MedeA: Fitting Data Manager

Contents

1 Introduction

Forcefield optimization as well as the generation of machine-learned potentials requires the management of the results of quantum mechanical calculations which form the scientific basis of such potentials. *MedeA* uses fitting training sets to handle data for molecular or periodic structures with the associated properties such as energies, forces, and stresses. The *MedeA* Fitting Data Manager can create and modify fitting training sets, which can be used by the *Forcefield Optimizer* and *Machine-Learned Potential Generator* in subsequent steps.

The *MedeA* Fitting Data Manager can be accessed from the main menu bar of *MedeA* in the File menu by clicking on Fitting Data Manager.

2 Data Manager Overview

The Fitting Data Manager allows you to browse and modify the content of a new or an existing fitting training set. As shown below, the main Data Manager component is a table of structures, with descriptive information displayed above the tabular view:

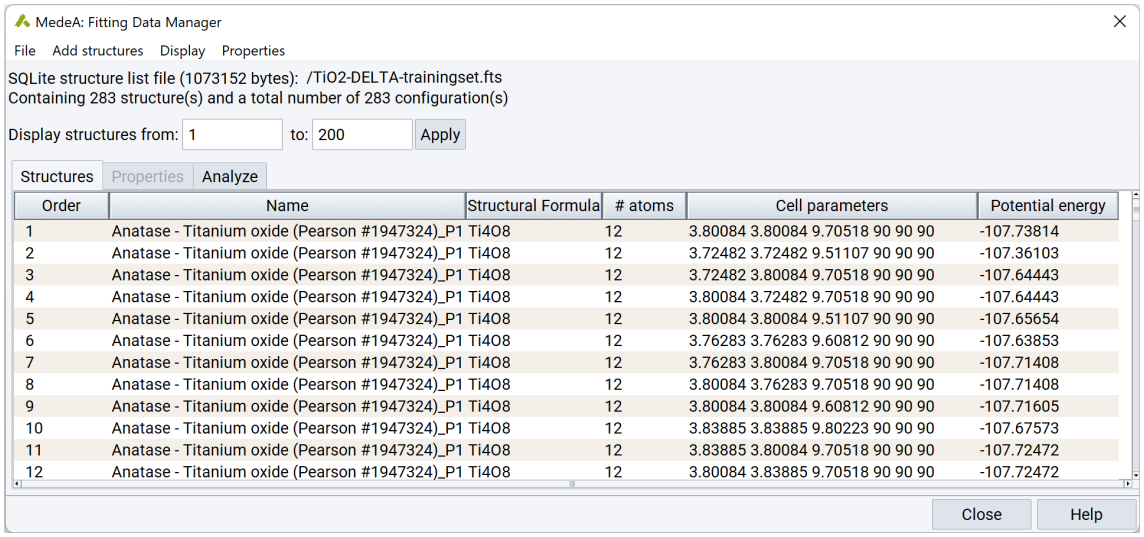
The descriptive part shows the file format, its size, and the full path. One can also control the number of displayed structures in the table, specifying the start and end index (clicking on the Apply button will update the view).

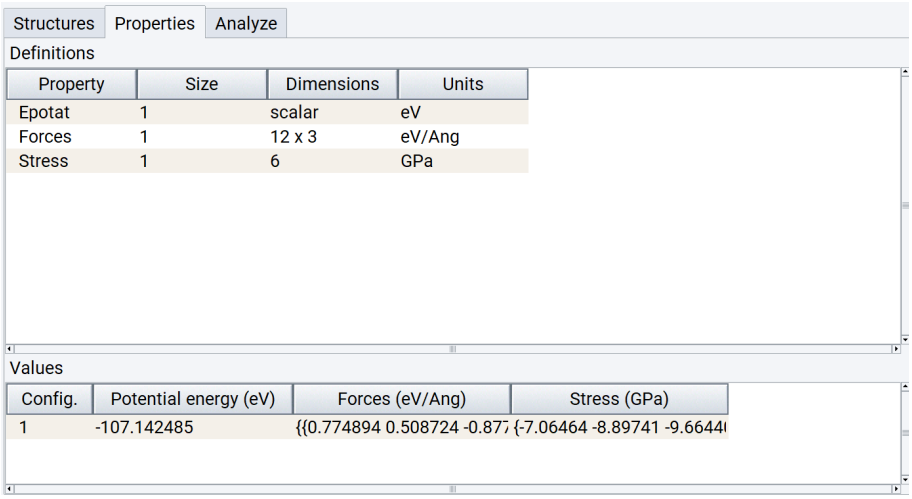
The Structures tab shows an overview of all structures in the structure list opened. The Properties tab shows all properties in a table for a given structure. It can be activated by selecting a structure in the Structures tab, right-clicking, and selecting Show structure # properties from the pull down menu. The property is shown:

The Analyze tab provides a graphical overview of the data in the fitting training set.

3 File Formats

A fitting training set file has one of the following two formats:





- Text format : data is written in plain ASCII format. One can edit the content with a simple text file editor but with a high risk of erroneous modification. Many molecular dynamics jobs will produce a trajectory file in that format (*Trajectory.dat* or files with *.traj* extension). One cannot apply all available operations with that format.
- SQLite format : information is organized with an SQLite database format, which offers a much higher level of performance when processing and updating lists containing a large number of structures.

By default, the SQLite format is used when a new structure list is created. It is recommended to convert to SQLite whenever possible, which can be done with the File >> Convert to SQLite/text format menu item. It will create a file of the other type than the one displayed.

4 Menus

The Fitting Data Manager has its own menu bar. The following menu items are present.

4.1 File

In the main window's File menu, the following actions can be invoked:

- New data set : will ask for the name and location to store the new empty fitting training set file. The file will be initialized allowing you to start adding structures.

- Open data set from disk : will ask to select an existing fitting training set file on the local file system and display its content.
- Open data set from job : will ask to select an existing fitting training set file on the active JobServer after which the storage location on the local file system can be selected. The content is then displayed.
- Convert to SQLite/Text format : Convert structure list to SQLite or text format.
- Convert symmetry to P1 : will ask for the name and location of a new structure list file and place a copy of all structures of the current list converted to P1 symmetry in this structure list. Not all data will be copied since some data, such as forces, would require a transformation.
- Close : Close the Fitting Data Manager.
- Export to CVS : Export the data to a CVS file.

4.2 Add Structure(s)

There are two ways to add data to the current training set. These data can be in the form of a trajectory, a structure list containing the relevant properties, which are, energies, forces, and stresses, and/or an existing fitting training set.

These two options are accessible from the Add structures menu:

From file : opens a window to select files on the local disk drive.

From job : opens a dialog to select files from the active JobServer.

When importing a structure list, fitting training set or a trajectory containing multiple entries a prompt will appear asking if the import should step over *n* configurations. In the case of a trajectory the user has, in addition, the option to import all the trajectory frames as separate structure entries by toggling the option Load configurations as structures .

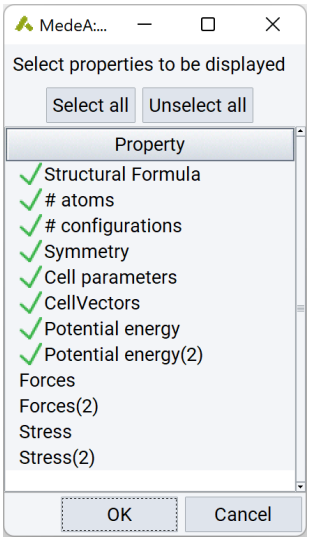
Note: A structure list containing the relevant properties, *i.e.*, energies, forces and stresses, can be generated in a flowchart using the **Save to List** stage after a *VASP* stage and ticking in the **Save to List** stage the option **Including FF fitting data**.

4.3 Display

The information appearing in the structure table can be controlled by the Display menu selectors. The columns Order and Name will always be displayed.

The information displayed can be the structural information (Structural Formula , #atoms : number of atoms, #configurations : number of configurations, Symmetry and Cell parameters for systems with periodic boundary conditions) or potential energies, forces and stresses. If a large number of properties is defined, only the first three will be displayed by default, others have to be selected from the Display menu to appear.

In case there are several configurations for a structure, the value shown in the table will correspond to the first configuration. Other configurations property values can be seen by choosing to show the structure properties in the Properties table.



5 Data Manager Table

5.1 Ordering and Numbering Structures

Entries are stored in a given order, which appears in the Order column of the table. It is possible to change the order by sorting rows (structures) according to a column from the context menu commands Sort Ascending or Sort Descending, on a column title. A given entry can be moved up or down by a right-click on its row with the corresponding command in the popup menu.

If the initial order of the rows in the table is changed by one of these methods, the actual order in the fitting training set file remains unchanged. It is possible to apply the new displayed order to update the internal ordering with the right-click popup menu command Save the order of the rows. This can be useful, for example, to change the processing order in a job.

5.2 Context Menu

Pressing the right mouse button on a table row in the **Structures** tab opens a context menu. This context menu contains the following items:

- Move up : Move structure one row up.
- Move down : Move structure one row down.
- View structure n : Open structure in row n in a viewer.
- Show structure n properties : Show properties of structure in row n in *Properties* tab.
- Export structure n : Export structure in row n to an individual structure file (*.sci* extension).
- Rename structure n : Rename the structure in row n.
- Delete selected structure(s) : Delete the structures in all selected rows.
- Save selected structure(s) to MD database : Save the structures in all selected rows in the Materials Design database.
- View selected structure(s) first configuration : Open viewers with the first configuration of all selected rows.
- Save all structure(s) to MD database : Save all structures at once in the Materials Design database.

- Save the order of the rows : Update the internal ordering with the displayed order
- Export all structures : Export all structures at once to individual structure files (*.sci* extension). A file name prefix is used and the structure number and configuration (if more than one) are used in the final file name.

6 Analyze

The Analyze tab contains functionality to provide an overview of the ranges of some properties of the structures in the list. It shows a graphing area and some check boxes which are based on the data available for the structures. If present the first check box can be used to display the distribution of energies. The other check boxes create distributions of all bond lengths, bond angles and torsion angles per element. Multiple distributions can be displayed simultaneously by selecting multiple check boxes. The graphs are created by pressing the Plot button. The calculation of distributions for large structure lists can be time consuming.