[WiP] Chemical Datasets

Objective

Structuring chemical datasets with a modern, secure, and scalable tech stack - grounded in information retrieval principles - transforms raw data into actionable knowledge. This foundation is essential for supporting experimentation, enabling scientific exploration, integrating workflows, and powering advanced simulations and machine learning models. Ultimately, it accelerates the discovery and development of new products, providing a strategic advantage in scientific innovation.

Data Pipelining and Governance

WiP

Data Exploration Tech Stack

WiP

Information Retrieval for Chemical Data

WiP

Eligible Datasets

Dataset Name	Description	Format(s)	Size	Update Frequency	Reference/Link
NIST TDE	Critically evaluated thermophysical and thermochemical property data for pure compounds, mixtures, and reactions, widely used for chemical engineering and materials science	CSV, XML, JSON, proprietary export formats	~GBs	Periodic (annually or as new data is available)	NIST TDE
OPoly26	Large-scale open dataset of 26M+ unique polymer structures with computed properties and rich metadata for polymer informatics and machine learning	CSV, JSON, SDF, HDF5, SMILES, SELFIES	~1.2 TB	Batch releases (every few months)	arXiv:2512.23117
PolyInfo	Polymer properties, structures, synthesis routes	CSV, SDF, JSON	~GBs	Periodic	https://polymer.nims.go.jp/
Polymer Genome	Polymer property predictions, descriptors	CSV, JSON	~GBs	Periodic	https://www.polymergenome.org/
Materials Project	Inorganic materials, properties, structures	JSON, CSV	~TBs	Weekly	https://materialsproject.org/
QM9	Small organic molecules, quantum properties	XYZ, CSV, JSON	~GBs	Static	https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/qm9.csv
Open Catalyst Project	Catalysts, surface reactions, DFT calculations	HDF5, JSON	~TBs	Periodic	https://opencatalystproject.org/
PubChem	Chemical structures, properties, bioactivity	SDF, CSV, JSON	~TBs	Daily	https://pubchem.ncbi.nlm.nih.gov/

References

In Search of Better Search https://dl.acm.org/doi/10.1145/3760247