

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-05-01	yes	Jet Classification	<a href="https://github.com/fastnucleonics/fastnucleonics/tree/main/jet-classify">https://github.com/fastnucleonics/fastnucleonics/tree/main/jet-classify</a>	Particle Physics	Real-time classification of particle jets using HL-LHC simulation features	classification, real-time ML, jet tagging, QKeras	This benchmark evaluates ML models for real-time classification of particle jets using high-level features derived from simulated LHC data. It includes both full-precision \nand quantized models optimized for FPGA deployment.	Classification	Real-time model performance

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-05-01	yes	Irregular Sensor Data Compression	<a href="https://github.com/fasterthanplain/sensor-data-compression">https://github.com/fasterthanplain/sensor-data-compression</a>	Particle Physics / fast	Real-time compression of sparse sensor data with autoencoders	compression, autoencoder, sparse data, irregular sampling	This benchmark addresses lossy compression of irregularly sampled sensor data from \nparticle detectors using real-time autoencoder architectures, targeting latency-critical \napplications in physics experiments.	Compression	Reconstruction compression

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-05-01	yes	Beam Control	<a href="https://github.com/fasacscience/tree/main/beamcontrol">https://github.com/fasacscience/tree/main/beamcontrol</a>	Accelerator Learning Magnets	Reinforcement learning control of accelerator beam position	RL, beam stabilization, control systems, simulation	Beam Control explores real-time reinforcement learning strategies for maintaining stable beam trajectories in particle accelerators. The benchmark is based on the BOOSTR environment for accelerator simulation.	Control	Policy per simulated control

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type	
2024-07-08	yes	Ultrafast jet classification at the HL-LHC	<a href="https://arxiv.org/pdf/2402.11876v1.pdf">https://arxiv.org/pdf/2402.11876v1.pdf</a>	2402.11876	Physics	FPGA-optimized real-time jet origin classification at the HL-LHC	jet classification, FPGA, quantization-aware training, Deep Sets, Interaction Networks	Demonstrates three ML models (MLP, Deep Sets, Interaction Networks) optimized for FPGA deployment with O(100 ns) inference using quantized models and hls4ml, targeting real-time jet tagging in the L1 trigger environment at the high-luminosity LHC. Data is available on Zenodo DOI:10.5281/zenodo.3602260.	Classification	Real-time inference on the FPGA

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-10-15	yes	Quench detection	<a href="https://indico.cern.ch/event/1387540/contributions/6153618/attachments/294844/5182075/fast_nof_magnets_2024.pdf">https://indico.cern.ch/event/1387540/contributions/6153618/attachments/294844/5182075/fast_nof_magnets_2024.pdf</a>	Avenio-1387540/Contributions/6153618/Magnets	Real-time detection of superconducting magnet quenches using ML	tion, autoencoder, anomaly detection, real-time	real-time quench detection using unsupervised and RL approaches, combining multi-modal sensor data (BPM, power supply, acoustic), operating on kHz-MHz streams with anomaly detection and frequency-domain features.	Quench localization	Real-time detection modal sens

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-10-15	yes	DUNE	<a href="https://indico.fnal.gov/event/6599/contributions/30442/">https://indico.fnal.gov/event/6599/contributions/30442/</a>	Particle Physics for DUNE DAQ time-series data	indico.fnal.gov	DUNE, time-series, real-time, trigger	ML methods to time-series data from DUNE detectors, exploring trigger-level anomaly detection and event selection with low latency constraints.	Trigger selection, anomaly detection	Low-latency detection

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2025-01-08	yes	Intelligent experiments through real-time AI	<a href="https://arxiv.org/pdf/2501.04845">https://arxiv.org/pdf/2501.04845</a>	Trigger classification and Detectors; Nuclear Physics; Particle Physics	Real-time FPGA-based triggering and detector control for sPHENIX and future EIC	FPGA, Graph Neural Network, hls4ml, real-time inference, detector control	Research and Development demonstrator for real-time processing of high-rate tracking data from the sPHENIX detector (RHIC) and future EIC systems. Uses GNNs with hls4ml for FPGA-based trigger generation to identify rare events (heavy flavor, DIS electrons) within 10 $\mu$ s latency. Demonstrated improved accuracy and latency on Alveo/FELIX platforms.	Trigger classification, Detector control, Real-time inference	Low-latency inference on

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2025-01-09	yes	Neural Architecture Codeign for Fast Physics Applications	<a href="https://arxiv.org/abs/2501.05511">https://arxiv.org/abs/2501.05511</a>	Physics; Materials Science; Particle Physics	Automated neural architecture search and hardware-efficient model codesign for fast physics applications	neural architecture search, FPGA deployment, quantization, pruning, hls4ml	Introduces a two-stage neural architecture codesign (NAC) pipeline combining global and local search, quantization-aware training, and pruning to design efficient models for fast Bragg peak finding and jet classification, synthesized for FPGA deployment with hls4ml. Achieves >30x reduction in BOPs and sub-100 ns inference latency on FPGA.	Classification, Peak finding	Hardware-optimization, latency inf



Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Typ
2024-06-24	yes	Smart Pixels for LHC	<a href="https://arxiv.org/abs/2406.14860">https://arxiv.org/abs/2406.14860</a>	Physics; Instrumentation and Detectors	On-sensor, in-pixel ML filtering for high-rate LHC pixel detectors	smart pixel, on-sensor inference, data reduction, trigger	Presents a 256x256-pixel ROIC in 28 nm CMOS with embedded 2-layer NN for cluster filtering at 25 ns, achieving 54-75% data reduction while maintaining noise and latency constraints. Prototype consumes ~300 $\mu$ W/pixel and operates in combinatorial digital logic.	Image Classification, Data filtering	On-chip, inference; da

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type	
2023-10-03	yes	HEDM BraggNN	<a href="https://arxiv.org/abs/2008.08198">https://arxiv.org/abs/2008.08198</a>	2008.08198	arXiv.org	Fast Bragg peak analysis using deep learning in diffraction microscopy	BraggNN, diffraction, peak finding, HEDM	Uses BraggNN, a deep neural network, for rapid Bragg peak localization in high-energy diffraction microscopy, achieving ~13x speedup compared to Voigt-based methods while maintaining sub-pixel accuracy.	Peak detection	High-throughput localization
2023-12-03	yes	4D-STEM	<a href="https://openreview.net/forum?id=Uj46%3A1-57jz8N0W">https://openreview.net/forum?id=Uj46%3A1-57jz8N0W</a>	Uj46%3A1-57jz8N0W	arXiv.org	Real-time ML for scanning transmission electron microscopy	4D-STEM, electron microscopy, real-time, image processing	Proposes ML methods for real-time analysis of 4D scanning transmission electron microscopy datasets; framework details in progress.	Image Classification, Streamed data inference	Real-time microscopy

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2023-12-05	yes	In-Situ High-Speed Computer Vision	<a href="https://arxiv.org/abs/2312.00128">https://arxiv.org/abs/2312.00128</a>	Real-time Plasma	Real-time image classification for in-situ plasma diagnostics	plasma, in-situ vision, real-time ML	Applies low-latency CNN models for image classification of plasma diagnostics streams; supports deployment on embedded platforms.	Image Classification	Real-time inference

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2020-01-01	yes	BenchCouncil AIBench	<a href="https://www.benchmarkcouncil.org/">https://www.benchmarkcouncil.org/</a>	<a href="https://www.benchmarkcouncil.org/AIBench/">https://www.benchmarkcouncil.org/AIBench/</a>	End-to-end AI benchmarking across micro, component, and application levels	benchmarking, AI systems, application-level evaluation	AIBench is a comprehensive benchmark suite that evaluates AI workloads at different levels (micro, component, application) across hardware systems—covering image generation, object detection, translation, recommendation, video prediction, etc.	Training, Inference, End-to-end AI workloads	System-level load performance

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2020-01-01	yes	BenchCouncil Big-DataBench	<a href="https://www.benchmarking.org/">https://www.benchmarking.org/</a>	<a href="https://www.benchmarking.org/BigDataBench/">https://www.benchmarking.org/BigDataBench/</a>	Big data and AI benchmarking across structured, semi-structured, and unstructured data workloads	big data, AI benchmarking, data analytics	BigDataBench provides benchmarks for evaluating big data and AI workloads with realistic datasets (13 sources) and pipelines across analytics, graph, warehouse, NoSQL, streaming, and AI.	Data preprocessing, Inference, End-to-end data pipelines	Data processing, model inference, performance at scale

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2021-10-20	yes	MLPerf HPC	<a href="https://github.com/mlcommons/hpc">https://github.com/mlcommons/hpc</a>	CosmoFlow, hpc Climate, Protein Structure, Catalysis	Scientific ML training and inference on HPC systems	HPC, training, inference, scientific ML	MLPerf HPC introduces scientific model benchmarks (e.g., CosmoFlow, DeepCAM) aimed at large-scale HPC evaluation with >10x performance scaling through system-level optimizations.	Training, Inference	Scaling training to accuracy of

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2023-06-01	yes	MLCommons Science	<a href="https://github.com/mlcommons/science">https://github.com/mlcommons/science</a>	Earthquake, Satellite Image, Drug Discovery, Electron Microscope, CFD	AI benchmarks for scientific applications including time-series, imaging, and simulation	science AI, benchmark, MLCommons, HPC	MLCommons Science assembles benchmark tasks with datasets, targets, and implementations across earthquake forecasting, satellite imagery, drug screening, electron microscopy, and CFD to drive scientific ML reproducibility.	Time-series analysis, Image classification, Simulation surrogate modeling	Inference simulation generalization

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2021-07-05	yes	LHC New Physics Dataset	<a href="https://arxiv.org/pdf/2107.02157v1.pdf">https://arxiv.org/pdf/2107.02157v1.pdf</a>	Physics; Real-time Triggering	Real-time LHC event filtering for anomaly detection using proton collision data	anomaly detection, proton collision, real-time inference, event filtering, unsupervised ML	A dataset of proton-proton collision events emulating a 40 MHz real-time data stream from LHC detectors, pre-filtered on electron or muon presence. Designed for unsupervised new-physics detection algorithms under latency/bandwidth constraints.	Anomaly detection, Event classification	Unsupervised detection tency and constraints



Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2023-07-17	yes	MLCommons Medical AI	<a href="https://github.com/mlcommons/medical-ai">https://github.com/mlcommons/medical-ai</a>	Healthcare, medical AI	Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data	medical AI, federated evaluation, privacy-preserving, fairness, healthcare benchmarks	The MLCommons Medical AI working group develops benchmarks, best practices, and platforms (MedPerf, GaNDLF, COFE) to accelerate robust, privacy-preserving AI development for healthcare. MedPerf enables federated testing of clinical models on diverse datasets, improving generalizability and equity while keeping data onsite.	Federated evaluation, Model validation	Clinical accuracy, generalizability, privacy compliance

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-10-28	yes	CaloChallenge 2022	<a href="http://arxiv.org/abs/2410.21611">http://arxiv.org/abs/2410.21611</a>	LHC21Calorimeter; Particle Physics	Fast generative-model-based calorimeter shower simulation evaluation	calorimeter simulation, generative models, surrogate modeling, LHC, fast simulation	The Fast Calorimeter Simulation Challenge 2022 assessed 31 generative-model submissions (VAEs, GANs, Flows, Diffusion) on four calorimeter shower datasets; benchmarking shower quality, generation speed, and model complexity.	Surrogate modeling	Simulation speed, efficiency

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
ongoing	yes	Papers With Code-SOTA Platform	<a href="https://paperswithcode.com/">https://paperswithcode.com/</a>	General ML; All domains	Open platform tracking state-of-the-art results, benchmarks, and implementations across ML tasks and papers	leaderboard, benchmarking, reproducibility, open-source	Papers With Code (PWC) aggregates benchmark suites, tasks, and code across ML research: 12,423 benchmarks, 5,358 unique tasks, and 154,766 papers with code links. It tracks SOTA metrics and fosters reproducibility.	Multiple (Classification, Detection, NLP, etc.)	Model across tasks (F1, BLEU, etc.)
2022-01-01	yes	Codabench	<a href="https://www.codabench.org/">https://www.codabench.org/</a>	General ML; Multiple	Open-source platform for organizing reproducible AI benchmarks and competitions	benchmark platform, code submission, competitions, meta-benchmark	Codabench (successor to CodaLab) is a flexible, easy-to-use, reproducible API platform for hosting AI benchmarks and code-submission challenges. It supports custom scoring, inverted benchmarks, and scalable public or private queues	Multiple	Model performance on datasets

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2021-09-27	yes	Sabath - SBI-FAIR	<a href="https://sbi-fair.github.io/docs/software/sabath/">https://sbi-fair.github.io/docs/software/sabath/</a>	Systems; Metadata	FAIR meta-data framework for ML-driven surrogate workflows in HPC systems	meta-benchmark, metadata, HPC, surrogate modeling	Sabath is a meta-data framework from the SBI-FAIR group (UTK, Argonne, Virginia) facilitating FAIR-compliant benchmarking and surrogate execution logging across HPC systems	Systems benchmarking	Metadata producible flows
2022-10-13	yes	PDEBench	<a href="https://github.com/pdebench/pdebench">https://github.com/pdebench/pdebench</a>	CFD; PDE Modeling	Benchmark suite for ML-based surrogates solving time-dependent PDEs	PDEs, CFD, scientific ML, surrogate modeling, NeurIPS	PDEBench offers forward/inverse PDE tasks with large ready-to-use datasets and baselines (FNO, U-Net, PINN), packaged via a unified API. It won the SimTech Best Paper Award 2023.	Supervised Learning	Time-dependent modeling; accuracy

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-12-03	yes	The Well	<a href="https://polymathic-ai.org/the_well/">https://polymathic-ai.org/the_well/</a>	biological systems, fluid dynamics, acoustic scattering, astrophysical MHD	Foundation model + surrogate dataset spanning 16 physical simulation domains	surrogate modeling, foundation model, physics simulations, spatiotemporal dynamics	A 15 TB collection of ML-ready physics simulation datasets (HDF5), covering 16 domains—from biology to astrophysical magnetohydrodynamic simulations—with unified API and metadata. Ideal for training surrogate and foundation models on scientific data.	Supervised Learning	Surrogate modeling, prediction
2024-10-31	yes	LLM-Inference-Bench	<a href="https://github.com/argunlcf/LLM-Inference-Bench">https://github.com/argunlcf/LLM-Inference-Bench</a>	LLM-HPC/inference	Hardware performance benchmarking of LLMs on AI accelerators	LLM, inference benchmarking, GPU, accelerator, throughput	A suite evaluating inference performance of LLMs (LLaMA, Mistral, Qwen) across diverse accelerators (NVIDIA, AMD, Intel, SambaNova) and frameworks (vLLM, DeepSpeed-MII, etc.), with an interactive dashboard and per-platform metrics.	Inference Benchmarking	Inference latency, hardware utilization

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2023-12-12	yes	SGLang Framework	<a href="https://github.com/sglang/sglang/tree/main/benchmark">https://github.com/sglang/sglang/tree/main/benchmark</a>	LLM Vision	Fast serving framework for LLMs and vision-language models	LLM serving, vision-language, RadixAttention, performance, JSON decoding	A high-performance open-source serving framework combining efficient backend runtime (RadixAttention, batching, quantization) and expressive frontend language, boosting LLM/VLM inference throughput up to ~3x over alternatives.	Model serving framework	Serving JSON/task latency

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2023-09-12	yes	vLLM Inference and Serving Engine	<a href="https://github.com/vllm-project/vllm/tree/main/benchmarks">https://github.com/vllm-project/vllm/tree/main/benchmarks</a>	LLM; Inference	High-throughput, memory-efficient inference and serving engine for LLMs	LLM inference, PagedAttention, CUDA graph, streaming API, quantization	vLLM is a fast, high-throughput, memory-efficient inference and serving engine for large language models, featuring PagedAttention, continuous batching, and support for quantized and pipelined model execution. Benchmarks compare it to TensorRT-LLM, SGLang, and others	Inference Benchmarking	Throughput, memory efficiency

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2022-06-22	yes	vLLM Performance Dashboard	<a href="https://simon-mo-workspace.observablehq.com/@vllm">https://simon-mo-workspace.observablehq.com/@vllm</a>	LLM; Inference	Interactive dashboard showing inference performance of vLLM	Dashboard, Throughput visualization, Latency analysis, Metric tracking	A live visual dashboard for vLLM showcasing throughput, latency, and other inference metrics across models and hardware configurations.	Performance visualization	Throughput hardware u
2022-04-01	yes	Nixtla Neural-Forecast	<a href="https://github.com/Nixtla">https://github.com/Nixtla</a>	Time-series forecasting; General ML	High-performance neural forecasting library with >30 models	time-series, neural forecasting, NBEATS, NHITS, TFT, probabilistic forecasting, usability	NeuralForecast offers scalable, user-friendly implementations of over 30 neural forecasting models (NBEATS, NHITS, TFT, DeepAR, etc.), emphasizing quality, usability, interpretability, and performance.	Time-series forecasting	Forecast a terpretabil



Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2023-06-01	yes	Nixtla Neural Forecast NHITS	<a href="https://github.com/Nixtla/nixtla-forecast">https://github.com/Nixtla/nixtla-forecast</a>	Time-series, General ML	Official NHITS implementation for long-horizon time series forecasting	NHITS, long-horizon forecasting, neural interpolation, time-series	NHITS (Neural Hierarchical Interpolation for Time Series) is a state-of-the-art model that improved accuracy by ~25% and reduced compute by 50x compared to Transformer baselines, using hierarchical interpolation and multi-rate sampling	Time-series forecasting	Accuracy, efficiency for
2023-10-03	yes	Nixtla Neural Forecast TimeLLM	<a href="https://github.com/Nixtla/nixtla-forecast">https://github.com/Nixtla/nixtla-forecast</a>	Time-series, General ML	Reprogramming LLMs for time series forecasting	Time-LLM, language model, time-series, reprogramming	Time-LLM uses reprogramming layers to adapt frozen LLMs for time series forecasting, treating forecasting as a language task.	Time-series forecasting	Model reuse, few-shot fo

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2023-10-05	yes	Nixtla Neural Forecast TimeGPT	<a href="https://github.com/Nixtla/TimeGPT">https://github.com/Nixtla/TimeGPT</a>	Time-series, Gener- eral ML	Time-series founda- tion model "TimeGPT" for forecasting and anomaly detection	TimeGPT, foundation model, time- series, genera- tive model	TimeGPT is a transformer-based generative pre- trained model on 100B+ time series data for zero-shot forecasting and anomaly detection via API.	Time-series forecasting, Anomaly detection	Zero-shot anomaly d
2025-03-03	yes	HDR ML Anomaly Challenge- Gravi- tational Waves	<a href="https://www.codabench.org/challenges/2025-03-03">https://www.codabench.org/challenges/2025-03-03</a>	Astronomy, Time-series	Detecting anomalous gravitational- wave sig- nals from LIGO/Virgo datasets	anomaly detec- tion, gravita- tional waves, astrophysics, time-series	A benchmark for detecting anoma- lous transient gravitational-wave signals, includ- ing "unknown- unknowns," using preprocessed LIGO time-series at 4096 Hz. Com- petitors submit inference models on Codabench for continuous 50 ms segments from dual interferometers.	Anomaly detection	Novel event in physical

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type	
2025-03-03	yes	HDR ML Anomaly Challenge- Butterfly	<a href="https://www.codabench.org/competitions/3704/">https://www.codabench.org/competitions/3704/</a>	Codabench/ Image/CV	Image/CV	detecting hybrid but- terflies via image anomaly detection in genomic- informed dataset	anomaly de- tection, com- puter vision, genomics, but- terfly hybrids	Image-based chal- lenge for detecting butterfly hybrids in microscopy- driven species data. Participants evaluate models on Codabench using image segmenta- tion/classification.	Anomaly detection	Hybrid det- ectological sys-
2025-03-03	yes	HDR ML Anomaly Challenge- Sea Level Rise	<a href="https://www.codabench.org/competitions/3228/">https://www.codabench.org/competitions/3228/</a>	Codabench/ Time-series, Im- age/CV	Image/CV	detecting anomalous sea- level rise and flooding events via time-series and satellite imagery	anomaly detec- tion, climate science, sea- level rise, time-series, remote sensing	A challenge com- bining North Atlantic sea-level time-series and satellite imagery to detect flooding anomalies. Models submitted via Codabench.	Anomaly detection	Detection mental and

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2025-01-24	yes	Single Qubit Readout on QICK System	<a href="https://github.com/facsimile-camp/quantum-readout">https://github.com/facsimile-camp/quantum-readout</a>	Quantum Computing	ml-Real-time single-qubit state classification using FPGA firmware	qubit readout, hls4ml, FPGA, QICK	Implements real-time ML models for single-qubit readout on the Quantum Instrumentation Control Kit (QICK), using hls4ml to deploy quantized neural networks on RFSoc FPGAs. Offers high-fidelity, low-latency quantum state discrimination.	Classification	Single-shot interference tasks
2023-11-20	yes	GPQA A Graduate Level Google Proof Question and Answer Benchmark	<a href="https://arxiv.org/abs/2311.12021">https://arxiv.org/abs/2311.12021</a>	Science (Biology, Physics, Chemistry)	Graduate-level, expert-validated multiple-choice questions hard even with web access	Google-proof, multiple-choice, expert reasoning, science QA	Contains 448 challenging questions written by domain experts, with expert accuracy at 65% (74% discounting clear errors) and non-experts reaching just 34%. GPT-4 baseline scores ~39%—designed for scalable oversight evaluation.	Multiple choice	Scientific knowledge

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-12-13	yes	SeafloorAI	<a href="https://neurips.cc/virtual/2024/poster/97432">https://neurips.cc/virtual/2024/poster/97432</a>	May/2024/Poster/97432 Vision-Language	Large-scale vision-language dataset for seafloor mapping and geological classification	sonar imagery, vision-language, seafloor mapping, segmentation, QA	A first-of-its-kind dataset covering 17,300 sq km of seafloor with 696K sonar images, 827K segmentation masks, and 696K natural-language descriptions plus ~7M QA pairs—designed for both vision and language-based ML models in marine science	Image segmentation, Vision-language QA	Geospatial standing, reasoning
2024-12-13	yes	SuperCon3D	<a href="https://neurips.cc/virtual/2024/poster/97553">https://neurips.cc/virtual/2024/poster/97553</a>	May/2024/Poster/97553 Superconductivity	Dataset and models for predicting and generating high-Tc superconductors using 3D crystal structures	superconductivity, crystal structures, equivariant GNN, generative models	SuperCon3D introduces 3D crystal structures with associated critical temperatures (Tc) and two deep-learning models: SODNet (equivariant graph model) and DiffCSP-SC (diffusion generator) designed to screen and synthesize high-Tc candidates.	Regression (Tc prediction), Generative modeling	Structure-prediction, generation

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-12-13	yes	GeSS	<a href="https://neurips.cc/virtual/2024/MLSC/97816">https://neurips.cc/virtual/2024/MLSC/97816</a>	Sub-2024 MLSC-97816 metric Deep Learning	Benchmark suite evaluating geometric deep learning models under real-world distribution shifts	geometric deep learning, distribution shift, OOD robustness, scientific applications	GeSS provides 30 benchmark scenarios across particle physics, materials science, and biochemistry, evaluating 3 GDL backbones and 11 algorithms under covariate, concept, and conditional shifts, with varied OOD access	Classification, Regression	OOD performance, scientific scenarios

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type	
2024-12-13	yes	Vocal Call Locator	<a href="https://neurips.cc/virtualization/poster/97470">https://neurips.cc/virtualization/poster/97470</a>	NeurIPS 2024 Bioacoustics		Benchmarking sound-source localization of rodent vocalizations from multi-channel audio	source localization, bioacoustics, time-series, SSL	The first large-scale benchmark (767K sounds across 9 conditions) for localizing rodent vocal calls using synchronized audio and video in standard lab environments, enabling systematic evaluation of sound-source localization algorithms in bioacoustics	Sound source localization	Source localization accuracy in settings

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-12-13	yes	MassSpecGym	<a href="https://neurips.cc/virtual/2024/poster/97823">https://neurips.cc/virtual/2024/poster/97823</a>	Chen/2024/poster/97823 Molecular Discovery	Benchmark suite for discovery and identification of molecules via MS/MS	mass spectrometry, molecular structure, de novo generation, retrieval, dataset	MassSpecGym curates the largest public MS/MS dataset with three standardized tasks—de novo structure generation, molecule retrieval, and spectrum simulation—using challenging generalization splits to propel ML-driven molecule discovery	De novo generation, Retrieval, Simulation	Molecular generation and inference from spectra
2024-12-13	yes	Urban Data Layer	<a href="https://neurips.cc/virtual/2024/poster/97837">https://neurips.cc/virtual/2024/poster/97837</a>	Urban/2024/poster/97837 Data Engineering	Unified data pipeline for multi-modal urban science research	data pipeline, urban science, multi-modal, benchmark	UrbanDataLayer standardizes heterogeneous urban data formats and provides pipelines for tasks like air quality prediction and land-use classification, enabling the rapid creation of multi-modal urban benchmarks.	Prediction, Classification	Multi-modal inference, generation



Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-12-13	yes	Delta Squared-DFT	<a href="https://neurips.cc/virtualization/poster/97788">https://neurips.cc/virtualization/poster/97788</a>	Chemistry; Materials Science	Benchmarking machine-learning corrections to DFT using Delta Squared-trained models for reaction energies	density functional theory, Delta Squared-ML correction, reaction energetics, quantum chemistry	Introduces the Delta Squared-ML paradigm—using ML corrections to DFT to predict reaction energies with accuracy comparable to CCSD(T), while training on small CC datasets. Evaluated across 10 reaction datasets covering organic and organometallic transformations.	Regression	High-accuracy prediction, reaction
2024-12-13	yes	LLMs for Crop Science	<a href="https://neurips.cc/virtualization/poster/97570">https://neurips.cc/virtualization/poster/97570</a>	Agriculture; Science; NLP	Evaluating LLMs on crop trait QA and textual inference tasks with domain-specific prompts	crop science, prompt engineering, domain adaptation, question answering	Establishes a benchmark of 3,500 expert-annotated prompts and QA pairs covering crop traits, growth stages, and environmental interactions. Tests GPT-style LLMs on accuracy and domain reasoning using in-context, chain-of-thought, and retrieval-augmented prompts.	Question Answering, Inference	Scientific crop reasoning

Date	Expiration	Valid	Name	URL	Domain	Focus	Keywords	Description	Task Type
2024-12-13	yes	SPIQA LLM	<a href="https://neurips.cc/virtual/2024/poster/97575">https://neurips.cc/virtual/2024/poster/97575</a>	multimodal scientific QA; Computer Vision	LLMs on image-based scientific paper figure QA tasks (LLM Adapter performance)	multimodal QA, scientific figures, image+text, chain-of-thought prompting	A workshop version of SPIQA comparing 10 LLM adapter methods on the SPIQA benchmark with scientific diagram/questions. Highlights performance differences between chain-of-thought and end-to-end adapter models.	Multimodal QA	Visual scientific figure understanding

## References

- [1] J. Duarte, N. Tran, B. Hawks, *et al.*, *Fastml science benchmarks: Accelerating real-time scientific edge machine learning*, 2022. arXiv: 2207.07958 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2207.07958>.
- [2] J. Duarte, N. Tran, B. Hawks, *et al.*, *Fastml science benchmarks: Accelerating real-time scientific edge machine learning*, 2022. arXiv: 2207.07958 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2207.07958>.
- [3] D. Kafkes and J. S. John, *Boostr: A dataset for accelerator control systems*, 2021. arXiv: 2101.08359 [physics.acc-ph]. [Online]. Available: <https://arxiv.org/abs/2101.08359>.
- [4] J. Duarte, N. Tran, B. Hawks, *et al.*, *Fastml science benchmarks: Accelerating real-time scientific edge machine learning*, 2022. arXiv: 2207.07958 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2207.07958>.
- [5] P. Odagiu, Z. Que, J. Duarte, *et al.*, “Ultrafast jet classification at the hl-lhc,” *Machine Learning: Science and Technology*, vol. 5, no. 3, p. 035017, Jul. 2024, ISSN: 2632-2153. DOI: 10.1088/2632-2153/ad5f10. [Online]. Available: <http://dx.doi.org/10.1088/2632-2153/ad5f10>.
- [6] J. Kvapil, G. Borca-Tasciuc, H. Bossi, *et al.*, *Intelligent experiments through real-time ai: Fast data processing and autonomous detector control for sphenix and future eic detectors*, 2025. arXiv: 2501.04845 [physics.ins-det]. [Online]. Available: <https://arxiv.org/abs/2501.04845>.
- [7] J. Weitz, D. Demler, L. McDermott, N. Tran, and J. Duarte, *Neural architecture codesign for fast physics applications*, 2025. arXiv: 2501.05515 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2501.05515>.
- [8] B. Parpillon, C. Syal, J. Yoo, *et al.*, *Smart pixels: In-pixel ai for on-sensor data filtering*, 2024. arXiv: 2406.14860 [physics.ins-det]. [Online]. Available: <https://arxiv.org/abs/2406.14860>.
- [9] Z. Liu, H. Sharma, J.-S. Park, *et al.*, *Braggnet: Fast x-ray bragg peak analysis using deep learning*, 2021. arXiv: 2008.08198 [eess.IV]. [Online]. Available: <https://arxiv.org/abs/2008.08198>.
- [10] S. Qin, J. Agar, and N. Tran, “Extremely noisy 4d-tem strain mapping using cycle consistent spatial transforming autoencoders,” in *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023. [Online]. Available: <https://openreview.net/forum?id=7yt3N0o0W9>.
- [11] Y. Wei, R. F. Forelli, C. Hansen, *et al.*, “Low latency optical-based mode tracking with machine learning deployed on fpgas on a tokamak,” *Review of Scientific Instruments*, vol. 95, no. 7, Jul. 2024, ISSN: 1089-7623. DOI: 10.1063/5.0190354. [Online]. Available: <http://dx.doi.org/10.1063/5.0190354>.
- [12] W. Gao, F. Tang, L. Wang, *et al.*, *Aibench: An industry standard internet service ai benchmark suite*, 2019. arXiv: 1908.08998 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/1908.08998>.
- [13] W. Gao, J. Zhan, L. Wang, *et al.*, *Bigdatabench: A scalable and unified big data and ai benchmark suite*, 2018. arXiv: 1802.08254 [cs.DC]. [Online]. Available: <https://arxiv.org/abs/1802.08254>.
- [14] S. Farrell, M. Emani, J. Balma, *et al.*, *Mlperf hpc: A holistic benchmark suite for scientific machine learning on hpc systems*, 2021. arXiv: 2110.11466 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2110.11466>.
- [15] M. S. W. Group, *Mlcommons science working group benchmarks*, 2023. [Online]. Available: <https://github.com/mlcommons/science>.
- [16] T. Aarrestad, E. Govorkova, J. Ngadiuba, E. Puljak, M. Pierini, and K. A. Wozniak, *Unsupervised new physics detection at 40 mhz: Training dataset*, version v2, Jun. 2021. DOI: 10.5281/zenodo.5046428. [Online]. Available: <https://doi.org/10.5281/zenodo.5046428>.
- [17] A. Karargyris, M. J. Sheller, *et al.*, “Federated benchmarking of medical artificial intelligence with med-perf,” *Nature Machine Intelligence*, 2023. [Online]. Available: <https://www.nature.com/articles/s42256-023-00652-2>.
- [18] C. Krause, M. F. Giannelli, G. Kasieczka, *et al.*, *Calochallenge 2022: A community challenge for fast calorimeter simulation*, 2024. arXiv: 2410.21611 [physics.ins-det]. [Online]. Available: <https://arxiv.org/abs/2410.21611>.

- [19] P. W. Code, *Papers with code: Open machine learning benchmarks and leaderboards*, 2025. [Online]. Available: <https://paperswithcode.com>.
- [20] Z. Xu, S. Escalera, *et al.*, “Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform,” *Patterns*, vol. 3, no. 7, p. 100543, 2022. DOI: 10.1016/j.patter.2022.100543.
- [21] P. Luszczek *et al.*, “Sabath: Fair metadata technology for surrogate benchmarks,” University of Tennessee, Tech. Rep., 2021.
- [22] M. Takamoto, T. Praditia, R. Leiteritz, *et al.*, *Pdebench: An extensive benchmark for scientific machine learning*, 2024. arXiv: 2210.07182 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2210.07182>.
- [23] R. Ohana, M. McCabe, L. Meyer, *et al.*, “The well: A large-scale collection of diverse physics simulations for machine learning,” *NeurIPS*, vol. 37, pp. 44989–45037, 2024.
- [24] K. T. Chitty-Venkata, S. Raskar, B. Kale, *et al.*, *Llm-inference-bench: Inference benchmarking of large language models on ai accelerators*, 2024. arXiv: 2411.00136 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2411.00136>.
- [25] L. Zheng, L. Yin, Z. Xie, *et al.*, *Sglang: Efficient execution of structured language model programs*, 2024. arXiv: 2312.07104 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2312.07104>.
- [26] W. Kwon *et al.*, “Efficient memory management for large language model serving with pagedattention,” in *SOSP 2023*, 2023.
- [27] S. Mo, *Vllm performance dashboard*, 2024. [Online]. Available: <https://simon-mo-workspace.observablehq.cloud/vllm-dashboard-v0/>.
- [28] K. G. Olivares, C. Challú, *et al.*, *Neuralforecast: User friendly state-of-the-art neural forecasting models*, PyCon US, 2022. [Online]. Available: <https://github.com/Nixtla/neuralforecast>.
- [29] C. Challu, K. G. Olivares, *et al.*, “Nhits: Neural hierarchical interpolation for time series forecasting,” in *AAAI 2023*, 2023.
- [30] M. Jin, S. Wang, L. Ma, *et al.*, *Time-llm: Time series forecasting by reprogramming large language models*, 2024. arXiv: 2310.01728 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2310.01728>.
- [31] A. Garza, C. Challu, and M. Mergenthaler-Canseco, *Timegpt-1*, 2024. arXiv: 2310.03589 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2310.03589>.
- [32] E. G. Campolongo, Y.-T. Chou, E. Govorkova, *et al.*, *Building machine learning challenges for anomaly detection in science*, 2025. arXiv: 2503.02112 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2503.02112>.
- [33] E. G. Campolongo, Y.-T. Chou, E. Govorkova, *et al.*, *Building machine learning challenges for anomaly detection in science*, 2025. arXiv: 2503.02112 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2503.02112>.
- [34] E. G. Campolongo, Y.-T. Chou, E. Govorkova, *et al.*, *Building machine learning challenges for anomaly detection in science*, 2025. arXiv: 2503.02112 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2503.02112>.
- [35] G. D. Guglielmo, B. Du, J. Campos, *et al.*, *End-to-end workflow for machine learning-based qubit readout with qick and hls4ml*, 2025. arXiv: 2501.14663 [quant-ph]. [Online]. Available: <https://arxiv.org/abs/2501.14663>.
- [36] D. Rein, B. L. Hou, A. C. Stickland, *et al.*, *Gpqa: A graduate-level google-proof q&a benchmark*, 2023. arXiv: 2311.12022 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2311.12022>.
- [37] K. X. Nguyen, F. Qiao, A. Trembanis, and X. Peng, *Seafloorai: A large-scale vision-language dataset for seafloor geological survey*, 2024. arXiv: 2411.00172 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2411.00172>.
- [38] Z. Zuo *et al.*, *Supercon3d: Learning superconductivity from ordered and disordered material structures*, NeurIPS Poster, 2024.

- [39] D. Zou, S. Liu, *et al.*, *Gess: Benchmarking geometric deep learning under scientific applications with distribution shifts*, NeurIPS Poster, 2024.
- [40] R. Peterson, A. Tanelus, *et al.*, *Vocal call locator benchmark for localizing rodent vocalizations*, NeurIPS Poster, 2024. [Online]. Available: <https://neurips.cc/virtual/2024/poster/97470>.
- [41] R. Bushuiev, A. Bushuiev, *et al.*, *Massspecgym: A benchmark for the discovery and identification of molecules*, NeurIPS Spotlight Poster, 2024. [Online]. Available: <https://neurips.cc/virtual/2024/poster/97823>.
- [42] Y. Wang, T. Wang, *et al.*, *Urbandatalayer: A unified data pipeline for urban science*, NeurIPS Poster, 2024. [Online]. Available: <https://neurips.cc/virtual/2024/poster/97837>.
- [43] W. Liu, R. Chen, *et al.*, *Delta squared-dft: Machine-learning corrected density functional theory for reaction energetics*, NeurIPS Poster, 2024. [Online]. Available: <https://neurips.cc/virtual/2024/poster/97788>.
- [44] D. Patel, L. Zhao, *et al.*, *Large language models for crop science: Benchmarking domain reasoning and qa*, NeurIPS Poster, 2024. [Online]. Available: <https://neurips.cc/virtual/2024/poster/97570>.
- [45] X. Zhong, Y. Gao, *et al.*, *Spiqa-llm: Evaluating llm adapters on scientific figure qa*, NeurIPS Poster, 2024. [Online]. Available: <https://neurips.cc/virtual/2024/poster/97575>.