

# MLCommons Science Working Group AI Benchmarks Collection

Gregor von Laszewski and Ben Hawks and Marco Colombo and  
Reece Shiraishi and Anjay Krishnan and  
Nhan Tran and Geoffrey C. Fox

November 4, 2025

## Abstract

This document provides an overview of various benchmarks, including their descriptions, URLs, domains, focus areas, keywords, task types, AI capabilities measured, metrics, models, and notes. Each benchmark

## Citation

```
@misc{www-las-mlcommons-benchmark-coolection,
author = {
    Gregor von Laszewski and
    Ben Hawks and
    Marco Colombo and
    Reece Shiraishi and
    Anjay Krishnan and
    Nhan Tran and
    Geoffrey C. Fox},
title = {MLCommons Science Working Group AI Benchmarks Collection},
url = {https://mlcommons-science.github.io/benchmark/benchmarks.pdf},
note = "Online Collection: \url={https://mlcommons-science.github.io/benchmark/}",
month = jun,
year = 2025,
howpublished = "GitHub"
}
```

# Contents

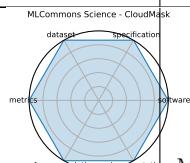
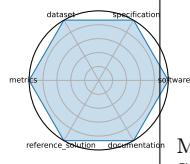
|  |           |
|--|-----------|
| <b>1 Benchmark Overview Table</b>  | <b>4</b>  |
| <b>2 Radar Chart Table</b>   | <b>38</b> |
| <b>3 Benchmark Details</b>   | <b>41</b> |
| 3.1 ClimateLearn - Weather Forcasting . . . . .  | 41        |
| 3.2 ClimateLearn - Downscaling . . . . .   | 42        |
| 3.3 ClimateLearn - Climate Projection . . . . .  | 43        |
| 3.4 MLCommons Science - CloudMask . . . . .  | 44        |
| 3.5 MLCommons Science - Earthquake . . . . .   | 45        |
| 3.6 MLCommons Science - Candle UNO . . . . .   | 46        |
| 3.7 MLCommons Science - STEMDL . . . . .   | 47        |
| 3.8 ARC-Challenge (Advanced Reasoning Challenge) . . . . .                             | 48        |
| 3.9 MOLGEN . . . . .   | 49        |
| 3.10 Open Graph Benchmark (OGB) - Biology . . . . .                                    | 50        |
| 3.11 LLMs for Crop Science . . . . .   | 51        |
| 3.12 SciCode . . . . .   | 52        |
| 3.13 CaloChallenge 2022 . . . . .  | 53        |
| 3.14 PDEBench . . . . .  | 54        |
| 3.15 Urban Data Layer (UDL) - PM2.5 Concentration Prediction . . . . .                 | 55        |
| 3.16 Urban Data Layer (UDL) - Built-up Area Classification . . . . .                   | 56        |
| 3.17 Urban Data Layer (UDL) - Administrative Boundaries Identification . . . . .       | 57        |
| 3.18 Urban Data Layer (UDL) - El Nino Anomaly Detection . . . . .                      | 58        |
| 3.19 SPIQA (LLM) . . . . .   | 59        |
| 3.20 MLCommons Medical AI - Pancreas Segmentation (DFCI) . . . . .                     | 60        |
| 3.21 MLCommons Medical AI - Brain Tumor Segmentation (BraTS) . . . . .                 | 61        |
| 3.22 MLCommons Medical AI - Surgical Workflow Phase Recognition (SurgMLCube) . . . . . | 62        |
| 3.23 SeafloorAI . . . . .  | 63        |
| 3.24 SeafloorGenAI . . . . .   | 64        |
| 3.25 GeSS - Track Pileup . . . . .   | 65        |
| 3.26 GeSS - Track Signal . . . . .   | 66        |
| 3.27 GeSS - DrugOOD . . . . .  | 67        |
| 3.28 GeSS - QMOF . . . . .   | 68        |
| 3.29 OCP (Open Catalyst Project) . . . . .   | 69        |
| 3.30 Jet Classification . . . . .  | 70        |
| 3.31 Irregular Sensor Data Compression . . . . .                                       | 71        |
| 3.32 MLPerf HPC - Cosmoflow . . . . .  | 72        |
| 3.33 MLPerf HPC - DeepCAM . . . . .  | 73        |
| 3.34 MLPerf HPC - Open Catalyst Project DimeNet++ . . . . .                            | 74        |
| 3.35 MLPerf HPC - OpenFold . . . . .   | 75        |
| 3.36 HDR ML Anomaly Challenge - Gravitational Waves . . . . .                          | 76        |
| 3.37 SuperCon3D - Property Prediction . . . . .  | 77        |
| 3.38 SuperCon3D - Inverse Crystal Structure Generation . . . . .                       | 78        |
| 3.39 BaisBench (Biological AI Scientist Benchmark) - Question Answering . . . . .      | 79        |
| 3.40 BaisBench (Biological AI Scientist Benchmark) - Cell Type Annotation . . . . .    | 80        |
| 3.41 The Well . . . . .  | 81        |
| 3.42 MMLU (Massive Multitask Language Understanding) . . . . .                         | 82        |
| 3.43 SatImgNet . . . . .   | 83        |
| 3.44 GPQA Diamond . . . . .  | 84        |
| 3.45 PRM800K . . . . .   | 85        |

|   |     |
|---|-----|
| 3.46 FEA Bench (Finite Element Analysis Benchmark): Evaluating Language Models on Multi-physics Reasoning Ability . . . . . | 86  |
| 3.47 Neural Architecture Codesign for Fast Physics Applications . . . . .   | 87  |
| 3.48 Delta Squared-DFT . . . . .  | 88  |
| 3.49 HDR ML Anomaly Challenge - Sea Level Rise . . . . .  | 89  |
| 3.50 Vocal Call Locator (VCL) . . . . .   | 90  |
| 3.51 MassSpecGym - De novo molecule generation . . . . .  | 91  |
| 3.52 MassSpecGym - Molecule Retrieval . . . . .   | 92  |
| 3.53 MassSpecGym - Spectrum Simulation . . . . .  | 93  |
| 3.54 SPIQA (Scientific Paper Image Question Answering) . . . . .  | 94  |
| 3.55 GPQA: A Graduate-Level Google-Proof Question and Answer Benchmark . . . . .  | 95  |
| 3.56 MedQA . . . . .  | 96  |
| 3.57 Single Qubit Readout on QICK System . . . . .  | 97  |
| 3.58 CFD Bench (Fluid Dynamics) . . . . .   | 98  |
| 3.59 CURIE (Scientific Long-Context Understanding, Reasoning and Information Extraction) . . . . .                          | 99  |
| 3.60 Smart Pixels for LHC . . . . .   | 100 |
| 3.61 LHC New Physics Dataset . . . . .  | 101 |
| 3.62 Quantum Computing Benchmarks (QML) . . . . .   | 102 |
| 3.63 Ultrafast jet classification at the HL-LHC . . . . .   | 103 |
| 3.64 HEDM (BraggNN) . . . . .   | 104 |
| 3.65 4D-STEM . . . . .  | 105 |
| 3.66 Beam Control . . . . .   | 106 |
| 3.67 Intelligent experiments through real-time AI . . . . .   | 107 |
| 3.68 HDR ML Anomaly Challenge - Butterfly . . . . .   | 108 |
| 3.69 DUNE . . . . .   | 109 |
| 3.70 FrontierMath . . . . .   | 110 |
| 3.71 AIME (American Invitational Mathematics Examination) . . . . .   | 111 |
| 3.72 Quench detection . . . . .   | 112 |
| 3.73 Materials Project . . . . .  | 113 |
| 3.74 In-Situ High-Speed Computer Vision . . . . .   | 114 |

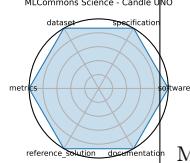
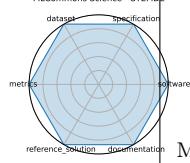
## 1 Benchmark Overview Table

| Ratings | Name                              | Domain                  | Focus                               | Keywords                                    | Task Types  | AI/ML Motif                     | AI Capability                        | Metrics                   | Models                         | Average Rating | Citation |
|---------|-----------------------------------|-------------------------|-------------------------------------|---|-------------|---------------------------------|--------------------------------------|---------------------------|--------------------------------|----------------|----------|
|         | ClimateLearn - Weather Forcasting | Climate & Earth Science | ML for weather and climate modeling | medium-range forecasting, ERA5, data-driven | Forecasting | Sequence Prediction/Forecasting | Global weather prediction (3-5 days) | RMSE, Anomaly correlation | CNN baselines, ResNet variants | <b>5.00</b>    | [1]      |
|         | ClimateLearn - Downscaling        | Climate & Earth Science | ML for weather and climate modeling | medium-range forecasting, ERA5, data-driven | Forecasting | Regression                      | Global weather prediction (3-5 days) | RMSE, Anomaly correlation | CNN baselines, ResNet variants | <b>5.00</b>    | [1]      |
|         | ClimateLearn - Climate Projection | Climate & Earth Science | ML for weather and climate modeling | medium-range forecasting, ERA5, data-driven | Forecasting | Regression                      | Global weather prediction (3-5 days) | RMSE, Anomaly correlation | CNN baselines, ResNet variants | <b>5.00</b>    | [1]      |

Continued on next page

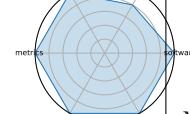
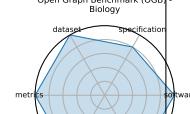
| Ratings   | Name                           | Domain                  | Focus  | Keywords                              | Task Types  | AI/ML Motif                     | AI Capability   | Metrics                              | Models                | Average Rating | Citation |
|---|--------------------------------|-------------------------|--|---------------------------------------|---|---------------------------------|---|--------------------------------------|-----------------------|----------------|----------|
|  | MLCommons Science - CloudMask  | Climate & Earth Science | AI benchmarks for scientific applications including time-series, imaging, and simulation | science AI, benchmark, MLCommons, HPC | Time-series analysis, Image classification, Simulation surrogate modeling | Classification                  | Inference accuracy, simulation speed-up, generalization | MAE, Accuracy, Speedup vs simulation | CNN, GNN, Transformer | <b>5.00</b>    | [2]      |
|  | MLCommons Science - Earthquake | Climate & Earth Science | AI benchmarks for scientific applications including time-series, imaging, and simulation | science AI, benchmark, MLCommons, HPC | Time-series analysis, Image classification, Simulation surrogate modeling | Sequence Prediction/Forecasting | Inference accuracy, simulation speed-up, generalization | MAE, Accuracy, Speedup vs simulation | CNN, GNN, Transformer | <b>5.00</b>    | [2]      |

Continued on next page

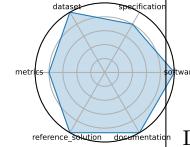
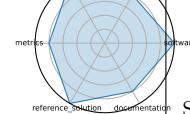
| Ratings   | Name                           | Domain             | Focus  | Keywords                              | Task Types  | AI/ML Motif    | AI Capability   | Metrics                              | Models                | Average Rating | Citation |
|---|--------------------------------|--------------------|--|---------------------------------------|---|----------------|---|--------------------------------------|-----------------------|----------------|----------|
|  | MLCommons Science - Candle UNO | Biology & Medicine | AI benchmarks for scientific applications including time-series, imaging, and simulation | science AI, benchmark, MLCommons, HPC | Time-series analysis, Image classification, Simulation surrogate modeling | Classification | Inference accuracy, simulation speed-up, generalization | MAE, Accuracy, Speedup vs simulation | CNN, GNN, Transformer | <b>5.00</b>    | [2]      |
|  | MLCommons Science - STEM DL    | Materials Science  | AI benchmarks for scientific applications including time-series, imaging, and simulation | science AI, benchmark, MLCommons, HPC | Time-series analysis, Image classification, Simulation surrogate modeling | Classification | Inference accuracy, simulation speed-up, generalization | MAE, Accuracy, Speedup vs simulation | CNN, GNN, Transformer | <b>5.00</b>    | [2]      |

Continued on next page

∞

| Ratings | Name   | Domain                     | Focus  | Keywords   | Task Types  | AI/ML Motif                     | AI Capability  | Metrics   | Models               | Average Rating | Citation |
|---------|--|----------------------------|--|--|---|---------------------------------|--|---|----------------------|----------------|----------|
|         |  ARC-Challenge (Advanced Reasoning Challenge) | Computational Science & AI | Glade-school science with reasoning emphasis | grade-school, science QA, challenge set, reasoning     | Multiple choice   | Reasoning & Generalization      | Commonsense and scientific reasoning                   | Accuracy  | GPT-4, Claude        | <b>4.83</b>    | [3]      |
|         |  MOLGEN                                       | Chemistry                  | Molecular generation and optimization        | SELFIES, GAN, property optimization                    | Distribution learning, Goal-oriented generation                               | Generative                      | Generation of valid and optimized molecular structures | Validity%, Novelty%, QED, Docking score, penalized logP | MolGen               | <b>4.83</b>    | [4]      |
|         |  Open Graph Benchmark (OGB) - Biology        | Biology & Medicine         | Biological graph property prediction         | node prediction, link prediction, graph classification | Node property prediction, Link property prediction, Graph property prediction | Sequence Prediction/Forecasting | Scalability and generalization in graph ML for biology | Accuracy, ROC-AUC                                       | GCN, Graph-SAGE, GAT | <b>4.83</b>    | [5]      |

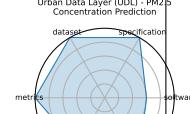
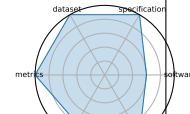
Continued on next page

| Ratings   | Name                  | Domain                     | Focus   | Keywords  | Task Types                    | AI/ML Motif                | AI Capability                           | Metrics            | Models   | Average Rating | Citation |
|---|-----------------------|----------------------------|---|---|-------------------------------|----------------------------|---|--------------------|--|----------------|----------|
|  | LLMs for Crop Science | Climate & Earth Science    | Evaluating LLMs on crop trait QA and textual inference tasks with domain-specific prompts | crop science, prompt engineering, domain adaptation, question answering | Question Answering, Inference | Reasoning & Generalization | Scientific knowledge, crop reasoning    | Accuracy, F1 score | GPT-3.5, GPT-4, Claude-3-opus, Qwen-max, LLama3-8B, InternLM2-7B, Qwen1.5-7B | <b>4.67</b>    | [6]      |
|  | SciCode               | Computational Science & AI | Scientific code generation and problem solving  | code synthesis, scientific computing, programming benchmark             | Coding                        | Generative                 | Program synthesis, scientific computing | Solve rate (%)     | Claude3.5-Sonnet   | <b>4.50</b>    | [7]      |

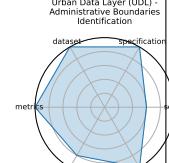
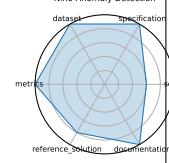
Continued on next page

| Ratings | Name               | Domain   | Focus  | Keywords  | Task Types          | AI/ML Motif | AI Capability                                  | Metrics  | Models  | Average Rating | Citation |
|---------|--------------------|--|--|---|---------------------|-------------|--|--|---|----------------|----------|
|         | CaloChallenge 2022 | High Energy Physics  | Fast generative-model-based calorimeter shower simulation evaluation | calorimeter simulation, generative models, surrogate modeling, LHC, fast simulation | Surrogate modeling  | Generative  | Simulation fidelity, speed, efficiency         | Histogram similarity, Classifier AUC, Generation latency | VAE variants, GAN variants, Normalizing flows, Diffusion models | <b>4.50</b>    | [8]      |
|         | PDEBench           | Computational Science & AI, Climate & Earth Science, Mathematics | Benchmark suite for ML-based surrogates solving time-dependent PDEs  | PDEs, CFD, scientific ML, surrogate modeling, NeurIPS                               | Supervised Learning | Regression  | Time-dependent PDE modeling; physical accuracy | RMSE, boundary RMSE, Fourier RMSE                        | FNO, U-Net, PINN, Gradient-Based inverse methods                | <b>4.50</b>    | [9]      |

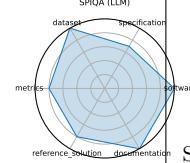
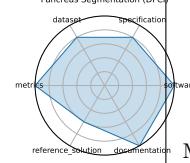
Continued on next page

| Ratings   | Name  | Domain                  | Focus  | Keywords   | Task Types                 | AI/ML Motif    | AI Capability                                | Metrics                        | Models                                       | Average Rating | Citation |
|---|---|-------------------------|--|--|----------------------------|----------------|--|--------------------------------|--|----------------|----------|
|  | Urban Data Layer (UDL) - PM2.5 Concentration Prediction | Climate & Earth Science | Unified data pipeline for multi-modal urban science research | data pipeline, urban science, multi-modal, benchmark | Prediction, Classification | Regression     | Multi-modal urban inference, standardization | Task-specific accuracy or RMSE | Baseline regression/classification pipelines | <b>4.50</b>    | [10]     |
|  | Urban Data Layer (UDL) - Built-up Area Classification   | Climate & Earth Science | Unified data pipeline for multi-modal urban science research | data pipeline, urban science, multi-modal, benchmark | Prediction, Classification | Classification | Multi-modal urban inference, standardization | Task-specific accuracy or RMSE | Baseline regression/classification pipelines | <b>4.50</b>    | [10]     |

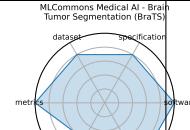
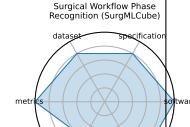
Continued on next page

| Ratings   | Name  | Domain                  | Focus  | Keywords   | Task Types                 | AI/ML Motif       | AI Capability                                | Metrics                        | Models                                       | Average Rating | Citation |
|---|---|-------------------------|--|--|----------------------------|-------------------|--|--------------------------------|--|----------------|----------|
|  | Urban Data Layer (UDL) - Administrative Boundaries Identification | Climate & Earth Science | Unified data pipeline for multi-modal urban science research | data pipeline, urban science, multi-modal, benchmark | Prediction, Classification | Classification    | Multi-modal urban inference, standardization | Task-specific accuracy or RMSE | Baseline regression/classification pipelines | <b>4.50</b>    | [10]     |
|  | Urban Data Layer (UDL) - El Nino Anomaly Detection                | Climate & Earth Science | Unified data pipeline for multi-modal urban science research | data pipeline, urban science, multi-modal, benchmark | Prediction, Classification | Anomaly Detection | Multi-modal urban inference, standardization | Task-specific accuracy or RMSE | Baseline regression/classification pipelines | <b>4.50</b>    | [10]     |

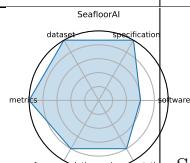
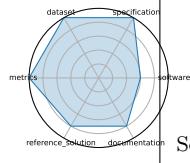
Continued on next page

| Ratings   | Name  | Domain                     | Focus  | Keywords  | Task Types                             | AI/ML Motif          | AI Capability   | Metrics                             | Models                                     | Average Rating | Citation |
|---|---|----------------------------|--|---|--|----------------------|---|-------------------------------------|--|----------------|----------|
|  | SPIQA (LLM)   | Computational Science & AI | Evaluating LLMs on image-based scientific paper figure QA tasks (LLM Adapter performance)          | multimodal QA, scientific figures, image+text, chain-of-thought prompting             | Multimodal QA                          | Multimodal Reasoning | Visual reasoning, scientific figure understanding                 | Accuracy, F1 score                  | LLaVA, MiniGPT-4, Owl-LLM adapter variants | 4.42           | [11]     |
|  | MLCommons Medical AI - Pancreas Segmentation (DFCI) | Biology & Medicine         | Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data | medical AI, federated evaluation, privacy-preserving, fairness, healthcare benchmarks | Federated evaluation, Model validation | Classification       | Clinical accuracy, fairness, generalizability, privacy compliance | ROC AUC, Accuracy, Fairness metrics | MedPerf-validated CNNs, GaN-DLF workflows  | 4.33           | [12]     |

Continued on next page

| Ratings   | Name   | Domain             | Focus  | Keywords  | Task Types                             | AI/ML Motif    | AI Capability   | Metrics                             | Models                                    | Average Rating | Citation |
|---|--|--------------------|--|---|--|----------------|---|-------------------------------------|---|----------------|----------|
|  | MLCommons Medical AI - Brain Tumor Segmentation (BraTS)                  | Biology & Medicine | Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data | medical AI, federated evaluation, privacy-preserving, fairness, healthcare benchmarks | Federated evaluation, Model validation | Classification | Clinical accuracy, fairness, generalizability, privacy compliance | ROC AUC, Accuracy, Fairness metrics | MedPerf-validated CNNs, GaN-DLF workflows | 4.33           | [12]     |
|  | MLCommons Medical AI - Surgical Workflow Phase Recognition (SurgML-Cube) | Biology & Medicine | Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data | medical AI, federated evaluation, privacy-preserving, fairness, healthcare benchmarks | Federated evaluation, Model validation | Classification | Clinical accuracy, fairness, generalizability, privacy compliance | ROC AUC, Accuracy, Fairness metrics | MedPerf-validated CNNs, GaN-DLF workflows | 4.33           | [12]     |

Continued on next page

| Ratings   | Name          | Domain                  | Focus  | Keywords   | Task Types                             | AI/ML Motif                | AI Capability                                  | Metrics                                  | Models                                  | Average Citation Rating |
|---|---------------|-------------------------|--|--|--|----------------------------|--|--|---|-------------------------|
|  | SeafloorAI    | Climate & Earth Science | Large-scale vision-language dataset for seafloor mapping and geological classification | sonar imagery, vision-language, seafloor mapping, segmentation, QA | Image segmentation, Vision-language QA | Classification             | Geospatial understanding, multimodal reasoning | Segmentation pixel accuracy, QA accuracy | SegFormer, ViLT-style multimodal models | 4.33 [13]               |
|  | SeafloorGenAI | Climate & Earth Science | Large-scale vision-language dataset for seafloor mapping and geological classification | sonar imagery, vision-language, seafloor mapping, segmentation, QA | Image segmentation, Vision-language QA | Reasoning & Generalization | Geospatial understanding, multimodal reasoning | Segmentation pixel accuracy, QA accuracy | SegFormer, ViLT-style multimodal models | 4.33 [13]               |

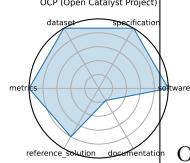
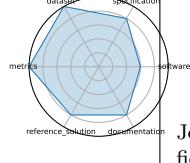
Continued on next page

| Ratings | Name              | Domain | Focus               | Keywords   | Task Types   | AI/ML Motif    | AI Capability  | Metrics                                | Models                               | Average Rating      | Citation  |
|---------|-------------------|--------|---------------------|--|--|----------------|----------------|--|--------------------------------------|---------------------|-----------|
|         | GeSS Track Pileup | -      | High Energy Physics | Benchmark suite evaluating geometric deep learning models under real-world distribution shifts | geometric deep learning, distribution shift, OOD robustness, scientific applications | Classification | Classification | OOD performance in scientific settings | Accuracy, RMSE, OOD robustness delta | GCN, EGN, DimeNet++ | 4.33 [14] |
|         | GeSS Track Signal | -      | High Energy Physics | Benchmark suite evaluating geometric deep learning models under real-world distribution shifts | geometric deep learning, distribution shift, OOD robustness, scientific applications | Classification | Classification | OOD performance in scientific settings | Accuracy, RMSE, OOD robustness delta | GCN, EGN, DimeNet++ | 4.33 [14] |

Continued on next page

| Ratings | Name           | Domain             | Focus  | Keywords   | Task Types                 | AI/ML Motif    | AI Capability                          | Metrics                              | Models              | Average Rating | Citation |
|---------|----------------|--------------------|--|--|----------------------------|----------------|--|--------------------------------------|---------------------|----------------|----------|
|         | GeSS - DrugOOD | Biology & Medicine | Benchmark suite evaluating geometric deep learning models under real-world distribution shifts | geometric deep learning, distribution shift, OOD robustness, scientific applications | Classification             | Classification | OOD performance in scientific settings | Accuracy, RMSE, OOD robustness delta | GCN, EGN, DimeNet++ | 4.33           | [14]     |
|         | GeSS - QM0F    | Materials Science  | Benchmark suite evaluating geometric deep learning models under real-world distribution shifts | geometric deep learning, distribution shift, OOD robustness, scientific applications | Classification, Regression | Regression     | OOD performance in scientific settings | Accuracy, RMSE, OOD robustness delta | GCN, EGN, DimeNet++ | 4.33           | [14]     |

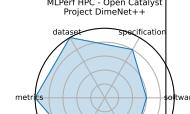
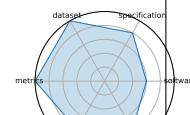
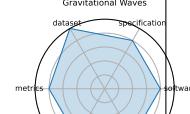
Continued on next page

| Ratings   | Name                        | Domain                       | Focus  | Keywords  | Task Types                          | AI/ML Motif    | AI Capability                                      | Metrics                   | Models                              | Average Rating | Citation  |
|---|-----------------------------|------------------------------|--|---|-------------------------------------|----------------|--|---------------------------|-------------------------------------|----------------|-----------|
|  | OCP (Open Catalyst Project) | Chemistry, Materials Science | Catalyst adsorption energy prediction                                      | DFT relaxations, adsorption energy, graph neural networks | Energy prediction, Force prediction | Regression     | Prediction of adsorption energies and forces       | MAE (energy), MAE (force) | CGCNN, SchNet, DimeNet++, GemNet-OC | 4.17           | [15]–[18] |
|  | Jet Classification          | High Energy Physics          | Real-time classification of particle jets using HL-LHC simulation features | classification, real-time ML, jet tagging, QKeras         | Classification                      | Classification | Real-time inference, model compression performance | Accuracy, AUC             | Keras DNN, QKeras quantized DNN     | 4.17           | [19]      |

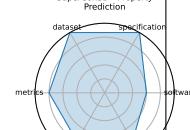
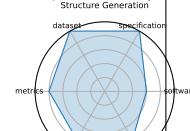
Continued on next page

| Ratings | Name                              | Domain                  | Focus   | Keywords  | Task Types          | AI/ML Motif    | AI Capability  | Metrics                                  | Models      | Average Rating | Citation |
|---------|-----------------------------------|-------------------------|---|---|---------------------|----------------|--|--|-------------|----------------|----------|
|         | Irregular Sensor Data Compression | High Energy Physics     | Real-time compression of sparse sensor data with autoencoders | compression, autoencoder, sparse data, irregular sampling | Compression         | Generative     | Reconstruction quality, compression efficiency           | MSE, Compression ratio                   | Autoencoder | 4.17           | [20]     |
|         | MLPerf HPC - Cosmoflow            | High Energy Physics     | Scientific ML training and inference on HPC systems           | HPC, training, inference, scientific ML                   | Training, Inference | Regression     | Scaling efficiency, training time, model accuracy on HPC | Training time, Accuracy, GPU utilization | CosmoFlow   | 4.17           | [21]     |
|         | MLPerf HPC - DeepCAM              | Climate & Earth Science | Scientific ML training and inference on HPC systems           | HPC, training, inference, scientific ML                   | Training, Inference | Classification | Scaling efficiency, training time, model accuracy on HPC | Training time, Accuracy, GPU utilization | DeepCAM     | 4.17           | [21]     |

Continued on next page

| Ratings  | Name   | Domain              | Focus   | Keywords  | Task Types          | AI/ML Motif                     | AI Capability  | Metrics                                  | Models                         | Average Rating | Citation |
|--|--|---------------------|---|---|---------------------|---------------------------------|--|--|--------------------------------|----------------|----------|
|   | MLPerf HPC - Open Catalyst Project DimeNet++   | Chemistry           | Scientific ML training and inference on HPC systems                     | HPC, training, inference, scientific ML                           | Training, Inference | Regression                      | Scaling efficiency, training time, model accuracy on HPC | Training time, Accuracy, GPU utilization | DeepCAM                        | 4.17           | [21]     |
|   | MLPerf HPC - OpenFold                          | Biology & Medicine  | Scientific ML training and inference on HPC systems                     | HPC, training, inference, scientific ML                           | Training, Inference | Sequence Prediction/Forecasting | Scaling efficiency, training time, model accuracy on HPC | Training time, Accuracy, GPU utilization | DeepCAM                        | 4.17           | [21]     |
|  | HDR ML Anomaly Challenge - Gravitational Waves | High Energy Physics | Detecting anomalous gravitational wave signals from LIGO/Virgo datasets | anomaly detection, gravitational waves, astrophysics, time-series | Anomaly Detection   | Anomaly Detection               | Novel event detection in physical signals                | ROC-AUC, Precision/Recall                | Deep latent CNNs, Autoencoders | 4.17           | [22]     |

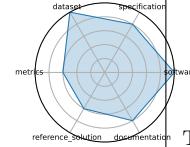
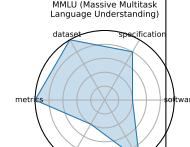
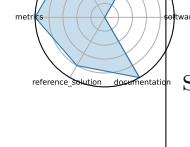
Continued on next page

| Ratings   | Name  | Domain            | Focus  | Keywords   | Task Types                                      | AI/ML Motif | AI Capability  | Metrics                                    | Models             | Average Rating | Citation |
|---|---|-------------------|--|--|---|-------------|--|--|--------------------|----------------|----------|
|  | SuperCon3D - Property Prediction                  | Materials Science | Dataset and models for predicting and generating high-Tc superconductors using 3D crystal structures | superconducting crystal structures, equivariant GNN, generative models | Regression (Tc prediction), Generative modeling | Regression  | Structure-to-property prediction, structure generation | MAE (Tc), Validity of generated structures | SODNet, DiffCSP-SC | 4.17           | [23]     |
|  | SuperCon3D - Inverse Crystal Structure Generation | Materials Science | Dataset and models for predicting and generating high-Tc superconductors using 3D crystal structures | superconducting crystal structures, equivariant GNN, generative models | Regression (Tc prediction), Generative modeling | Generative  | Structure-to-property prediction, structure generation | MAE (Tc), Validity of generated structures | SODNet, DiffCSP-SC | 4.17           | [23]     |

Continued on next page

| Ratings | Name   | Domain             | Focus                          | Keywords  | Task Types                            | AI/ML Motif                | AI Capability                               | Metrics                          | Models                        | Average Rating | Citation |
|---------|--|--------------------|--------------------------------|---|---------------------------------------|----------------------------|---|----------------------------------|-------------------------------|----------------|----------|
|         | BaisBench (Biological AI Scientist Benchmark) - Question Answering   | Biology & Medicine | Omics-driven AI research tasks | single-cell annotation, biological QA, autonomous discovery | Cell type annotation, Multiple choice | Reasoning & Generalization | Autonomous biological research capabilities | Annotation accuracy, QA accuracy | LLM-based AI scientist agents | 4.00           | [24]     |
|         | BaisBench (Biological AI Scientist Benchmark) - Cell Type Annotation | Biology & Medicine | Omics-driven AI research tasks | single-cell annotation, biological QA, autonomous discovery | Cell type annotation, Multiple choice | Classification             | Autonomous biological research capabilities | Annotation accuracy, QA accuracy | LLM-based AI scientist agents | 4.00           | [24]     |

Continued on next page

| Ratings   | Name  | Domain  | Focus  | Keywords   | Task Types           | AI/ML Motif                     | AI Capability                                   | Metrics                      | Models                                  | Average Rating | Citation |
|---|---|---|--|--|----------------------|---------------------------------|---|------------------------------|---|----------------|----------|
|    | The Well  | Biology & Medicine, Computational Science & AI, High Energy Physics | Foundation model + surrogate dataset spanning 16 physical simulation domains | surrogate modeling, foundation model, physics simulations, spatiotemporal dynamics | Supervised Learning  | Sequence Prediction/Forecasting | Surrogate modeling, physics-based prediction    | Dataset size, Domain breadth | FNO baselines, U-Net baselines          | 4.00           | [25]     |
|    | MMLU (Massive Multitask Language Understanding) | Computational Science & AI  | Academic knowledge and reasoning across 57 subjects                          | multitask, multiple-choice, zero-shot, few-shot, knowledge probing                 | Multiple choice      | Reasoning & Generalization      | General reasoning, subject-matter understanding | Accuracy                     | GPT-4o, Gemini 1.5 Pro, 01, DeepSeek-R1 | 3.83           | [26]     |
|  | SatImgNet                                       | Climate & Earth Science   | Satellite imagery classification   | land-use, zero-shot, multi-task  | Image classification | Multimodal Reasoning            | Zero-shot land-use classification               | Accuracy                     | CLIP, BLIP, ALBEF                       | 3.83           | [27]     |

Continued on next page

| Ratings | Name   | Domain   | Focus                                   | Keywords   | Task Types                         | AI/ML Motif                | AI Capability                                | Metrics                | Models          | Average Rating | Citation |
|---------|--|--|---|--|------------------------------------|----------------------------|--|------------------------|-----------------|----------------|----------|
|         | GPQA Diamond   | Biology & Medicine, Chemistry, High Energy Physics | Graduate-level scientific reasoning     | Google-proof, graduate-level, science QA, chemistry, physics | Multiple choice, Multi-step QA     | Reasoning & Generalization | Scientific reasoning, deep knowledge         | Accuracy               | o1, DeepSeek-R1 | 3.83           | [28]     |
|         | PRM800K  | Mathematics  | Math reasoning generalization           | calculus, algebra, number theory, geometry                   | Problem solving                    | Reasoning & Generalization | Math reasoning and generalization            | Accuracy               | GPT-4           | 3.83           | [29]     |
|         | FEABench (Finite Element Analysis Benchmark): Evaluating Language Models on Multiphysics Reasoning Ability | Mathematics  | FEA simulation accuracy and performance | finite element, simulation, PDE                              | Simulation, Performance evaluation | Reasoning & Generalization | Numerical simulation accuracy and efficiency | Solve time, Error norm | FEniCS, deal.II | 3.83           | [30]     |

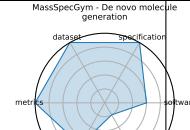
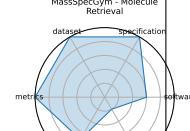
Continued on next page

| Ratings | Name   | Domain                       | Focus   | Keywords   | Task Types                   | AI/ML Motif    | AI Capability  | Metrics   | Models  | Average Rating | Citation |
|---------|--|------------------------------|---|--|------------------------------|----------------|--|---|---|----------------|----------|
|         | Neural Architecture Codesign for Fast Physics Applications | High Energy Physics          | Automated neural architecture search and hardware-efficient model codesign for fast physics applications  | neural architecture search, FPGA deployment, quantization, pruning, hls4ml                     | Classification, Peak finding | Classification | Hardware-aware model optimization; low-latency inference | Accuracy, Latency, Resource utilization           | NAC-based BraggNN, NAC-optimized Deep Sets (jet)              | 3.83           | [31]     |
|         | Delta Squared-DFT  | 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 | Regression                   | Regression     | High-accuracy energy prediction, DFT correction          | Mean Absolute Error (eV), Energy ranking accuracy | Delta Squared-ML correction networks, Kernel ridge regression | 3.83           | [32]     |

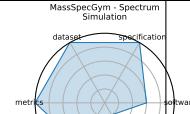
Continued on next page

| Ratings   | Name                                      | Domain                  | Focus  | Keywords  | Task Types                | AI/ML Motif       | AI Capability  | Metrics                                   | Models                   | Average Rating | Citation |
|---|---|-------------------------|--|---|---------------------------|-------------------|--|---|--------------------------|----------------|----------|
|  | HDR ML Anomaly Challenge - Sea Level Rise | Climate & Earth Science | Detecting anomalous sea-level rise and flooding events via time-series and satellite imagery | anomaly detection, climate science, sea-level rise, time-series, remote sensing | Anomaly Detection         | Anomaly Detection | Detection of environmental anomalies                 | ROC-AUC, Precision/Recall                 | CNNs, RNNs, Transformers | 3.83           | [33]     |
|  | Vocal Call Locator (VCL)                  | Biology & Medicine      | Benchmarking sound-source localization of rodent vocalizations from multi-channel audio      | source localization, bioacoustics, time-series, SSL                             | Sound source localization | Regression        | Source localization accuracy in bioacoustic settings | Localization error (cm), Recall/Precision | CNN-based SSL models     | 3.83           | [34]     |

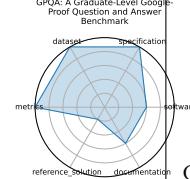
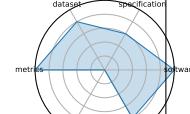
Continued on next page

| Ratings   | Name                                      | Domain    | Focus   | Keywords   | Task Types                                | AI/ML Motif | AI Capability  | Metrics   | Models   | Average Rating | Citation |
|---|---|-----------|---|--|---|-------------|--|---|--|----------------|----------|
|  | MassSpecGym - De novo molecule generation | Chemistry | Benchmark suite for discovery and identification of molecules via MS/MS | mass spectrometry, molecular structure, de novo generation, retrieval, dataset | De novo generation, Retrieval, Simulation | Generative  | Molecular identification and generation from spectral data | Structure accuracy, Retrieval precision, Simulation MSE | Graph-based generative models, Retrieval baselines | 3.75           | [35]     |
|  | MassSpecGym - Molecule Retrieval          | Chemistry | Benchmark suite for discovery and identification of molecules via MS/MS | mass spectrometry, molecular structure, de novo generation, retrieval, dataset | De novo generation, Retrieval, Simulation | Regression  | Molecular identification and generation from spectral data | Structure accuracy, Retrieval precision, Simulation MSE | Graph-based generative models, Retrieval baselines | 3.75           | [35]     |

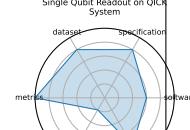
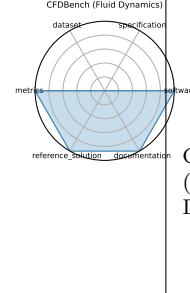
Continued on next page

| Ratings   | Name  | Domain                     | Focus   | Keywords   | Task Types   | AI/ML Motif          | AI Capability  | Metrics   | Models   | Average Rating | Citation |
|---|---|----------------------------|---|--|--|----------------------|--|---|--|----------------|----------|
|  | MassSpecGym - Spectrum Simulation                 | Chemistry                  | Benchmark suite for discovery and identification of molecules via MS/MS | mass spectrometry, molecular structure, de novo generation, retrieval, dataset | De novo generation, Retrieval, Simulation                      | Regression           | Molecular identification and generation from spectral data | Structure accuracy, Retrieval precision, Simulation MSE | Graph-based generative models, Retrieval baselines | 3.75           | [35]     |
|  | SPIQA (Scientific Paper Image Question Answering) | Computational Science & AI | Multimodal QA on scientific figures                                     | multimodal QA, figure understanding, table comprehension, chain-of-thought     | Question answering, Multimodal QA, Chain-of-Thought evaluation | Multimodal Reasoning | Visual-textual reasoning in scientific contexts            | Accuracy, F1 score                                      | Chain-of-Thought models, Multimodal QA systems     | 3.67           | [36]     |

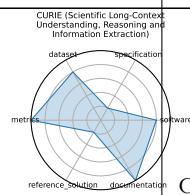
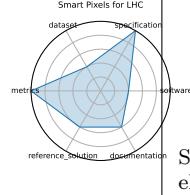
Continued on next page

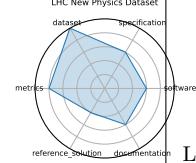
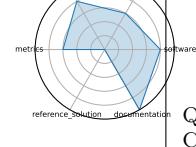
| Ratings   | Name  | Domain   | Focus  | Keywords  | Task Types      | AI/ML Motif                | AI Capability                             | Metrics  | Models                                    | Average Rating | Citation |
|---|---|--|--|---|-----------------|----------------------------|---|----------|---|----------------|----------|
|  | GPQA: A Graduate-Level Google-Proof Question and Answer Benchmark | Biology & Medicine, High Energy Physics, Chemistry | Graduate-level, expert-validated multiple-choice questions hard even with web access | Google-proof, multiple-choice, expert reasoning, science QA | Multiple choice | Reasoning & Generalization | Scientific reasoning, knowledge probing   | Accuracy | GPT-4 baseline                            | 3.67           | [37]     |
|  | MedQA   | Biology & Medicine                                 | Medical board exam QA  | USMLE, diagnostic QA, medical knowledge, multilingual       | Multiple choice | Reasoning & Generalization | Medical diagnosis and knowledge retrieval | Accuracy | Neural reader, Retrieval-based QA systems | 3.50           | [38]     |

Continued on next page

| Ratings   | Name                                | Domain                     | Focus   | Keywords                             | Task Types         | AI/ML Motif    | AI Capability                               | Metrics           | Models               | Average Rating | Citation |
|---|-------------------------------------|----------------------------|---|--------------------------------------|--------------------|----------------|---|-------------------|----------------------|----------------|----------|
|  | Single Qubit Readout on QICK System | Computational Science & AI | Real-time single-qubit state classification using FPGA firmware | qubit readout, hls4ml, FPGA, QICK    | Classification     | Classification | Single-shot fidelity, inference latency     | Accuracy, Latency | hls4ml quantized NN  | 3.50           | [39]     |
|  | CFDBench (Fluid Dynamics)           | Mathematics                | Neural operator surrogate modeling                              | neural operators, CFD, FNO, DeepONet | Surrogate modeling | Regression     | Generalization of neural operators for PDEs | L2 error, MAE     | FNO, DeepONet, U-Net | 3.33           | [40]     |

Continued on next page

| Ratings   | Name  | Domain   | Focus  | Keywords  | Task Types   | AI/ML Motif                | AI Capability                                       | Metrics                              | Models           | Average Rating | Citation |
|---|---|--|--|---|--|----------------------------|---|--------------------------------------|------------------|----------------|----------|
|  | CURIE (Scientific Long-Context Understanding, Reasoning and Information Extraction) | Materials Science, High Energy Physics, Biology & Medicine, Chemistry, Climate & Earth Science | Long-context scientific reasoning                                  | long-context, information extraction, multi-modal         | Information extraction, Reasoning, Concept tracking, Aggregation, Algebraic manipulation, Multimodal comprehension | Reasoning & Generalization | Long-context understanding and scientific reasoning | Accuracy                             | unkown           | 3.33           | [41]     |
|  | Smart Pixels for LHC  | High Energy Physics  | On-sensor, in-pixel ML filtering for high-rate LHC pixel detectors | smart pixel, on-sensor inference, data reduction, trigger | Image Classification, Data filtering   | Classification             | On-chip, low-power inference; data reduction        | Data rejection rate, Power per pixel | 2-layer pixel NN | 3.33           | [42]     |

| Ratings   | Name                               | Domain                     | Focus   | Keywords   | Task Types                                 | AI/ML Motif       | AI Capability   | Metrics                       | Models   | Average Rating | Citation |
|---|------------------------------------|----------------------------|---|--|--|-------------------|---|-------------------------------|--|----------------|----------|
|  | LHC New Physics Dataset            | High Energy Physics        | Real-time LHC event filtering for anomaly detection using proton collision data | anomaly detection, proton collision, real-time inference, event filtering, unsupervised ML | Anomaly Detection, Event classification    | Anomaly Detection | Unsupervised signal detection under latency and bandwidth constraints | ROC-AUC, Detection efficiency | Autoencoder, Variational autoencoder, Isolation forest | 8.33           | [43]     |
|  | Quantum Computing Benchmarks (QML) | Computational Science & AI | Quantum algorithm performance evaluation  | quantum circuits, state preparation, error correction                                      | Circuit benchmarking, State classification | Classification    | Quantum algorithm performance and fidelity                            | Fidelity, Success probability | IBM Q, IonQ, AQT@LBNL                                  | 3.17           | [44]     |

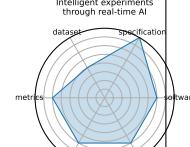
Continued on next page

| Ratings | Name                                       | Domain              | Focus  | Keywords   | Task Types     | AI/ML Motif    | AI Capability                              | Metrics                                 | Models                              | Average Rating | Citation |
|---------|--|---------------------|--|--|----------------|----------------|--|---|-------------------------------------|----------------|----------|
|         | Ultrafast jet classification at the HL-LHC | High Energy Physics | FPGA-optimized real-time jet origin classification at the HL-LHC       | jet classification, FPGA, quantization-aware training, Deep Sets, Interaction Networks | Classification | Classification | Real-time inference under FPGA constraints | Accuracy, Latency, Resource utilization | MLP, Deep Sets, Interaction Network | 3.17           | [45]     |
|         | HEDM (BraggNN)                             | Materials Science   | Fast Bragg peak analysis using deep learning in diffraction microscopy | BraggNN, diffraction, peak finding, HEDM   | Peak detection | Classification | High-throughput peak localization          | Localization accuracy, Inference time   | BraggNN                             | 3.17           | [46]     |

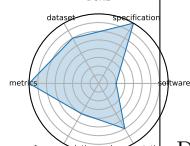
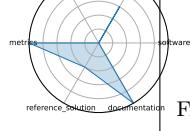
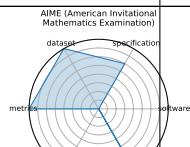
Continued on next page

| Ratings | Name         | Domain              | Focus   | Keywords  | Task Types                                    | AI/ML Motif                    | AI Capability                                       | Metrics                             | Models                 | Average Rating | Citation   |
|---------|--------------|---------------------|---|---|---|--------------------------------|---|-------------------------------------|------------------------|----------------|------------|
|         | 4D-STEM      | Materials Science   | Real-time ML for scanning transmission electron microscopy  | 4D-STEM, electron microscopy, real-time, image processing | Image Classification, Streamed data inference | Classification                 | Real-time large-scale microscopy inference          | Classification accuracy, Throughput | CNN models (prototype) | 3.17           | [47]       |
|         | Beam Control | High Energy Physics | Reinforcement learning control of accelerator beam position | RL, beam stabilization, control systems, simulation       | Control                                       | Reinforcement Learning/Control | Policy performance in simulated accelerator control | Stability, Control loss             | DDPG, PPO (planned)    | 3.00           | [48], [49] |

Continued on next page

| Ratings   | Name   | Domain              | Focus  | Keywords  | Task Types  | AI/ML Motif       | AI Capability                          | Metrics  | Models   | Average Rating | Citation |
|---|--|---------------------|--|---|---|-------------------|--|--|--|----------------|----------|
|  | Intelligent experiments through real-time AI | High Energy Physics | Real-time FPGA-based triggering and detector control for sPHENIX and future EIC      | FPGA, Graph Neural Network, hls4ml, real-time inference, detector control | Trigger classification, Detector control, Real-time inference | Classification    | Low-latency GNN inference on FPGA      | Accuracy (charm and beauty detection), Latency (micros), Resource utilization (LUT/FF) | Bipartite Graph Network with Set Transformers (BGN-ST), GarNet (edge-classifier) | 3.00           | [50]     |
|  | HDR ML Anomaly Challenge - Butterfly         | Biology & Medicine  | Detecting hybrid butterflies via image anomaly detection in genomic-informed dataset | anomaly detection, computer vision, genomics, butterfly hybrids           | Anomaly Detection   | Anomaly Detection | Hybrid detection in biological systems | Classification accuracy, F1 score  | CNN-based detectors  | 3.00           | [51]     |

Continued on next page

| Ratings  | Name   | Domain              | Focus                                       | Keywords   | Task Types                                       | AI/ML Motif                | AI Capability                                | Metrics                       | Models              | Average Rating | Citation |
|--|--|---------------------|---|--|--|----------------------------|--|-------------------------------|---------------------|----------------|----------|
|   | DUNE   | High Energy Physics | Real-time ML for DUNE DAQ time-series data  | DUNE, time-series, real-time, trigger                                  | Trigger selection, Time-series anomaly detection | Anomaly Detection          | Low-latency event detection                  | Detection efficiency, Latency | CNN, LSTM (planned) | 2.83           | [52]     |
|   | FrontierMath   | Mathematics         | Challenging advanced mathematical reasoning | symbolic reasoning, number theory, algebraic geometry, category theory | Problem solving                                  | Reasoning & Generalization | Symbolic and abstract mathematical reasoning | Accuracy                      | unknown             | 2.50           | [53]     |
|  | AIME (American Invitational Mathematics Examination) | Mathematics         | Pre-college advanced problem solving        | algebra, combinatorics, number theory, geometry                        | Problem solving                                  | Reasoning & Generalization | Mathematical problem-solving and reasoning   | Accuracy                      | unknown             | 2.33           | [54]     |

Continued on next page

| Ratings | Name                               | Domain              | Focus   | Keywords  | Task Types                             | AI/ML Motif       | AI Capability  | Metrics                    | Models                                     | Average Rating | Citation |
|---------|------------------------------------|---------------------|---|---|--|-------------------|--|----------------------------|--|----------------|----------|
|         | Quench detection                   | High Energy Physics | Real-time detection of superconducting magnet quenches using ML | quench detection, autoencoder, anomaly detection, real-time | Anomaly Detection, Quench localization | Anomaly Detection | Real-time anomaly detection with multi-modal sensors | ROC-AUC, Detection latency | Autoencoder RL agents (in development)     | 2.17           | [55]     |
|         | Materials Project                  | Materials Science   | DFT-based property prediction                                   | DFT, materials genome, high-throughput                      | Property prediction                    | Regression        | Prediction of inorganic material properties          | MAE, R^2                   | Automatmine, Crystal Graph Neural Networks | 0.2            | [56]     |
|         | In-Situ High-Speed Computer Vision | High Energy Physics | Real-time image classification for in-situ plasma diagnostics   | plasma, in-situ vision, real-time ML                        | Image Classification                   | Classification    | Real-time diagnostic inference                       | Accuracy, FPS              | CNN  | 1.50           | [57]     |

## 2 Radar Chart Table

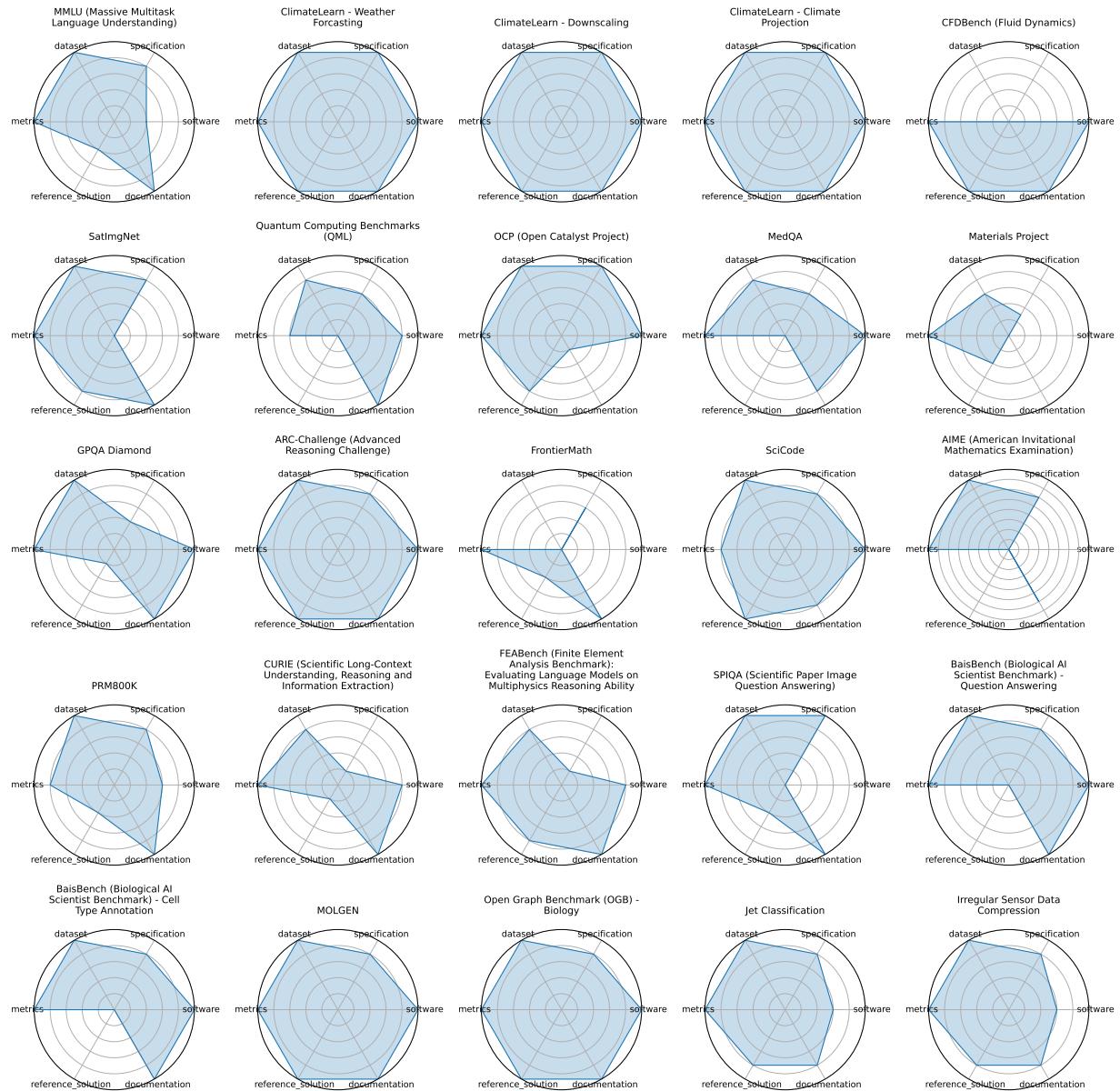


Figure 1: Radar chart overview (page 1)

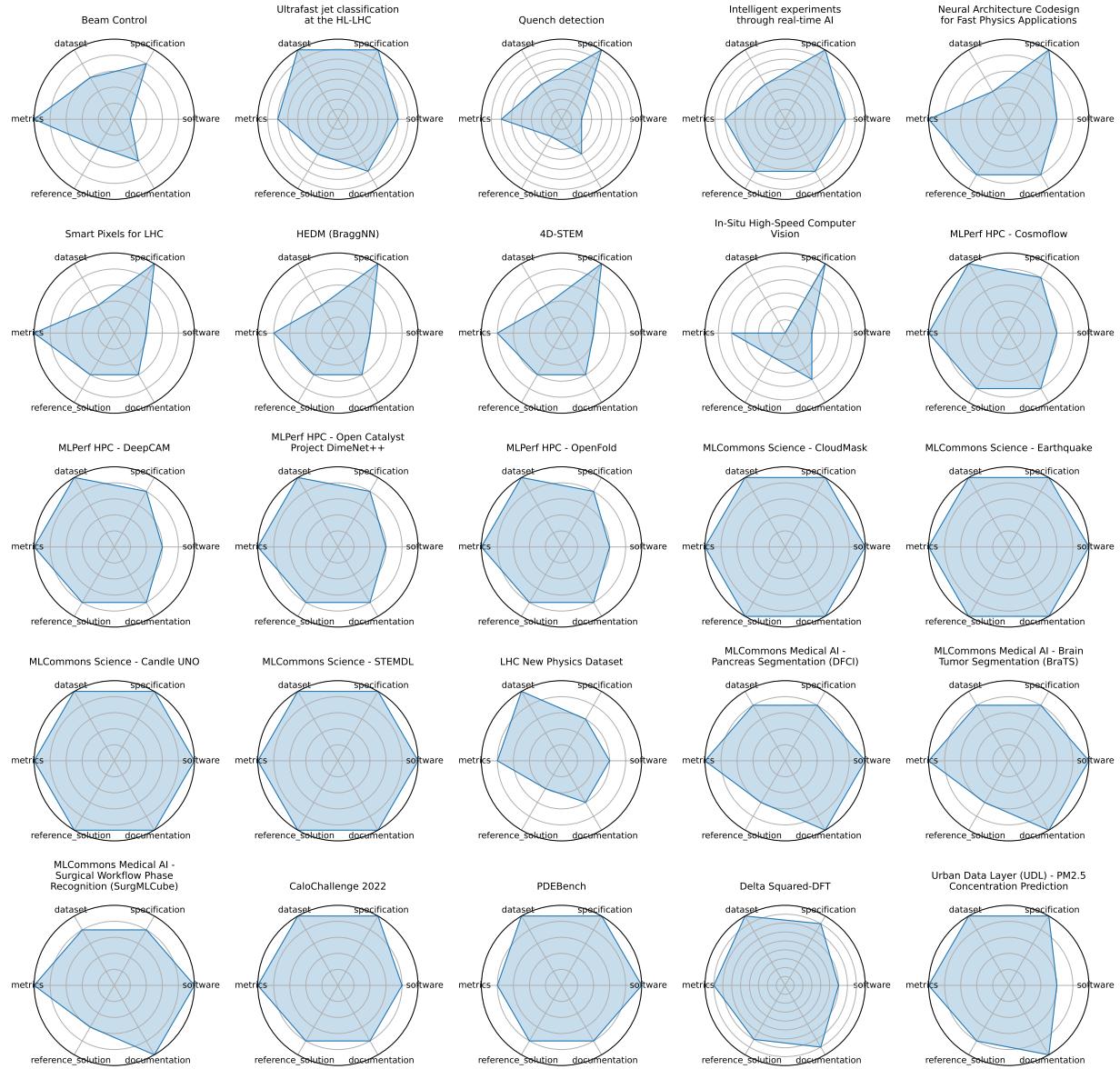


Figure 2: Radar chart overview (page 2)

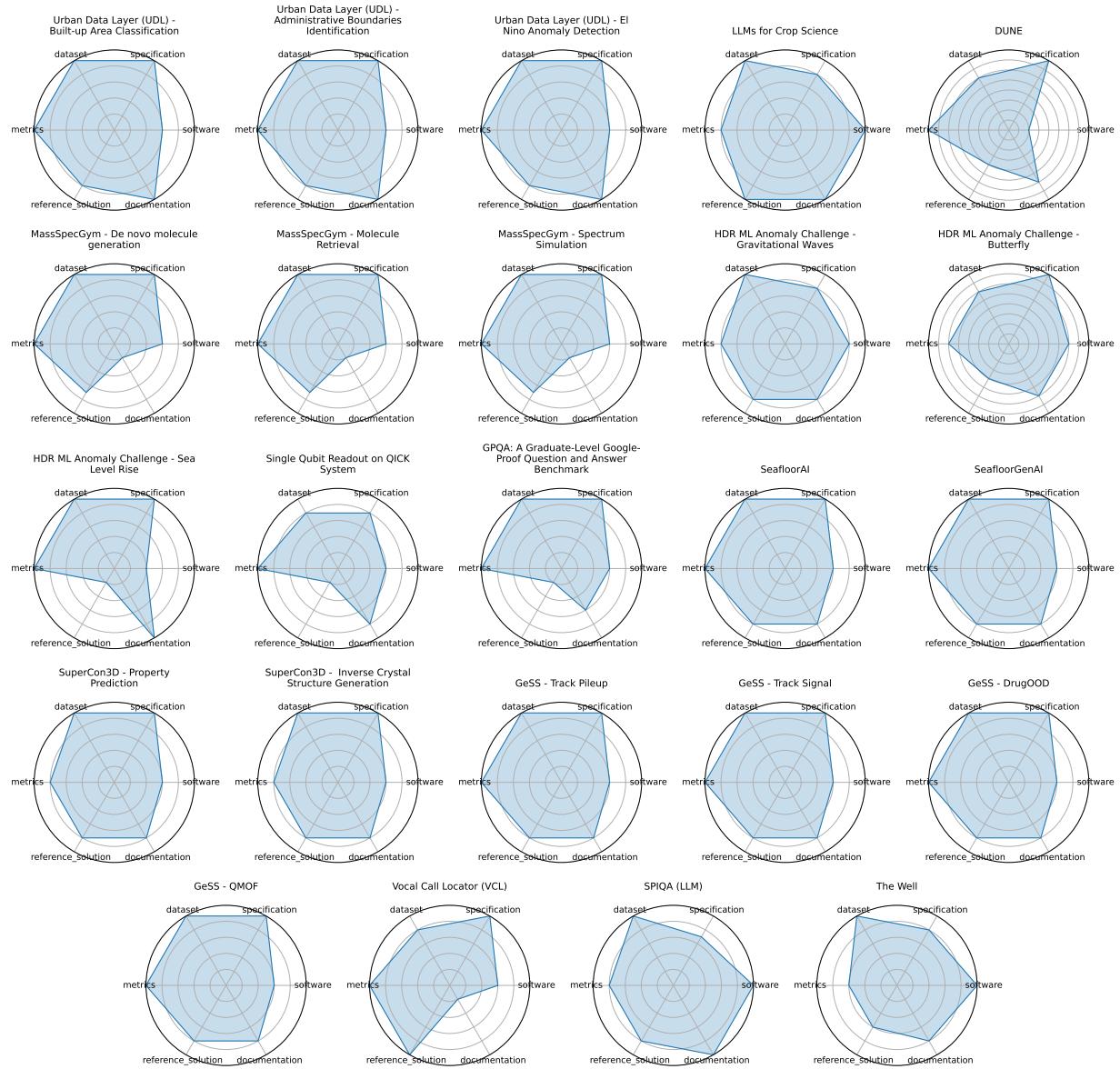


Figure 3: Radar chart overview (page 3)

### 3 Benchmark Details

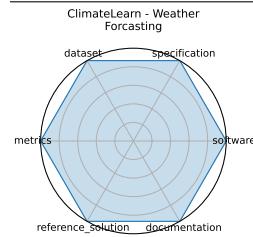
#### 3.1 ClimateLearn - Weather Forcasting

ClimateLearn provides standardized datasets and evaluation protocols for machine learning models in medium-range weather and climate forecasting using ERA5 reanalysis.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2023-07-19  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2023-07-19  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2023-07-19  |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2307.01909">https://arxiv.org/abs/2307.01909</a>                             |
| <b>doi:</b>                    | 10.48550/arXiv.2307.01909   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | ML for weather and climate modeling   |
| <b>keywords:</b>               | - medium-range forecasting - ERA5 - data-driven   |
| <b>licensing:</b>              | CC-BY-4.0   |
| <b>task_types:</b>             | - Forecasting   |
| <b>ai_capability_measured:</b> | - Global weather prediction (3-5 days)  |
| <b>metrics:</b>                | - RMSE - Anomaly correlation  |
| <b>models:</b>                 | - CNN baselines - ResNet variants   |
| <b>ml_motif:</b>               | - Sequence Prediction/Forecasting   |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | Multiple baseline models provided   |
| <b>notes:</b>                  | Includes physical and ML baselines.   |
| <b>contact.name:</b>           | Jason Jewik   |
| <b>contact.email:</b>          | jason.jewik@ucla.edu  |
| <b>datasets.links.name:</b>    | ClimateLearn GitHub Repository (data loaders and processing)  |
| <b>datasets.links.url:</b>     | <a href="https://github.com/aditya-grover/climate-learn">https://github.com/aditya-grover/climate-learn</a> |
| <b>results.links.name:</b>     | ClimateLearn Paper (results section)  |
| <b>results.links.url:</b>      | <a href="https://arxiv.org/abs/2307.01909">https://arxiv.org/abs/2307.01909</a>                             |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | climatelearn_-weather_forcasting  |
| <b>Citations:</b>              | [1]   |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Provides standardized access to ERA5 and other reanalysis datasets, with ML-ready splits, metadata, and Xarray-compatible formats; versioned and fully FAIR-compliant.                            |
| documentation      | 5     | Explained in the benchmark's paper.   |
| metrics            | 5     | ACC and RMSE are standard, quantitative, and appropriate for climate forecasting; well-integrated into the benchmark, though interpretation across domains may vary.                              |
| reference_solution | 5     | A Quickstart notebook is provided that uses ResNet as a baseline model  |
| software           | 5     | Quickstart notebook makes for easy usage  |
| specification      | 5     | Task framing (medium-range climate forecasting), input/output formats, and evaluation windows are clearly defined; benchmark supports both physical and learned models with detailed constraints. |



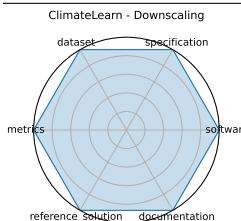
## 3.2 ClimateLearn - Downscaling

ClimateLearn provides standardized datasets and evaluation protocols for machine learning models in medium-range weather and climate forecasting using ERA5 reanalysis.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2023-07-19  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2023-07-19  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2023-07-19  |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2307.01909">https://arxiv.org/abs/2307.01909</a>                             |
| <b>doi:</b>                    | 10.48550/arXiv.2307.01909   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | ML for weather and climate modeling   |
| <b>keywords:</b>               | - medium-range forecasting - ERA5 - data-driven   |
| <b>licensing:</b>              | CC-BY-4.0   |
| <b>task_types:</b>             | - Forecasting   |
| <b>ai_capability_measured:</b> | - Global weather prediction (3-5 days)  |
| <b>metrics:</b>                | - RMSE - Anomaly correlation  |
| <b>models:</b>                 | - CNN baselines - ResNet variants   |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | Multiple baseline models provided   |
| <b>notes:</b>                  | Includes physical and ML baselines.   |
| <b>contact.name:</b>           | Jason Jewik   |
| <b>contact.email:</b>          | jason.jewik@ucla.edu  |
| <b>datasets.links.name:</b>    | ClimateLearn GitHub Repository (data loaders and processing)  |
| <b>datasets.links.url:</b>     | <a href="https://github.com/aditya-grover/climate-learn">https://github.com/aditya-grover/climate-learn</a> |
| <b>results.links.name:</b>     | ClimateLearn Paper (results section)  |
| <b>results.links.url:</b>      | <a href="https://arxiv.org/abs/2307.01909">https://arxiv.org/abs/2307.01909</a>                             |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | climatelearn_-_downscaling  |
| <b>Citations:</b>              | [1]   |

### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Provides standardized access to ERA5 and other reanalysis datasets, with ML-ready splits, metadata, and Xarray-compatible formats; versioned and fully FAIR-compliant.                            |
| documentation      | 5     | Explained in the benchmark's paper.   |
| metrics            | 5     | ACC and RMSE are standard, quantitative, and appropriate for climate forecasting; well-integrated into the benchmark, though interpretation across domains may vary.                              |
| reference_solution | 5     | A Quickstart notebook is provided that uses ResNet as a baseline model  |
| software           | 5     | Quickstart notebook makes for easy usage  |
| specification      | 5     | Task framing (medium-range climate forecasting), input/output formats, and evaluation windows are clearly defined; benchmark supports both physical and learned models with detailed constraints. |



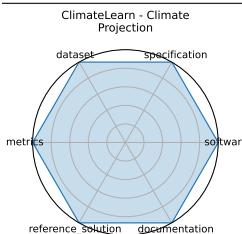
### 3.3 ClimateLearn - Climate Projection

ClimateLearn provides standardized datasets and evaluation protocols for machine learning models in medium-range weather and climate forecasting using ERA5 reanalysis.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2023-07-19  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2023-07-19  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2023-07-19  |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2307.01909">https://arxiv.org/abs/2307.01909</a>                             |
| <b>doi:</b>                    | 10.48550/arXiv.2307.01909   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | ML for weather and climate modeling   |
| <b>keywords:</b>               | - medium-range forecasting - ERA5 - data-driven   |
| <b>licensing:</b>              | CC-BY-4.0   |
| <b>task_types:</b>             | - Forecasting   |
| <b>ai_capability_measured:</b> | - Global weather prediction (3-5 days)  |
| <b>metrics:</b>                | - RMSE - Anomaly correlation  |
| <b>models:</b>                 | - CNN baselines - ResNet variants   |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | Multiple baseline models provided   |
| <b>notes:</b>                  | Includes physical and ML baselines.   |
| <b>contact.name:</b>           | Jason Jewik   |
| <b>contact.email:</b>          | jason.jewik@ucla.edu  |
| <b>datasets.links.name:</b>    | ClimateLearn GitHub Repository (data loaders and processing)  |
| <b>datasets.links.url:</b>     | <a href="https://github.com/aditya-grover/climate-learn">https://github.com/aditya-grover/climate-learn</a> |
| <b>results.links.name:</b>     | ClimateLearn Paper (results section)  |
| <b>results.links.url:</b>      | <a href="https://arxiv.org/abs/2307.01909">https://arxiv.org/abs/2307.01909</a>                             |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | climatelearn_-_climate_projection   |
| <b>Citations:</b>              | [1]   |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Provides standardized access to ERA5 and other reanalysis datasets, with ML-ready splits, metadata, and Xarray-compatible formats; versioned and fully FAIR-compliant.                            |
| documentation      | 5     | Explained in the benchmark's paper.   |
| metrics            | 5     | ACC and RMSE are standard, quantitative, and appropriate for climate forecasting; well-integrated into the benchmark, though interpretation across domains may vary.                              |
| reference_solution | 5     | A Quickstart notebook is provided that uses ResNet as a baseline model  |
| software           | 5     | Quickstart notebook makes for easy usage  |
| specification      | 5     | Task framing (medium-range climate forecasting), input/output formats, and evaluation windows are clearly defined; benchmark supports both physical and learned models with detailed constraints. |



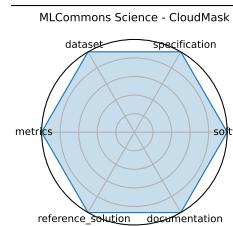
### 3.4 MLCommons Science - CloudMask

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.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2023-06-01  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2023-06   |
| <b>expired:</b>                | no  |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2023-06-01  |
| <b>url:</b>                    | <a href="https://github.com/mlcommons/science">https://github.com/mlcommons/science</a>   |
| <b>doi:</b>                    | 10.1007/978-3-031-23220-6_4   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | AI benchmarks for scientific applications including time-series, imaging, and simulation  |
| <b>keywords:</b>               | - science AI - benchmark - MLCommons - HPC  |
| <b>licensing:</b>              | Apache License 2.0  |
| <b>task_types:</b>             | - Time-series analysis - Image classification - Simulation surrogate modeling   |
| <b>ai_capability_measured:</b> | - Inference accuracy - simulation speed-up - generalization   |
| <b>metrics:</b>                | - MAE - Accuracy - Speedup vs simulation  |
| <b>models:</b>                 | - CNN - GNN - Transformer   |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Framework   |
| <b>ml_task:</b>                | - NA  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Joint effort under Apache-2.0 license.  |
| <b>contact.name:</b>           | MLCommons Science Working Group   |
| <b>contact.email:</b>          | science-chairs@mlcommons.org  |
| <b>datasets.links.name:</b>    | CANDLE UNO  |
| <b>datasets.links.url:</b>     | <a href="https://github.com/mlcommons/science/tree/main/benchmarks/uno#data-description">https://github.com/mlcommons/science/tree/main/benchmarks/uno#data-description</a> |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | mlcommons_science_-_cloudmask   |
| <b>Citations:</b>              | [2]   |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Public scientific datasets are used with defined splits. At least 4 FAIR principles are followed.  |
| documentation      | 5     | Thorough documentation exists covering the task, background, motivation, evaluation criteria, and includes a supporting paper.   |
| metrics            | 5     | Clearly defined metrics such as accuracy, training time, and GPU utilization are used. These metrics are explained and effectively capture solution performance.                                       |
| reference_solution | 5     | A reference implementation is available, well-documented, trainable/open, and includes full metric evaluation and software/hardware details.   |
| software           | 5     | Actively maintained GitHub repository available at <a href="https://github.com/mlcommons/science">https://github.com/mlcommons/science</a> with implementations, scripts, and reproducibility support. |
| specification      | 5     | All five specification aspects are covered: system constraints, task, dataset format, benchmark inputs, and outputs.   |



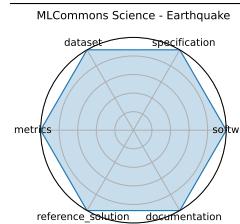
### 3.5 MLCommons Science - Earthquake

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.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2023-06-01  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2023-06   |
| <b>expired:</b>                | no  |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2023-06-01  |
| <b>url:</b>                    | <a href="https://github.com/mlcommons/science">https://github.com/mlcommons/science</a>   |
| <b>doi:</b>                    | 10.1007/978-3-031-23220-6_4   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | AI benchmarks for scientific applications including time-series, imaging, and simulation  |
| <b>keywords:</b>               | - science AI - benchmark - MLCommons - HPC  |
| <b>licensing:</b>              | Apache License 2.0  |
| <b>task_types:</b>             | - Time-series analysis - Image classification - Simulation surrogate modeling   |
| <b>ai_capability_measured:</b> | - Inference accuracy - simulation speed-up - generalization   |
| <b>metrics:</b>                | - MAE - Accuracy - Speedup vs simulation  |
| <b>models:</b>                 | - CNN - GNN - Transformer   |
| <b>ml_motif:</b>               | - Sequence Prediction/Forecasting   |
| <b>type:</b>                   | Framework   |
| <b>ml_task:</b>                | - NA  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Joint effort under Apache-2.0 license.  |
| <b>contact.name:</b>           | MLCommons Science Working Group   |
| <b>contact.email:</b>          | science-chairs@mlcommons.org  |
| <b>datasets.links.name:</b>    | CANDLE UNO  |
| <b>datasets.links.url:</b>     | <a href="https://github.com/mlcommons/science/tree/main/benchmarks/uno#data-description">https://github.com/mlcommons/science/tree/main/benchmarks/uno#data-description</a> |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | mlcommons_science_-_earthquake  |
| <b>Citations:</b>              | [2]   |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Public scientific datasets are used with defined splits. At least 4 FAIR principles are followed.  |
| documentation      | 5     | Thorough documentation exists covering the task, background, motivation, evaluation criteria, and includes a supporting paper.   |
| metrics            | 5     | Clearly defined metrics such as accuracy, training time, and GPU utilization are used. These metrics are explained and effectively capture solution performance.                                       |
| reference_solution | 5     | A reference implementation is available, well-documented, trainable/open, and includes full metric evaluation and software/hardware details.   |
| software           | 5     | Actively maintained GitHub repository available at <a href="https://github.com/mlcommons/science">https://github.com/mlcommons/science</a> with implementations, scripts, and reproducibility support. |
| specification      | 5     | All five specification aspects are covered: system constraints, task, dataset format, benchmark inputs, and outputs.   |



### 3.6 MLCommons Science - Candle UNO

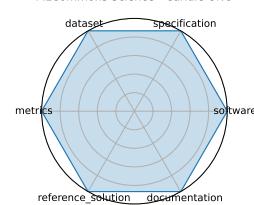
MLCommons Science assembles benchmark tasks with datasets, targets, and implementations across earthquake forecasting, satellite imagery, drug screening, electron microscopy, and CFD to drive scientific ML reproducibility.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2023-06-01  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2023-06   |
| <b>expired:</b>                | no  |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2023-06-01  |
| <b>url:</b>                    | <a href="https://github.com/mlcommons/science">https://github.com/mlcommons/science</a>   |
| <b>doi:</b>                    | 10.1007/978-3-031-23220-6_4   |
| <b>domain:</b>                 | - Biology & Medicine  |
| <b>focus:</b>                  | AI benchmarks for scientific applications including time-series, imaging, and simulation  |
| <b>keywords:</b>               | - science AI - benchmark - MLCommons - HPC  |
| <b>licensing:</b>              | Apache License 2.0  |
| <b>task_types:</b>             | - Time-series analysis - Image classification - Simulation surrogate modeling   |
| <b>ai_capability_measured:</b> | - Inference accuracy - simulation speed-up - generalization   |
| <b>metrics:</b>                | - MAE - Accuracy - Speedup vs simulation  |
| <b>models:</b>                 | - CNN - GNN - Transformer   |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Framework   |
| <b>ml_task:</b>                | - NA  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Joint effort under Apache-2.0 license.  |
| <b>contact.name:</b>           | MLCommons Science Working Group   |
| <b>contact.email:</b>          | science-chairs@mlcommons.org  |
| <b>datasets.links.name:</b>    | CANDLE UNO  |
| <b>datasets.links.url:</b>     | <a href="https://github.com/mlcommons/science/tree/main/benchmarks/uno#data-description">https://github.com/mlcommons/science/tree/main/benchmarks/uno#data-description</a> |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | mlcommons_science_-_candle_uno  |
| <b>Citations:</b>              | [2]   |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Public scientific datasets are used with defined splits. At least 4 FAIR principles are followed.  |
| documentation      | 5     | Thorough documentation exists covering the task, background, motivation, evaluation criteria, and includes a supporting paper.   |
| metrics            | 5     | Clearly defined metrics such as accuracy, training time, and GPU utilization are used. These metrics are explained and effectively capture solution performance.                                       |
| reference_solution | 5     | A reference implementation is available, well-documented, trainable/open, and includes full metric evaluation and software/hardware details.   |
| software           | 5     | Actively maintained GitHub repository available at <a href="https://github.com/mlcommons/science">https://github.com/mlcommons/science</a> with implementations, scripts, and reproducibility support. |
| specification      | 5     | All five specification aspects are covered: system constraints, task, dataset format, benchmark inputs, and outputs.   |

MLCommons Science - Candle UNO



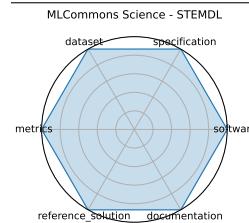
### 3.7 MLCommons Science - STEM DL

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.

**date:** 2023-06-01  
**version:** v1.0  
**last\_updated:** 2023-06  
**expired:** no  
**valid:** yes  
**valid\_date:** 2023-06-01  
**url:** <https://github.com/mlcommons/science>  
**doi:** 10.1007/978-3-031-23220-6\_4  
**domain:** - Materials Science  
**focus:** AI benchmarks for scientific applications including time-series, imaging, and simulation  
**keywords:** - science AI - benchmark - MLCommons - HPC  
**licensing:** Apache License 2.0  
**task\_types:** - Time-series analysis - Image classification - Simulation surrogate modeling  
**ai\_capability\_measured:** - Inference accuracy - simulation speed-up - generalization  
**metrics:** - MAE - Accuracy - Speedup vs simulation  
**models:** - CNN - GNN - Transformer  
**ml\_motif:** - Classification  
**type:** Framework  
**ml\_task:** - NA  
**solutions:** 0  
**notes:** Joint effort under Apache-2.0 license.  
**contact.name:** MLCommons Science Working Group  
**contact.email:** science-chairs@mlcommons.org  
**datasets.links.name:** A Database of Convergent Beam Electron Diffraction Patterns for Machine Learning of the Structural Properties of Materials  
**datasets.links.url:** <https://doi.ccs.ornl.gov/dataset/7aed61eb-e44c-5b14-82ea-07917d1b2d3b>  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** mlcommons\_science -- stemdl  
**Citations:** [2]

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Public scientific datasets are used with defined splits. At least 4 FAIR principles are followed.  |
| documentation      | 5     | Thorough documentation exists covering the task, background, motivation, evaluation criteria, and includes a supporting paper.   |
| metrics            | 5     | Clearly defined metrics such as accuracy, training time, and GPU utilization are used. These metrics are explained and effectively capture solution performance.                                       |
| reference_solution | 5     | A reference implementation is available, well-documented, trainable/open, and includes full metric evaluation and software/hardware details.   |
| software           | 5     | Actively maintained GitHub repository available at <a href="https://github.com/mlcommons/science">https://github.com/mlcommons/science</a> with implementations, scripts, and reproducibility support. |
| specification      | 5     | All five specification aspects are covered: system constraints, task, dataset format, benchmark inputs, and outputs.   |



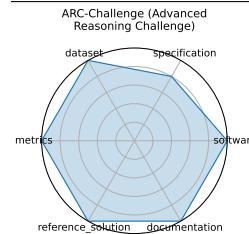
### 3.8 ARC-Challenge (Advanced Reasoning Challenge)

The AI2 Reasoning Challenge (ARC) Challenge set comprises 7,787 natural, grade-school science questions that retrieval-based and word co-occurrence algorithms both fail, requiring advanced reasoning over a 14-million-sentence corpus.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2018-03-14  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2018-03-14  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2018-03-14  |
| <b>url:</b>                    | <a href="https://allenai.org/data/arc">https://allenai.org/data/arc</a>                                       |
| <b>doi:</b>                    | 10.48550/arXiv.1803.05457   |
| <b>domain:</b>                 | - Computational Science & AI  |
| <b>focus:</b>                  | Grade-school science with reasoning emphasis  |
| <b>keywords:</b>               | - grade-school - science QA - challenge set - reasoning   |
| <b>licensing:</b>              | Apache 2.0 License  |
| <b>task_types:</b>             | - Multiple choice   |
| <b>ai_capability_measured:</b> | - Commonsense and scientific reasoning  |
| <b>metrics:</b>                | - Accuracy  |
| <b>models:</b>                 | - GPT-4 - Claude  |
| <b>ml_motif:</b>               | - Reasoning & Generalization  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Good  |
| <b>contact.name:</b>           | unknown   |
| <b>contact.email:</b>          | unknown   |
| <b>datasets.links.name:</b>    | Hugging Face  |
| <b>datasets.links.url:</b>     | <a href="https://huggingface.co/datasets/allenai/ai2_arc">https://huggingface.co/datasets/allenai/ai2_arc</a> |
| <b>results.links.name:</b>     | ARC-Solvers   |
| <b>results.links.url:</b>      | <a href="https://github.com/allenai/arc-solvers">https://github.com/allenai/arc-solvers</a>                   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | arc-challenge_advanced_reasoning_challenge  |
| <b>Citations:</b>              | [3]   |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Data accessible, offers instructions on how to download the data via CLI tools. Splits provided on Huggingface |
| documentation      | 5     | Explains all necessary information inside a paper  |
| metrics            | 5     | All questions in the dataset are multiple choice, all have a correct answer                                    |
| reference_solution | 5     | Reference solution is available and containerized  |
| software           | 5     | Code is available and well documented for evaluation.  |
| specification      | 4     | Task is clear and inputs/outputs are provided along with format on dataset card.                               |



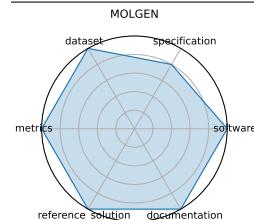
### 3.9 MOLGEN

MolGen is a pre-trained molecular language model that generates chemically valid molecules using SELFIES and reinforcement learning, guided by chemical feedback to optimize properties such as logP, QED, and docking score.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-17  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2023-01-26  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2023-01-26  |
| <b>url:</b>                    | <a href="https://github.com/zjunlp/MolGen">https://github.com/zjunlp/MolGen</a>                     |
| <b>doi:</b>                    | 10.48550/arXiv.2301.11259   |
| <b>domain:</b>                 | - Chemistry   |
| <b>focus:</b>                  | Molecular generation and optimization   |
| <b>keywords:</b>               | - SELFIES - GAN - property optimization   |
| <b>licensing:</b>              | MIT License   |
| <b>task_types:</b>             | - Distribution learning - Goal-oriented generation  |
| <b>ai_capability_measured:</b> | - Generation of valid and optimized molecular structures  |
| <b>metrics:</b>                | - Validity% - Novelty% - QED - Docking score - penalized logP                                       |
| <b>models:</b>                 | - MolGen  |
| <b>ml_motif:</b>               | - Generative  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  |   |
| <b>contact.name:</b>           | zhangningyu@zju.edu.cn  |
| <b>contact.email:</b>          | Ningyu Zhang  |
| <b>datasets.links.name:</b>    | MolGen: A Pre-trained Molecular Language Model  |
| <b>datasets.links.url:</b>     | <a href="https://github.com/zjunlp/MolGen/tree/main">https://github.com/zjunlp/MolGen/tree/main</a> |
| <b>results.links.name:</b>     | Domain-Agnostic Molecular Generation with Chemical Feedback   |
| <b>results.links.url:</b>      | <a href="https://arxiv.org/abs/2301.11259">https://arxiv.org/abs/2301.11259</a>                     |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | molgen  |
| <b>Citations:</b>              | [4]   |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Dataset and train/test splits are available through the github repo, as well as mentions of source datasets in the paper. |
| documentation      | 5     | All necessary information is provided in the paper and github repo  |
| metrics            | 5     | Metrics are well defined and appropriate for the task   |
| reference_solution | 5     | A pretrained model is provided, as well as training code and instructions   |
| software           | 5     | Code is available on the github repo, along with instructions to run the model and reproduce results.                     |
| specification      | 4     | Task, dataset format, and input/output formats are well specified. No system constraints are mentioned.                   |



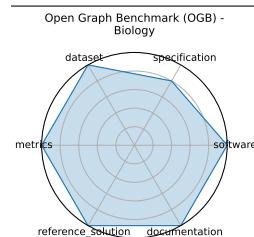
### 3.10 Open Graph Benchmark (OGB) - Biology

OGB-Biology is a suite of large-scale biological network datasets (protein-protein interaction, drug-target, etc.) with standardized splits and evaluation protocols for node, link, and graph property prediction tasks.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2020-05-02  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2020-05-02  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2020-05-02  |
| <b>url:</b>                    | <a href="https://ogb.stanford.edu/docs/home/">https://ogb.stanford.edu/docs/home/</a>                         |
| <b>doi:</b>                    | 10.48550/arXiv.2005.00687   |
| <b>domain:</b>                 | - Biology & Medicine  |
| <b>focus:</b>                  | Biological graph property prediction  |
| <b>keywords:</b>               | - node prediction - link prediction - graph classification  |
| <b>licensing:</b>              | MIT License   |
| <b>task_types:</b>             | - Node property prediction - Link property prediction - Graph property prediction                             |
| <b>ai_capability_measured:</b> | - Scalability and generalization in graph ML for biology  |
| <b>metrics:</b>                | - Accuracy - ROC-AUC  |
| <b>models:</b>                 | - GCN - GraphSAGE - GAT   |
| <b>ml_motif:</b>               | - Sequence Prediction/Forecasting   |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Community-driven updates  |
| <b>contact.name:</b>           | OGB Team  |
| <b>contact.email:</b>          | ogb@cs.stanford.edu   |
| <b>datasets.links.name:</b>    | OGB Webpage   |
| <b>datasets.links.url:</b>     | <a href="https://ogb.stanford.edu/docs/dataset_overview/">https://ogb.stanford.edu/docs/dataset_overview/</a> |
| <b>results.links.name:</b>     | unknown   |
| <b>results.links.url:</b>      | unknown   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | open_graph_benchmark_ogb_-_biology  |
| <b>Citations:</b>              | [5]   |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Fully FAIR- datasets are versioned, split, and accessible via a standardized API; extensive metadata and documentation are included.             |
| documentation      | 5     | All necessary information is included in a paper.  |
| metrics            | 5     | Reproducible, quantitative metrics (e.g., ROC-AUC, accuracy) that are tightly aligned with the tasks.  |
| reference_solution | 5     | Multiple baselines implemented and documented (GCN, GAT, GraphSAGE).   |
| software           | 5     | All necessary information is provided on the Github  |
| specification      | 4     | Tasks (node/link/graph property prediction) are clearly specified with input/output formats and standardized protocols; splits are well-defined. |



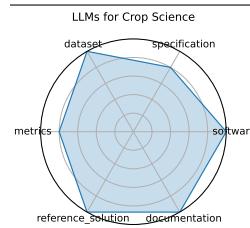
### 3.11 LLMs for Crop Science

Establishes a benchmark of over 5000 expert-annotated QA pairs and prompts in Chinese and English, 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.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-11-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-11   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-11-13  |
| <b>url:</b>                    | <a href="https://openreview.net/forum?id=hMj6jZ6JWU#discussion">https://openreview.net/forum?id=hMj6jZ6JWU#discussion</a> |
| <b>doi:</b>                    | N/A   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | Evaluating LLMs on crop trait QA and textual inference tasks with domain-specific prompts                                 |
| <b>keywords:</b>               | - crop science - prompt engineering - domain adaptation - question answering  |
| <b>licensing:</b>              | CC-BY-NC-4.0  |
| <b>task_types:</b>             | - Question Answering - Inference  |
| <b>ai_capability_measured:</b> | - Scientific knowledge - crop reasoning   |
| <b>metrics:</b>                | - Accuracy - F1 score   |
| <b>models:</b>                 | - GPT-3.5 - GPT-4 - Claude-3-opus - Qwen-max - LLama3-8B - InternLM2-7B - Qwen1.5-7B                                      |
| <b>ml_motif:</b>               | - Reasoning & Generalization  |
| <b>type:</b>                   | Dataset   |
| <b>ml_task:</b>                | - QA, inference   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.   |
| <b>notes:</b>                  | Includes examples with retrieval-augmented and chain-of-thought prompt templates; supports few-shot adaptation.           |
| <b>contact.name:</b>           | Deepak Patel  |
| <b>contact.email:</b>          | unknown   |
| <b>datasets.links.name:</b>    | CROP Benchmark (Test Split)   |
| <b>datasets.links.url:</b>     | <a href="https://huggingface.co/datasets/AI4Agr/CROP-benchmark">https://huggingface.co/datasets/AI4Agr/CROP-benchmark</a> |
| <b>results.links.name:</b>     | Empowering and Assessing the Utility of Large Language Models in Crop Science - Experiments                               |
| <b>results.links.url:</b>      | <a href="https://renqichen.github.io/The_Crop/">https://renqichen.github.io/The_Crop/</a>                                 |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | llms_for_crop_science   |
| <b>Citations:</b>              | [6]   |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Dataset adheres to all FAIR principles, is well-documented, and publicly available on Hugging Face. Train/Test splits are provided across two Huggingface datasets. |
| documentation      | 5     | The benchmark is well documented with a detailed paper, README, and webpage. Instructions for reproducing results are clear.  |
| metrics            | 4     | Accuracy is mentioned in the README and webpage as an evaluation metric,  |
| reference_solution | 5     | A reference solution is available and well documented. Training code is provided for multiple open weight models.   |
| software           | 5     | Code for evaluation and training of multiple models is available and well documented. Environment details are provided.   |
| specification      | 4     | Tasks are clearly defined (QA, inference) with structured input/output formats, though no system constraints are provided.  |



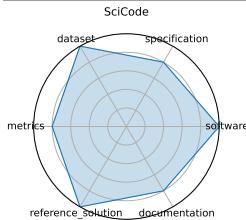
### 3.12 SciCode

SciCode is a scientist-curated coding benchmark with 338 subproblems derived from 80 real research tasks across 16 scientific subfields, evaluating models on knowledge recall, reasoning, and code synthesis for scientific computing tasks.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-07-18  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2024-07-18  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-07-18  |
| <b>url:</b>                    | <a href="https://scicode-bench.github.io/">https://scicode-bench.github.io/</a>                                 |
| <b>doi:</b>                    | 10.48550/arXiv.2407.13168   |
| <b>domain:</b>                 | - Computational Science & AI  |
| <b>focus:</b>                  | Scientific code generation and problem solving  |
| <b>keywords:</b>               | - code synthesis - scientific computing - programming benchmark   |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Coding  |
| <b>ai_capability_measured:</b> | - Program synthesis, scientific computing   |
| <b>metrics:</b>                | - Solve rate (%)  |
| <b>models:</b>                 | - Claude3.5-Sonnet  |
| <b>ml_motif:</b>               | - Generative  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | unknown   |
| <b>notes:</b>                  | Good  |
| <b>contact.name:</b>           | Minyang Tian  |
| <b>contact.email:</b>          | mtian8@illinois.edu   |
| <b>datasets.links.name:</b>    | SciCode on Huggingface  |
| <b>datasets.links.url:</b>     | <a href="https://huggingface.co/datasets/SciCode1/SciCode">https://huggingface.co/datasets/SciCode1/SciCode</a> |
| <b>results.links.name:</b>     | SciCode Learderboard  |
| <b>results.links.url:</b>      | <a href="https://scicode-bench.github.io/leaderboard/">https://scicode-bench.github.io/leaderboard/</a>         |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | scicode   |
| <b>Citations:</b>              | [7]   |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Dataset meets all FAIR principles, test and validation splits are available (no train split)             |
| documentation      | 4     | Paper containing all needed info except for evlauation criteria  |
| metrics            | 4     | Metrics stated, grading guidelines are provided in repo (problems are pass/fail)                         |
| reference_solution | 5     | Code to evaluate is available and well documented. Baseline models include closed and open weight models |
| software           | 5     | Code to run exists on github repo  |
| specification      | 4     | Expected outputs and broad types of inputs stated. Few details on output grading. No HW constraints.     |



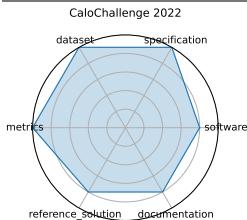
### 3.13 CaloChallenge 2022

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 .

|                                |  |
|--------------------------------|--|
| <b>date:</b>                   | 2024-10-28   |
| <b>version:</b>                | v1.0   |
| <b>last_updated:</b>           | 2024-10  |
| <b>expired:</b>                | unknown  |
| <b>valid:</b>                  | yes  |
| <b>valid_date:</b>             | 2024-10-28   |
| <b>url:</b>                    | <a href="http://arxiv.org/abs/2410.21611">http://arxiv.org/abs/2410.21611</a>  |
| <b>doi:</b>                    | 10.48550/arXiv.2410.21611  |
| <b>domain:</b>                 | - High Energy Physics  |
| <b>focus:</b>                  | Fast generative-model-based calorimeter shower simulation evaluation   |
| <b>keywords:</b>               | - calorimeter simulation - generative models - surrogate modeling - LHC - fast simulation                              |
| <b>licensing:</b>              | Via Fermilab   |
| <b>task_types:</b>             | - Surrogate modeling   |
| <b>ai_capability_measured:</b> | - Simulation fidelity - speed - efficiency   |
| <b>metrics:</b>                | - Histogram similarity - Classifier AUC - Generation latency   |
| <b>models:</b>                 | - VAE variants - GAN variants - Normalizing flows - Diffusion models   |
| <b>ml_motif:</b>               | - Generative   |
| <b>type:</b>                   | Dataset  |
| <b>ml_task:</b>                | - Surrogate Modeling   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.  |
| <b>notes:</b>                  | The most comprehensive survey to date on ML-based calorimeter simulation; 31 submissions over different dataset sizes. |
| <b>contact.name:</b>           | Claudius Krause (CaloChallenge Lead)   |
| <b>contact.email:</b>          | unknown  |
| <b>datasets.links.name:</b>    | Four LHC calorimeter shower datasets   |
| <b>datasets.links.url:</b>     | various voxel resolutions  |
| <b>results.links.name:</b>     | ChatGPT LLM  |
| <b>fair.reproducible:</b>      | Yes  |
| <b>fair.benchmark_ready:</b>   | Yes  |
| <b>id:</b>                     | calochallenge_   |
| <b>Citations:</b>              | [8]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Four well-structured calorimeter datasets are provided, with different voxel resolutions, open access, signal/background separation, and metadata. FAIR principles are well covered.           |
| documentation      | 4     | Accompanied by a detailed paper and dataset description. Reproduction of pipelines may require additional setup or familiarity with the model submissions.                                     |
| metrics            | 5     | Metrics like histogram similarity, classifier AUC, and generation latency are well defined and relevant for simulation quality, fidelity, and performance.                                     |
| reference_solution | 4     | Several baselines (GANs, VAEs, flows, diffusion models) are documented and evaluated. Some are available via community repos, though not all are fully standardized or bundled.                |
| software           | 4     | Community GitHub repos and model implementations are available for the 31 submissions. While not fully unified in one place, the software is accessible and reproducible.                      |
| specification      | 5     | The task—evaluating fast generative calorimeter simulations—is clearly defined with benchmarking protocols, constraints like latency and model complexity, and structured evaluation criteria. |



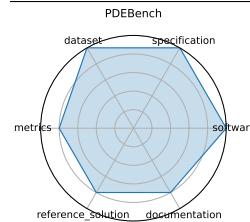
### 3.14 PDEBench

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 .

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2022-10-13  |
| <b>version:</b>                | v0.1.0  |
| <b>last_updated:</b>           | 2025-05   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2022-10-13  |
| <b>url:</b>                    | <a href="https://github.com/pdebench/PDEBench">https://github.com/pdebench/PDEBench</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2210.07182   |
| <b>domain:</b>                 | - Computational Science & AI - Climate & Earth Science - Mathematics                    |
| <b>focus:</b>                  | Benchmark suite for ML-based surrogates solving time-dependent PDEs                     |
| <b>keywords:</b>               | - PDEs - CFD - scientific ML - surrogate modeling - NeurIPS                             |
| <b>licensing:</b>              | Other   |
| <b>task_types:</b>             | - Supervised Learning   |
| <b>ai_capability_measured:</b> | - Time-dependent PDE modeling; physical accuracy  |
| <b>metrics:</b>                | - RMSE - boundary RMSE - Fourier RMSE   |
| <b>models:</b>                 | - FNO - U-Net - PINN - Gradient-Based inverse methods                                   |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Framework   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.                   |
| <b>notes:</b>                  | Datasets hosted on DaRUS (DOI:10.18419/darus-2986); contact maintainers by email        |
| <b>contact.name:</b>           | Makoto Takamoto (makoto.takamoto@neclab.eu)   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | pdebench  |
| <b>Citations:</b>              | [9]   |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Diverse PDE datasets (synthetic and real-world) hosted on DaRUS with DOIs. Datasets are well-documented, structured, and follow FAIR practices.   |
| documentation      | 4     | Strong documentation on GitHub including examples, configs, and usage instructions. Some model-specific details and tutorials could be further expanded.  |
| metrics            | 4     | Includes RMSE, boundary RMSE, and Fourier-domain RMSE. These are well-suited to PDE problems, though rationale behind metric choices could be expanded in some cases.   |
| reference_solution | 4     | Baselines (FNO, U-Net, PINN, etc.) are available and documented, but not every model includes full training and evaluation reproducibility out-of-the-box.  |
| software           | 5     | GitHub repository ( <a href="https://github.com/pdebench/PDEBench">https://github.com/pdebench/PDEBench</a> ) is actively maintained and includes training pipelines, data loaders, and evaluation scripts. Installation and usage are well-documented. |
| specification      | 5     | Clearly defined tasks for forward and inverse PDE problems, with structured input/output formats, system constraints, and task specifications.  |



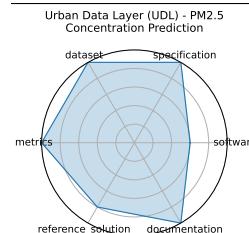
### 3.15 Urban Data Layer (UDL) - PM2.5 Concentration Prediction

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 .

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97837">https://neurips.cc/virtual/2024/poster/97837</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | Unified data pipeline for multi-modal urban science research  |
| <b>keywords:</b>               | - data pipeline - urban science - multi-modal - benchmark   |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Prediction - Classification   |
| <b>ai_capability_measured:</b> | - Multi-modal urban inference - standardization   |
| <b>metrics:</b>                | - Task-specific accuracy or RMSE  |
| <b>models:</b>                 | - Baseline regression/classification pipelines  |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Framework   |
| <b>ml_task:</b>                | - Prediction, classification  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Source code available on GitHub (SJTU-CILAB/udl); promotes reusable urban-science foundation models .   |
| <b>contact.name:</b>           | Yiheng Wang   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | urban_data_layer_udl_-_pm_concentration_prediction  |
| <b>Citations:</b>              | [10]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Large, multi-modal urban datasets are open-source, well-documented, and support reproducible research.   |
| documentation      | 5     | GitHub repository and conference poster provide comprehensive code and reproducibility instructions.   |
| metrics            | 5     | Uses task-specific accuracy and RMSE metrics appropriate for prediction and classification.  |
| reference_solution | 4     | Baseline models available but not exhaustive; community adoption and extensions expected.  |
| software           | 3     | Source code is publicly available on GitHub; baseline regression and classification pipelines are included but framework maturity is moderate. |
| specification      | 5     | Multiple urban science tasks like prediction and classification are well specified with clear input/output and evaluation criteria.            |



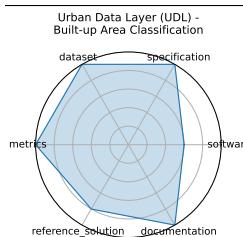
### 3.16 Urban Data Layer (UDL) - Built-up Area Classification

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 .

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97837">https://neurips.cc/virtual/2024/poster/97837</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | Unified data pipeline for multi-modal urban science research  |
| <b>keywords:</b>               | - data pipeline - urban science - multi-modal - benchmark   |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Prediction - Classification   |
| <b>ai_capability_measured:</b> | - Multi-modal urban inference - standardization   |
| <b>metrics:</b>                | - Task-specific accuracy or RMSE  |
| <b>models:</b>                 | - Baseline regression/classification pipelines  |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Framework   |
| <b>ml_task:</b>                | - Prediction, classification  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Source code available on GitHub (SJTU-CILAB/udl); promotes reusable urban-science foundation models .   |
| <b>contact.name:</b>           | Yiheng Wang   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | urban_data_layer_udl_-built-up_area_classification  |
| <b>Citations:</b>              | [10]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Large, multi-modal urban datasets are open-source, well-documented, and support reproducible research.   |
| documentation      | 5     | GitHub repository and conference poster provide comprehensive code and reproducibility instructions.   |
| metrics            | 5     | Uses task-specific accuracy and RMSE metrics appropriate for prediction and classification.  |
| reference_solution | 4     | Baseline models available but not exhaustive; community adoption and extensions expected.  |
| software           | 3     | Source code is publicly available on GitHub; baseline regression and classification pipelines are included but framework maturity is moderate. |
| specification      | 5     | Multiple urban science tasks like prediction and classification are well specified with clear input/output and evaluation criteria.            |



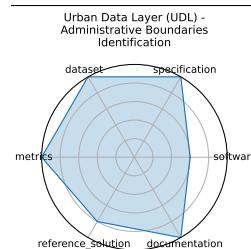
### 3.17 Urban Data Layer (UDL) - Administrative Boundaries Identification

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 .

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97837">https://neurips.cc/virtual/2024/poster/97837</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | Unified data pipeline for multi-modal urban science research  |
| <b>keywords:</b>               | - data pipeline - urban science - multi-modal - benchmark   |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Prediction - Classification   |
| <b>ai_capability_measured:</b> | - Multi-modal urban inference - standardization   |
| <b>metrics:</b>                | - Task-specific accuracy or RMSE  |
| <b>models:</b>                 | - Baseline regression/classification pipelines  |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Framework   |
| <b>ml_task:</b>                | - Prediction, classification  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Source code available on GitHub (SJTU-CILAB/udl); promotes reusable urban-science foundation models .   |
| <b>contact.name:</b>           | Yiheng Wang   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | urban_data_layer_udl_-_administrative_boundaries_identification   |
| <b>Citations:</b>              | [10]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Large, multi-modal urban datasets are open-source, well-documented, and support reproducible research.   |
| documentation      | 5     | GitHub repository and conference poster provide comprehensive code and reproducibility instructions.   |
| metrics            | 5     | Uses task-specific accuracy and RMSE metrics appropriate for prediction and classification.  |
| reference_solution | 4     | Baseline models available but not exhaustive; community adoption and extensions expected.  |
| software           | 3     | Source code is publicly available on GitHub; baseline regression and classification pipelines are included but framework maturity is moderate. |
| specification      | 5     | Multiple urban science tasks like prediction and classification are well specified with clear input/output and evaluation criteria.            |



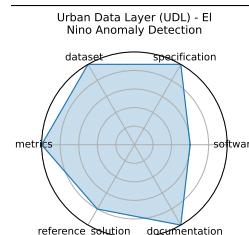
### 3.18 Urban Data Layer (UDL) - El Nino Anomaly Detection

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 .

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97837">https://neurips.cc/virtual/2024/poster/97837</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | Unified data pipeline for multi-modal urban science research  |
| <b>keywords:</b>               | - data pipeline - urban science - multi-modal - benchmark   |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Prediction - Classification   |
| <b>ai_capability_measured:</b> | - Multi-modal urban inference - standardization   |
| <b>metrics:</b>                | - Task-specific accuracy or RMSE  |
| <b>models:</b>                 | - Baseline regression/classification pipelines  |
| <b>ml_motif:</b>               | - Anomaly Detection   |
| <b>type:</b>                   | Framework   |
| <b>ml_task:</b>                | - Prediction, classification  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Source code available on GitHub (SJTU-CILAB/udl); promotes reusable urban-science foundation models .   |
| <b>contact.name:</b>           | Yiheng Wang   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | urban_data_layer_udl_-_el_nino_anomaly_detection  |
| <b>Citations:</b>              | [10]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Large, multi-modal urban datasets are open-source, well-documented, and support reproducible research.   |
| documentation      | 5     | GitHub repository and conference poster provide comprehensive code and reproducibility instructions.   |
| metrics            | 5     | Uses task-specific accuracy and RMSE metrics appropriate for prediction and classification.  |
| reference_solution | 4     | Baseline models available but not exhaustive; community adoption and extensions expected.  |
| software           | 3     | Source code is publicly available on GitHub; baseline regression and classification pipelines are included but framework maturity is moderate. |
| specification      | 5     | Multiple urban science tasks like prediction and classification are well specified with clear input/output and evaluation criteria.            |



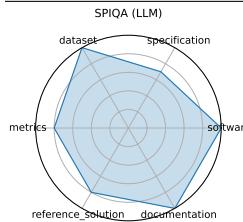
### 3.19 SPIQA (LLM)

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.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97575">https://neurips.cc/virtual/2024/poster/97575</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2407.09413   |
| <b>domain:</b>                 | - Computational Science & AI  |
| <b>focus:</b>                  | Evaluating LLMs on image-based scientific paper figure QA tasks (LLM Adapter performance)               |
| <b>keywords:</b>               | - multimodal QA - scientific figures - image+text - chain-of-thought prompting                          |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Multimodal QA   |
| <b>ai_capability_measured:</b> | - Visual reasoning - scientific figure understanding  |
| <b>metrics:</b>                | - Accuracy - F1 score   |
| <b>models:</b>                 | - LLaVA - MiniGPT-4 - Owl-LLM adapter variants  |
| <b>ml_motif:</b>               | - Multimodal Reasoning  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Multimodal QA   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.                                   |
| <b>notes:</b>                  | Companion to SPIQA main benchmark; compares adapter strategies using same images and QA pairs.          |
| <b>contact.name:</b>           | Xiaoyan Zhong   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | spiqa_llm   |
| <b>Citations:</b>              | [11]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Full dataset available on Hugging Face with train/test/valid splits.  |
| documentation      | 5     | Full paper available  |
| metrics            | 4     | Reports accuracy and F1; fair but no visual reasoning-specific metric.  |
| reference_solution | 4     | 10 LLM adapter baselines; results included without constraints.   |
| software           | 5     | Well-documented codebase available on Github  |
| specification      | 3.5   | Task of QA over scientific figures is sufficient but not fully formalized in input/output terms. No hardware constraints. |



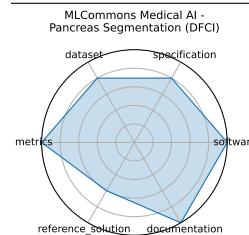
### 3.20 MLCommons Medical AI - Pancreas Segmentation (DFCI)

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 .

|                                |  |
|--------------------------------|--|
| <b>date:</b>                   | 2023-07-17   |
| <b>version:</b>                | v1.0   |
| <b>last_updated:</b>           | 2023-07  |
| <b>expired:</b>                | unknown  |
| <b>valid:</b>                  | yes  |
| <b>valid_date:</b>             | 2023-07-17   |
| <b>url:</b>                    | <a href="https://github.com/mlcommons/medical">https://github.com/mlcommons/medical</a>            |
| <b>doi:</b>                    | 10.1038/s42256-023-00652-2   |
| <b>domain:</b>                 | - Biology & Medicine   |
| <b>focus:</b>                  | Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data |
| <b>keywords:</b>               | - medical AI - federated evaluation - privacy-preserving - fairness - healthcare benchmarks        |
| <b>licensing:</b>              | Apache License 2.0   |
| <b>task_types:</b>             | - Federated evaluation - Model validation  |
| <b>ai_capability_measured:</b> | - Clinical accuracy - fairness - generalizability - privacy compliance                             |
| <b>metrics:</b>                | - ROC AUC - Accuracy - Fairness metrics  |
| <b>models:</b>                 | - MedPerf-validated CNNs - GaNDLF workflows  |
| <b>ml_motif:</b>               | - Classification   |
| <b>type:</b>                   | Platform   |
| <b>ml_task:</b>                | - NA   |
| <b>solutions:</b>              | 0  |
| <b>notes:</b>                  | Open-source platform under Apache-2.0; used across 20+ institutions and hospitals .                |
| <b>contact.name:</b>           | Alex Karargyris (MLCommons Medical AI)   |
| <b>contact.email:</b>          | unknown  |
| <b>datasets.links.name:</b>    | Multi-institutional clinical datasets, radiology   |
| <b>results.links.name:</b>     | ChatGPT LLM  |
| <b>fair.reproducible:</b>      | Yes  |
| <b>fair.benchmark_ready:</b>   | Yes  |
| <b>id:</b>                     | mlcommons_medical_ai_-_pancreas_segmentation_dfcf  |
| <b>Citations:</b>              | [12]   |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 4     | Multi-institutional datasets used in federated settings; real-world data is handled privately onsite, but some FAIR aspects (e.g., accessibility and metadata) are implicit.  |
| documentation      | 5     | Extensive documentation, papers, and community support exist. Clear examples and usage instructions are provided in GitHub and publications.  |
| metrics            | 5     | Metrics such as ROC AUC, accuracy, and fairness are clearly specified and directly support goals like generalizability and equity.  |
| reference_solution | 3     | GaNDLF workflows and MedPerf-validated CNNs are referenced, but not all baseline models are centrally documented or easily reproducible.  |
| software           | 5     | GitHub repository ( <a href="https://github.com/mlcommons/medical">https://github.com/mlcommons/medical</a> ) provides actively maintained open-source tools like MedPerf and GaNDLF for federated medical AI evaluation. |
| specification      | 4     | The platform defines federated tasks and model evaluation scenarios. Some clinical and system-level constraints are implied but not uniformly formalized across all use cases.  |



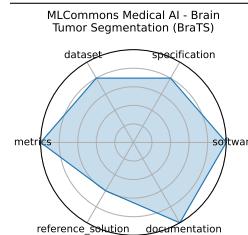
### 3.21 MLCommons Medical AI - Brain Tumor Segmentation (BraTS)

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 .

|                                |  |
|--------------------------------|--|
| <b>date:</b>                   | 2023-07-17   |
| <b>version:</b>                | v1.0   |
| <b>last_updated:</b>           | 2023-07  |
| <b>expired:</b>                | unknown  |
| <b>valid:</b>                  | yes  |
| <b>valid_date:</b>             | 2023-07-17   |
| <b>url:</b>                    | <a href="https://github.com/mlcommons/medical">https://github.com/mlcommons/medical</a>            |
| <b>doi:</b>                    | 10.1038/s42256-023-00652-2   |
| <b>domain:</b>                 | - Biology & Medicine   |
| <b>focus:</b>                  | Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data |
| <b>keywords:</b>               | - medical AI - federated evaluation - privacy-preserving - fairness - healthcare benchmarks        |
| <b>licensing:</b>              | Apache License 2.0   |
| <b>task_types:</b>             | - Federated evaluation - Model validation  |
| <b>ai_capability_measured:</b> | - Clinical accuracy - fairness - generalizability - privacy compliance                             |
| <b>metrics:</b>                | - ROC AUC - Accuracy - Fairness metrics  |
| <b>models:</b>                 | - MedPerf-validated CNNs - GaNDLF workflows  |
| <b>ml_motif:</b>               | - Classification   |
| <b>type:</b>                   | Platform   |
| <b>ml_task:</b>                | - NA   |
| <b>solutions:</b>              | 0  |
| <b>notes:</b>                  | Open-source platform under Apache-2.0; used across 20+ institutions and hospitals .                |
| <b>contact.name:</b>           | Alex Karargyris (MLCommons Medical AI)   |
| <b>contact.email:</b>          | unknown  |
| <b>datasets.links.name:</b>    | Multi-institutional clinical datasets, radiology   |
| <b>results.links.name:</b>     | ChatGPT LLM  |
| <b>fair.reproducible:</b>      | Yes  |
| <b>fair.benchmark_ready:</b>   | Yes  |
| <b>id:</b>                     | mlcommons_medical_ai_-_brain_tumor_segmentation_brats  |
| <b>Citations:</b>              | [12]   |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 4     | Multi-institutional datasets used in federated settings; real-world data is handled privately onsite, but some FAIR aspects (e.g., accessibility and metadata) are implicit.  |
| documentation      | 5     | Extensive documentation, papers, and community support exist. Clear examples and usage instructions are provided in GitHub and publications.  |
| metrics            | 5     | Metrics such as ROC AUC, accuracy, and fairness are clearly specified and directly support goals like generalizability and equity.  |
| reference_solution | 3     | GaNDLF workflows and MedPerf-validated CNNs are referenced, but not all baseline models are centrally documented or easily reproducible.  |
| software           | 5     | GitHub repository ( <a href="https://github.com/mlcommons/medical">https://github.com/mlcommons/medical</a> ) provides actively maintained open-source tools like MedPerf and GaNDLF for federated medical AI evaluation. |
| specification      | 4     | The platform defines federated tasks and model evaluation scenarios. Some clinical and system-level constraints are implied but not uniformly formalized across all use cases.  |



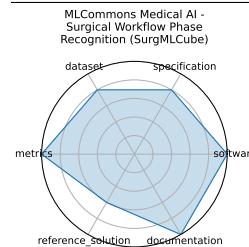
## 3.22 MLCommons Medical AI - Surgical Workflow Phase Recognition (SurgMLCube)

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 .

|                                |  |
|--------------------------------|--|
| <b>date:</b>                   | 2023-07-17   |
| <b>version:</b>                | v1.0   |
| <b>last_updated:</b>           | 2023-07  |
| <b>expired:</b>                | unknown  |
| <b>valid:</b>                  | yes  |
| <b>valid_date:</b>             | 2023-07-17   |
| <b>url:</b>                    | <a href="https://github.com/mlcommons/medical">https://github.com/mlcommons/medical</a>            |
| <b>doi:</b>                    | 10.1038/s42256-023-00652-2   |
| <b>domain:</b>                 | - Biology & Medicine   |
| <b>focus:</b>                  | Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data |
| <b>keywords:</b>               | - medical AI - federated evaluation - privacy-preserving - fairness - healthcare benchmarks        |
| <b>licensing:</b>              | Apache License 2.0   |
| <b>task_types:</b>             | - Federated evaluation - Model validation  |
| <b>ai_capability_measured:</b> | - Clinical accuracy - fairness - generalizability - privacy compliance                             |
| <b>metrics:</b>                | - ROC AUC - Accuracy - Fairness metrics  |
| <b>models:</b>                 | - MedPerf-validated CNNs - GaNDLF workflows  |
| <b>ml_motif:</b>               | - Classification   |
| <b>type:</b>                   | Platform   |
| <b>ml_task:</b>                | - NA   |
| <b>solutions:</b>              | 0  |
| <b>notes:</b>                  | Open-source platform under Apache-2.0; used across 20+ institutions and hospitals .                |
| <b>contact.name:</b>           | Alex Karargyris (MLCommons Medical AI)   |
| <b>contact.email:</b>          | unknown  |
| <b>datasets.links.name:</b>    | Multi-institutional clinical datasets, radiology   |
| <b>results.links.name:</b>     | ChatGPT LLM  |
| <b>fair.reproducible:</b>      | Yes  |
| <b>fair.benchmark_ready:</b>   | Yes  |
| <b>id:</b>                     | mlcommons_medical_ai_-__surgical_workflow_phase_recognition_surgmlcube                             |
| <b>Citations:</b>              | [12]   |

### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 4     | Multi-institutional datasets used in federated settings; real-world data is handled privately onsite, but some FAIR aspects (e.g., accessibility and metadata) are implicit.  |
| documentation      | 5     | Extensive documentation, papers, and community support exist. Clear examples and usage instructions are provided in GitHub and publications.  |
| metrics            | 5     | Metrics such as ROC AUC, accuracy, and fairness are clearly specified and directly support goals like generalizability and equity.  |
| reference_solution | 3     | GaNDLF workflows and MedPerf-validated CNNs are referenced, but not all baseline models are centrally documented or easily reproducible.  |
| software           | 5     | GitHub repository ( <a href="https://github.com/mlcommons/medical">https://github.com/mlcommons/medical</a> ) provides actively maintained open-source tools like MedPerf and GaNDLF for federated medical AI evaluation. |
| specification      | 4     | The platform defines federated tasks and model evaluation scenarios. Some clinical and system-level constraints are implied but not uniformly formalized across all use cases.  |



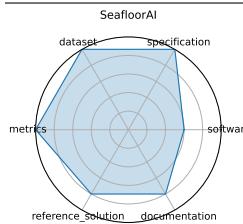
### 3.23 SeafloorAI

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

|                                |  |
|--------------------------------|--|
| <b>date:</b>                   | 2024-12-13   |
| <b>version:</b>                | v1.0   |
| <b>last_updated:</b>           | 2024-12  |
| <b>expired:</b>                | unknown  |
| <b>valid:</b>                  | yes  |
| <b>valid_date:</b>             | 2024-12-13   |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97432">https://neurips.cc/virtual/2024/poster/97432</a>  |
| <b>doi:</b>                    | 10.48550/arXiv.2411.00172  |
| <b>domain:</b>                 | - Climate & Earth Science  |
| <b>focus:</b>                  | Large-scale vision-language dataset for seafloor mapping and geological classification                   |
| <b>keywords:</b>               | - sonar imagery - vision-language - seafloor mapping - segmentation - QA                                 |
| <b>licensing:</b>              | unknown  |
| <b>task_types:</b>             | - Image segmentation - Vision-language QA  |
| <b>ai_capability_measured:</b> | - Geospatial understanding - multimodal reasoning  |
| <b>metrics:</b>                | - Segmentation pixel accuracy - QA accuracy  |
| <b>models:</b>                 | - SegFormer - ViLT-style multimodal models   |
| <b>ml_motif:</b>               | - Classification   |
| <b>type:</b>                   | Dataset  |
| <b>ml_task:</b>                | - Segmentation, QA   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.                                    |
| <b>notes:</b>                  | Data processing code publicly available, covering five geological layers; curated with marine scientists |
| <b>contact.name:</b>           | Kien X. Nguyen   |
| <b>contact.email:</b>          | unknown  |
| <b>datasets.links.name:</b>    | Sonar imagery + annotations  |
| <b>datasets.links.url:</b>     | unknown  |
| <b>results.links.name:</b>     | ChatGPT LLM  |
| <b>results.links.url:</b>      | unknown  |
| <b>fair.reproducible:</b>      | Yes  |
| <b>fair.benchmark_ready:</b>   | Yes  |
| <b>id:</b>                     | seafloorai   |
| <b>Citations:</b>              | [13]   |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Large-scale, well-annotated sonar imagery dataset with segmentation masks and natural language descriptions; curated with domain experts.  |
| documentation      | 4     | Dataset description and data processing instructions are provided, but tutorials and benchmark usage guides are limited.                   |
| metrics            | 5     | Standard segmentation pixel accuracy and QA accuracy metrics are clearly specified and appropriate for the tasks.                          |
| reference_solution | 4     | Some baseline models (e.g., SegFormer, ViLT-style) are mentioned, but reproducible code or pretrained weights are not fully available yet. |
| software           | 3     | Data processing code is publicly available, but no full benchmark framework or runnable model implementations are provided yet.            |
| specification      | 5     | Tasks (image segmentation and vision-language QA) are clearly defined with geospatial and multimodal objectives well specified.            |



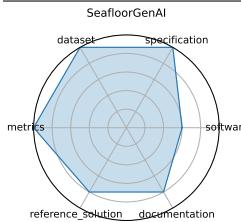
### 3.24 SeafloorGenAI

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

|                                |  |
|--------------------------------|--|
| <b>date:</b>                   | 2024-12-13   |
| <b>version:</b>                | v1.0   |
| <b>last_updated:</b>           | 2024-12  |
| <b>expired:</b>                | unknown  |
| <b>valid:</b>                  | yes  |
| <b>valid_date:</b>             | 2024-12-13   |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97432">https://neurips.cc/virtual/2024/poster/97432</a>  |
| <b>doi:</b>                    | 10.48550/arXiv.2411.00172  |
| <b>domain:</b>                 | - Climate & Earth Science  |
| <b>focus:</b>                  | Large-scale vision-language dataset for seafloor mapping and geological classification                   |
| <b>keywords:</b>               | - sonar imagery - vision-language - seafloor mapping - segmentation - QA                                 |
| <b>licensing:</b>              | unknown  |
| <b>task_types:</b>             | - Image segmentation - Vision-language QA  |
| <b>ai_capability_measured:</b> | - Geospatial understanding - multimodal reasoning  |
| <b>metrics:</b>                | - Segmentation pixel accuracy - QA accuracy  |
| <b>models:</b>                 | - SegFormer - ViLT-style multimodal models   |
| <b>ml_motif:</b>               | - Reasoning & Generalization   |
| <b>type:</b>                   | Dataset  |
| <b>ml_task:</b>                | - Segmentation, QA   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.                                    |
| <b>notes:</b>                  | Data processing code publicly available, covering five geological layers; curated with marine scientists |
| <b>contact.name:</b>           | Kien X. Nguyen   |
| <b>contact.email:</b>          | unknown  |
| <b>datasets.links.name:</b>    | Sonar imagery + annotations  |
| <b>datasets.links.url:</b>     | unknown  |
| <b>results.links.name:</b>     | ChatGPT LLM  |
| <b>results.links.url:</b>      | unknown  |
| <b>fair.reproducible:</b>      | Yes  |
| <b>fair.benchmark_ready:</b>   | Yes  |
| <b>id:</b>                     | seafloorgenai  |
| <b>Citations:</b>              | [13]   |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Large-scale, well-annotated sonar imagery dataset with segmentation masks and natural language descriptions; curated with domain experts.  |
| documentation      | 4     | Dataset description and data processing instructions are provided, but tutorials and benchmark usage guides are limited.                   |
| metrics            | 5     | Standard segmentation pixel accuracy and QA accuracy metrics are clearly specified and appropriate for the tasks.                          |
| reference_solution | 4     | Some baseline models (e.g., SegFormer, ViLT-style) are mentioned, but reproducible code or pretrained weights are not fully available yet. |
| software           | 3     | Data processing code is publicly available, but no full benchmark framework or runnable model implementations are provided yet.            |
| specification      | 5     | Tasks (image segmentation and vision-language QA) are clearly defined with geospatial and multimodal objectives well specified.            |



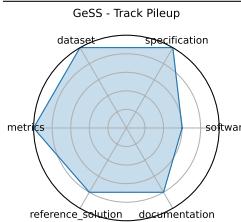
### 3.25 GeSS - Track Pileup

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 .

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97816">https://neurips.cc/virtual/2024/poster/97816</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - High Energy Physics   |
| <b>focus:</b>                  | Benchmark suite evaluating geometric deep learning models under real-world distribution shifts          |
| <b>keywords:</b>               | - geometric deep learning - distribution shift - OOD robustness - scientific applications               |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Classification  |
| <b>ai_capability_measured:</b> | - OOD performance in scientific settings  |
| <b>metrics:</b>                | - Accuracy - RMSE - OOD robustness delta  |
| <b>models:</b>                 | - GCN - EGNN - DimeNet++  |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Classification, Regression  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Includes no-OOD, unlabeled-OOD, and few-label scenarios .   |
| <b>contact.name:</b>           | Deyu Zou  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | gess - track_pileup   |
| <b>Citations:</b>              | [14]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Curated datasets of 3D crystal structures and material properties are included and publicly available for reproducible research.                                      |
| documentation      | 4     | Paper and poster provide solid explanation of benchmarks and scientific motivation; more extensive user documentation forthcoming.                                    |
| metrics            | 5     | Uses well-established metrics such as MAE and structural validity for materials modeling, plus accuracy and OOD robustness deltas.                                    |
| reference_solution | 4     | Two reference models (SODNet, DiffCSP-SC) are reported with results, code expected to be released soon.   |
| software           | 3     | Reference code expected post-conference; current public software availability limited. Benchmark infrastructure partially described but not fully released yet.       |
| specification      | 5     | Benchmark clearly defines OOD robustness scenarios with classification and regression tasks in scientific domains, though no explicit hardware constraints are given. |



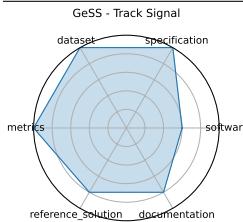
### 3.26 GeSS - Track Signal

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 .

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97816">https://neurips.cc/virtual/2024/poster/97816</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - High Energy Physics   |
| <b>focus:</b>                  | Benchmark suite evaluating geometric deep learning models under real-world distribution shifts          |
| <b>keywords:</b>               | - geometric deep learning - distribution shift - OOD robustness - scientific applications               |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Classification  |
| <b>ai_capability_measured:</b> | - OOD performance in scientific settings  |
| <b>metrics:</b>                | - Accuracy - RMSE - OOD robustness delta  |
| <b>models:</b>                 | - GCN - EGNN - DimeNet++  |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Classification, Regression  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Includes no-OOD, unlabeled-OOD, and few-label scenarios .   |
| <b>contact.name:</b>           | Deyu Zou  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | gess - track signal   |
| <b>Citations:</b>              | [14]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Curated datasets of 3D crystal structures and material properties are included and publicly available for reproducible research.                                      |
| documentation      | 4     | Paper and poster provide solid explanation of benchmarks and scientific motivation; more extensive user documentation forthcoming.                                    |
| metrics            | 5     | Uses well-established metrics such as MAE and structural validity for materials modeling, plus accuracy and OOD robustness deltas.                                    |
| reference_solution | 4     | Two reference models (SODNet, DiffCSP-SC) are reported with results, code expected to be released soon.   |
| software           | 3     | Reference code expected post-conference; current public software availability limited. Benchmark infrastructure partially described but not fully released yet.       |
| specification      | 5     | Benchmark clearly defines OOD robustness scenarios with classification and regression tasks in scientific domains, though no explicit hardware constraints are given. |



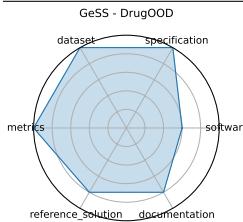
### 3.27 GeSS - DrugOOD

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 .

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97816">https://neurips.cc/virtual/2024/poster/97816</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Biology & Medicine  |
| <b>focus:</b>                  | Benchmark suite evaluating geometric deep learning models under real-world distribution shifts          |
| <b>keywords:</b>               | - geometric deep learning - distribution shift - OOD robustness - scientific applications               |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Classification  |
| <b>ai_capability_measured:</b> | - OOD performance in scientific settings  |
| <b>metrics:</b>                | - Accuracy - RMSE - OOD robustness delta  |
| <b>models:</b>                 | - GCN - EGNN - DimeNet++  |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Classification, Regression  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Includes no-OOD, unlabeled-OOD, and few-label scenarios .   |
| <b>contact.name:</b>           | Deyu Zou  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | gess_-_drugood  |
| <b>Citations:</b>              | [14]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Curated datasets of 3D crystal structures and material properties are included and publicly available for reproducible research.                                      |
| documentation      | 4     | Paper and poster provide solid explanation of benchmarks and scientific motivation; more extensive user documentation forthcoming.                                    |
| metrics            | 5     | Uses well-established metrics such as MAE and structural validity for materials modeling, plus accuracy and OOD robustness deltas.                                    |
| reference_solution | 4     | Two reference models (SODNet, DiffCSP-SC) are reported with results, code expected to be released soon.   |
| software           | 3     | Reference code expected post-conference; current public software availability limited. Benchmark infrastructure partially described but not fully released yet.       |
| specification      | 5     | Benchmark clearly defines OOD robustness scenarios with classification and regression tasks in scientific domains, though no explicit hardware constraints are given. |



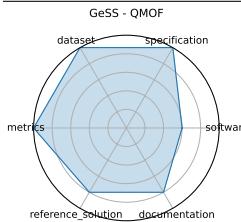
### 3.28 GeSS - QMOF

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 .

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97816">https://neurips.cc/virtual/2024/poster/97816</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Materials Science   |
| <b>focus:</b>                  | Benchmark suite evaluating geometric deep learning models under real-world distribution shifts          |
| <b>keywords:</b>               | - geometric deep learning - distribution shift - OOD robustness - scientific applications               |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Classification - Regression   |
| <b>ai_capability_measured:</b> | - OOD performance in scientific settings  |
| <b>metrics:</b>                | - Accuracy - RMSE - OOD robustness delta  |
| <b>models:</b>                 | - GCN - EGNN - DimeNet++  |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Classification, Regression  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Includes no-OOD, unlabeled-OOD, and few-label scenarios .   |
| <b>contact.name:</b>           | Deyu Zou  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | gess - _ qmof   |
| <b>Citations:</b>              | [14]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Curated datasets of 3D crystal structures and material properties are included and publicly available for reproducible research.                                      |
| documentation      | 4     | Paper and poster provide solid explanation of benchmarks and scientific motivation; more extensive user documentation forthcoming.                                    |
| metrics            | 5     | Uses well-established metrics such as MAE and structural validity for materials modeling, plus accuracy and OOD robustness deltas.                                    |
| reference_solution | 4     | Two reference models (SODNet, DiffCSP-SC) are reported with results, code expected to be released soon.   |
| software           | 3     | Reference code expected post-conference; current public software availability limited. Benchmark infrastructure partially described but not fully released yet.       |
| specification      | 5     | Benchmark clearly defines OOD robustness scenarios with classification and regression tasks in scientific domains, though no explicit hardware constraints are given. |



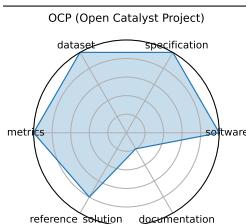
### 3.29 OCP (Open Catalyst Project)

The Open Catalyst Project (OC20 and OC22) provides DFT-calculated catalyst-adsorbate relaxation datasets, challenging ML models to predict energies and forces for renewable energy applications.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2020-10-20  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2020-10-20  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2020-10-20  |
| <b>url:</b>                    | <a href="https://opencatalystproject.org/">https://opencatalystproject.org/</a>   |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Chemistry - Materials Science   |
| <b>focus:</b>                  | Catalyst adsorption energy prediction   |
| <b>keywords:</b>               | - DFT relaxations - adsorption energy - graph neural networks   |
| <b>licensing:</b>              | OCP Terms of Use  |
| <b>task_types:</b>             | - Energy prediction - Force prediction  |
| <b>ai_capability_measured:</b> | - Prediction of adsorption energies and forces  |
| <b>metrics:</b>                | - MAE (energy) - MAE (force)  |
| <b>models:</b>                 | - CGCNN - SchNet - DimeNet++ - GemNet-OC  |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Public leaderboards; active community development   |
| <b>contact.name:</b>           | unknown   |
| <b>contact.email:</b>          | unknown   |
| <b>datasets.links.name:</b>    | OCP Dataset   |
| <b>datasets.links.url:</b>     | <a href="https://fair-chem.github.io/catalysts/datasets/summary">https://fair-chem.github.io/catalysts/datasets/summary</a> |
| <b>results.links.name:</b>     | OCP Pretrained Models   |
| <b>results.links.url:</b>      | <a href="https://fair-chem.github.io/catalysts/models.html">https://fair-chem.github.io/catalysts/models.html</a>           |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | ocp_open_catalyst_project   |
| <b>Citations:</b>              | [15], [16], [17], [18]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Fully FAIR- OC20, per-adsorbate trajectories, and OC22 are versioned; datasets come with standardized splits, metadata, and are downloadable.       |
| documentation      | 1     | Paper exists, but content is behind a paywall.  |
| metrics            | 5     | MAE (energy and force) are standard and reproducible.   |
| reference_solution | 4     | Multiple baselines (GemNet-OC, DimeNet++, etc.) implemented and evaluated. No hardware listed.  |
| software           | 5     | Data provided in Github links   |
| specification      | 5     | Tasks (energy and force prediction) are clearly defined with explicit I/O specifications, constraints, and physical relevance for renewable energy. |



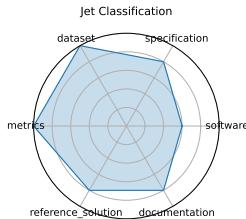
### 3.30 Jet Classification

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 and quantized models optimized for FPGA deployment.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-05-01  |
| <b>version:</b>                | v0.2.0  |
| <b>last_updated:</b>           | 2024-05   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-05-01  |
| <b>url:</b>                    | <a href="https://github.com/fastmachinelearning/fastml-science/tree/main/jet-classify">https://github.com/fastmachinelearning/fastml-science/tree/main/jet-classify</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2207.07958   |
| <b>domain:</b>                 | - High Energy Physics   |
| <b>focus:</b>                  | Real-time classification of particle jets using HL-LHC simulation features  |
| <b>keywords:</b>               | - classification - real-time ML - jet tagging - QKeras  |
| <b>licensing:</b>              | Apache License 2.0  |
| <b>task_types:</b>             | - Classification  |
| <b>ai_capability_measured:</b> | - Real-time inference - model compression performance   |
| <b>metrics:</b>                | - Accuracy - AUC  |
| <b>models:</b>                 | - Keras DNN - QKeras quantized DNN  |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.   |
| <b>notes:</b>                  | Includes both float and quantized models using QKeras   |
| <b>contact.name:</b>           | Jules Muhizi  |
| <b>contact.email:</b>          | unknown   |
| <b>datasets.links.name:</b>    | JetClass  |
| <b>datasets.links.url:</b>     | <a href="https://zenodo.org/record/6619768">https://zenodo.org/record/6619768</a>   |
| <b>results.links.name:</b>     | unknown   |
| <b>results.links.url:</b>      | unknown   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | jet_classification  |
| <b>Citations:</b>              | [19]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | None  |
| documentation      | 4     | Full reproducibility requires manual setup                                |
| metrics            | 5     | None  |
| reference_solution | 4     | HW/SW requirements missing; Reference not bundled as official starter kit |
| software           | 3     | Not containerized; Setup automation/documentation could be improved       |
| specification      | 4     | System constraints missing  |



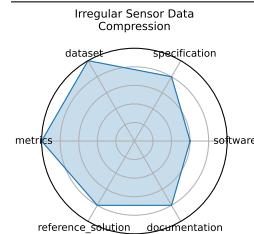
### 3.31 Irregular Sensor Data Compression

This benchmark addresses lossy compression of irregularly sampled sensor data from particle detectors using real-time autoencoder architectures, targeting latency-critical applications in physics experiments.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-05-01  |
| <b>version:</b>                | v0.2.0  |
| <b>last_updated:</b>           | 2024-05   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-05-01  |
| <b>url:</b>                    | <a href="https://github.com/fastmachinelearning/fastml-science/tree/main/sensor-data-compression">https://github.com/fastmachinelearning/fastml-science/tree/main/sensor-data-compression</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2207.07958   |
| <b>domain:</b>                 | - High Energy Physics   |
| <b>focus:</b>                  | Real-time compression of sparse sensor data with autoencoders   |
| <b>keywords:</b>               | - compression - autoencoder - sparse data - irregular sampling  |
| <b>licensing:</b>              | Apache License 2.0  |
| <b>task_types:</b>             | - Compression   |
| <b>ai_capability_measured:</b> | - Reconstruction quality - compression efficiency   |
| <b>metrics:</b>                | - MSE - Compression ratio   |
| <b>models:</b>                 | - Autoencoder - Quantized autoencoder   |
| <b>ml_motif:</b>               | - Generative  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Unsupervised Learning   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.   |
| <b>notes:</b>                  | Based on synthetic but realistic physics sensor data  |
| <b>contact.name:</b>           | Ben Hawks, Nhan Tran  |
| <b>contact.email:</b>          | unknown   |
| <b>datasets.links.name:</b>    | Custom synthetic irregular sensor dataset   |
| <b>datasets.links.url:</b>     | <a href="https://github.com/fastmachinelearning/fastml-science/tree/main/sensor-data-compression">https://github.com/fastmachinelearning/fastml-science/tree/main/sensor-data-compression</a> |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | irregular_sensor_data_compression   |
| <b>Citations:</b>              | [20]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | All criteria met   |
| documentation      | 4     | Setup for deployment (e.g., FPGA pipeline) requires familiarity with tooling |
| metrics            | 5     | All criteria met   |
| reference_solution | 4     | Not fully documented or automated for reproducibility                        |
| software           | 3     | Not containerized; Full automation and documentation could be improved       |
| specification      | 4     | Exact latency or resource constraints not numerically specified              |



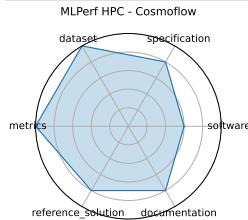
### 3.32 MLPerf HPC - Cosmoflow

MLPerf HPC introduces scientific model benchmarks (e.g., CosmoFlow, DeepCAM) aimed at large-scale HPC evaluation with >10x performance scaling through system-level optimizations.

**date:** 2021-10-20  
**version:** v1.0  
**last\_updated:** 2021-10  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2021-10-20  
**url:** <https://github.com/mlcommons/hpc>  
**doi:** 10.48550/arXiv.2110.11466  
**domain:** - High Energy Physics  
**focus:** Scientific ML training and inference on HPC systems  
**keywords:** - HPC - training - inference - scientific ML  
**licensing:** Apache License 2.0  
**task\_types:** - Training - Inference  
**ai\_capability\_measured:** - Scaling efficiency - training time - model accuracy on HPC  
**metrics:** - Training time - Accuracy - GPU utilization  
**models:** - CosmoFlow - DeepCAM - OpenCatalyst  
**ml\_motif:** - Regression  
**type:** Framework  
**ml\_task:** - NA  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Shared framework with MLCommons Science; reference implementations included.  
**contact.name:** Steven Farrell (MLCommons)  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** mlperf\_hpc\_-\_cosmoflow  
**Citations:** [21]

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Not all data is independently versioned or comes with standardized FAIR metadata.                                    |
| documentation      | 4     | Central guidance is available but requires domain-specific effort to replicate results across systems.               |
| metrics            | 5     | None   |
| reference_solution | 4     | Reproducibility and environment tuning depend on system configuration; baseline models not uniformly bundled.        |
| software           | 3     | Reference implementations exist but containerization and environment setup require manual effort across HPC systems. |
| specification      | 4     | Hardware constraints and I/O formats are not fully defined for all scenarios.  |



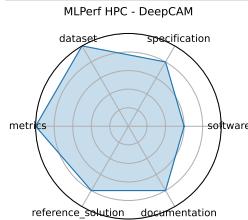
### 3.33 MLPerf HPC - DeepCAM

MLPerf HPC introduces scientific model benchmarks (e.g., CosmoFlow, DeepCAM) aimed at large-scale HPC evaluation with >10x performance scaling through system-level optimizations.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2021-10-20  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2021-10   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2021-10-20  |
| <b>url:</b>                    | <a href="https://github.com/mlcommons/hpc">https://github.com/mlcommons/hpc</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2110.11466   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | Scientific ML training and inference on HPC systems                             |
| <b>keywords:</b>               | - HPC - training - inference - scientific ML                                    |
| <b>licensing:</b>              | Apache License 2.0  |
| <b>task_types:</b>             | - Training - Inference  |
| <b>ai_capability_measured:</b> | - Scaling efficiency - training time - model accuracy on HPC                    |
| <b>metrics:</b>                | - Training time - Accuracy - GPU utilization                                    |
| <b>models:</b>                 | - DeepCAM   |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Framework   |
| <b>ml_task:</b>                | - NA  |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.           |
| <b>notes:</b>                  | Shared framework with MLCommons Science; reference implementations included.    |
| <b>contact.name:</b>           | Steven Farrell (MLCommons)  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | mlperf_hpc_-_deepcam  |
| <b>Citations:</b>              | [21]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Not all data is independently versioned or comes with standardized FAIR metadata.                                    |
| documentation      | 4     | Central guidance is available but requires domain-specific effort to replicate results across systems.               |
| metrics            | 5     | None   |
| reference_solution | 4     | Reproducibility and environment tuning depend on system configuration; baseline models not uniformly bundled.        |
| software           | 3     | Reference implementations exist but containerization and environment setup require manual effort across HPC systems. |
| specification      | 4     | Hardware constraints and I/O formats are not fully defined for all scenarios.  |



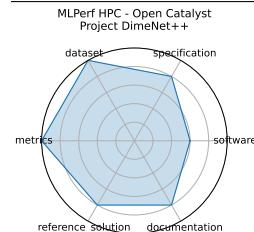
### 3.34 MLPerf HPC - Open Catalyst Project DimeNet++

MLPerf HPC introduces scientific model benchmarks (e.g., CosmoFlow, DeepCAM) aimed at large-scale HPC evaluation with >10x performance scaling through system-level optimizations.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2021-10-20  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2021-10   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2021-10-20  |
| <b>url:</b>                    | <a href="https://github.com/mlcommons/hpc">https://github.com/mlcommons/hpc</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2110.11466   |
| <b>domain:</b>                 | - Chemistry   |
| <b>focus:</b>                  | Scientific ML training and inference on HPC systems                             |
| <b>keywords:</b>               | - HPC - training - inference - scientific ML                                    |
| <b>licensing:</b>              | Apache License 2.0  |
| <b>task_types:</b>             | - Training - Inference  |
| <b>ai_capability_measured:</b> | - Scaling efficiency - training time - model accuracy on HPC                    |
| <b>metrics:</b>                | - Training time - Accuracy - GPU utilization                                    |
| <b>models:</b>                 | - DeepCAM   |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Framework   |
| <b>ml_task:</b>                | - NA  |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.           |
| <b>notes:</b>                  | Shared framework with MLCommons Science; reference implementations included.    |
| <b>contact.name:</b>           | Steven Farrell (MLCommons)  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | mlperf_hpc--open_catalyst_project_dimenet_                                      |
| <b>Citations:</b>              | [21]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Not all data is independently versioned or comes with standardized FAIR metadata.                                    |
| documentation      | 4     | Central guidance is available but requires domain-specific effort to replicate results across systems.               |
| metrics            | 5     | None   |
| reference_solution | 4     | Reproducibility and environment tuning depend on system configuration; baseline models not uniformly bundled.        |
| software           | 3     | Reference implementations exist but containerization and environment setup require manual effort across HPC systems. |
| specification      | 4     | Hardware constraints and I/O formats are not fully defined for all scenarios.  |



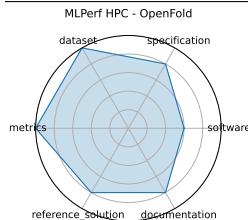
### 3.35 MLPerf HPC - OpenFold

MLPerf HPC introduces scientific model benchmarks (e.g., CosmoFlow, DeepCAM) aimed at large-scale HPC evaluation with >10x performance scaling through system-level optimizations.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2021-10-20  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2021-10   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2021-10-20  |
| <b>url:</b>                    | <a href="https://github.com/mlcommons/hpc">https://github.com/mlcommons/hpc</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2110.11466   |
| <b>domain:</b>                 | - Biology & Medicine  |
| <b>focus:</b>                  | Scientific ML training and inference on HPC systems                             |
| <b>keywords:</b>               | - HPC - training - inference - scientific ML                                    |
| <b>licensing:</b>              | Apache License 2.0  |
| <b>task_types:</b>             | - Training - Inference  |
| <b>ai_capability_measured:</b> | - Scaling efficiency - training time - model accuracy on HPC                    |
| <b>metrics:</b>                | - Training time - Accuracy - GPU utilization                                    |
| <b>models:</b>                 | - DeepCAM   |
| <b>ml_motif:</b>               | - Sequence Prediction/Forecasting   |
| <b>type:</b>                   | Framework   |
| <b>ml_task:</b>                | - NA  |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.           |
| <b>notes:</b>                  | Shared framework with MLCommons Science; reference implementations included.    |
| <b>contact.name:</b>           | Steven Farrell (MLCommons)  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | mlperf_hpc_-_openfold   |
| <b>Citations:</b>              | [21]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Not all data is independently versioned or comes with standardized FAIR metadata.                                    |
| documentation      | 4     | Central guidance is available but requires domain-specific effort to replicate results across systems.               |
| metrics            | 5     | None   |
| reference_solution | 4     | Reproducibility and environment tuning depend on system configuration; baseline models not uniformly bundled.        |
| software           | 3     | Reference implementations exist but containerization and environment setup require manual effort across HPC systems. |
| specification      | 4     | Hardware constraints and I/O formats are not fully defined for all scenarios.  |



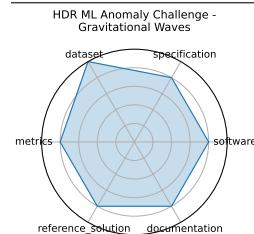
### 3.36 HDR ML Anomaly Challenge - Gravitational Waves

A benchmark for detecting anomalous transient gravitational-wave signals, including "unknown-unknowns," using preprocessed LIGO time-series at 4096 Hz. Competitors submit inference models on Codabench for continuous 50 ms segments from dual interferometers.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2025-03-03  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2025-03   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2025-03-03  |
| <b>url:</b>                    | <a href="https://www.codabench.org/competitions/2626/">https://www.codabench.org/competitions/2626/</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2503.02112   |
| <b>domain:</b>                 | - High Energy Physics   |
| <b>focus:</b>                  | Detecting anomalous gravitational-wave signals from LIGO/Virgo datasets                                 |
| <b>keywords:</b>               | - anomaly detection - gravitational waves - astrophysics - time-series                                  |
| <b>licensing:</b>              | NA  |
| <b>task_types:</b>             | - Anomaly Detection   |
| <b>ai_capability_measured:</b> | - Novel event detection in physical signals   |
| <b>metrics:</b>                | - ROC-AUC - Precision/Recall  |
| <b>models:</b>                 | - Deep latent CNNs - Autoencoders   |
| <b>ml_motif:</b>               | - Anomaly Detection   |
| <b>type:</b>                   | Dataset   |
| <b>ml_task:</b>                | - Anomaly Detection   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.                                   |
| <b>notes:</b>                  | NSF HDR A3D3 sponsored; prize pool and starter kit provided on Codabench.                               |
| <b>contact.name:</b>           | HDR A3D3 Team   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | hdr_ml_anomaly_challenge--gravitational_waves   |
| <b>Citations:</b>              | [22]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Uses preprocessed LIGO/Virgo time series data at 4096 Hz, publicly available and standard in astrophysics.  |
| documentation      | 4     | Documentation includes challenge instructions, starter kit details, and baseline descriptions, but could benefit from more thorough tutorials and code walkthroughs.                            |
| metrics            | 4     | ROC-AUC, precision, and recall metrics are clearly specified and appropriate for anomaly detection.   |
| reference_solution | 4     | Baseline deep latent CNNs and autoencoders are provided and reproducible, but not extensively documented.   |
| software           | 4     | Benchmark platform provided on Codabench with starter kits and submission infrastructure. Code and baseline models are publicly accessible but not extensively maintained beyond the challenge. |
| specification      | 4     | Well-defined anomaly detection task on gravitational-wave time series with clear input/output expectations and challenge constraints.   |



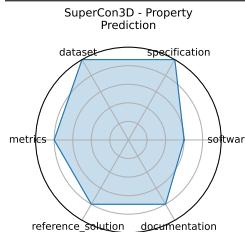
### 3.37 SuperCon3D - Property Prediction

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 .

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97553">https://neurips.cc/virtual/2024/poster/97553</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Materials Science   |
| <b>focus:</b>                  | Dataset and models for predicting and generating high-Tc superconductors using 3D crystal structures    |
| <b>keywords:</b>               | - superconductivity - crystal structures - equivariant GNN - generative models                          |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Regression (Tc prediction) - Generative modeling  |
| <b>ai_capability_measured:</b> | - Structure-to-property prediction - structure generation   |
| <b>metrics:</b>                | - MAE (Tc) - Validity of generated structures   |
| <b>models:</b>                 | - SODNet - DiffCSP-SC   |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Dataset + Models  |
| <b>ml_task:</b>                | - Regression, Generation  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Demonstrates advantage of combining ordered and disordered structural data in model design              |
| <b>contact.name:</b>           | Zhong Zuo   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | supercond_-_property_prediction   |
| <b>Citations:</b>              | [23]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Dataset contains 3D crystal structures and associated properties; well-curated but not fully released publicly at this time.                            |
| documentation      | 4     | Paper and GitHub provide good metadata and data processing descriptions; tutorials and user guides could be expanded.                                   |
| metrics            | 4     | Metrics such as MAE for Tc prediction and validity checks for generated structures are appropriate and clearly described.                               |
| reference_solution | 4     | Paper provides model architecture details and some training insights, but no complete open-source reference implementations yet.                        |
| software           | 3     | Baseline models (SODNet, DiffCSP-SC) are described in the paper; however, fully reproducible code and pretrained models are not publicly available yet. |
| specification      | 5     | Tasks for regression (Tc prediction) and generative modeling with clear input/output structures and domain constraints are well defined.                |



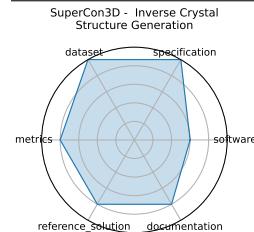
### 3.38 SuperCon3D - Inverse Crystal Structure Generation

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 .

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97553">https://neurips.cc/virtual/2024/poster/97553</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Materials Science   |
| <b>focus:</b>                  | Dataset and models for predicting and generating high-Tc superconductors using 3D crystal structures    |
| <b>keywords:</b>               | - superconductivity - crystal structures - equivariant GNN - generative models                          |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Regression (Tc prediction) - Generative modeling  |
| <b>ai_capability_measured:</b> | - Structure-to-property prediction - structure generation   |
| <b>metrics:</b>                | - MAE (Tc) - Validity of generated structures   |
| <b>models:</b>                 | - SODNet - DiffCSP-SC   |
| <b>ml_motif:</b>               | - Generative  |
| <b>type:</b>                   | Dataset + Models  |
| <b>ml_task:</b>                | - Regression, Generation  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Demonstrates advantage of combining ordered and disordered structural data in model design              |
| <b>contact.name:</b>           | Zhong Zuo   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | supercond_--inverse_crystal_structure_generation  |
| <b>Citations:</b>              | [23]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Dataset contains 3D crystal structures and associated properties; well-curated but not fully released publicly at this time.                            |
| documentation      | 4     | Paper and GitHub provide good metadata and data processing descriptions; tutorials and user guides could be expanded.                                   |
| metrics            | 4     | Metrics such as MAE for Tc prediction and validity checks for generated structures are appropriate and clearly described.                               |
| reference_solution | 4     | Paper provides model architecture details and some training insights, but no complete open-source reference implementations yet.                        |
| software           | 3     | Baseline models (SODNet, DiffCSP-SC) are described in the paper; however, fully reproducible code and pretrained models are not publicly available yet. |
| specification      | 5     | Tasks for regression (Tc prediction) and generative modeling with clear input/output structures and domain constraints are well defined.                |



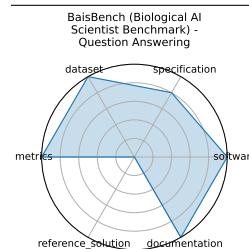
### 3.39 BaisBench (Biological AI Scientist Benchmark) - Question Answering

BaisBench evaluates AI scientists' ability to perform data-driven biological research by annotating cell types in single-cell datasets and answering MCQs derived from biological study insights, measuring autonomous scientific discovery.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2025-05-13  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2025-05-13  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2025-05-13  |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2505.08341">https://arxiv.org/abs/2505.08341</a>         |
| <b>doi:</b>                    | 10.48550/arXiv.2505.08341   |
| <b>domain:</b>                 | - Biology & Medicine  |
| <b>focus:</b>                  | Omics-driven AI research tasks  |
| <b>keywords:</b>               | - single-cell annotation - biological QA - autonomous discovery                         |
| <b>licensing:</b>              | MIT License   |
| <b>task_types:</b>             | - Cell type annotation - Multiple choice  |
| <b>ai_capability_measured:</b> | - Autonomous biological research capabilities   |
| <b>metrics:</b>                | - Annotation accuracy - QA accuracy   |
| <b>models:</b>                 | - LLM-based AI scientist agents   |
| <b>ml_motif:</b>               | - Reasoning & Generalization  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Underperforms human experts; aims to advance AI-driven discovery                        |
| <b>contact.name:</b>           | Xuegong Zhang   |
| <b>contact.email:</b>          | <a href="mailto:zhangxg@mail.tsinghua.edu.cn">zhangxg@mail.tsinghua.edu.cn</a>          |
| <b>datasets.links.name:</b>    | Github  |
| <b>datasets.links.url:</b>     | <a href="https://github.com/EperLuo/BaisBench">https://github.com/EperLuo/BaisBench</a> |
| <b>results.links.name:</b>     | unknown   |
| <b>results.links.url:</b>      | unknown   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | baisbench_biological_ai_scientist_benchmark_question_answering                          |
| <b>Citations:</b>              | [24]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Uses public scRNA-seq datasets linked in paper appendix; structured and accessible, though versioning and full metadata not formalized per FAIR standards. |
| documentation      | 5     | Dataset and paper accessible; IPYNB files for setup are available on the github repo.  |
| metrics            | 5     | Includes precise and interpretable metrics (annotation and QA accuracy); directly aligned with task outputs and benchmarking goals.                        |
| reference_solution | 0     | Model evaluations and LLM agent results discussed; however, no fully packaged, runnable baseline confirmed yet.  |
| software           | 5     | Instructions for environment setup available   |
| specification      | 4     | Task clearly defined-cell type annotation and biological QA; input/output formats are well-described; system constraints are not quantified.               |



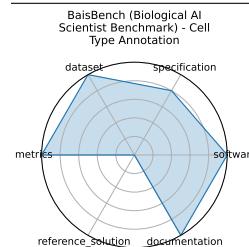
### 3.40 BaisBench (Biological AI Scientist Benchmark) - Cell Type Annotation

BaisBench evaluates AI scientists' ability to perform data-driven biological research by annotating cell types in single-cell datasets and answering MCQs derived from biological study insights, measuring autonomous scientific discovery.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2025-05-13  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2025-05-13  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2025-05-13  |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2505.08341">https://arxiv.org/abs/2505.08341</a>         |
| <b>doi:</b>                    | 10.48550/arXiv.2505.08341   |
| <b>domain:</b>                 | - Biology & Medicine  |
| <b>focus:</b>                  | Omics-driven AI research tasks  |
| <b>keywords:</b>               | - single-cell annotation - biological QA - autonomous discovery                         |
| <b>licensing:</b>              | MIT License   |
| <b>task_types:</b>             | - Cell type annotation - Multiple choice  |
| <b>ai_capability_measured:</b> | - Autonomous biological research capabilities   |
| <b>metrics:</b>                | - Annotation accuracy - QA accuracy   |
| <b>models:</b>                 | - LLM-based AI scientist agents   |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Underperforms human experts; aims to advance AI-driven discovery                        |
| <b>contact.name:</b>           | Xuegong Zhang   |
| <b>contact.email:</b>          | <a href="mailto:zhangxg@mail.tsinghua.edu.cn">zhangxg@mail.tsinghua.edu.cn</a>          |
| <b>datasets.links.name:</b>    | Github  |
| <b>datasets.links.url:</b>     | <a href="https://github.com/EperLuo/BaisBench">https://github.com/EperLuo/BaisBench</a> |
| <b>results.links.name:</b>     | unknown   |
| <b>results.links.url:</b>      | unknown   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | baisbench_biological_ai_scientist_benchmark_cell_type_annotation                        |
| <b>Citations:</b>              | [24]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Uses public scRNA-seq datasets linked in paper appendix; structured and accessible, though versioning and full metadata not formalized per FAIR standards. |
| documentation      | 5     | Dataset and paper accessible; IPYNB files for setup are available on the github repo.  |
| metrics            | 5     | Includes precise and interpretable metrics (annotation and QA accuracy); directly aligned with task outputs and benchmarking goals.                        |
| reference_solution | 0     | Model evaluations and LLM agent results discussed; however, no fully packaged, runnable baseline confirmed yet.  |
| software           | 5     | Instructions for environment setup available   |
| specification      | 4     | Task clearly defined-cell type annotation and biological QA; input/output formats are well-described; system constraints are not quantified.               |



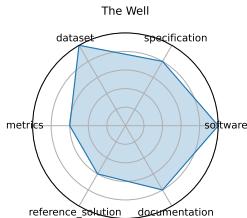
### 3.41 The Well

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.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-03  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2025-06   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-03  |
| <b>url:</b>                    | <a href="https://polymathic-ai.org/the_well/">https://polymathic-ai.org/the_well/</a>   |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Biology & Medicine - Computational Science & AI - High Energy Physics   |
| <b>focus:</b>                  | Foundation model + surrogate dataset spanning 16 physical simulation domains  |
| <b>keywords:</b>               | - surrogate modeling - foundation model - physics simulations - spatiotemporal dynamics   |
| <b>licensing:</b>              | BSD 3-Clause License  |
| <b>task_types:</b>             | - Supervised Learning   |
| <b>ai_capability_measured:</b> | - Surrogate modeling - physics-based prediction   |
| <b>metrics:</b>                | - Dataset size - Domain breadth   |
| <b>models:</b>                 | - FNO baselines - U-Net baselines   |
| <b>ml_motif:</b>               | - Sequence Prediction/Forecasting   |
| <b>type:</b>                   | Dataset   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | 1   |
| <b>notes:</b>                  | Includes unified API and dataset metadata; see 2025 NeurIPS paper for full benchmark details.<br>Size: 15 TB. "Benchmarks" are nominally baseline models that were trained on different parts of the dataset, all of them being time series prediction tasks. |
| <b>contact.name:</b>           | Ruben Ohana   |
| <b>contact.email:</b>          | <a href="mailto:rohana@flatironinstitute.org">rohana@flatironinstitute.org</a>  |
| <b>datasets.links.name:</b>    | 16 simulation datasets  |
| <b>datasets.links.url:</b>     | HDF5) via PyPI/GitHub   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | the_well  |
| <b>Citations:</b>              | [25]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | 15 TB of ML-ready HDF5 datasets across 16 physics domains. Public, well-structured, richly annotated, and designed with FAIR principles in mind.                            |
| documentation      | 4     | The GitHub repo and NeurIPS paper provide detailed guidance on dataset use, structure, and training setup. Tutorials and walkthroughs could be expanded further.            |
| metrics            | 3     | Domain breadth and dataset size are emphasized. Standardized quantitative metrics for model evaluation (e.g., RMSE, accuracy) are not uniformly applied across all domains. |
| reference_solution | 3     | Includes FNO and U-Net baselines, but does not yet provide fully trained, reproducible models or scripts across all datasets.   |
| software           | 5     | BSD-licensed software and unified API are available via GitHub and PyPI. Supports loading and manipulating large HDF5 datasets across 16 domains.                           |
| specification      | 4     | The benchmark includes clearly defined surrogate modeling tasks, data structure, and metadata. However, constraints and formal task specs vary slightly across domains.     |



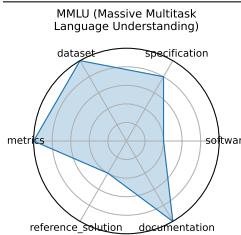
### 3.42 MMLU (Massive Multitask Language Understanding)

Measuring Massive Multitask Language Understanding (MMLU) is a benchmark of 57 multiple-choice tasks covering elementary mathematics, US history, computer science, law, and more, designed to evaluate a model's breadth and depth of knowledge in zero-shot and few-shot settings.

**date:** 2020-09-07  
**version:** 1  
**last\_updated:** 2020-09-07  
**expired:** false  
**valid:** yes  
**valid\_date:** 2025-07-28  
**url:** <https://huggingface.co/datasets/cais/mmlu>  
**doi:** 10.48550/arXiv.2009.03300  
**domain:** - Computational Science & AI  
**focus:** Academic knowledge and reasoning across 57 subjects  
**keywords:** - multitask - multiple-choice - zero-shot - few-shot - knowledge probing  
**licensing:** MIT License  
**task\_types:** - Multiple choice  
**ai\_capability\_measured:** - General reasoning, subject-matter understanding  
**metrics:** - Accuracy  
**models:** - GPT-4o - Gemini 1.5 Pro - o1 - DeepSeek-R1  
**ml\_motif:** - Reasoning & Generalization  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 1  
**notes:** Good  
**contact.name:** Dan Hendrycks  
**contact.email:** dan (at) safe.ai  
**datasets.links.name:** Huggingface Dataset  
**datasets.links.url:** <https://huggingface.co/datasets/cais/mmlu>  
**results.links.name:** Measuring Massive Multitask Language Understanding - Test Leaderboard  
**results.links.url:** <https://github.com/hendrycks/test?tab=readme-ov-file#test-leaderboard>  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** mmlu\_massive\_multitask\_language\_understanding  
**Citations:** [26]

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Meets all FAIR principles and properly versioned.   |
| documentation      | 5     | Well-explained in a provided paper.   |
| metrics            | 5     | Fully defined, represents a solution's performance.   |
| reference_solution | 2     | Reference models are available (i.e. GPT-3), but are not trainable or publicly documented     |
| software           | 2     | Some code is available on github to reproduce results via OpenAI API, but not well documented |
| specification      | 4     | No system constraints   |



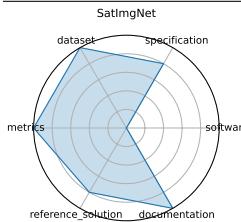
### 3.43 SatImgNet

SATIN (sometimes referred to as SatImgNet) is a multi-task metadataset of 27 satellite imagery classification datasets evaluating zero-shot transfer of vision-language models across diverse remote sensing tasks.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2023-04-23  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2023-04-23  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2023-04-23  |
| <b>url:</b>                    | <a href="https://satinbenchmark.github.io/">https://satinbenchmark.github.io/</a>   |
| <b>doi:</b>                    | 10.48550/arXiv.2304.11619   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | Satellite imagery classification  |
| <b>keywords:</b>               | - land-use - zero-shot - multi-task   |
| <b>licensing:</b>              | CC-BY-4.0   |
| <b>task_types:</b>             | - Image classification  |
| <b>ai_capability_measured:</b> | - Zero-shot land-use classification   |
| <b>metrics:</b>                | - Accuracy  |
| <b>models:</b>                 | - CLIP - BLIP - ALBEF   |
| <b>ml_motif:</b>               | - Multimodal Reasoning  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | Numerous, evaluated via leaderboard   |
| <b>notes:</b>                  | Public leaderboard available  |
| <b>contact.name:</b>           | Jonathan Roberts  |
| <b>contact.email:</b>          | j.roberts@cs.ox.ac.uk   |
| <b>datasets.links.name:</b>    | SatImgNet on Hugging Face   |
| <b>datasets.links.url:</b>     | <a href="https://huggingface.co/datasets/jonathan-roberts1/SATIN">https://huggingface.co/datasets/jonathan-roberts1/SATIN</a> |
| <b>results.links.name:</b>     | SatImgNet Leaderboard   |
| <b>results.links.url:</b>      | <a href="https://satinbenchmark.github.io/_pages/leaderboard/">https://satinbenchmark.github.io/_pages/leaderboard/</a>       |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | satimgnet   |
| <b>Citations:</b>              | [27]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Hosted on Hugging Face, versioned, FAIR-compliant with rich metadata; covers many well-known remote sensing datasets unified under one metadataset, though documentation depth varies slightly across tasks.          |
| documentation      | 5     | Paper provides all required information   |
| metrics            | 5     | Accuracy of classification is an appropriate metric   |
| reference_solution | 4     | Baselines like CLIP, BLIP, ALBEF evaluated in the paper; no constraints specified   |
| software           | 0     | No scripts or environment information provided  |
| specification      | 4     | Tasks (image classification across 27 satellite datasets) are clearly defined with multi-task and zero-shot framing; input/output structure is mostly standard but some task-specific nuances require interpretation. |



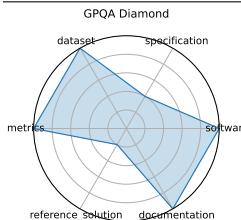
### 3.44 GPQA Diamond

GPQA is a dataset of 448 challenging, multiple-choice questions in biology, physics, and chemistry, written by domain experts. It is Google-proof - experts score 65% (74% after error correction) while skilled non-experts with web access score only 34%. State-of-the-art LLMs like GPT-4 reach around 39% accuracy.

**date:** 2023-11-20  
**version:** 1  
**last\_updated:** 2023-11-20  
**expired:** false  
**valid:** yes  
**valid\_date:** 2023-11-20  
**url:** <https://arxiv.org/abs/2311.12022>  
**doi:** 10.48550/arXiv.2311.12022  
**domain:** - Biology & Medicine - Chemistry - High Energy Physics  
**focus:** Graduate-level scientific reasoning  
**keywords:** - Google-proof - graduate-level - science QA - chemistry - physics  
**licensing:** unknown  
**task\_types:** - Multiple choice - Multi-step QA  
**ai\_capability\_measured:** - Scientific reasoning, deep knowledge  
**metrics:** - Accuracy  
**models:** - o1 - DeepSeek-R1  
**ml\_motif:** - Reasoning & Generalization  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** Good  
**contact.name:** Julian Michael  
**contact.email:** julianjm@nyu.edu  
**datasets.links.name:** unknown  
**datasets.links.url:** unknown  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** gpqa\_diamond  
**Citations:** [28]

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Easily able to access dataset. Comes with predefined splits as mentioned in the paper         |
| documentation      | 5     | All information is listed in the associated paper   |
| metrics            | 5     | Each question has a correct answer, representing the tested model's performance.              |
| reference_solution | 1     | Common models such as GPT-3.5 were compared. They are not open and don't provide requirements |
| software           | 5     | Python version and requirements specified on Github site                                      |
| specification      | 2     | No system constraints or I/O specified  |



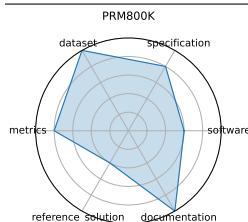
### 3.45 PRM800K

PRM800K is a process supervision dataset containing 800,000 step-level correctness labels for model-generated solutions to problems from the MATH dataset.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2023-05-30  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2023-05-30  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2023-05-30  |
| <b>url:</b>                    | <a href="https://github.com/openai/prm800k/tree/main">https://github.com/openai/prm800k/tree/main</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2305.20050   |
| <b>domain:</b>                 | - Mathematics   |
| <b>focus:</b>                  | Math reasoning generalization   |
| <b>keywords:</b>               | - calculus - algebra - number theory - geometry   |
| <b>licensing:</b>              | MIT License   |
| <b>task_types:</b>             | - Problem solving   |
| <b>ai_capability_measured:</b> | - Math reasoning and generalization   |
| <b>metrics:</b>                | - Accuracy  |
| <b>models:</b>                 | - GPT-4   |
| <b>ml_motif:</b>               | - Reasoning & Generalization  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Reasoning   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Math problems & Annotated reasoning steps based off of Dan Hendrycks' MATH dataset                    |
| <b>contact.name:</b>           | Karl Cobbe  |
| <b>contact.email:</b>          | karl@openai.com   |
| <b>datasets.links.name:</b>    | PRM800K: A Process Supervision Dataset  |
| <b>datasets.links.url:</b>     | <a href="https://github.com/openai/prm800k/tree/main">https://github.com/openai/prm800k/tree/main</a> |
| <b>results.links.name:</b>     | Let's Verify Step by Step   |
| <b>results.links.url:</b>      | <a href="https://arxiv.org/abs/2305.20050">https://arxiv.org/abs/2305.20050</a>                       |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | prm800k   |
| <b>Citations:</b>              | [29]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Dataset follows all FAIR Principles. Train/Test splits are available in the PRM800K repo  |
| documentation      | 5     | Documentation is present in the PRM800K repo and "Lets Verify Step by Step" paper.  |
| metrics            | 4     | Correctness is used as the primary metric, with grading guidelines provided.  |
| reference_solution | 2     | A reference solution is mentioned in the "Lets Verify Step by Step" paper, but the model is not open-sourced.   |
| software           | 3     | Code is provided in the PRM800K Repo for evaluation and grading, documentation is present but no environment details, baseline model, or training code is given |
| specification      | 4     | Task is well specified, format, inputs, and outputs are mentioned. No system constraints are provided.  |



### 3.46 FEA Bench (Finite Element Analysis Benchmark): Evaluating Language Models on Multiphysics Reasoning Ability

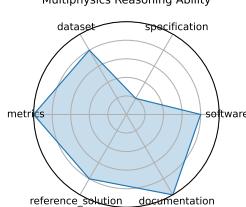
N/A

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2023-01-26  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2023-01-26  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | no  |
| <b>valid_date:</b>             | 2023-01-26  |
| <b>url:</b>                    | <a href="https://github.com/google/feabench">https://github.com/google/feabench</a>   |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Mathematics   |
| <b>focus:</b>                  | FEA simulation accuracy and performance   |
| <b>keywords:</b>               | - finite element - simulation - PDE   |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Simulation - Performance evaluation   |
| <b>ai_capability_measured:</b> | - Numerical simulation accuracy and efficiency  |
| <b>metrics:</b>                | - Solve time - Error norm   |
| <b>models:</b>                 | - FEniCS - deal.II  |
| <b>ml_motif:</b>               | - Reasoning & Generalization  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | unknown   |
| <b>notes:</b>                  | OK  |
| <b>contact.name:</b>           | unknown   |
| <b>contact.email:</b>          | unknown   |
| <b>datasets.links.name:</b>    | FEABench Github   |
| <b>datasets.links.url:</b>     | <a href="https://github.com/google/feabench?tab=readme-ov-file#datasets">https://github.com/google/feabench?tab=readme-ov-file#datasets</a> |
| <b>results.links.name:</b>     | unknown   |
| <b>results.links.url:</b>      | unknown   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | feabench_finite_element_analysis_benchmark_evaluating_language_models_on_multiphysics_reasoning_ability                                     |
| <b>Citations:</b>              | [30]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 4     | Available, but not split into sets                         |
| documentation      | 5     | In associated paper  |
| metrics            | 5     | Fully defined metrics                                      |
| reference_solution | 4     | Three open-source models were used. No system constraints. |
| software           | 4     | Code is available, but poorly documented                   |
| specification      | 1     | Output is defined and task clarity is questionable         |

FEABench (Finite Element Analysis Benchmark): Evaluating Language Models on Multiphysics Reasoning Ability



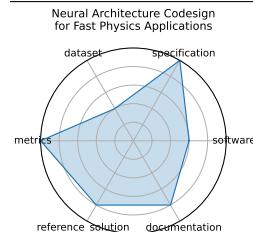
### 3.47 Neural Architecture Codesign for Fast Physics Applications

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.

|                                |  |
|--------------------------------|--|
| <b>date:</b>                   | 2025-01-09   |
| <b>version:</b>                | v1.0   |
| <b>last_updated:</b>           | 2025-01  |
| <b>expired:</b>                | unknown  |
| <b>valid:</b>                  | yes  |
| <b>valid_date:</b>             | 2025-01-09   |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2501.05515">https://arxiv.org/abs/2501.05515</a>                          |
| <b>doi:</b>                    | 10.48550/arXiv.2501.05515  |
| <b>domain:</b>                 | - High Energy Physics  |
| <b>focus:</b>                  | Automated neural architecture search and hardware-efficient model codesign for fast physics applications |
| <b>keywords:</b>               | - neural architecture search - FPGA deployment - quantization - pruning - hls4ml                         |
| <b>licensing:</b>              | Via Fermilab   |
| <b>task_types:</b>             | - Classification - Peak finding  |
| <b>ai_capability_measured:</b> | - Hardware-aware model optimization; low-latency inference   |
| <b>metrics:</b>                | - Accuracy - Latency - Resource utilization  |
| <b>models:</b>                 | - NAC-based BraggNN - NAC-optimized Deep Sets (jet)  |
| <b>ml_motif:</b>               | - Classification   |
| <b>type:</b>                   | Framework  |
| <b>ml_task:</b>                | - Supervised Learning  |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.                                    |
| <b>notes:</b>                  | Demonstrated two case studies (materials science, HEP); pipeline and code open-sourced.                  |
| <b>contact.name:</b>           | Jason Weitz (UCSD), Nhan Tran (FNAL)   |
| <b>contact.email:</b>          | unknown  |
| <b>results.links.name:</b>     | ChatGPT LLM  |
| <b>fair.reproducible:</b>      | Yes (nac-opt, hls4ml)  |
| <b>fair.benchmark_ready:</b>   | No   |
| <b>id:</b>                     | neural_architecture_codesign_for_fast_physics_applications   |
| <b>Citations:</b>              | [31]   |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 2     | Simulated datasets referenced but not publicly available or FAIR-compliant                     |
| documentation      | 4     | Detailed paper and tools described; open repo planned but not yet complete                     |
| metrics            | 5     | Clear, quantitative metrics aligned with task goals and hardware evaluation                    |
| reference_solution | 4     | Models tested on hardware with source code references; full training pipeline not yet released |
| software           | 3     | Toolchain (hls4ml, nac-opt) described but not yet containerized or fully packaged              |
| specification      | 5     | Fully specified task with constraints and target deployment; includes hardware context         |



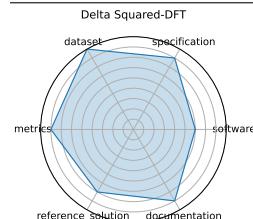
### 3.48 Delta Squared-DFT

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.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97788">https://neurips.cc/virtual/2024/poster/97788</a>       |
| <b>doi:</b>                    | 10.48550/arXiv.2406.14347   |
| <b>domain:</b>                 | - Chemistry - Materials Science   |
| <b>focus:</b>                  | Benchmarking machine-learning corrections to DFT using Delta Squared-trained models for reaction energies     |
| <b>keywords:</b>               | - density functional theory - Delta Squared-ML correction - reaction energetics - quantum chemistry           |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Regression  |
| <b>ai_capability_measured:</b> | - High-accuracy energy prediction - DFT correction  |
| <b>metrics:</b>                | - Mean Absolute Error (eV) - Energy ranking accuracy  |
| <b>models:</b>                 | - Delta Squared-ML correction networks - Kernel ridge regression  |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Dataset + Benchmark   |
| <b>ml_task:</b>                | - Regression  |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.   |
| <b>notes:</b>                  | Demonstrates CC-level accuracy with ~1% of high-level data. Benchmarks publicly included for reproducibility. |
| <b>contact.name:</b>           | Wei Liu   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | delta_squared-dft   |
| <b>Citations:</b>              | [32]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 4.5   | Multi-modal quantum chemistry datasets are standardized and accessible; repository available.                             |
| documentation      | 4     | Source code supports pipeline reuse, but formal evaluation splits may vary.   |
| metrics            | 4     | Uses standard regression metrics like MAE and energy ranking accuracy; appropriate for task.                              |
| reference_solution | 3.5   | Includes baseline regression and kernel ridge models; implementations are reproducible.                                   |
| software           | 3     | Source code and baseline models available for ML correction to DFT; framework maturity is moderate.                       |
| specification      | 4     | Benchmark focuses on reaction energy prediction with clear goals, though some task specifics could be formalized further. |



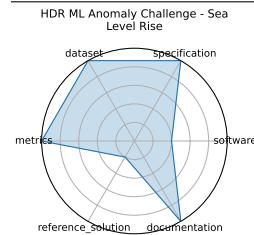
### 3.49 HDR ML Anomaly Challenge - Sea Level Rise

A challenge combining North Atlantic sea-level time-series and satellite imagery to detect flooding anomalies. Models submitted via Codabench.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2025-03-03  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2025-03   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2025-03-03  |
| <b>url:</b>                    | <a href="https://www.codabench.org/competitions/3223/">https://www.codabench.org/competitions/3223/</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2503.02112   |
| <b>domain:</b>                 | - Climate & Earth Science   |
| <b>focus:</b>                  | Detecting anomalous sea-level rise and flooding events via time-series and satellite imagery            |
| <b>keywords:</b>               | - anomaly detection - climate science - sea-level rise - time-series - remote sensing                   |
| <b>licensing:</b>              | NA  |
| <b>task_types:</b>             | - Anomaly Detection   |
| <b>ai_capability_measured:</b> | - Detection of environmental anomalies  |
| <b>metrics:</b>                | - ROC-AUC - Precision/Recall  |
| <b>models:</b>                 | - CNNs, RNNs, Transformers  |
| <b>ml_motif:</b>               | - Anomaly Detection   |
| <b>type:</b>                   | Dataset   |
| <b>ml_task:</b>                | - Anomaly Detection   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.                                   |
| <b>notes:</b>                  | Sponsored by NSF HDR; integrates sensor and satellite data.   |
| <b>contact.name:</b>           | HDR A3D3 Team   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | hdr_ml_anomaly_challenge--sea_level_rise  |
| <b>Citations:</b>              | [33]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | Uses preprocessed, public, and well-structured sensor and satellite data for the North Atlantic sea-level rise region.                 |
| documentation      | 5     | Challenge page, starter kits, and related papers offer strong guidance for participants.   |
| metrics            | 5     | Standard metrics such as ROC-AUC, precision, and recall are specified and suitable for the anomaly detection tasks.                    |
| reference_solution | 1     | No starter models or baseline implementations linked or provided publicly.   |
| software           | 2     | Benchmark platform exists on Codabench, but no baseline code or maintained repository for reference solutions provided yet.            |
| specification      | 5     | Well-defined anomaly detection task combining satellite imagery and time-series data, with clear physical and domain-specific framing. |



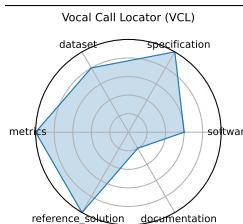
### 3.50 Vocal Call Locator (VCL)

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.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97470">https://neurips.cc/virtual/2024/poster/97470</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Biology & Medicine  |
| <b>focus:</b>                  | Benchmarking sound-source localization of rodent vocalizations from multi-channel audio                 |
| <b>keywords:</b>               | - source localization - bioacoustics - time-series - SSL  |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Sound source localization   |
| <b>ai_capability_measured:</b> | - Source localization accuracy in bioacoustic settings  |
| <b>metrics:</b>                | - Localization error (cm) - Recall/Precision  |
| <b>models:</b>                 | - CNN-based SSL models  |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Dataset   |
| <b>ml_task:</b>                | - Anomaly Detection / localization  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Dataset spans real, simulated, and mixed audio; supports benchmarking across data types .               |
| <b>contact.name:</b>           | Ralph Peterson  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | vocal_call_locator_vcl  |
| <b>Citations:</b>              | [34]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 4     | Large-scale audio dataset covering real and simulated data with standardized splits, though exact data formats are not fully detailed.    |
| documentation      | 1     | Methodology and paper are thorough, but setup instructions and runnable code are not publicly provided, limiting user onboarding.         |
| metrics            | 5     | Includes localization error, precision, recall, and other relevant metrics for robust evaluation.   |
| reference_solution | 5     | Multiple baselines evaluated over diverse models and architectures, supporting reproducibility of benchmark comparisons.                  |
| software           | 3     | Some baseline CNN models for sound source localization are reported, but no publicly available or fully integrated runnable codebase yet. |
| specification      | 5     | Well-defined localization tasks with multiple scenarios and real-world environment conditions; input/output formats clearly described.    |



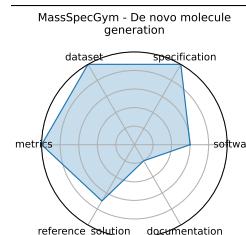
### 3.51 MassSpecGym - De novo molecule generation

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.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97823">https://neurips.cc/virtual/2024/poster/97823</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Chemistry   |
| <b>focus:</b>                  | Benchmark suite for discovery and identification of molecules via MS/MS                                 |
| <b>keywords:</b>               | - mass spectrometry - molecular structure - de novo generation - retrieval - dataset                    |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - De novo generation - Retrieval - Simulation   |
| <b>ai_capability_measured:</b> | - Molecular identification and generation from spectral data  |
| <b>metrics:</b>                | - Structure accuracy - Retrieval precision - Simulation MSE   |
| <b>models:</b>                 | - Graph-based generative models - Retrieval baselines   |
| <b>ml_motif:</b>               | - Generative  |
| <b>type:</b>                   | Dataset, Benchmark  |
| <b>ml_task:</b>                | - Generation, retrieval, simulation   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Dataset $\sim >1M$ spectra; open-source GitHub repo; widely cited as a go-to benchmark for MS/MS tasks. |
| <b>contact.name:</b>           | Roman Bushuiev  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | massspecgym_-_de_novo_molecule_generation   |
| <b>Citations:</b>              | [35]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Largest public MS/MS dataset with extensive annotations; minor point deducted for lack of explicit train/validation/test splits.          |
| documentation      | 1     | Paper and poster describe benchmark goals and design, but documentation and user guides are minimal and repo status uncertain.            |
| metrics            | 5     | Well-defined metrics such as structure accuracy, retrieval precision, and simulation MSE used consistently.                               |
| reference_solution | 3.5   | CNN-based baselines are referenced, but pretrained weights and comprehensive training pipelines are not fully documented.                 |
| software           | 3     | Open-source GitHub repository available; baseline models and training code partially provided but overall framework maturity is moderate. |
| specification      | 5     | Clearly defined tasks including molecule generation, retrieval, and spectrum simulation, scoped for MS/MS molecular identification.       |



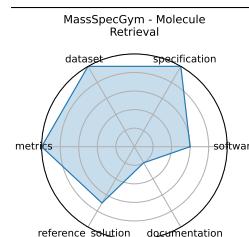
## 3.52 MassSpecGym - Molecule Retrieval

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.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97823">https://neurips.cc/virtual/2024/poster/97823</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Chemistry   |
| <b>focus:</b>                  | Benchmark suite for discovery and identification of molecules via MS/MS                                 |
| <b>keywords:</b>               | - mass spectrometry - molecular structure - de novo generation - retrieval - dataset                    |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - De novo generation - Retrieval - Simulation   |
| <b>ai_capability_measured:</b> | - Molecular identification and generation from spectral data  |
| <b>metrics:</b>                | - Structure accuracy - Retrieval precision - Simulation MSE   |
| <b>models:</b>                 | - Graph-based generative models - Retrieval baselines   |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Dataset, Benchmark  |
| <b>ml_task:</b>                | - Generation, retrieval, simulation   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Dataset $\sim >1M$ spectra; open-source GitHub repo; widely cited as a go-to benchmark for MS/MS tasks. |
| <b>contact.name:</b>           | Roman Bushuiev  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | massspecgym_-_molecule_retrieval  |
| <b>Citations:</b>              | [35]  |

### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Largest public MS/MS dataset with extensive annotations; minor point deducted for lack of explicit train/validation/test splits.          |
| documentation      | 1     | Paper and poster describe benchmark goals and design, but documentation and user guides are minimal and repo status uncertain.            |
| metrics            | 5     | Well-defined metrics such as structure accuracy, retrieval precision, and simulation MSE used consistently.                               |
| reference_solution | 3.5   | CNN-based baselines are referenced, but pretrained weights and comprehensive training pipelines are not fully documented.                 |
| software           | 3     | Open-source GitHub repository available; baseline models and training code partially provided but overall framework maturity is moderate. |
| specification      | 5     | Clearly defined tasks including molecule generation, retrieval, and spectrum simulation, scoped for MS/MS molecular identification.       |



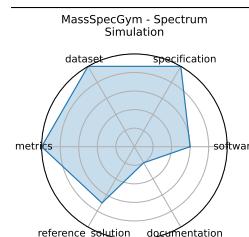
### 3.53 MassSpecGym - Spectrum Simulation

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.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-12-13  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-12-13  |
| <b>url:</b>                    | <a href="https://neurips.cc/virtual/2024/poster/97823">https://neurips.cc/virtual/2024/poster/97823</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Chemistry   |
| <b>focus:</b>                  | Benchmark suite for discovery and identification of molecules via MS/MS                                 |
| <b>keywords:</b>               | - mass spectrometry - molecular structure - de novo generation - retrieval - dataset                    |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - De novo generation - Retrieval - Simulation   |
| <b>ai_capability_measured:</b> | - Molecular identification and generation from spectral data  |
| <b>metrics:</b>                | - Structure accuracy - Retrieval precision - Simulation MSE   |
| <b>models:</b>                 | - Graph-based generative models - Retrieval baselines   |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Dataset, Benchmark  |
| <b>ml_task:</b>                | - Generation, retrieval, simulation   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Dataset $\sim >1M$ spectra; open-source GitHub repo; widely cited as a go-to benchmark for MS/MS tasks. |
| <b>contact.name:</b>           | Roman Bushuiev  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | massspecgym_-_spectrum_simulation   |
| <b>Citations:</b>              | [35]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Largest public MS/MS dataset with extensive annotations; minor point deducted for lack of explicit train/validation/test splits.          |
| documentation      | 1     | Paper and poster describe benchmark goals and design, but documentation and user guides are minimal and repo status uncertain.            |
| metrics            | 5     | Well-defined metrics such as structure accuracy, retrieval precision, and simulation MSE used consistently.                               |
| reference_solution | 3.5   | CNN-based baselines are referenced, but pretrained weights and comprehensive training pipelines are not fully documented.                 |
| software           | 3     | Open-source GitHub repository available; baseline models and training code partially provided but overall framework maturity is moderate. |
| specification      | 5     | Clearly defined tasks including molecule generation, retrieval, and spectrum simulation, scoped for MS/MS molecular identification.       |



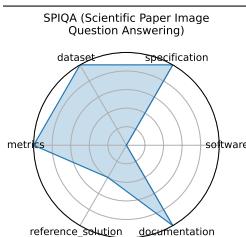
### 3.54 SPIQA (Scientific Paper Image Question Answering)

SPIQA assesses AI models' ability to interpret and answer questions about figures and tables in scientific papers by integrating visual and textual modalities with chain-of-thought reasoning.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-07-12  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2024-07-12  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-07-12  |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2407.09413">https://arxiv.org/abs/2407.09413</a>                         |
| <b>doi:</b>                    | 10.48550/arXiv.2407.09413   |
| <b>domain:</b>                 | - Computational Science & AI  |
| <b>focus:</b>                  | Multimodal QA on scientific figures   |
| <b>keywords:</b>               | - multimodal QA - figure understanding - table comprehension - chain-of-thought                         |
| <b>licensing:</b>              | Apache 2.0 License  |
| <b>task_types:</b>             | - Question answering - Multimodal QA - Chain-of-Thought evaluation                                      |
| <b>ai_capability_measured:</b> | - Visual-textual reasoning in scientific contexts   |
| <b>metrics:</b>                | - Accuracy - F1 score   |
| <b>models:</b>                 | - Chain-of-Thought models - Multimodal QA systems   |
| <b>ml_motif:</b>               | - Multimodal Reasoning  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Good  |
| <b>contact.name:</b>           | Subhashini Venugopalan  |
| <b>contact.email:</b>          | vsubhashini@google.com  |
| <b>datasets.links.name:</b>    | Hugging Face  |
| <b>datasets.links.url:</b>     | <a href="https://huggingface.co/datasets/google/spiqa">https://huggingface.co/datasets/google/spiqa</a> |
| <b>results.links.name:</b>     | unknown   |
| <b>results.links.url:</b>      | unknown   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | spiqa_scientific_paper_image_question_answering   |
| <b>Citations:</b>              | [36]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Dataset is available (via paper/appendix), includes train/test/valid split. FAIR-compliant with minor gaps in versioning or access standardization. |
| documentation      | 5     | All information provided in paper   |
| metrics            | 5     | Uses quantitative metrics (Accuracy, F1) aligned with the task  |
| reference_solution | 2     | Multiple model results (e.g., GPT-4V, Gemini) reported; baselines exist, but full runnable code not confirmed for all.                              |
| software           | 0     | Not provided  |
| specification      | 5     | Task administration clearly defined; prompt instructions explicitly given, no ambiguity in format or scope.   |



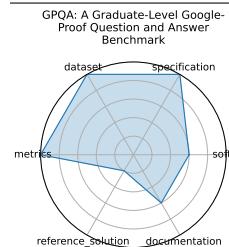
### 3.55 GPQA: A Graduate-Level Google-Proof Question and Answer Benchmark

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.

|                                |  |
|--------------------------------|--|
| <b>date:</b>                   | 2023-11-20   |
| <b>version:</b>                | v1.0   |
| <b>last_updated:</b>           | 2023-11  |
| <b>expired:</b>                | unknown  |
| <b>valid:</b>                  | yes  |
| <b>valid_date:</b>             | 2023-11-20   |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2311.12022">https://arxiv.org/abs/2311.12022</a>      |
| <b>doi:</b>                    | 10.48550/arXiv.2311.12022  |
| <b>domain:</b>                 | - Biology & Medicine - High Energy Physics - Chemistry                               |
| <b>focus:</b>                  | Graduate-level, expert-validated multiple-choice questions hard even with web access |
| <b>keywords:</b>               | - Google-proof - multiple-choice - expert reasoning - science QA                     |
| <b>licensing:</b>              | NA   |
| <b>task_types:</b>             | - Multiple choice  |
| <b>ai_capability_measured:</b> | - Scientific reasoning - knowledge probing   |
| <b>metrics:</b>                | - Accuracy   |
| <b>models:</b>                 | - GPT-4 baseline   |
| <b>ml_motif:</b>               | - Reasoning & Generalization   |
| <b>type:</b>                   | Benchmark  |
| <b>ml_task:</b>                | - Multiple choice  |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.                |
| <b>notes:</b>                  | Google-proof, supports oversight research.   |
| <b>contact.name:</b>           | David Rein (NYU)   |
| <b>contact.email:</b>          | unknown  |
| <b>datasets.links.name:</b>    | GPQA dataset   |
| <b>datasets.links.url:</b>     | zip/HuggingFace  |
| <b>results.links.name:</b>     | ChatGPT LLM  |
| <b>fair.reproducible:</b>      | Yes  |
| <b>fair.benchmark_ready:</b>   | Yes  |
| <b>id:</b>                     | gpqa_a_graduate-level_google-proof_question_and_answer_benchmark                     |
| <b>Citations:</b>              | [37]   |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 5     | The GPQA dataset is publicly released, well curated, with metadata and clearly documented splits.  |
| documentation      | 3     | Documentation includes dataset description and benchmark instructions, but lacks detailed usage tutorials or pipelines.  |
| metrics            | 5     | Accuracy is the primary metric and is clearly defined and appropriate for multiple-choice QA.  |
| reference_solution | 1     | No baseline implementations or starter code are linked or provided for reproduction.   |
| software           | 3     | Dataset and benchmark materials are publicly available via HuggingFace and GitHub, but no integrated runnable code or software framework is provided.                |
| specification      | 5     | Task is clearly defined as a multiple-choice benchmark requiring expert-level scientific reasoning. Input/output formats and evaluation criteria are well described. |



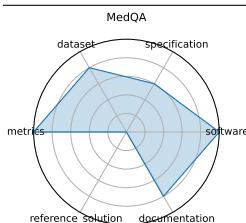
### 3.56 MedQA

MedQA is a large-scale multiple-choice dataset drawn from professional medical board exams (e.g., USMLE), testing AI systems on diagnostic and medical knowledge questions in English and Chinese.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2020-09-28  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2020-09-28  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2020-09-28  |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2009.13081">https://arxiv.org/abs/2009.13081</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2009.13081   |
| <b>domain:</b>                 | - Biology & Medicine  |
| <b>focus:</b>                  | Medical board exam QA   |
| <b>keywords:</b>               | - USMLE - diagnostic QA - medical knowledge - multilingual                      |
| <b>licensing:</b>              | Under Association for the Advancement of Artificial Intelligence                |
| <b>task_types:</b>             | - Multiple choice   |
| <b>ai_capability_measured:</b> | - Medical diagnosis and knowledge retrieval                                     |
| <b>metrics:</b>                | - Accuracy  |
| <b>models:</b>                 | - Neural reader - Retrieval-based QA systems                                    |
| <b>ml_motif:</b>               | - Reasoning & Generalization  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Multilingual (English, Simplified and Traditional Chinese)                      |
| <b>contact.name:</b>           | Di Jin  |
| <b>contact.email:</b>          | jindi15@mit.edu   |
| <b>datasets.links.name:</b>    | Github  |
| <b>datasets.links.url:</b>     | <a href="https://github.com/jind11/MedQA">https://github.com/jind11/MedQA</a>   |
| <b>results.links.name:</b>     | unknown   |
| <b>results.links.url:</b>      | unknown   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | medqa   |
| <b>Citations:</b>              | [38]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 4     | Dataset is publicly available (GitHub, paper, Hugging Face), well-structured. However, versioning and metadata could be more standardized to fully meet FAIR criteria.                 |
| documentation      | 4     | Paper is available. Evaluation criteria are not mentioned.   |
| metrics            | 5     | Uses clear, quantitative metric (accuracy), standard for multiple-choice benchmarks; easily comparable across models.  |
| reference_solution | 0     | No reference solution mentioned.   |
| software           | 5     | All code available on the github   |
| specification      | 3     | Task is clearly defined as multiple-choice QA for medical board exams; input and output formats are explicit; task scope is rigorous and structured. System constraints not specified. |



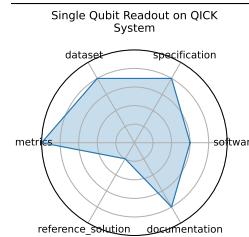
### 3.57 Single Qubit Readout on QICK System

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. :contentReference[oaicite:0]{index=0}

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2025-01-24  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2025-02   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2025-01-24  |
| <b>url:</b>                    | <a href="https://github.com/fastmachinelearning/ml-quantum-readout">https://github.com/fastmachinelearning/ml-quantum-readout</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2501.14663   |
| <b>domain:</b>                 | - Computational Science & AI  |
| <b>focus:</b>                  | Real-time single-qubit state classification using FPGA firmware   |
| <b>keywords:</b>               | - qubit readout - hls4ml - FPGA - QICK  |
| <b>licensing:</b>              | NA  |
| <b>task_types:</b>             | - Classification  |
| <b>ai_capability_measured:</b> | - Single-shot fidelity - inference latency  |
| <b>metrics:</b>                | - Accuracy - Latency  |
| <b>models:</b>                 | - hls4ml quantized NN   |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.   |
| <b>notes:</b>                  | Achieves ~96% fidelity with ~32 ns latency and low FPGA resource utilization.   |
| <b>contact.name:</b>           | Javier Campos, Giuseppe Di Guglielmo  |
| <b>contact.email:</b>          | unknown   |
| <b>datasets.links.name:</b>    | Zenodo: ml-quantum-readout dataset  |
| <b>datasets.links.url:</b>     | <a href="zenodo.org/records/14427490">zenodo.org/records/14427490</a>   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | single_qubit_readout_on_qick_system   |
| <b>Citations:</b>              | [39]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 4     | Dataset hosted on Zenodo with structured data; however, detailed documentation on image acquisition and labeling pipeline is limited.  |
| documentation      | 4     | Codabench task page and GitHub repo provide descriptions and usage instructions, but detailed API or deployment tutorials are limited.   |
| metrics            | 5     | Standard classification metrics (accuracy, latency) are used and directly relevant to the quantum readout task.  |
| reference_solution | 1     | No baseline or starter models with runnable code are linked publicly.  |
| software           | 3     | Code and FPGA firmware available on GitHub; integration with hls4ml demonstrated. Some deployment details and examples are provided but overall software maturity is moderate. |
| specification      | 4     | Task clearly defined: real-time single-qubit state classification with latency and fidelity constraints. Labeling and ground truth definitions could be more explicit.         |



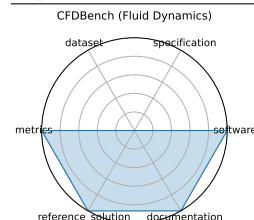
### 3.58 CFD Bench (Fluid Dynamics)

CFDBench provides large-scale CFD data for four canonical fluid flow problems, assessing neural operators' ability to generalize to unseen PDE parameters and domains.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-10-01  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2024-10-01  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-10-01  |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2310.05963">https://arxiv.org/abs/2310.05963</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2310.05963   |
| <b>domain:</b>                 | - Mathematics   |
| <b>focus:</b>                  | Neural operator surrogate modeling  |
| <b>keywords:</b>               | - neural operators - CFD - FNO - DeepONet                                       |
| <b>licensing:</b>              | CC-BY-4.0   |
| <b>task_types:</b>             | - Surrogate modeling  |
| <b>ai_capability_measured:</b> | - Generalization of neural operators for PDEs                                   |
| <b>metrics:</b>                | - L2 error - MAE  |
| <b>models:</b>                 | - FNO - DeepONet - U-Net  |
| <b>ml_motif:</b>               | - Regression  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | Numerous, as it's a benchmark for ML models                                     |
| <b>notes:</b>                  | 302K frames across 739 cases  |
| <b>contact.name:</b>           | Yining Luo  |
| <b>contact.email:</b>          | yining.luo@mail.utoronto.ca   |
| <b>datasets.links.name:</b>    | unknown   |
| <b>datasets.links.url:</b>     | unknown   |
| <b>results.links.name:</b>     | unknown   |
| <b>results.links.url:</b>      | unknown   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | cfdbench_fluid_dynamics   |
| <b>Citations:</b>              | [40]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 0     | Not given   |
| documentation      | 5     | Associated paper gives all necessary information.   |
| metrics            | 5     | Quantitative metrics (L2 error, MAE, relative error) are clearly defined and align with regression task objectives. |
| reference_solution | 5     | Baseline models like FNO and DeepONet are implemented, hardware specified.  |
| software           | 5     | The benchmark provides Python scripts for data loading, preprocessing, and model training/evaluation                |
| specification      | 0     | Not listed  |



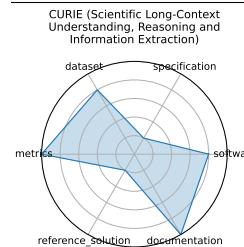
### 3.59 CURIE (Scientific Long-Context Understanding, Reasoning and Information Extraction)

CURIE is a benchmark of 580 problems across six scientific disciplines-materials science, quantum computing, biology, chemistry, climate science, and astrophysics- designed to evaluate LLMs on long-context understanding, reasoning, and information extraction in realistic scientific workflows.

**date:** 2024-04-02  
**version:** 1  
**last\_updated:** 2024-04-02  
**expired:** false  
**valid:** yes  
**valid\_date:** 2024-04-02  
**url:** <https://arxiv.org/abs/2503.13517>  
**doi:** 10.48550/arXiv.2503.13517  
**domain:** - Materials Science - High Energy Physics - Biology & Medicine - Chemistry - Climate & Earth Science  
**focus:** Long-context scientific reasoning  
**keywords:** - long-context - information extraction - multimodal  
**licensing:** Apache 2.0 License  
**task\_types:** - Information extraction - Reasoning - Concept tracking - Aggregation - Algebraic manipulation  
- Multimodal comprehension  
**ai\_capability\_measured:** - Long-context understanding and scientific reasoning  
**metrics:** - Accuracy  
**models:** - unkown  
**ml\_motif:** - Reasoning & Generalization  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** Good  
**contact.name:** Subhashini Venugopalan  
**contact.email:** vsubhashini@google.com  
**datasets.links.name:** unknown  
**datasets.links.url:** unknown  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** curie\_scientific\_long-context\_understanding\_reasoning\_and\_information\_extraction  
**Citations:** [41]

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 4     | Dataset is available via Github, but hard to find   |
| documentation      | 5     | Associated paper explains all criteria  |
| metrics            | 5     | Quantitative metrics such as ROUGE-L and F1 used. Metrics are tailored to the specific problem. |
| reference_solution | 1     | Exists, but is not open   |
| software           | 4     | Code is available, but not well documented  |
| specification      | 1     | Explains types of problems in detail, but does not state exactly how to administer them.        |



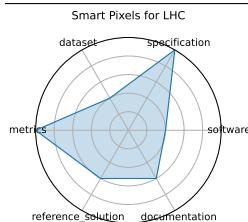
## 3.60 Smart Pixels for LHC

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 microW/pixel and operates in combinatorial digital logic.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-06-24  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-06   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-06-24  |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2406.14860">https://arxiv.org/abs/2406.14860</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2406.14860   |
| <b>domain:</b>                 | - High Energy Physics   |
| <b>focus:</b>                  | On-sensor, in-pixel ML filtering for high-rate LHC pixel detectors              |
| <b>keywords:</b>               | - smart pixel - on-sensor inference - data reduction - trigger                  |
| <b>licensing:</b>              | Via Fermilab  |
| <b>task_types:</b>             | - Image Classification - Data filtering   |
| <b>ai_capability_measured:</b> | - On-chip - low-power inference; data reduction                                 |
| <b>metrics:</b>                | - Data rejection rate - Power per pixel   |
| <b>models:</b>                 | - 2-layer pixel NN  |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Image Classification  |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.           |
| <b>notes:</b>                  | Prototype in CMOS 28 nm; proof-of-concept for Phase III pixel upgrades.         |
| <b>contact.name:</b>           | Lindsey Gray; Jennet Dickinson  |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes (Zenodo:7331128)  |
| <b>id:</b>                     | smart_pixels_for_lhc  |
| <b>Citations:</b>              | [42]  |

### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 2     | No dataset links; not publicly hosted or FAIR-compliant  |
| documentation      | 3     | Paper contains detailed descriptions, but no repo or external guide for reproducing results        |
| metrics            | 5     | None   |
| reference_solution | 3     | In-pixel 2-layer NN described and evaluated, but reproducibility and source files are not released |
| software           | 2     | No packaged code or setup scripts available; replication depends on hardware description and paper |
| specification      | 5     | None   |



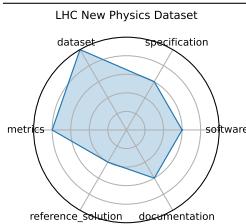
### 3.61 LHC New Physics Dataset

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.

|                                |  |
|--------------------------------|--|
| <b>date:</b>                   | 2021-07-05   |
| <b>version:</b>                | v1.0   |
| <b>last_updated:</b>           | 2021-07  |
| <b>expired:</b>                | unknown  |
| <b>valid:</b>                  | yes  |
| <b>valid_date:</b>             | 2021-07-05   |
| <b>url:</b>                    | <a href="https://arxiv.org/pdf/2107.02157">https://arxiv.org/pdf/2107.02157</a>                      |
| <b>doi:</b>                    | unknown  |
| <b>domain:</b>                 | - High Energy Physics  |
| <b>focus:</b>                  | Real-time LHC event filtering for anomaly detection using proton collision data                      |
| <b>keywords:</b>               | - anomaly detection - proton collision - real-time inference - event filtering - unsupervised ML     |
| <b>licensing:</b>              | unknown  |
| <b>task_types:</b>             | - Anomaly Detection - Event classification   |
| <b>ai_capability_measured:</b> | - Unsupervised signal detection under latency and bandwidth constraints                              |
| <b>metrics:</b>                | - ROC-AUC - Detection efficiency   |
| <b>models:</b>                 | - Autoencoder - Variational autoencoder - Isolation forest   |
| <b>ml_motif:</b>               | - Anomaly Detection  |
| <b>type:</b>                   | Framework  |
| <b>ml_task:</b>                | - NA   |
| <b>solutions:</b>              | 0  |
| <b>notes:</b>                  | Includes electron/muon-filtered background and black-box signal benchmarks; 1M events per black box. |
| <b>contact.name:</b>           | Ema Puljak   |
| <b>contact.email:</b>          | ema.puljak@cern.ch   |
| <b>datasets.links.name:</b>    | Zenodo stores, background + 3 black-box signal sets. 1M events each                                  |
| <b>results.links.name:</b>     | ChatGPT LLM  |
| <b>fair.reproducible:</b>      | Yes  |
| <b>fair.benchmark_ready:</b>   | Yes  |
| <b>id:</b>                     | lhc_new_physics_dataset  |
| <b>Citations:</b>              | [43]   |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 5     | Large-scale dataset hosted on Zenodo, publicly available, well-documented, with defined train/test structure. Appears to follow at least 4 FAIR principles.   |
| documentation      | 3     | Some description in papers and dataset metadata exists, but lacks a unified guide, README, or training setup in a central location.   |
| metrics            | 4     | Uses reasonable metrics (ROC-AUC, detection efficiency) that capture performance but lacks full explanation and standard evaluation tools.  |
| reference_solution | 2     | Baselines are described across multiple papers but lack centralized, reproducible implementations and hardware/software setup details.  |
| software           | 3     | While not formally evaluated in the previous version, Zenodo and paper links suggest available code for baseline models (e.g., autoencoders, GANs), though they are scattered and not unified in a single repository. |
| specification      | 3     | The task and context are clearly described, but system constraints and formal inputs/outputs are not fully specified.   |



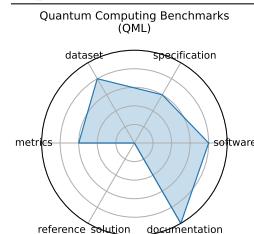
## 3.62 Quantum Computing Benchmarks (QML)

A suite of benchmarks evaluating quantum hardware and algorithms on tasks such as state preparation, circuit optimization, and error correction across multiple platforms.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2022-02-22  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2022-02-22  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2022-02-22  |
| <b>url:</b>                    | <a href="https://github.com/XanaduAI/qml-benchmarks">https://github.com/XanaduAI/qml-benchmarks</a>   |
| <b>doi:</b>                    | 10.48550/arXiv.2403.07059   |
| <b>domain:</b>                 | - Computational Science & AI  |
| <b>focus:</b>                  | Quantum algorithm performance evaluation  |
| <b>keywords:</b>               | - quantum circuits - state preparation - error correction   |
| <b>licensing:</b>              | Apache-2.0  |
| <b>task_types:</b>             | - Circuit benchmarking - State classification   |
| <b>ai_capability_measured:</b> | - Quantum algorithm performance and fidelity  |
| <b>metrics:</b>                | - Fidelity - Success probability  |
| <b>models:</b>                 | - IBM Q - IonQ - AQT@LBNL   |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | Varies per benchmark  |
| <b>notes:</b>                  | Hardware-agnostic, application-level metrics. The citation may not be correct.  |
| <b>contact.name:</b>           | Xanadu AI   |
| <b>contact.email:</b>          | <a href="mailto:support@xanadu.ai">support@xanadu.ai</a>  |
| <b>datasets.links.name:</b>    | PennyLane QML Benchmarks Datasets   |
| <b>datasets.links.url:</b>     | <a href="https://pennylane.ai/datasets/collection/qml-benchmarks">https://pennylane.ai/datasets/collection/qml-benchmarks</a>                         |
| <b>results.links.name:</b>     | QML Benchmarks GitHub Repository (Results section)  |
| <b>results.links.url:</b>      | <a href="https://github.com/XanaduAI/qml-benchmarks#results-and-leaderboards">https://github.com/XanaduAI/qml-benchmarks#results-and-leaderboards</a> |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | quantum_computing_benchmarks_qml  |
| <b>Citations:</b>              | [44]  |

### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 4     | Datasets are accessible, but not split.   |
| documentation      | 5     | Paper is available with all required information.   |
| metrics            | 3     | Partially defined, somewhat inferrable metrics. Unknown whether a system's performance is captured.     |
| reference_solution | 0     | Not provided  |
| software           | 4     | Software is built upon multiple common frameworks for simulation, training, and benchmarking workflows. |
| specification      | 3     | No system constraints. Task clarity and dataset format are not clearly specified.                       |



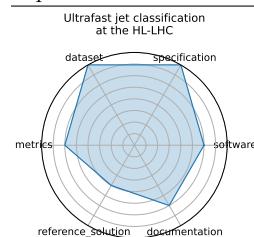
### 3.63 Ultrafast jet classification at the HL-LHC

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.

**date:** 2024-07-08  
**version:** v1.0  
**last\_updated:** 2024-07  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-07-08  
**url:** <https://arxiv.org/pdf/2402.01876>  
**doi:** 10.48550/arXiv.2402.01876  
**domain:** - High Energy Physics  
**focus:** FPGA-optimized real-time jet origin classification at the HL-LHC  
**keywords:** - jet classification - FPGA - quantization-aware training - Deep Sets - Interaction Networks  
**licensing:** CC-BY  
**task\_types:** - Classification  
**ai\_capability\_measured:** - Real-time inference under FPGA constraints  
**metrics:** - Accuracy - Latency - Resource utilization  
**models:** - MLP - Deep Sets - Interaction Network  
**ml\_motif:** - Classification  
**type:** Model  
**ml\_task:** - Supervised Learning  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Uses quantization-aware training; hardware synthesis evaluated via hls4ml  
**contact.name:** Patrick Odagiu  
**contact.email:** podagiu@ethz.ch  
**datasets.links.name:** Zenodo dataset  
**datasets.links.url:** <https://zenodo.org/records/3602260>  
**results.links.name:** ChatGPT LLM  
**results.links.url:** [https://docs.google.com/document/d/1gDf1CIYtfmfZ9urv1jCRZMYz\\_3WwEETkugUC65OZBdw](https://docs.google.com/document/d/1gDf1CIYtfmfZ9urv1jCRZMYz_3WwEETkugUC65OZBdw)  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** No  
**id:** ultrafast\_jet\_classification\_at\_the\_hl-lhc  
**Citations:** [45]

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 4     | FAIR metadata limited; no clear mention of dataset format or splits  |
| documentation      | 3     | No linked GitHub repo or setup instructions; paper provides partial guidance only                          |
| metrics            | 3     | Metrics exist (accuracy, latency, utilization), but formal definitions and evaluation guidance are limited |
| reference_solution | 2     | Reference implementations not fully reproducible; no evaluation pipeline or training setup provided        |
| software           | 3     | Not containerized; Setup and automation incomplete   |
| specification      | 4     | Hardware constraints are referenced but not fully detailed or standardized                                 |



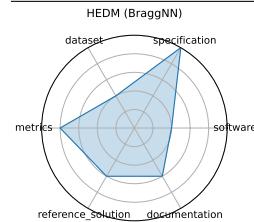
### 3.64 HEDM (BraggNN)

Uses BraggNN, a deep neural network, for rapid Bragg peak localization in high-energy diffraction microscopy, achieving about 13x speedup compared to Voigt-based methods while maintaining sub-pixel accuracy.

**date:** 2023-10-03  
**version:** v1.0  
**last\_updated:** 2023-10  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2023-10-03  
**url:** <https://arxiv.org/abs/2008.08198>  
**doi:** 10.48550/arXiv.2008.08198  
**domain:** - Materials Science  
**focus:** Fast Bragg peak analysis using deep learning in diffraction microscopy  
**keywords:** - BraggNN - diffraction - peak finding - HEDM  
**licensing:** DOE Public Access Plan  
**task\_types:** - Peak detection  
**ai\_capability\_measured:** - High-throughput peak localization  
**metrics:** - Localization accuracy - Inference time  
**models:** - BraggNN  
**ml\_motif:** - Classification  
**type:** Framework  
**ml\_task:** - Peak finding  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Enables real-time HEDM workflows; basis for NAC case study.  
**contact.name:** Jason Weitz (UCSD)  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** hedm\_braggnn  
**Citations:** [46]

#### Ratings:

| Rating                 | Value | Reason  |
|------------------------|-------|---|
| dataset                | 2     | No dataset links or FAIR metadata; unclear public access  |
| documentation          | 3     | Paper is clear, but lacks a GitHub repo or full reproducibility pipeline                              |
| metrics                | 4     | Only localization accuracy and inference time mentioned; not formally benchmarked with scripts        |
| reference_solution     | 3     | BraggNN model is described and evaluated, but no direct implementation or inference scripts available |
| software_specification | 2     | No standalone code repository or setup instructions provided  |
| specification          | 5     | None  |



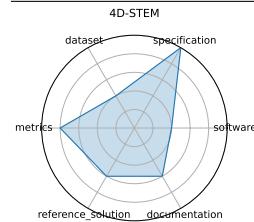
### 3.65 4D-STEM

Proposes ML methods for real-time analysis of 4D scanning transmission electron microscopy datasets; framework details in progress.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2023-12-03  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2023-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2023-12-03  |
| <b>url:</b>                    | <a href="https://openreview.net/pdf?id=7yt3N0o0W9">https://openreview.net/pdf?id=7yt3N0o0W9</a> |
| <b>doi:</b>                    | unknown   |
| <b>domain:</b>                 | - Materials Science   |
| <b>focus:</b>                  | Real-time ML for scanning transmission electron microscopy                                      |
| <b>keywords:</b>               | - 4D-STEM - electron microscopy - real-time - image processing                                  |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Image Classification - Streamed data inference  |
| <b>ai_capability_measured:</b> | - Real-time large-scale microscopy inference  |
| <b>metrics:</b>                | - Classification accuracy - Throughput  |
| <b>models:</b>                 | - CNN models (prototype)  |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Model   |
| <b>ml_task:</b>                | - Image Classification  |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | In-progress; model design under development.  |
| <b>contact.name:</b>           | Shuyu Qin   |
| <b>contact.email:</b>          | shq219@lehigh.edu   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | in progress   |
| <b>fair.benchmark_ready:</b>   | No  |
| <b>id:</b>                     | d-stem  |
| <b>Citations:</b>              | [47]  |

#### Ratings:

| Rating                 | Value | Reason  |
|------------------------|-------|---|
| dataset                | 2     | No dataset links or FAIR metadata; unclear public access  |
| documentation          | 3     | Paper is clear, but lacks a GitHub repo or full reproducibility pipeline                              |
| metrics                | 4     | Only localization accuracy and inference time mentioned; not formally benchmarked with scripts        |
| reference_solution     | 3     | BraggNN model is described and evaluated, but no direct implementation or inference scripts available |
| software_specification | 2     | No standalone code repository or setup instructions provided  |
|                        | 5     | None  |



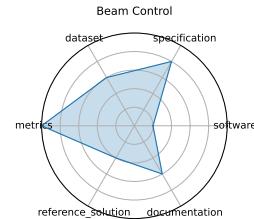
## 3.66 Beam Control

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.

**date:** 2024-05-01  
**version:** v0.2.0  
**last\_updated:** 2024-05  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-05-01  
**url:** <https://github.com/fastmachinelearning/fastml-science/tree/main/beam-control>  
**doi:** 10.48550/arXiv.2207.07958  
**domain:** - High Energy Physics  
**focus:** Reinforcement learning control of accelerator beam position  
**keywords:** - RL - beam stabilization - control systems - simulation  
**licensing:** Apache License 2.0  
**task\_types:** - Control  
**ai\_capability\_measured:** - Policy performance in simulated accelerator control  
**metrics:** - Stability - Control loss  
**models:** - DDPG - PPO (planned)  
**ml\_motif:** - Reinforcement Learning/Control  
**type:** Benchmark  
**ml\_task:** - Reinforcement Learning  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Environment defined, baseline RL implementation is in progress  
**contact.name:** Ben Hawks, Nhan Tran  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** in progress  
**fair.benchmark\_ready:** in progress  
**id:** beam\_control  
**Citations:** [48], [49]

### Ratings:

| Rating                 | Value | Reason  |
|------------------------|-------|---|
| dataset                | 3     | Not findable (no DOI/indexing); Not interoperable (format/schema unspecified)                 |
| documentation          | 3     | Setup instructions and pretrained model details are missing                                   |
| metrics                | 5     | All criteria met  |
| reference_solution     | 2     | HW/SW requirements missing; Metrics not evaluated with reference; Baseline not trainable/open |
| software_specification | 1     | Code not documented; Incomplete setup and not containerized                                   |
| specification          | 4     | Latency/resource constraints not fully quantified   |



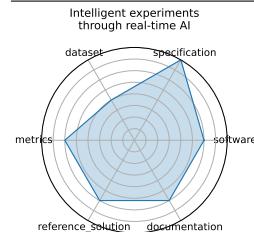
### 3.67 Intelligent experiments through real-time AI

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 micros latency. Demonstrated improved accuracy and latency on Alveo/FELIX platforms.

|                                |  |
|--------------------------------|--|
| <b>date:</b>                   | 2025-01-08   |
| <b>version:</b>                | v1.0   |
| <b>last_updated:</b>           | 2025-01  |
| <b>expired:</b>                | unknown  |
| <b>valid:</b>                  | yes  |
| <b>valid_date:</b>             | 2025-01-08   |
| <b>url:</b>                    | <a href="https://arxiv.org/pdf/2501.04845">https://arxiv.org/pdf/2501.04845</a>  |
| <b>doi:</b>                    | 10.48550/arXiv.2501.04845  |
| <b>domain:</b>                 | - High Energy Physics  |
| <b>focus:</b>                  | Real-time FPGA-based triggering and detector control for sPHENIX and future EIC  |
| <b>keywords:</b>               | - FPGA - Graph Neural Network - hls4ml - real-time inference - detector control  |
| <b>licensing:</b>              | CC BY-NC-ND 4.0  |
| <b>task_types:</b>             | - Trigger classification - Detector control - Real-time inference  |
| <b>ai_capability_measured:</b> | - Low-latency GNN inference on FPGA  |
| <b>metrics:</b>                | - Accuracy (charm and beauty detection) - Latency (micros) - Resource utilization (LUT/FF/BRAM/DSP)                                    |
| <b>models:</b>                 | - Bipartite Graph Network with Set Transformers (BGN-ST) - GarNet (edge-classifier)  |
| <b>ml_motif:</b>               | - Classification   |
| <b>type:</b>                   | Model  |
| <b>ml_task:</b>                | - Supervised Learning  |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.  |
| <b>notes:</b>                  | Achieved ~97.4% accuracy for beauty decay triggers; sub-10 micros latency on Alveo U280; hit-based FPGA design via hls4ml and FlowGNN. |
| <b>contact.name:</b>           | Jakub Kvapil   |
| <b>contact.email:</b>          | Jakub.Kvapil@lanl.gov  |
| <b>datasets.links.name:</b>    | Internal simulated tracking data (sPHENIX and EIC DIS-electron tagger)   |
| <b>results.links.name:</b>     | ChatGPT LLM  |
| <b>fair.reproducible:</b>      | Yes  |
| <b>fair.benchmark_ready:</b>   | No   |
| <b>id:</b>                     | intelligent_experiments_through_real-time_ai   |
| <b>Citations:</b>              | [50]   |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 2     | Dataset is internal and not publicly available or FAIR-compliant      |
| documentation      | 3     | No public GitHub or complete pipeline documentation                   |
| metrics            | 3     | Metrics relevant but not supported by evaluation scripts or baselines |
| reference_solution | 3     | No public or reproducible implementation released                     |
| software           | 3     | No containerized or open-source setup provided                        |
| specification      | 4     | Architectural/system specifications are incomplete                    |



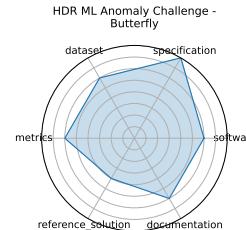
### 3.68 HDR ML Anomaly Challenge - Butterfly

Image-based challenge for detecting butterfly hybrids in microscopy-driven species data. Participants evaluate models on Codabench using image segmentation/classification.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2025-03-03  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2025-03   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2025-03-03  |
| <b>url:</b>                    | <a href="https://www.codabench.org/competitions/3764/">https://www.codabench.org/competitions/3764/</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2503.02112   |
| <b>domain:</b>                 | - Biology & Medicine  |
| <b>focus:</b>                  | Detecting hybrid butterflies via image anomaly detection in genomic-informed dataset                    |
| <b>keywords:</b>               | - anomaly detection - computer vision - genomics - butterfly hybrids                                    |
| <b>licensing:</b>              | NA  |
| <b>task_types:</b>             | - Anomaly Detection   |
| <b>ai_capability_measured:</b> | - Hybrid detection in biological systems  |
| <b>metrics:</b>                | - Classification accuracy - F1 score  |
| <b>models:</b>                 | - CNN-based detectors   |
| <b>ml_motif:</b>               | - Anomaly Detection   |
| <b>type:</b>                   | Dataset   |
| <b>ml_task:</b>                | - Anomaly Detection   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.                                   |
| <b>notes:</b>                  | Hybrid detection benchmarks hosted on Codabench   |
| <b>contact.name:</b>           | Imageomics/HDR Team   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | hdr_ml_anomaly_challenge_-_butterfly  |
| <b>Citations:</b>              | [51]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 3     | Dataset consists of real detector data with synthetic anomaly injections; access is restricted and requires NDA, limiting openness and FAIR compliance. |
| documentation      | 3     | Challenge website provides basic descriptions and evaluation metrics but lacks comprehensive tutorials or example workflows.                            |
| metrics            | 3     | Standard metrics (ROC, F1, precision) are used; evaluation protocols are clear but not deeply elaborated.   |
| reference_solution | 2     | Baselines are partially described but lack public code or reproducible execution scripts.   |
| software           | 3     | Codabench platform provides submission infrastructure but no fully maintained code repository or reproducible baseline implementations.                 |
| specification      | 4     | Task is clearly described with domain-specific anomaly detection objectives and relevant physics motivation.  |



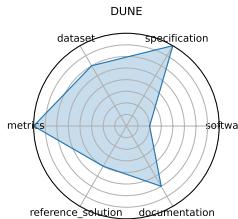
### 3.69 DUNE

Applying real-time ML methods to time-series data from DUNE detectors, exploring trigger-level anomaly detection and event selection with low latency constraints.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-10-15  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2024-10   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-10-15  |
| <b>url:</b>                    | <a href="https://indico.fnal.gov/event/66520/contributions/301423/attachments/182439/250508/fast_ml_dunedaq_sonic.pdf">https://indico.fnal.gov/event/66520/contributions/301423/attachments/182439/250508/fast_ml_dunedaq_sonic.pdf</a> |
| <b>doi:</b>                    | 10.48550/arXiv.2103.13910   |
| <b>domain:</b>                 | - High Energy Physics   |
| <b>focus:</b>                  | Real-time ML for DUNE DAQ time-series data  |
| <b>keywords:</b>               | - DUNE - time-series - real-time - trigger  |
| <b>licensing:</b>              | Via Fermilab  |
| <b>task_types:</b>             | - Trigger selection - Time-series anomaly detection   |
| <b>ai_capability_measured:</b> | - Low-latency event detection   |
| <b>metrics:</b>                | - Detection efficiency - Latency  |
| <b>models:</b>                 | - CNN - LSTM (planned)  |
| <b>ml_motif:</b>               | - Anomaly Detection   |
| <b>type:</b>                   | Benchmark (in progress)   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.   |
| <b>notes:</b>                  | Prototype models demonstrated on SONIC platform   |
| <b>contact.name:</b>           | Andrew J. Morgan  |
| <b>contact.email:</b>          | unknown   |
| <b>datasets.links.name:</b>    | DUNE SONIC data   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>fair.reproducible:</b>      | no  |
| <b>fair.benchmark_ready:</b>   | No  |
| <b>id:</b>                     | dune  |
| <b>Citations:</b>              | [52]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 3     | Dataset lacks a public URL; FAIR metadata and versioning are missing                        |
| documentation      | 3     | Documentation exists only in slides/GDocs; no implementation guide or structured release    |
| metrics            | 4     | Metrics are relevant but no benchmark baseline or detailed evaluation guidance is provided  |
| reference_solution | 2     | Autoencoder prototype exists but is not reproducible; RL model still in development         |
| software           | 1     | Code not available; no containerization or setup provided                                   |
| specification      | 4     | Constraints like latency thresholds are described qualitatively but not numerically defined |



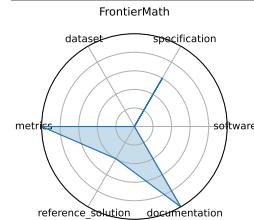
### 3.70 FrontierMath

FrontierMath is a benchmark of hundreds of expert-vetted mathematics problems spanning number theory, real analysis, algebraic geometry, and category theory, measuring LLMs ability to solve problems requiring deep abstract reasoning.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2024-11-07  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2024-11-07  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2024-11-07  |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2411.04872">https://arxiv.org/abs/2411.04872</a>                                     |
| <b>doi:</b>                    | 10.48550/arXiv.2411.04872   |
| <b>domain:</b>                 | - Mathematics   |
| <b>focus:</b>                  | Challenging advanced mathematical reasoning   |
| <b>keywords:</b>               | - symbolic reasoning - number theory - algebraic geometry - category theory   |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Problem solving   |
| <b>ai_capability_measured:</b> | - Symbolic and abstract mathematical reasoning  |
| <b>metrics:</b>                | - Accuracy  |
| <b>models:</b>                 | - unknown   |
| <b>ml_motif:</b>               | - Reasoning & Generalization  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | More information available at <a href="https://epoch.ai/frontiermath/about">https://epoch.ai/frontiermath/about</a> |
| <b>contact.name:</b>           | FrontierMath team   |
| <b>contact.email:</b>          | math_evals@epochai.org  |
| <b>datasets.links.name:</b>    | unknown   |
| <b>datasets.links.url:</b>     | unknown   |
| <b>results.links.name:</b>     | unknown   |
| <b>results.links.url:</b>      | unknown   |
| <b>fair.reproducible:</b>      | No  |
| <b>fair.benchmark_ready:</b>   | No  |
| <b>id:</b>                     | frontiermath  |
| <b>Citations:</b>              | [53]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 0     | Only samples of dataset exist, not publicly available  |
| documentation      | 5     | All necessary information is in the paper and website  |
| metrics            | 5     | All questions in the dataset have a correct answer   |
| reference_solution | 2     | Displays result of leading models on the benchmark, but none are trainable or list constraints         |
| software           | 0     | No publicaly available code to run the benchmark   |
| specification      | 3     | Well-specified process for asking questions and receiving answers. No software or hardware constraints |



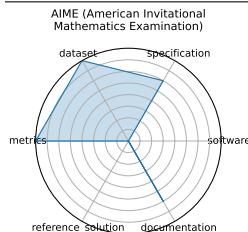
### 3.71 AIME (American Invitational Mathematics Examination)

The AIME is a 15-question, 3-hour exam for high-school students featuring challenging short-answer math problems in algebra, number theory, geometry, and combinatorics, assessing depth of problem-solving ability.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2025-03-13  |
| <b>version:</b>                | 1   |
| <b>last_updated:</b>           | 2025-03-13  |
| <b>expired:</b>                | false   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2025-03-13  |
| <b>url:</b>                    | <a href="https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions">https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions</a> |
| <b>doi:</b>                    | NA  |
| <b>domain:</b>                 | - Mathematics   |
| <b>focus:</b>                  | Pre-college advanced problem solving  |
| <b>keywords:</b>               | - algebra - combinatorics - number theory - geometry  |
| <b>licensing:</b>              | unknown   |
| <b>task_types:</b>             | - Problem solving   |
| <b>ai_capability_measured:</b> | - Mathematical problem-solving and reasoning  |
| <b>metrics:</b>                | - Accuracy  |
| <b>models:</b>                 | - unknown   |
| <b>ml_motif:</b>               | - Reasoning & Generalization  |
| <b>type:</b>                   | Benchmark   |
| <b>ml_task:</b>                | - Supervised Learning   |
| <b>solutions:</b>              | 0   |
| <b>notes:</b>                  | Designed for human test-takers  |
| <b>contact.name:</b>           | unknown   |
| <b>contact.email:</b>          | unknown   |
| <b>datasets.links.name:</b>    | AoPS website  |
| <b>datasets.links.url:</b>     | <a href="https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions">https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions</a> |
| <b>results.links.name:</b>     | unknown   |
| <b>results.links.url:</b>      | unknown   |
| <b>fair.reproducible:</b>      | Yes   |
| <b>fair.benchmark_ready:</b>   | Yes   |
| <b>id:</b>                     | aime_american_invitational_mathematics_examination  |
| <b>Citations:</b>              | [54]  |

#### Ratings:

| Rating             | Value | Reason  |
|--------------------|-------|---|
| dataset            | 4     | Easily accessible data with problems and solutions, but no splits   |
| documentation      | 3     | Some background and other information is provided, but it is not comprehensive. No info on how to run an evaluation |
| metrics            | 4     | Correctness is measured, but no grading guidelines are provided.  |
| reference_solution | 0     | Not given. Human performance stats exist, but no mentions of AI performance   |
| software           | 0     | No code available   |
| specification      | 3     | Task and Inputs/Outputs are well specified. No system constraints or dataset format is mentioned                    |



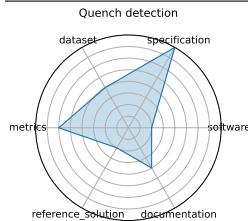
## 3.72 Quench detection

Exploration of 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.

**date:** 2024-10-15  
**version:** v1.0  
**last\_updated:** 2024-10  
**expired:** no  
**valid:** yes  
**valid\_date:** 2024-10-15  
**url:** [https://indico.cern.ch/event/1387540/contributions/6153618/attachments/2948441/5182077/fast\\_ml\\_magnets.pdf](https://indico.cern.ch/event/1387540/contributions/6153618/attachments/2948441/5182077/fast_ml_magnets.pdf)  
**doi:** NA  
**domain:** - High Energy Physics  
**focus:** Real-time detection of superconducting magnet quenches using ML  
**keywords:** - quench detection - autoencoder - anomaly detection - real-time  
**licensing:** Via Fermilab  
**task\_types:** - Anomaly Detection - Quench localization  
**ai\_capability\_measured:** - Real-time anomaly detection with multi-modal sensors  
**metrics:** - ROC-AUC - Detection latency  
**models:** - Autoencoder - RL agents (in development)  
**ml\_motif:** - Anomaly Detection  
**type:** Benchmark  
**ml\_task:** - Reinforcement, Unsupervised Learning  
**solutions:** 0  
**notes:** Precursor detection in progress; multi-modal and dynamic weighting methods  
**contact.name:** Maira Khan  
**contact.email:** unknown  
**datasets.links.name:** BPM and power supply data from BNL  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** in progress  
**fair.benchmark\_ready:** No  
**id:** quench\_detection  
**Citations:** [55]

### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 2     | Dataset URL is missing; FAIR principles largely unmet  |
| documentation      | 2     | Only a conference slide deck is available; lacks detailed instructions or repository for reproduction                                      |
| metrics            | 3     | ROC-AUC and latency are mentioned, but metric definitions and formal evaluation setup are missing  |
| reference_solution | 1     | No baseline or reproducible model implementation available   |
| software           | 1     | Code not provided; no evidence of documentation or containerization  |
| specification      | 4     | Real-time detection task is clearly described, but exact constraints, inputs/outputs, and evaluation protocol are only partially specified |



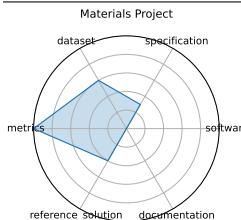
### 3.73 Materials Project

The Materials Project provides an open-access database of computed properties for inorganic materials via high-throughput density functional theory (DFT), accelerating materials discovery.

**date:** 2011-10-01  
**version:** 1  
**last\_updated:** 2011-10-01  
**expired:** false  
**valid:** yes  
**valid\_date:** 2011-10-01  
**url:** <https://materialsproject.org/>  
**doi:** unknown  
**domain:** - Materials Science  
**focus:** DFT-based property prediction  
**keywords:** - DFT - materials genome - high-throughput  
**licensing:** <https://next-gen.materialsproject.org/about/terms>  
**task\_types:** - Property prediction  
**ai\_capability\_measured:** - Prediction of inorganic material properties  
**metrics:** - MAE -  $R^2$   
**models:** - Automatminer - Crystal Graph Neural Networks  
**ml\_motif:** - Regression  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** Core component of the Materials Genome Initiative  
**contact.name:** unknown  
**contact.email:** unknown  
**datasets.links.name:** Materials Project Catalysis Explorer  
**datasets.links.url:** <https://next-gen.materialsproject.org/catalysis>  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** materials\_project  
**Citations:** [56]

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 3     | API key required to access data. No predefined splits.   |
| documentation      | 0     | No explanations or paper provided  |
| metrics            | 5     | Uses numerical metrics like MAE and $R^2$  |
| reference_solution | 2     | Numerous models (e.g., Automatminer, CGCNN) trained on the database, but no constraints or documentation listed.   |
| software           | 0     | No instructions available  |
| specification      | 1.5   | The platform offers a wide range of material property prediction tasks, but task framing and I/O formats vary by API use and are not always standardized across use cases. |



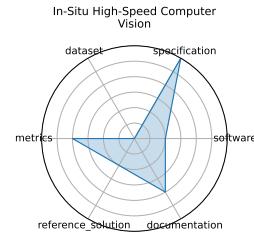
### 3.74 In-Situ High-Speed Computer Vision

Applies low-latency CNN models for image classification of plasma diagnostics streams; supports deployment on embedded platforms.

|                                |   |
|--------------------------------|---|
| <b>date:</b>                   | 2023-12-05  |
| <b>version:</b>                | v1.0  |
| <b>last_updated:</b>           | 2023-12   |
| <b>expired:</b>                | unknown   |
| <b>valid:</b>                  | yes   |
| <b>valid_date:</b>             | 2023-12-05  |
| <b>url:</b>                    | <a href="https://arxiv.org/abs/2312.00128">https://arxiv.org/abs/2312.00128</a>   |
| <b>doi:</b>                    | 10.48550/arXiv.2312.00128   |
| <b>domain:</b>                 | - High Energy Physics   |
| <b>focus:</b>                  | Real-time image classification for in-situ plasma diagnostics   |
| <b>keywords:</b>               | - plasma - in-situ vision - real-time ML  |
| <b>licensing:</b>              | Via Fermilab  |
| <b>task_types:</b>             | - Image Classification  |
| <b>ai_capability_measured:</b> | - Real-time diagnostic inference  |
| <b>metrics:</b>                | - Accuracy - FPS  |
| <b>models:</b>                 | - CNN   |
| <b>ml_motif:</b>               | - Classification  |
| <b>type:</b>                   | Model   |
| <b>ml_task:</b>                | - Image Classification  |
| <b>solutions:</b>              | Solution details are described in the referenced paper or repository.   |
| <b>notes:</b>                  | Embedded/deployment details in progress.  |
| <b>contact.name:</b>           | unknown   |
| <b>contact.email:</b>          | unknown   |
| <b>results.links.name:</b>     | ChatGPT LLM   |
| <b>results.links.url:</b>      | <a href="https://docs.google.com/document/d/1EqkRHuQs1yQqMvZs_L6p9JAY2vKX5OCTubzttFBuRoQ/edit?usp=sharing">https://docs.google.com/document/d/1EqkRHuQs1yQqMvZs_L6p9JAY2vKX5OCTubzttFBuRoQ/edit?usp=sharing</a> |
| <b>fair.reproducible:</b>      | in progress   |
| <b>fair.benchmark_ready:</b>   | No  |
| <b>id:</b>                     | in-situ_high-speed_computer_vision  |
| <b>Citations:</b>              | [57]  |

#### Ratings:

| Rating             | Value | Reason   |
|--------------------|-------|--|
| dataset            | 0     | Dataset not provided or described in any formal way                        |
| documentation      | 2     | Some insight via papers, but no working repo, setup, or replication path   |
| metrics            | 2     | Throughput and accuracy mentioned, but not defined or benchmarked          |
| reference_solution | 1     | Prototype CNNs described; no code, baseline, or training details available |
| software           | 1     | No public implementation or containerized setup released                   |
| specification      | 3     | No standardized I/O, latency constraint, or complete framing               |



## References

- [1] T. Nguyen, J. Jewik, H. Bansal, P. Sharma, and A. Grover, *Climatelearn: Benchmarking machine learning for weather and climate modeling*, 2023. arXiv: 2307.01909 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2307.01909>.
- [2] J. Thiyyagalingam, G. von Laszewski, J. Yin, *et al.*, “Ai benchmarking for science: Efforts from the mlcommons science working group,” in *High Performance Computing. ISC High Performance 2022 International Workshops*, H. Anzt, A. Bienz, P. Luszczek, and M. Baboulin, Eds., Cham: Springer International Publishing, 2022, pp. 47–64, ISBN: 978-3-031-23220-6.
- [3] P. Clark, I. Cowhey, O. Etzioni, *et al.*, “Think you have solved question answering? try arc, the ai2 reasoning challenge,” *arXiv:1803.05457v1*, 2018. DOI: 10.48550/arXiv.1803.05457.
- [4] Y. Fang, N. Zhang, Z. Chen, L. Guo, X. Fan, and H. Chen, *Domain-agnostic molecular generation with chemical feedback*, 2024. arXiv: 2301.11259 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2301.11259>.
- [5] W. Hu, M. Fey, M. Zitnik, *et al.*, *Open graph benchmark: Datasets for machine learning on graphs*, 2021. arXiv: 2005.00687 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2005.00687>.
- [6] H. Zhang, J. Sun, R. Chen, *et al.*, “Empowering and assessing the utility of large language models in crop science,” in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. [Online]. Available: <https://openreview.net/forum?id=hMj6jZ6JWU>.
- [7] M. Tian, L. Gao, S. D. Zhang, *et al.*, *Scicode: A research coding benchmark curated by scientists*, 2024. arXiv: 2407.13168 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2407.13168>.
- [8] C. Krause, M. F. Giannelli, G. Kasiczka, *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>.
- [9] 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>.
- [10] Y. Wang, T. Wang, Y. Zhang, *et al.*, “Urbandalayer: A unified data pipeline for urban science,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, *et al.*, Eds., vol. 37, Curran Associates, Inc., 2024, pp. 7296–7310. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/0db7f135f6991e8cec5e516ecc66bfba - Paper - Datasets\\_and\\_Benchmarks\\_Track.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/0db7f135f6991e8cec5e516ecc66bfba-Paper-Datasets_and_Benchmarks_Track.pdf).
- [11] S. Pramanick, R. Chellappa, and S. Venugopalan, *Spiqa: A dataset for multimodal question answering on scientific papers*, 2025. arXiv: 2407.09413 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2407.09413>.
- [12] A. Karargyris, R. Umeton, M. J. Sheller, *et al.*, “Federated benchmarking of medical artificial intelligence with medperf,” *Nature Machine Intelligence*, vol. 5, no. 7, pp. 799–810, Jul. 2023. DOI: 10.1038/s42256-023-00652-2. [Online]. Available: <https://doi.org/10.1038/s42256-023-00652-2>.
- [13] 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>.
- [14] D. Zou, S. Liu, S. Miao, V. Fung, S. Chang, and P. Li, “Gess: Benchmarking geometric deep learning under scientific applications with distribution shifts,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, *et al.*, Eds., vol. 37, Curran Associates, Inc., 2024, pp. 92499–92528. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/a8063075b00168dc39bc81683619f1a8 - Paper - Datasets\\_and\\_Benchmarks\\_Track.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/a8063075b00168dc39bc81683619f1a8-Paper-Datasets_and_Benchmarks_Track.pdf).

- [15] L. Chanussot, A. Das, S. Goyal, *et al.*, “The open catalyst 2020 (oc20) dataset and community challenges,” *ACS Catalysis*, vol. 11, no. 10, pp. 6059–6072, 2021. DOI: 10.1021/acscatal.0c04525. [Online]. Available: <https://pubs.acs.org/doi/10.1021/acscatal.0c04525>.
- [16] R. Tran, J. Lan, M. Shuaibi, *et al.*, “The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts,” *ACS Catalysis*, vol. 13, no. 5, pp. 3066–3084, 2023. DOI: 10.1021/acscatal.2c05426. [Online]. Available: <https://pubs.acs.org/doi/10.1021/acscatal.2c05426>.
- [17] L. Chanussot, A. Das, S. Goyal, *et al.*, “Open catalyst 2020 (oc20) dataset and community challenges,” *ACS Catalysis*, vol. 11, no. 10, pp. 6059–6072, 2021. DOI: 10.1021/acscatal.0c04525. eprint: <https://doi.org/10.1021/acscatal.0c04525>. [Online]. Available: <https://doi.org/10.1021/acscatal.0c04525>.
- [18] R. Tran, J. Lan, M. Shuaibi, *et al.*, “The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts,” *ACS Catalysis*, vol. 13, no. 5, pp. 3066–3084, Feb. 2023, ISSN: 2155-5435. DOI: 10.1021/acscatal.2c05426. [Online]. Available: <http://dx.doi.org/10.1021/acscatal.2c05426>.
- [19] 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>.
- [20] 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>.
- [21] 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>.
- [22] 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>.
- [23] P. Chen, L. Peng, R. Jiao, *et al.*, “Learning superconductivity from ordered and disordered material structures,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, *et al.*, Eds., vol. 37, Curran Associates, Inc., 2024, pp. 108902–108928. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/c4e3b55ed4ac9ba52d7df11f8bddbf4-Paper-Datasets\\_and\\_Benchmarks\\_Track.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/c4e3b55ed4ac9ba52d7df11f8bddbf4-Paper-Datasets_and_Benchmarks_Track.pdf).
- [24] E. Luo, J. Jia, Y. Xiong, *et al.*, *Benchmarking ai scientists in omics data-driven biological research*, 2025. arXiv: 2505.08341 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2505.08341>.
- [25] R. Ohana, M. McCabe, L. Meyer, *et al.*, “The well: A large-scale collection of diverse physics simulations for machine learning,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, *et al.*, Eds., vol. 37, Curran Associates, Inc., 2024, pp. 44989–45037. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/4f9a5acd91ac76569f2fe291b1f4772b-Paper-Datasets\\_and\\_Benchmarks\\_Track.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/4f9a5acd91ac76569f2fe291b1f4772b-Paper-Datasets_and_Benchmarks_Track.pdf).
- [26] D. Hendrycks, C. Burns, and S. Kadavath, *Measuring massive multitask language understanding*, 2021. [Online]. Available: <https://arxiv.org/abs/2009.03300>.
- [27] J. Roberts, K. Han, and S. Albanie, “Satin: A multi-task metadataset for classifying satellite imagery using vision-language models,” *ICCV Workshop: Towards the Next Generation of Computer Vision Datasets*, Mar. 2023. DOI: 10.48550/arXiv.2304.11619.
- [28] D. Rein, B. L. Hou, and A. C. Stickland, *Gpqa: A graduate-level google-proof q and a benchmark*, 2023. [Online]. Available: <https://arxiv.org/abs/2311.12022>.
- [29] H. Lightman, V. Kosaraju, Y. Burda, *et al.*, “Let’s verify step by step,” *arXiv preprint arXiv:2305.20050*, 2023. DOI: 10.48550/arXiv.2305.20050. eprint: [arXiv:2305.20050](https://arxiv.org/abs/2305.20050).

- [30] N. Mudur, H. Cui, S. Venugopalan, P. Raccuglia, M. P. Brenner, and P. Norgaard, *Feabench: Evaluating language models on multiphysics reasoning ability*, 2025. arXiv: 2504.06260 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2504.06260>.
- [31] 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>.
- [32] K. Khrabrov, A. Ber, A. Tsypin, *et al.*, *Delta-squared dft: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials*, 2024. arXiv: 2406.14347 [physics.chem-ph]. [Online]. Available: <https://arxiv.org/abs/2406.14347>.
- [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] R. E. Peterson, A. Tanelus, C. Ick, *et al.*, “Vocal call locator benchmark (vcl) for localizing rodent vocalizations from multi-channel audio,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, *et al.*, Eds., vol. 37, Curran Associates, Inc., 2024, pp. 106370–106382. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/c00d37d6b04d73b870b963a4d70051c1-Paper-Datasets\\_and\\_Benchmarks\\_Track.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/c00d37d6b04d73b870b963a4d70051c1-Paper-Datasets_and_Benchmarks_Track.pdf).
- [35] R. Bushuiev, A. Bushuiev, N. F. de Jonge, *et al.*, “Massspecgym: A benchmark for the discovery and identification of molecules,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, *et al.*, Eds., vol. 37, Curran Associates, Inc., 2024, pp. 110010–110027. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/c6c31413d5c53b7d1c343c1498734b0f-Paper-Datasets\\_and\\_Benchmarks\\_Track.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/c6c31413d5c53b7d1c343c1498734b0f-Paper-Datasets_and_Benchmarks_Track.pdf).
- [36] X. Zhong, Y. Gao, and S. Gururangan, *Spiqa: Scientific paper image question answering*, 2024. [Online]. Available: <https://arxiv.org/abs/2407.09413>.
- [37] D. Rein, B. L. Hou, A. C. Stickland, *et al.*, *Gpqa: A graduate-level google-proof q and a benchmark*, 2023. arXiv: 2311.12022 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2311.12022>.
- [38] D. Jin, E. Pan, N. Oufattolle, W.-H. Weng, H. Fang, and P. Szolovits, *What disease does this patient have? a large-scale open domain question answering dataset from medical exams*, 2020. arXiv: 2009.13081 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2009.13081>.
- [39] 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>.
- [40] Y. Luo, Y. Chen, and Z. Zhang, *Cfdbench: A large-scale benchmark for machine learning methods in fluid dynamics*, 2024. [Online]. Available: <https://arxiv.org/abs/2310.05963>.
- [41] H. Cui, Z. Shamsi, G. Cheon, *et al.*, *Curie: Evaluating llms on multitask scientific long context understanding and reasoning*, 2025. arXiv: 2503.13517 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2503.13517>.
- [42] 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>.
- [43] T. Arrestad, E. Govorkova, J. Ngadiuba, E. Puljak, M. Pierini, and K. A. Wozniak, *Unsupervised new physics detection at 40 mhz: Training dataset*, 2021. DOI: 10.5281/ZENODO.5046389. [Online]. Available: <https://zenodo.org/record/5046389>.
- [44] J. Bowles, S. Ahmed, and M. Schuld, *Better than classical? the subtle art of benchmarking quantum machine learning models*, 2024. arXiv: 2403.07059 [quant-ph]. [Online]. Available: <https://arxiv.org/abs/2403.07059>.
- [45] P. Odagiu, Z. Que, J. Duarte, *et al.*, *Ultrafast jet classification on fpgas for the hl-lhc*, 2024. DOI: <https://doi.org/10.1088/2632-2153/ad5f10>. arXiv: 2402.01876 [hep-ex]. [Online]. Available: <https://arxiv.org/abs/2402.01876>.

- [46] Z. Liu, H. Sharma, J.-S. Park, *et al.*, *Bragggnn: Fast x-ray bragg peak analysis using deep learning*, 2021. arXiv: 2008.08198 [eess.IV]. [Online]. Available: <https://arxiv.org/abs/2008.08198>.
- [47] 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>.
- [48] 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>.
- [49] 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>.
- [50] 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>.
- [51] 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>.
- [52] A. A. Abud, B. Abi, R. Acciarri, *et al.*, *Deep underground neutrino experiment (dune) near detector conceptual design report*, 2021. arXiv: 2103.13910 [physics.ins-det]. [Online]. Available: <https://arxiv.org/abs/2103.13910>.
- [53] E. Glazer, E. Erdil, T. Besiroglu, *et al.*, *Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai*, 2024. arXiv: 2411.04872 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2411.04872>.
- [54] TBD, *Aime*, [Online accessed 2025-06-24], Mar. 2025. [Online]. Available: <https://www.vals.ai/benchmarks/aime-2025-03-13>.
- [55] M. Khan, S. Krave, V. Marinozzi, J. Ngadiuba, S. Stoynev, and N. Tran, “Benchmarking and interpreting real time quench detection algorithms,” in *Fast Machine Learning for Science Conference 2024*, Purdue University, IN: indico.cern.ch, Oct. 2024. [Online]. Available: [https://indico.cern.ch/event/1387540/contributions/6153618/attachments/2948441/5182077/fast\\_ml\\_magnets\\_2024\\_final.pdf](https://indico.cern.ch/event/1387540/contributions/6153618/attachments/2948441/5182077/fast_ml_magnets_2024_final.pdf).
- [56] A. Jain, S. P. Ong, G. Hautier, *et al.*, “The materials project: A materials genome approach,” *APL Materials*, vol. 1, no. 1, 2013. DOI: 10.1063/1.4812323. [Online]. Available: <https://materialsproject.org/>.
- [57] Y. Wei, R. F. Forelli, C. Hansen, *et al.*, *Low latency optical-based mode tracking with machine learning deployed on fpgas on a tokamak*, 2024. DOI: <https://doi.org/10.1063/5.0190354>. arXiv: 2312.00128 [physics.plasm-ph]. [Online]. Available: <https://arxiv.org/abs/2312.00128>.