
LONG-RANGE-FIT

A PROGRAM TO STUDY THE LONG-RANGE INTERACTIONS BETWEEN
ATOMS AND MOLECULES

USER'S GUIDE

ADRIAN BATISTA-PLANAS
ERNESTO QUINTAS-SÁNCHEZ
RICHARD DAWES

*Department of Chemistry
Missouri University of Science & Technology
Rolla 65401 MO, USA*

SOFTWARE VERSION: v5.0

Table of Contents

1	Introduction	3
1.1	About LRF	3
1.2	Contact Information	5
1.3	Documentation and Related Resources	5
2	Getting Started	6
2.1	How to Obtain LRF	6
2.2	How to Install LRF	6
3	How to use LRF	9
3.1	Initial Considerations	9
3.1.1	LRF capabilities and workflow	9
3.1.2	Internal coordinates and units	10
3.1.3	Orientation of the symmetry elements	11
3.1.4	The input data file	11
3.2	Graphical User Interface	12
3.2.1	<i>Setup</i> tab	12
3.2.2	<i>Expansion</i> tab	14
3.2.3	<i>Dashboard</i> tab	16
3.2.4	<i>Logs</i>	18
3.2.5	<i>Model Visualization</i>	23
3.2.6	<i>Fitting Options</i>	25
3.2.7	<i>Workspace Portability</i>	31
3.2.8	<i>Coefficients file & FORTRAN subroutine</i>	31
4	Test Cases & Examples	33
4.1	Charged Molecules: $\text{CF}^+ - \text{H}_2$	33
4.2	Anisotropic Molecules: $\text{C}_{10}\text{H}^- - \text{H}_2$	38
4.3	Identical Molecules: CO-Dimer	41
4.4	Beyond linear molecules : O_3 - Ar	45

Copyright © 2024
Missouri University of Science & Technology
All Rights Reserved

The program Long-Range-Fit (LRF) is provided on an 'as is' basis, without any expressed or implied warranties. Users accept all risks associated with the use of this software. No guarantee of fitness for any specific purpose is provided. Individuals are permitted to create copies of LRF for academic and nonprofit purposes, including modification of said copies. However, distribution of original or modified copies of the source material, executables, or documentation, to users at any location other than their group's research environment is prohibited. Prior to utilization of this software, it is imperative to thoroughly review the complete LRF license, and to complete, sign, and return the associated license agreement. Any commercial utilization of LRF is subject to a distinct agreement outlined within the comprehensive LRF license.

version 5.0
November-2024

Introduction

1.1 About LRF

Long-Range-Fit (LRF) is a MATLAB-based tool designed to facilitate the construction and study of potential energy surfaces (PESs) corresponding to the long-range interaction region of atomic and molecular systems. In most foreseeable applications, LRF will be used to complement an *ab initio* study of a system of interest, for which a dataset of energies from electronic structure calculations has been produced and fitted into a PES-function describing the close interaction region. Many approaches to fitting the close interaction region have been developed, including our automated AUTOSURF method. However, despite the importance of long-range interactions, and the sensitivity of quantities such as low temperature scattering cross sections to its accuracy, too often the long-range of the PES is treated as an after-thought, represented by just one or two leading terms from the multipole expansion.

LRF offers the user many capabilities. Based on the charge and point group symmetry for each of the interacting partners, LRF can first produce tables of all relevant interaction terms classified as Electrostatic, Induction, and Dispersion — up to 15th order in each type, where the order specifies the distance dependence. For each fragment, an exhaustive list of more than 50 point group symmetries is included, covering nearly every conceivable case. This analysis provides insight into

the nature of the most important interactions. If enough quantities — multipole and polarizability terms — are accurately known for each fragment, then those can be directly inputted as fixed parameters and LRF can render a usable representation of the long-range interactions, truncated to any chosen order. An efficient implementation has been derived to include all of the interactions up to the maximum 15th order, some of which individually involve hundreds or thousands of algebraic terms. For any constructed representation, 1- and 2-D plots of various cuts through the PES can be rendered directly in the app, highlighting the contribution of various individual or combined terms.

More commonly, the molecular parameters are not known accurately, if at all. These quantities, even the lowest order ones such as dipoles, are difficult and expensive to compute accurately using the highest levels of quantum chemistry. This is where LRF shines. Given a set of raw *ab initio* data for the combined interacting system (distributed in arbitrary fashion over the long-range region of the PES), LRF will fit the data using all of the relevant terms up to a user-specified maximum order, correctly symmetry-adapted for each particular system. LRF will also “zero” the PES representation at infinite separation of the fragments, and export a FORTRAN subroutine for external evaluation of the potential. The LRF function for the long-range can then be easily and smoothly connected to any preferred representation of the close interaction region, providing a global description of the PES. In the overall scope of a PES-construction project, using LRF to handle the long-range adds very little to the computational cost. For most applications, a few hundred data points scattered across the long-range region of the PES is sufficient to well-determine a physically correct representation with exquisite accuracy. A few example applications — for which the final fitting error in the long-range is on the order of a few thousandths of a wavenumber — are provided with the code and discussed in the last section of this manual.

LRF is freely available for Windows and Linux machines and both versions are precisely equivalent, with the same user experience via a MATLAB-based Graphic User Interface (GUI). Though based on MATLAB, the use of LRF does not require installation of MATLAB or even a MATLAB license. Instead, a freely available MATLAB Runtime library is used and is all that is needed to provide the full functionality of the app. Moreover, if not already installed on the user’s machine, the compatible version of the MATLAB Runtime library is automatically downloaded, installed, and linked by the provided LRF installer. MATLAB Runtime is a freely-available collection of shared libraries, MATLAB code, and other relevant files needed to run compiled MATLAB applications on a target system without a licensed copy of MATLAB; all relevant documentation can be found at MATLAB’s [website](#).

1.2 Contact Information

Team Members	Email
Richard Dawes*	dawesr@mst.edu
Ernesto Quintas-Sánchez*	quintassancheze@mst.edu
Adrian Batista-Planas*	albgzz@mst.edu

* Department of Chemistry, MS&T, Rolla, USA

1.3 Documentation and Related Resources

Official documentation about the theoretical framework behind LRF can be found in our GitHub [repository](#). LRF publications and other Dawes' research group projects can be found on the group's website: <https://web.mst.edu/~dawesr/>.

Getting Started

2.1 How to Obtain LRF

For non-commercial purposes, LONG-RANGE-FIT (LRF) is freely available for academic users — research institutes, universities, individuals — upon signing a License Agreement. The License form should be requested from the authors. Upon receipt of the completed and signed document, a link for downloading the latest version of LRF will be provided. The following sections contain instructions detailing the LRF installation process.

2.2 How to Install LRF

A list of suggested hardware specifications for the installation and use of LRF is given in Table 1. Most applications of LRF are not very resource intensive and can be performed on any modern PC including laptops. Note however that as discussed in the following sections, the time-to-solution for the non-linear fitting algorithms employed in LRF depends strongly on the size of the data set and especially on the number of fitted terms, which generally increases rapidly with order. For typical applications, fits converge within a few seconds to a few minutes on a laptop. If desired, the user can specify parallel processing parameters employing additional available cores on the machine. Finally, as will be demonstrated in the example

	Prerequisites
Operating System	Windows 10 [version 21H2 or higher] Windows 11 Windows Server 2022 Ubuntu [versions 20.04, 22.04, and 24.04] Debian 12 Red Hat Enterprise Linux [versions 8.4 (or higher) and 9] SUSE Linux Enterprise 15
Co-installed Software	MATLAB Runtime 2024a (Windows installation) MATLAB Runtime 2024b (Linux installation)
Processor	any x86–64 processor
RAM	Minimum: 4 GB
Internet Connection	Only necessary during installation of MATLAB Runtime

Table 1: Recommended minimum hardware specifications.

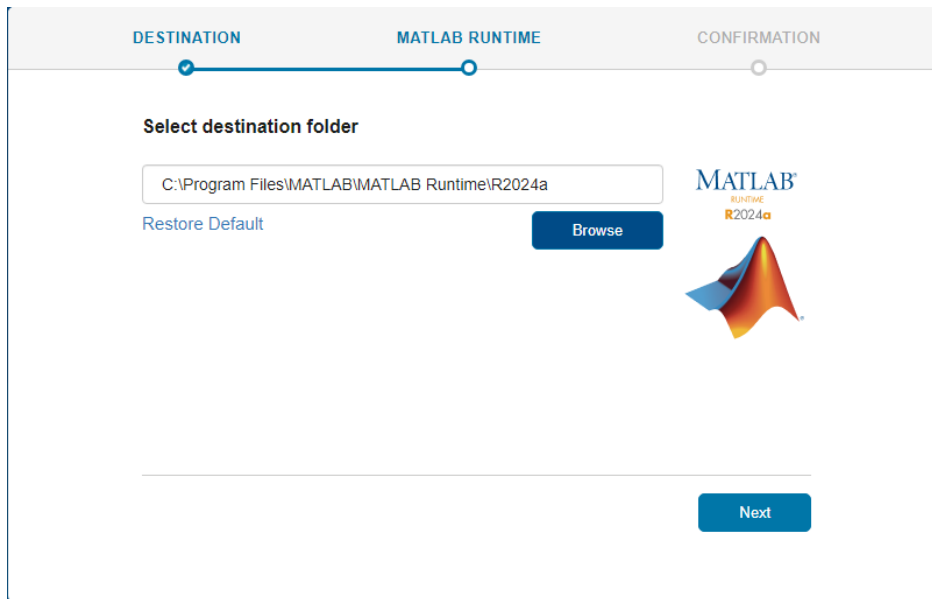
test cases, it is recommended to start by fitting a few leading terms, and then add additional terms sequentially by order, updating the fit at each stage. This allows the user to monitor fit quality by tracking error residuals and other statistics and hence decide at which order to stop.

The installation process is very simple. After returning the signed license agreement, the user will receive an email with a link to download a compressed ZIP file, which upon extraction contains an executable named ‘LRF.exe’. Then, follow the steps described below:

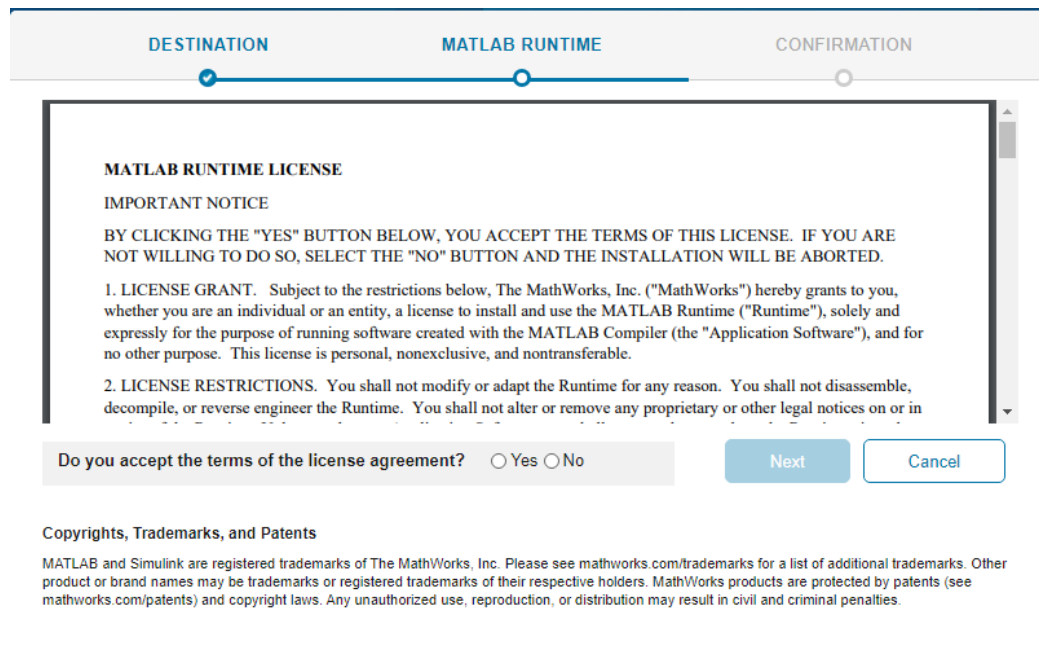
For Windows users

1. **Run the Installer:** Double-click the installer file (‘LRF.exe’) to start the installation process.
2. **Choose Installation Location:** Use the default path, or click “Browse” to select a different folder, Fig. 1 (a).
3. **Accept License Agreement & Install:** Carefully read the terms and conditions. The MATLAB License Agreement must be accepted to continue, and click “Next”, Fig. 1 (b). The installation could take several minutes.
4. **Check Installation:** After the installation is complete, launch the LRF app using the Desktop icon and confirm that it launches as expected.

Administrator privileges are required to execute the MATLAB Runtime installer. During the installation, warning dialogues may appear due to system settings and antivirus software, as the application is not listed in the Microsoft Store.



(a) Installation Location



(b) MATLAB License Agreement

Figure 1: Steps in the Installer Window

For Linux users

How to use LRF

3.1 Initial Considerations

3.1.1 LRF capabilities and workflow

The LRF model for the multipole interactions assumes that both (rigid) molecules/fragments are in their electronic ground state. Also, the user's *ab initio* data should be located in the long-range region of the PES, where the overlap between the charge distribution clouds of each fragment is minimal.

The flowchart shown in Fig. 2 illustrates the five main steps in the LRF workflow:

1. *System Definition*: Specify the point group symmetry and net charge for each molecule, and optionally, initial coefficient values like dipoles or quadrupoles

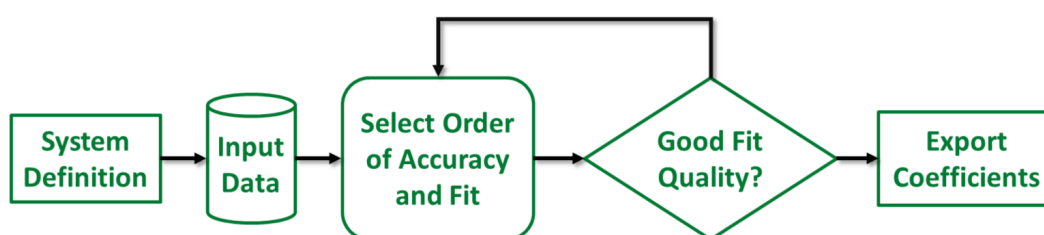


Figure 2: LRF flowchart.

if available. If the molecules share the same symmetry, the user may also specify whether they are identical, or enantiomers.

2. *Input Data*: Input the *ab initio* data to be used in the fit.
3. *Select Accuracy and Fit*: Select the expansion order for the fit — this, combined with the system symmetry, define the terms to be included in the analytical function describing the long-range interactions — then, executes the fitting process.
4. *Check Fit Quality*: Upon completion of the fit, LRF offers different methods and tools to assess the fit's quality, helping determine if the desired accuracy has been reached, ensuring it is adequate for the next phase.
5. *Export Coefficients*: Should the fit be considered satisfactory, the coefficients can be exported into an output file. This file, combined with an available FORTRAN subroutine, allows for a convenient and portable method for evaluating this fit of the PES's long-range region, or integrate it with other codes..?

3.1.2 Internal coordinates and units

Once the molecular system is set, the user can tell to **LRF** what to expect when reading the input data file. The input data file must contain a matrix with geometric coordinates and the corresponding value of the interaction energy. For this version, **LRF** supports as internal coordinates: *Autosurf* internal coordinates, *Euler* angles employing ZXZ and ZYZ convention, and *Spherical* (only available for nonlinear molecules atom systems).

Format	Internal Coordinates
<i>Autosurf</i>	$\{R, \alpha, \cos(\beta_1), \cos(\beta_2), \gamma_1, \gamma_2\}$
<i>Euler Angles (ZXZ) or (ZYZ)</i>	$\{R, \alpha, \beta_1, \beta_2, \gamma_1, \gamma_2\}$
<i>Spherical</i>	$\{R, \theta, \varphi\}$

Table 2

The column number and the unit of each internal coordinate can be specified after entering the input file location in block 3. ???

	Units
R	Angstrom, Bohr Radius
α	Radians, Degrees
β_1 or θ	Cosine, Radians, Degrees
β_2 or θ	Cosine, Radians, Degrees
γ_1 or φ	Radians, Degrees
γ_2 or φ	Radians, Degrees
Energy	Kilocalorie Per Mole ($kcal/mol$), Wavenumbers (cm^{-1}), Hartree (<i>hartree</i>), ElectronVolt (eV), KiloJoule Per Mole (kJ/mol), Kelvin (K), Gigahertz (GHz)

Table 3

3.1.3 Orientation of the symmetry elements

The generation of the right analytical expression describing the long-range interactions for any particular system, relies on the determination of the right molecular symmetry for each of the fragments of the system. Therefore, it's crucial to pay attention when defining the proper orientation of each molecule's symmetry elements — *i.e.*, the orientation of the initial non-rotated fragments must match the convention used by LRF for any particular symmetry element to be used. For example: (i) for C_s molecules, LRF assumes the xy plane as the plane of symmetry; (ii) for C_n molecules, z should be the the rotation symmetry axis; (iii) chiral partners must be converted into one another via a reflection across the xz plane, leading to the following charge distribution: $\rho^B(x, y, z) = \rho^A(x, -y, z)$; and so on. The complete set of conventions for symmetry elements of the non-rotated molecules can be seen in Table 4. Note that not every possible symmetry element is listed in the table, but only the minimal number of elements needed to guarantee that the right symmetry has been chosen for each molecule, and its orientation matches the LRF assumed convention.

3.1.4 The input data file

structure and formatting...

Extensions allowed:

Text Files: .txt, .dat, and .csv

Spreadsheet files: .xls, .xlsb, .xlsm, .xlsx, .xltm, .xltx, and .ods

Group	C_n^z	C_2^x	σ_{xy}	σ_{xz}	i	S_n^z	C_3^*	C_5^*
$C_{\infty v}$	✓	—	—	—	—	—	—	—
$D_{\infty h}$	✓	—	✓	—	—	—	—	—
C_1	—	—	—	—	—	—	—	—
C_s	—	—	✓	—	—	—	—	—
C_i	—	—	—	—	✓	—	—	—
C_n	✓	—	—	—	—	—	—	—
C_{nh}	✓	—	✓	—	—	—	—	—
C_{nv}	✓	—	—	✓	—	—	—	—
D_n	✓	✓	—	—	—	—	—	—
D_{nh}	✓	✓	✓	✓	—	—	—	—
D_{nd}	✓	✓	—	—	—	✓	—	—
S_{2n}	—	—	—	—	—	✓	—	—
T	✓	✓	—	—	—	—	✓	—
T_h	✓	✓	✓	✓	✓	—	✓	—
T_d	✓	✓	—	—	—	✓	✓	—
O	✓	✓	—	—	—	—	✓	—
O_h	✓	✓	✓	✓	✓	—	✓	—
I	✓	✓	—	—	—	—	✓	✓
I_h	✓	✓	✓	✓	✓	—	✓	✓

Table 4: Table of symmetry elements corresponding to each symmetry point group. C_n^z : C_n rotation axis around z , C_2^x : C_2 rotation axis around x (notice all C_4^x also imply C_2^x), σ_{xy} : Symmetry Plane in xy plane, σ_{xz} : Symmetry Plane in xz plane, i : Inversion center at the origin, S_n^z : S_n roto-reflection axis around z , C_3^* : C_3 rotation axis around $x + y + z = 1$ (Tetrahedral, Octahedral and Icosahedral symmetries only), C_5^* : C_5 rotation axis around $\frac{1}{2}(1 + \sqrt{5})x + y = 0$ (Icosahedral symmetry only).

3.2 Graphical User Interface

LRF main window is divided into four tabs: **Setup**, **Expansion**, **Dashboard**, and **Manual**, as can be seen on the left side of Fig. 3. In the rest of this section, we will navigate the first three tabs, describing available options and explaining all relevant aspects and settings. For convenience, the **Manual** tab contains an embedded copy of this PDF document for future reference.

3.2.1 Setup tab

As shown in Fig. 3, the **Setup** tab allows the user to define the molecular system and enter the input data (Steps 1 and 2 in the Workflow, see Fig. 2). The tab is divided into three numerated sections.

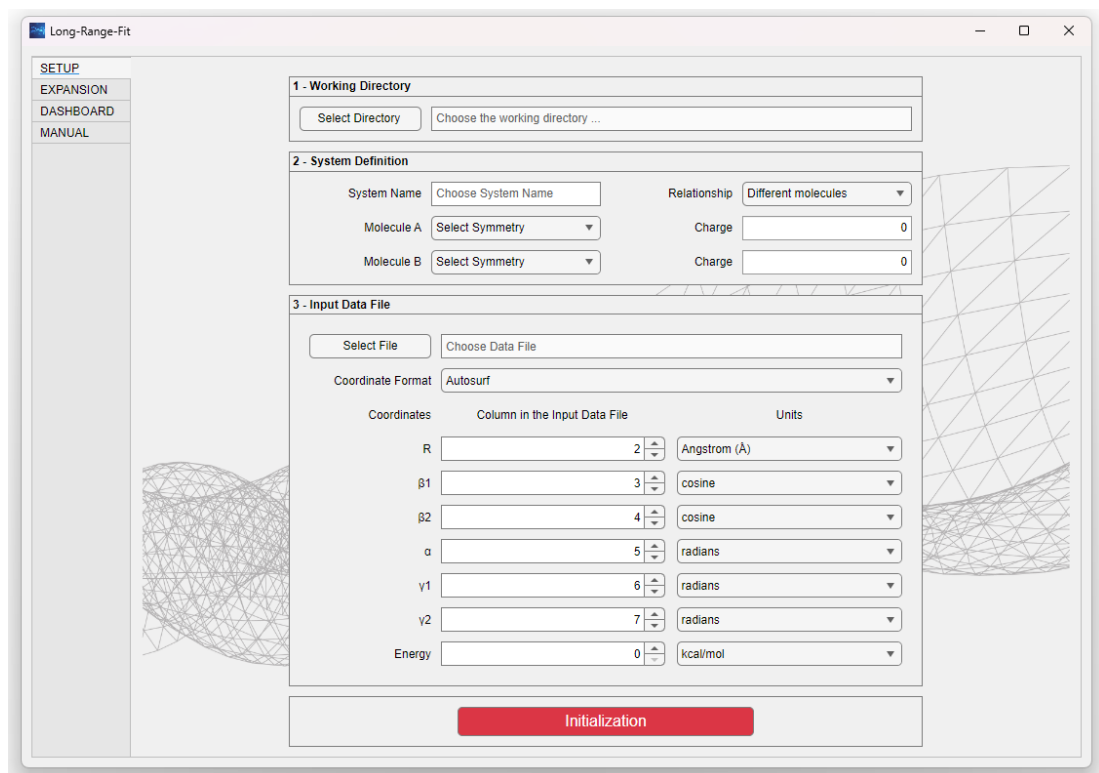


Figure 3: LRF *Setup* tab.

First, the user (1-) choose the “*Working Directory*” — where all files exported by LRF will be saved. After that, the user (2-) select a name/label for the system, and also the symmetry and net charge of each separate fragment. The list of available symmetries includes all point groups listed in Table 4, with a rotation axis (C_n and S_n^z) up to $n = 8$. Also, if both molecules have the same symmetry, the “*Relationship*” option becomes available, and the user can specify the particular cases where the molecules are identical or chiral partners.

Next, (3-) a “*Input Data File*” containing the *ab initio* data to be used in the fit should be selected — the input file does not necessary have to be located in the “*Working Directory*” specified before. See Section XXX for relevant information concerning the structure and formatting of the input file. After that, the “*Input Coordinates Format and Units*” should be chosen. The complete list of available possibilities is presented in Tables 2 and 3 (see Section ??).

Finally, by clicking the “Initialization” button the program will import the *ab initio* data and generate analytical functions to describe the long-range interactions of the system. Depending on the size of the input data and the computer hardware, this process could take from seconds up to a couple of minutes.

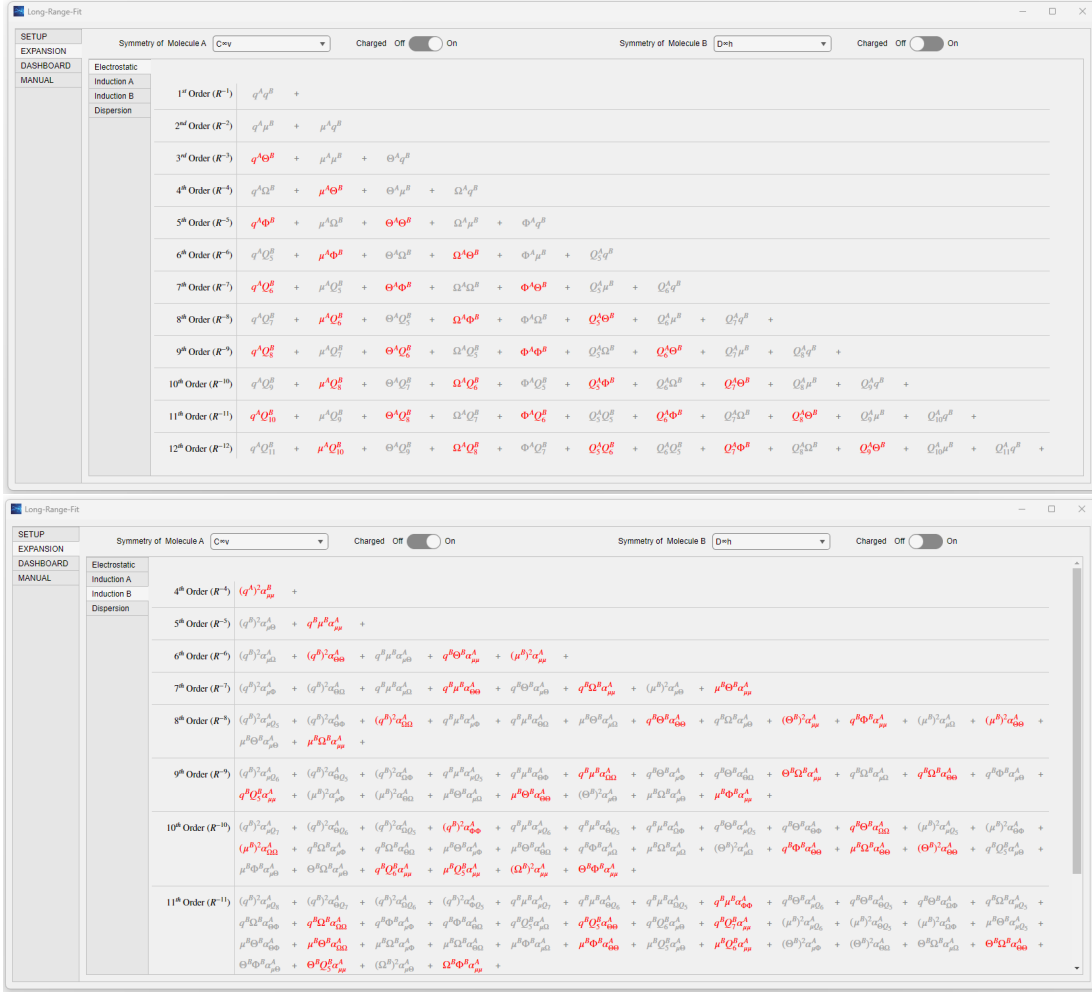


Figure 4: LRF *Expansion* tab. Electrostatic interaction expansion (upper panel), Induction expansion of A over B (center panel) for a non-centrosymmetric, linear, charged molecule A that interacts with a centrosymmetric, linear, neutral molecule B. See the text for details.

3.2.2 Expansion tab

For the system specified in the previous tab, the complete multipole expansion describing long-range interactions is represented in the *Expansion* tab, as shown in Fig. 4. Here, all expansion terms up to the 15th order (R^n with $n \leq 15$) are included, separated by interactions: Electrostatic (upper panel), Induction (central panel), and Dispersion (lower panel).

All coefficients are presented in their spherical form, with a superindex A or B denoting the associated molecule. In the Electrostatic expansion, the 2^n -multipole moments are symbolized by Q_n , with the exception of the net charge (q), the dipole (μ), the quadrupole (Θ), the octupole (Ω), and the hexadecapole (Φ); which are represented by their specific symbols. The polarizability coefficient for the interaction between 2^n - and 2^m -multipole moments is denoted by $\alpha_{Q_n Q_m}$ (for example, dipole-dipole polarizability would be $\alpha_{\mu\mu}$). The dispersion coefficients are given as $D_{Q_n Q_m - Q_p Q_q}$, where the first two and last two subscripts come from the imaginary

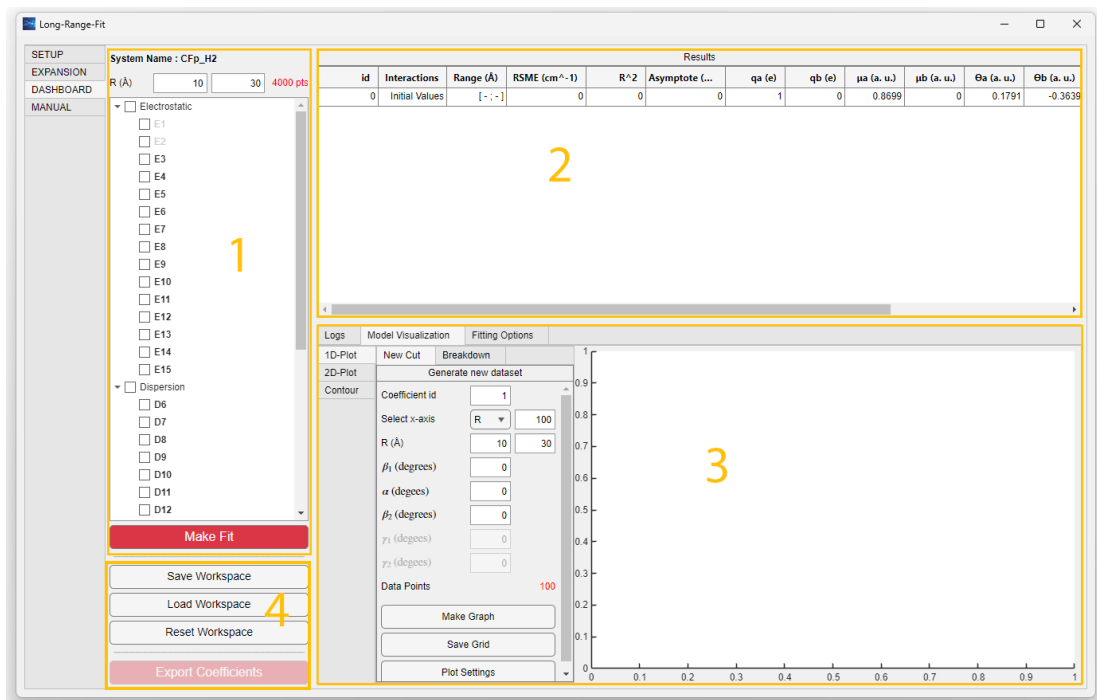


Figure 5: LRF *Dashboard* tab.

polarizability of molecules A and B, respectively (e.g., the dipole-dipole – dipole-dipole dispersion coefficient is denoted by $D_{\mu\mu-\mu\mu}$).

The notation used for each term denote the condensed representation of their elements' summation — e.g., the dipole-quadrupole term $\mu^A\Theta^B$, a order term of the Electrostatic interaction expansion (Fig. 4, upper panel), contains a total of 15 terms, combining the three dipole components and the five quadrupole components of each molecule:

$$\mu^A\Theta^B = \sum_{\kappa,\kappa'} \mu_{1\kappa}^A \cdot \Theta_{2\kappa'}^B \cdot T_{1\kappa,2\kappa'} , \quad (1)$$

where $T_{1\kappa,2\kappa'}$ represents ... angular dependence ... Defined by the symmetrical properties of the fragments, the terms marked in red in Fig. 4 have at least one nonzero element in this sum, while those in gray are constrained to zero.

Initially — after the initialization of the program — the multipole expansion represented in the *Expansion* tab is set to the symmetry of the system (specified by the symmetry of each fragment, as defined in the *Setup* tab); but for convenience, any other available symmetry could be also specified in the left side of the tab, allowing to explore different possibilities without the need to restart the app or load a new set of *ab initio* data.

The expansion representation provides a concise overview, facilitating the identification of dominant terms and non-zero orders on the expansion.

3.2.3 *Dashboard* tab

Fig. 5 shows the *Dashboard* tab. As underlined and highlighted in yellow in the figure, to facilitate the description of each set of options we have divided the tab into four sections.

Figure 6: Fit configuration (*Dashboard*, Section 1). In this example, for the "CFp_H2" system, the asymptote, as well as the 3rd and 4th order of the electrostatic interaction, will be fitted using 4000 data points distributed from 10 to 30 Å. Interaction orders that have not any non-zero terms appears in light gray like the 1st and 2nd orders. See the text.

1- System Name : CFp_H2

2- R (Å) 10 30 4000 pts

3- ☒ Electrostatic

- ☐ E1
- ☐ E2
- ☒ E3
- ☒ E4
- ☐ E5
- ☐ E6
- ☐ E7
- ☐ E8
- ☐ E9
- ☐ E10
- ☐ E11
- ☐ E12
- ☐ E13
- ☐ E14
- ☐ E15

☐ Dispersion

- ☐ D6
- ☐ D7
- ☐ D8
- ☐ D9
- ☐ D10
- ☐ D11
- ☐ D12

4- Make Fit

Section 1

In the first section the final parameters are set and the fit executed. In Fig. 6, the different components of the section are enumerated:

1. System name, as defined by the user in the *Setup* tab. This label will be included in the output files produced by LRF. It is recommended not to use special characters nor white spaces.
2. Filter for the R coordinate, so only the data points in the specified R -interval are considered in the fit. The values of R_{min} (left) and R_{max} (right) must be given in Angströms. Initially, these values are automatically set to include all the data in the input file. In red, the number of data points to be used in the fit after applying the R -filter is shown.

3. Once the symmetry of the system is defined, LRF generates the corresponding analytical expression for each long-range interaction (Electrostatic, Induction, and Dispersion) as shown in the **Expansion** tab. Each expression is an infinite sum of terms that can be grouped by the powers of R^{-n} (i.e., by the "order" of the term). Here, from the available terms — that go up to $n = 15$ for each interaction — the user must select specifically which ones should be included in the fit.
4. After selecting the terms to be used, the fit can be performed by clicking the red button, "Make Fit".

Section 2

Each time a new fit is completed, a new line with relevant information is added to the "Results" table in this section, as shown in Fig. 7. Here a description of the table elements by column:

- **id:** provides an identification consecutive number of each fit.
- **Interactions:** indicate the orders included on each fit (separated by "-") for each interaction (separated by "/"), identified by a preceding capital letter ("E" = electrostatic, "I" = induction, "D" = dispersion).
- **Range:** minimum and maximum values of R (in Å) considered for the fit.
- **RMSE:** global Root Mean Square Error of the fit (in cm^{-1}).
- **R²:** Coefficient of determination. correlation???
- **Asymptote:** Asymptotic energy (in cm^{-1}).
- **Multipoles Coefficients:** fitted coefficient values (Net charge, Dipole, Quadrupole, Octupole, and Hexadecapole) for each molecule, in a.u. Small variation of the lower-order multipoles between consecutive fits constitutes a good indicator of convergence.

Section 3

The third section of the **Dashboard** tab is devoted to managing all elements related to the fit and its outcomes, by providing different tools to assist and/or manipulate the fit, and analyze their quality. The section is divided into three panels: **Logs**, **Model Visualization**, and **Fitting Options**.

This panel displays how the number of *ab initio* points and the basic fit statistics are distributed in different intervals of R . The first table shows the stats in the whole fitting space while the second table show the same analysis across evenly divided intervals. The number of intervals is controlled by the parameter “Splitting”. The first column shows the R limit of the region, the second column shows the number of points, followed by RMSE (Root Mean Square Error) in wave numbers, R^2 and RMAE (Relative Mean Absolute Error) calculated as:

$$RMAE = 100 \times \frac{1}{N} \sum_{i=1}^N \left| \frac{E_{fit}(i) - E_{data}(i)}{E_{data}(i) - E_{Asymptote}} \right|$$

(see figure 8.)

Histogram of Residuals panel

In an ideal scenario, the residuals of a fit should exhibit a normal distribution with a mean of zero in any coordinate subspace. The **Histogram of Residuals** presents the distribution of the residuals and its corresponding fitted normal distribution, including its mean and standard deviation. This plot will be shown for any subspace defined by the user.

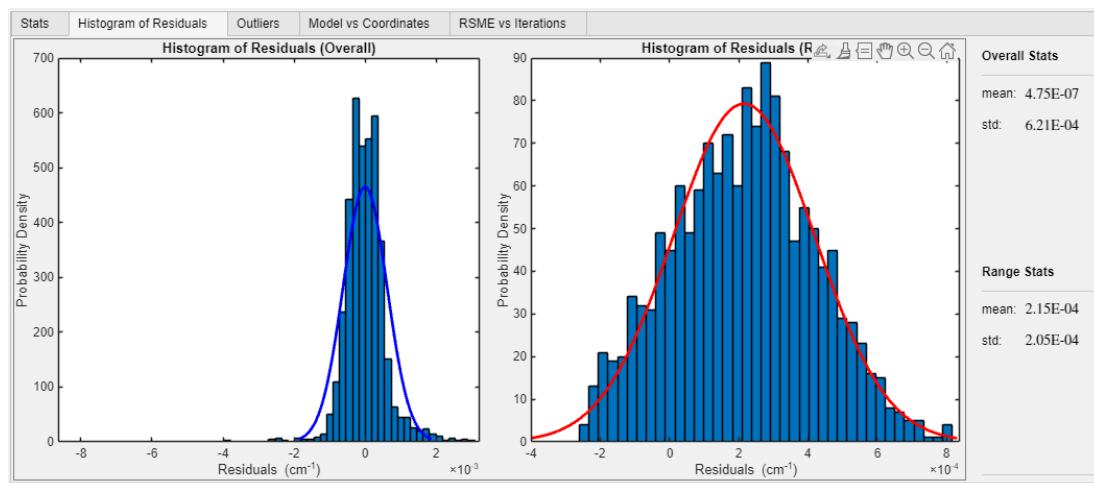


Figure 9: Histogram of Residuals panel. In this example, the logs correspond to the 7th fit performed. On the left, it is shown the residual’s histogram corresponding to the whole fitting space ($R \in [10\text{\AA}; 30\text{\AA}]$). On the right, the residual’s histogram corresponding to the coordinate subspace defined ($R \in [20\text{\AA}; 30\text{\AA}]$). As the model improves with larger R values, the residual distribution is expected to align more closely with a normal distribution in the ultra long-range.

Outliers panel

This panel shows on the left the distributions of residuals (Raw Residuals, *res*) with respect to any coordinate. The panel on the right side allows the user to change

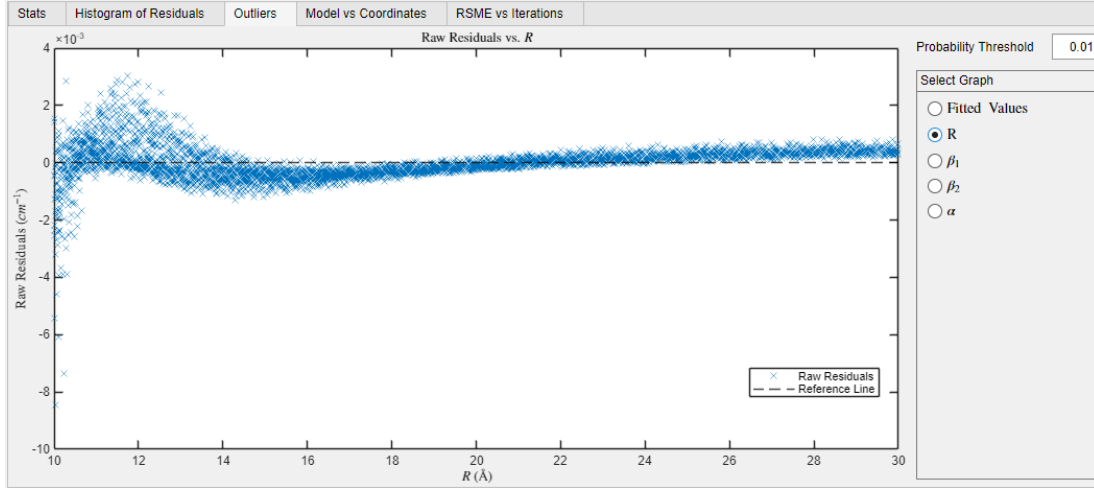


Figure 10: Residuals Distribution by R In this example, the logs correspond to the 9th fit performed. On the left, the **Outliers** panel shows a typical distribution of residuals by R.

the x-axis (the coordinate). Typically, the distribution by R is characterized by a tendency to larger errors for smaller values of R that decrease monotonically to zero for larger values of R (see Figure 10), while the angular distribution of the residuals should be more homogeneous.

Data sets may occasionally contain erroneous data points due to inadequate convergence in quantum calculations, but identifying these outliers is not always straightforward. Larger absolute errors don't necessarily indicate a poor data point, as they are anticipated at lower R values. The relative error is also an not absolute indicator since these tend to be larger in the points close to the asymptote. Taking these facts into account, we have constructed the weight function:

$$w_i = \sqrt{\left[\frac{R_i}{R_{max}}\right]^2 + \left[1 - \left|\frac{E_i^{fit}}{E_{max}^{fit}}\right|\right]^2}$$

, which highlights all points close to the asymptote with larger absolute errors and larger R. Also notice that $R_{min} \leq R_i \leq R_{max}$ and $0 \leq \left|\frac{E_i^{fit}}{E_{max}^{fit}}\right| \leq 1$, so $\frac{R_{min}}{R_{max}} \leq w_i \leq 1 + \frac{R_{min}}{R_{max}}$.

The Weighted Residuals ($wres_i = w_i \cdot res_i$) are assumed to be normal distributed, so the worst points are located in the tails of the distribution function. The parameter “Probability Threshold” P_{th} is located on the right side of the **Outliers** panel is employed to classify the outliers, so the data point i is marked as a potential outlier if

$$P(wres_i) \leq P_{th} \quad \text{or} \quad P(wres_i) \geq 1 - P_{th}$$

The plot “Weighted Residuals vs. Energy Fitted” shows the Weighted Residuals distribution. Outliers are generally far from the main cluster of points (see Figures 10 and 12), and excluding them from the data is expected to significantly affect the fit

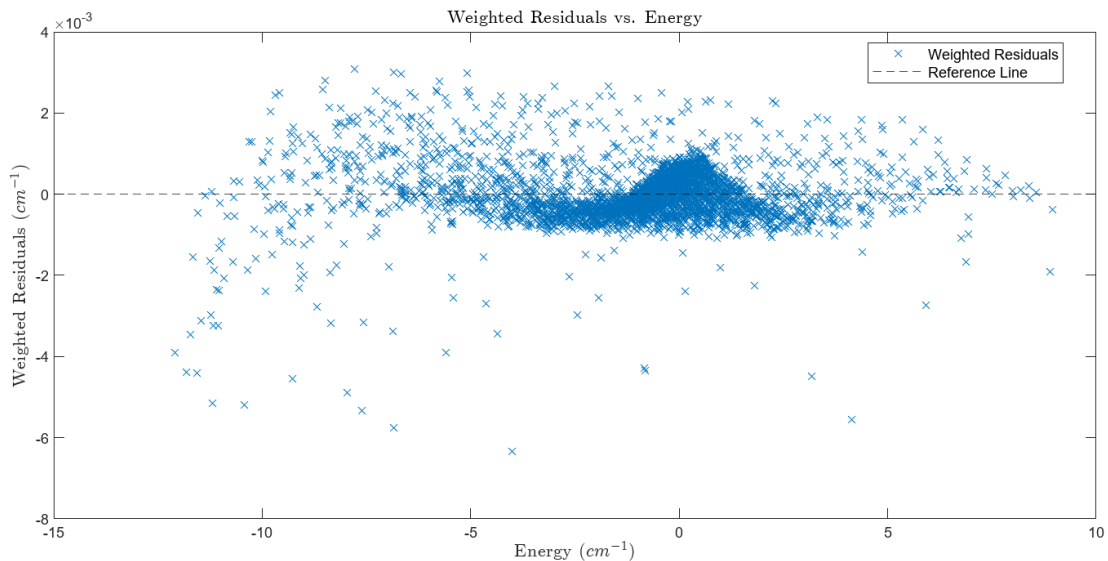


Figure 11: Weighted residuals distribution for a fit including all non-zero orders up to 8th order for the system $CF^+ - H_2$. It does not show any potential outlier since all the points are close to each other

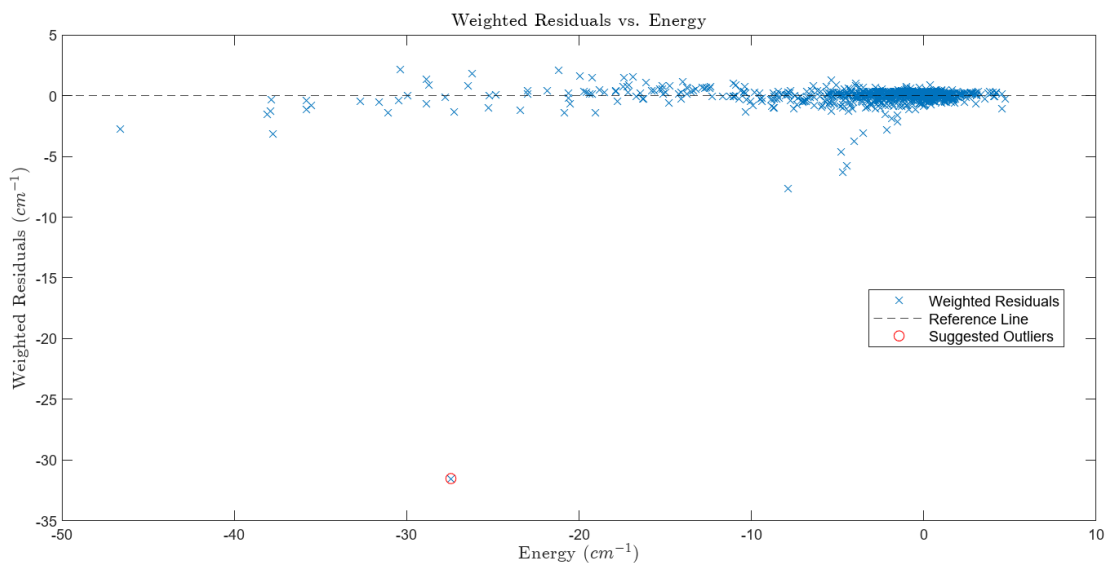


Figure 12: Weighted residuals distribution for a fit including all non-zero orders up to 8th order for the system $C_{10}^- - H_2$. There is at least one point separated from the cluster.

statistics. Suggested outliers for the defined “Probability Threshold” parameter will be marked with a red circle. By hovering over specific data points in the graphs, users can view the precise coordinate values, input energy, raw and weighted residuals, and outlier classification.

Model vs User Data panel

This tab enables users to verify input data sampling, assess the model, and compare it to user input energy for any coordinate. On the right side, the user can change the coordinate (Figure 13).

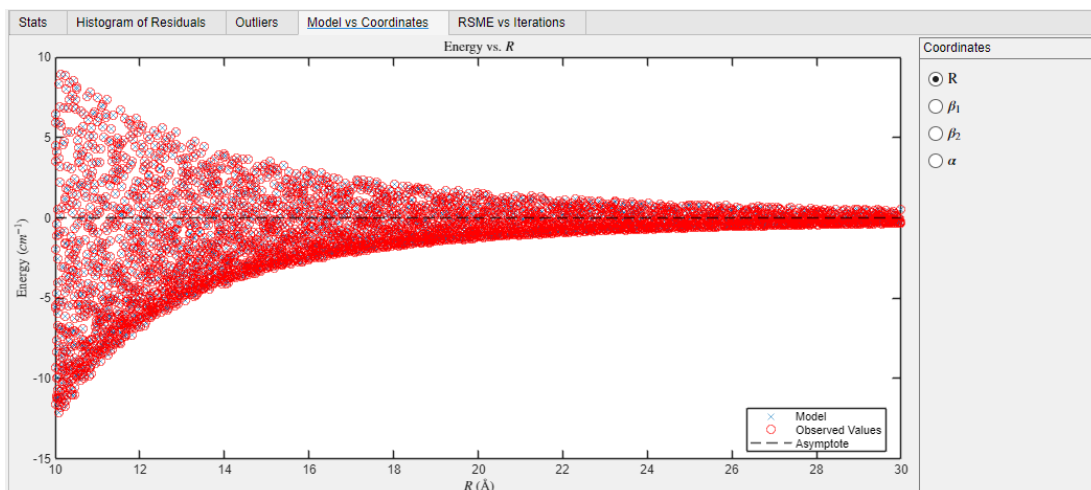


Figure 13: Model vs User Data panel. On the left, the Energies calculated by the model and the user input data vs R . On the right side, the coordinate list where the user can change the x-axis of the graph.

RSME vs Iterations panel

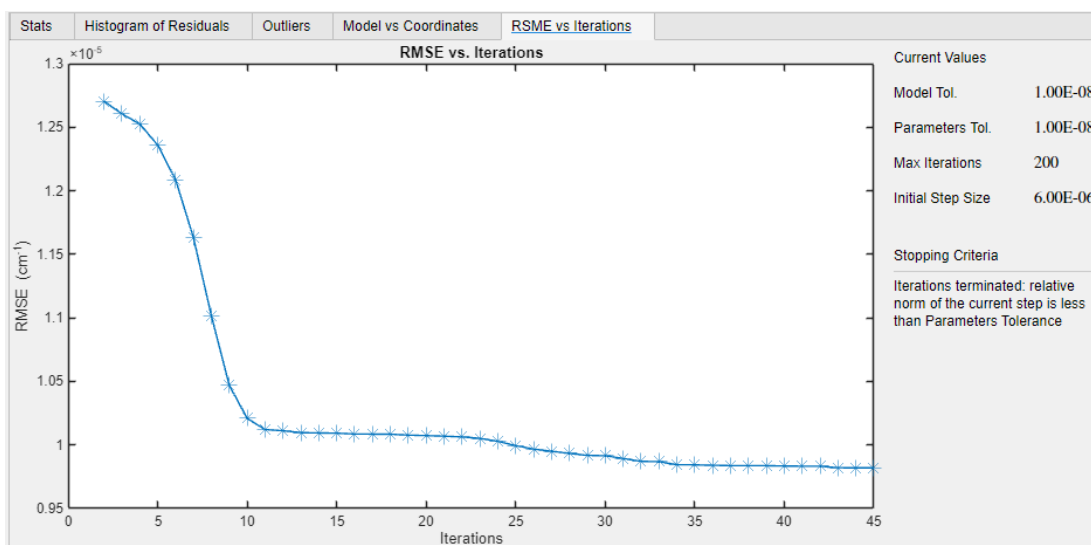


Figure 14: RSME vs Iterations panel. In this example, the logs correspond to the 7th fit performed. On the left, it is shown the evolution of RMSE during the fitting. On the right, the current parameter of the fit engine.

A good indicator of fit convergence is having a flat tail in the *RSME vs. Iterations* plot, which means that the global error does not improve, so the fit engine is close to a local minimum. An early stop could be due to three reasons: the maximum number of iterations is too small, or the goal error of either individual parameters of the model or the global model has been reached. For more details, see the Stopping Criteria on the right. The user can also modify the stop parameter on the **Fitting Options** tab.

3.2.5 Model Visualization

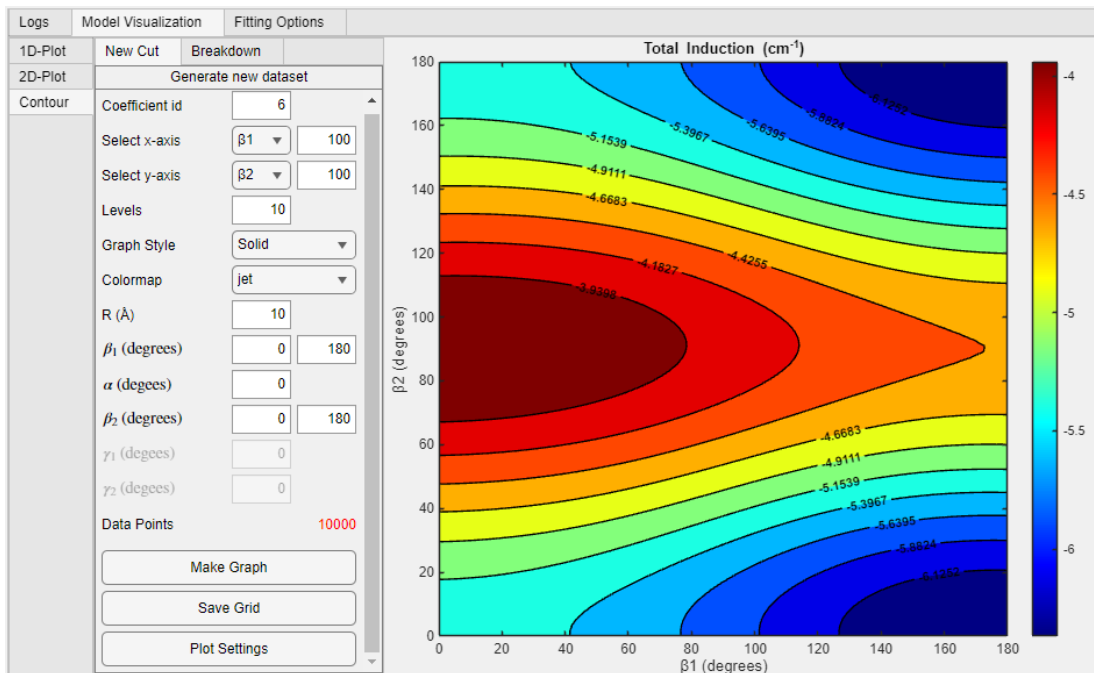


Figure 15: Contour tab. On the left, the user control panel showing the “New Cut” tab. On the right, the canvas is rendering the Total Induction Interaction corresponding to the fit **id** = **9** vs β_1 and β_2 , at $R = 10$ and $\alpha = 0$

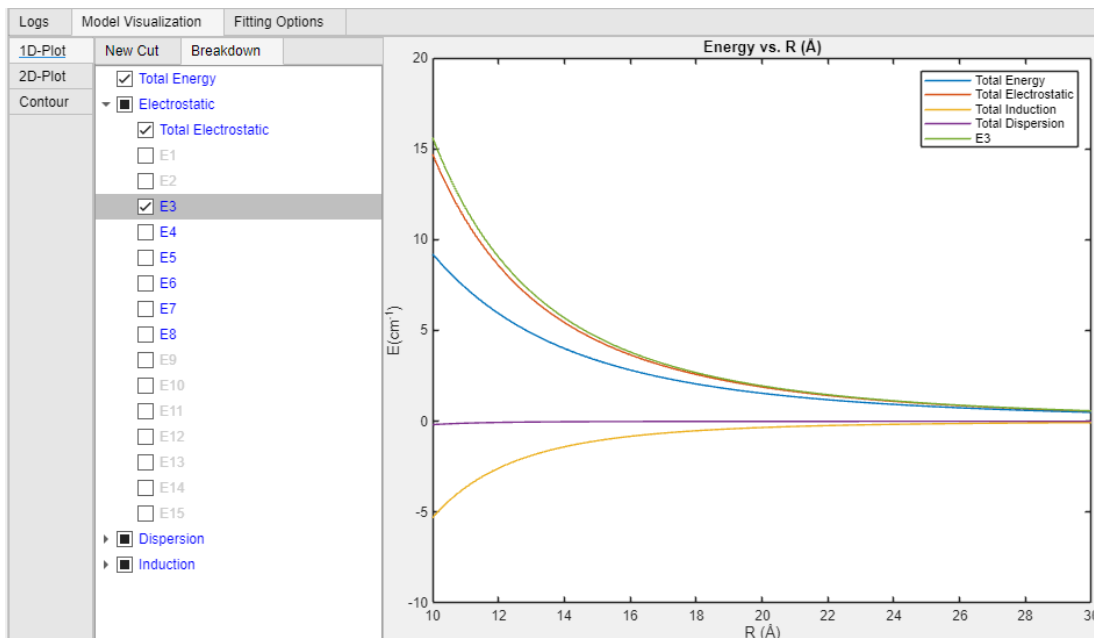


Figure 16: It shows a 1D-cut corresponding to the fit **id** = **9** vs R at $\beta_1 = 0$ $\beta_2 = 0$ and $\alpha = 0$ (the linear molecules are contained in the same plane, are parallel between them and perpendicular to the line that joins the center of masses). On the Breakdown tab, the user can also see along the total energy, the leading term $q^A\Theta^B$ (electrostatic third order, E_3), the total electrostatic, induction and dispersion interactions.

Provides functionalities for plotting any 1D (Figure 16) or 2D (Figures 17 and 15) and the total energy, the total interaction energies involved, and the specific interaction order of them. The internal coordinates of the system in the graph are the distance between the center of the masses R measured in Å, and the Euler angles defined under the ZXZ convention in degrees. For 2D-cuts, the user can choose between a surface and contour plots. Each plot tab is divided into two sections: the user control panel and the graph canvas. The user control panel has two tabs: "New Cut" (Figure 18a), and "Breakdown" (Figure 18b). The canvas section will dynamically change unless the grid of the cut changes (coordinates on the grid or numbers of points in each coordinate). On right upper corner of the canvas and also in any figure displayed in the application, there is a graph toolbar with seven options: "export", "brush", "datacursor", "rotate", "pan", "zoomin", "zoomout", and "restoreview" (see Figure 17). They allow the user to interact dynamically with the graph and copy or save the modified graph (see the description in Table 5).

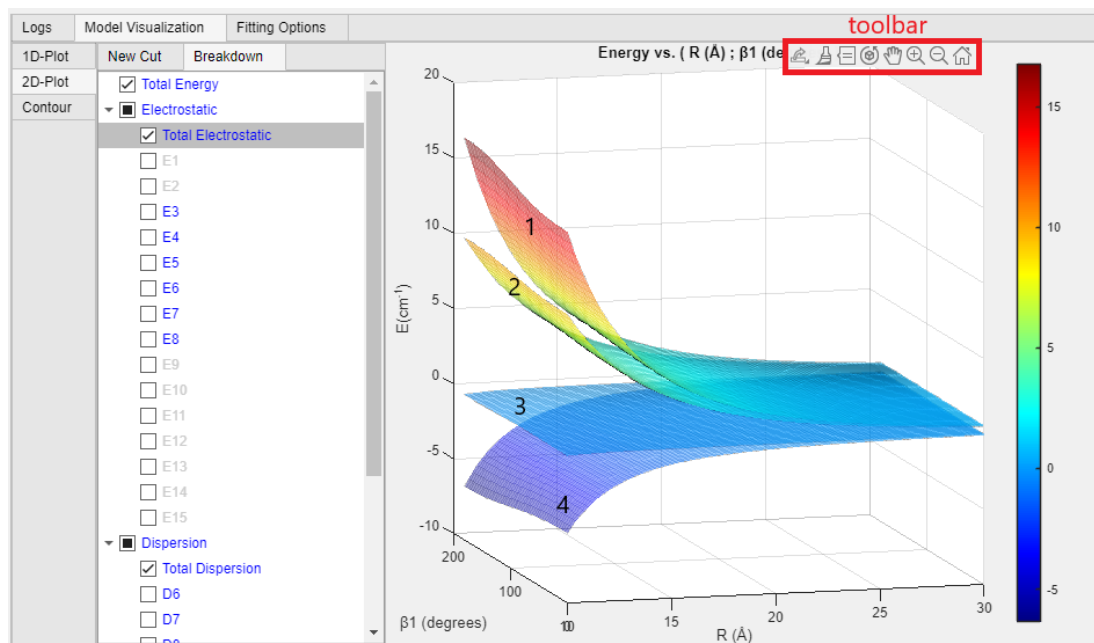


Figure 17: It shows a 2D-cut corresponding to the fit **id = 9** vs R and β_1 , fixing $\beta_2 = 0$ and $\alpha = 0$. Employing the Breakdown tab, the user can also see along the total energy (2), the total electrostatic (1), induction (3) and dispersion (4). Hovering the graph, the toolbar will show up in the right upper corner.

The "New Cut" feature offers options for selecting the model ID, choosing axis variables for the plot, defining the plotted subspace, and adjusting various plot settings. Each axis variable has three defining elements: minimum boundary, maximum boundary, and grid partition count, while non-axis variables maintain fixed values. The Data Points label shows the total grid points. Variables in light gray are inactive for the system. For instance, 2D systems (like a linear molecule with an atom) have two active variables, R and β_1 . After configuring the system, users can either render the plot using the "Make Graph" button or save the geometric grid












Symbol	Option	Description
	export	 Save the content as a image or PDF.  Copy the content as an image.  Copy the content as a vector graphic.
	brush	Enables you to highlight data points in a chart with a color of your choosing. You can select individual or multiple points with a selection rectangle and they will remain until removed.
	datacursor	Interactively creating and editing data tips (small text boxes that display information about individual data points).
	rotate	Interactively rotating the 3-D view of the axes.
	pan	Interactively panning the view of the axes.
	zoomin	Interactively zoom in the view of the axes.
	zoomout	Interactively zoom out the view of the axes.
	restoreview	Restore original view of axes or tiled chart layout.

Table 5: Description of the toolbar options

and model-derived energy in a ".txt" file. The panel also provides graph style options via "Plot Setting" button which display all available options given by Matlab to customize the canvas plot (Figure 18c).

The "Breakdown" panel (Figure 18b) features a checkbox tree that includes the Total Energy (sum of all interactions), along with the net electrostatic, induction, and dispersion energies, as well as their respective components. The components and net energies that are active in the model are highlighted in blue. This checkbox list enables the user to plot multiple interactions at once, offering a unique perspective on the contribution of each component (see Figures 16 and 17).

3.2.6 Fitting Options

Next, as shown in Figs.19 and 20, the **Fitting Options** panel contains different tuning options that allow the user to control both the coefficients of the model and the fitting engine.

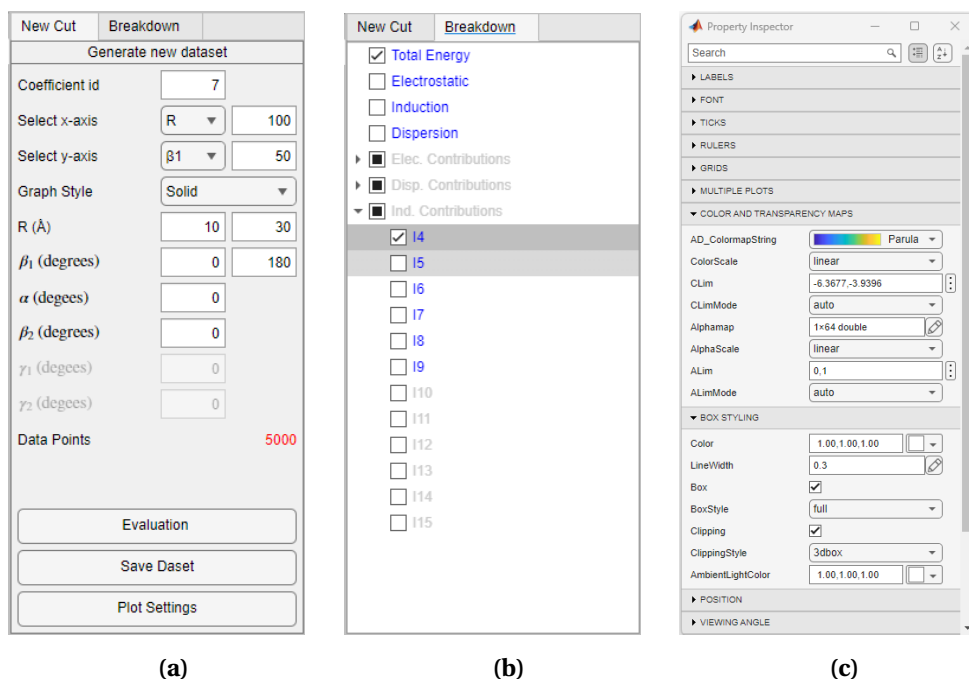


Figure 18: (a) “Generate New Cut”. (b): “Interactions Breakdown”. (c) “Plot Setting”

Coefficients tab

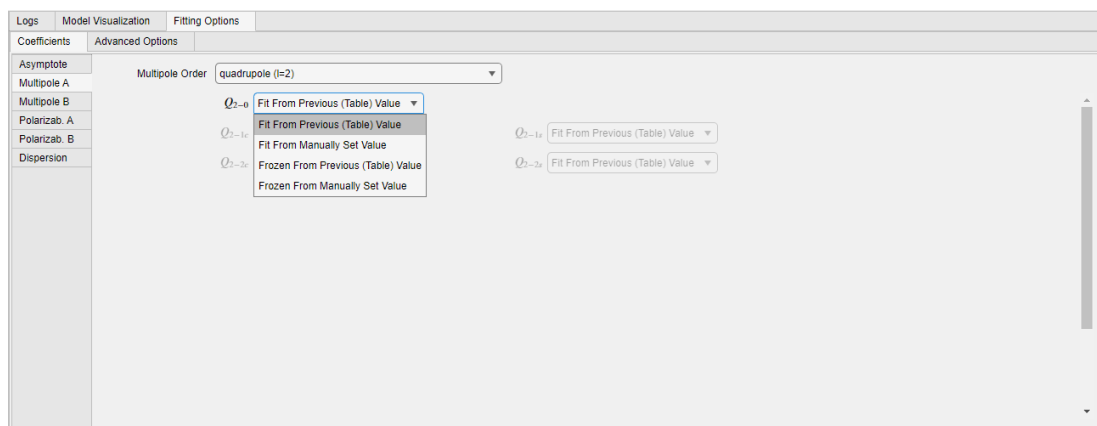


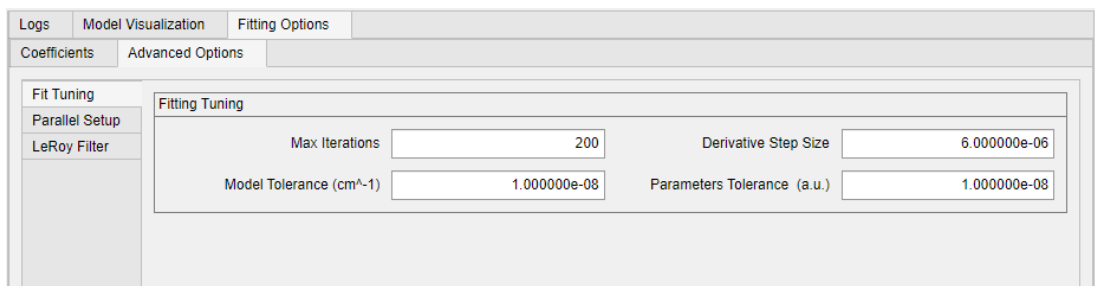
Figure 19: It presents the multipole components options for CF^+ . The molecule has a non-zero quadrupole along the z-axis, hence displaying the fitting options drop-down menu. Since other components are zero, their fitting options will become inaccessible.

The Coefficients section lists all available coefficients, separating them into six sections (Asymptote, Multipoles of A, Multipoles of B, Polarizabilities of A, Polarizabilities of B, and Dispersion). Each tab contains a drop-down menu that shows only non-zero orders (for example, centrosymmetric molecules like H_2 do not have odd multipole moments, so they will not appear in the multipole section). After selecting the multipole, polarizability, or dispersion order, the panel shows their spherical components. The zero components are inactive and colored with a light gray, while the non-zero ones are colored in black and show four options:

- *Fit From Previous (Table) Value:* The coefficient **will be fitted** using as an initial guess the values of the prior fit. If no prior fit exists, **LRF** includes a built-in algorithm to set a reasonable initial estimate for the first three nonzero orders of multipoles and a random value between -1 and 1 for the rest.
- *Fit From Manually Set Value:* The coefficient **will be fitted** using as an initial guess a value set by the user. It will be particularly useful when some coefficients are known so it can help the fitting engine to converge quicker and closer to the real values
- *Frozen From Previous (Table) Value:* The coefficient **will NOT be fitted** and the value will be taken from the prior fit.
- *Frozen From Manually Set Value:* The coefficient **will NOT be fitted** and the value set by the user.

Advanced Options tab

◆ Fit Tuning



Fitting Tuning	
Max Iterations	200
Derivative Step Size	6.000000e-06
Model Tolerance (cm ⁻¹)	1.000000e-08
Parameters Tolerance (a.u.)	1.000000e-08

Figure 20: Fit Tuning

- *Max. Iterations:* The allowable maximum number of iterations, a positive integer.,
- *Derivative Step Size:* The relative variance employed in finite difference calculations of derivatives. It can be a positive scalar or a vector of positive scalars matching the size of the parameter vector estimated by the Statistics and Machine Learning Toolbox function via the options structure.
- *Termination Tolerance:* The termination tolerance for the objective function's value, a positive scalar.
- *Termination for the parameters:* The termination tolerance pertaining to the parameters, a positive scalar.

◆ Parallel Setup

Certain internal operations can be executed in parallel, accelerating the fitting process. However, it has to be used with caution since not always parallelization improves the sequential execution time since parallelization requires that each node has a copy of the data. In our experience, it is relevant when fitting higher orders of a system composed of two nonlinear molecules that have low symmetry. The exact parallel performance depends on the machine hardware, the molecular system, the size of the user data, and the interaction orders involved in the fitting. **LRF** have implemented a build-in algorithm that estimates the performance of both “Parallel” and “Non-Parallel” approaches (see Figure ??)

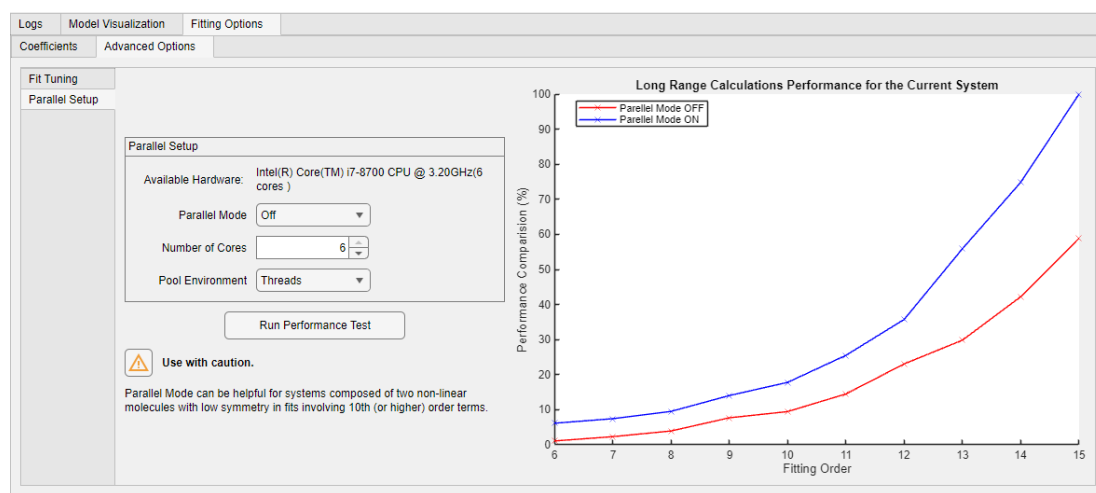


Figure 21: Parallel Setup Tab for $CF^+ - H_2$ in a test machine for a dataset with 1000 data-points. Notice that the parallel approach is always worse than the non-parallel approach

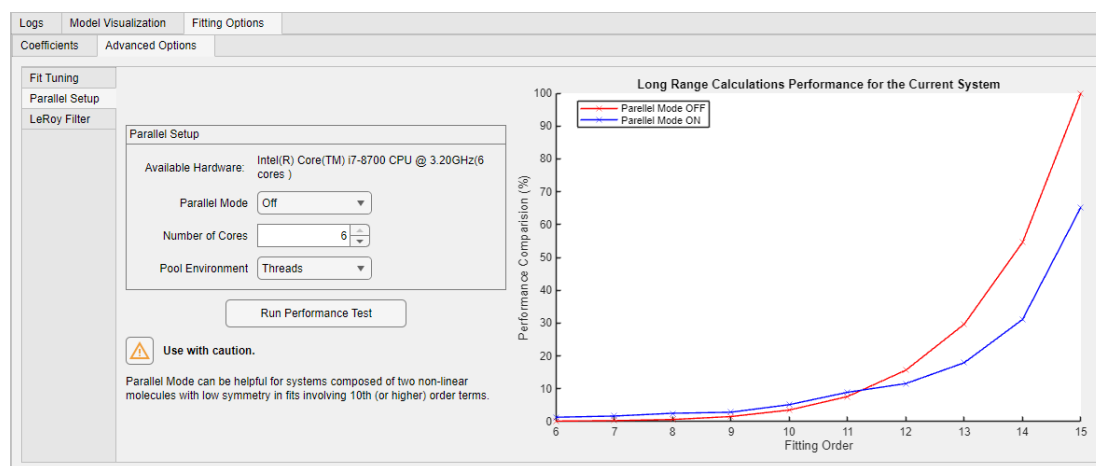


Figure 22: Parallel Setup Tab for two molecules with C_1 symmetry in a test machine for a dataset with 1000 datapoints. Notice that the non-parallel approach is slightly better up to 10th order but then worse for higher orders. It is important to highlight that in the higher order the fit takes much more time than a the fit including lower order only, so a 20% improvement could means a lot of time.

- *Available Hardware:* List all CPU hardware available and their number of cores.

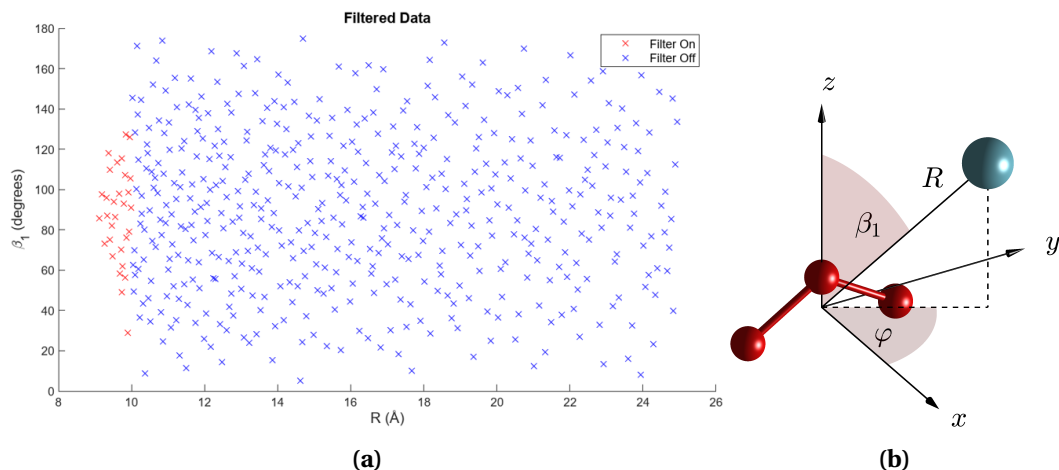


Figure 23: Caption place holder

- *Parallel Mode:* .
- *Pool Environment:* The MATLAB Parallel Computing Toolbox facilitates the execution of parallel code in two separate environments known as thread-based and process-based. When running operations on a local machine, it is advisable to employ the "Threads" environment. In contrast, when working within a cluster configuration composed of multiple computational nodes, the recommended setup to utilize is the "Process" environment.
- *Number of Cores:* Size of the parallel pool, specified as a positive integer greater than or equal to 2. The default number of cores is the maximum number of physical cores available, up to 32 cores (limit of core usability imposed by MATLAB).

◆ LeRoy Filter

The long-range starts when there is no overlapping between the electronic clouds associated with each molecule, however; those clouds actually cover the whole space, so there is always an overlapping. Since they decay exponentially, at a sufficient fragments separation, the overlapping is minimal, so typically the long-range is established at some fixed value $R = R_0$. Unless both molecules are atoms, the long range starts at different values of R , depending on their relative orientation. For example, for the system $O_3 - Ar$ (see figure 23 (b)), the Argon atom will enter to the short range region quicker if it is approaching the ozone atom along the z -axis ($\beta_1 = 0$) than if it is approaching along the x -axis. However, consider the long range as isotropic by fixing R_0 does have much impact in the long range data. For example, if we consider the overlapping at $R_0 = 10\text{\AA}$, then only 6% of the data-points has $R < 10\text{\AA}$ (red datapoints in figure 23 (a)) all close to $\beta_1 = 90$.

However, this approximation is not valid for systems where at least one of the molecules is very anisotropic such as $C_{10}H^{-}H_2$ (see figure 24 (b)). In this case, when the two linear molecules are parallel ($\beta_1 = 90$ and $\beta_2 = 90$) the long range could be as

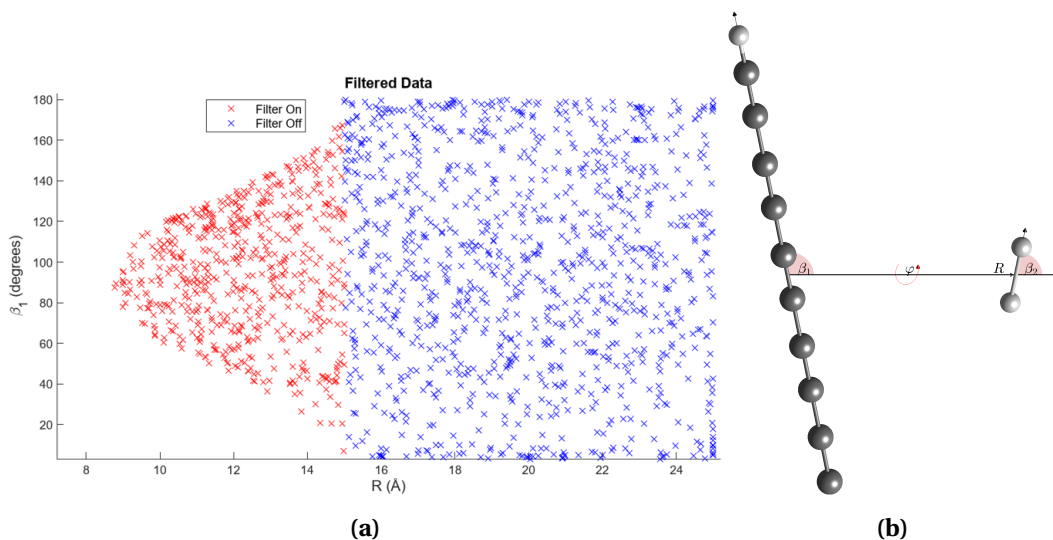


Figure 24: Caption place holder

close as 9Å, while at the same distance but in collinear shape, they are in the repulsive wall of the PES (for this configuration the long range starts around 15Å). Therefore, if we consider our long-range at 9Å for all directions, the long-range model will be forced to fit a big number of datapoints that belongs to the short range region. On the other hand, if the long range is set at 15Å for all directions, the data will miss around the 40% of the points in the long-range region. To solve this issue and inspired in the LeRoy radius, we implemented a specialized filter to recover all the points that belongs to the long-range at lower values of R.

The LeRoy radius, derived by Robert J. LeRoy, is defined as:

$$R_{LR} = 2[r_A^{2/3} + r_B^{2/3}]$$

where r_A and r_B denote the atomic radii of the two atoms. For $R \leq R_{LR}$, the inter-nuclear potential can be reasonably approximated by charge independent atomic distributions, so the their electronic molecular cloud overlapping is negligible. For two given atoms this formula can be seen as placing each atom in an sphere of radii $[r_A^{2/3} + r_B^{2/3}]$. This way long-range is determined whether or not the spheres are overlapping. For molecules with more than one atom, the expected value of R_{LR} actually depends on the angles that define the relative orientation between the fragment, since the cloud does not have spherical symmetry. However, we can reuse this idea and approximate the shape of the molecular cloud to the shape given by concentric spheres centered at each atom. Therefore, the long-range region will be given by the overlapping between the spheres. The radii of this sphere will be defined by the user, when he set the R_{min} on the left upper corner of the Dashboard Tab and enter the Cartesian coordinates of both molecules in the Leroy Filter Tab. The Leroy Filter Algorithm will find the minimum distance between every pair of atoms formed by one atom of molecule A and one atom of molecule B for all possi-

ble orientations at $R = R_{min}$, forming a matrix of our version of the Leroy radius for a given R_{min} ($LM = LeroyMatrix(R_{min})$). Now, for points in the data with $R < R_{min}$, the algorithm calculates the distance between all pairs of atoms for the given orientation ($D = distance(R, \Omega^A, \Omega^B)$) and compare it with $LM(R_{min})$. If for all i, j , $D_{i,j} \leq LM_{i,j}$, then this datapoint belongs to the long-range.

Section 4

Finally, the fourth section is dedicated to exporting the whole working environment to reproduce the exact same fit later and the data file with the fitted coefficients.

3.2.7 Workspace Portability

The LRF Workspace includes variables generated during fitting or employed to generate the complementary graphs and statistics. These variables are not retained post-closure unless manually saved with the “Save Workspace” button, creating a compressed “.mat” file in the Working Directory. The saved LRF environment can be loaded with the "Load Workspace" button. On the other hand, "Reset Workspace" enables the user to initiate a new fit within the same session without closing the LRF windows.

3.2.8 Coefficients file & FORTRAN subroutine

After the user has confirmed satisfaction with the quality of the fit, the set of coefficients can be exported by clicking the “Export Coefficients” button. This exported data can subsequently be utilized by the corresponding FORTRAN subroutine to assess the long-range section of the Potential Energy Surface (PES). Moreover, it is possible to integrate this long-range region with the short-range region, developed using a different methodology. To evaluate the PES in a FORTRAN environment, the user needs our subroutine package LRF.f90 and coefficient set exported by the Matlab app **LRF**. The file LRF.f90 contains a Subroutine *evaluateLR*, which requires four input arguments to return the energy in wave-numbers(cm^{-1}):

- **Energy (REAL, output):** energy in wave-numbers(cm^{-1})
- **coordinates (REAL, input):** a real array containing the coordinates of a single geometrical configuration employing the distance between the molecules and set of angles. R must be in angstroms and the angles in degrees.
- **angular coordinate format (CHAR, input):** The user can choose between: *Euler_ZYZ*, *Euler_ZXZ* and *Spherical*.

- **xdim (INTEGER, input):** The numbers of internal coordinates or the dimension of the system depend on the system and follow the order described in the table 6.
- **file path (CHAR, input):** local path to the coefficient file exported by the Matlab app LRF.

System	Dime.	Euler Angles	Spherical
Atom - Atom	1	R	-
Linear - Atom	2	R, β_1	-
Non-Linear - Atom	3	R, β_1, γ_1	R, θ, φ
Linear - Linear	4	$R, \beta_1, \beta_2, \alpha$	-
Non-Linear - Linear	5	$R, \beta_1, \beta_2, \alpha, \gamma_1$	-
Non-Linear - Non-Linear	6	$R, \beta_1, \beta_2, \alpha, \gamma_1, \gamma_2$	-

Table 6: Coordinates array order. Linear and Non-Linear makes reference to Linear and Non-Linear molecules

The file “*LRF90” is part of the installation package and can also be obtained via email upon request. Below is a simple example demonstrating the application of our FORTRAN subroutine:

```

1 PROGRAM min_example
2   IMPLICIT NONE
3   integer :: xdim=4
4   real*8  :: coordinates(4)
5   real*8  :: Energy
6
7   coordinates(1) = 10.0d0  !R
8   coordinates(2) = 30d0    !beta1
9   coordinates(3) = 20d0    !beta2
10  coordinates(4) = 120d0   !alpha
11
12  call evaluateLR( Energy,&
13                  coordinates,&
14                  "Euler_ZYZ",&
15                  xdim,&
16                  "CFp_H2_coefficients.txt")
17
18 END PROGRAM min_example

```

Test Cases & Examples

4.1 Charged Molecules: $\text{CF}^+ - \text{H}_2$

The CF^+ molecule is regarded a crucial species in the exploration of fluorine chemistry within the interstellar medium (ISM). The recent identification of this molecule, along with its potential application as a tracer for atomic fluorine in the ISM, has intensified interest in investigating the physical and chemical attributes of this cation. Accurately measuring CF^+ concentrations in the ISM necessitates advanced models that consider its excitation, influenced by radiative and collisional interactions, primarily with atomic and molecular hydrogen.

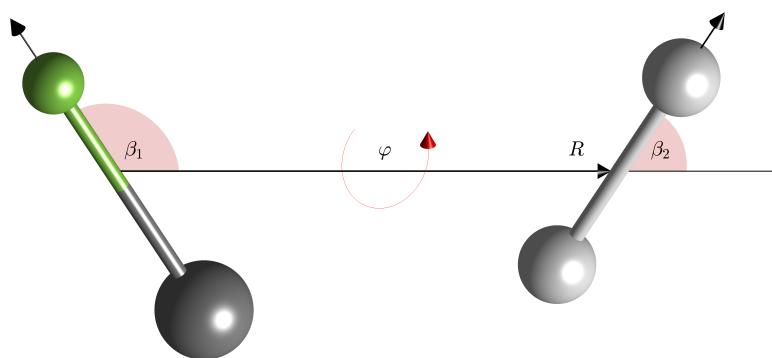


Figure 25: Representation of the internal coordinates for the $\text{CF}^+ - \text{H}_2$ system.

This system is composed of two linear molecules, the cation CF^+ with symmetry $C_{\infty v}$ and H_2 is a neutral linear molecule with symmetry $D_{\infty h}$. Our data file, “CFp-H2-Abinitio.dat” consists about 4000 rows and six columns, the number of rows and R, the distance between the centers of the masses, the cosine of β_1 and β_2 , the torsion angle φ and the energy of the configuration (see Figure 25). The value of R ranges from 10 Å to 30 Å, ensuring that only the long-range forces persist. More details of the system and the *ab-initio* calculations employed in this example can be found in [J. Phys. Chem. A 123, 9637–9643 (2019) DOI: 10.1021/acs.jpca.9b05538]. Let us begin the discussion of this system by opening the LRF “Setup” tab and follow the System Setup steps below.

Coordinates	Column in the Input Data File	Units
R	2	Angstrom (Å)
β_1	3	cosine
β_2	4	cosine
α	5	radians
γ_1	0	radians
γ_2	0	radians
Energy	6	kcal/mol

Figure 26: Two linear molecules, the cation CF^+ with symmetry $C_{\infty v}$ interacting with the H_2 with symmetry $D_{\infty h}$

1. Working Directory

- Choose the Working Directory

2. System Definition

- System Name: CFp-H2
- Symmetry of Molecule A : $C_{\infty v}$
- Charge of Molecule A : +1
- Symmetry of Molecule B : $D_{\infty h}$
- Charge of Molecule B : 0

3. Input Data File

- Select File: $\text{CFp-H2-Abinitio.dat}$
- Coordinate Format : *Autosurf* (column 1: index, column 2: R in Å, column 3: $\cos(\beta_1)$, column 4: $\cos(\beta_2)$, column 5: α in radians, and column 6: Interaction energy in $\frac{\text{kcal}}{\text{mol}}$. Check figure 26)

4. Press the “Initialization” button.

Since CF^+ is charged and the first nonzero multipole of H_2 is the quadrupole, the electrostatic, induction and dispersion leading terms of expansion are the charge-quadrupole ($q^A \Theta^B \sim R^{-3}$), the charge-dipolar polarizability ($(q^A)^2 \alpha_{\mu\mu}^B \sim R^{-4}$) and

Results													
id	Interactions	Range (Å)	RSME (cm ⁻¹)	R ²	Asymptote (cm ⁻¹)	qa (e)	qb (e)	μa (a. u.)	μb (a. u.)	Θa (a. u.)	Θb (a. u.)	Ωa (a. u.)	Ωb (a. u.)
0	Initial Values	[- ; -]	0	0	-30427239.937937	1	0	-0.4496	0	0.7388	0.4679	1.2551	0
1	E-3//	[10 ; 30]	1.2119	0.8141	-30427240.9180393	1	0	-0.4496	0	0.7388	0.4481	1.2551	0
2	E-3-4//	[10 ; 30]	1.2072	0.8156	-30427240.9180873	1	0	-0.4193	0	0.7388	0.4483	1.2551	0
3	E-3-4-5//	[10 ; 30]	1.2074	0.8156	-30427240.9179734	1	0	-0.4189	0	0.7757	0.4483	1.2551	0
4	E-3-4-5-6//	[10 ; 30]	1.2075	0.8156	-30427240.9179734	1	0	-0.4189	0	0.7757	0.4483	1.2551	0
5	E-3-4-5-6//I-4//	[10 ; 30]	0.0816	0.9992	-30427239.8430282	1	0	-0.4015	0	0.3056	0.4831	1.9904	0
6	E-3-4-5-6//I-4-5//	[10 ; 30]	0.0109	1.0000	-30427239.8426099	1	0	-0.4454	0	0.3269	0.4829	2.2158	0
7	E-3-4-5-6//I-4-5-6//	[10 ; 30]	0.0054	1.0000	-30427239.8508675	1	0	-0.4556	0	0.4138	0.4888	2.1274	0
8	E-3-4-5-6//I-4-5-6//D-6	[10 ; 30]	0.0047	1.0000	-30427239.8508871	1	0	-0.4552	0	0.3810	0.4884	2.0969	0
9	E-3-4-5-6-7//I-4-5-6-7//D-6	[10 ; 30]	0.0009	1.0000	-30427239.852478	1	0	-0.4427	0	0.3646	0.4804	2.3273	0
10	E-3-4-5-6-7//I-4-5-6-7//D-6-7	[10 ; 30]	0.0009	1.0000	-30427239.8524792	1	0	-0.4425	0	0.3724	0.4804	2.2537	0

Figure 27: Two linear molecules, the cation CF^+ with symmetry $C_{\infty v}$ interacting with the H_2 with symmetry $D_{\infty h}$

the always present sixth order of dispersion ($D_{\mu\mu-\mu\mu} \sim R^{-6}$). The fitting coefficients are the asymptote, multipoles, polarizability, and nonzero dispersion coefficients of A and B. The number of new coefficients added every time we add a new interaction order can go from few (lower order) to tens of them (higher order), so in our experience the best strategy is to add just one new interaction term every time we run a new fit. Let's start the fit adding the electrostatic interactions up to sixth order. As shown in 27, the electrostatic leading term has a quite good start, however; E₄, E₅ and E₆ (fits id 2,3 and 4) do not considerably improve the fit. On the other hand, the induction terms I₄, I₅ and I₆ (fits id 5,6 and 7) have a great impact in both the global RMSE and R² while dispersion leading term ($\sim R^{-6}$) has a smaller impact. Notice that the last fits show small variations of the first multipole values (dipoles, quadrupoles, and octupoles), and they are close to the real values, which is a very good sign of the fit convergence. However, in not all scenarios the coefficients are close to the real values; for example, the fit forced the coefficients to absorb the missing terms to obtain the best possible fit. In some systems, the coefficients cannot be absolutely determinate. Consider two linear, neutral, and polar molecules with dipoles oriented along the z-axis; here, the primary term is the dipole-dipole interaction expressed as $E_3 = (\mu_{10}^A \mu_{10}^B) T_{10,10}$. When fitting is performed considering only this main term, the fit will be meaningful solely for the product $\mu_{10}^A \cdot \mu_{10}^B$ since there's no constraint ensuring each coefficient separately gets a real value. In contrast, in systems such as charged ones, all coefficients are well defined due to a cascade effect; known charge values determine the related multipoles in the electrostatic expansion, such as the leading term $q^A \Theta_{20}^B \cdot T_{20,0}(R, \alpha, \beta_1, \beta_2)$, which in turn defines others until all multipoles are determined. Once the multipoles are known, the polarizability coefficients become determinable in the induction terms. For multi-atom fragments, a similar effect can be achieved by knowing and fixing at least one multipole coefficient during the fitting process. For systems involving neutral atoms, the cascade may not be achievable as is shown in *****nonlinear molecule system/ *****.

Another notable drop in RMSE comes with I₇ but not when adding dispersion D₇, reasonable since the electrostatic and induction leading terms decay as $\sim R^{-3}$ and $\sim R^{-4}$ and therefore make sense that dispersion may be less relevant in this

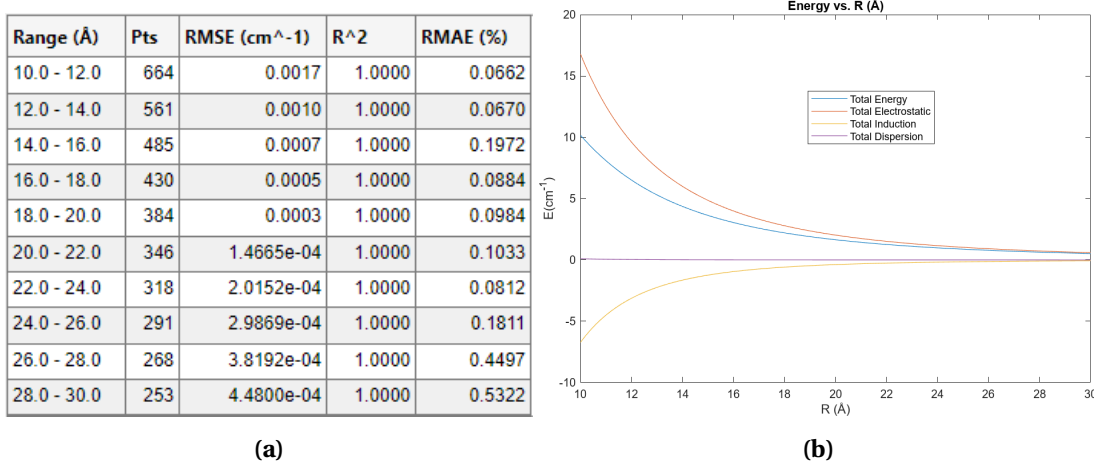


Figure 28: Caption place holder

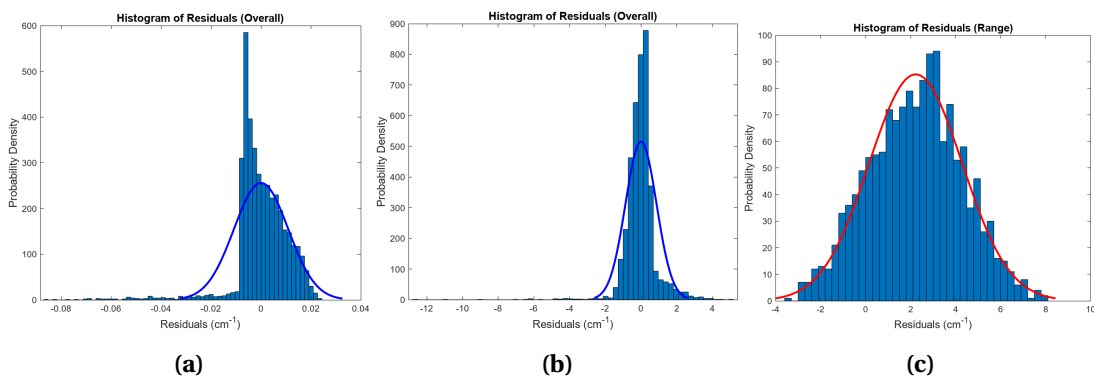


Figure 29: Caption place holder

system. There is no universal criterion to define the best fit. **LRF** comes with different tools to assess and determine the best fit. In general, we look at the following features:

- Global *RMSE* and R^2 .
- Pattern of *RMSE* and *RMAE* by range.
- Patterns of the residuals.
- Number of parameters.

The global *RMSE* and R^2 are excellent measurements of the quality of the fit but are not conclusive. The model also must exhibit a well-behavior in any region of the fitting space. *RMSE* is expected to be higher at lower R and decay toward zero in the ultra long-range. *RMAE* is expected to be lower for small values of R and higher in the ultra long range (see Figure (a)).

On the other hand, residuals show to be near-normally distributed. For example, the fit with $id = 6$, has reasonable $RMSE = 0.01$, an perfect $R^2 = 1$ and includes the

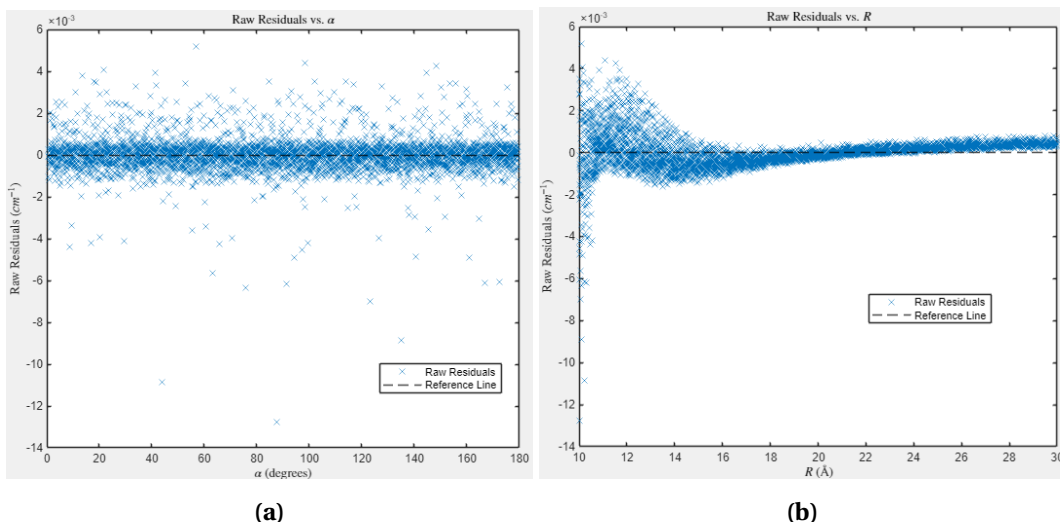
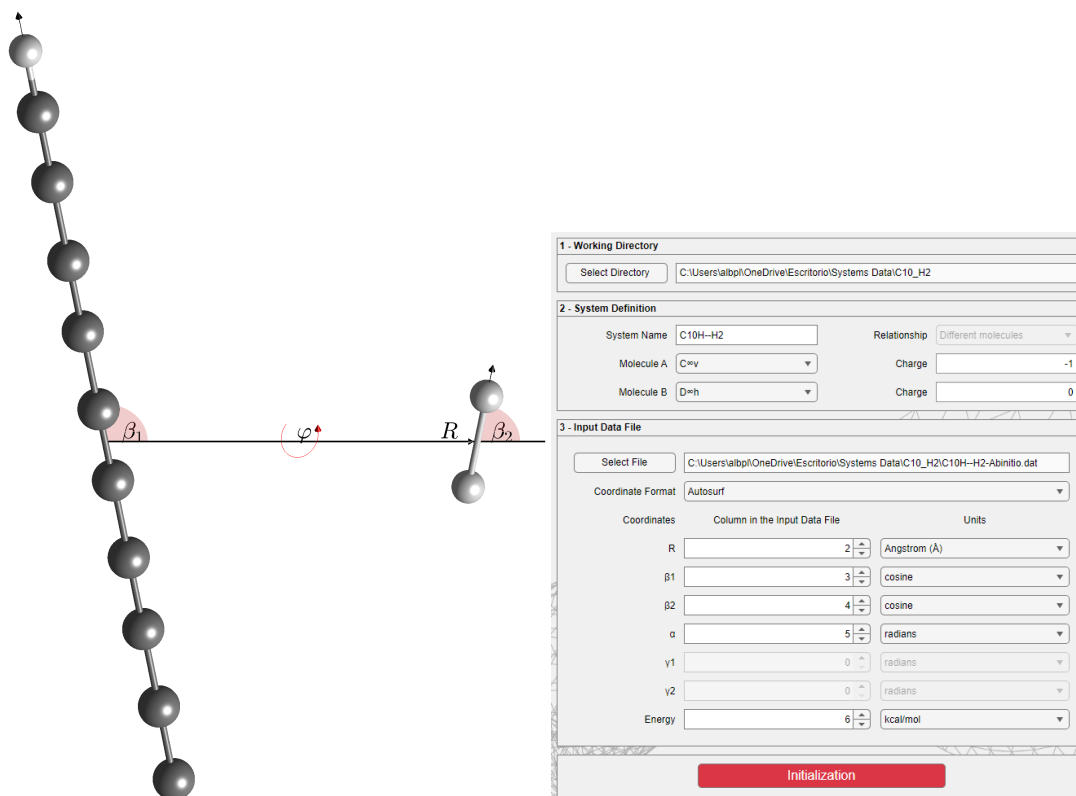


Figure 30: Caption place holder

most important terms in the expansion (E_3 , I_4 and I_5), since dispersion is about 20 times smaller than induction and 100 times than electrostatic. However, the residual histogram is absolutely asymmetric and far away from a normal distribution (see figure 29(a)). The histogram is expected to resemble a normal distribution more closely with the inclusion of higher-order terms in the fit (see histogram for $id=10$, figure 29(b)). Furthermore, filtering out the lowest R values should enhance the normal shape in any fit subspace, particularly noticeable when lower values are taken off in the histogram ($id=10$ with $R \geq 20\text{\AA}$ (Figure 29(c)). Occasionally, the ultra long-range appears asymmetric because the asymptotic value slightly deviates from the actual value; however, this can be corrected by adhering to the asymptote calculation guidelines presented in the subsequent example. Furthermore, the residual should display an angular distribution that highlights the directions of higher and lower energies and an R distribution that decays following a power law of R (Figure 30).

The model's number of independent parameters should be significantly fewer than the number of data points; otherwise, the fitting can be overly flexible, resulting in a minimal $RMSE$ and an excellent R^2 . The number of parameters input into the model (number of parameters reported in **LRF** Logs tab) is an upper quote of the number of independent parameters (if only the dipole-dipole interaction ($\mu^A \cdot \mu^B$) is fitted, **LRF** will indicate two parameters (one per dipole), yet only the product of the dipoles will remain independent) so it can be used as the worst estimation of the number of data points per model parameter. In this case, we have 22 parameters in the last fit, which gives us 181 data points per parameter. Moreover, parameter count provides insight into the evaluation speed of the potential, as more parameters mean more terms to compute.



(a) Representation of the internal coordinates for the $\text{C}_{10}\text{H}^- - \text{H}_2$ system.

(b)

4.2 Anisotropic Molecules: $\text{C}_{10}\text{H}^- - \text{H}_2$

Carbon-chain anions have been identified in the interstellar medium (ISM) fairly recently, offering crucial insights into the physical and chemical states in different astrophysical contexts. However, local thermodynamic equilibrium (LTE) conditions are rarely met in these regions. Carbon-chain anions are crucial due to their impact on ISM chemical dynamics, although they are much less abundant than their neutral counterparts. Current chemical models fail to accurately replicate the observed anion-to-neutral ratios. In cool, low-density areas where anions are found, LTE assumptions generally do not hold, and accurate emission line modeling must consider radiative and collisional rates with H_2 , the most common molecular species.

This system is composed of two linear molecules, the anion C_{10}H^- with symmetry $C_{\infty v}$ and H_2 is a neutral linear molecule with symmetry $D_{\infty h}$. The anion is long sting of 11 atoms which the potential is extremely anisotropic, where the overlap in parallel position is negligible about 8.5\AA where in T-shape positions there is a well about 12\AA . Our data file, “CFp-H2-Abinitio.dat” consists about 9370 rows and six columns, the number of rows and R, the distance between the centers of the masses, the cosine of β_1 and β_2 , the torsion angle φ and the energy of the configuration (see Figure 25). The value of R ranges from 2.2\AA to 30\AA . Let us begin the discussion of this system by opening the LRF “Setup” tab and follow the System

Results													
id	Interactions	Range (Å)	RSME (cm ⁻¹)	R ²	Asymptote (cm ⁻¹)	qa (e)	qb (e)	μa (a. u.)	μb (a. u.)	Θa (a. u.)	Θb (a. u.)	Ωa (a. u.)	Ωb (a. u.)
0	Initial Values	[- ; -]	0	0	-83833525.3271793	-1	0	13.8329	0	-207.0476	0.5794	-1.2184	0
1	E-3//	[15 ; 29.9]	1.9406	0.5824	-83833525.8591208	-1	0	13.8329	0	-207.0476	0.8021	-1.2184	0
2	E-3-4//	[15 ; 29.9]	1.3959	0.7841	-83833525.8551377	-1	0	8.0633	0	-207.0476	0.7125	-1.2184	0
3	E-3-4-5//	[15 ; 29.9]	1.1729	0.8479	-83833525.7404686	-1	0	10.5272	0	-119.6814	0.4988	-1.2184	0
4	E-3-4-5-6//	[15 ; 29.9]	1.0682	0.8739	-83833525.7591425	-1	0	6.3302	0	-108.3518	0.5095	1.4812e+03	0
5	E-3-4-5-6//I-4/	[15 ; 29.9]	0.9872	0.8925	-83833525.107692	-1	0	6.6799	0	-116.7550	0.4573	1.7664e+03	0
6	E-3-4-5-6//I-4-5/	[15 ; 29.9]	0.9345	0.9037	-83833525.1493144	-1	0	5.7990	0	-94.5571	0.5938	1.2956e+03	0
7	E-3-4-5-6//I-4-5-6/	[15 ; 29.9]	0.9031	0.9103	-83833525.1729891	-1	0	6.0668	0	-103.8740	0.5554	1.3340e+03	0
8	E-3-4-5-6//I-4-5-6/D-6	[15 ; 29.9]	0.9032	0.9104	-83833525.1906553	-1	0	6.2277	0	-115.7205	0.5490	1.3337e+03	0
9	E-3-4-5-6-7//I-4-5-6/D-6	[15 ; 29.9]	0.8907	0.9131	-83833525.2047508	-1	0	6.8048	0	-116.4574	0.5627	1.2039e+03	0
10	E-3-4-5-6-7//I-4-5-6-7/D-6	[15 ; 29.9]	0.8855	0.9142	-83833525.2193171	-1	0	6.6012	0	-122.4784	0.5525	1.3425e+03	0
11	E-3-4-5-6-7//I-4-5-6-7/D-6-7	[15 ; 29.9]	0.8867	0.9142	-83833525.2193171	-1	0	6.6012	0	-122.4784	0.5525	1.3425e+03	0

Figure 32: Two linear molecules, the cation $C_{10}H^-$ with symmetry $C_{\infty v}$ interacting with the H_2 with symmetry $D_{\infty h}$

Setup steps below.

1. Working Directory

- Choose the Working Directory

2. System Definition

- System Name: *C10H-H2*
- Symmetry of Molecule A : $C_{\infty v}$
- Charge of Molecule A : -1
- Symmetry of Molecule B : $D_{\infty h}$
- Charge of Molecule B : 0

3. Input Data File

- Select File: *C10H-H2-Abinitio.dat*
- Coordinate Format : *Autosurf* (column 1: index, column 2: R in Å, column 3: $\cos(\beta_1)$, column 4: $\cos(\beta_2)$, column 5: α in radians, and column 6: Interaction energy in $\frac{kcal}{mol}$. Check figure ??(b))

4. Press the “Initialization” button.

Since the anion C_{10}^- is charged and the first nonzero multipole of H_2 is the quadrupole, the electrostatic, induction and dispersion leading terms of expansion are the charge-quadrupole ($q^A \Theta^B \sim R^{-3}$), the charge-dipolar polarizability ($(q^A)^2 \alpha_{\mu\mu}^B \sim R^{-4}$) and the always present sixth order of dispersion ($D_{\mu\mu-\mu\mu} \sim R^{-6}$). As shown in 27, the electrostatic leading term has a quite good start, however; E_4 , E_5 and E_6 (fits id 2,3 and 4) do not considerably improve the fit. On the other hand, the induction terms I_4 , I_5 and I_6 (fits id 5,6 and 7) have a great impact in both the global $RMSE$ and R^2 while the dispersion leading term ($\sim R^{-6}$) has a smaller impact. Notice that the

Range (Å)	Pts	RMSE (cm ⁻¹)	R ²	RMAE (%)
15.0 - 17.5	283	0.3294	0.9953	34.6557
17.5 - 20.0	259	0.5682	0.9349	56.4057
20.0 - 22.4	259	0.2090	0.9815	57.3410
22.4 - 24.9	233	1.7598	0.2860	48.3015
24.9 - 27.4	50	0.1380	0.9992	59.3605
27.4 - 29.9	15	0.1077	0.9993	53.7660

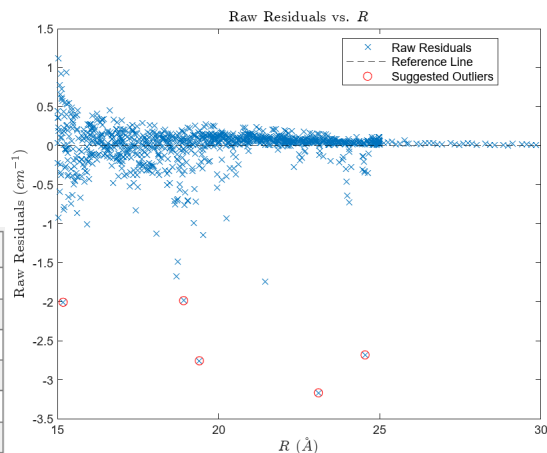
Range (Å)	Pts	RMSE (cm ⁻¹)	R ²	RMAE (%)
15.0 - 17.5	283	0.3278	0.9953	28.2946
17.5 - 20.0	258	0.3532	0.9746	80.0913
20.0 - 22.4	259	0.1715	0.9820	60.7832
22.4 - 24.9	229	0.2967	0.9040	36.0532
24.9 - 27.4	50	0.0503	0.9992	23.1586
27.4 - 29.9	15	0.0178	0.9991	22.2041

(a) Representation of the internal coordinates for the C₁₀H⁻-H₂ system.

(b) Representation of the internal coordinates for the C₁₀H⁻-H₂ system.

Range (Å)	Pts	RMSE (cm ⁻¹)	R ²	RMAE (%)
15.0 - 17.5	283	0.3294	0.9953	34.6557
17.5 - 20.0	259	0.5682	0.9349	56.4057
20.0 - 22.4	259	0.2090	0.9815	57.3410
22.4 - 24.9	233	1.7598	0.2860	48.3015
24.9 - 27.4	50	0.1380	0.9992	59.3605
27.4 - 29.9	15	0.1077	0.9993	53.7660

(a) Representation of the internal coordinates for the C₁₀H⁻-H₂ system.



(b) Representation of the internal coordinates for the C₁₀H⁻-H₂ system.

last fits show small variations of the first multipole values (dipoles, quadrupoles, and octupoles), and they are close to the real values, which is a very good sign of the fit convergence. However, in not all scenarios the coefficients are close to the real values; for example, the fit forced the coefficients to absorb the missing terms to obtain the best possible fit. In some systems, the coefficients cannot be absolutely determinate. Consider two linear, neutral, and polar molecules with dipoles oriented along the z-axis; here, the primary term is the dipole-dipole interaction expressed as $E_3 = (\mu_{10}^A \mu_{10}^B) T_{10,10}$. When fitting is performed considering only this main term, the fit will be meaningful solely for the product $\mu_{10}^A \cdot \mu_{10}^B$ since there is no constraint ensuring each coefficient separately gets a real value. In contrast, in systems such as charged ones, all coefficients are well defined due to a cascade effect; known charge values determine the related multipoles in the electrostatic expansion, such as the Θ_{20}^B leading term $q^A \Theta^B$, which in turn defines others until all multipoles are determined. Once the multipoles are known, the polarizability coefficients become determinable in the induction terms, which also define the dispersion coefficients. A similar effect can be achieved by knowing and fixing at least one multipole (or polarizability in case neutral atoms) coefficient during the fitting process.

For numerous reasons, the ab initio calculations may not converge, introducing bad points into our data set that have a harmful impact on R^2 and $RMSE$. If this assumption is true, then the model stats must have a big improvement after removing

those few outliers from the data. This is the case for our dataset for this system. If we look at the fit table, fitting with all possible terms up to seventh order, $R \approx 0.91$ and $RMSE \approx 0.89$, which are unexpectedly bad considering that the fit includes up to four orders beyond the leading terms (see figure 32). If we take a closer look at the stats by range, as shown in figure ?? (a), (in the app, navigate to **Logs > Stats**, set “Splitting” to 6 and press “Filter”), we can see that the model fits well in most ranges, but has poor performance for $[17.5\text{\AA} - 20\text{\AA}]$ and extremely poor for $[22.4\text{\AA} - 24.9\text{\AA}]$. If we go to the “Outliers” tab and set the “Probability Threshold” to 0.5, we can see the five worst points in the data, which are located in those regions (see figure ?? (b)). If we hover over the outliers points we can see some information about them including the row number where they are located in the input file, as well as their value of R and Energy. Let us make a copy of our input file (*C10H-H2-Abinitio.dat*) and delete the outliers, and then press the “reset workspace” button and load the outlier free data file and repeat the fit up to seventh order. Notice that the range stats improves considerably in the poor performance regions observed before but still there are clear outliers on the data. Let’s repeat the same procedure two more times until we have removed, first, the worst 5 and then the worst 3 points on the data. It is important to remove few outliers in every iteration since the fit is corrupted those outliers so fitted model could be far away from the real fit in order to improve the general RMSE.

4.3 Identical Molecules: CO-Dimer

The asymptote is fitted every time a new fit is done as a default behavior, however; if there is not enough data points in the ultra long-range, the asymptote could have been displaced to better accommodate the data closer to the short range, since the energy in this region is bigger. If we have data with R ridiculously bigger, we can take this energy as the asymptote and fix it in the app or fit the asymptote with the data in the ultra long-range just employing the leading terms (the only ones that survive at such long distances).

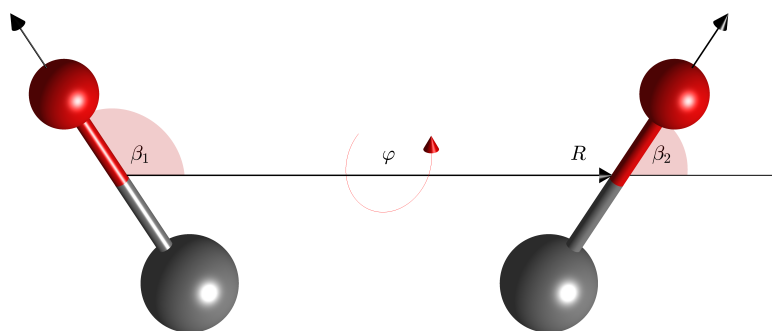


Figure 35: Representation of the internal coordinates for the CO dimer system.

Another special case is when both molecules are identical, for example, a CO dimer system. In astrochemistry, the CO molecule holds significant importance be-

cause of its prevalence in interstellar space, serving as an essential tracer for identifying molecular clouds and star-forming regions, thereby revealing formation conditions. It also plays a vital role as a precursor in forming complex molecules on dust grain surfaces, functioning as a foundational element in interstellar chemistry [?].

1 - Working Directory

Select Directory: C:\Users\albp\OneDrive\Escritorio\System Data\CO_Dimer

2 - System Definition

System Name: CO_Dimer Relationship: Identical

Molecule A: $C_{\infty v}$ Charge: 0

Molecule B: $C_{\infty v}$ Charge: 0

3 - Input Data File

Select File: C:\Users\albp\OneDrive\Escritorio\System Data\CO_Dimer\AbINITIO_CO-dimer.dat

Coordinate Format: Autosurf

Coordinates	Column in the Input Data File	Units
R	2	Angstrom (Å)
β_1	3	cosine
β_2	4	cosine
α	5	radians
γ_1	0	radians
γ_2	0	radians
Energy	6	kcal/mol

Initialization

Figure 36: Two linear molecules, the cation CF^+ with symmetry $C_{\infty v}$ interacting with the H_2 with symmetry $D_{\infty h}$

Let's start in the **LRF Setup** tab and follow the System Setup steps below:

1. Working Directory

- Choose the Working Directory

2. System Definition

- System Name: *CO-Dimer*
- Symmetry of Molecule A : $C_{\infty v}$
- Charge of Molecule A : 0
- Symmetry of Molecule B : $C_{\infty v}$
- Charge of Molecule B : 0
- Relationship: *Identical*

Results											
id	Interactions	Range (Å)	RSME (cm ⁻¹)	R ²	Asymptote (cm ⁻¹)	qa (e)	qb (e)	μa (a. u.)	μb (a. u.)	Θa (a. u.)	Θb (a. u.)
0	Initial Values	[- ; -]	0	0	-49740543.8612685	0	0	0.1052	0.1052	1.2369	1.2369
1	E-3//	[6 ; 14.9]	3.6563	0.0811	-49740545.962355	0	0	0.1327	0.1327	1.2369	1.2369
2	E-3-4//	[6 ; 14.9]	3.5731	0.1241	-49740546.0296789	0	0	0.1215	0.1215	0.8010	0.8010
3	E-3-4-5//	[6 ; 14.9]	3.2932	0.2573	-49740546.1639718	0	0	0.0833	0.0833	1.1728	1.1728
4	E-3-4-5-6//6/	[6 ; 14.9]	0.9265	0.9415	-49740543.7404493	0	0	0.0997	0.0997	0.7921	0.7921
5	E-3-4-5-6//6/D-6	[6 ; 14.9]	0.8677	0.9488	-49740543.6587832	0	0	0.0905	0.0905	0.9753	0.9753
6	E-3-4-5-6-7//6-7/D-6-7	[6 ; 14.9]	0.2661	0.9952	-49740543.618979	0	0	0.1021	0.1021	1.2025	1.2025
7	E-3-4-5-6-7-8//6-7-8/D-6-7-8	[6 ; 14.9]	0.1033	0.9993	-49740543.6959369	0	0	0.1107	0.1107	0.6428	0.6428
8	E-3-4-5-6-7-8-9//6-7-8-9/D-6-7-8-9	[6 ; 14.9]	0.0684	0.9997	-49740543.6962519	0	0	0.1114	0.1114	0.7407	0.7407
9	E-3-4-5-6-7-8-9-10//6-7-8-9-10/D-6-7-8-9-10	[6 ; 14.9]	0.0279	1.0000	-49740543.7172106	0	0	0.1118	0.1118	0.6058	0.6058
10	E-3-4-5-6-7-8-9-10-11//6-7-8-9-10-11/D-6-7-8-9-10-11	[6 ; 14.9]	0.0236	1.0000	-49740543.7192003	0	0	0.1116	0.1116	0.7000	0.7000

Figure 37: Two linear molecules, the cation CF^+ with symmetry $C_{\infty v}$ interacting with the H_2 with symmetry $D_{\infty h}$

3. Input Data File

- Select File: *CO-Dimer-Abinitio.dat*
- Coordinate Format : *Autosurf* (column 1: index, column 2: R in Å, column 3: $\cos(\beta_1)$, column 4: $\cos(\beta_2)$, column 5: α in radians, and column 6: Interaction energy in $\frac{kcal}{mol}$. Check figure 36)

4. Press the “Initialization” button.

The leading terms of expansion are the dipole-dipole in the third order of electrostatic (μ^2), the dipole-dipole polarizability in induction in the sixth order ($\mu^2\alpha_{\mu\mu}$) and dispersion in the sixth order (see tab expansion). The Ab initio data contain 2208 points spread between 2.7 and 14.9 Å, but only a few points are located in the very long range. In those cases, we recommend setting the minimum R value as low as the model can give out without losing much precision. In this case, we can set the minimum R = 6 Å, so **LRF** a subset of 540 points. Let us start the fitting procedure by adding the first three none zero orders of electrostatic (E_3, E_4 , and E_5), once at a time. As shown in the figure (id =1,2,3), the model does not perform well using only electrostatic terms. Next, add E_6 and the first non-zero term of induction I_6 which will show a huge improvement in the *RMSE* and R^2 of the model. The addition of D_6 shows a more modest improvement in the fit. Keep adding terms of the same order E_n, I_n and D_n , all at once, up to eleventh order (adding too many variables at once could lead the fit engine to find a local minimum and get stuck in it when adding new terms in further steps of the fitting procedure; however, since both molecules are linear and identical, the number of independent coefficients is reduced considerably).

Observe that each incremental order of terms consistently enhances the fit, with the exception of the final term. Hence, we consider the optimal fit to be the model identified by id = 9. This model incorporates all nonzero terms up to the 10th order, resulting in a total of 52 estimated independent coefficients. Due to the necessity of extending our data collection to 6Å, it was imperative that the model include higher-order terms, as these are more significant at smaller values of R. The model exhibits a global *RMSE* of approximately $2 \times 10^{-2} \text{ cm}^{-1}$, alongside an ideal R^2 value of 1.

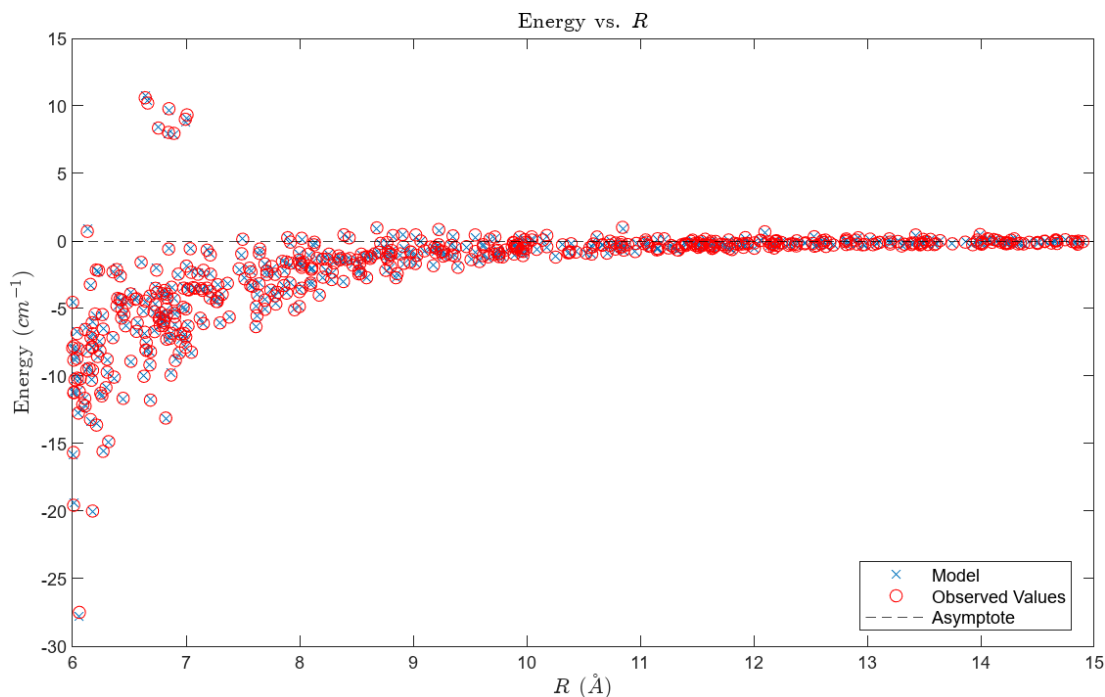


Figure 38: Two linear molecules, the cation CF^+ with symmetry $C_{\infty v}$ interacting with the H_2 with symmetry $D_{\infty h}$

The residuals adhere to an almost perfectly normal distribution, as illustrated in figure 40. Both the $RMSE$ and R^2 maintain excellent performance even within the narrower segments of R , specifically between 6-15Å. Within this range, a discernible pattern emerges: the $RMSE$ diminishes, whereas the $RMAE$ grows as R increases. This trend occurs because the energy value approaches zero, as demonstrated in figure 39.

Range (Å)	Pts	RMSE (cm ⁻¹)	R ²	RMAE (%)
6.0 - 6.9	115	0.0358	1.0000	0.5971
6.9 - 7.8	69	0.0270	0.9999	0.9601
7.8 - 8.7	63	0.0250	0.9997	2.1871
8.7 - 9.6	48	0.0231	0.9993	3.4721
9.6 - 10.4	48	0.0278	0.9971	6.1885
10.4 - 11.3	37	0.0309	0.9944	11.5664
11.3 - 12.2	59	0.0190	0.9939	14.8915
12.2 - 13.1	35	0.0158	0.9939	8.6444
13.1 - 14.0	29	0.0175	0.9963	19.5957
14.0 - 14.9	37	0.0139	0.9912	20.8277

Figure 39: E vs R for $\beta_1 = 90$ and $\varphi = 0$

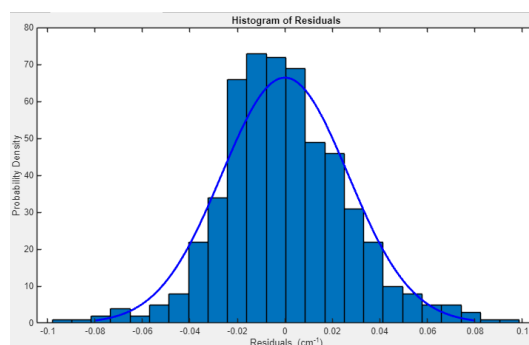


Figure 40: E vs R for $\beta_1 = 0$ and $\varphi = 0$

Over large separations, the electrostatic component characterized by the leading term ($\sim R^{-3}$) predominantly determines the overall interaction energy due to the more rapid decay of other contributions ($\sim R^{-6}$). Conversely, at closer proximities, the minor dipole and quadrupole moments of CO diminish the influence of the electrostatic interaction, thereby elevating the roles of induction and especially

dispersion forces as the primary contributors to the total interaction energy, as illustrated in figures 41 and 42. An additional intriguing aspect of this system is that when CO assumes a parallel alignment, the potential energy is attractive. However, when CO is in an antiparallel orientation, the potential energy shifts to a repulsive regime, resulting in the formation of a potential barrier (figures ??).

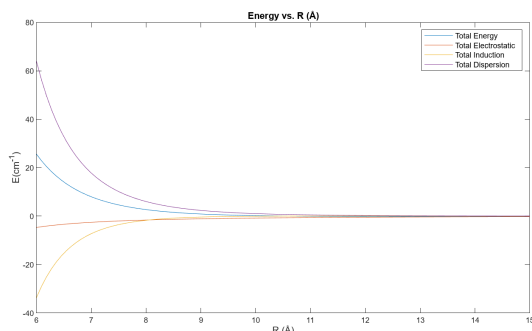


Figure 41: E vs R for CO-CO in parallel ($\beta_1 = 0$, $\beta_2 = 0$ and $\varphi = 0$)

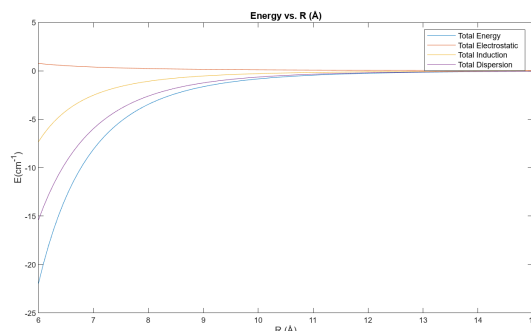


Figure 42: E vs R for CO-CO in T shape ($\beta_1 = 90$, $\beta_2 = 0$ and $\varphi = 0$)

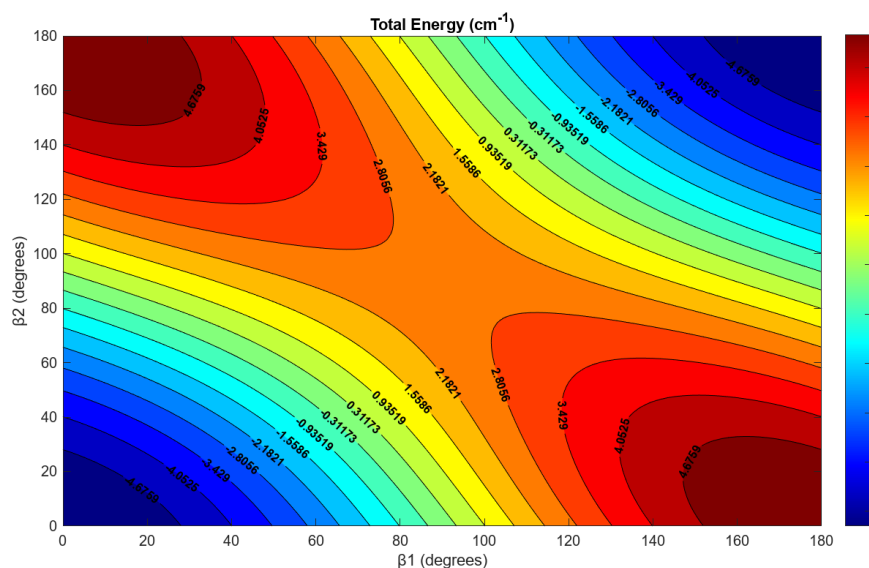


Figure 43: Two linear molecules, the cation CF^+ with symmetry $C_{\infty v}$ interacting with the H_2 with symmetry $D_{\infty h}$

4.4 Beyond linear molecules : O_3 - Ar

LRF is also applicable in modeling systems that consist of atoms or nonlinear molecules. To illustrate this, consider the scenario that involves the interaction between ozone (O_3), a non-linear molecule that is notably recognized as a key role in the chemistry of our atmosphere, and a helium atom [?].

Let's start in the **LRF Setup** tab and follow the System Setup steps below:

1. Working Directory

- *Choose the Working Directory*

2. System Definition

- System Name: O_3-He
- Symmetry of Molecule A : C_{2v} (The ab initio data points were computed for the ozone molecule considering this specific symmetry)
- Charge of Molecule A : 0
- Symmetry of Molecule B : *Spherical*
- Charge of Molecule B : 0

3. Input Data File

- Select File: *O3-Ar-Abinitio.dat*
- Coordinate Format : *Autosurf* (column 1: index, column 2: R in Å, column 3: $\cos(\beta_1)$, column 4: $\gamma_1 = \varphi$ in radians, and column 5: Interaction energy in *kcal/mol*. Check figure 45)

4. Press the “Initialization” button.

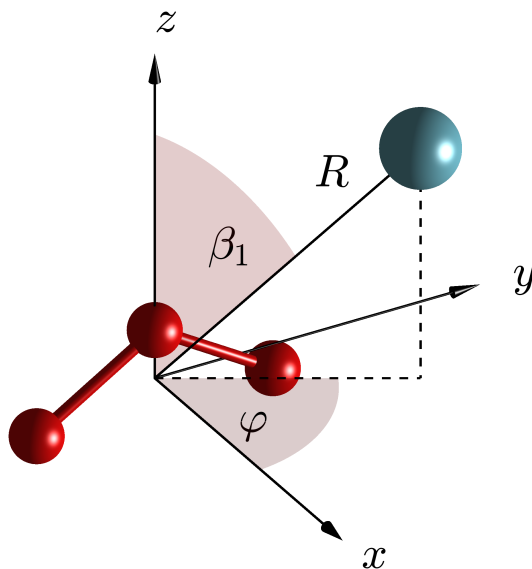


Figure 44: Caption for Figure 1

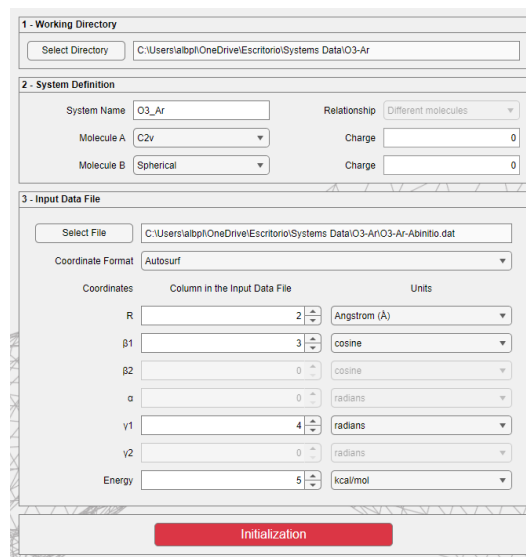


Figure 45: Caption for Figure 2

Because the atoms have spherical symmetry, all of its multipole moments vanish, so the electrostatic and induction interactions of A (due to the distortion of multipoles of B over A) are zero. The leading terms of expansion are the sixth-order induction B ($\mu^{A^2} \cdot \alpha_{\mu\mu}^B$) and dispersion. The Ab initio data contain 600 points spread between 8 and 25 Å, where the energies are just a tiny fraction of the depth of the well.

Results											
id	Interactions	Range (Å)	RSME (cm ⁻¹)	R ²	Asymptote (cm ⁻¹)	qa (e)	qb (e)	μa (a. u.)	μb (a. u.)	Θa (a. u.)	Θb (a. u.)
0	Initial Values	[- ; -]	0	0	0	0	0	-0.7411	0	0.8096	0
1	/I-6/	[8 ; 24.9]	0.2597	0.6704	-165116513.805004	0	0	-15.0614	0	0.8096	0
2	/I-6/D-6	[8 ; 24.9]	0.0265	0.9966	-165116513.742109	0	0	-9.7587	0	0.8096	0
3	/I-6-7/D-6	[8 ; 24.9]	0.0236	0.9973	-165116513.741829	0	0	-9.5342	0	3.9010	0
4	/I-6-7/D-6-7	[8 ; 24.9]	0.0202	0.9980	-165116513.741809	0	0	-8.7451	0	7.7112	0
5	/I-6-7-8/D-6-7	[8 ; 24.9]	0.0201	0.9981	-165116513.742088	0	0	-8.7183	0	8.2371	0
6	/I-6-7-8/D-6-7-8	[8 ; 24.9]	0.0203	0.9981	-165116513.742083	0	0	-8.7181	0	8.2479	0
7	/I-6-7-8-9/D-6-7-8	[8 ; 24.9]	0.0055	0.9999	-165116513.753952	0	0	-7.0733	0	1.5160	0
8	/I-6-7-8-9/D-6-7-8-9	[8 ; 24.9]	0.0055	0.9999	-165116513.753952	0	0	-7.0733	0	1.5160	0
9	/I-6-7-8-9-10/D-6-7-8-9	[8 ; 24.9]	0.0050	0.9999	-165116513.75384	0	0	-7.4790	0	1.7175	0
10	/I-6-7-8-9-10/D-6-7-8-9-10	[8 ; 24.9]	0.0050	0.9999	-165116513.753842	0	0	-7.4783	0	1.7398	0
11	/I-6-7-8-9-10-11/D-6-7-8-9-10-11	[8 ; 24.9]	0.0050	0.9999	-165116513.753844	0	0	-7.4783	0	1.7383	0

Figure 46: Two linear molecules, the cation CF^+ with symmetry $C_{\infty v}$ interacting with the H_2 with symmetry $D_{\infty h}$

The fitting strategy will be to add one iteration order at a time and switch from induction (I) to dispersion (D) starting with I_6 . Following this procedure, we should obtain the pattern of $RMSE$ and R^2 shown in figure 46. The big drops in the global $RMSE$ are seen after adding D_6 (id 2) and I_9 (id 7). Contributions in tenth order or higher are negligible, so they do not improve the fit. The dispersion is dominant when the position of the atom is close to the xy plane ($\beta_1 \approx \frac{\pi}{2}$) while induction occurs for β_1 is close to 0 and π (see figures 47 and 48)

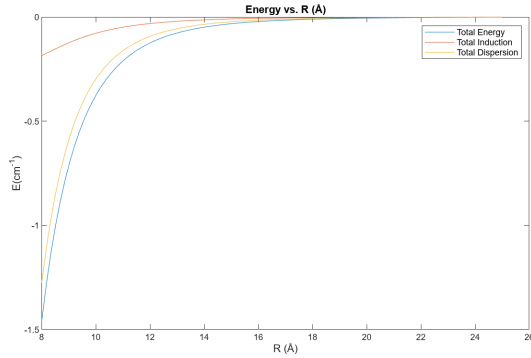


Figure 47: E vs R for $\beta_1 = 90$ and $\varphi = 0$

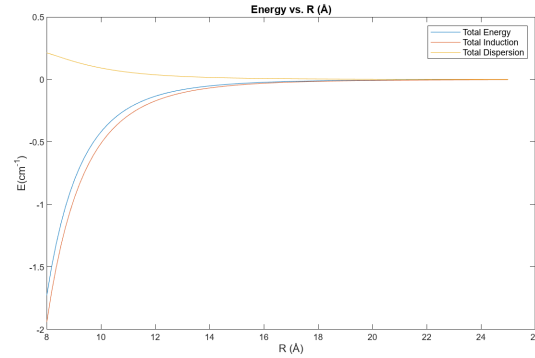


Figure 48: E vs R for $\beta_1 = 0$ and $\varphi = 0$

The residual distribution shows a balanced histogram showing that the expected behavior has larger errors at lower R with no evident presence of outliers in the data (no points are marked with a red circle). Its difficult to imagine before the fit the key role played by I_9 particularly being a so small interaction ($10^{-4} cm^{-1}$). The estimated number of parameters used by the fit $id = 7$ is 22, with global $RMSE$ at $id = 7$ is $5.5 \times 10^{-3} cm^{-1}$, R^2 close to 1 which makes this fit a great candidate to be consider the best fit coefficient set.

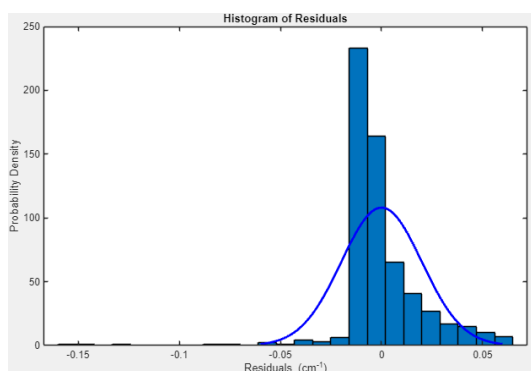


Figure 49: fit id = 6 Unbalanced residual histogram before adding I_9

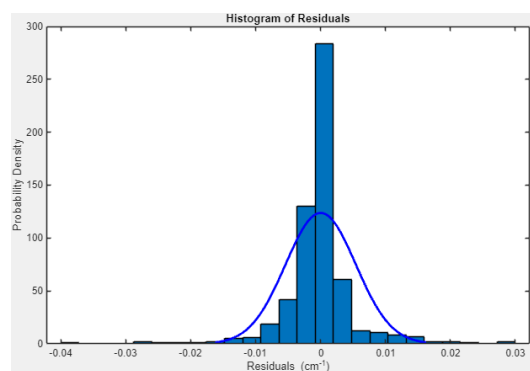


Figure 50: fit id = 7 Unbalanced residual histogram after adding I_9