



# Physical Sciences Data Infrastructure (PSDI)

Physical Chemistry Properties Data Collection  
PSDI Report

Author: Matthew Partridge, Jeremy Frey

Report Date: 13/05/2025

PSDI-Dataset-Series:Report\_001

## Publishing Information

Physical Chemistry Properties Data Collection  
PSDI-Dataset-Series:Report\_001  
Report Date: 13/05/2025  
DOI: 10.5281/zenodo.15656342  
Published by University of Southampton

## Funding Information

### Physical Sciences Data Infrastructure

PSDI acknowledges the funding support by the EPSRC grants EP/X032701/1, EP/X032663/1 and EP/W032252/1

Title: *Physical Sciences Data Infrastructure Phase 1b EP/X032701/1*

Principal Investigator: *Professor Simon Coles*

Other Investigator: *Professor Jeremy Frey*

Co-Investigators: *Dr Nicola Knight & Dr Samantha Kanza*

Title: *PSDI Phase 1b EP/X032663/1*

Principal Investigator: *Dr Juan Bicarregui*

Other Investigator: *Dr Brian Mathews, Dr Vasily Bunakov, Dr Barbara Montanari*

Co-Investigators: *Dr Abraham Nieva de la Hidalga*

## Project Details

Project Name	Physical Chemistry Properties Data Collection
Project Dates	24/06/2024-30/04/2025
Website	<a href="#">PSDI Website</a>

## Project Team

Please list all members of the project team as Name, Affiliation, ORCID

Author Name	Affiliation	ORCID
Matthew Partridge	University of Southampton	<a href="#">0000-0001-5280-8309</a>
William Poole	University of Southampton	<a href="#">0009-0003-2441-8794</a>
Samuel Munday	University of Southampton	<a href="#">0000-0001-5404-6934</a>
Ashley Unitt	University of Southampton	<a href="#">0009-0007-0037-0035</a>
Thomas Allam	University of Southampton	<a href="#">0009-0009-9897-333X</a>
Joshua Cheung	University of Southampton	<a href="#">0009-0003-9952-3468</a>
Joanna Grundy	University of Southampton	<a href="#">0000-0003-2583-5680</a>
Jeremy Frey	University of Southampton	<a href="#">0000-0003-0842-4302</a>

## Project Description

Constructing and aggregating data collections is a key underpinning service for research, powering discovery, understanding and prediction. However, the majority of data measured in practice does not get incorporated into any form of usable collection. This pathfinder explored and developed methods to build, store, manage and access collections for the different types of data, such as institutional data deposited by numerous groups into a single repository, facilities data collected at beam lines on a range of different types of samples, legacy data extracted from papers and proprietary digital sources, and orphaned data such as collections generated for a specific one-off purpose. Spectroscopy is a key technique in environmental, life and physical sciences providing not only characterisation of samples but also in-situ analysis of dynamic systems and will provide the testbed for much of this work. This work was conducted with the University of Cambridge and Imperial College and developed standards, based on IUPAC FAIRSpec, and infrastructure components to drive tools and processes to capture, manage, analyse and reuse spectroscopic data from across the sciences.

## Project Data & Materials

As part of this project, we have compiled a structured data collection from a variety of literature sources. While the original data sources themselves are **not available** due to rights restrictions.

The data collection includes physical chemistry properties such as melting points, boiling points, aqueous solubility (LogS), Henry's Law constants, and miscibility. All data types are defined in accordance with IUPAC Goldbook standards. The sources used to generate these datasets include:

- **NSRDS-NBS 36** – Critical Micelle Concentrations of Aqueous Surfactant Systems (1966)
- **BioQuest** – Experimental melting point and boiling point data
- **Bergstrom 2003** – Solubility data from oral drug bioavailability studies
- **DDB 2023** – Data extracted from the Dortmund Data Bank
- **IUPAC** – Solubility data from IUPAC evaluated sources
- **Meng 2022** – Solubility data compiled from recent literature
- **Delaney 2004** – Data derived from aqueous solubility prediction studies
- **Sander 2023** – Henry’s Law constant values from a critically evaluated dataset

From these sources, the following datasets have been created

- Compounds with both melting point and boiling point data
- Compounds with LogS (aqueous solubility) data
- Compounds with both solubility and miscibility data
- Compounds without canonical names
- Compounds with multiple solubility records in the same solvent where entries differ by more than 0.5 log units
- Compounds with either LogS or Henry’s Law Constant data
- Compounds with both LogS and Henry’s Law Constant data

These structured datasets are suitable for downstream analysis, modelling, and data-driven research applications in chemical informatics.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Methodology</b>	<b>2</b>
2.1	Digitisation of core datasets . . . . .	2
2.2	Data auditing . . . . .	2
2.3	Development of the data collection . . . . .	4
2.4	Data ingestion . . . . .	4
2.5	Dataset export . . . . .	8
<b>3</b>	<b>Results</b>	<b>8</b>
3.1	Digitisation of core datasets . . . . .	8
3.2	Data auditing . . . . .	9
3.3	Creation of the data collection . . . . .	9
3.4	Dataset export . . . . .	10
<b>4</b>	<b>Conclusions &amp; Future Work</b>	<b>10</b>
<b>5</b>	<b>Outputs, Data &amp; Software Links</b>	<b>11</b>

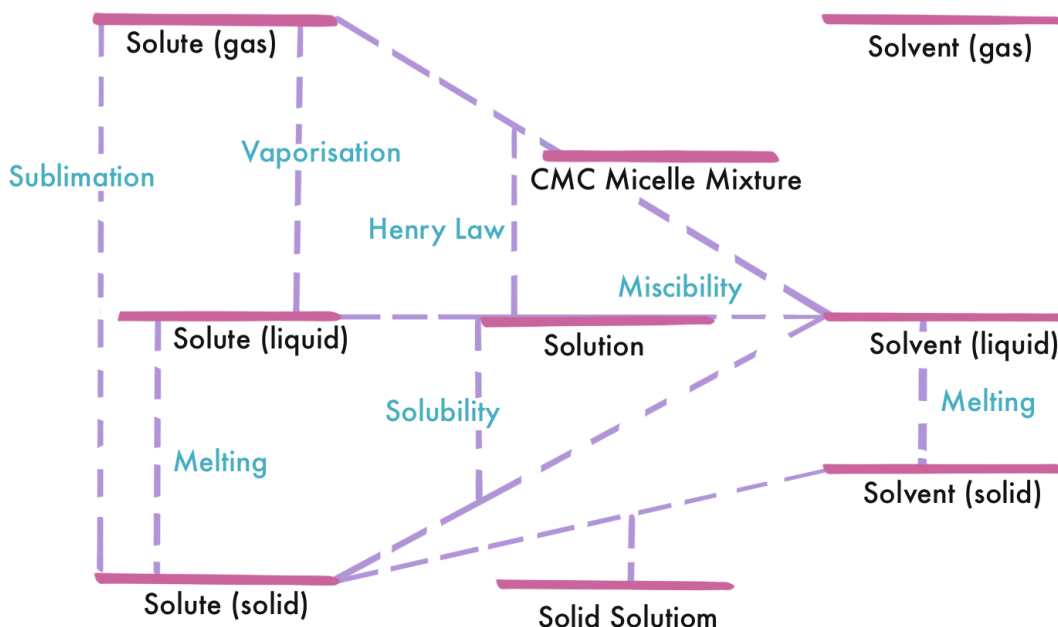


Figure 1: The interconnectedness of the thermodynamic cycle graphically outlined

## 1 Introduction

The accurate prediction and use of physical chemical property data is foundational to modelling chemical behaviour in environmental, industrial, and biological contexts. One of the most widely used properties in these settings is aqueous solubility, which, along with related thermodynamic and phase behaviour data (e.g., Henry’s Law constants, melting points, and miscibility), forms the basis for many computational and regulatory models. Recent work, such as that by Lowe et al. (2023), has emphasised the importance of high-quality, curated datasets to support the development of transparent and reliable quantitative structure–activity/property relationships (QSAR/QSPR). These efforts align with the OECD principles for model validation and have highlighted both the potential and the limitations of using machine learning tools with chemical property data.

The motivation for this project emerged from the need to better integrate and structure a diverse collection of experimental data on solubility and related properties. Much of this data exists in legacy formats—static tables, disparate spreadsheets, and unstructured documents—which hinders its reuse, discoverability, and application to modern computational methods. Furthermore, related properties such as Henry’s Law constants, melting points, and miscibility are often treated in isolation, despite being linked through well-established thermodynamic cycles. The opportunity to better organise and connect these properties represents both a scientific and infrastructural challenge.

The primary aim of this project was to create a harmonised, structured collection of compound-level physical chemistry data that spans aqueous solubility, Henry’s Law constants, melting points, boiling points, and miscibility. A key objective was to support downstream use in modelling—both for developing new models and validating existing ones—by ensuring data provenance, consistency, and accessibility. We focused on making the data machine-actionable by assigning unique compound identifiers, capturing metadata, and outputting datasets that can be used in modelling.

The scope of this project includes:

- Digitisation of core datasets (e.g., NSRDS-NBS 36),
- Data auditing and suitability testing of core datasets,
- Structuring datasets for database ingestion and reuse,
- Identifying inconsistencies (e.g., multiple solubility values for the same solvent) and flagging them for further review.
- Exporting combined datasets for use in modelling and making these available to the public.

## 2 Methodology

In this section we lay out the approach we have taken to digitisation of the datasets, data auditing, development of the data collection, and export of the datasets.

### 2.1 Digitisation of core datasets

The digitisation of the CMC (Critical Micelle Concentration) data required a tailored approach due to the format and variability of the original sources. As the CMC data originated from scanned documents and legacy print materials, we employed a multi-stage pipeline that combined OCR (Optical Character Recognition), heuristics-based corrections, and LLM-assisted column identification to extract and regularise data into a machine-actionable format. A schema of this pipeline is included in figure 2.

The initial step involved OCR to identify tabular structures, which were then parsed into a JSON schema. Using bounding box metadata and known section markers (e.g. compound names and headers), we reconstructed table structure and merged fragmented elements. Tables were regularised by inferring missing or malformed headers and aligning rows across inconsistencies caused by split or merged cells.

Special cases such as multi-line headers, additive-only rows, and duplicate entries were corrected using a combination of visual layout analysis and content-based heuristics. LLMs were employed to classify and normalise column headers, correct OCR misclassifications, and resolve ambiguities arising from merged or malformed columns. Unit conversions and data continuity were preserved through explicit detection and merging processes, ensuring harmonised entries for CMC values reported in multiple units.

The final processed data was output as a structured CSV file, designed for ingestion into the PSDI core dataset framework. The pipeline also retained all metadata needed to trace back to original sources, ensuring transparency and reproducibility

### 2.2 Data auditing

Before integrating the various data sources into a unified collection, we carried out a comprehensive data auditing process to assess the consistency, completeness, and reliability of the records. This step was essential to identify and flag potential issues such as conflicting measurements, miss represented data, and duplicate data sources. The outcomes of this auditing process informed decisions about which source data to include in the data collection in this first version. Any additional source data that is added into the collection will go through a similar data audit process. For each provided data source, the initial step was to trace the data back to its originating publication. Each source under audit was then associated with its originating DOI (or URL, for databases), date of first publication, citation for use (where provided), and

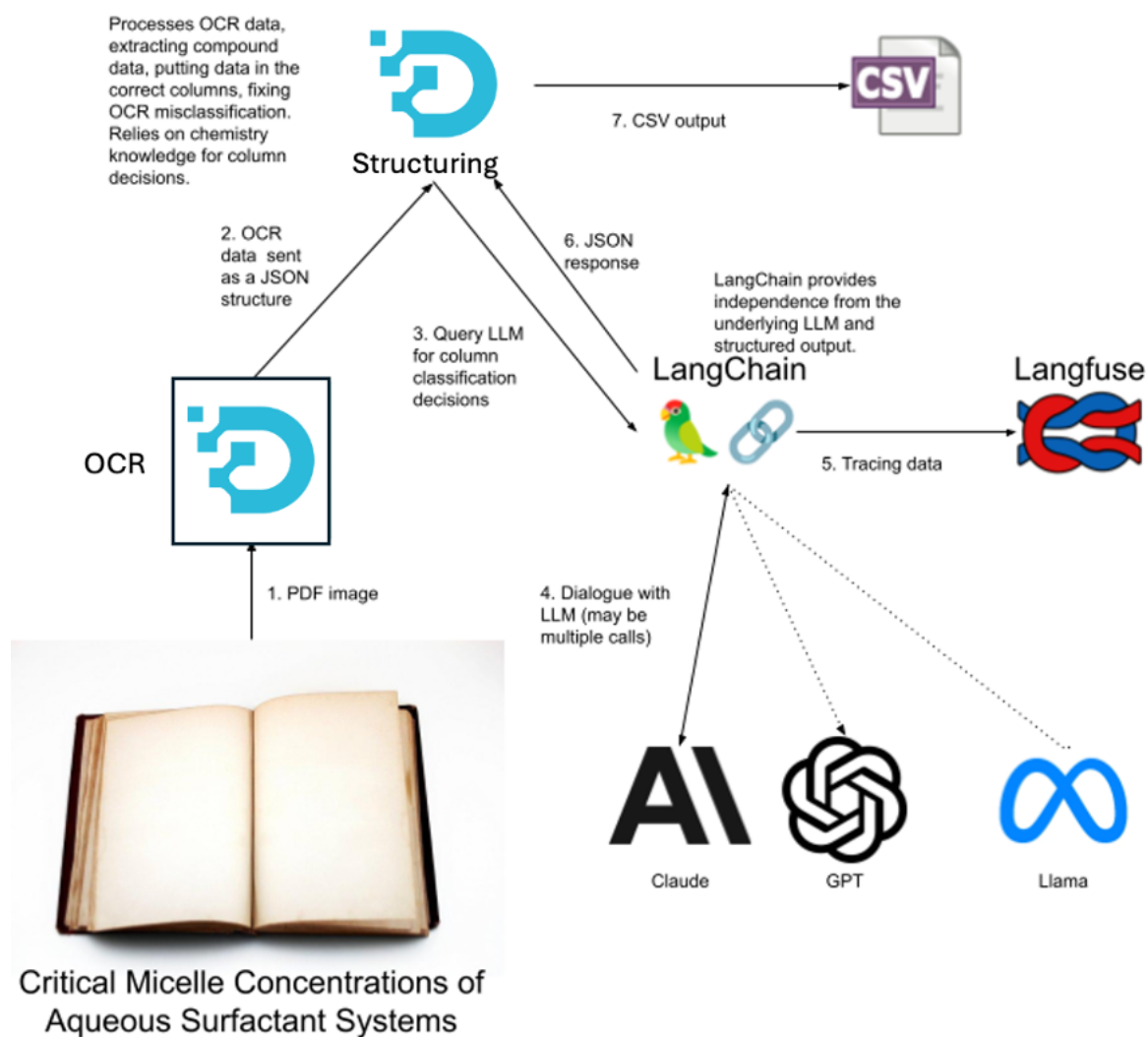


Figure 2: Data Revival Schema



applicable terms and conditions. Next, each data source was assessed for content, specifically examining its size, structure, format, and whether it constituted primary data or included secondary data (i.e., data derived from other sources).

To be included in the data collection, each dataset was required to meet the following criteria during the audit:

- Have a CC0 licence (or have obtained explicit permission for use in the data collection).
- Include InChI or SMILES identifiers to allow proper identification and integration with other datasets.
- Contain experimental data.
- Contain primary data (or clearly identifiable secondary data).

There were also additional considerations around the quality and provenance of the data that was ingested. For example, one dataset (Meng2022) has data that is skewed by the measurement method they used in their work. For this reason, all data included in the collection maintains the link to its original source to enable researchers to make their own determinations about the best way to put that data to use.

## 2.3 Development of the data collection

A number of methodologies were assessed for use in assembling and holding the data collection. The eventual structure was chosen to both allow for longevity but also allow for later restructuring and reshaping as future use cases may determine.

The Physical Chemistry Properties Data Collection was developed by consolidating and structuring data from multiple sources into a unified resource. Each data source underwent rigorous curation, including the standardisation of chemical identifiers such as InChI and SMILES, and the assignment of unique PSDI IDs for each compound. These identifiers facilitated accurate integration and maintenance of a master ID list, ensuring the data was consistently structured and each chemical uniquely represented. Each data source was imported systematically into a series structured CSV records which are stored in a ZIP file to ensure ease of access and compression.

This curated data resource provides the foundation from which targeted datasets can be dynamically generated for specific analyses or applications

.The structure of the data collection is shown below.

```
Physical_Chemistry_Properties_Data_Collection/  
|  
|- master_list.csv  
|  
|--> record_CSVs.zip  
    |000  
        |- PSDI000001.csv  
        |- PSDI000002.csv  
        |- PSDI000003.csv  
        |- ...
```

## 2.4 Data ingestion

The Data Ingestion process involved systematically incorporating and standardising diverse physical chemistry property data into a unified collection.

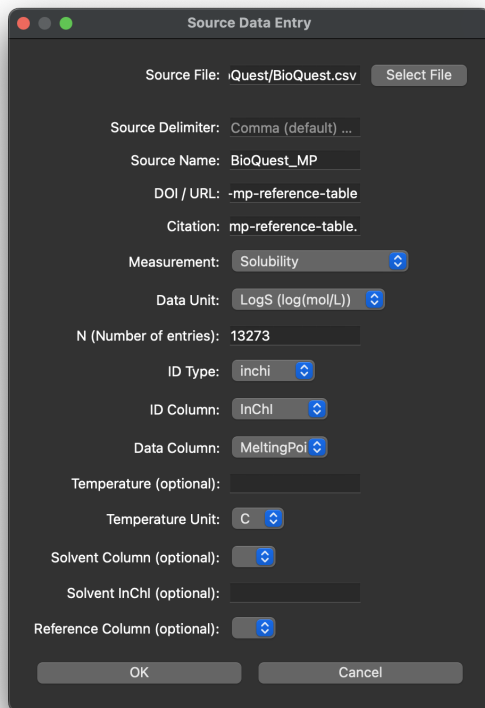


Figure 3: Screenshot of the user interface for the data ingestion user interface

The first step in this process is to assess the data source that is to be imported and ensure that the source type, format, and structure is logged in a file `source_data_info.csv`. This is done via a helper script and user interface, which prompts the user to provide all the information that is required for the source data to be ingested. A screenshot of this user input interface is included below.

When run this user interface saves all the relevant details in the source data info log and also copies the data source from the user into a folder called `ingested_source_data`. This data is now ready for the second stage of data ingestion which is best outlined in the following steps and run using the script `main.py`.

1. **Data ID normalisation.** Each dataset underwent identifier normalisation—assigning or validating key chemical identifiers such as InChI, SMILES.
2. **PSID ID assignment.** Each unique compound in a data source is given a unique 6 digit hex identifier.
3. **Additional molecular information calculation.** Using a number of RDKit python scripts, we add additional molecular information that may be useful in the database.
4. **Master list update.** All unique compounds are added to the `master_list.csv` file with their molecular information. This file also contains a summary of the number of records held.
5. **Compound specific CSV record update.** Finally, the individual record for each compound is updated with the additional data. If no record exists, a new one is created with the master list IDs.

Version: 5.0.0 Last Updated: 2025-02-28 Processed Sources: BioQuest_MF_OIGR2021, BioQuest003, IUPAC_Ming022		PSDI_ID	Canonical name	InChiKey	SMILES	Tautomers	Isomers	n_MeltingPoint	n_BoilingPoint	n_HLC	n_LogS	n_Miscibility	Molecular Weight	CLogP	Heavy Atom Count	Hydrogen Bond Acceptors	Hydrogen Bond Donors	Rotatable Bonds	Rings	Hetero Aromatic Rings	Aromatic Rings	Topological Polar Surface Area	Quantitative Estimation of Drug-likeness
PSDI000001		1HCH-15C18H18O11-18237-4-5-1033-6-7981011-17-81314-4921-1-393	UHFHACIOYH-N	CC1CCCC(C)C(C)C(C)C2O	1	8	1	0	1	1	0	0	154.14	2.19	11	1	1	0	2	0	0	26.23	0.57
PSDI000002		1HCH-15C18H18O11-17238-5-4-8239-1081157-1314-4921-1-393	UHFHACIOYH-N	CC1CCCC(C)C(C)C(C)C1	1	8	2	1	0	0	0	0	156.15	2.44	11	1	1	1	1	0	0	26.23	0.62
PSDI000003		1HCH-15C18H18O11-15-6-6-13-1543-6-158-17-13-12105	UHFHACIOYH-N	CC1CCCC(C)C(C)C(C)C2C	1	16	1	0	0	0	0	0	220.18	3.94	16	1	0	0	3	0	0	12.53	0.44
PSDI000004		1HCH-15C18H18O11-17238-5-4-8239-1081157-814-4921-1-393	UHFHACIOYH-N	CC1CCCC(C)C(C)C(C)C1	3	4	1	1	1	4	0	0	154.14	2.65	11	1	0	1	1	0	0	17.07	0.57
PSDI000005		1HCH-15C18H18O11-17238-5-4-8239-1081157-814-4921-1-393	UHFHACIOYH-N	CC1CCCC(C)C(C)C(C)C1	3	1	1	0	0	0	0	0	286.1	5.07	22	2	2	1	4	0	4	40.46	0.51
PSDI000006		1HCH-15C18H18O11-17238-5-4-8239-1081157-814-4921-1-393	UHFHACIOYH-N	CC1CCCC(C)C(C)C(C)C1	4	1	1	0	0	0	0	0	270.08	3.2	20	4	0	3	2	0	2	52.6	0.63

Figure 4: Example master list file

Data normalisation—assigning or validating key chemical identifiers such as InChI, SMILES, and PSDI ID—was to ensure consistency and facilitate integration. New entries were compared against an existing master list to prevent duplication, with unique compounds receiving sequentially generated identifiers. The ingested data was then structured into individual compound-specific CSV records, capturing all relevant properties and metadata to support streamlined retrieval and future analysis.

An extract of the master\_list.csv (updated in step 4) is shown in figure 4. The master list table then contains unique records for all the compounds in the collection. At the top of the master list is three lines of metadata which is as follows:

1. **Version.** A sequential version number that helps track the master list as it is updated.
2. **Last Updated.** The date of the last version number change.
3. **Processed Sources.** A list of the sources incorporated in this master list version. These sources are detailed in source data info log

Following this metadata The main master list table has the following headings:

1. **PSDI.ID.** Unique hexadecimal identifier assigned to each compound.
2. **Canonical name.** Standardised chemical name following IUPAC conventions.
3. **InChi.** International Chemical Identifier, a textual representation of chemical structures.
4. **InChiKey.** Short hashed version of the InChi, optimised for database searches.
5. **SMILES.** Simplified text-based notation to represent molecular structures.
6. **Tautomers.** Count of isomers differing by proton or electron positioning.
7. **Isomers.** Count of structural or stereochemical variations of the compound.
8. **n\_MeltingPoint.** Available melting point data records for the compound.
9. **n\_BoilingPoint.** Available boiling point data records for the compound.
10. **n\_HLC.** Number of available Henry’s Law constant data entries.
11. **n\_LogS.** Number of aqueous solubility (LogS) data entries available.
12. **n\_Miscibility.** Available miscibility data records.
13. **Molecular Weight.** Total atomic weight of all atoms within the molecule.
14. **CLogP.** Predicted measure of hydrophobicity (octanol-water partition coefficient).
15. **Heavy Atom Count.** Total number of non-hydrogen atoms.

```

;PSDI_ID: PSDI000A0D
;Canonical Name: Unknown
;InChI: InChI=1S/C10H16N2O3/c1-3-5-6-10(4-2)7(13)11-9(15)12-8(10)14/h3-6H2
;InChIKey: STDBAQMTJLUMFW-UHFFFAOYSA-N
;SMILES: CCCCC1(CC)C(O)=NC(=O)N=C1O
;Tautomers: 5.0
;Isomers: 1.0
;Molecular Weight: 212.12
;CLogP: 2.62
;Heavy Atom Count: 15.0
;Hydrogen Bond Acceptors: 1.0
;Hydrogen Bond Donors: 2.0
;Rotatable Bonds: 4.0
;Rings: 1.0
;Hetero Aromatic Rings: 0.0
;Aromatic Rings: 0.0
;Topological Polar Surface Area: 82.25
;Quantitative Estimation of Drug-likeness: 0.75
;Date Created: 2025-02-28
;Master Version upon Creation: 0.0.1

```

Source	Melting Point (C)	Boiling Point (C)	LogS (log(mol/L))	LogS Temperature (C)	LogS Solvent InChi	Misc. MoleFraction	Misc. Temperature (C)	Misc. Solvent InChi	HLC (mol m-3 Pa-1)	HLC Temperature (C)	HLC Solvent InChi	Citation	Date Added	Master Version
BioQuest_MP	128.5											Quest Database-Boilin	28/02/2025	0.0.1
Meng2022			-1.65	25	InChi=1S/H2O/h1H2							Meng, J., Chen, P., Wah	28/02/2025	0.0.6
Meng2022			-1.64	25	InChi=1S/H2O/h1H2							Meng, J., Chen, P., Wah	28/02/2025	0.0.6
Delaney2004			-1.66	25	InChi=1S/H2O/h1H2							J. Chem. Inf. Comput. S	28/02/2025	0.0.7

Figure 5: Example record CSV with three data entries

16. **Hydrogen Bond Acceptors.** Count of atoms capable of accepting hydrogen bonds.
17. **Hydrogen Bond Donors.** Count of atoms capable of donating hydrogen bonds.
18. **Rotatable Bonds.** Number of bonds allowing molecular flexibility.
19. **Rings.** Count of cyclic structures present.
20. **Hetero Aromatic Rings.** Number of aromatic rings containing atoms other than carbon.
21. **Aromatic Rings.** Total number of aromatic ring structures.
22. **Topological Polar Surface Area.** Molecular surface area contributed by polar atoms.
23. **Quantitative Estimation of Drug-likeness.** Numerical estimate of similarity to known drug compounds.

These same headings are included a metadata in the individual record\_CSV files for each of the compounds with the addition of Date of Creation and Master Version upon Creation.

An example of the record CSV is shown in figure 5. In these records, any data imported from source data is appended to the table with all the relevant details. The table headings for a record CSV file are as follows:

1. **Source** – The origin or dataset from which the data point was obtained.
2. **Melting Point (C)** – The temperature in Celsius at which the compound transitions from solid to liquid.
3. **Boiling Point (C)** – The temperature in Celsius at which the compound transitions from liquid to gas.
4. **LogS (log(mol/L))** – The logarithmic value of aqueous solubility in moles per litre.
5. **LogS Temperature (C)** – The temperature at which the solubility (LogS) was measured.
6. **LogS Solvent InChi** – The InChI identifier for the solvent used in the solubility measurement.
7. **Misc. MoleFraction** – A miscibility mole fraction value associated with the measurement.

8. **Misc. Temperature (C)** – A temperature in Celsius related to miscibility.
9. **Misc. Solvent InChi** – The InChi identifier for the solvent involved in the miscibility measurement.
10. **HLC** ( $\text{mol m}^{-3} \text{ Pa}^{-1}$ ) – Henry’s Law Constant, indicating gas solubility in  $\text{mol/m}^3/\text{Pa}$ .
11. **HLC Temperature (C)** – The temperature at which the Henry’s Law Constant was determined.
12. **HLC Solvent InChi** – The InChi identifier for the solvent used in the Henry’s Law Constant determination.
13. **Citation** – The bibliographic reference that should be used with that data row.
14. **Date Added** – The date the entry was incorporated into the collection.
15. **Master Version** – The version of the collection at the time the entry was added.

## 2.5 Dataset export

Once assembled the final stage of the developing the Physical Chemistry Properties Data Collection is to produce tools and systems to allow users to export datasets.

In this version the collection itself will remain private but the the datasets are the public facing output that is intended for use in research. Within this project we have develop a number of example scripts that show how the data can be extracted from the database and example datasets that can be downloaded and used

The example datasets have been chosen by project partners and selected by the PSDI team as sets that may have immediate value and interest. They are outlined further in our results section.

The next stage of the project will begin to open this data collection up to the public and allow for custom data exports.

## 3 Results

### 3.1 Digitisation of core datasets

The digitisation of the CMC dataset presented a number of technical and structural challenges arising from the nature of the source material. Originally captured as scanned pages from a legacy print data book, the compound data was fragmented across inconsistently formatted tables, with frequent issues in text alignment, column labelling, and row merging. Despite these limitations, a high-quality structured dataset was successfully produced through an iterative, multi-step extraction process.

A total of over 600 compound-specific sections were processed. However, it was noted during validation that at least two compounds—compounds 180 and 649—were missing from the original printed source and, as a result, do not appear in the final digitised dataset. Their absence has been documented, and the pipeline is capable of ingesting them should additional scans or supplementary material become available.

Several notable challenges required the development of targeted solutions:

- **Header inconsistency and multi-line formatting:** Many compound tables contained headers split across multiple lines, which OCR systems often interpreted as disjointed or interleaved. This was resolved using bounding box data to identify the spatial structure of the headers, enabling reconstruction into single coherent header lines.

- Column ambiguity and misclassification: OCR often misinterpreted vertical or narrow text columns, resulting in merged or fragmented column data (e.g., the combination of "Source" and "Evaluation" into a single field). To correct this, a language model-assisted step was introduced that analysed the unique values of each column and reclassified them using examples and contextual descriptions.
- Handling of duplicated or partial rows: Some tables included summary rows such as "X ENTRIES FOR COMPOUND" or rows containing only additive information. These were systematically identified and removed or merged with the correct preceding entries to ensure completeness and accuracy.
- Multi-unit data entries: Several entries for the same experiment were found with values reported in multiple units (e.g., mg/L and mol/L). These were flagged, parsed, and, where possible, merged into a unified format while retaining original units for traceability.
- Missing or misclassified table sections: In some cases, table data was not recognised at all by the extraction process due to layout irregularities. A heuristic based on the expectation of one table per page was applied to detect and incorporate such orphaned data.

As a result of these interventions, the digitised dataset maintains a high degree of fidelity to the original data while being fully structured, searchable, and suitable for integration into the wider PSDI infrastructure. The cleaned dataset has been exported as CSV and retains traceability to original page and section references for validation or reprocessing in future iterations.

### 3.2 Data auditing

Initially 16 data sources were identified by the project partners. Of these 9 were chosen for initial ingest into the collection. These datasets are outlined in the following table. The sources ingested were chosen as they all passed the data auditing checks outlined in methods. The remaining datasets have not been included at this stage because of either licensing or duplication. Those that have not been included due to licensing may be ingested at a later date as licences are being sought for those sources. Those not included due to duplication may also be included at a later date, once they have been checked. A number of the sources themselves are collections of data from multiple sources (which often overlap). Further detailed data tracing and development of data filtering and duplication checking code will resolve these issues in the future.

### 3.3 Creation of the data collection

The data collection was created by ingesting one source at a time as per our outlined method. Version 1 of the data collection was formed using the data sources identified as a result of our data audit. This resulted in a first version of the data collection with the following attributes:

- Total number of individual compounds: **110,804**
- Number of compounds with melting point data: 8,811
- Number of compounds with boiling point data: 6,544
- Number of compounds with HLC data: 1,224
- Number of compounds with LogS data: 102,927
- Number of molecules with Miscibility data: 236

Of this original data sources identified, this represents about 1/3 of the data, which is a significant achievement for the first version of the database. As discussed in data audit, it is intended that future version will have additional source data and grow this collection.

Source name	Property	Entries	Ingested
<a href="#">Cherqaoui1994</a>	BP	134	
<a href="#">Kim2024</a>	BP	1,748	
<a href="#">BioQuest</a>	BP, MP	13,273	Y
<a href="#">WikiData</a>	BP, MP	1,565	
<a href="#">Bergstrom2003</a>	MP	277	Y
<a href="#">Williams2015</a>	MP	228,174	Y
<a href="#">BigSolDB</a>	Sol	54,273	Y
<a href="#">DDB2023</a>	Mis	4,842	Y
<a href="#">IUPAC</a>	Mis	4,982	Y
<a href="#">SolProp</a>	Sol	4,953	
<a href="#">Boobier2020</a>	Sol	1,575	
<a href="#">Delaney2004</a>	Sol	1,144	Y
<a href="#">Llompert2024</a>	Sol	12,233	
<a href="#">Lowe2023</a>	Sol	39,671	
<a href="#">Meng2022</a>	Sol	127,949	Y
<a href="#">Sander2023</a>	HLC	12,255	Y

Table 1: Summary of data sources and ingestion status. Source names are linked to their respective URLs. Property types are Solubility (Sol), Boiling Point (BP), Melting Point (MP), Miscibility (Mis) and Henry Law Constants (HLC)

### 3.4 Dataset export

Each dataset was exported from the data collection using a specific 'dataset module' these are designed to act as examples for users to develop their own data collection export modules. In total, six of these modules were developed with an additional 'GUI' based module which allowed for the export of a single data type (e.g. all compounds with melting points). The resulting datasets are found in our GitHub repository detailed at the end of this report.

## 4 Conclusions & Future Work

The Physical Chemistry Properties Data Collection project has successfully created a comprehensive, structured data collection that significantly advances the integration and accessibility of essential physical chemistry properties, including aqueous solubility, melting and boiling points, Henry’s Law constants, and miscibility.

Through meticulous digitisation, auditing, and curation of data from diverse sources, the project has established robust and machine-actionable datasets, compliant with international standards such as IUPAC Goldbook and FAIR principles. These datasets are now readily available for immediate research applications and offer substantial value for chemical informatics, computational modelling, environmental analysis, and regulatory frameworks. The structured nature of this dataset not only facilitates enhanced reproducibility and transparency in scientific research but also serves as a foundational resource for the development and validation of advanced predictive models.

Looking forward, future phases of this project should focus on three key aspects.

Firstly, expanding the database infrastructure to enhance data accessibility further. Implementing robust database technologies such as SQL or MongoDB would support dynamic data retrieval, efficient data management, and advanced query capabilities. This would signific-

antly improve user interaction with the data collection, enabling researchers to quickly extract relevant information for their specific analytical or modelling needs.

Secondly, improving record ingestion processes to address historical data duplication issues is essential. Developing more sophisticated algorithms and scripts for identifying and resolving duplicates and inconsistencies across multiple data sources would enhance data quality and reliability. Additionally, the integration of further licensed datasets would enrich the diversity and scope of the collection, thereby expanding its applicability across various research fields.

Finally, developing interactive tools such as a web-hosted data browser, alongside more sophisticated data export modules, would greatly enhance the usability and accessibility of the dataset. These tools would enable researchers from diverse disciplines to interact more intuitively with the data, facilitating easier exploration, visualisation, and extraction of data tailored to their specific research questions. Such developments would broaden the user base and encourage greater interdisciplinary collaboration, ultimately driving further innovation and discovery in chemical informatics and related areas.

This data collection is already a highly valuable resource, which has brought together disparate data sources for the first time. A future which see this data collection expanded further can only grow its value and utility to researchers.

## 5 Outputs, Data & Software Links

Due to licensing issues, only the datasets are publically available at this time in the project. It is envisaged that the main collection will be available in the near future once licensing issues have been resolved.

The datasets exported from the Data collection and from the CMC data are available in our GitHub repository.

- <https://github.com/PSDI-UK/psdi-datasets/>