UNIVERSITY OF TORONTO

Foundation Models for Science Workshop

Hackathon Instructions

Schmidt Sciences

FOUNDATION MODELS for SCIENCE

Hackathon Milestones

- **By End of Day-1:** Know your teammates and start planning your work
- **By Afternoon Break of Day-2:** Finalize your plan and set goals and expectations
- **By end of Day-2:** Preliminary code pushed to GitHub and share your repo in the group chat
- **By Lunch time Day-3:** Discuss your progress with the tutors
- **By Afternoon Break of Day-3:** Work completed, and results finalized
- **Post Afternoon Break of Day-3:** Make 1 slide summary, present the results and prizes!

Goals

Use a foundation model with a dataset to explore one or more of the following ideas:

- Impact of data cleaning
- Effect of fine tuning
- Aleatoric and Epistemic uncertainties of model predictions
- Representations of features inside the model

Choose a project that takes **~5 hours** of teamwork.
Focus on **finishing** one single result/demonstration.

Evaluation Criteria

We are looking for submissions that demonstrate **strong collaboration** and **good scientific communication**, such as:

- **Project documentation**
  - Documentation, website, repository wiki, well documented Jupyter notebooks, commented code, etc.

- **Creative project results**
  - Models and datasets from different scientific domains combined in new and unusual ways, team names, “coolness”, etc.

- **GitHub pull requests and most GitHub commits**
  - Syncing code to GitHub is good practice and helps collaboration. We want to see it! Committing to GitHub tracks the contributions of individuals, this is a great way to highlight your contributions to the team. Share the link to the Github repo on the WhatsApp group!

Evaluation Criteria

We are looking for submissions that demonstrate **strong collaboration** and **good scientific communication**, such as:

- **Collaborations within the workshop**
  - "Cheating" is encouraged. Talk to other hackathon teams, share code and ideas, pull from other people's repos, and cite your sources of inspiration.
- **Visualizations**
  - You will have one (1) slide to present your project; make those figures beautiful and meaningful to our diverse audience.
- **60 Second Summary**
  - Sometimes you only have a few seconds to get your point across. Your presentation is a good opportunity to practice this.

Computing Support

A Jupyter-based cloud instance:

- Individual links and passwords will be shared.
- Environment already setup to run tutorials. You can install any additional software for your hackathon project.
- 1x A100 40 GB,14 vCPUs, 110 GB RAM, 1.7 TB storage.
- More **details in Tutorial-1**.
- Reach out to Alex Olson in person or via “Tech Support” group chat for any technical assistance.
- **Backup all code and data at the end of Day-3!**

Alex Olson

We acknowledge the support of Denvr Dataworks for providing cloud computing services for this workshop.

DENVR

Suggested Resources

Materials datasets from HF for hackathon

- https://huggingface.co/datasets/SciKnowOrg/ontolearner-materials_science_and_engineering
- https://huggingface.co/datasets/cgeorgiaw/WyFormer-Symmetric-Crystals
- https://huggingface.co/blog/lematerial
- https://huggingface.co/datasets/kanhatakeyama/material-properties
- https://huggingface.co/datasets/Allanatrix/Materials

HuggingFace foundation models for materials

- https://huggingface.co/collections/ibm/materials-673465deacbdf38d9c0c6303

(Has multiple FMs from IBM on materials property prediction tasks based on both SMILES and SELFIES)

- https://huggingface.co/ibm-research/MoLFormer-XL-both-10pct
- https://huggingface.co/datasets/RSE-Group11/Hugging-Ligand-embeddings
- https://huggingface.co/fabikru/MolEncoder
- https://huggingface.co/liuganghuggingface/Llamole-Pretrained-GraphDiT
- https://huggingface.co/MS-ML/SpecTUS_pretrained_only

Suggested Resources

Drug discovery type datasets from HF for hackathon

- https://huggingface.co/datasets/molport/Drug-like-Compound-Library
- https://huggingface.co/datasets/molport/In-stock-Screening-Compound-Database
- https://huggingface.co/datasets/maomlab/TDC
- https://huggingface.co/datasets/OpenMol/DD100
- https://huggingface.co/datasets/jablonkagroup/chempile-lift

Some UQ Resources

- https://www.ibm.com/think/topics/uncertainty-quantification
- A survey of uncertainty in deep neural networks
- ICLR 2025 Workshop: Quantify Uncertainty and Hallucination in Foundation Models
- Aleatoric and epistemic uncertainty: https://arxiv.org/pdf/1910.09457
- Calibration and temperature scaling: https://geoffpleiss.com/blog/nn_calibration.html
- MC dropout: https://arxiv.org/pdf/1506.02142
- Heteroscedastic model with MC dropout: https://arxiv.org/pdf/1703.04977
- Conformal prediction: https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/conformal.pdf
- Conformalized quantile regression: https://arxiv.org/pdf/1905.03222

Suggested Resources

Bio datasets on huggingface

- Uniprot protein sequences: https://huggingface.co/datasets/damlab/uniprot
- Protein sequence to structure: https://huggingface.co/datasets/lamm-mit/protein_secondary_structure_from_PDB
- Protein function: https://huggingface.co/datasets/Protein-FN/Protein-FN
- Enzyme binary classification: https://huggingface.co/datasets/graphs-datasets/PROTEINS
- Proteingym: https://huggingface.co/datasets/ICML2022/ProteinGym
- Protein binding: https://huggingface.co/datasets/ronig/protein_binding_sequences
- cellXgene single cell datasets: https://cellxgene.cziscience.com/datasets
- Gene coding sequence tasks: https://huggingface.co/datasets/InstaDeepAI/true-cds-protein-tasks

Some foundation models for bio

- ESM-2 (protein): https://huggingface.co/facebook/esm2_t36_3B_UR50D
- Prot_bert: https://huggingface.co/Rostlab/prot_bert
- EVO-2 (DNA): https://huggingface.co/arcinstitute/evo2_40b
- Biobert (biomedical LLM): https://huggingface.co/dmis-lab/biobert-v1.1
- scGPT (sc gene expression): https://huggingface.co/tdc/scGPT
- Geneformer (gene expression): https://huggingface.co/ctheodoris/Geneformer

Suggested Resources

Astronomy Datasets

- https://huggingface.co/MultimodalUniverse
- https://dawn-cph.github.io/dja/
- https://datalab.noirlab.edu/

Astronomy Foundation Models

- https://huggingface.co/polymathic-ai/aion-base
- https://huggingface.co/Smith42/astroPT

Physics Foundation Models

- https://huggingface.co/flwi/Physics-Foundation-Model
- https://huggingface.co/sid22669/TinyLlama-Physics

Tabular Foundation Models

- https://huggingface.co/Prior-Labs/TabPFN-v2-clf
- https://huggingface.co/jingang/TabICL-clf

Let’s get started!

FOUNDATION MODELS *for* SCIENCE

Schedule and event details: https://ai-for-science.org