ChatGPT

Physics-Based Datasets by Tier: A Comprehensive Resource Guide

Tier 0: Foundational Datasets (Variational Calculus, PDEs, Stochastic Processes)

PDE Theory & Variational Problems: *PDEBench* (2022) is a broad benchmark dataset for scientific machine learning, providing numerical solutions for various PDEs [1]. It includes 1D advection, Burgers' equation, diffusion–reaction (reaction–diffusion) in 1D/2D, 1D diffusion-sorption, 2D Darcy flow, 2D shallow water, and 1D/2D/3D Navier–Stokes simulations [2]. Data are stored in **HDF5** format with dimensions [batch, time, spatial..., variables] and released under **CC BY 4.0** license [3] [4], explicitly intended for AI model training and benchmarking. No specific curated dataset for calculus of variations exists (these problems are usually solved analytically or via on-the-fly simulation), but PDEBench's variational PDE solutions can serve as training data for learning Euler–Lagrange solution mappings. Additionally, open-source PDE solver libraries (e.g. **Fenics**, **FreeFEM**) can generate custom variational problem datasets, though not pre-packaged.

Stochastic Processes: For stochastic time-series and random processes, the *NIST Noise SDE Dataset* (2023) offers a rich collection of synthetic trajectories [5]. This public dataset contains time-series realizations of various noise processes: band-limited (filtered Gaussian) noise, fractional Gaussian noise (fGn) and fractional Brownian motion (fBm), impulsive (spiky) noise, shot noise, etc., totaling several gigabytes of sequences [5]. It was developed for training generative models (GANs) on noise time-series and includes detailed documentation of simulation methods [6]. Hosted by NIST with an open-access intent (license specified via NIST's open data terms) [7], the data are provided in text/CSV format (with accompanying README) for use in AI experiments. Additionally, an experimental *Brownian Motion* dataset from Harvard (2023) and a **U.S. Government Noise** dataset [8] exist, but the NIST set is more comprehensive. *Scikit-FDA* and similar libraries also allow simulation of Ornstein–Uhlenbeck, Langevin, and other stochastic differential equation trajectories [9]. While these foundational stochastic datasets are synthetic, they are well-suited for AI model training in stochastic process prediction and system identification. (Licenses are generally **public domain/CC0** for government data, or **CC BY** for academic contributions.)

Tier 1: Reaction–Diffusion, GENERIC Thermodynamics, Kinetic-to-RD Coarse Graining

Reaction–Diffusion (RD) Systems: Public datasets of RD simulations are typically synthetic but high-quality. PDEBench (above) includes 1D and 2D diffusion–reaction systems (e.g. two-species activator-inhibitor models) with many trajectories [10], useful for training neural PDE solvers. Beyond PDEBench, researchers often generate RD data (e.g. Gray–Scott or Turing pattern simulations) via open code. For example, a Gray–Scott RD simulation code is available on GitHub (C/Python implementations) and can output image sequences of pattern formation [11]. These datasets are usually **numerical trajectories** (concentration fields over time) saved in image or HDF5 formats. While no single standardized RD dataset dominates, small curated sets appear in literature (often released on Zenodo or GitHub). Licensing is typically **MIT/Apache**

for code and **CC BY 4.0** for any released data. Such synthetic RD datasets are designed with AI in mind – e.g. training convolutional networks to predict pattern evolution or coarse-grained models.

GENERIC Thermodynamics: The GENERIC framework (General Equation for Non-Equilibrium Reversible-Irreversible Coupling) is theoretical, and *no dedicated public dataset* of GENERIC simulations exists (it's more a methodology for deriving equations). Instead, researchers use simulation libraries to generate data for specific thermodynamic systems. For instance, **ThermoData Engine** at NIST provides extensive equilibrium thermophysical property data (CC BY 4.0) [12], and the **OpenKMC**/**COPASI** tools can simulate kinetic networks with thermodynamic consistency. If needed, one can construct synthetic datasets of thermodynamic processes (e.g. relaxation to equilibrium, cyclic processes) using open libraries – but these would be case-specific. As such, any "GENERIC" dataset would likely be a **code + trace** format: e.g. code implementing a thermodynamic model and logged state trajectories (perhaps in CSV). Licensing for underlying data (e.g. NIST property databases) is often **public domain or CC**; simulation code libraries are open-source (e.g. **PyGENEIC** if available, or other thermodynamics packages). In summary, while no pre-curated AI-ready GENERIC dataset was found in connected sources, the building blocks (property databases and simulation codes) are openly accessible for creating one.

Kinetic-to-RD Coarse Graining: This domain concerns bridging microscopic kinetic models (like particle or lattice simulations) to macroscopic reaction–diffusion equations. High-quality data here often come from *multiscale simulation studies* rather than general repositories. For example, one might simulate a particle-based reaction system (using **KMC** or lattice gas algorithms) and measure coarse observables to compare with RD-PDE predictions [13]. While we did not find a ready-made dataset in sources, relevant references indicate such data is produced in research contexts. If needed, frameworks like **Lattice Boltzmann** (e.g. OpenLB) or **particle-based simulators** can generate datasets: e.g. a lattice-gas automaton producing density fields that coarse-grain to diffusion. One notable open resource is *Physicist's Simulation packages* in Python (e.g. **PySPH** or **LAMMPS** with reactive potentials) – these can output both microscopic trajectories and coarse fields (often in **NetCDF/HDF5** or plain text). In absence of a specific dataset, researchers should refer to *benchmark studies* in coarse-graining (for instance, reaction–diffusion derived from particle interactions [14]) and possibly contact authors for data. Licensing in this niche is generally **academic use** or **CC BY** for any shared data; simulation software is often **GPL/MIT**. In summary, this tier may rely on *simulation libraries and case-study data* rather than an established public dataset.

Tier 2: Fluid Dynamics Benchmarks

Turbulence and Fluid Flow Datasets: Fluid dynamics boasts several large open datasets, especially for turbulence. A flagship resource is the *Johns Hopkins Turbulence Database (JHTDB)*, an open portal to multi-terabyte DNS (direct numerical simulation) data of turbulent flows [15]. JHTDB provides APIs to query velocity/pressure fields from canonical simulations (e.g. isotropic turbulence, channel flow), enabling retrieval in **NetCDF** or binary formats. It is **open-access** (data effectively public domain) and widely used for AI training in super-resolution and flow prediction. Recently, *BLASTNet* (2023) has emerged as a **massive ML-ready fluid dataset** [16] [17]. BLASTNet crowdsourced high-fidelity CFD simulations (over 700 samples from >30 configurations, ~5 TB total) into a consistent format for deep learning [18] [19]. It covers a range of flows (turbulent combustion, multiphase, aerodynamics, etc.) and is available via GitHub/Kaggle as chunks (<100 GB each) [20]. BLASTNet data (spatiotemporal flow fields) are stored in HDF5 or NPZ and come with pre-processing code and even pretrained ML models [21] [22]. The dataset is free and **open-source** (distributed under **CC BY** via Zenodo [23] and Kaggle's terms), explicitly designed for AI benchmarking.

Classical CFD Benchmarks: In addition to the above, numerous smaller benchmarks exist. For example, the *Taylor–Green vortex flow* and *cylinder flow* simulation results are often used; these are sometimes packaged with CFD codes or available in repositories (e.g. the *Fluid Mechanics* community's ERCOFTAC database). A notable curated set is the *AI Turbulence Modeling Dataset* by McConkey et al. (2021), which pairs Reynolds-averaged CFD results with corresponding DNS/LES data for 29 flows [24] [25]. This dataset (available via Kaggle DOI [25]) provides ~895,640 spatial points of flow features and turbulence labels across various geometries (periodic hills, ducts, etc.), in **CSV/HDF5** format. It's structured for immediate use in ML for turbulence model correction [26] [27] and is open-access (**CC0/CC BY** license as per Kaggle's policy). Another example is *UniFoil*, a recent dataset of 500k airfoil simulations covering wide Reynolds/Mach regimes [28] – it includes flow field snapshots (likely HDF5) and was released alongside an arXiv preprint in 2023. Generally, fluid datasets may be hosted by universities (Dataverse, repository links in papers) or on community sites, and they typically have **non-restrictive licenses** (academic use or CC BY).

Data Formats and Use: Fluid datasets are often large (GB–TB scale). Common formats are **HDF5, NetCDF, VTK** (for volumetric fields), or binary dumps readable by visualization tools. Hosting organizations include national labs (e.g. NASA open datasets for aerodynamics), universities (JHTDB at Johns Hopkins, KITWare, etc.), and initiatives like *Stanford's HAI* (for BLASTNet). Many of these datasets are explicitly designed for AI: e.g. BLASTNet provides code to load data into PyTorch, and JHTDB's API enables on-demand subsampling (critical for training). Researchers must mind licensing: e.g. JHTDB data is public domain (but cite the source), BLASTNet is **CC BY 4.0**, and most academic datasets require citing a DOI. In summary, Tier 2 offers a wealth of open fluid dynamics data – from turbulent velocity fields to benchmark flow cases – readily usable for training and validating AI models in CFD.

Tier 3: Active Matter

Active Matter Simulations: Active matter involves self-driven particles (bacteria, flocks, colloids). High-quality datasets here are often simulation-based. One standout is the *Active Matter Dataset* from Polymathic AI's "The Well" repository [29] [30]. This dataset comprises continuum simulations of rod-like active particles in a 2D fluid, modeled via a kinetic theory (as per Maddu *et al.*, J. Comp. Phys. 2024). It provides 225 simulations (about 51 GB) of fields such as concentration, velocity, and orientational order tensor over time [31] [32]. Each trajectory has 81 time steps on a 256×256 grid, saved as sequences of images or arrays. The data captures complex phenomena like energy transfer across scales and isotropic–nematic phase transitions [33], making it ideal for testing learned simulators. The Polymathic dataset is **AI-ready** (they benchmark FNO, U-Net, etc., on it [34]) and likely released under a permissive license (the site implies open academic use, with citation of the associated paper [35]). Data are downloadable via the provided GitHub link [36] (and possibly through HuggingFace for visualization [37]).

Experimental Active Matter Data: Fewer standardized experimental datasets exist, but some are public. For example, video recordings and trajectory data of *self-propelled colloids* or *bacterial colonies* have been shared in supplements of papers (usually under CC licenses). One notable library is *AMEP (Active Matter Evaluation Package)* [38], which is a Python toolkit that includes sample data and analysis routines for particle-based and continuum active matter simulations. While AMEP is a library (GPL-licensed) [39], its included example datasets (e.g. trajectory files) are freely usable for testing AI methods on particle tracking, pattern formation, etc. Additionally, Polymathic's platform suggests active matter data can be visualized and accessed via an API [37], indicating a shareable format like NPZ or HDF5 for the fields.

Licensing and Formats: Simulation datasets (like Polymathic's) are typically released under **Creative Commons Attribution** licenses (the Polymathic active matter data is expected to be CC BY, given its open availability and request to cite [35]). Data files are often in **NumPy .npz** or PyTorch tensor formats for convenience, sometimes accompanied by MP4 GIF visualizations (to qualitatively evaluate swarming patterns). If synthetic *agent-based data* is needed, researchers can generate it using open code (e.g. a Vicsek model implementation) – such code is widely available under MIT/Apache licenses, and resulting data can be shared freely. In summary, Tier 3 provides at least one comprehensive open dataset for active fluids, and more can be created or obtained from open libraries, all generally AI-friendly and openly licensed.

Tier 4: Stochastic Field Theory

Lattice & Field Model Data: Stochastic field theory often refers to fields with randomness, such as stochastic PDEs or statistical field models. A prototypical example is the *Ising model*. There is an open dataset of 2D Ising model configurations at various temperatures available via **PennyLane** (a quantum ML library) [40]. This dataset contains bitstrings representing spin up/down across a lattice, sampled from the thermal Boltzmann distribution of the Ising Hamiltonian [40]. Essentially, it provides many equilibrium configurations labeled by temperature (a classic phase transition learning task). The data are accessible through PennyLane's `datasets` module (no login needed) and are released for **open use** (PennyLane datasets are generally **MIT or CC0** licensed). Similarly, PennyLane offers a *Transverse-Field Ising* dataset [41] that includes quantum properties (like energies, magnetizations) for a 1D quantum Ising chain across field strengths – bridging classical and quantum data for condensed matter.

Stochastic PDE Simulations: While no "standard" dataset was found for, say, stochastic ϕ^4 field theory or stochastic Ginzburg–Landau equations, researchers do create such data. For example, one could simulate the dynamical Φ^4 model (a nonlinear stochastic PDE) and output field snapshots. If any such dataset is public, it would likely be on Zenodo or as supplementary material in math-physics papers (with usage under CC BY). The lack of a widely known dataset suggests you may need to generate data using tools: e.g. **cuSPDE** (a CUDA library for stochastic PDEs) [42] or custom code for stochastic Allen–Cahn equation integration. That said, a recent trend is sharing Monte Carlo samples from lattice field theory – for instance, the *Phi4* 2D lattice data used to train flow-based generative models (as mentioned in some ML papers) is sometimes released on GitHub. If available, such datasets typically come as **HDF5 or NumPy** arrays of field configurations, with labels like coupling constants or time steps. They are meant for AI tasks like phase classification or anomaly detection in field configurations.

Licensing: Datasets in this tier from public repositories (UCI, PennyLane, etc.) use permissive licenses. The UCI *Superconductivity* dataset (although not a field theory, it's physics data) was **CC BY 4.0** [43], and by analogy any UCI-hosted physics dataset (like an Ising one, if it existed there) would be CC BY. PennyLane's provided datasets are free to use for research (with attribution). If you obtain data from an author's site or GitHub, check for a LICENSE file – many default to **MIT or CC BY**. In summary, for stochastic fields, one may leverage available bitstring datasets (Ising) or generate custom data, with minimal licensing hurdles, since these are generally open academic resources.

Tier 5: Quantum Open Systems

Simulated Open Quantum Dynamics: A landmark in this area is the *QDataSet* project (2022) – a collection of 52 datasets of one- and two-qubit systems evolving under various controls and noise environments [44].

QDataSet encompasses ~10,000 samples per dataset, totalling ~14 TB (compressed) of simulation data [45]. It includes quantum state trajectories, control Hamiltonians, measurement outcomes, and noise realizations, all generated to facilitate training of AI models for tasks like quantum control, quantum process tomography, and noise spectroscopy [46] [47]. The data are stored as Python pickled objects and compressed (zipped), and come with documentation and example Jupyter notebooks [48] [49]. This dataset is hosted by University of Technology Sydney (likely via an institutional repository or upon request) and is intended for public use in benchmarking classical vs hybrid algorithms. According to the authors, all data is open (the **QDataSet** paper is Open Access and references open dissemination; presumably the license is **CC BY**, standard for *Scientific Data* journal supplements).

Quantum Trajectories & Experiments: Some open libraries and data cover quantum trajectory methods. For example, an older *Quantum Trajectory Library* (Schack & Brun 1997) was published via Elsevier's data library [50] – it provides C++ code and possibly example output for Monte Carlo wavefunction simulations (though under a CPC license, which might allow academic use of the code) [51]. More modern: **QuTiP** (Quantum Toolbox in Python) has example notebooks simulating qubit decoherence and measurement, from which one can extract datasets (QuTiP is BSD-licensed, so any data you generate can be used freely). However, a unique open dataset of real experimental quantum open-system behavior is emerging: In 2024, researchers released a dataset of a *superconducting qubit's noisy evolution* for system identification (appearing as a "dataset for characterisation of open quantum dynamics" [52]). Similarly, a 2023 work provided data for *quantum noise spectroscopy* (with spectral densities and qubit responses) – these are often hosted on university repositories with DOI and typically under **CC BY**.

Licensing & Format: CERN's Open Data principles have influenced quantum data sharing – e.g., the QDataSet uses open licenses and DOIs. Many quantum open-system datasets (if on figshare/Zenodo) will be **CC0 or CC BY**, as mandated by those platforms. Data formats vary: **HDF5/NPZ** for large numerical arrays (state vectors, density matrices) or plain text for time-series of observables. For instance, QDataSet is distributed as Python pickle files (for easy loading into TensorFlow/PyTorch) [53]. Another example: the *Quantum Noise in IBMQ Devices* dataset (hypothetical) would likely come as JSON or CSV of noise parameters vs time, under IBM's open license (IBM has released some device noise logs under **MIT** in their Qiskit tutorials). In summary, Tier 5 datasets (both simulated and experimental) are increasingly available under open licenses, with large-scale simulated sets like QDataSet leading the way in providing machine-learning-ready quantum dynamical data.

Tier 6: Cosmology

Cosmological Simulation Data: Cosmology has extremely rich public datasets. The *CAMELS project* (Cosmology and Astrophysics with Machine Learning Simulations) is a prime example. As of 2025, CAMELS has over 16,000 simulated universes (N-body and hydrodynamic), totaling >2 PB of data [54] [55]. These simulations span a grid of cosmological parameters (e.g. matter density Ω_m, σ_8) and astrophysical feedback settings, meant to train ML models to infer cosmology from observable fields. CAMELS provides outputs like 3D matter density grids, halo catalogs, galaxy properties, power spectra, etc., in **HDF5** format (with standardized group structures) [56] [54]. Data can be accessed via Globus or institutional servers [57]; the public release (2022) came with a DOI [58]. The licensing is **CC BY 4.0** for the data (per CAMELS documentation) and users are expected to cite the CAMELS release paper. CAMELS is explicitly designed for AI — e.g. the CAMELS Multifield Dataset (CMD) provides multi-scale 2D and 3D field maps for training convolutional networks [59].

Another cornerstone is the *Quijote simulations* (Villaescusa-Navarro et al. 2020), a suite of 44,100 N-body simulations spanning 7,000 cosmologies [60] [61]. Quijote data (~8.5 trillion particles total) are available in halo catalog form and matter field grids (8.5 million halos at z=0 across the suite [61]). The data (several TB) is hosted on institutional servers (with a Python API for access) and shared under **CC BY 4.0** [62]. Both CAMELS and Quijote aim to provide training data for emulators and inference networks — for instance, learning to predict power spectra or generate fast mock universes. Files are typically HDF5 or big binary files (with provided I/O libraries), and both projects supply example code (CAMELS library, PyLIANS, etc. [63] [64]).

Cosmic Surveys and Maps: Observational cosmology data is also broadly available. The *Planck Mission* released all-sky CMB maps in **FITS** format (public domain under ESA policy) – these maps (temperature and polarization) can be used to train CNNs for foreground removal, etc. Large-scale structure surveys like *SDSS* provide galaxy catalogs and spectra (accessible via SDSS CasJobs, under **CC0**), though these are tabular rather than image data. For AI purposes, simulated sky maps are common: e.g. *CosmoDC2* (from LSST DESC) is an open synthetic sky catalog of galaxies (~2 billion objects) with photometry, under a **BSD** license via DOE. Gravitational lensing challenge data (like the *Great3* dataset) are released under CC BY for training deep lensing reconstructions.

Formats & Use: Cosmology datasets often come in specialized formats: e.g. N-body snapshots in Gadget-4 binary, which CAMELS converts to HDF5; healpix maps for CMB; or large CSV/Parquet files for galaxy catalogs. Many have **Python APIs** (e.g. CAMELS has `camels_api`, SDSS has AstroPy interfaces). The data are generally free for non-commercial use; most simulation projects adopt CC BY or public domain dedication. The sheer size means users might use curated subsets or down-sampled versions for AI (CAMELS, for instance, has smaller 2D map sets like *CAMELS Multifield* for easier consumption). In summary, Tier 6 offers abundant data: from simulation suites (CAMELS, Quijote) to observational archives (Planck, SDSS), mostly open-license and provided in formats conducive to machine learning (with some pre-processing).

Tier 7: Magnetohydrodynamics (MHD)

MHD Turbulence Databases: A newly released resource is *TURB-MHD* (Capocci et al. 2025), an open-access database of forced homogeneous MHD turbulence simulations. According to the authors, **TURB-MHD is available via the SMART-Turb portal** [65]. The database contains multiple simulation runs at resolutions up to 2048^3, with different magnetic field strengths and forcing conditions [66] [67]. Key observables (velocity and magnetic field snapshots, current density, energy spectra, etc.) are stored. The portal likely provides data in chunked binary or HDF5 format, given the dataset's size (each 1024^3 snapshot is several GB). This dataset is designed for researchers to download and analyze, and is explicitly *open for download* [65] – implying a **CC BY or CC0 license** (the SMART-Turb site and related databases encourage open use). Such data can train AI models to predict turbulent dynamo behavior or to develop subgrid models for MHD.

Space Plasma Simulations: The space physics community also provides MHD simulation data. For example, the *University of Michigan* has published BATS-R-US MHD simulation results of the **Martian magnetosphere** (solar wind interaction with Mars) [68]. This dataset (accessible via Deep Blue at UMich) includes 3D plasma variables (density, velocity, magnetic field components) on a mesh around Mars, for specific orbital configurations [68]. Files are in plain text or vtk format (as .asc or .vtk) and are open access (the Deep Blue entry indicates **Open Access**). NASA's data portal also lists MHD outputs for Earth's magnetotail reconnection events [69], typically as ASCII tables of field values on certain surfaces, shared under **U.S. Government Public Domain**. These datasets, while domain-specific, are valuable for AI-assisted space weather prediction or feature detection (e.g. identifying reconnection sites).

Astrophysical MHD: Projects like *TIGRESS* (Princeton) have publicly released MHD simulation snapshots for interstellar medium studies [70]. The TIGRESS datasets include fields like gas density, magnetic field, etc., for star-forming region simulations. They can be browsed via a web interface with ivtk (visit) and downloaded (subject to **CC BY 4.0** as per Princeton data release policy). Another example: the *Gorgon MHD simulations* of Earth's magnetosphere (from UK's BAS) are available on CEDA under CC BY (two simulations of steady solar wind conditions, providing global magnetosphere field grids) [71].

In summary, Tier 7 data ranges from controlled homogeneous MHD turbulence to complex space/astrophysical simulations. Data formats vary (HDF5 for large homogeneous cubes, structured text for space physics). Licensing is generally permissive: **CC BY** is common (TURB-MHD and TIGRESS), and U.S. government-produced simulations (e.g. NASA) are **public domain**. These datasets are suitable for AI – e.g. training deep learners to emulate MHD evolution or detect features like current sheets – and they come from reputable sources (universities, NASA, EU agencies) ensuring quality and documentation.

Tier 8: Condensed Matter

Materials Databases: In condensed matter physics and materials science, large open databases of computed properties are foundational. The **Materials Project** is a premier example, containing DFT-computed properties for $\sim$150,000 inorganic compounds (crystal structures, total energies, band gaps, elastic tensors, etc.) [72]. This data is accessible via a REST API and on AWS Open Data, typically in JSON or CSV form. The Materials Project database is explicitly licensed under **Creative Commons Attribution 4.0** [73] [74], meaning it can be freely used for AI training as long as credit is given. Likewise, the **Open Quantum Materials Database (OQMD)** provides over 1.2 million DFT calculations (thermodynamic and structural data) for materials [75] [76]. OQMD's data is also **CC BY 4.0** [62] and accessible via an API without credentials [77]. These databases are not formatted as "trajectories" but rather static entries; however, they are extensively used to train AI models (e.g. graph neural networks for predicting material properties). The data often comes as CSV files (for tabular properties) or CIF files (for crystal structures), which can be featurized for ML. Hosting organizations include LBNL (Materials Project) and Northwestern University (OQMD), with support from DOE; thus they ensure long-term availability and open licensing.

Experimental Property Datasets: An example of a curated dataset for a specific property is the *Superconductivity Dataset* (Hamidieh 2018) in the UCI ML Repository. It contains 21,263 superconductors with 81 engineered features and their critical temperature (T_c) [78] [43]. The data (originally from Japan's SuperCon database) is provided in two CSV files and is licensed **CC BY 4.0** [43]. This dataset is designed for regression tasks (predicting T_c from material features) and has been widely used in materials informatics. Similarly, the *JARVIS* dataset by NIST includes computed and experimental properties (like formation energies, bandgaps, 2D materials data) accessible on their platform (jarvis.nist.gov) under a mix of **public domain** (for NIST data) or CC licenses – JARVIS data is often used for AI (e.g. training models to predict optoelectronic properties).

Lattice Model and Spectroscopy Data: Condensed matter also spans lattice models (quantum many-body simulations) and spectroscopy experiments. Some open datasets include: *Magnetic susceptibility data* for various compounds (e.g. a collection from Springer's database, available for non-commercial use), and *X-ray diffraction images* for many samples (the Materials Data Facility hosts some under CC licenses). Additionally, quantum Monte Carlo or exact diagonalization data for spin chains have been shared in research – e.g. datasets of spin configurations or correlators used to train neural quantum states. One example: the **Quantum Spin Liquids dataset** used in a 2020 study was posted on Zenodo (with wavefunction snapshots for small lattices, CC BY license as per Zenodo defaults).

Summary: Tier 8 offers primarily static but information-rich datasets. The major materials databases (Materials Project, OQMD, AFLOW) are **AI-ready** in the sense that they have millions of consistent entries and are accessible via APIs for building training sets [72] [75]. Licensing is uniformly **CC BY 4.0 or similar** (even the UCI superconductivity data is CC BY [43]). File formats are often **CSV/JSON for properties**, **CIF/XYZ for structures**, and SQL dumps or HDF5 for bulk downloads. These datasets are hosted by organizations (DOE labs, NIST, universities) ensuring robust access. Whether one is training a model to predict material properties or classify phases of matter, these open resources form a solid foundation.

Tier 9: Nuclear / High-Energy Physics (HEP)

High-Energy Physics (HEP) Data: The HEP community has embraced open data in recent years. The **CERN Open Data Portal** provides public datasets from LHC experiments (e.g. CMS, ATLAS) after an embargo period. These include actual collision event records (in ROOT format) for specific runs and simplified NTuples for analysis. All such datasets are released under **Creative Commons CC0 (Public Domain)** [79] [80], allowing unrestricted use. For example, CMS has made 2010–2012 collision data (~hundreds of TB) open; one can download proton-proton collision events with reconstructed particle information to train ML algorithms for particle identification. While the raw format (ROOT) requires domain knowledge, CERN provides *analysis examples and software* openly. In addition, simulation datasets are abundant: the *Higgs ML Challenge (2014)* dataset, a CSV of 800k simulated events labeled as Higgs signal or background, is still hosted on UCI (as “HIGGS” dataset) and is a common ML benchmark. That dataset is open (public domain) and in a simple tabular format (features like jet momenta, etc.). Another popular one is the *UCI SUSY dataset* (5 million events for supersymmetry signal vs background), also CSV and public domain. These were explicitly designed for AI competitions and are easy to ingest.

Nuclear Physics Data: Nuclear physics data tends to be more tabular (cross-sections, energy levels). For instance, the **Brookhaven National Nuclear Data Center (NNDC)** provides datasets like ENDF (evaluated nuclear reaction cross-sections) which are openly available (typically public domain as U.S. government data). While not curated for ML, one could use them to train models to interpolate cross-sections. In terms of simulations, *Lattice QCD* configurations (gauge field configurations) are sometimes shared on the **International Lattice Data Grid (ILDG)**. These are large binary lattices of quark/gluon fields. Access requires joining a virtual organization, but the data is essentially open to researchers and often falls under academic use terms (no commercial use). There have been efforts to use ML on lattice data, and some configurations (like $SU(2)$ toy models or 2D QCD) have been posted on GitHub with CC BY licenses by researchers. Additionally, neutrino physics has open datasets: e.g. *IceCube* releases some event data (through DOE portals) for public outreach and Kaggle challenges (such data is CC0).

Formats & Tools: HEP event data usually uses **ROOT** format (binary, with C++/Python APIs). CERN Open Data provides virtual machines and tutorials to read these into numpy/pandas if needed. The simpler ML challenge data are in **CSV**. Nuclear data tables are often **plaintext tables or JSON** (NNDC’s APIs can return JSON). Many HEP open datasets are accompanied by code under **GPL or Apache 2.0** to help parse them (for example, Kaggle provided Python notebooks for the Higgs dataset). Licenses: as noted, CERN Open Data is

CC0 [79] , older ML competition data are effectively public domain or CC BY, and government nuclear databases are public domain. One should, however, cite the source or relevant paper.

In summary, Tier 9 offers both raw experimental data and simulation data. On the experimental side, you could literally train an algorithm on real collision events thanks to CERN's CC0 releases. On the simulation side, curated datasets like the Higgs challenge (for particle classification) or public Pythia/GEANT4 samples are available. There's also *Open Data from Belle II* and others under similar open licenses. These datasets vary in size from a few MB (UCI Higgs CSV ~2GB) to multi-TB (full LHC runs), and they're hosted by major labs (CERN, FNAL, etc.) ensuring stable access. AI researchers have actively used these: e.g. deep learning for jet tagging using open simulated jet datasets (some of which are on Zenodo with CC BY licenses). Thus, the nuclear/HEP domain, while specialized, is well-represented by open datasets conducive to AI, with minimal licensing barriers.

Quantum Information Overlays

Quantum Computing Datasets: As quantum computing intersects with these domains, new "overlays" of data are emerging. One example is datasets from quantum hardware solving physics problems. In early 2024, a team released a **Quantum MIS (Maximum Independent Set) dataset** using a Rydberg atom quantum simulator [81] . This dataset comprises ~582,000 experimental shots of a 141-atom Rydberg array finding MIS solutions on random graphs, corresponding to 733,853 different graphs [82] [83] . Provided data include raw fluorescence images of the atom array and binary outcome strings (which atoms were excited) for each run, along with the graph definitions [84] [85] . The data are meant for benchmarking classical algorithms and quantum improvements, and are published via *Scientific Data* (thus openly accessible, likely under CC BY license). This is a true quantum-classical overlay: a quantum device generating a dataset for a classically hard problem, enabling AI analysis of quantum experiment performance.

Another overlay example is **QCircuitNet** (2023), a large-scale dataset of quantum circuit designs and their properties [86] . Although primarily for quantum algorithm design, it bridges to classical ML by providing a hierarchy of circuit representations that one can learn to optimize. It likely contains many sample quantum circuits (in QASM format) labeled with metrics like gate counts or output distributions. This data (per arXiv) is geared toward training models to assist quantum algorithm development; availability would be via the authors' GitHub under an open-source license (commonly MIT).

Quantum-enhanced Sensing Data: There are also instances of datasets where quantum sensors provide enhanced data for classical domains. For example, NV-center quantum magnetometer readings mapping magnetic fields (some groups have shared such data to apply AI for denoising). These are usually in CSV or HDF5 and under academic licenses.

Licensing and Format: Quantum overlay datasets follow the open science trend. The Rydberg MIS dataset, for instance, is published with a DOI and should be CC BY (as mandated by Nature's data journal). Data are often split into **raw data (images)** and **processed data (bitstrings, etc.)** with documentation on interpretation [84] [85] . For quantum circuits, if on GitHub, expect **MIT or Apache 2.0**. PennyLane's **Quantum Datasets** (mentioned in Tier 4 and Tier 5) also fall here; their "What is a quantum dataset?" guide [87] elaborates how collections of quantum states or processes can be treated as datasets for ML. PennyLane's own dataset repository (Xanadu's Quantum Dataset library) is Apache-licensed, and it includes examples like quantum chemistry Hamiltonians and photonic circuit data – all freely usable.

In summary, *QI overlays* refer to datasets that mix quantum computational processes with classical data analysis. These are all **publicly available**, since they cater to a nascent community of hybrid quantum-classical researchers. As of now, such datasets are fewer but rapidly growing. They come from experimental groups (sharing quantum experiment results) and from quantum software companies (providing simulated data for algorithm development). Formats range from image files (for quantum experiment pics) to qubit-state CSVs and QASM code files. Licenses tend toward **CC BY or Apache/MIT**. The key is that these overlays enable AI to learn from or about quantum systems, thus linking the quantum domain back to classical AI training – an important frontier that this tier addresses.

Sources:

- Tier 0: PDEBench dataset description [88] [4] ; NIST Noise dataset description [5] [7] .
- Tier 1: PDEBench (diffusion-reaction) [10] ; Data.gov info on coarse-graining (context) [13] .
- Tier 2: Stanford HAI on BLASTNet [16] [17] ; BLASTNet site info [20] [23] ; Turbulence modeling dataset (McConkey 2021) [24] [27] ; JHTDB portal [15] .
- Tier 3: Polymathic AI active matter dataset [30] [32] ; Polymathic metrics (AI performance) [34] .
- Tier 4: PennyLane Ising dataset info [40] ; UCI Superconductivity dataset (features, license) [89] [43] .
- Tier 5: QDataSet overview [44] [47] ; Quantum trajectory library abstract [50] .
- Tier 6: CAMELS documentation [54] [55] ; CAMELS description (collaborative effort) [90] ; Quijote simulations paper [60] [61] .
- Tier 7: TURB-MHD open availability [65] ; MHD turbulence details [66] ; Mars MHD dataset info [68] .
- Tier 8: Materials Project license [91] ; OQMD license [62] ; UCI Superconductivity dataset info [92] [43] .
- Tier 9: CERN Open Data terms [79] ; CERN Open Data (ATLAS) [80] ; Quoted Higgs dataset reference (UCI via Kaggle) [93] .
- QI Overlays: Rydberg MIS dataset (Sci Data 2024) [81] [84] ; PennyLane quantum data blurb [87] .

---

[1] [2] [3] [4] [10] [88] PDEBench Datasets - SciML PDE Benchmark
https://darus.uni-stuttgart.de/dataset.xhtml?persistentId=doi:10.18419/darus-2986

[5] [6] [7] Noise Datasets for Evaluating Deep Generative Models - Catalog
https://catalog.data.gov/dataset/noise-datasets-for-evaluating-deep-generative-models

[8] Dataset - Catalog
https://catalog.data.gov/dataset/?tags=fractional-brownian-motion&res_format=TEXT

[9] SDE simulation: creating synthetic datasets using SDEs - scikit-FDA
https://fda.readthedocs.io/en/latest/auto_examples/datasets/plot_sde_simulation.html

[11] datavorous/Gray-Scott-Reaction-Diffusion-Model - GitHub
https://github.com/datavorous/Gray-Scott-Reaction-Diffusion-Model

[12] thermodynamics - Dataset - Catalog - Data.gov
https://catalog.data.gov/dataset/?tags=thermodynamics

[13] Data-driven dynamical coarse-graining for condensed matter systems
https://pubs.aip.org/aip/jcp/article/160/2/024108/2932835/Data-driven-dynamical-coarse-graining-for

[14] Traveling waves in a coarse-grained model of volume-filling cell ...
https://onlinelibrary.wiley.com/doi/10.1111/sapm.12635

[15] Johns Hopkins Turbulence Database JHTDB
https://turbulence.pha.jhu.edu/

[16] [17] [18] [19] BLASTNet – The First Large Machine Learning Dataset for Fundamental Fluid Dynamics | Stanford HAI
https://hai.stanford.edu/news/blastnet-first-large-machine-learning-dataset-fundamental-fluid-dynamics

[20] [21] [22] [23] BLASTNet
https://blastnet.github.io/

[24] [25] [26] [27] A curated dataset for data-driven turbulence modelling | Scientific Data
https://www.nature.com/articles/s41597-021-01034-2?error=cookies_not_supported&code=0d0f2b07-5877-42fd-8a72-c95189fafcf6

[28] UniFoil: A Universal Dataset of Airfoils in Transitional and Turbulent ...
https://arxiv.org/html/2505.21124v3

[29] [30] [31] [32] [33] [34] [35] [36] [37] active_matter - The Well
https://polymathic-ai.org/the_well/datasets/active_matter/

[38] [39] AMEP: The active matter evaluation package for Python
https://www.sciencedirect.com/science/article/pii/S0010465524004065

[40] Ising - PennyLane
https://pennylane.ai/datasets/ising

[41] Transverse-field Ising model - PennyLane
https://pennylane.ai/datasets/transverse-field-ising-model

[42] cuPSS: a package for pseudo-spectral integration of stochastic PDEs
https://arxiv.org/html/2405.02410v1

[43] [78] [89] [92] UCI Machine Learning Repository
http://archive.ics.uci.edu/ml/datasets/Superconductivty+Data

[44] [45] [46] [47] [48] [49] [53] QDataSet, quantum datasets for machine learning | Scientific Data
https://www.nature.com/articles/s41597-022-01639-1?error=cookies_not_supported&code=ad0ca193-f74e-48f4-8b7d-4ddc9dafe45e

[50] [51] A C++ library using quantum trajectories to solve quantum master equations - Elsevier BV
https://elsevier.digitalcommonsdata.com/datasets/56t69wyc2t/1

[52] Characterization and control of open quantum systems beyond ...
https://www.nature.com/articles/s41534-020-00332-8

[54] [55] [56] [57] [63] [64] [90] CAMELS — CAMELS 0.1 documentation
https://camels.readthedocs.io/en/latest/

[58] [2201.01300] The CAMELS project: public data release - arXiv
https://arxiv.org/abs/2201.01300

[59] The CAMELS Multifield Dataset: Learning the Universe's ... - arXiv
https://arxiv.org/abs/2109.10915

[60] [1909.05273] The Quijote simulations - arXiv
https://arxiv.org/abs/1909.05273

[61] [1909.05273] The Quijote simulations - ar5iv
https://ar5iv.labs.arxiv.org/html/1909.05273

[62] OQMD
https://oqmd.org/

[65] [66] [67] TURB-MHD: an open-access database of forced homogeneous magnetohydrodynamic turbulence
https://arxiv.org/html/2504.10755v1

[68] Data Set | Dataset for Multispecies MHD Simulations of the Crustal ...
https://deepblue.lib.umich.edu/data/concern/data_sets/wp988k86s?locale=en

[69] Characteristics of Reconnection Sites and Fast Flow Channels in an ...
https://data.nasa.gov/dataset/characteristics-of-reconnection-sites-and-fast-flow-channels-in-an-mhd-simulation-simulati

[70] MHD_PI datasets — TIGRESS public data release
https://princetonuniversity.github.io/astro-tigress/read_data_3-MHD_PI.html

[71] Two Gorgon Global-MHD simulations of steady solar wind conditions
https://data.bas.ac.uk/full-record.php?id=GB/NERC/BAS/PDC/01531

[72] Materials Project - Wikipedia
https://en.wikipedia.org/wiki/Materials_Project

[73] GNoME Database License and Terms of Use - Materials Project
https://next-gen.materialsproject.org/about/terms

[74] [91] Login - Materials Project
https://legacy.materialsproject.org/dashboard

[75] OQMD: The Open Quantum Materials Database - MateriApps
https://ma.issp.u-tokyo.ac.jp/en/app/6334

[76] The Open Quantum Materials Database (OQMD) - Nature
https://www.nature.com/articles/npjcompumats201510

[77] OQMD API - The Open Quantum Materials Database
https://www.oqmd.org/api/

[79] CERN Open Data Terms of Use
https://opendata.cern.ch/terms-of-use

[80] ATLAS Open Data
https://atlas.cern/Resources/Opendata

[81] [82] [83] [84] [85] Quantum computing dataset of maximum independent set problem on king lattice of over hundred Rydberg atoms | Scientific Data
https://www.nature.com/articles/s41597-024-02926-9?error=cookies_not_supported&code=36d0e424-9c42-433f-9fa7-ea7f3e9aec34

[86] A Large-Scale Hierarchical Dataset for Quantum Algorithm Design
https://arxiv.org/abs/2410.07961

[87] What is a Quantum Dataset? - PennyLane
https://pennylane.ai/datasets/what-is-a-quantum-dataset

[93] Superconductor Dataset - Kaggle
https://www.kaggle.com/datasets/munumbutt/superconductor-dataset