Towards a general approach to model ion-molecule reactions at low temperature

A Data Management Plan created using DMPonline.be

Creators: Jenne Van Veerdeghem https://orcid.org/0000-0001-7753-8931, n.n. n.n.

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Jenne Van Veerdeghem https://orcid.org/0000-0001-7753-8931

Grant number / URL: 11PHB24N

ID: 204532

Start date: 01-11-2023

End date: 31-10-2027

Project abstract:

Ion-molecule reactions play a key role in the chemistry of many environments, from interstellar space to industrial plasmas. The corresponding rate coefficients are often calculated using theories based on capture models, but these are known to be inadequate in many instances. In addition, they do not allow the calculation of state-to-state rate coefficients (that are resolved in the ro-vibrational states of the reactants and the products), which is crucial to model environments that are not at local thermodynamic equilibrium, such as those that commonly arise in astrophysical media. In this project we will develop a general approach that relies on machine learning tools to construct multidimensional potential energy surfaces combined with quasi-classical trajectory calculations to study the reaction dynamics. A major aspect of this project will be the collaboration with experimentalists at the University of Liverpool, who have recently started investigating ion-molecule reactions below room temperature, which led to some unexpected findings.

This project will lead to new insights into the underlying mechanisms of ion-molecule reactions at low temperature, with direct applications to astrochemistry and to the interpretation of state-of-the-art experiments.

Last modified: 03-04-2024

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Generate new data

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Ab initio calculation input	Input files for quantum chemical calculations performed using e.g. Molpro.	Generate new data	Digital	Other	.inp	<1GB	
Ab initio calculation results	Results from quantum chemical calculations performed using e.g. Molpro. Raw datasets will be stored in .csv files. Processed datasets will be stored in .pkl files.	Generate new data	Digital	Simulation data	.out .csv .pkl	<1TB	
Machine learning model	Machine learning models created using software like PyTorch which are stored in pickle files.	Generate new data	Digital	Simulation data	.pth .pkl	<1TB	

Serialized machine learning model	Serialized machine learning models allow for OS independent operation and deployment	Generate new data	Digital	Simulation data	.pt	<1 GB	
Results from dynamics calculations		Generate new data	Digital	Simulation data	.csv	<1TB	
Scripts	Scripts to train ML models, generate input files, preprocess data, visualize data,...	Generate new data	Digital	Software	.py .f or f.95 or .f03 .sh	<1GB	
Jupyter Notebook	Notebooks containing results and explanations.	Generate new data	Digital	Software	.ipynb	<1GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

NA

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

The directory structure will be layed out in such a way to understand the origin of the datasets, different models, and type of code it contains.

The project's python code and scripts will be stored in a Github repository which is accompanied by an environment.yml file describing the software environment used for the project. This allows for exact recreation of the processed datasets from the raw datasets as well as creation of the machine learning models.

Both Python and Fortran code will be described by in-file comments and documentation to explain the purpose and use of the code.

Machine Learning models are saved in pickle (.pkl or .pth) files which save the object state directly.

Imcluded in the pickle files besided the model, are the dataset the model is trained on, the model parameters, model definition, the validation dataset, and validation results.

The model file name will include the Git version number of the repository at the time of creation and describe the model parameters.

Processed datasets from *ab initio* calculations are stored in pickle files as well. Included are the processed dataset, the raw dataset, the preprocessing pipeline used, information on units, and the Git version number of the repository at the time of creation.

Results from dynamics calculations will be accompanied by README.md files to describe the dataset.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

Datasets will be uploaded to the Zenodo repository which uses the DataCite metadata schema.

3. Data storage & back-up during the research project

Where will the data be stored?

Scripts and code are stored locally, on the Github servers, and on a personal OneDrive via KU Leuven Microsoft 365.

Small datasets such as those containing ab initio calculation results are also stored locally and on the Github servers, and on a personal OneDrive via KU Leuven Microsoft 365.

Machine learning models are too large to upload to Github. Instead these data are stored locally and on a personal OneDrive via KU Leuven Microsoft 365.

Ab initio files are kept on the HPC cluster owned and operated by the Quantum Chemistry division, which receives regular backups for all data in the home directory.

How will the data be backed up?

The Git repositories are backed up every three days to the personal OneDrive via KU Leuven Microsoft 365.
The same data will be backed up monthly on an external hard drive.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

Storage and backup is currently sufficient.
If the complexity of the machine learning models and the size of the datasets increases, this might not be the case in the future.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Local data storage is password protected.
Data on OneDrive are secured by KU Leuven two-factor authentication.
Data on the HPC cluster are secured by an SSH keypair.
The external hard drive is kept in a locked drawer.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The costs for backup on the HPC cluster are covered by the PhD promotor.
If other costs arise, they will be covered by the FWO bench fee of the principal investigator

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

Data that will be kept are machine learning models and final processed datasets, as well as the raw data it was created from, and all code and scripts that are at the basis of publications and the PhD thesis will be kept.

Where will these data be archived (stored and curated for the long-term)?

For long term storage, data will be kept in the KU Leuven RDR and ManGO.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

The costs for backup on the HPC cluster and ManGO are covered by the PhD promotor.
If other costs arise they will be covered by the PhD promotor.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

Of the ML models that are used for publication, their training sets will be made available for reuse.

If access is restricted, please specify who will be able to access the data and under what conditions.

NA

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

The data will be made available in Zenodo or KU Leuven RDR.

When will the data be made available?

After the end of the PhD, currently planned on 2027/10/31, or upon publication of research results.

Which data usage licenses are you going to provide? If none, please explain why.

Dynamics code written for the project will be implemented in currently unpublished code. The license will be discussed with the original author upon publication of the code.
Training datasets will be published as Public Domain Mark (PD).

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

What are the expected costs for data sharing? How will these costs be covered?

There are no expected costs at the moment.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Jenne Van Veerdeghem

Who will manage data storage and backup during the research project?

Jenne Van Veerdeghem

Who will manage data preservation and sharing?

Jérôme Loreau - Project supervisor

Who will update and implement this DMP?

Jenne Van Veerdeghem