ThermoML-fair Project Summary

Executive Summary

ThermoML-fair: FAIR Data Parser for Thermophysical Properties

ThermoML-fair unlocks thousands of peer-reviewed experimental datasets published in leading journals and archived by NIST. By converting complex XML into structured, machine learning–ready tables, it makes validated, high-quality thermophysical property data usable for machine learning — a critical step toward trustworthy, reproducible models in materials discovery.

What I Built

- Automates extraction of tabular datasets from 1,100+ ThermoML XML files.
- Normalizes property and variable names, preserving uncertainty and mixture compositions.
- Generates CSVs and pandas DataFrames directly consumable for ML pipelines.
- Provides CLI interface and schema validation for robust, repeatable use.

Impact

- Bridges raw NIST XML data to ML-ready training sets.
- Unlocks property prediction modeling for conductivity, viscosity, and beyond.
- Supports reproducible research aligned with FAIR data standards.
- Provides ML access to rigorously peer-reviewed data spanning journals such as Journal of Chemical & Engineering Data, Journal of Chemical Thermodynamics, and Fluid Phase Equilibria.
- Facilitates benchmarking and reproducibility by ensuring every data point is linked to its source publication.

Technical Overview

The parser integrates `xmlschema` with custom mapping logic to extract structured data from ThermoML files. Features include:

- Handling PureOrMixtureData sections, phases, and components
- Normalizing variables, properties, and units
- Serializing outputs into CSV and pandas DataFrames
- Modular design with CLI + library usage

Key Stats

- 1,100+ NIST ThermoML files parsed
- 2.6M+ property entries extracted
- Multi-property support: thermal conductivity, viscosity, heat capacity, and more

Visuals

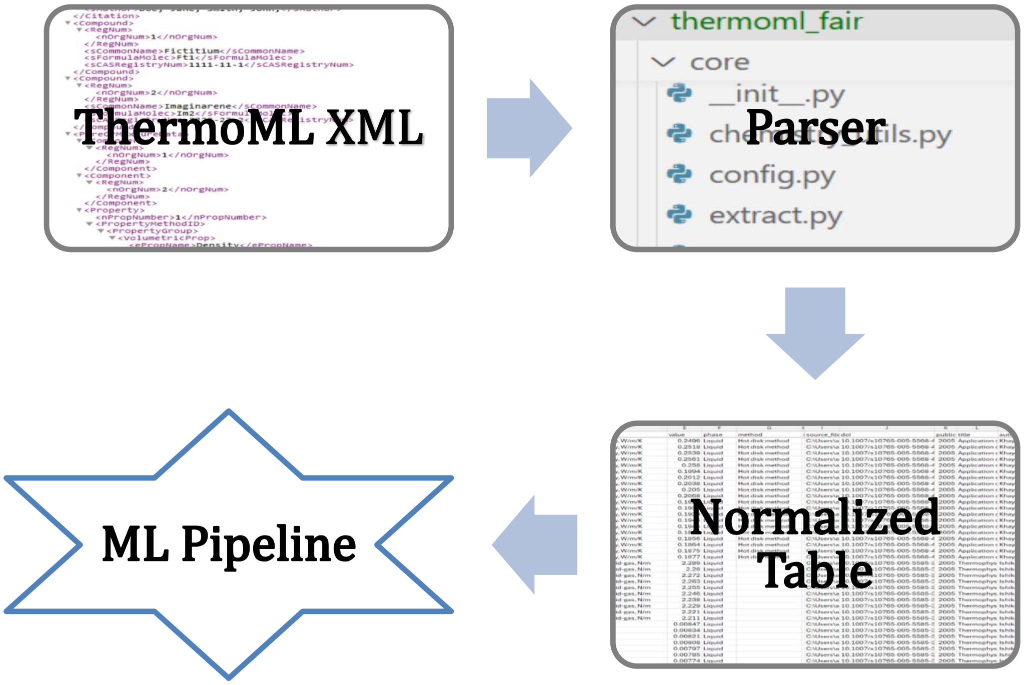
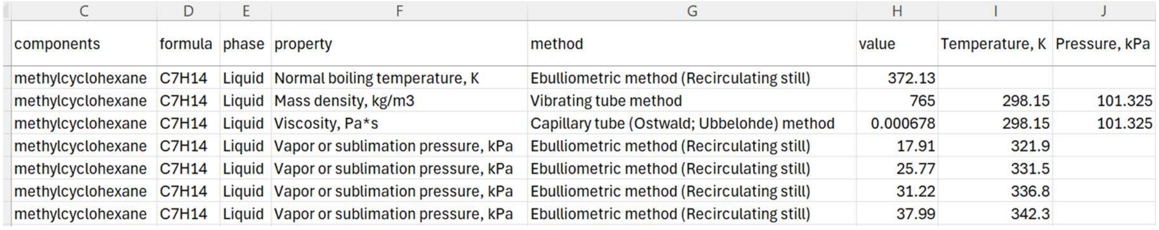
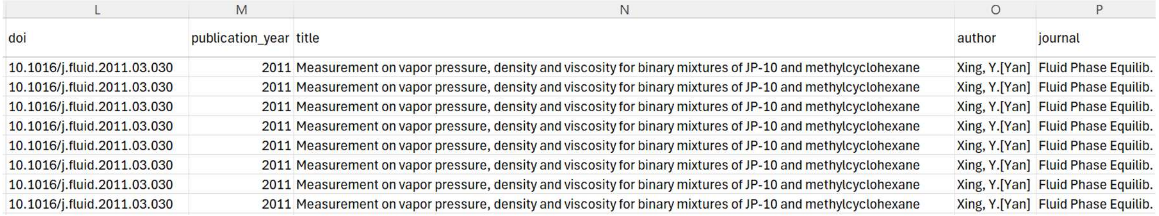


Figure 1 System architecture showing how ThermoML XML files are parsed, normalized, and converted into ML-ready DataFrames. Highlights end-to-end reproducibility and FAIR compliance.



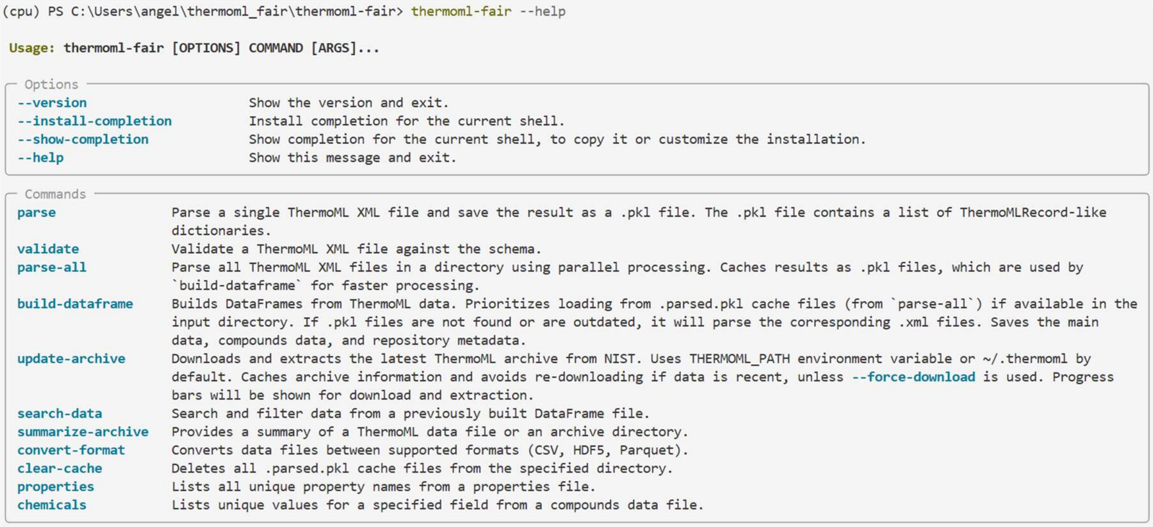
C	D	E	F	G	H	I	J
components	formula	phase	property	method	value	Temperature, K	Pressure, kPa
methylcyclohexane	C7H14	Liquid	Normal boiling temperature, K	Ebulliometric method (Recirculating still)	372.13		
methylcyclohexane	C7H14	Liquid	Mass density, kg/m3	Vibrating tube method	765	298.15	101.325
methylcyclohexane	C7H14	Liquid	Viscosity, Pa*s	Capillary tube (Ostwald; Ubbelohde) method	0.000678	298.15	101.325
methylcyclohexane	C7H14	Liquid	Vapor or sublimation pressure, kPa	Ebulliometric method (Recirculating still)	17.91	321.9	
methylcyclohexane	C7H14	Liquid	Vapor or sublimation pressure, kPa	Ebulliometric method (Recirculating still)	25.77	331.5	
methylcyclohexane	C7H14	Liquid	Vapor or sublimation pressure, kPa	Ebulliometric method (Recirculating still)	31.22	336.8	
methylcyclohexane	C7H14	Liquid	Vapor or sublimation pressure, kPa	Ebulliometric method (Recirculating still)	37.99	342.3	

Figure 2 Example parsed DataFrame output. Each row corresponds to a property measurement under defined experimental conditions (e.g., temperature, pressure, composition).



L	M	N	O	P
doi	publication_year	title	author	journal
10.1016/j.fluid.2011.03.030	2011	Measurement on vapor pressure, density and viscosity for binary mixtures of JP-10 and methylcyclohexane	Xing, Y.[Yan]	Fluid Phase Equilib.
10.1016/j.fluid.2011.03.030	2011	Measurement on vapor pressure, density and viscosity for binary mixtures of JP-10 and methylcyclohexane	Xing, Y.[Yan]	Fluid Phase Equilib.
10.1016/j.fluid.2011.03.030	2011	Measurement on vapor pressure, density and viscosity for binary mixtures of JP-10 and methylcyclohexane	Xing, Y.[Yan]	Fluid Phase Equilib.
10.1016/j.fluid.2011.03.030	2011	Measurement on vapor pressure, density and viscosity for binary mixtures of JP-10 and methylcyclohexane	Xing, Y.[Yan]	Fluid Phase Equilib.
10.1016/j.fluid.2011.03.030	2011	Measurement on vapor pressure, density and viscosity for binary mixtures of JP-10 and methylcyclohexane	Xing, Y.[Yan]	Fluid Phase Equilib.
10.1016/j.fluid.2011.03.030	2011	Measurement on vapor pressure, density and viscosity for binary mixtures of JP-10 and methylcyclohexane	Xing, Y.[Yan]	Fluid Phase Equilib.
10.1016/j.fluid.2011.03.030	2011	Measurement on vapor pressure, density and viscosity for binary mixtures of JP-10 and methylcyclohexane	Xing, Y.[Yan]	Fluid Phase Equilib.
10.1016/j.fluid.2011.03.030	2011	Measurement on vapor pressure, density and viscosity for binary mixtures of JP-10 and methylcyclohexane	Xing, Y.[Yan]	Fluid Phase Equilib.

Figure 3 FAIR principle in practice: every measurement is linked back to its peer-reviewed source, ensuring data transparency and reproducibility.

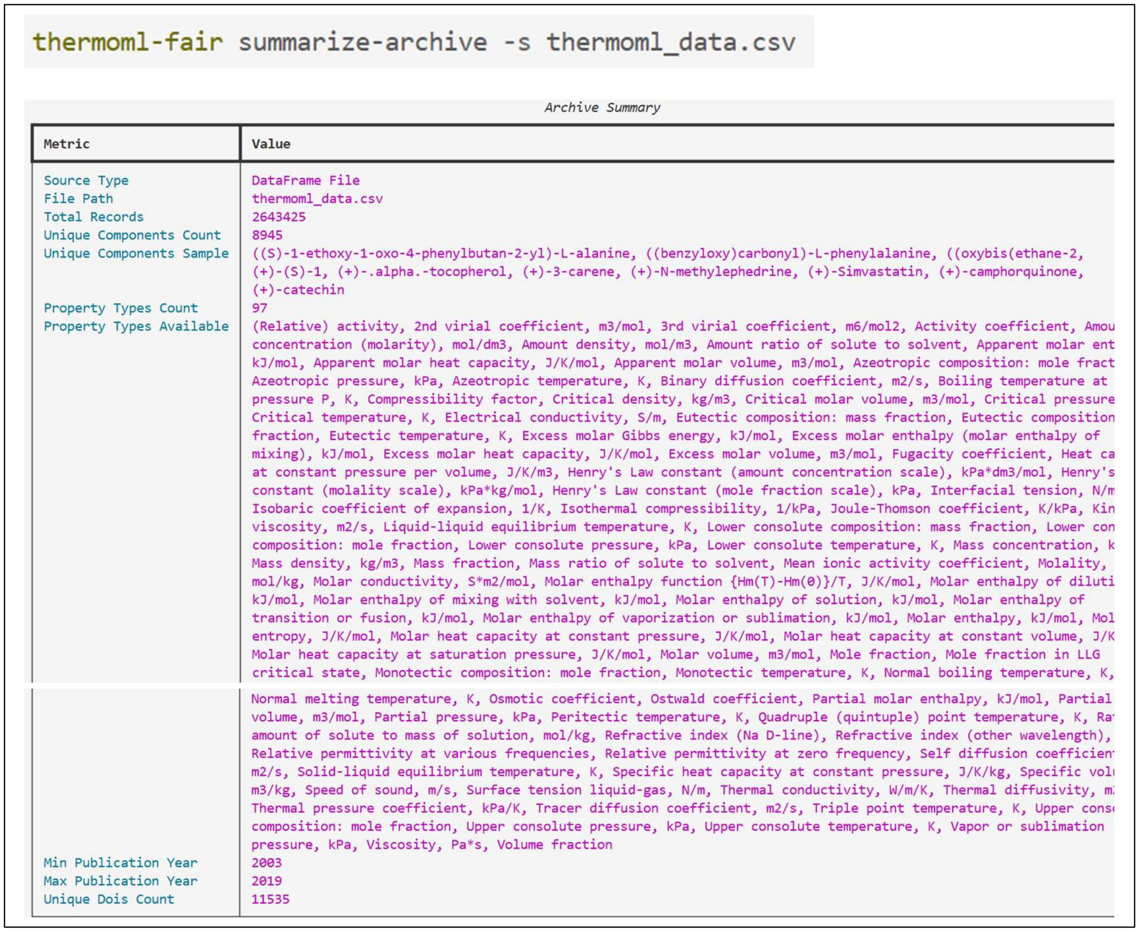


```
(cpu) PS C:\Users\angel\thermoml_fair\thermoml-fair> thermoml-fair --help

 Usage: thermoml-fair [OPTIONS] COMMAND [ARGS]...

╭─ Options ──────────────────────────────────────────────────────────────────────────────────────────────╮
│ --version                     Show the version and exit.                                               │
│ --install-completion          Install completion for the current shell.                                │
│ --show-completion             Show completion for the current shell, to copy it or customize the installation. │
│ --help                        Show this message and exit.                                              │
╰────────────────────────────────────────────────────────────────────────────────────────────────────────╯
╭─ Commands ─────────────────────────────────────────────────────────────────────────────────────────────╮
│ parse               Parse a single ThermoML XML file and save the result as a .pkl file. The .pkl file contains a list of ThermoMLRecord-like │
│                     dictionaries.                                                                      │
│ validate            Validate a ThermoML XML file against the schema.                                   │
│ parse-all           Parse all ThermoML XML files in a directory using parallel processing. Caches results as .pkl files, which are used by │
│                     `build-dataframe` for faster processing.                                           │
│ build-dataframe     Builds DataFrames from ThermoML data. Prioritizes loading from .parsed.pkl cache files (from `parse-all`) if available in the │
│                     input directory. If .pkl files are not found or are outdated, it will parse the corresponding .xml files. Saves the main │
│                     data, compounds data, and repository metadata.                                     │
│ update-archive      Downloads and extracts the latest ThermoML archive from NIST. Uses THERMOML_PATH environment variable or ~/.thermoml by │
│                     default. Caches archive information and avoids re-downloading if data is recent, unless --force-download is used. Progress │
│                     bars will be shown for download and extraction.                                    │
│ search-data         Search and filter data from a previously built DataFrame file.                     │
│ summarize-archive   Provides a summary of a ThermoML data file or an archive directory.                │
│ convert-format      Converts data files between supported formats (CSV, HDF5, Parquet).                │
│ clear-cache         Deletes all .parsed.pkl cache files from the specified directory.                  │
│ properties          Lists all unique property names from a properties file.                            │
│ chemicals           Lists unique values for a specified field from a compounds data file.              │
╰────────────────────────────────────────────────────────────────────────────────────────────────────────╯
```

Figure 4 Command-line interface (CLI) features, supporting robust parsing, validation, and DataFrame construction with progress bars and error handling.



```
thermoml-fair summarize-archive -s thermoml_data.csv
```

```
                                            Archive Summary
┏━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
┃ Metric                   ┃ Value
┡━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
│ Source Type              │ DataFrame File
│ File Path                │ thermoml_data.csv
│ Total Records            │ 2643425
│ Unique Components Count  │ 8945
│ Unique Components Sample │ ((S)-1-ethoxy-1-oxo-4-phenylbutan-2-yl)-L-alanine, ((benzyloxy)carbonyl)-L-phenylalanine, ((oxybis(ethane-2,
│                          │ (+)-(S)-1, (+)-.alpha.-tocopherol, (+)-3-carene, (+)-N-methylephedrine, (+)-Simvastatin, (+)-camphorquinone,
│                          │ (+)-catechin
│ Property Types Count     │ 97
│ Property Types Available │ (Relative) activity, 2nd virial coefficient, m3/mol, 3rd virial coefficient, m6/mol2, Activity coefficient, Amou
│                          │ concentration (molarity), mol/dm3, Amount density, mol/m3, Amount ratio of solute to solvent, Apparent molar ent
│                          │ kJ/mol, Apparent molar heat capacity, J/K/mol, Apparent molar volume, m3/mol, Azeotropic composition: mole fract
│                          │ Azeotropic pressure, kPa, Azeotropic temperature, K, Binary diffusion coefficient, m2/s, Boiling temperature at
│                          │ pressure P, K, Compressibility factor, Critical density, kg/m3, Critical molar volume, m3/mol, Critical pressure
│                          │ Critical temperature, K, Electrical conductivity, S/m, Eutectic composition: mass fraction, Eutectic composition
│                          │ fraction, Eutectic temperature, K, Excess molar Gibbs energy, kJ/mol, Excess molar enthalpy (molar enthalpy of
│                          │ mixing), kJ/mol, Excess molar heat capacity, J/K/mol, Excess molar volume, m3/mol, Fugacity coefficient, Heat ca
│                          │ at constant pressure per volume, J/K/m3, Henry's Law constant (amount concentration scale), kPa*dm3/mol, Henry's
│                          │ constant (molality scale), kPa*kg/mol, Henry's Law constant (mole fraction scale), kPa, Interfacial tension, N/m
│                          │ Isobaric coefficient of expansion, 1/K, Isothermal compressibility, 1/kPa, Joule-Thomson coefficient, K/kPa, Kin
│                          │ viscosity, m2/s, Liquid-liquid equilibrium temperature, K, Lower consolute composition: mass fraction, Lower con
│                          │ composition: mole fraction, Lower consolute pressure, kPa, Lower consolute temperature, K, Mass concentration, k
│                          │ Mass density, kg/m3, Mass fraction, Mass ratio of solute to solvent, Mean ionic activity coefficient, Molality,
│                          │ mol/kg, Molar conductivity, S*m2/mol, Molar enthalpy function {Hm(T)-Hm(0)}/T, J/K/mol, Molar enthalpy of diluti
│                          │ kJ/mol, Molar enthalpy of mixing with solvent, kJ/mol, Molar enthalpy of solution, kJ/mol, Molar enthalpy of
│                          │ transition or fusion, kJ/mol, Molar enthalpy of vaporization or sublimation, kJ/mol, Molar enthalpy, kJ/mol, Mol
│                          │ entropy, J/K/mol, Molar heat capacity at constant pressure, J/K/mol, Molar heat capacity at constant volume, J/K
│                          │ Molar heat capacity at saturation pressure, J/K/mol, Molar volume, m3/mol, Mole fraction, Mole fraction in LLG
│                          │ critical state, Monotectic composition: mole fraction, Monotectic temperature, K, Normal boiling temperature, K,

│                          │ Normal melting temperature, K, Osmotic coefficient, Ostwald coefficient, Partial molar enthalpy, kJ/mol, Partial
│                          │ volume, m3/mol, Partial pressure, kPa, Peritectic temperature, K, Quadruple (quintuple) point temperature, K, Ra
│                          │ amount of solute to mass of solution, mol/kg, Refractive index (Na D-line), Refractive index (other wavelength),
│                          │ Relative permittivity at various frequencies, Relative permittivity at zero frequency, Self diffusion coefficien
│                          │ m2/s, Solid-liquid equilibrium temperature, K, Specific heat capacity at constant pressure, J/K/kg, Specific vol
│                          │ m3/kg, Speed of sound, m/s, Surface tension liquid-gas, N/m, Thermal conductivity, W/m/K, Thermal diffusivity, m
│                          │ Thermal pressure coefficient, kPa/K, Tracer diffusion coefficient, m2/s, Triple point temperature, K, Upper cons
│                          │ composition: mole fraction, Upper consolute pressure, kPa, Upper consolute temperature, K, Vapor or sublimation
│                          │ pressure, kPa, Viscosity, Pa*s, Volume fraction
│ Min Publication Year     │ 2003
│ Max Publication Year     │ 2019
│ Unique Dois Count        │ 11535
└──────────────────────────┴──────────────────────────────────────────────────────────────────────────────────────────────────────────────────
```

Figure 5 CLI summary view showing aggregated counts of parsed properties, compounds, and measurements across the ThermoML archive.

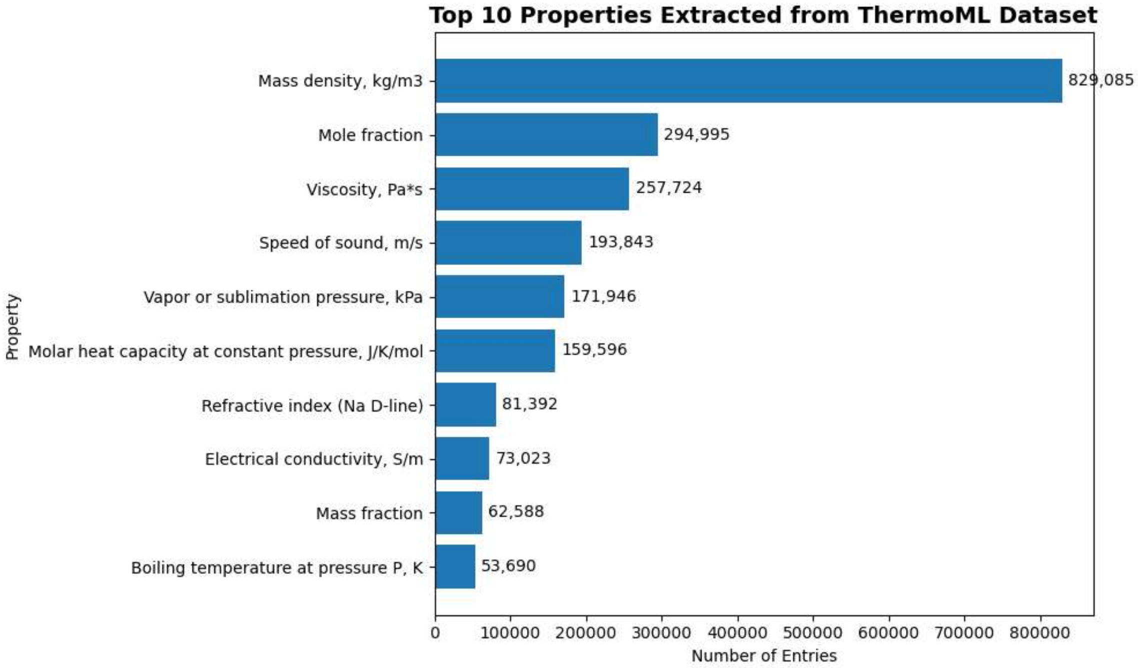
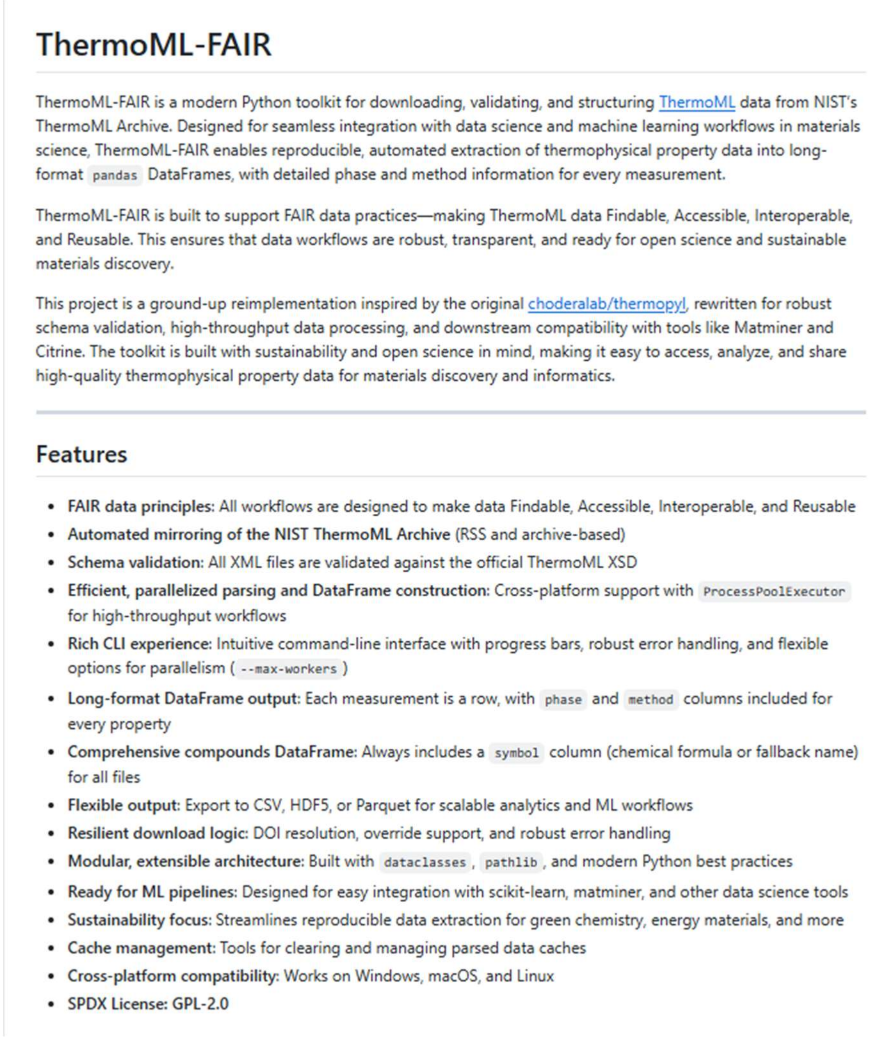


Figure 6 Distribution of the top 10 most frequently reported thermophysical properties. Demonstrates dataset breadth (97 unique properties, 2.6M+ entries).



ThermoML-FAIR

ThermoML-FAIR is a modern Python toolkit for downloading, validating, and structuring ThermoML data from NIST's ThermoML Archive. Designed for seamless integration with data science and machine learning workflows in materials science, ThermoML-FAIR enables reproducible, automated extraction of thermophysical property data into long-format `pandas` DataFrames, with detailed phase and method information for every measurement.

ThermoML-FAIR is built to support FAIR data practices—making ThermoML data Findable, Accessible, Interoperable, and Reusable. This ensures that data workflows are robust, transparent, and ready for open science and sustainable materials discovery.

This project is a ground-up reimplementation inspired by the original choderalab/thermopyl, rewritten for robust schema validation, high-throughput data processing, and downstream compatibility with tools like Matminer and Citrine. The toolkit is built with sustainability and open science in mind, making it easy to access, analyze, and share high-quality thermophysical property data for materials discovery and informatics.

Features

- **FAIR data principles:** All workflows are designed to make data Findable, Accessible, Interoperable, and Reusable
- **Automated mirroring of the NIST ThermoML Archive** (RSS and archive-based)
- **Schema validation:** All XML files are validated against the official ThermoML XSD
- **Efficient, parallelized parsing and DataFrame construction:** Cross-platform support with `ProcessPoolExecutor` for high-throughput workflows
- **Rich CLI experience:** Intuitive command-line interface with progress bars, robust error handling, and flexible options for parallelism (`--max-workers`)
- **Long-format DataFrame output:** Each measurement is a row, with `phase` and `method` columns included for every property
- **Comprehensive compounds DataFrame:** Always includes a `symbol` column (chemical formula or fallback name) for all files
- **Flexible output:** Export to CSV, HDF5, or Parquet for scalable analytics and ML workflows
- **Resilient download logic:** DOI resolution, override support, and robust error handling
- **Modular, extensible architecture:** Built with `dataclasses` , `pathlib` , and modern Python best practices
- **Ready for ML pipelines:** Designed for easy integration with scikit-learn, matminer, and other data science tools
- **Sustainability focus:** Streamlines reproducible data extraction for green chemistry, energy materials, and more
- **Cache management:** Tools for clearing and managing parsed data caches
- **Cross-platform compatibility:** Works on Windows, macOS, and Linux
- **SPDX License:** GPL-2.0

Figure 7 Screenshot of the GitHub README top section, showing professional project documentation and open-source accessibility on GitHub.

ThermoML-FAIR transforms decades of peer-reviewed thermophysical research into machine learning–ready data — making reproducible, sustainable materials discovery possible at scale.

Links:

GitHub: https://github.com/acfdavis/thermoml-fair

Portfolio: https://acfdavis.github.io

v1.0 – August 2025

Author

Created by **Angela C. Davis**, Materials Scientist and Data Innovator, passionate about sustainable materials discovery and data-driven reproducibility.