

Tutorial 1: Uploading molecular dynamics data and examining the metadata

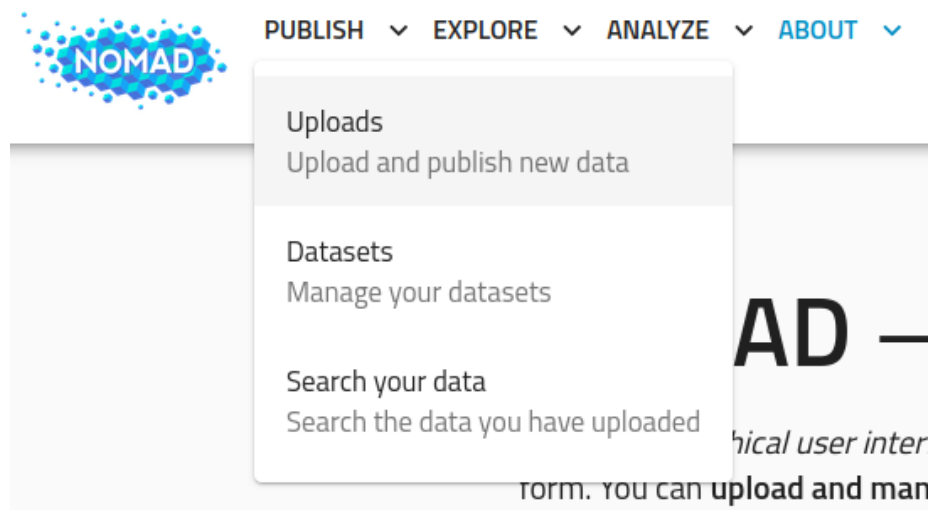
NOMAD currently supports 2 molecular dynamics codes: Gromacs and LAMMPS. Here, we will walk you through uploading data from a series of simulations run in Gromacs.

In the current directory (`Tutorial-1_Uploading_MD_Data/`), unzip the file `data/water_workflow.zip`. Take a minute to explore the directory structure. If you are familiar with Gromacs you will immediately see the input/output from 3 simulations: an energy minimization (`Emin/`), an NPT equilibration (`Equil-NPT/`), and an NVT production run (`Prod-NVT/`). In the main directory, you will also see a `.yaml` file, which contains the NOMAD schema for connecting these 3 simulations into a workflow. This will be covered in Tutorial 2.

In your browser, go to the NOMAD home page. At the bottom of the page, click on the NOMAD beta button:



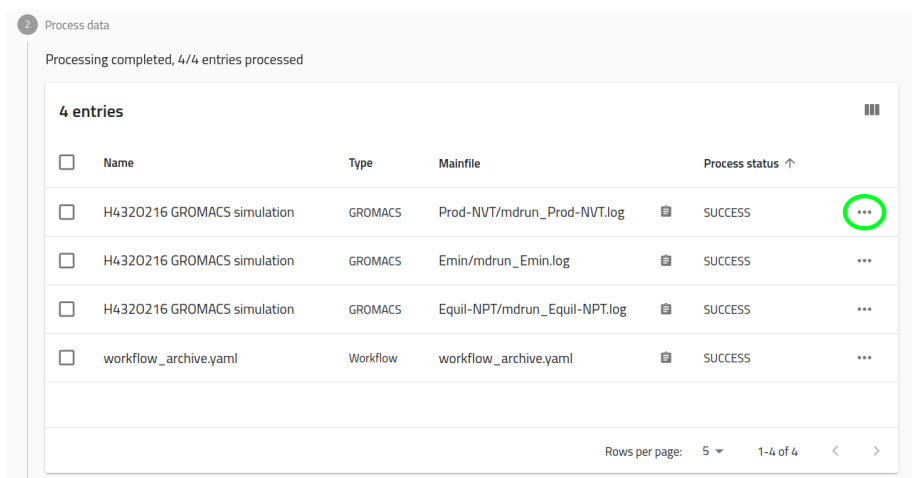
Use the button on the top right to login or, if you are a new user, register for a NOMAD account. Once you are logged in, at the top left of the page, go to **PUBLISH** → **UPLOADS**:



Click **CREATE A NEW UPLOAD**, and then you can either drop the `water_workflow.zip` file directly or click **CLICK OR DROP FILES** to find the file within your file system. Note that there is currently a 32GB file size limit

per upload. This will likely be expanded in the future, and exceptions can be received by contacting the NOMAD team. NOMAD's approach to the large data sets generated by molecular dynamics simulations is still under development. A deeper discussion of this topic is beyond the scope of this initial tutorial series, but will be addressed in the future.

At this point, NOMAD will unzip your upload and automatically detect all possible "main" files corresponding to output from supported codes (in this case, the `.log` file from each Gromacs simulation). NOMAD will then proceed to parse all the relevant simulation files and store the simulation data and *metadata* within the NOMAD *MetaInfo* schema. In this case, the parsing should take ~ 30 seconds. You should now see the successfully processed data overview:



2 Process data

Processing completed, 4/4 entries processed

4 entries

<input type="checkbox"/>	Name	Type	Mainfile	Process status ↑
<input type="checkbox"/>	H4320216 GROMACS simulation	GROMACS	Prod-NVT/mdrun_Prod-NVT.log	SUCCESS
<input type="checkbox"/>	H4320216 GROMACS simulation	GROMACS	Emin/mdrun_Emin.log	SUCCESS
<input type="checkbox"/>	H4320216 GROMACS simulation	GROMACS	Equil-NPT/mdrun_Equil-NPT.log	SUCCESS
<input type="checkbox"/>	workflow_archive.yaml	Workflow	workflow_archive.yaml	SUCCESS

Rows per page: 5 1-4 of 4

Let's examine the production simulation by clicking the 3 dots to the right of the entry labeled `Prod-NVT/mdrun_Prod-NVT.log`, circled in green in the above image. You will now be on the **OVERVIEW** page for this entry, which aims to provide a simple description of this entry through visualizations of the system itself, some key observables, and some of the overarching metadata. The **OVERVIEW** page will be examined in detail in Tutorial 2. For now, we will focus on how the uploaded data is stored within the NOMAD repository. In addition to the **OVERVIEW** tab, there are 3 other tabs at the top of the page: **FILES**, **DATA**, and **LOGS**.

Click on the **FILES** tab. Here you will find all the raw data that was uploaded via the `.zip` file, retained within the original file system structure. The raw files are stored in the repository and can be downloaded at any time.

Now click on the **LOGS** tab. Here you will find some technical information about the data processing along with any warnings or errors that were raised by the NOMAD software.

Finally, click on the **DATA** tab. Here you can navigate through the NOMAD

Metainfo for this entry, i.e., the processed and normalized version of the simulation data and metadata.

Navigate to section **run** → **program**, where we find some basic information about the simulation code (name and version):

The screenshot shows a three-pane interface. The left pane is titled 'Entry' and lists sub-sections: results, metadata, workflow2, workflow, run (highlighted), and REFERENCED BY (closed). The middle pane is titled 'Run' and lists sub-sections: program (highlighted), time_run, method, system (expanded to show a list of numbers 0-10), and calculation. The right pane is titled 'Program' and shows quantities: name = GROMACS and version = 5.0.8-dev-20160207-b53b2f6, along with a REFERENCED BY (closed) link.

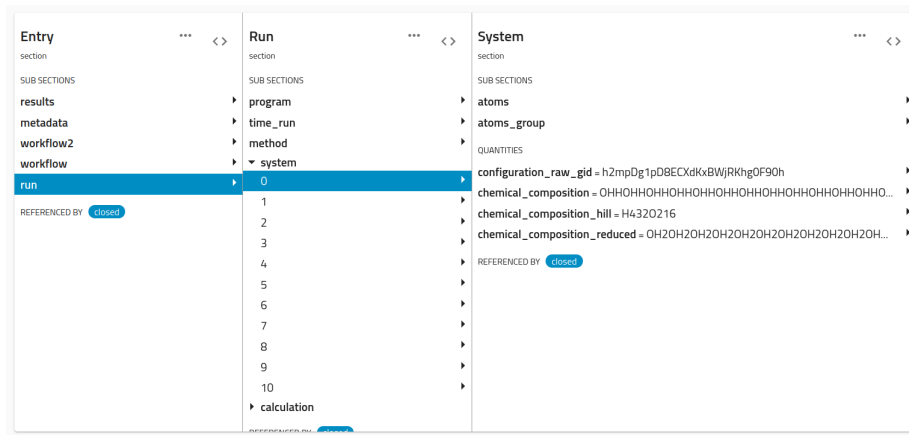
Now navigate to section **run** → **method**:

The screenshot shows a three-pane interface. The left pane is titled 'Entry' and lists sub-sections: results, metadata, workflow2, workflow, run (highlighted), and REFERENCED BY (closed). The middle pane is titled 'Run' and lists sub-sections: program, time_run, method (highlighted), system, and calculation. The right pane is titled 'Method' and lists sub-sections: force_field and atom_parameters (highlighted), along with a REFERENCED BY (closed) link.

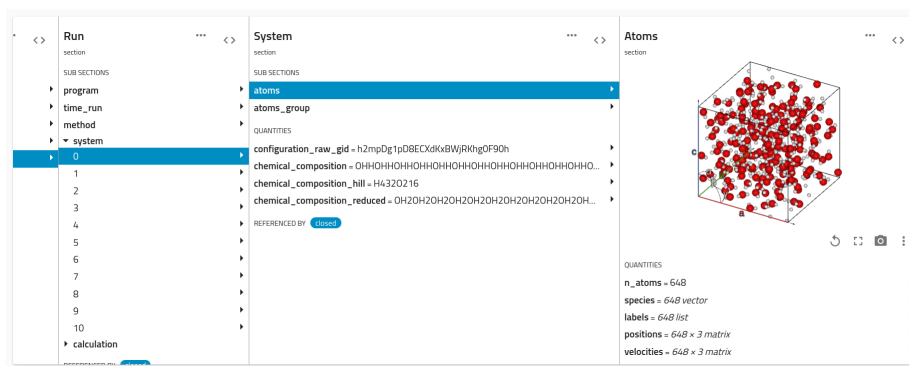
Under **atom_parameters** you will find a list of force-field-dependent atom attributes such as mass and charge. Under **force_field** → **model** → **contribu-**

tions you will find a list of intramolecular interactions defined for this simulation. The development of metadata for a more comprehensive storage of force fields in NOMAD is currently underway. Under **force_field** → **force_calculations** you will find some basic input parameters dealing with the calculation of forces in this simulation.

Navigate to section **run** → **system** → **0**:



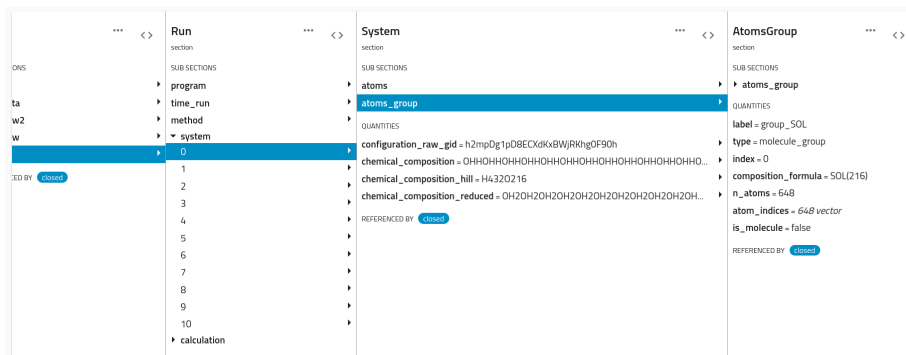
The **system** section holds the configurational information from this entry. The first item in this section, **0**, corresponds to the first saved configuration within the simulation trajectory. Now click on the atoms section:



There will be a prompt to ask if you would like to visualize the system. If you click yes, the particles within the simulation box will be displayed. (The visualizer can also be accessed with additional features in the **OVERVIEW** page). In the **atoms** section, various atom attributes are stored, e.g., the positions and velocities of each atom for this frame. Notice that the simulation trajectory is stored as a list of repeating dictionaries, as opposed to the typical dictionary of lists that may be found in trajectory analysis software, such as *MDAnalysis*. In short, this is done for consistency with data from other methods that are

stored in the NOMAD repository. However, when working with the data from a particular entry, there are tools in NOMAD to easily convert the *MetaInfo* to a more convenient format for analysis. More about this in Tutorial 3.

Now click on the **atoms_group** section:



This section holds a hierarchical organization of the system (denoted in the following as the *topology*), based on the *bonds* defined in the force field used for this simulation. By default, the first level of this hierarchy contains *molecule groups*, which group together all molecules of the same type. Now, by clicking the **atoms_group** subsection of this group, you reach the next level down, which contains each molecule within the current molecule group. Similarly to these molecule groups and molecule levels, monomer groups and monomer levels will be defined for polymer systems. The individual atoms within each molecule are not explicitly stored within this hierarchy, but are simply referenced via their indices within the **atoms** section. Take a few minutes to examine this hierarchy and the stored quantities in more detail. Note that at the moment the hierarchy is only stored in the first entry of **system**.

Now go back and navigate to section **run** → **calculation** → **0**:

Entry	Run	Calculation
section	section	section
SUB SECTIONS	SUB SECTIONS	SUB SECTIONS
results	program	energy
metadata	time_run	forces
workflow2	method	QUANTITIES
workflow	system	system_ref = reference ...
run	calculation	method_ref = reference ...
REFERENCED BY closed	0	pressure = -0.00765 GPa
	1	temperature = 304.94113 K
	2	step = 0
	3	time = 0 fs
	4	REFERENCED BY closed
	5	
	6	
	7	
	8	
	9	
	10	
	REFERENCED BY closed	

The **calculation** section contains any saved thermodynamic quantities that are a function of a single configuration, e.g., energy, pressure, temperature, etc., as well as any saved force information for the atoms within each configuration.

Exercises

1. What are the oxygen and hydrogen atom types used in the force field for this simulation?
2. What is the step number of the last saved configuration of this simulation? What is the corresponding time for this configuration?
3. (CHALLENGE) Which thermostat is used for temperature coupling in this simulation? What is the frequency of temperature coupling?