Plan Overview

A Data Management Plan created using DMPonline

Title: Advanced Computer Simulations for Material and Biomolecular Science

Creator: Alexander Lyubartsev

Principal Investigator: Alexander Lyubartsev

Data Manager: Alexander Lyubartsev, Fredrik Grote, Mikhail Ivanov, Maxim Posysoev

Contributor: Fredrik Grote, Mikhail Ivanov, Maxim Posysoev

Affiliation: Stockholm University

Funder: Swedish Research Council

Template: Swedish Research Council Template

ORCID iD: 0000-0002-9390-5719

Project abstract:

The overall aim of the project is to address to several challenges in the modern molecular modeling having primary application area in biomolecular and nanomaterials scienc:

- development of the multiscale modelling methodology in which models operating on a longer length- and time- scale (up to micrometers-milliseconds) are deduced from atomistic and ab-initio quantum-chemical simulations using combination of physics-based and machine learning approaches.
- development of force fields for atomistic simulations of metal oxide surfaces and nanoparticles in contact with aqueous media and biomolecules, based on results of ab-initio simulations.
- development of advanced sampling techniques to study phase transformations
- addressing to several actual problems of biomolecular and material science in order to validate the developed methodologies and to demonstrate their power by making substantial progress in our understanding of the phenomena: interaction of inorganic nanoparticles with biomolecules to understand molecular origin of possible toxic effects of nanoparticles; folding of DNA in chromatin, prediction of polymorphic forms of drugs and their stability

ID: 92999

Start date: 01-01-2022

End date: 31-12-2025

Last modified: 25-01-2023

Grant number / URL: 2021-04474

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Advanced Computer Simulations for Material and Biomolecular Science

General Information

Project Title

Advanced Computer Simulations for Material and Biomolecular Science

Project Leader

Alexander Lyubartsev

Registration number/corresponding, date and version of the data management plan

v 1.0

Version

v. 1.0

Date

2022-01-31

Description of data – reuse of existing data and/or production of new data

How will data be collected, created or reused?

The overall project is divided into a number of subprojects (further referenced as "projects") each corresponding to one planned publication. In certain cases 2 or more related publications can be united into a single project. All the data related to a project will be collected in a specific directory (folder). After finishing the project the data will be curated and transferred to the archive.

Data to be collected and archived describe molecular computer simulations / computations carried out within the project.

Data sets include the following components:

- input data describing simulated systems and algorithms
- data produced during the simulations (raw data)

- data produced during analysis of simulations (final data)
- eventually: developed own software, codes or scripts.

Input data: produced manually and/or with the use of software (Avogadro, antechamber, etc)

Raw data: produced by simulation engines (e.g., Gromacs, CP2K)

Final data: produced by the analysing software

The data will be collected during the project, and will be kept for possibility of further reuse. Exception is molecular dynamics trajectories which will be archived selectively.

What types of data will be created and/or collected, in terms of data format and amount/volume of data?

I. Input data:

1) Data describing simulated system (composition; chemical structure of the components; force field)

Example for atomistic molecular dynamics simulations with Gromacs: .itp files describing chemical structure of the simulated molecules; .top file describing system composition; .itp files describing force field; .gro file describing initial system configuration

2) Files describing methodology and workflow:

Workflow is described as one of (or combination of) following:

- Jupyther Notebooks
- README or other text files
- MODA template
- Files with simulation parameters (depending on the software). For Gromacs: .mdp files for each step in the workflow

Input data will be saved and archived, in order to have full control of the parameters and methodology used in the simulations. Input data should contain all the information necessary to reproduce the whole simulation

Data format: ASCII text files

Data volume: minor, within 1 Gb

II. Raw data produced during the simulations

Data produced during simulations may include:

- molecular dynamics trajectories and snapshots
- log files
- electron density / wave functions information from ab-initio simulations/computations
- other auxilary information (e.g., .edr files from Gromacs, COLVAR from Metadynamics, etc)

The data will be saved for the duration of the project. Archiving of very large files (e.g, molecular dynamics simulation trajectories) will be made selectively, by assessing åpossibility or need of further reuse.

Data format: format of the used software

Data volume: up to 100 TB

III Final postprocessed data

Data produced by the analysis of the raw data from the simulations. This particularly includes all the data presented in Figures and Tables of the publications based on the calculations/simulations within the project.

Data format: text based files; Open Document spreadshits (.ods)

Data volume: within 1 TB

IV Self-produced software / utilities /scripts used to prepare input data, make computations and analysis

Data format: text files - software codes
Data volume: within 1 Gb

Documentation and data quality

How will the material be documented and described, with associated metadata relating to structure, standards and format for descriptions of the content, collection method, etc.?

The data will be documented in the following way:
- Original paper or manuscript with eventual supporting information describing research and methodology in scientific terms.
- README file in the top directory of each project describing main purposes of the study, the data structure, folders and file names, types of data . The file can be written in .md format with possibility for further referencing.
- Jupyther Notebooks (.jpynb) describing the sequence of actions / commands during simulations set-up, run and post-processing.

We will follow development of the accepted standards to describe simulation and modeling metadata and provide corresponding files, e.g in the format of MODA templates.

How will data quality be safeguarded and documented (for example repeated measurements, validation of data input, etc.)?

The data will be subject to control for consistency and reproducible before the publication. For example, control of the content of input file and reported simulation parameters in the log file. Furthermore, Jupyther notebook will be used both as a mean of documentation and as a tool for reproduciblity control.

Storage and backup

How is storage and backup of data and metadata safeguarded during the research process?

During the research all the working data including metadata will be backuped on local external hard disks, and later on the institutional large storage facility (to be in operation after summer 2022).

How is data security and controlled access to data safeguarded, in relation to the handling of sensitive data and personal data, for example?

The data are secured by usual user authentication. Since the data do not handle sensitive / personal

information, no additional security control is required.

Legal and ethical aspects

How is data handling according to legal requirements safeguarded, e.g. in terms of handling of personal data, confidentiality and intellectual property rights?

The data do not handle sensitive / personal information.
Copyrights and IPR for data belongs to the researches participating in acquiring and collection of the data, who are also included as authors of the manuscript.
The data are supposed to become open after publication of the paper according to one of standard open licenses, such as Creative Commons.

How is correct data handling according to ethical aspects safeguarded?

N/A

Accessibility and long-term storage

How, when and where will research data or information about data (metadata) be made accessible? Are there any conditions, embargoes and limitations on the access to and reuse of data to be considered?

The data will become open after publication of the paper based on the performed research

In what way is long-term storage safeguarded, and by whom? How will the selection of data for long-term storage be made?

The data will be archived at the institutional large storage facility. After publication of each paper, relevant data will be collected in a folder, complemented with necessary metadata and uploaded to the storage server. Reference to the data will be posted to the Web page of the research group.

Will specific systems, software, source code or other types of services be necessary in order to understand, partake of or use/analyse data in the long term?

All the metadata and final data will be saved in open, generally available formats (text; Open Document Format). Raw data will be saved in format of the used simulation / modeling software (e.g, Gromacs, CP2K, LAMMPS).

How will the use of unique and persistent identifiers, such as a Digital Object Identifier (DOI), be safeguarded?

DOI identifier of the published paper will be used

Responsibility and resources

Who is responsible for data management and (possibly) supports the work with this while the research project is in progress? Who is responsible for data management, ongoing management and long-term storage after the research project has ended?

The first author of the publication is responsible for collecting and organazing data in a folder for archiving.
The PI is responsible for long-term data management and access to the storage.

What resources (costs, labour input or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)? What resources will be needed to ensure that data fulfil the FAIR principles?

The PI and the first author of a publication are responsible to ensure that data fulfill FAIR principles. Resources for backup, access and long-term storage will be provided by the Department after setting up the storage facility