terms

The three-body CHIPR function of Eq. (27) has the form of

$$V^{(3)}(\mathbf{r}) = \mathcal{D}^{(3)} \sum_{i,j,k=0}^{L} C_{i,j,k} \left\{ \sum^{g\in G} \mathcal{P}_g^{(i,j,k)} \left[y_1^i y_2^j y_3^k \right] \right\}, \tag{38}$$

where $C_{i,j,k}$ are the expansion coefficients subject to several constraints depending on the molecule's species, y_p ($p = 1, 2, 3$) are the coordinates from Eq.(29), g is the permutation element, and G is a subgroup of the $\mathcal{S}_3$ symmetric group; therefore, $\mathcal{P}_g^{(i,j,k)}$ are the operators that reflect the action of the atom permutation g on y_p. Table I in Rocha & Varandas [?] shows the expansion coefficients' constraints and the permutation elements. The additional factor which serves to dampen the three-body terms is the case $n = 3$ of

$$\mathcal{D}^{(n)}(\mathbf{R}) = \left[\prod_{i=1}^{\tau} h(R_i)\right]^{\xi}, \tag{39}$$

where $\mathbf{R} = \{R_1 := r_{12}, R_2 := r_{13}, R_3 := r_{23}, ...\}$, and $h(R_i)$ is expressed as

$$h(R_i) = (1 + \tanh(\kappa(R_i - R_0)))/2, \tag{40}$$

with the chosen parameters are

$$\begin{aligned} R_0 &= 0.5\mathrm{a}_0, \\ \kappa &= 100\mathrm{a}_0^{-1}, \\ \xi &= 10. \end{aligned} \tag{41}$$

The relevant states for H_xO_y are

$$\begin{aligned} &\mathrm{H_2O}(X^1A_1), \\ &\mathrm{HO_2}(X^2A''), \\ &\mathrm{HO_2^+}(X^3A''), \mathrm{HO_2^+}(1^3A''). \end{aligned} \tag{42}$$

2.2.5 Four-body terms

For general tetratomic molecule, the four-body CHIPR function is

$$V^{(4)}(\mathbf{r}) = \mathcal{D}^{(4)} \sum_{i,j,k,l,m,n=0}^{L} C_{i,j,k,l,m,n} \left\{ \sum^{g\in G} \mathcal{P}_g^{(i,j,k,l,m,n)} \left[y_1^i y_2^j y_3^k y_4^l y_5^m y_6^n \right] \right\}, \tag{43}$$

similar to the three-body form, the constraints which depend on the molecular species, are provided in the Table II by Rocha & Varandas [?]. The damping-term $\mathcal{D}^{(4)}$ is also analogous to the three-body's version, with exactly the same quantities to Eq.(39), except the product's end index is now 6. The relevant states of H_xO_y for four-body terms are

$$\mathrm{H_2O_2}(X^1A). \tag{44}$$

In the tetratomic molecule case, in some instances, such as in Coelho & Brando [?] the **reference geometries** are used to better fit the curve around the regions of interest, such as the minima, transition state configurations, and other regions which play important role in the molecular dynamics. The form of the reference geometry is

$$\sum_{rg} P(\mathbf{R})\Theta_{rg}(\mathbf{R}) \tag{45}$$

where the sum is over all reference geometries rg; $P(\mathbf{R})$ is a polynomial function which is expanded by using the reference geometries as the centres, shown as

$$P(\mathbf{R}) = \sum_{k=0}^{M} C_k(\mathbf{R} - \mathbf{R}_{rg})^k, \tag{46}$$

where C_k is the fitted coefficient and $\mathbf{R}_{rg}$ is the reference geometry;

$$\Theta(\mathbf{R}) = \prod_{i=1}^{d} \exp(\frac{\eta_i v_i - v_i^2}{8\lambda_i}) \tag{47}$$

is an exponential range factor term;

$$v_i = (\vec{\psi} \cdot \vec{u_i}) = \sum_{j=1}^{d} \psi_j \times u_{ij}, \tag{48}$$

where the appearing $\vec{u_i}$ is the eigenvector, λ_i is the eigenvalue, η_i is the eigenvector direction, ψ is the displacement integrity factor, and d is the total number of the displacement integrity basis (the complete form and explicit values are shown in Table 2 and Eq.(17) of Coelho & Brando [**?**]).

2.3 CHIPR for HO_2^+ (Xavier et al 2019)

The explicit usage of CHIPR from Subsection 2.2 is demonstrated by Xavier et al [**?**] for HO_2^+. The functional form which represents the total interaction potential is

$$V(\mathbf{r}) = V^{(2)}_{O_2^+}(r_{23}) + V^{(2)}_{OH^+}(r_{12}) + V^{(2)}_{OH^+}(r_{13}) + V^{(3)}_{HO_2^+}(r_{23}, r_{12}, r_{13}), \tag{49}$$

where r_{23} is the distance between oxygen atoms, and r_{12} and r_{13} are the distances between hydrogen and oxygen atom. The diatomic potential's explicit form is the realization of Eq.(36) shown as

$$V^{(2)}(r_{ij}) = \frac{Z_i Z_j}{r_{ij}}(C_1 y_1 + C_2 y_1^2 + C_3 y_1^3), \tag{50}$$

where N_i and N_j are the nuclear charges of atom i and j respectively, C_1, C_2, C_3 are parameters to be fitted. According to Eq.(27), with $L = 3$, without the damping term, the explicit form of the three-body interaction energy is

$$\begin{aligned} V^{(3)}(\mathbf{r}) = {} & C_{110} y_1(y_2 + y_3) + 2C_{011} y_2 y_3 + 2C_{111} y_1 y_2 y_3 \\ & + C_{120} y_1(y_2^2 + y_3^2) + C_{210} y_1^2(y_2 + y_3) + C_{012}(y_2 y_3^2 + y_2^2 y_3^2), \end{aligned} \tag{51}$$

where C_{i_1,i_2,i_3} are the coefficients to be fitted. It can be noted that since $C_{110} = C_{101}$ (permutational symmetry), only one is chosen, analogous argument applies to terms with higher y powers.

The appearing y factors are the contracted basis with general form described in Eq.(29). The y_i is defined as

$$y_i = \begin{cases} \begin{cases} \sum_{k=1}^{4} c_{1k}\ \mathrm{sech}^{\eta_{1k}}(\gamma_{1k}(\left[R_1 - \zeta_1 (R_1^{\mathrm{ref}})^{k-1}\right])) & \text{for } V^{(2)}_{\mathrm{OH}^+}, \\ \sum_{k=1}^{4} c_{1k}(s_k R_1 + t_k)\ \mathrm{sech}^{\eta_{1k}}(\gamma_{1k}(\left[R_1 - \zeta_1 (R_1^{\mathrm{ref}})^{k-1}\right])) & \text{for } V^{(2)}_{\mathrm{O}_2^+}, \\ \sum_{k=1}^{3} c_{1k}(R_1^{\mathrm{ref}})^{1-k}((\tanh\left[\beta_1(R_1 - \zeta_1 (R_1^{\mathrm{ref}})^{k-1})\right]) & \text{for } V^{(3)}_{\mathrm{HO}_2^+}, \\ \quad \times\ \mathrm{sech}^{\eta_{1k}}(\gamma_{1k}\left[R_1 - \zeta_1 (R_1^{\mathrm{ref}})^{k-1}\right])), & \end{cases} & \text{if } i = 1, \\ \sum_{k=1}^{3} c_{ik}((\tanh\left[\beta_i(R_i - \zeta_i (R_i^{\mathrm{ref}})^{k-1})\right]) & \\ \quad \times \mathrm{sech}^{\eta_{ik}}(\gamma_{ik}\left[R_i - \zeta_i (R_i^{\mathrm{ref}})^{k-1}\right])) \times (R_i^{\mathrm{ref}})^{1-k} & \text{for all } V^{(2)} \text{ and } V^{(3)}, \end{cases} \quad \text{for } i = 2, 3 \tag{52}$$

where $R_1 := r_{23}$, $R_2 := r_{12}$, $R_3 := r_{13}$; The coefficients $c_{ik}, \gamma_{ik}, \beta_i, \eta_{ik}, \zeta_i, s_k, t_k$ are obtained from fitting. The complete coefficients' values found are provided in the supplementary materials of Xavier et al [**?**].

2.4 2 x 2 diabatic matrix for HO_2^+ dissociation (Xavier & Varandas 2021)

Employed by Xavier & Varandas [**?**]. The dissociation scheme of HO_2^+ for the double-sheeted PES which will determine the many-body terms is

$$\mathrm{HO}_2^+(X^3A'') \rightarrow \begin{cases} \mathrm{O}_2^+(X^2\Pi_g) + \mathrm{H}(^2S), \\ \mathrm{OH}^+(X^3\Sigma^-) + \mathrm{O}(^3P), \end{cases} \tag{53}$$

$$\mathrm{HO}_2^+(X^3A'') \rightarrow \begin{cases} \mathrm{O}_2(X^3\Sigma_g^-) + \mathrm{H}^+, \\ \mathrm{OH}^+(X^3\Sigma^-) + \mathrm{O}(^1\mathrm{D}), \end{cases} \tag{54}$$

$$\left.\begin{matrix} \mathrm{HO}_2^+(X^3A'') \\ \mathrm{HO}_2^+(1^3A'') \end{matrix}\right\} \rightarrow \mathrm{O}(^3P) + \mathrm{O}^+(^4S) + \mathrm{H}(^2S). \tag{55}$$

The function to model a cusp in the PES is the eigenvalue expression of a 2×2 diabatic matrix, shown as

$$V_d = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \tag{56}$$

where the main diagonal elements are the diabatic functions, while the off-diagonal elements $V_{12} = V_{21} \neq 0$ are coupling potentials. When V_d is diagonalized (numerically), it yields two solutions which are the eigenvalues (the adiabatic energies) in the adiabatic matrix ($V_{12} = V_{21} = 0$). The form of the adiabatic potentials is

$$V_{\pm} = \frac{V_{11} + V_{22}}{2} \pm \frac{1}{2}\sqrt{(V_{11} - V_{22})^2 + 4V_{12}^2}, \tag{57}$$

where $V_{\pm} = \{V_-, V_+\}$ is the set of eigenvalues of matrix described in Eq.(56), V_- is the ground state potential and V_+ is the excited state potential. The V_{ij} terms are formed from the basis of **Double Many Body Expansion (DMBE)** theory as the following:

$$V_{11} = V^{(2)}_{O_2}(X^3\Sigma^-_g; R_1) + V^{(2)}_{OH^+}(X^3\Sigma^-; R_2) + V^{(2)}_{OH^+}(X^3\Sigma^-; R_3) + V^{(3)}_{HO^+_2}(X^3A''; R_1, R_2, R_3), \tag{58}$$

$$V_{22} = V^{(2)}_{O^+_2}(X^2\Pi_g; R_1) + V^{(2)}_{OH^+}(X^3\Sigma^-; R_2) + V^{(2)}_{OH^+}(X^3\Sigma^-; R_3) + V^{(3)}_{HO^+_2}(1^3A''; R_1, R_2, R_3), \tag{59}$$

$$V_{12} = V_{21} = \left[\left(\frac{R_1^2 + R_3^2 - R_2^2}{2R_1R_3}\right)^2 - \left(\frac{R_1^2 + R_2^2 - R_3^2}{2R_1R_2}\right)^2\right] S^2, \tag{60}$$

where $V_{O_2}, V_{O^+_2}$ and V_{OH^+} are two-body terms of diatomic potential modelled by CHIPR (explained in Subsection 2.2). Each term within the summation has the general form of $V(\text{state}; R_i)$, the appearing molecular states can be referred to the dissociation scheme shown in Eq.(53) and (55); R_i $(i = 1, 2, 3)$ represents the atomic distances, where R_1 is the distance between two oxygens (O^1 – O^2), R_2 is the distance between the first oxygen and the hydrogen (O^1 – H), and R_3 is the distance of the hydrogen and the second oxygen (H – O^2). The three-body term ($V_{HO^+_2}$) when modelled by CHIPR, typically has the following form:

$$\begin{aligned} V^{(3)}_{HO^+_2} = \quad & C_{110}y_1(y_2 + y_3) + 2C_{011}y_2y_3 + 2C_{111}y_1y_2y_3 \\ & + C_{120}y_1(y_2^2 + y_3^2) + C_{210}y_1^2(y2 + y3) \\ & + C_{012}(y_1y_3^2 + y_2^2y_3) + C_{121}y_1(y_1^2y_3 + y_2y_3^2) \\ & + ..., \end{aligned} \tag{61}$$

where y_i, $(i = 1, 2, 3)$ are the contracted basis; C_{jkl} are the coefficients obtained from fitting, in this case, a maximum of 10 for the overall power of the polynomial expansion is considered, which results in 34 coefficients; S is CHIPR properly decaying polynomial-type form.

2.5 3 x 3 diabatic matrix for H_2O_2 dissociation (Coelho & Brando 2017)

The PES functional in the form of a 3 × 3 symmetric matrix was employed by Coelho & Brando [?] for tetratomic dissociation of H_2O_2. Consider the dissociation channels as follows:

$$H_2O_2(X,^1A) \rightarrow \begin{cases} OH(X^2\Pi) + OH(X^2\Pi) \\ H_2(X^1\Sigma^+_g) + O_2(a^1\Delta_g) \\ O(^1D) + H_2O(X^1A_1) \\ H(^2S) + HO_2(X^2A'') \\ 2H(^2S) + O_2(^3\Sigma^-_g) \\ H_2(X^1\Sigma^+_g) + 2O(^3P) \\ H(^2S) + O(^3P) + OH(X^2\Pi) \\ 2H(^2S) + 2O(^3P), \end{cases} \tag{62}$$

similar to Subsection (2.4), the channels will determine the summation for each term presents in the matrix. The 3×3 symmetric matrix has the following form:

$$V_l = V_l^T = \begin{bmatrix} V_{11} & V_{12} & 0 \\ V_{21} & V_{22} & 0 \\ 0 & 0 & V_{33} \end{bmatrix}, \tag{63}$$

$$V_{12} = V_{21}, \tag{64}$$

where only the lowest eigenvalue of V_l is considered. Simplifications are assumed where each diagonal term consists of only one, two, and three body terms. The diagonal terms has the following forms:

$$\begin{aligned} V_{11} = \; & V_{\mathrm{O}}^{(1)}(^1D) + V_{\mathrm{O}}^{(1)}(^1D) + V_{\mathrm{O_2}}^{(2)}(^3\Sigma_g^-; R_1) \\ & + V_{\mathrm{H_2}}^{(2)}(^1\Sigma_g^+; R_2) + V_{\mathrm{OH}}^{(2)}(^2\Sigma; R_3) \\ & + V_{\mathrm{OH}}^{(2)}(^2\Sigma; R_4) + V_{\mathrm{OH}}^{(2)}(^2\Sigma; R_5) + V_{\mathrm{OH}}^{(2)}(^2\Sigma; R_6) \\ & + V_{\mathrm{HO_2}}^{(3)}(^2A''; R_1, R_3, R_5) + V_{\mathrm{HO_2}}^{(3)}(^2A''; R_1, R_4, R_6) \\ & + V_{\mathrm{H_2O}}^{(3)}(^1A_1; R_2, R_3, R_6) + V_{\mathrm{H_2O}}^{(3)}(^1A_1; R_2, R_4, R_5), \end{aligned} \tag{65}$$

$$\begin{aligned} V_{22} = \; & V_{\mathrm{O_2}}^{(2)}(^3\Sigma_g^-; R_1) + V_{\mathrm{H_2}}^{(2)}(^3\Sigma_u^+; R_2) \\ & + V_{\mathrm{OH}}^{(2)}(^2\Pi; R_3) + V_{\mathrm{OH}}^{(2)}(^2\Pi; R_4) + V_{\mathrm{OH}}^{(2)}(^2\pi; R_5) + V_{\mathrm{OH}}^{(2)}(^2\pi; R_6) \\ & + V_{\mathrm{HO_2}}^{(3)}(^2A''; R_1, R_3, R_5) + V_{\mathrm{HO_2}}^{(3)}(^2A''; R_1, R_4, R_6) \\ & + V_{\mathrm{H_2O}}^{(3)}(^1A_1; R_1, R_4, R_6) + V_{\mathrm{H_2O}}^{(3)}(^1A_1; R_2, R_4, R_5), \end{aligned} \tag{66}$$

$$\begin{aligned} V_{33} = \; & V_{\mathrm{O_2}}^{(2)}(^1\Delta_g; R_1) + V_{\mathrm{H_2}}^{(2)}(^1\Sigma_g^+; R_2) \\ & + V_{\mathrm{OH}}^{(2)}(^2\Pi; R_3) + V_{\mathrm{OH}}^{(2)}(^2\Pi; R_4) + V_{\mathrm{OH}}^{(2)}(^2\Pi; R_5) + V_{\mathrm{OH}}^{(2)}(^2\Pi; R_6) \\ & + V_{\mathrm{HO_2}}^{(3)}(^2A''; R_1, R_3, R_5) + V_{\mathrm{HO_2}}^{(3)}(^2A''; R_1, R_4, R_6) \\ & + V_{\mathrm{H_2O}}^{(3)}(^3A'; R_1, R_4, R_6) + V_{\mathrm{H_2O}}^{(3)}(^3A'; R_2, R_4, R_5). \end{aligned} \tag{67}$$

The V_{12} term describes the diabatic crossings within the $H_2O(^1A_1)$ molecule is given by

$$V_{12} = V_{\mathrm{H_2O}}^{(3)}(^1\mathrm{A}_1; R_2, R_3, R_6) + V_{\mathrm{H_2O}}^{(3)}(^1\mathrm{A}_1; R_2, R_4, R_5). \tag{68}$$

Each term within the summation for each V_{ij} is possible to be analogous to Subsection (2.4), even though originally in Coelho & Brando [**?**] the authors refer to older methods, particularly: $H_2O(^1A_1)$ by Brando & Rio [**?**] (Subsection 2.6), $H_2O(^3A')$ by Brando et al [**?**] (Subsection 2.6), and $HO_2(^2A'')$ by Brando et al [**?**] (Subsection 2.7).

A total of 21 reference geometries (Table 3 of Coelho & Brando [**?**]) are used by Eq.(31) to indicate the deviation of the coordinates; several of the reference geometries used are 8 stationary points (Table 1 of Coelho & Brando [**?**]); two polynomial centres are used to warrant a good description of the insertion and abstraction entrance channels for $O(^1D) + H_2O$, named as O_insertion and H_abstraction; a reference geometry H_approach is used to describe the approach of a hydrogen atom to HO_2; 10 of the reference points are generated by k-means algorithm to cover all configurational space of the computed points.

2.6 EHF and long range term for $H_2O(^1A_1)$ (Brando & Rio 2003)

2.6.1 Elements of the matrix

Functional form for the double-valued PES for $H_2O(^1A_1)$ was employed by Brando & Rio [**?**]. The dissociation channels present in $H_2O(^1A_1)$ is

$$H_2O(\widetilde{X}^1A') \rightarrow \begin{cases} H_2(\widetilde{X}^1\Sigma_g^+) + O(^1D) \rightarrow 2H(^2S) + O(^1D) \\ H_2(\widetilde{a}^1\Sigma_u^+) + O(^3P) \rightarrow 2H(^2S) + O(^3P) \\ OH(\widetilde{A}^2\Sigma) + H(^2S) \rightarrow 2H(^2S) + O(^1D) \\ OH(\widetilde{X}^2\Pi) + H(^2S) \rightarrow 2H(^2S) + O(^3P). \end{cases} \tag{69}$$

The double-valued functional form is represented as the eigenvalues of a 2×2 matrix, analogous to Subsection 2.4. Each term within the matrix is represented as:

$$V_{11} = V_O^{(1)}(^1D) + V_{OH}^{(2)}(^2\Sigma; R_2) + V_{OH}^{(2)}(^2\Sigma; R_3) + V_{HH}^{(2)}(^1\Sigma_g^+; R_1) + V_{11(LR)}^{(3)}(\mathbf{R}) + V_{11(EHF,nele)}^{(3)}(\mathbf{R}), \tag{70}$$

$$V_{22} = V_{OH}^{(2)}(^2\Pi; R_2) + V_{OH}^{(2)}(^2\Pi; R_3) + V_{HH}^{(2)}(^3\Sigma_u^+; R_1) + V_{22(LR)}^{(3)}(\mathbf{R}) + V_{22(LR,nele)}^{(3)}(\mathbf{R}), \tag{71}$$

$$V_{12} = V_{12(EHF,\ nele)}^{(3)}(\mathbf{R}). \tag{72}$$

Brando et al [**?**] employs the **Extended Hartree-Fock** (EHF) and long range term for $H_2O(^3A')$ analogously, however, instead of using the eigenvalues of a matrix, the functional form is single energy valued:

$$V = V_{HH}^{(2)}(^1\Sigma_g^+, R_1) + V_{OH}^{(2)}(^2\Pi, R_2) + V_{OH}^{(2)}(^2\Pi, R_3) + V_{(LR)}^{(3)}(\mathbf{R}) + V_{(EHF,nele)}^{(3)}(\mathbf{R}), \tag{73}$$

with different coefficients' and constants' values to build the functional (the explicit values are described in Brando et al [**?**]).

2.6.2 Diatomic terms

The diatomic terms $V_k^{(2)}(k = \text{OH, HH})$ is the sum of the dispersion correlation term and the (EHF) term, which is shown as the following:

$$V^{(2)}(R) = V_{dc}^{(2)}(R) + V_{EHF}^{(2)}(R), \tag{74}$$

where the form of the dispersion correlation energy is

$$V_{dc}^{(2)}(R) = -\sum_{n=6,8,10...} C_n \chi_n(R) R^{-n}, \tag{75}$$

with the damping function's formula shown as

$$\chi_n(R) = \left[1 - \exp\left(-\frac{A_n R}{\rho} - \frac{B_n R^2}{\rho^2}\right)\right]^n, \tag{76}$$

with coefficients' values as the following:

$$\begin{aligned} A_n &= 25.9528 n^{-1.1868}, \\ B_n &= 15.7381 \exp(-0.09729 n), \\ \rho &= \frac{1}{2}(R_m + 2.5 R_0), \end{aligned} \tag{77}$$

where R_m is the equilibrium diatomic geometry and R_0 is the **Le Roy parameter** introduced by R. J. Le Roy [**?**] shown as the following:

$$R_0(X-Y) = 2\left(\langle r_X^2\rangle^{\frac{1}{2}} + \langle r_Y^2\rangle^{\frac{1}{2}}\right). \tag{78}$$

The EHF term varies depending on the diatomic form, in the case of $H_2(^3\Sigma_u^+)$, the expression is shown as

$$V_{\text{EHF}}^{(2)}(R) = -DR^m\left(1+\sum_{i=1}^{3} a_i X^i\right)\exp\left[-\sum_{i=1}^{3} e_i X^i\right] + \chi_6(R)V^a(R), \tag{79}$$

with the appearing X defined as

$$X = R - R_m, \tag{80}$$

and V^a denotes the asymptotic exchange energy, which has the following expression:

$$V^a(R) = 0.805R^{2.5}\exp(-2R). \tag{81}$$

The explicit values of the coefficients and constants appearing in the $V_{\text{dc}}^{(2)}(R)$ and $V_{\text{EHF}}^{(2)}(R)$ term for $H_2(^3\Sigma_u^+)$ are listed in the Table II of Brando & Rio [**?**], the aforementioned coefficients and constants are: D, $a_i(i=1,2,3)$, $e_i(i=1,2,3)$, R_0, R_m, $C_i(i=6,8,10,11,12,13,14,15,16)$ and ε. The EHF term of OH for $^2\Pi$ and $^2\Sigma$ states has the following expression:

$$V_{\text{EHF}}^{(2)}(R) = -DR^m\left(1+\sum_{i=1}^{3} a_i X^i\right)\exp\left[-\gamma(R)X\right] + \chi_6(R)V^a(R), \tag{82}$$

with the factor inside the exponential is described as

$$\gamma(R) = \gamma_0\left[1+\gamma_1\tanh(\gamma_2 X)\right], \tag{83}$$

and a different V^a expression, shown as

$$V^a(R) = -\tilde{A}R^{\tilde{\alpha}}(1+\tilde{a}R)\exp(-\tilde{\gamma}R), \tag{84}$$

where the values of the coefficients and constants are described in Table III of Brando & Rio [**?**], the aforementioned coefficients and constants are: D, $a_i(i=1,2,3)$, $\gamma_0(i=0,1,2)$, m, $\tilde{A}$, $\tilde{\alpha}$, $\tilde{a}$, $\tilde{\gamma}$, R_0, R_m, $C_i(i=6,8,10)$, and ε.

2.6.3 Triatomic terms

The functional form to represent $V_{i(\text{EHF, nele})}^{(3)}$ (nele here denotes non-electrostatic energy), where $i=11,22,12$, is shown as the following:

$$V_{i(\text{EHF, nele})}^{(3)}(R_1,R_2,R_3) = P_i^{(3)}(Q_1,Q_2,Q_3)D_i^{(3)}. \tag{85}$$

The appearing $P_i^{(3)}(Q_1,Q_2,Q_3)$, $(i=11,22,12)$ is a polynomial function, it has the following pattern:

$$P_i^{(3)}(Q_1,Q_2,Q_3) = \sum_j^{n_c} c_j Q_1^k Q_2^l Q_3^m S_{2a}^{2p} S_{2b}^{2q} S_3^{3r}, \tag{86}$$

where n_c indicates the number of the c_j coefficients (implies the number of terms within the polynomial function); the $Q_1, Q_3, S_{2a}^2, S_{2b}^2$, and S_3^2 are the integrity bases of the S_2 permutation symmetry; $k, m, p, q, r \in \mathbb{N}^0$, and in this case $l = 0$ (the complete values of n_c, k, m, p, q, r are available in Eqs.(20),(21), and (22) of Brando & Rio [**?**]). The symmetry coordinates ($Q_1, Q_2,$ and Q_3) has the following form:

$$\begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \end{bmatrix} = \begin{bmatrix} \sqrt{1/3} & \sqrt{1/3} & \sqrt{1/3} \\ 0 & \sqrt{1/2} & -\sqrt{1/2} \\ \sqrt{2/3} & -\sqrt{1/6} & -\sqrt{1/6} \end{bmatrix} \begin{bmatrix} R_1 - R_1^{(i)} \\ R_2 - R_2^{(i)} \\ R_3 - R_3^{(i)} \end{bmatrix}, \tag{87}$$

where $R_j^{(i)} (j = 1, 2, 3)$ refers to the reference geometries (the explicit values are available in Table V of Brando & Rio [**?**]); the rest of the integrity bases have the following form:

$$\begin{aligned} S_{2a}^2 &= Q_2^2 + Q_3^2, \\ S_{2b}^2 &= Q_2^2 - Q_3^2, \\ S_3^3 &= Q_3^3 - 3Q_2^2 Q_3. \end{aligned} \tag{88}$$

The considered degrees for the polynomials are: 8 (95 terms) for $i = 11$, 5 (34 terms) for $i = 22$, and 3 (13 terms) for $i = 12$. The complete coefficients' values are available in Table VI, VII, and VIII in Brando & Rio [**?**] for $i = 11, 22$ and 12 respectively. The range factor is expressed as the product of hyperbolic tangents, except for $i = 12$, the complete form is shown as the following:

$$\begin{aligned} D_{11}^{(3)} = &\left\{1 - \tanh\left[c_{96}(R_1 - R_1^{(11)})\right]\right\}\left\{1 - \tanh\left[c_{97}(R_2 - R_2^{(11)})\right]\right\} \\ &\left\{1 - \tanh\left[c_{96}(R_3 - R_3^{(11)})\right]\right\}, \end{aligned} \tag{89}$$

$$\begin{aligned} D_{22}^{(3)} = &\left\{1 - \tanh\left[c_{132}(R_1 - R_1^{(22)})\right]\right\}\left\{1 - \tanh\left[c_{133}(R_2 - R_2^{(22)})\right]\right\} \\ &\left\{1 - \tanh\left[c_{133}(R_3 - R_3^{(22)})\right]\right\}, \end{aligned} \tag{90}$$

$$\begin{aligned} D_{12}^{(3)} = &\sin(\angle\mathrm{HOH})\exp\left[-100(V_{11} - V_{22})^2\right]\exp\left[(-c_{147}(R_1 - R_1^{(12)})^{(2)})\right] \\ &\exp\left[(-c_{148}(R_2 + R_3 - 2R_3^{(12)})^{(2)})\right], \end{aligned} \tag{91}$$

where the value of $\angle$HOH is mentioned in Table IX in in Brando & Rio [**?**], which also describes the minimum C_{2v} for water molecule. The $V_{i(\mathrm{LR})}^{(3)}, (i = 11, 22)$ terms appearing in Eq.(70-71) are the long-range terms (described by Brando & Rio [**?**]), represented as the sum of electrostatic energy, induction energy, and dispersion correlation energy:

$$V_{i(\mathrm{LR})}^{(3)} = V_{(\mathrm{ele})}^{(3)} + V_{(\mathrm{ind})}^{(3)} + V_{(\mathrm{dc})}^{(3)}, \tag{92}$$

with

$$V_{(\mathrm{ele})}^{(3)} = \frac{3}{4}\Theta_{\mathrm{H}_2}(R)\Theta_{\mathrm{O}}\mathcal{A}(\omega)\chi_5(r)r^{-5} \tag{93}$$

as the electrostatic energy (each factor and its coefficients are explained in more detail by Varandas & Brando [**?**] and Varandas & Pais [**?**]); the induction energy is

$$V_{(\mathrm{ind})}^{(3)} = -\sum_{i=2}^{3} \mu_{\mathrm{OH}_i}^2(R_i)\alpha_{\mathrm{H}}(3\cos^2\theta_i + 1)\chi_6(r_i)/(2r_i^6); \tag{94}$$

the dispersion correlation energy is

$$\begin{aligned} V_{(\mathrm{dc})}^{(3)} = & \sum_{i=1}^{3} S(R_i, r_i) \sum_{n=6}^{10} C_n^i(R_i, \theta_i) \chi_n(r_i) r_i^{-n} \\ & + \sum_{i=1}^{3} \left[\prod_{j \neq i} (1 - S(R_j, r_j)^2) \right] \sum_{n=6}^{10} C_n^i \chi_n(R_i) R_i^{-n}, \end{aligned} \tag{95}$$

with

$$S(R, r) = \frac{1}{2} \left\{ 1 + \tanh \left[\gamma_1 \left(\frac{r}{R} - \gamma_2 \right) \right] \right\}, \tag{96}$$

where R is diatom separation, r is atom-diatom distance, and the index i specifies the atom-diatom combination; γ_1 and γ_2 are calibration parameters, where in this case $\gamma_1 = \gamma_2 = 2$.

2.7 V_R potentials for HO_2 (Brando et al 2009)

Brando et al [?] employed a functional form for HO_2, which has the following expression:

$$V_{R_{\mathrm{spect}}} = T \exp(-D), \tag{97}$$

where T is a third degree polynomial function with 13 adjustable coefficients (the values are obtainable from Table VII in Brando et al [?]), shown as the following:

$$\begin{aligned} T = \; & c_1 + c_2 R_{1d} + c_3 S_{1d} + c_4 R_{1d}^2 + c_5 S_{1d}^2 + c_6 R_{1d} S_{1d} + c_7 S_{2d} + c_6 R_{1d}^3 + c_7 S_{2d} + c_8 R_{1d}^3 \\ & + c_9 R_{1d} S_{1d}^2 + c_{10} R_{1d}^2 S_{1d} + c_{11} R_{1d} S_{2d} + c_{12} S_{1d}^3 + c_{13} S_{1d} S_{2d}, \end{aligned} \tag{98}$$

and the appearing D in Eq.(97) is a decay term with fixed $c_{i_{\mathrm{fix}}} (i = 1, 2, 3, 4)$ (Brando et al [?] set all $c_{i_{\mathrm{fix}}}$ values to 2), shown as the following:

$$D = c_{1_{\mathrm{fix}}} R_{1d}^2 + c_{2_{\mathrm{fix}}} S_{1d}^2 + c_{3_{\mathrm{fix}}} R_{1d} S_{1d} + c_{4_{\mathrm{fix}}} S_{2d}^2, \tag{99}$$

R_2 and R_3 denotes the O^i–H $(i = 1, 2)$ interatomic distances; the R_{1d}, S_{1d}, and S_{2d} are the displacements coordinates, shown as the following:

$$\begin{aligned} R_{1d} &= R_1 - R_{1_{\mathrm{eq}}}, \\ S_{1d} &= \frac{1}{\sqrt{2}} \left[(R_2 + R_3) - (R_{2_{\mathrm{eq}}} + R_{3_{\mathrm{eq}}}) \right], \\ S_{2d} &= \frac{1}{2} \left[(R_2 - R_3)^2 - (R_{2_{\mathrm{eq}}} - R_{3_{\mathrm{eq}}})^2 \right], \end{aligned} \tag{100}$$

where the values of the experimental equilibrium geometry $(R_{i_{\mathrm{eq}}} (i = 1, 2, 3))$ are set as the following:

$$\begin{aligned} R_{1_{\mathrm{eq}}} &= 2.5143 \mathrm{a}_0, \\ R_{2_{\mathrm{eq}}} &= 1.8346 \mathrm{a}_0, \\ R_{3_{\mathrm{eq}}} &= 3.4592 \mathrm{a}_0. \end{aligned} \tag{101}$$

2.8 MBE/DC PES for H_2O (Galvo & Rodrigues 2008)

Employed by Galvo & Rodrigues [?]. The functional form is

$$G^{(3)}(\mathbf{R}) = \sum_{i=1}^{37} P_i(\mathbf{R}) \exp\left[-\sum_{j=1}^{3} b_{ij}(R_j - R_{ij}^0)^2\right], \tag{102}$$

where the polynomials (three interatomic coordinates) are written in the terms of D_{3h} symmetry coordinates, which is shown as

$$\begin{pmatrix} Q_1 \\ Q_2 \\ Q_3 \end{pmatrix} = \begin{pmatrix} \sqrt{1/3} & \sqrt{1/3} & \sqrt{1/3} \\ 0 & \sqrt{1/2} & -\sqrt{1/2} \\ \sqrt{2/3} & -\sqrt{1/6} & -\sqrt{1/6} \end{pmatrix} \begin{pmatrix} R_3 \\ R_2 \\ R_1 \end{pmatrix}, \tag{103}$$

where R_1 and R_2 are OH bond distances while R_3 is HH. Hence, the general form of $P_i(\mathbf{r})$ is

$$P_i(\mathbf{r}) = c_{i0} + c_{i1}Q_2^2 + c_{i2}Q_2^4 + c_{i3}Q_2^6 + c_{i4}Q_1 + c_{i5}Q_1^2 + c_{i6}Q_3 + c_{i7}Q_3^2. \tag{104}$$

To warrant proper symmetry in a permutation of R_1 and R_2, no odd powers of variable Q_2 are used to construct the polynomials. In total, the function has 212 adjustable parameters, including the polynomial and Gaussian coefficients.

2.9 Neural network for H_3O (Chen et al 2013)

A PES model in the form of feed-forward **neural network** (NN) with two hidden layers is employed by Chen et al [?]. Formally, each layer is

$$y_{k,j} = f_k\left(b_{k,j} + \sum_{i=1}^{N_{k-1}} (w_{k,j,i} \times y_{k-1,i})\right), \quad j = 1, 2, ..., N_k, \tag{105}$$

where N_k indicates the number of neurons in the kth layer; $y_{k,j}$ is the output on the kth layer and jth node, in the case of $k = 0$, the $y_{0,i}(i = 1, ..., N_0)$ is the input value, which is one value from the set of the molecular distances; $b_{k,j} \in \mathbb{R}$ is the bias on the kth layer and jth node; $w_{k,j,i} \in \mathbb{R}$ is the weight on the kth layer and jth node which connects the ith neuron of the $(k-1)$th layer; f_k is the activation function of the kth layer, in this case tanh (the hyperbolic tangent) is chosen. Consequently, the final output neuron of the output layer $y_{\bar{k}}$ follows analogously to Eq.(105), however it is not necessary to apply the activation function f_k, it is formally given as

$$y_{\bar{k}} = b_{\bar{k}} + \sum_{i=1}^{N_{\bar{k}-1}} (w_{\bar{k},j,i} \times y_{\bar{k}-1,i}). \tag{106}$$

In this case of H_3O system, the permuted set of distances $\{r_{OH_1}, r_{OH_2}, r_{OH_3}, r_{H_1H_2}, r_{H_1H_3}, r_{H_2H_3}\}$, satisfying $r_{OH_1} \leq r_{OH_2} \leq r_{OH_3}$, is fed into the input layer.

3 Fitting the potentials

3.1 CHIPR fitting

In CHIPR, generally the fitting is done by utilizing the weighted least-square method. The objective function for diatomic energy to be minimized is

$$\chi^2 = \sum_{h=1}^{N} \left\{ W_h \left[V^{(2)}(R_h; C_k, c_\alpha, R^{\text{ref}}, \zeta) - E(R_h) \right] \right\}^2, \tag{107}$$

where N is the total *ab initio* points, W_h is the least-squares weight of the hth calculated point at R_h, and $E(R_h)$ is the corresponding *ab initio* energy.

The three-body (triatomic molecules) energy for fitting, $\epsilon^{(3)}$, is determined by

$$\epsilon^{(3)}(\mathbf{r}) = E(\mathbf{r}) - \mathcal{V}_S^{(2)}(\mathbf{r}) - \mathcal{V}_S^{(1)}. \tag{108}$$

In order to obtain the optimum basis set parameters y_p, first, the objective function to be minimized is

$$\chi_p^2 = \sum_{h=1}^{N_{\text{basis}}} \left\{ W_h \left[y_p(R_p; c_\alpha, R_p^{\text{ref}}, \zeta) - \epsilon_h^{(3)} \right] \right\}, \tag{109}$$

where $p = \{1\}, \{1,2\}, \{1,2,3\}$ for A_3, AB_2, and ABC molecular species respectively. After obtaining the optimum basis set parameters, the next objective function to be minimized is

$$\chi^2 = \sum_{h=1}^{N_{\text{pol}}} \left\{ W_h \left[V^{(3)}(\mathbf{R}; C_I, c_\alpha, R_p^{\text{ref}}, \zeta) \times \mathcal{D}^{(3)} - \epsilon_h^{(3)} \right] \right\}^2, \tag{110}$$

where $I = \{i, j, k\}$ set of indices.

Analogous to the triatomic energy, the tetratomic energy is obtained from the subtraction of the total energy with the $(n < 4)$-body terms, its form is

$$\begin{aligned} \epsilon^{(4)}(\mathbf{r}) &= E(\mathbf{r}) - \mathcal{V}_S^{(2+3)}(\mathbf{r}) \\ &= E(\mathbf{r}) - \left[\sum_{i<j} V_{ij}^{(2)}(r_{ij}) + \sum_{i<j<k} V_{ijk}^{(3)}(r_{ik}, r_{jk}, r_{ij}) \right]. \end{aligned} \tag{111}$$

The objective function for calibrating y_p follows from Eq.(109), with $I = \{i, j, k, l, m, n\}$; $\epsilon^{(4)}(\mathbf{R})$ replaces $\epsilon^{(3)}(\mathbf{R})$ and $p = \{1\}$, $\{1,4\}$, $\{1,3,6\}$, $\{1,2,3,6\}$, $\{1,2,3,4,5,6\}$ for A_4, AB_3, A_2B_2, ABC_2, and ABCD molecular species respectively; as shown in Eq.(112) and (113) which are analogous to Eq.(109) and (110) respectively.

$$\chi_p^2 = \sum_{h=1}^{N_{\text{basis}}} \left\{ W_h \left[y_p(R_p; c_\alpha, R_p^{\text{ref}}, \zeta) - \epsilon_h^{(4)} \right] \right\}, \tag{112}$$

$$\chi^2 = \sum_{h=1}^{N_{\text{pol}}} \left\{ W_h \left[V^{(4)}(\mathbf{R}; C_I, c_\alpha, R_p^{\text{ref}}, \zeta) \times \mathcal{D}^{(4)} - \epsilon_h^{(4)} \right] \right\}^2. \tag{113}$$

3.2 Neural Network fitting

The weights and biases of NN employed by CHEN et al [?] are optimized during training by using the **Levenberg-Marquardt algorithm** [?] with the RMSE as the objective function. Additional mechanism is applied due to the non-symmetric behaviour of the NN with respect to the permutation of H atoms, the PES is not continuous at a configuration with two or three equal OH distances. The aforementioned mechanism is an exchange scheme as the following: during the training stage, if the difference between two OH distances for a data point in the training set is less than 0.05 bohr, e.g., $|r_{\mathrm{OH}_1} - r_{\mathrm{OH}_2}| < 0.05$, then the training set is augmented by both $\{r_{\mathrm{OH}_1}, r_{\mathrm{OH}_2}, r_{\mathrm{OH}_3}\}$ and the permuted set $\{r_{\mathrm{OH}_2}, r_{\mathrm{OH}_1}, r_{\mathrm{OH}_3}\}$; in this way the fitting space is extended. During the evaluation, if the difference of two OH bond is close enough, e.g., $|r_{\mathrm{OH}_1} - r_{\mathrm{OH}_2}| < 0.1$, then the final energy is calculated as

$$V = w \times y_{\bar{k}} + (1 - w) \times y_{\bar{k}}, \tag{114}$$

where $y_{\bar{k}}$ is the output of the NN described in Eq.(106); w is a switch function given as

$$w = (1 + \sin(5\pi(r_{\mathrm{OH}_1} - r_{\mathrm{OH}_2})))/2. \tag{115}$$

4 Pair potentials for $\mathbf{H}_x\mathbf{O}_y$

4.1 Lennard-Jones-like pair potential (Deiters & Neumaier 2016)

A modification of the **Lennard-Jones** (LJ) potential [?] is proposed by DEITERS & NEUMAIER [?] for the PES of Argon. The form of the well-known LJ potential is

$$V(R) = 4\epsilon \left[\left(\frac{\sigma}{R}\right)^{12} - \left(\frac{\sigma}{R}\right)^{6} \right], \tag{116}$$

where R is the interatomic distance, σ is the collision diameter, and ϵ is the depth of the "well" in the potential energy curve. It has been noted by DEITERS & NEUMAIER [?] that the short distance repulsion of LJ potential is less realistic when compared to the experimental results, hence the new form of the pair potential is proposed. The earliest attempt on devising a more realistic pair potential was in the form of BUCKINGHAM [?] potential

$$V(R) = c_1 e^{-\alpha R} - \frac{c_2}{R^6}, \tag{117}$$

the drawback of the Buckingham potential is that it is infinitely attractive, i.e., it is infinitely attractive for $R \to 0$. PATHAK & THAKKAR [?] state that for monoatomic gases, the united-atom perturbation theory implies

$$V(R) = c_1 \frac{Z^2}{R} + c_2 + O(R^2) \quad \text{for } R \ll R_e, \tag{118}$$

where Z is the nuclear charge, R_e is the equilibrium distance, $c_1 = 1$ (in Hartree unit), and c_2 depends on the substance. With exponential screening of the Coulomb potential, the functional form becomes

$$V(R) = e^{-\alpha R} \left(c_1 \frac{Z^2}{R} \left(1 - \frac{(\alpha R)^2}{2} \right) + c_2(1 + \alpha R) \right) \quad \text{for } R \ll R_e. \tag{119}$$

This implies that pair potentials fitted to data at larger R should also include the correct short distance behaviour such as in Eq.(119). This can be implemented by matching c_2 and α to the value $V_s = V(R_s)$ and the slope $V'_s = V'(R_s)$ of the potential at small switching radius R_s. It can be verified that

$$\begin{aligned} &\alpha^2 R V(R) + (1+\alpha R)V'(R) \\ &= e^{-\alpha R} c_1 \frac{Z^2}{2R^2}(1+\alpha R)(2+2\alpha R+3(\alpha R)^2-(\alpha R)^3-(\alpha R)^4) \end{aligned} \tag{120}$$

when evaluated at R_s gives a nonlinear equation to be solved for α, which then c_2 can be found from the equation for $V(R)$. By varying R_s, one can obtain different variants of the same potential, e.g., to assess the effect of the repulsive part.

The newly proposed form by Deiters & Neumaier [**?**] is used to correct the R^{-6} behavior at large distances, the desired short-range behavior, and to avoid unrealistic (infinitely attractive) pole at $R = 0$. The form of the modified potential is

$$V(R) = c_1 e^{-2\alpha R}\left(\alpha + \frac{1}{R}\right) - \frac{c_2 R^2}{c_3 + R^8} + c_4, \tag{121}$$

where α, $c_1 ... c_3$ are chosen non-negative parameters, and c_4 is the energy shift parameter; the explicit parameters' values are shown in Table 1 of Deiters & Neumaier [**?**].

4.2 Pair potential for noble gases (Deiters & Sadus 2019)

Deiters & Sadus [**?**] proposed a simplified *ab initio* atomic potential (SAAP) for the noble gases (He, Ne, Ar, Kr, Xe). The functional form is

$$V(R) = \frac{\left(\frac{c_0}{R}\right) e^{c_1 R + c_6 R^2} + c_2 e^{c_3 R} + c_4}{1 + c_5 R^6} + c_7, \tag{122}$$

where $c_1 ... c_4 < 0$, $c_0, c_5 > 0$, $c_6 = 0$ except for He where $a_6 < 0$ to represent the repulsive behaviour in extremely small interatomic distances, and c_7 is the energy shift parameter. The complete parameters are summarized in Table 1 of Deiters & Sadus [**?**]. The form is proposed in order to address the correct potential curve with minimal computational cost, the considerations of the functional form are:

- The potential curve should pass the abscissa once at σ, which flips the sign from positive to negative of $V(R)$ for $R > \sigma$.
- The domination of dispersion interaction at large R, which is represented by R^{-6}, with V monotonously converge to zero.
- The **Pauli repulsion** at short distances should be represented, in particular, the $\lim_{R\to 0} V(R) = \infty$, in this case it is described by $\frac{e^R}{R}$ term.

4.3 Proposed diatomic potential

4.3.1 Polynomial fit

The proposed function for diatomic potential takes the concept of monomial expansion from Eq.(22), the specific form for diatomic is

$$V = \sum_{k=0}^{M} C_k y^k, \tag{123}$$

where M is the maximum polynomial degree; C_k is the coefficient which corresponds to the kth power which needs to be fitted; in this case y is the Morse potential described in Eq.(21). For example, for $M = 3$, the explicit form is

$$V = C_0 + C_1 y + C_2 y^2 + C_3 y^3, \tag{124}$$

which results in four coefficients to be fitted. An additional note, any polynomials are evaluated by using the **Horner's scheme** [**?**] to maximize the computation's efficiency.

4.3.2 RATPOT1

A proposed functional form which contains Coulomb-like term, repulsion term, attraction term, and energy shift constant; generalized with M-degrees of polynomial, which in total $3M + 1$ parameters are adjustable; the functional form is

$$V = c_0 + \frac{P(R)}{S(R)}, \tag{125}$$

where

$$P(R) = Z_{ij}\left(\frac{1}{R} + c_1 R\right) + c_2 + c_3 R^2 + ... + c_{2M-1} R^{2M-2}, \tag{126}$$

$$Q(R) = 1 + (c_{2M} + c_{2M+1} R + ... + c_{3M} R^M) R, \tag{127}$$

$$S(R) = 1 + c_1 (RQ(R))^2, \tag{128}$$

and

$$c_1 > 0, \tag{129}$$

where

$$Z_{ij} := Z_i Z_j, \tag{130}$$

Z_k is the kth atom's nuclear charge, e.g., $Z_{\mathrm{OH}} = Z_{\mathrm{O}} Z_{\mathrm{H}} = 1 \times 8 = 8$. For efficient computation, the polynomials are evaluated using Horner's scheme analogous to Eq.(123) and the rest of the operations are executed in certain order; Eq.(125 - 129) are evaluated algorithmically as the following:

Function: Diatomic General M
Input: C, R, Z, M
Output: V
if $C_1 < 0$:
  $C_1 := -C_1$

end if
$R_2 := R^2$
$K := [C_3...C_{2M-1}]$
$y := \text{horner}(R, K)$
$a := Z(1/R + C_1 R) + C_2 + y$
$K := [C_{2M}...C_{3M}]$
$y := \text{horner}(R, K)$
$b := 1 + yR$
$c := 1 + C_1(Rb)^2$
$V := C_0 + a/c,$

where the Horner's scheme is

Function: horner
Input: x, C
Output: y
$n :=$ length of C
$y := C_n$
for $i = n-1, n-2...0$, do:
$y := yx + C_i$
end for.

4.3.3 RATPOT2

This version of RATPOT has a total of $4M + 7$ tuning parameters, which are the vectors a, b, c, and d. RATPOT2 does not have the nuclear charge as a dependency, it is replaced by one tuning parameter, this should allow more degree of freedom in terms of the data fitting. The functional form is

$$\begin{aligned} V_m(R) &= c_0 + P/Q, \\ P &= c_1 \prod_{i=1}^{m}((R-a_i)^2 + b_i R), \\ Q &= R(R+d_1)\prod_{i=2}^{m+3}((R-c_i)^2 + d_i R), \end{aligned} \tag{131}$$

where $m > 0$ and $b_i \geq 0$ for $i > 1$ and $d_i \geq 0$ for $i > 0$.

4.3.4 RATPOT3

Similar to RATPOT2, RATPOT3 has a, b, c, and d as the tuning parameters' vectors, with a total of $4M + 8$ tuning parameters. The nuclear charge coefficient is included back. The functional form is

$$\begin{aligned} P &= Z(1/R - 1/R_0)\prod_{i=1}^{m}((1-R/a_i)^2 + R/b_i), \\ Q &= \prod_{i=1}^{m+3}((1-R/c_i)^2 + R/d_i), \end{aligned} \tag{132}$$

where $m \geq 0$ and all $b_i, d_i \geq 0$.

4.3.5 Linear RATPOT

An ansatz in order to solve pair potentials linearly was devised. The main idea is that the pair potentials can be used as a basis function for potentials of molecules containing more atoms. Hence this requires a cheap-to-compute ansatz, or even only needed to be computed once in the beginning (pre-computation) without requiring the parameters to be tuned, which reduces the complexity of the optimization problem.

In the first step, the diatomic distances need to be scaled

$$\rho_{ij} := R_{ij}/R_{xy}, \tag{133}$$

where R_{ij} is the diatomic distance between atom i of type x and atom j of type y, R_{xy} is the **equilibrium distance** of the pair potential between atom type x and y. Atomic equilibrium distance is the diatomic distance which corresponds to the lowest pair potential energy, i.e., when the atoms neither repel nor attract each other. For example, the equilibrium distance of H_2 molecule is $R_{\mathrm{HH}} \approx 1.4172946$ Bohr.

Then the equilibrium distance needs to be shifted to the center, in this case, 0 is chosen as the center, and the $[\min(\rho), \max(\rho)]$ are shifted to $[-1, 1]$, this is done by

$$q := \frac{1-\rho}{1+\rho} \in [-1, 1]. \tag{134}$$

Finally, the pair potential form is

$$V_k(R) = \frac{u(q)}{\rho + \rho^k}, \tag{135}$$

where $k = 1 : 6$ is a hyperparameter which denotes the asymptotic behavior of the transformed distance, where $k = 6$ is the theoretically expected best power since it complies with R^{-6}. The numerator $u(q)$ is the linear combination of Chebyshev polynomials of degree d (which means it has $d + 1$ parameters)

$$u(q) = \sum p_l(q)\theta_l. \tag{136}$$

Although it is possible to fit this pair potential ansatz directly with sufficiently high accuracy (low RMSE) and reasonable speed, the main idea is to solve this linearly, in order to do this, (136) needs to be transformed into a system of linear equations. With a total number of n data, and d degrees of polynomials, the system of the linear equation is

$$P\theta = V, \tag{137}$$

where $P \in \mathbb{R}^{n \times d+1}$ is a matrix containing the Chebyshev polynomials entries $p_l(q) \in \mathbb{R}^n$ for each column, where the first column is a $\vec{1}$ column vector since $p_0(q) = 1$, which then each entry is scaled (elementwise division) by

$$P := P/(\rho + \rho^k) \in \mathbb{R}^{n \times d+1}; \tag{138}$$

$\theta \in \mathbb{R}^{d+1}$ is the vector containing the tuning parameters; and $V \in \mathbb{R}^n$ is a vector containing the potential energy data points. In this manner, θ can be obtained by solving

(137) linearly. For example in programming language such as Julia and MATLAB, this can be done simply by

Possible ansatz to scale the RATPOT for solving it linearly (other than (135)) are

$$u(q) \approx (\rho + \rho^k) V_k(R), \tag{139}$$

and

$$\frac{u(q)}{1+\rho^{k-1}} \approx \rho V_k(R). \tag{140}$$

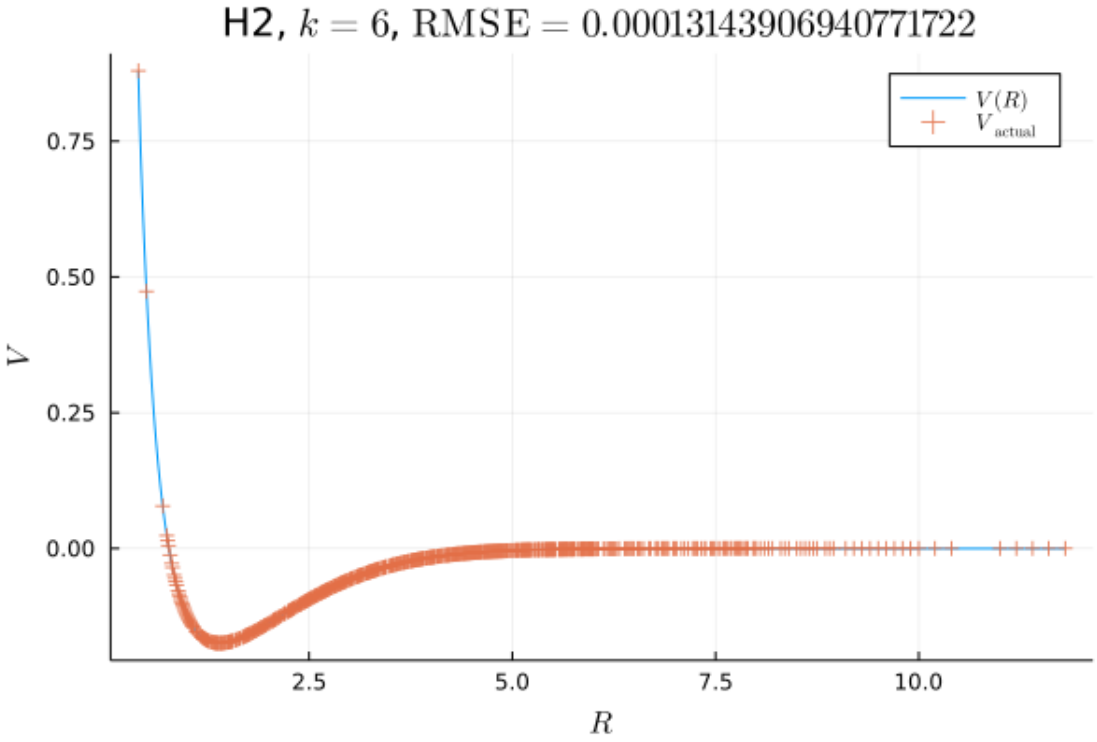

Figure 1: Nonlinear leastsquares fitting applied for linear RATPOT.

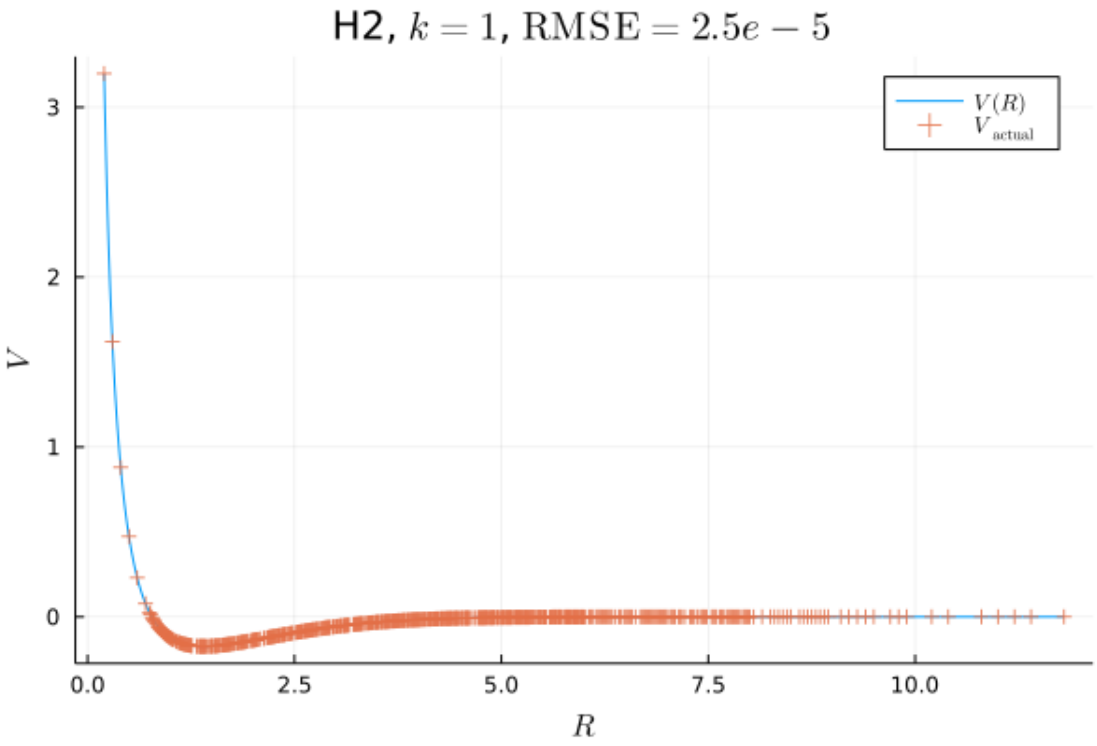

Figure 2: Directly solving linear RATPOT.

4.4 Fitting method

There exist several possible objective function forms which can be used to fit the diatomic functional forms. One possible objective function form is the RMSE in Eq.(15); another one would be the sum of squared residuals

$$\sum_{i=1}^{n} \left[V(R_i) - V_{i,\text{ab initio}}\right]^2 . \tag{141}$$

There are two types of optimization routines for nonlinear least squares form, one requires the objective function to return a residual vector, another one is the standard optimization routine which requires the objective function to return a scalar. In the case of overdetermined system (more data than tuning parameters), **Levenberg-Marquardt** (LM) algorithm [**?**] is employable; otherwise, local search such as **trust region reflective** (TRF) [**?**] and **dogleg** algorithm with rectangular trust regions [**?**] are employable. This rule is molecule-size-agnostic, i.e., applicable to any molecule size.

4.4.1 Fitting dissociated energy

For any functional form of potential energy, it is possible to scale the curve in a way that the repulsive and attractive part of the potential curve is more visible. This is done by scaling both the functional form and the data by the dissociation energy

$$\Delta = V_l - V_{\text{min}}, \tag{142}$$

where V_l is the potential energy at the largest distance and V_{min} is the minimum observed potential energy. This is done during the fitting (also applies to direct solvers):

$$\theta_{\text{min}} = f(V_{\text{ab initio}}/|V_{\text{ab initio}}| + \Delta, V(R)/|V(R)| + \Delta), \tag{143}$$

where θ_{min} is a vector containing optimized tuning parameters, here f is an objective function for data-fitting, for example (141); and $V(R)$ is the functional form. Similarly, an adjusted RMSE for dissociated energy is

$$\text{aRMSE} := \Delta\text{RMSE}(\delta), \tag{144}$$

where

$$\delta := \epsilon/(|V| + \Delta), \tag{145}$$

and

$$\epsilon = V_{\text{ab initio}} - V(R). \tag{146}$$

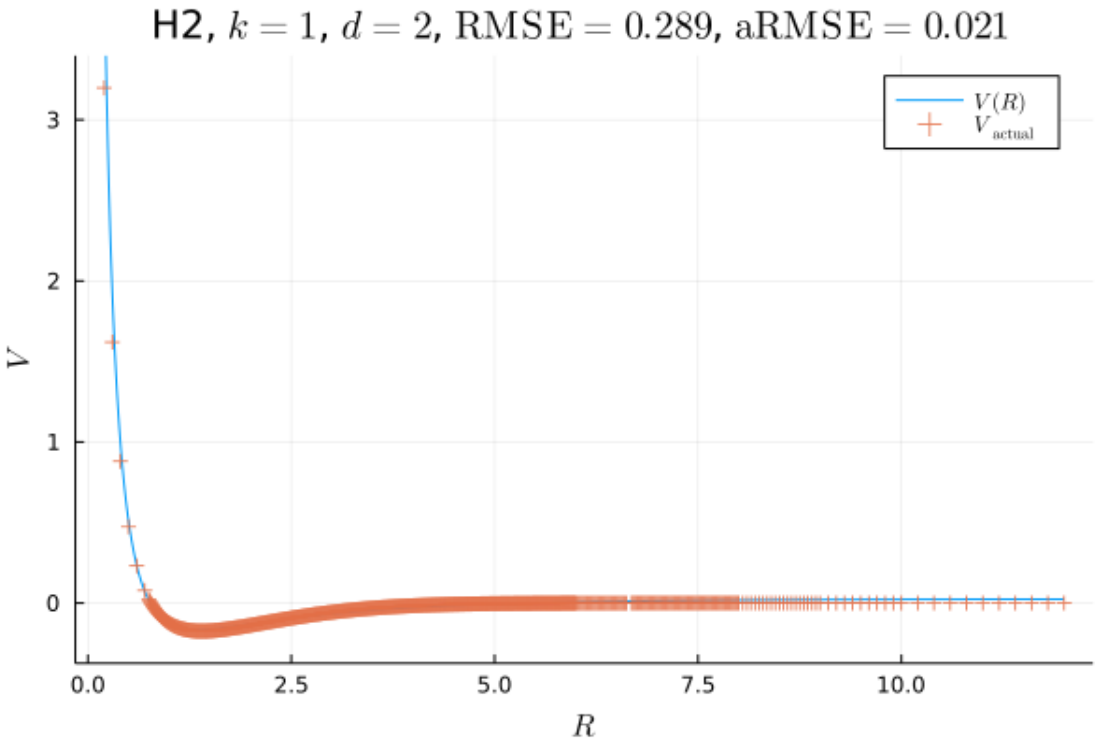

Figure 3: Direct linear solve for linear dissociated RATPOT with 5 initial data points and 5 end data points as validation data.

5 High-level summary of potentials with bonding features

Several bonding features and models are applied for H_3 molecule. Each feature is a basis function, each model is composed of the combinations of multiplications and summations of several basis functions with different polynomial degrees and wrapped in several forms which are then multiplied by tuning coefficients.

5.1 Features

5.1.1 Bonding function

The bonding function b is formed by the **Chebyshev polynomials** which defines the polynomial degree of the **bond strength**. The bond strength is a piecewise continuous smooth function which transforms each diatomic distance depending on a lower bound cut off and an upper bound cut off parameters, if the distance is below the lower bound cut off parameter then the bond strength is constant, if the distance is in between the lower and upper bound cut off parameters then the bond strength is equal to some smooth function, otherwise the bond strength is 0. Bonding function b is relevant for almost all of the features used, the only exception is the **reference energy** feature.

Given a total number of atoms n_{atom} (which implies degree of freedom $f_{\text{free}} := 0.5(n_{\text{atom}}^2 - n_{\text{atom}})$), polynomial degree d_{deg}, and total number of data points n_{data}, the bonding function is collected in an array with $(d_{\text{deg}}, f_{\text{free}}, n_{\text{data}})$ shape.

5.1.2 Reference energy

The **reference energy** U is a diatomic energy function which sums each energy of the possible combinations of the diatomic distances. For each atom i, it is the sum of energies of all of the possible distances which includes atom i, this depends on the degree of freedom of the molecule. For example, for triatomic molecules, the reference energy of each atom is composed of 2 terms of diatomic energy function to be summed; as a concrete example, for $i = 1, 2, 3$, then $U := (U[1] := V(R_{12}) + V(R_{13}), U[2] := V(R_{21}) + V(R_{23}), U[3] := V(R_{31}) + V(R_{32}))$. Since diatomic distance is permutationally invariant, then $R_{ij} = R_{ji}$, which means the indices can be re-arranged into $\{[(1,2),(1,3)],[(1,2),(2,3)],[(1,3),(2,3)]\}$, this type of indexing would be more efficient given ordered vector of distances $\mathbf{R}$, since it is possible to map each tuple of distance index into a scalar index, which needs only to be computed once at the beginning. Since for each atom i, $U[i]$ sums the diatomic energies which depend on the degree-of-freedom indices, then given n_{atom} and n_{data}, U is an array with $(n_{\text{atom}}, n_{\text{data}})$ shape. This indexing rule is, in fact, applicable to any molecule with $n_{\text{atom}} > 2$.

The reference energy has various diatomic terms depending on the model (the models will be explained in later sections). For model with trainable reference pair potential, for each atom, the diatomic energy function is similar to the bond strength function in Section 5.1.1, it is a non-smooth non-continuous function depending on a lower bound cut off parameter and upper bound cut off parameter. For model with fixed H_2 contribution, it uses a smooth and continuous rational polynomial function where the denominator has at least degree 6.

5.1.3 Coordination vector

For atom i, the **coordination vector** $Y[i] = (Y_1[i], Y_2[i], ..., Y_{d_{\text{deg}}}[i])$, is the sum of bonding function for atom i where the summation indices are equal to the summation indices of the reference energy in Section 5.1.2. All components of $Y[i]$ are invariant under translation, rotation, and reflection due to each being the sum of the degree of freedom of each atom. When considering all atoms, Y is an array with $(n_{\text{atom}}, d_{\text{deg}}, n_{\text{data}})$ shape.

5.1.4 Orientation vector

The **orientation vector** $r[i] = (r_1[i], r_2[i], ..., r_{d_{\text{deg}}}[i])$ is the sum of the multiplication between bonding function and difference of two atomic coordinates; similar to the previous section, the summation uses the indexing rule in Section 5.1.2. However this does not imply complete invariant (translation, rotation, and reflection), due to the difference of two coordinates only represents translation invariant property; e.g., let x_k be the atomic coordinate of atom k in three-dimensional Cartesian coordinate system, then $x_i - x_j \neq x_j - x_i$, which breaks the rotation and reflection invariant properties. Given n_{data}, d_{deg} and f_{free} of the bonding function, the dimension of the coordinate from x_k for all k (which in molecular system case it is always 3), and n_{atom} obtainable from the indexing routine, then r is an array with $(n_{\text{atom}}, d_{\text{deg}}, n_{\text{data}}, 3)$ shape.

Additionally, since the available *ab initio* data points throughout the literatures usually only provides the tuples of vector of distances and energy $[(\mathbf{R}_1, V_1), ..., (\mathbf{R}_{n_{\text{data}}}, V_{n_{\text{data}}})]$, then it would be necessary to transform the distance vectors into coordinate matrices whenever the coordinate matrices are not available from the data sources.

5.1.4.1 Converting distances to coordinates

Algorithm. To calculate the Gram matrix G with entries $G_{ik} := x_i \cdot x_k$ in a coordinate system where the center of the mass is 0, proceed as follows:

$$q_{ik} := r_{ik}^2, \tag{147}$$

$$\gamma_i := \frac{1}{N} \sum_k q_{ik}, \tag{148}$$

$$\gamma := \frac{1}{2N} \sum_i \gamma_i, \tag{149}$$

$$G_{ii} = \gamma_i - \gamma, \tag{150}$$

$$G_{ik} = \frac{1}{2}(G_{ii} + G_{kk} - q_{ik}), \tag{151}$$

with N defines the number of atom points.

Given G, calculate its eigenvalues and eigenvectors. Since G is a positive semi definite matrix, all eigenvalues are non-negative, any near 0 negative eigenvalue should be replaced by 0. Finally, multiply the corresponding 3 eigenvectors by the square root of the corresponding eigenvalues. This defines the 3 components of the coordinate vectors x_i.

Proof. The definition in (147) is required to be expanded in the terms of coordinates, it is written as

$$r_{ik}^2 = (x_i - x_k) \cdot (x_i - x_k).$$

Since the center of the mass is 0, we have

$$\sum_k x_k = 0.$$

Here and later in this section, summations run over all atom points,

$$k = 1, .., N.$$

Using this, we find

$$\begin{aligned}
\gamma_i &= \frac{1}{N} \sum_k q_{ik} = \frac{1}{N} \sum_k r_{ik}^2, \\
&= \frac{1}{N} \sum_k (x_i - x_k) \cdot (x_i - x_k), \\
&= \frac{1}{N} \left(\sum_k x_i^2 - \sum_k 2x_i \cdot x_k + \sum_k x_k^2 \right), \\
&= \frac{1}{N} \left(N(x_i^2) - 2x_i \cdot \sum_k x_k + \sum_k x_k^2 \right), \\
&= x_i^2 + \frac{1}{N} \sum_k x_k^2, \\
\gamma &= \frac{1}{2N} \sum_i \gamma_i, \\
&= \frac{1}{2N} \sum_i x_i^2 + \frac{1}{2N^2} N \sum_k x_k^2, \\
&= \frac{1}{N} \sum_k x_k^2.
\end{aligned}$$

Therefore

$$G_{ii} = \gamma_i - \gamma = x_i^2 + \frac{1}{N} \sum_k x_k^2 - \frac{1}{N} \sum_k x_k^2 = x_i^2$$

validates (150). Using this we find

$$G_{ik} = \frac{1}{2}(x_i^2 + x_k^2 - (x_i - x_k) \cdot (x_i - x_k)) = x_i \cdot x_k$$

which validates (151).

5.1.5 Gram matrix

For each atom i, **Gram matrix** $G[i]$ is a symmetric and positive semidefinite matrix formed by the dot product of the orientation vectors. As a feature, it is invariant to translation, rotation, and reflection. Given n_{atom}, n_{data}, and orientation vectors with maximum polynomial degree of d_{deg}, then G is collected in an array with 4 indices with $(n_{\text{atom}}, d_{\text{deg}}, d_{\text{deg}}, n_{\text{data}})$ shape.

5.1.6 Neighbourhood matrix

In the case of triatomic systems, the neighbourhood matrix $\Theta[i]$ is the Gram matrix $G[i]$ subtracted by the two-body terms, hence the neighbourhood matrix only represents the three-body terms. Programmatically, the shape of Θ is equal to the shape of G.

5.2 Models

There are two types of model employed for H_3 molecule, one model makes use of trainable reference pair potential, and the other uses fixed H_2 contribution. All of the models employ the previously mentioned features (basis functions) as monomials. Each feature is collected in $\Phi = \Phi(U, Y, H)$, where H is chosen either G or Θ for each model. For example, for the vector of basis functions up to degree 1: $\Phi(1) = (U, Y_1)$, for basis functions up to degree 2: $\Phi(2) = (UY_1, Y_1^2, Y_2, G_{11})$; then these vectors of basis functions of different degrees of polynomials are arranged in a contiguous way within a single vector $\Phi = (U, Y_1, UY_1, Y_1^2, Y_2, G_{11}, ...)$. For H_3 molecule, there exists a total of 59 basis functions for all of the combinations of monomials up to degree 5. Programmatically, given n_{atom}, n_{data}, and total number of basis functions n_{basis}, the shape of array Φ is $(n_{\text{atom}}, n_{\text{basis}}, n_{\text{data}})$.

5.2.1 Model with trainable reference pair potential

For the U monomial, this model uses a non-continuous non-smooth piecewise function with lower bound and upper bound cut off tuning parameters similar to the bonding function. There are a total of 7 tuning parameters from the bonding function and the U monomial combined, collected in a vector $\rho := (\rho_1, \rho_2, ...\rho_7)$, where $\rho_2 \leq \rho_3 \leq ... \leq \rho_7$ relations need to be satisfied.

The potential energy of the molecular system is the sum of all atomic energies, each atomic energy is composed of 3 different partial energy terms, and each partial energy term is in the form of rational potential where each of the numerator and denominator has n_{basis} tuning coefficients; in total there are $6 \times n_{\text{basis}}$ tuning coefficients for this model. For H_3 molecule, there is a maximum of 59 basis functions (where $H = G$, up to 5 degree of monomial combinations), this results in $6 \times 59 + 7 = 361$ tuning parameters.

5.2.2 Model with fixed H_2

In this model, a smooth and continuous rational potential function which has a polynomial of at least degree 6 in the denominator is used for the U monomial with fixed parameters. The differences with the model in Section 5.2.1 are, each atomic energy now is also summed with U; the numerator of each partial energy term now has $2 \times n_{\text{basis}}$ tuning coefficients; and the number of partial energy terms can be increased by adding more elementary terms to each of the partial energy term. Here $H = \Theta$ with a total of 64 basis functions up to polynomial degree of 6, hence a total of at least $9 \times 64 + 3 = 579$ tuning parameters overall are present.

5.3 Objective function

The objective function has the form of nonlinear least squares. The objective function considers the possibility of fitting multiple different datasets at once, this is compensated by the addition of two tuning parameters for each nonlinear least squares associated to a particular dataset: energy shift μ and the order of accuracy σ.

In case of the model with trainable pair potential in Section 5.2.1, the first 7 tuning parameters ρ are slightly changed to compensate the constraints, they become $\pi := (\pi_1 = \log(\rho_1)/20, \pi_2 = \log(\rho_2)/20, \pi_3 = \sqrt{\rho_3}, \pi_4 = \sqrt{\rho_4 - \rho_3}, \pi_5 = \sqrt{\rho_5 - \rho_4}, \pi_6 = \sqrt{\rho_6 - \rho_5}, \pi_7 = \sqrt{\rho_7 - \rho_6})$ as the input of the optimization routine; which is then inverted into $\rho := (\exp(20\pi_1), \exp(20\pi_2), \pi_3^2, \pi_4^2 + \rho_3, \pi_5^2 + \rho_4, \pi_6^2 + \rho_5, \pi_7^2 + \rho_6)$ for the potential evaluation. For model with fixed H_2, the tuning coefficients of partial energy terms are scaled by the expected magnitude of Φ.

5.4 High level programming details

To compute the bonding features for fitting PES (function evaluator), a wrapper function which takes in all of the tuning parameters alongside the hyperparameters, and computes all of the features in a certain order, which then returns the energy value at the final step is required. Suppose is the function to evaluate the energy given a matrix of atomic distances $\mathbf{R} \in \mathbb{R}^{n_{\text{data}} \times f_{\text{free}}}$, array of atomic coordinates $\mathbf{X}$ with the shape of $(n_{\text{data}}, n_{\text{atom}}, n_{\text{dim}})$, parameter vector ρ, and hyperparameter vector σ; with a vector of energy $\mathbf{V} \in \mathbb{R}^{n_{\text{data}}}$ as the returned quantity. The features (and sub features) within the function are computed in this order:

- Reference energy U.
- Bond strength b (sub feature for coordination vector Y).
- Coordination vector Y.
- Difference array Δ (sub feature for orientation vector r).
- Orientation vector r.
- Gram matrix G.
- Array of basis functions Φ.
- Partial (atomic) energy term $\epsilon[i]$, which finally sums into the energy value.

Additionally, index computation is done pre-function evaluation once for each atomic system, this indexer is passed into the function evaluation as an input. Suppose that the indexing function is this returns the n-body indices of molecular system containing n-atoms, let it be stored in a matrix $D \in \mathbb{Z}^{n_{\text{atom}} \times n_{\text{atom}}-1}$. For example, with $n_{\text{atom}} = 4$, the indexer has entries this is useful for vectorization of any arithmetic operations which make use the degree of freedom indices given an ordered vector of distances $\mathbf{R}_k \in \mathbf{R}, k = 0, 1, 2, ...n_{\text{data}}$, for example, for 4-body terms summation (using the indexer in the example above): $V_k = \sum_{j \neq i} \mathbf{R}_{k,i,j}$ can be represented by This indexer allows computations on vectors of distances without needing to transform matrix of distances

into an array of adjacency matrices with 0 at the diagonals (in graph-terms) with $(n_{\text{data}}, n_{\text{atom}}, n_{\text{atom}})$ shape, which in turn reduces the space complexity. The indexer is used in reference energy U, coordination vector Y, and orientation vector r.

Reference energy U. First the function decides the diatomic energy value within certain ranges for each element of $\mathbf{R}$ matrix. This is done using a sub function containing simple switch-statement where $C, R_h, R_C, R_0 \in \rho$ and $g \in \sigma$, then let $V^{\text{diatomic}} \in \mathbb{R}^{n_{\text{data}} \times f_{\text{free}}}$ be the matrix containing all of the returned values from above sub function. Next, the degree of freedom of each $V_k^{\text{diatomic}} \in \mathbb{R}^{f_{\text{free}}}$ is summed by using the indexer (essentially reducing the f_{free} index to n_{atom}), hence the array U has $(n_{\text{atom}}, n_{\text{data}})$ shape.

Bond strength b. The bond strength function computes the bond strength of each element of $\mathbf{R}$ matrix, where $\overline{R}, R_m, \underline{R} \in \rho$ and $d_{\text{deg}}, e \in \sigma$. First t_0 is obtained from a function which computes a smooth functional (polynomial with at least degree 2 in each numerator and denominator) then matrix $t \in \mathbb{R}^{n_{\text{data}} \times f_{\text{free}}}$ is computed using the same function by passing the matrix $\mathbf{R}$ at the R_m argument position. Both t_0 and t are used for the bond strength computation, which is essentially a similar switch sub function as the reference energy U (notice the slight differences of the inequality relations) then the result is stored in $s \in \mathbb{R}^{n_{\text{data}} \times f_{\text{free}}}$ matrix. On the next step, the bond strength s will need to be wrapped in the terms of polynomials so that s can have different degree of powers, in order to do this, the polynomials are evaluated recursively in terms of Tchebyshev polynomials hence b is an array with $(d_{\text{deg}}, n_{\text{data}}, f_{\text{free}})$ shape.

Coordination vector Y. The function for coordination vector Y simply computes the sum of the bond strength b by using the indexer in the same manner with the second half of U computation. Hence Y has the shape of $(n_{\text{atom}}, d_{\text{deg}}, n_{\text{data}})$.

Difference array Δ. The delta array Δ with shape $(n_{\text{data}}, f_{\text{free}}, n_{\text{dim}})$ (specifically in molecular systems, $n_{\text{dim}} = 3$ always) contains $\mathbf{X}_{:,j} - \mathbf{X}_{:,i}, i < j$, where $\mathbf{X}$ with the shape of $(n_{\text{data}}, n_{\text{atom}}, n_{\text{dim}})$ is the matrix containing molecular coordinates. The loop goes over the n_{atom} index hence this is sort of the "inverse" of the indexer, since this transforms n_{atom} into f_{free}.

Orientation vector r. To compute the orientation vector r, first the (elementwise) array multiplication between the $(n_{\text{data}}, f_{\text{free}})$ indices of b and $(n_{\text{data}}, f_{\text{free}}, 3)$ indices of Δ needs to be computed for $d = 1, 2.., d_{\text{deg}}$, then let the result be stored in an array η with shape $(d_{\text{deg}}, n_{\text{data}}, f_{\text{free}}, 3)$. Finally η is summed using the indexer rule, which the final result is stored in array r with the shape of $(n_{\text{atom}}, d_{\text{deg}}, n_{\text{data}}, 3)$.

Gram matrix G. For each atom $i = 1, 2, ...n_{\text{atom}}$, for each $d_1 = 1, 2, ...d_{\text{deg}}$, and for each $d_2 = 1, 2, ...d_{\text{deg}}$: this means that the elementwise array multiplication is done on the $(n_{\text{data}}, 3)$ indices, then the last index is summed (denoted by axis $= -1$ argument), hence the last index disappears, this represents the dot product between vectors. In the end G is an array with $(n_{\text{atom}}, d_{\text{deg}}, d_{\text{deg}}, n_{\text{data}})$ shape.

Array of basis functions Φ. The array Φ with shape $(n_{\text{atom}}, n_{\text{basis}}, n_{\text{data}})$, consists of basis functions which are composed of the feature array which have been calculated previously, simply for each degree, all possible polynomials are enumerated, each polynomial is the combination of each feature array, for example which shows that $\Phi = (U, Y_1, UY_1, Y_1^2, Y_2, G_{1,1}, ...)$.

Atomic energy terms $\epsilon[i]$. The atomic energy terms $\epsilon[i]$ are the sums of partial energy terms, where the partial energy terms are smooth rational functionals which depend on the tuning parameter matrices $A, B, C \in \mathbb{R}^{2\times n_{\text{basis}}}$. Technically each element of each tuning parameter vectors are multiplied by Φ_k, in simpler terms where A, B, or C is an input argument in place of E. Then the partial energies are combined in a certain way, for example $\epsilon[i] = V_{\text{part}}(A) - \sqrt{V_{\text{part}}(B) - V_{\text{part}}(C)}$. Finally the total energy of the system is $V = \sum_{i=1}^{n_{\text{atom}}} \epsilon[i]$.

5.5 Feature statistics

The statistics of all features described within this section are displayed in Table 4 and Table 5. The features are used to compute the energy of 6032 data points of H_3 dataset.

Table 4: Statistics of each feature per atom index using pre-optimization parameters.

feature[atom]	max	min	mean	std
$U[1]$	2.481150e+24	7.461384e+22	6.254963e+23	5.593566e+23
$U[2]$	2.435484e+24	4.340160e+22	5.548411e+23	5.030545e+23
$U[3]$	1.682283e+24	5.973577e+21	2.307342e+23	2.703311e+23
$Y[1]$	2.000000e+00	-2.234326e-02	3.997134e-01	7.999023e-01
$Y[2]$	2.000000e+00	-8.430624e-02	3.992064e-01	7.997078e-01
$Y[3]$	2.000000e+00	-1.055023e-01	3.989212e-01	7.996413e-01
$G[1]$	7.428453e-02	0.000000e+00	3.107268e-03	1.070428e-02
$G[2]$	1.743001e+00	7.438274e-16	3.483051e-02	1.402622e-01
$G[3]$	2.486132e+00	2.166612e-15	5.631022e-02	2.146574e-01
$\Phi[1]$	3.969833e+25	-1.901475e+22	3.286893e+23	1.990261e+24
$\Phi[2]$	3.896758e+25	-2.559325e+22	2.923480e+23	1.776411e+24
$\Phi[3]$	2.691653e+25	-7.797651e+20	1.213176e+23	8.457775e+23

Table 5: Statistics of each feature per atom index using post-optimization parameters, with RMSE = 0.069 across all H_3 data.

feature[atom]	max	min	mean	std
$U[1]$	1.450460e-276	-2.679209e-278	3.114115e-277	0.000000
$U[2]$	1.422420e-276	-4.622908e-278	2.679955e-277	0.000000
$U[3]$	9.601608e-277	-6.980936e-278	6.897668e-278	0.000000
$Y[1]$	1.999660e+00	-4.937911e-01	3.189007e-01	0.644286
$Y[2]$	1.999115e+00	-4.923274e-01	2.960542e-01	0.580598
$Y[3]$	1.999115e+00	-4.978102e-01	2.358900e-01	0.533017
$G[1]$	4.256561e+00	0.000000e+00	3.034012e-01	0.463637
$G[2]$	3.609277e+00	5.088141e-06	3.312816e-01	0.419666
$G[3]$	5.362877e+00	2.099490e-06	4.990154e-01	0.741347
$\Phi[1]$	3.197281e+01	-1.031954e+00	8.880903e-01	3.839208
$\Phi[2]$	3.192923e+01	-1.024145e+00	6.708816e-01	3.008128
$\Phi[3]$	3.192923e+01	-1.031435e+00	6.205682e-01	2.927580

6 Primitive bonding features

6.1 Coordination function

A spline-like C^2 function if $t^2 < 1$ and 0 otherwise, in the form of

$$z(t) = a - bt - ct^5 + dt^3, \tag{152}$$

where $t \in \mathbb{R}$. The coordination function is formed by fulfilling these conditions:

$$\begin{aligned} z(-1) &= 1, \\ z(1) &= 0, \\ z'(\pm 1) &= 0, \\ z''(\pm 1) &= 0. \end{aligned} \tag{153}$$

In order to obtain the correct a, b, c, d, the conditions in (153) need to be expanded, which results in a linear system of equations

$$\begin{aligned} z(-1) &= a + b + c - d = 1, \\ z(1) &= a - b - c + d = 0, \\ z'(\pm 1) &= -b - 5c + 3d = 0, \\ z''(\pm 1) &= 20c - 6d = 0. \end{aligned} \tag{154}$$

By solving (154) and writing the coefficients explicitly for the $t^2 < 1$ condition, the coordination function is

$$z(t) = \begin{cases} 1, & \text{if } t \le -1 \\ 0, & \text{if } t \ge 1 \\ \frac{1}{2} - t\left(\frac{15}{16} + s\left(\frac{3}{16}s - \frac{10}{16}\right)\right). & \text{if } s = t^2 < 1 \end{cases} \tag{155}$$

(155) can be verified by differentiating $z(t)$ directly and expanding the second last condition in (153)

$$\begin{aligned} z'(t) &= \frac{d}{dt}\left[\frac{1}{2} - t\left(\frac{15}{16} + t^2\left(\frac{3}{16}t^2 - \frac{10}{16}\right)\right)\right] = -\frac{15}{16}(t + t^4 - 2t^2), \\ z'(\pm 1) &= 0. \end{aligned} \tag{156}$$

6.2 Bump functions

The bump functions, has the form

$$f_k(t) = (1 - (t-k)^2)_+^3 \tag{157}$$

where $k \in \mathbb{Z}$, is a C^2 function everywhere except when $t \geq k+1$ or $t \leq k-1$, in which f_k is 0. The transformed $r \in \mathbf{R}$ has the form of

$$t = c_{xy}(r^2 - r_{xy}^2), \tag{158}$$

where

$$c_{xy} = \frac{N}{\bar{r}_{xy}^2 - \underline{r}_{xy}^2} \tag{159}$$

is a coefficient which depends on the largest atomic distance $\bar{r}_{xy}^2$ and smallest atomic distance $\underline{r}_{xy}^2$. While r_{xy} itself denotes the equilibrium distance of xy atomic pair (i.e., the closest distance where $dV/dr \approx 0$), for example, $r_{\mathrm{HH}} \approx 1.4172946$ Bohr.

From (158) we can define the upper bound and lower bound of t as

$$\begin{aligned} \underline{t} &= c_{xy}(\underline{r}_{xy}^2 - r_{xy}^2), \\ \bar{t} &= c_{xy}(\bar{r}_{xy}^2 - r_{xy}^2), \end{aligned} \tag{160}$$

when $\bar{t}$ and $\underline{t}$ are rounded to the nearest integers, this tells

$$N = \bar{t} - \underline{t}, \tag{161}$$

which is a hyperparameter that describes the number of the primitive functions needed to approximate the associated pair-potential. The list of required k for the total number of $N-1$ bump functions f_k can be obtained by

$$k \in \mathbb{Z} \cap \underline{t} < k < \bar{t}. \tag{162}$$

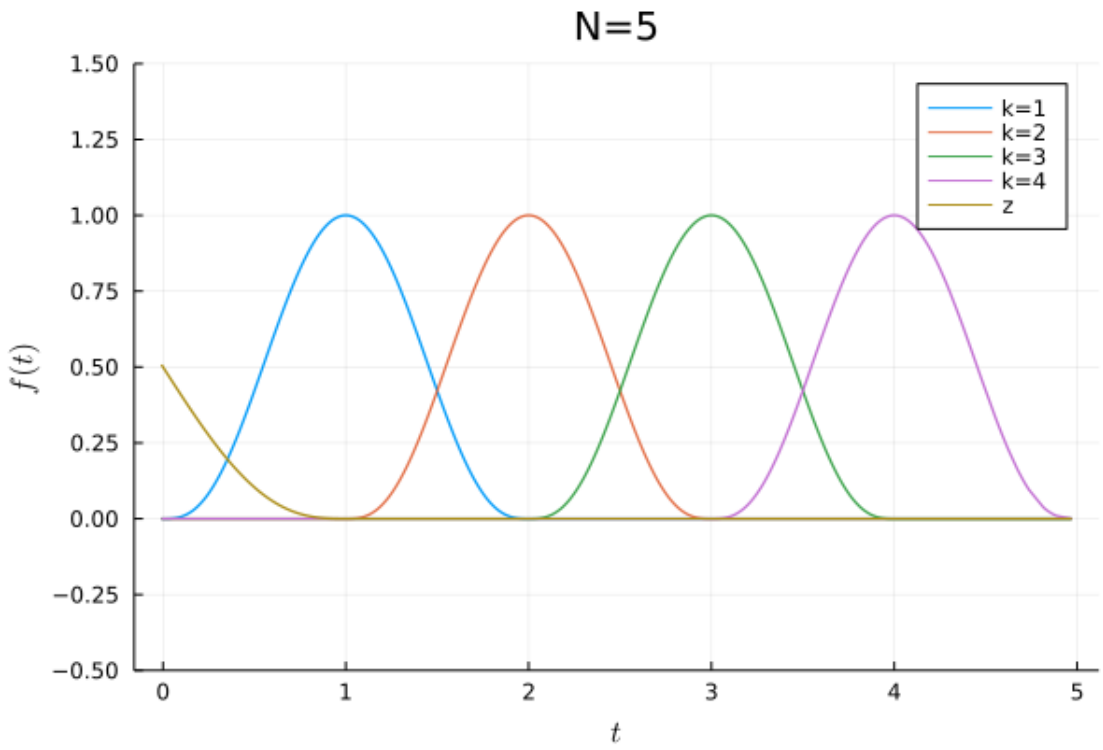

Figure 4: $z(t)$ and $f_k(t)$ for $k = 1, 2, ...4$ computed from $\mathbf{R}_{:,3}$ of H_3 dataset.

6.3 Linearizable bump functions

Total params: $2N + 2$. Scaler:

$$q := \frac{N}{1+\rho} \in (0, N] \tag{163}$$

Basic bump:

$$h_k := [1 - (q-k)^2]_+^3 \tag{164}$$

Basis LC:

$$u(q) = \sum_{k=0}^{N} (\theta_k(q-k) + \theta_{k+N+1}) h_k \tag{165}$$

Bumpsum:

$$w(q) = \sum h_k \tag{166}$$

Since most h_k vanish:

$$\begin{aligned} &i := q \geq 1, \\ &\epsilon := i - q \in [0, 1) \implies q = i - \epsilon \in (i-1, i] \end{aligned} \tag{167}$$

Simplified expressions:

$$\begin{aligned} \alpha &:= (\epsilon(2-\epsilon))^3 \\ \beta &:= (1-\epsilon^2)^3 \end{aligned} \tag{168}$$

Simplified bump:

$$h_k = \begin{cases} \alpha & \text{if } k = i-1, \\ \beta & \text{if } k = i, \\ 0 & \text{otherwise} \end{cases} \tag{169}$$

Simplified basis :C:

$$u(q) = (\theta_i(q-i+1) + \theta_{i+N})\alpha + (\theta_i(q-i) + \theta_{i+N+1})\beta \tag{170}$$

Simplified bumpsum:

$$w(q) = \alpha + \beta \tag{171}$$

Basic features:

$$x := h_k(q)/w(q), \tag{172}$$

$$y := (q-k)h_k(q)/w(q) \tag{173}$$

Linearized bumps:

$$\begin{aligned} u(q) &= \sum \left(\frac{h_k(q)}{w(q)}\theta_k + \frac{(q-k)h_k(q)}{w(q)}\theta_{k+N+1} \right) \\ &= A\theta = \sum A_{ik}\theta_k, \text{ if } q = q_i. \end{aligned} \tag{174}$$

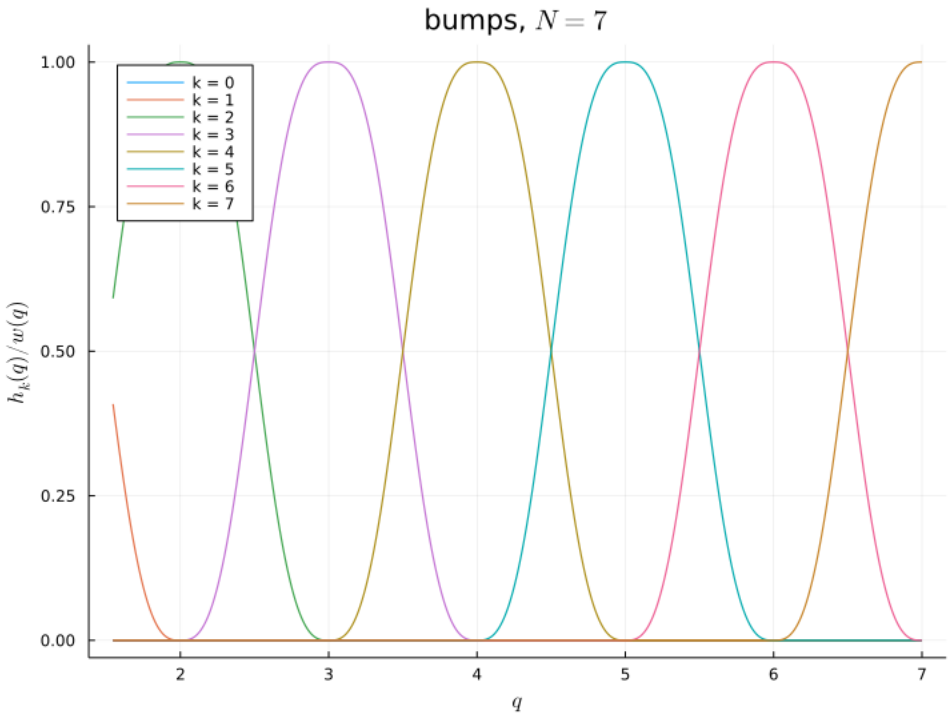

Figure 5: Bumps from $h_k(q)/w(q)$ for $N = 7$.

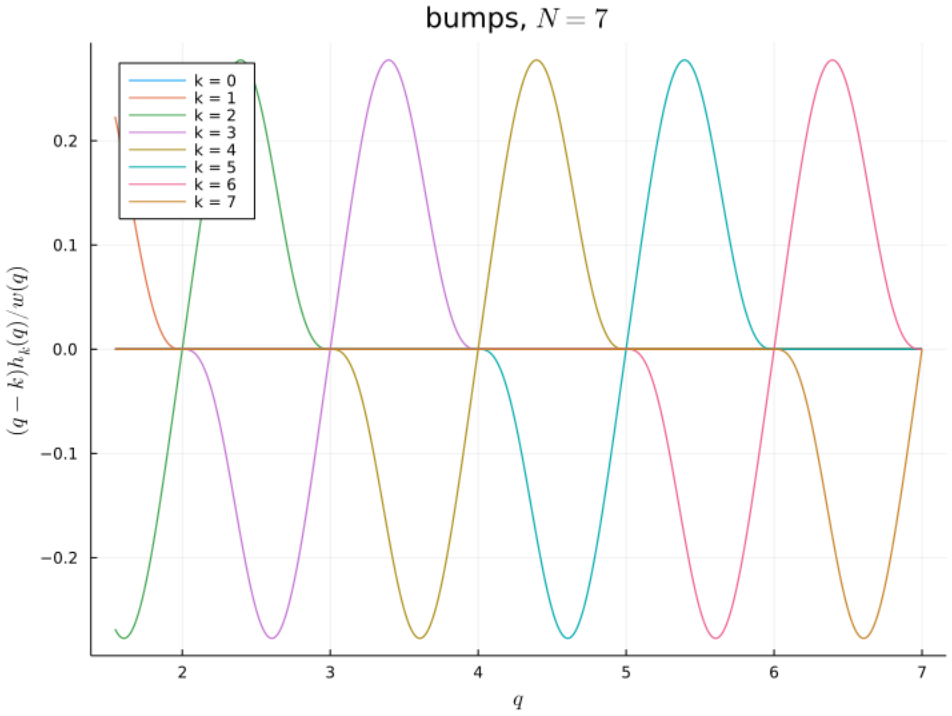

Figure 6: Bumps from $(q-k)h_k(q)/w(q)$ for $N = 7$.

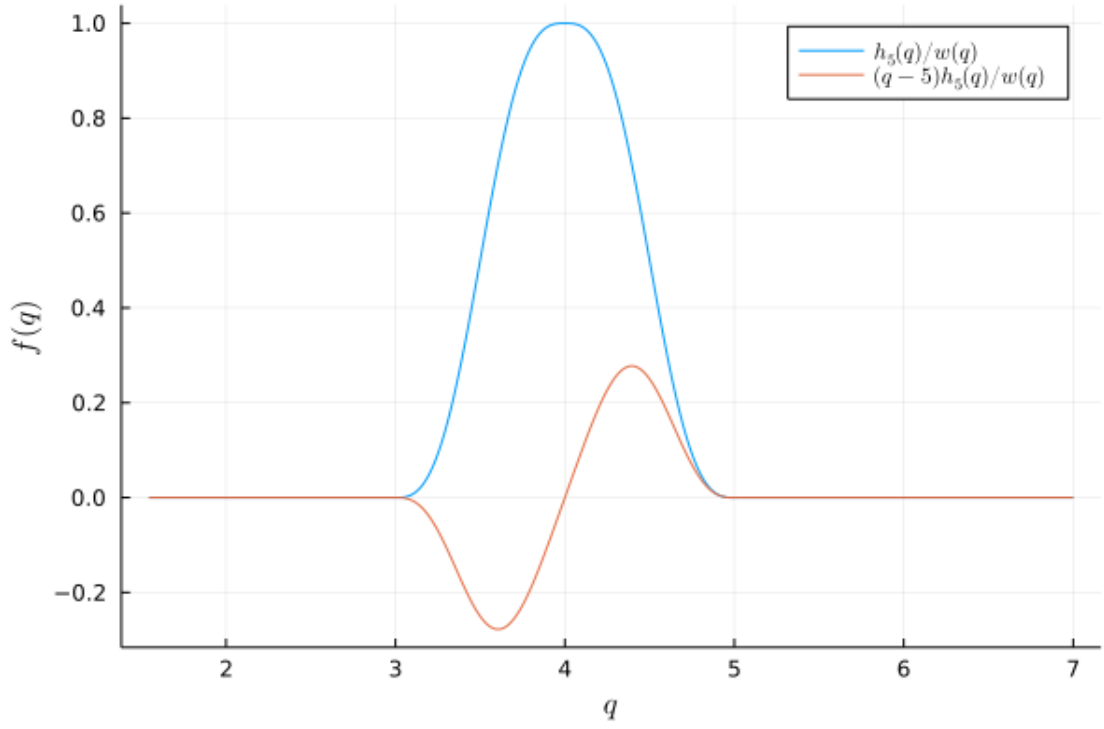

Figure 7: Bumps for x and y when $k = 5$.

7 Datasets

7.1 H_2 data

The data points for ground state H_2 (Figure 8) are obtained from Hager-Roiser [**?**], it contains:

- Data sizes:
  - Kolos & Wolniewicz (1965) [**?**]: 87 data points.
  - Wolniewicz (1993) [**?**]: 670 data points.
- The energy values of Wolniewicz (1993) are shifted by $V_i := V_i - V_{670}$, $i = 1, 2, ..., 670$, to accommodate the asymptotic energy $\lim_{R\to\infty} V(R) = 0$.

7.1.1 H_2^+ data

The H_2^+ datasets are obtained from Sharp (1971) [**?**], with the following information:

- Data sizes:
  - $X^2\Sigma_g^+ 1s\sigma_g$ state: 29 data points.
  - $\sigma_g 1s$ state: 29 data points.
  - $2p\sigma_u$ state: 29 data points.
  - $\sigma_u 1s$ state: 29 data points.
- The energy and distance units are converted from (Angstrom, eV) to (Bohr, Hartree).
- The data of $\sigma_u 1s$ state is corrected by removing the (R_{25}, V_{25}) data point.
- For each state, the data are shifted by the first ionization energy of H, $V_{\text{ion}} = 1312$ kj/mol.
- Only $X^2\Sigma_g^+ 1s\sigma_g$ and $2p\sigma_u$ states are used while $\sigma_g 1s$ and $\sigma_u 1s$ states are excluded due to redundancy.

7.2 O_2 data

The data points for O_2 (Figure 9) are provided by Bytautas et al (2010)[**?**], which describes:

- Data sizes:
  - $^3\Sigma_g^-$ state: 26 data points.
  - $^1\Delta_g$ state: 26 data points.
  - $^1\Sigma_g^+$ state: 26 data points.
- The energy and distance units are converted from (Angstrom, millihartree) to (Bohr, Hartree) (1 Hartree = 1000 millihartree).

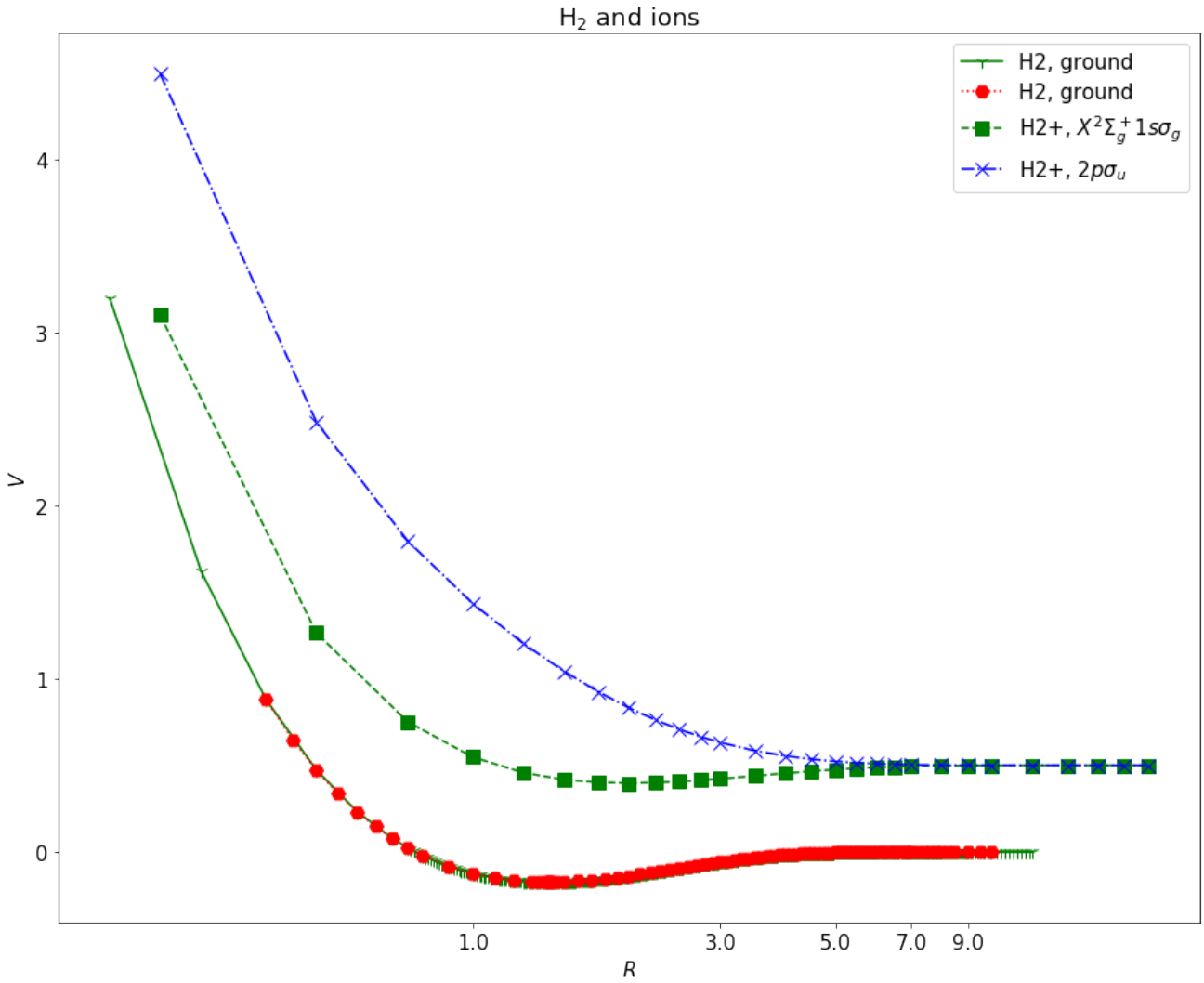

Figure 8: Potential energy curves of H_2 and its ions.

7.2.1 O_2^+ data

The data points for O_2^+ are provided by Xavier et al (2018) [?], the information contains:

- Data sizes:
  - $X^2\Pi_g$ state:
    - AV5Z method: 51
    - AV6Z method: 51
    - CBS method: 51
- The unit of R is converted from Angstrom to Bohr.
- The V is shifted by the first ionization energy of O, $V_{\text{ion}} = 1313.9$ kj/mol.
- Only one dataset is chosen due to the fact that each dataset is identical to each other (e.g., only the dataset from CBS method is picked).

7.3 OH data

The obtained OH datasets (Figure 10) are described by:

- Data sizes:
  - $X^2\Pi$ state:
    - Pradhan (1995) [?] (from Shizgal (1999) [?]): 29 data points.
    - Nemukhin & Grigorenko (1997) [?] (from Shizgal (1999) [?]): 32 data points.
    - van Dishoeck et al (1983) [?] (from Shizgal (1999) [?]): 24 data points.
    - Hodges (1993) [?] (from Shizgal (1999) [?]): 32 data points.
    - Chu et al (1974) [?] using MCSCF method: 22 data points.
    - Chu et al (1974) [?] using CI method: 22 data points.
    - Werner et al (1983) [?], using MCSCF-SCEP method: 17 data points.
    - Werner et al (1983) [?], using SCEP-CEPA: 12 data points.
  - $A^2\Sigma^+$ state:
    - Chu et al (1974) [?], using MCSCF method: 22 data points.
    - Chu et al (1974) [?], using CI method: 22 data points.
  - $^4\Sigma^-$ state, obtained from Shizgal (1999) [?]:
    - van Dishoeck & Dalgarno (1983) [?]: 23 data points.
    - Hodges (1993) [?]: 23 data points.
    - Cooper et al (1984) [?]: 23 data points.
  - $^2\Sigma^-$ state, obtained from Shizgal (1999) [?]:
    - Pradhan (1995) [?]: 21 data points.

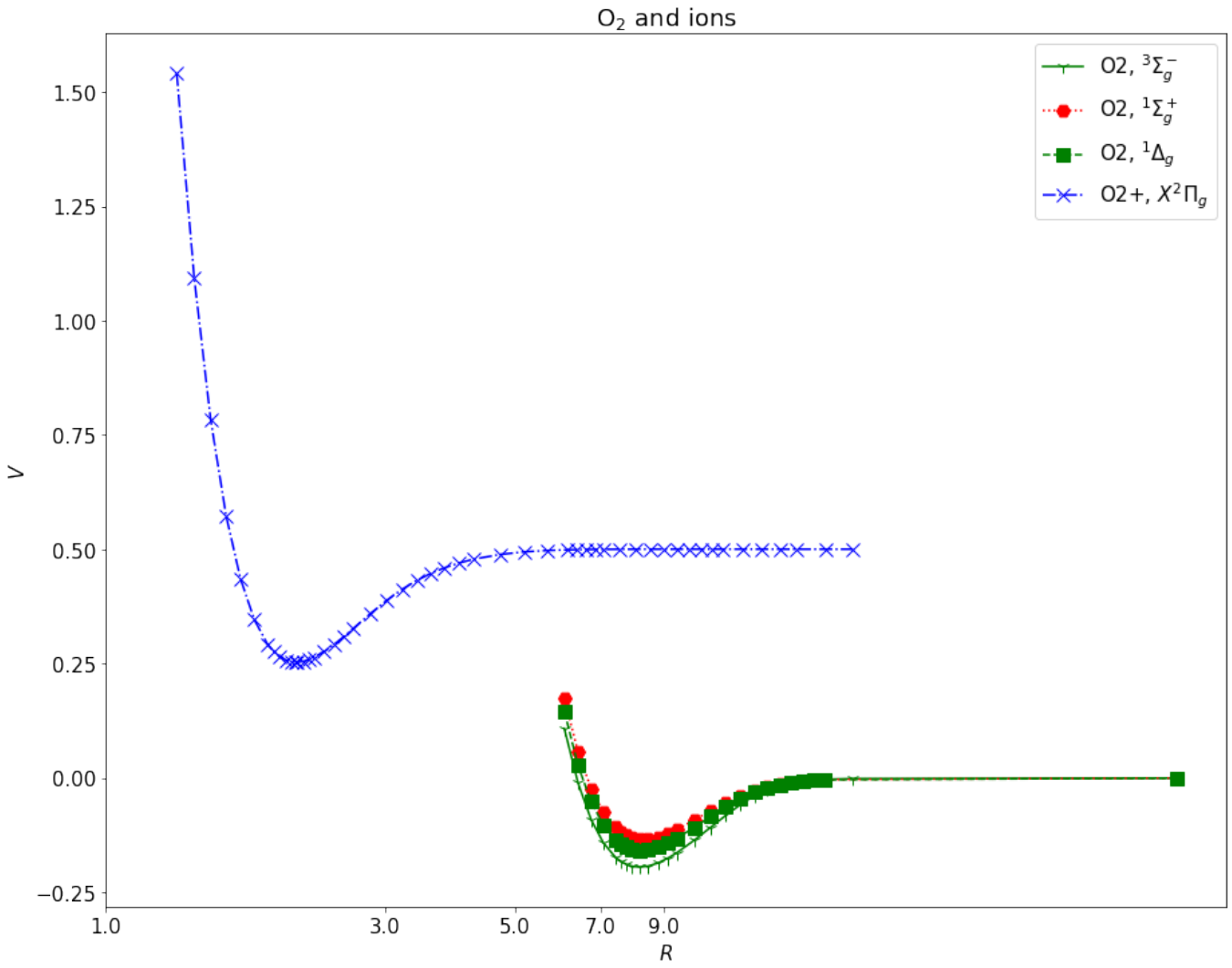

Figure 9: Potential energy curves of O_2 and its ions.

  - ∗ van Dishoeck et al (1983) [?]: 14 data points.
  - ∗ Hodges (1993) [?]: 24 data points.
- – $^4\Pi$ state, obtained from Shizgal (1999) [?]:
  - ∗ van Dishoeck & Dalgarno [?]: 20 data points.
  - ∗ Hodges (1993) [?]: 20 data points.
  - ∗ Cooper et al (1984) [?]: 20 data points.

- Data shifts:
  - – V values for each dataset of Chu et al (1974) [?] for $A^2\Sigma^+$ and $X^2\Pi$ states are shifted by its corresponding V_n.
  - – V values on each dataset of Werner et al (1983) [?] for $X^2\Pi$ state are shifted by its corresponding V_n.

- Data corrections:

  - Data of Chu et al (1974) [**?**] for $X^2\Pi$ state is corrected by removing (R_1, V_1).
  - Data of van Dishoeck & Dalgarno [**?**] for $^4\Pi$ state is corrected by removing (R_1, V_1).
  - Data of van Dishoeck & Dalgarno [**?**] for $X^2\Pi$ state is corrected by removing (R_4, V_4).

- The datasets used for fitting are:
  - $\{X^2\Pi$, Pradhan(1995)$\}$.
  - $\{X^2\Pi$, Nemukhin and Grigorenko(1997)$\}$.
  - $\{X^2\Pi$, vanDishoeck et al(1983)$\}$.
  - $\{^2\Sigma^-$, Pradhan(1995)$\}$.
  - $\{^2\Sigma^-$, vanDishoeck et al(1983)$\}$.
  - $\{^2\Sigma^-$, Hodges(1993)$\}$.
  - $\{^4\Sigma^-$, vanDishoeck and Dalgarno(1983)$\}$.
  - $\{^4\Sigma^-$, Hodges(1993)$\}$.
  - $\{^4\Pi$, vanDishoeck and Dalgarno(1983)$\}$.
  - $\{^4\Pi$, Hodges(1993)$\}$.
  - $\{X^2\Pi$, Chu et al (1974)$\}$.
  - $\{A^2\Sigma^+$, Chu et al (1974) $\}$.
  - $\{A^2\Sigma^+$, Fallon et al(1960)$\}$.

- Below datasets are excluded due to disagreeing with the majority of the datasets in the same molecular state:
  - $\{^4\Sigma^-$, Cooper et al(1984)$\}$.
  - $\{^4\Pi$, Cooper et al(1984)$\}$.
  - $\{A^2\Sigma^+$, Chu et al (1974)$\}$.
  - $\{X^2\Pi$, Hodges(1993)$\}$.
  - $\{X^2\Pi$, Werner et al(1983)$\}$.
  - $\{X^2\Pi$, Chu et al (1974)$\}$.

7.3.1 OH^+ data

The OH^+ data is obtained from Werner et al (1983) [**?**] and Xavier et al (2018) [**?**], the contained information is:

- Data sizes:
  - $X^3\Sigma^-$ state:
    - Xavier et al (2018) [**?**]:
      - AV5Z method: 32

        - AV6Z method: 32
        - CBS method: 32
    - Werner et al (1983):
        - MCSCF-SCEP method: 12 data points.
        - SCEP-CEPA method: 12 data points.

- The unit of Xavier et al (2018) [**?**] R is converted from Angstrom to Bohr.
- The V values are shifted by the first ionization energy of O, $V_{\text{ion}} = 1313.9$ kj/mol.
- Only the datasets of Xavier et al [**?**] with AV5Z method and Werner et al (1983) with SCEP-CEPA method are included, meanwhile the excluded datasets (in the form of {state, author}) are:
    - $\{X^3\Sigma^-$, Werner et al(1983)$\}$ due to disagreeing with other datasets.
    - $\{X^3\Sigma^-$, Xavier(2018) - AV6Z method$\}$ due to redundancy.
    - $\{X^3\Sigma^-$, Xavier(2018) - CBS method$\}$ due to redundancy.

7.3.2 OH^- data

The OH^- data is obtained from Werner et al (1983) [**?**], the contained information is:

- Data sizes:
    - $X^1\Sigma^+$ state:
        - MCSCF-SCEP method: 17 data points.
        - SCEP-CEPA method: 10 data points.

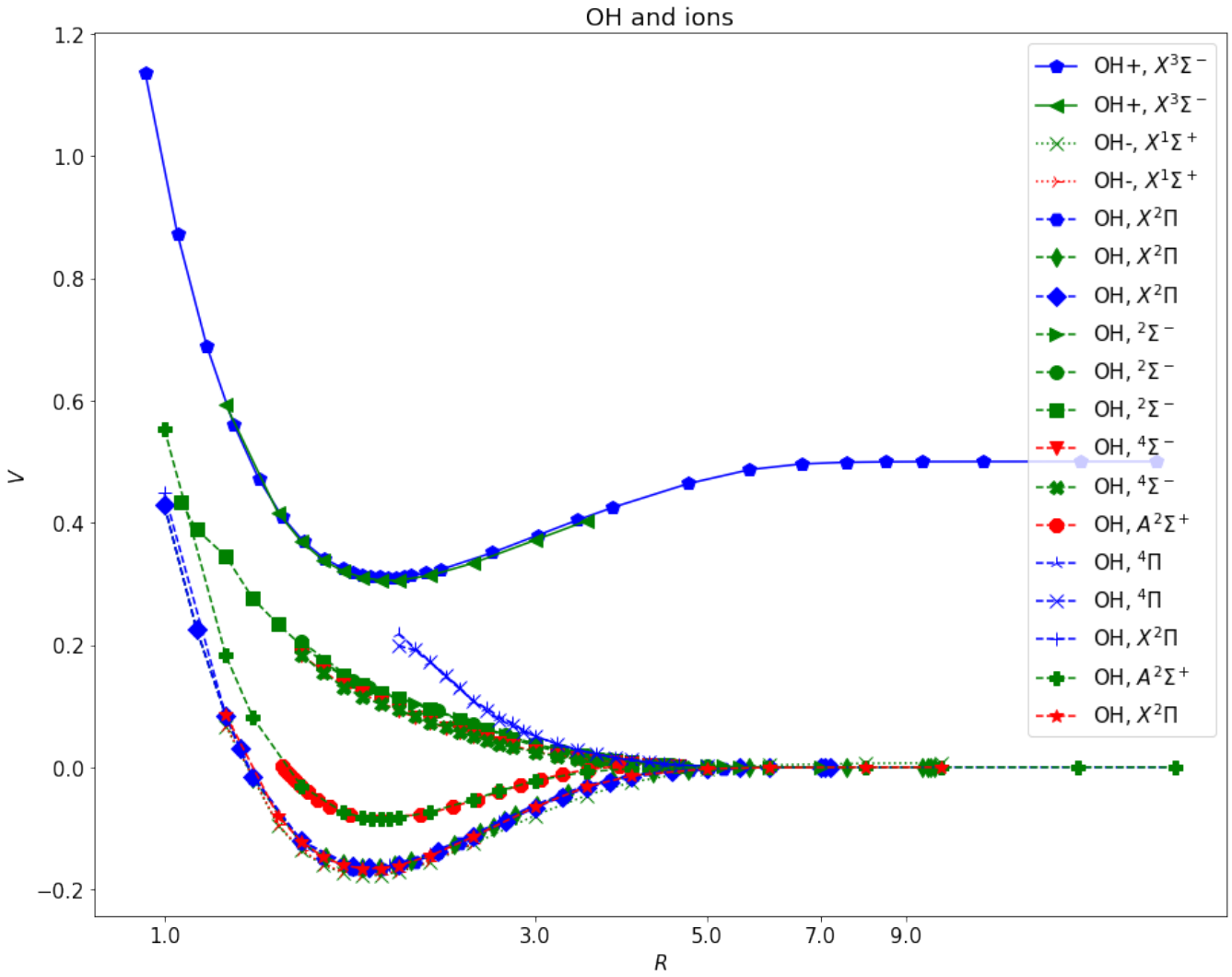

Figure 10: Potential energy curves of OH and its ions.