SolProp-mix: Predicting the solubility curve in solvent mixtures

A Data Management Plan created using DMPonline.be

Creator: Florence Vermeire

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: G021924N

ID: 208948

Start date: 01-01-2024

End date: 31-12-2027

Project abstract:

As chemical engineers our goal is to find ways to make processes safer, greener, and cheaper. One ongoing challenge we face is figuring out how well different chemicals dissolve in various solvents. Experimental data on the solubility for pharmaceuticals in organic solvents is scarce and current models lack predictive capabilities. Fast predictions would help us to choose the best solvent for a given application and make process development more cost-efficient. Recently, we developed a new software tool called SolProp that is the first to predict the solubility curve of pharmaceuticals for a broad range of organic solvents. We addressed this challenge by using a hybrid model that combines machine learning and rigid thermodynamics. With this project, we aim to expand SolProp's capabilities to predict how well pharmaceuticals dissolve in solvent mixtures.

In the pharmaceutical industry, solvent mixtures are often used to adjust solubility. It would be a huge breakthrough to be able to predict and optimize the composition of solvent mixtures to reduce energy usage, emissions, and waste streams. The machine learning architecture of SolProp will be extended to embed mixtures of molecules in an invariant, smooth, and consistent manner, thereby innovating chemical property prediction with machine learning to a whole new range of applications. The model will be validated using experimental solubility data, and data curation will focus on characterizing uncertainty.

Last modified: 03-07-2024

SolProp-mix: Predicting the solubility curve in solvent mixtures
FWO DMP (Flemish Standard DMP)

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.

				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	
SolProp-mix-exp	solubility and solvation free energies of solutes in solvent mixtures	reused	digital	compiled	csv (or txt)	<100 MB	
SolProp-mix-QM	QM calculations for solvation free energies	new	digital	simulation data	.pickle (pandas dataframe)	<100 GB	
ML models	trained machine learning models	new	digital	simulation data	.pt	<100 GB	
SolProp-mix	python code to train models and make predictions	reused with adaptions	digital	software	.py	<100 MB	

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:

Existing data will be compiled from different literature sources to create the experimental database. Regarding solubility measurements, important sources are those found for solubility in pure solvents as compiled in the SolProp database. https://doi.org/10.1021/jacs.2c01768
The software developed for this project will continue to work on the SolProp_ML software. https://github.com/fhvermei/SolProp_ML

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

Datasets do not have potential towards valorisation, however the software and the machine learning models trained could be used in follow-up projects at higher TRL.

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.

- Yes

The COSMOtherm software license could limit the publication of data calculated with the software in the SolProp-mix-QM database. However, the terms of the agreement are unclear and currently they seem to allow publication of newly generated data as is the case for this project.

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).

README.txt files will be added to the dataset to describe the columns. In addition a thorough explanation on data collection and curation will be available in the manuscript.
The python code will be documented according to standard python documentation, while changes to the code will be tracked through git.

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

The databases will be stored according to the format of common property prediction software. For the molecular representation, we will make use of SMILES and InChI format as defined by IUPAC. The data will next be made available on Zenodo (or KU Leuven RDR) with a doi and version assigned to the database.

The trained machine learning models will be stored as .pt files, as proposed by PyTorch.
The software developed for this project will be maintained on KULeuven GitLab to keep track of changes and versions.

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

Where will the data be stored?

During the research project, data will be stored on the OneDrive of the researchers. After the research project, a back-up of all data will be stored on the Vermeire group MSTeams channel.

How will the data be backed up?

On the MSTeams channel of the Vermeire group.

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

5TB on the MS Teams of the Vermeire group. Only ±10 GB will be stored for this project

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

MS Teams is secured by KU Leuven due factor authentification.

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

NA.

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...).

The final databases will be stored.
The intermediate calculations for the QM database will be stored.
The software and trained ML models will be stored and maintained through further projects.

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

MS Teams of the Vermeire group.

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

NA

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

Databases and trained ML models will be made available through Zenodo (or KU Leuven RDR).
Software will be made available through KU Leuven GitLab

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

Question not answered.

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.

see above

When will the data be made available?

After publication of research results

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

MIT license

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

DOI generated by e.g. Zenodo

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

NA

6. Responsibilities

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

PhD researcher assigned to the project, Simona Buzzi

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

PhD researcher assigned to the project

Who will manage data preservation and sharing?

prof. Florence Vermeire, supervisor of the project

Who will update and implement this DMP?

prof. Florence Vermeire, supervisor of the project