# mlf-core: A framework for deterministic machine learning

Edmund Miller

2021-03-09 Wed

# Introduction

# Reproducing Machine Learning

- Collberg and Proebsting in 2016 evaluated 402 computational experimental papers and could only reproduce 48.3% even when communicating with the authors

# Inspired by nf-core

Correspondence | Published: 13 February 2020

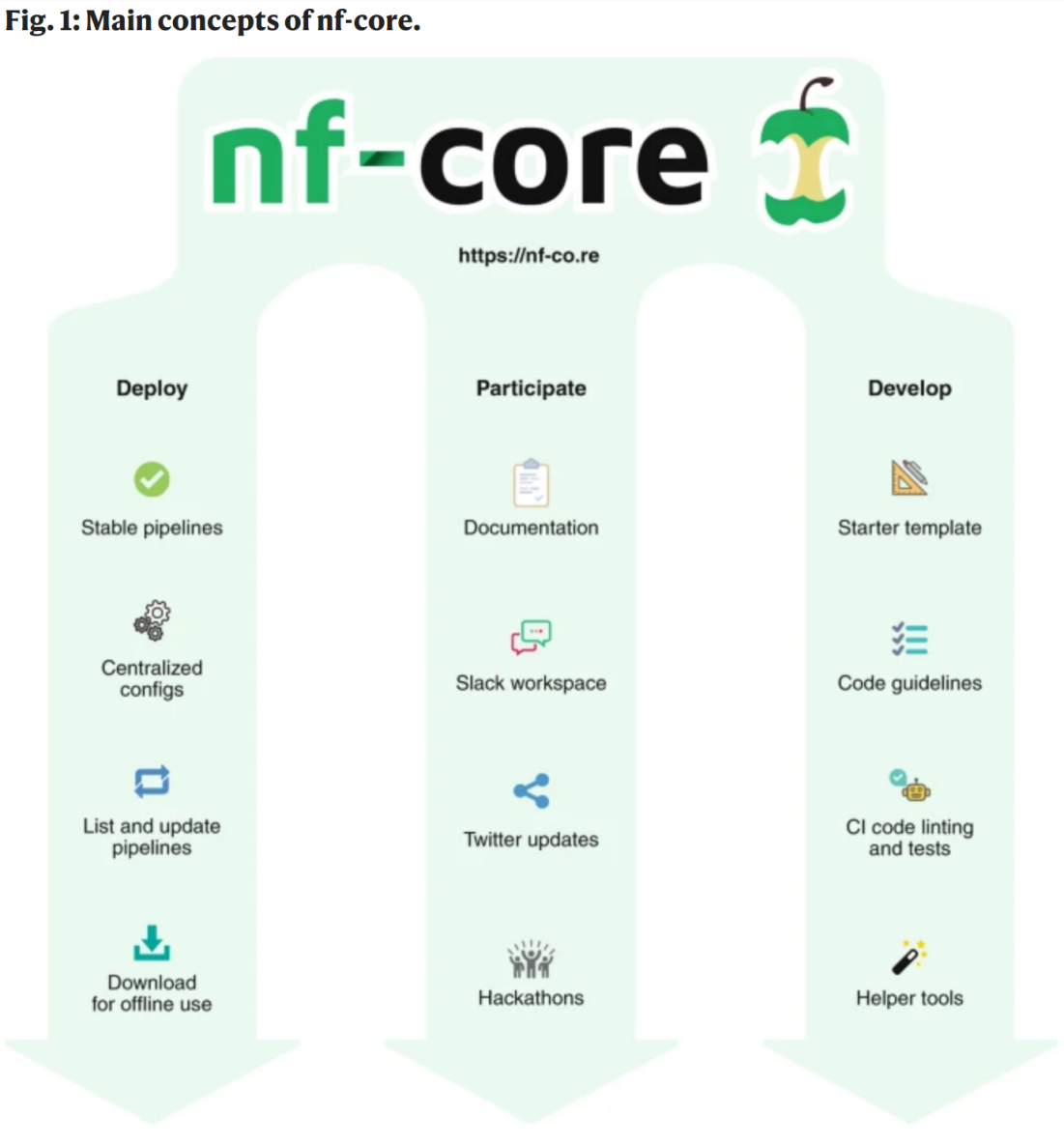
## The nf-core framework for community-curated bioinformatics pipelines

Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso & Sven Nahnsen ✉

*Nature Biotechnology* **38**, 276–278(2020) | Cite this article

# Inspired by nf-core

**Fig. 1: Main concepts of nf-core.**



# MLFlow

**Why not nextflow?**

1. Hyperparameter tracking
2. Experiment grouping
3. Model deplyoment
4. Interactive viz (Tower is for processes)
5. Model repository

   *Just throwing stuff into containers is absolutely not sufficient for reproducible machine learning. Especially with GPUs. Nextflow does (on its own) NOT solve the reproducibility issue of ML.*

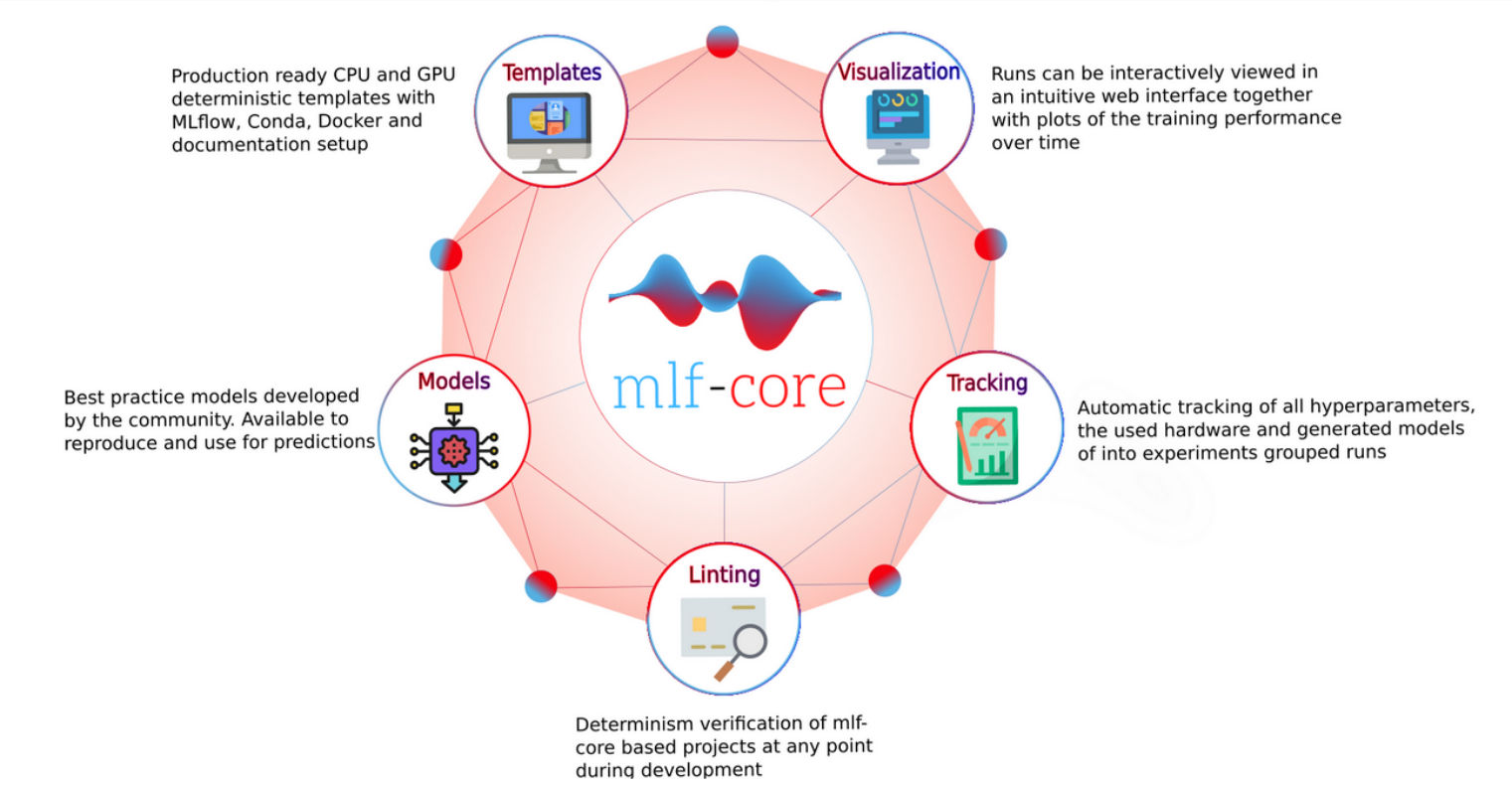
## Comparison of Frameworks

| Framework | Tracking | Visualization | Container | Hardware | Determinism |
|---|---|---|---|---|---|
| Polyaxon | Full | Dashboard | Docker | No | No |
| Guild AI | Full | Dashboard | No | No | No |
| Sacred | Full with dependencies | Dashboard | No | Limited | No |
| MLflow | Full with models | Dashboard | Conda, Docker | No | No |

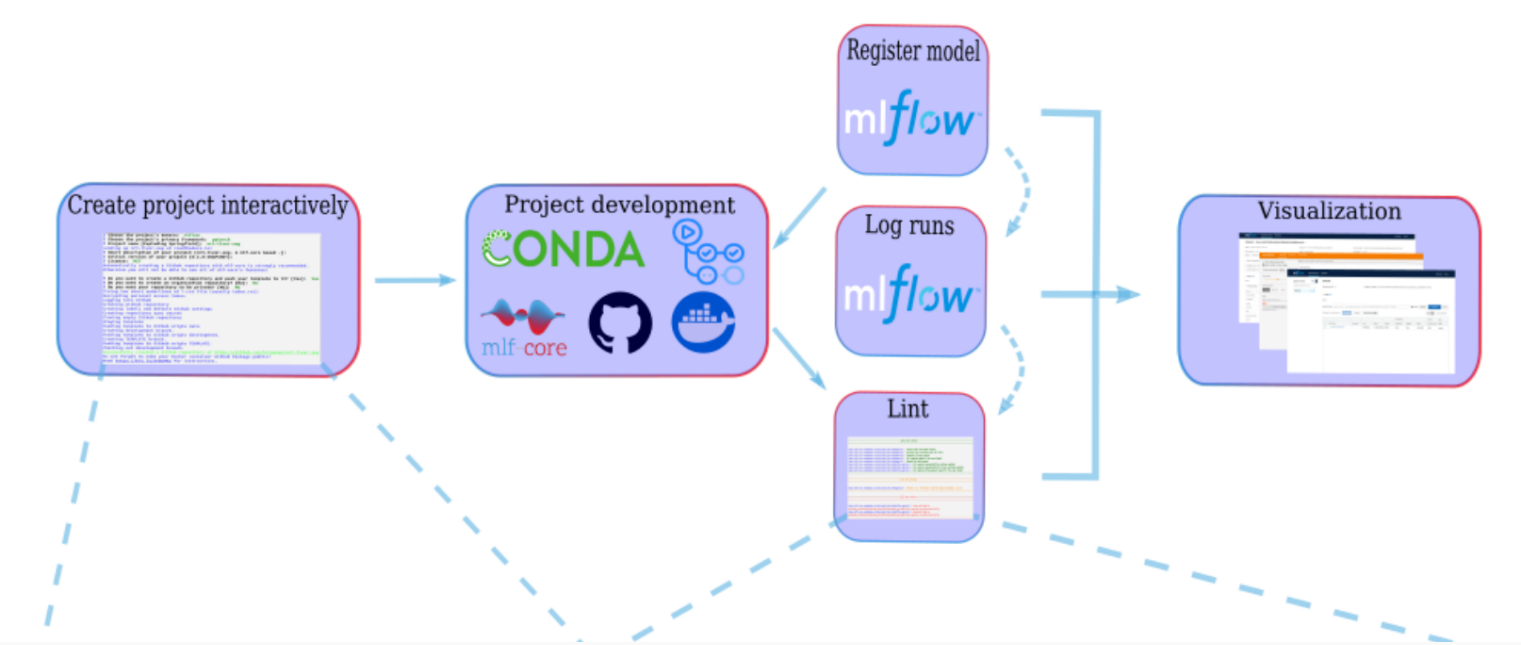
# MLFlow

- Platform to manage ML lifecycle
  - Experimentation Tracking
  - Reproducibility of runs on any platform
  - Deployment of models
  - A central model registry

# mlf-core ecosystem

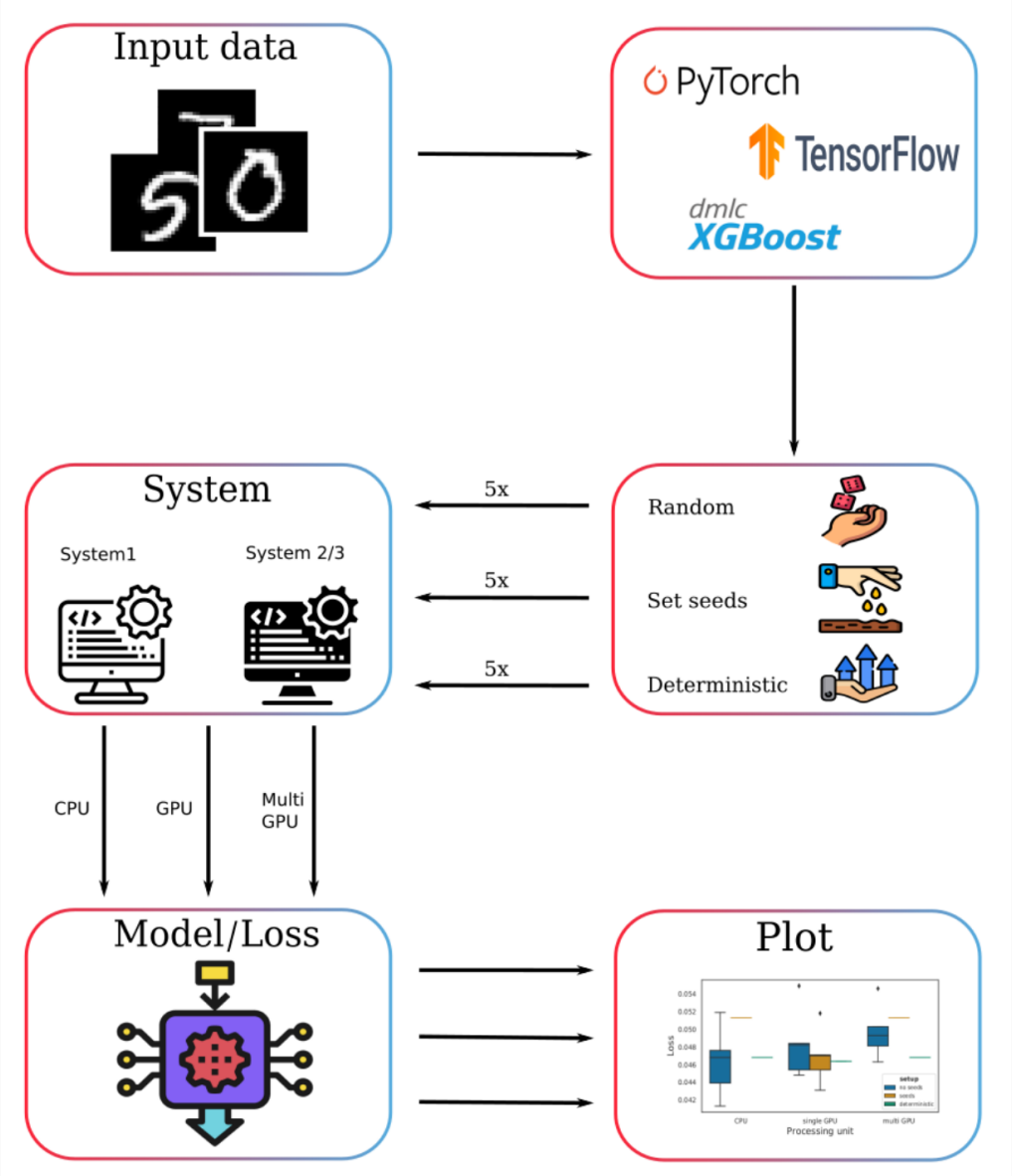


# mlf-core workflow



# Results

# Experimental Setup

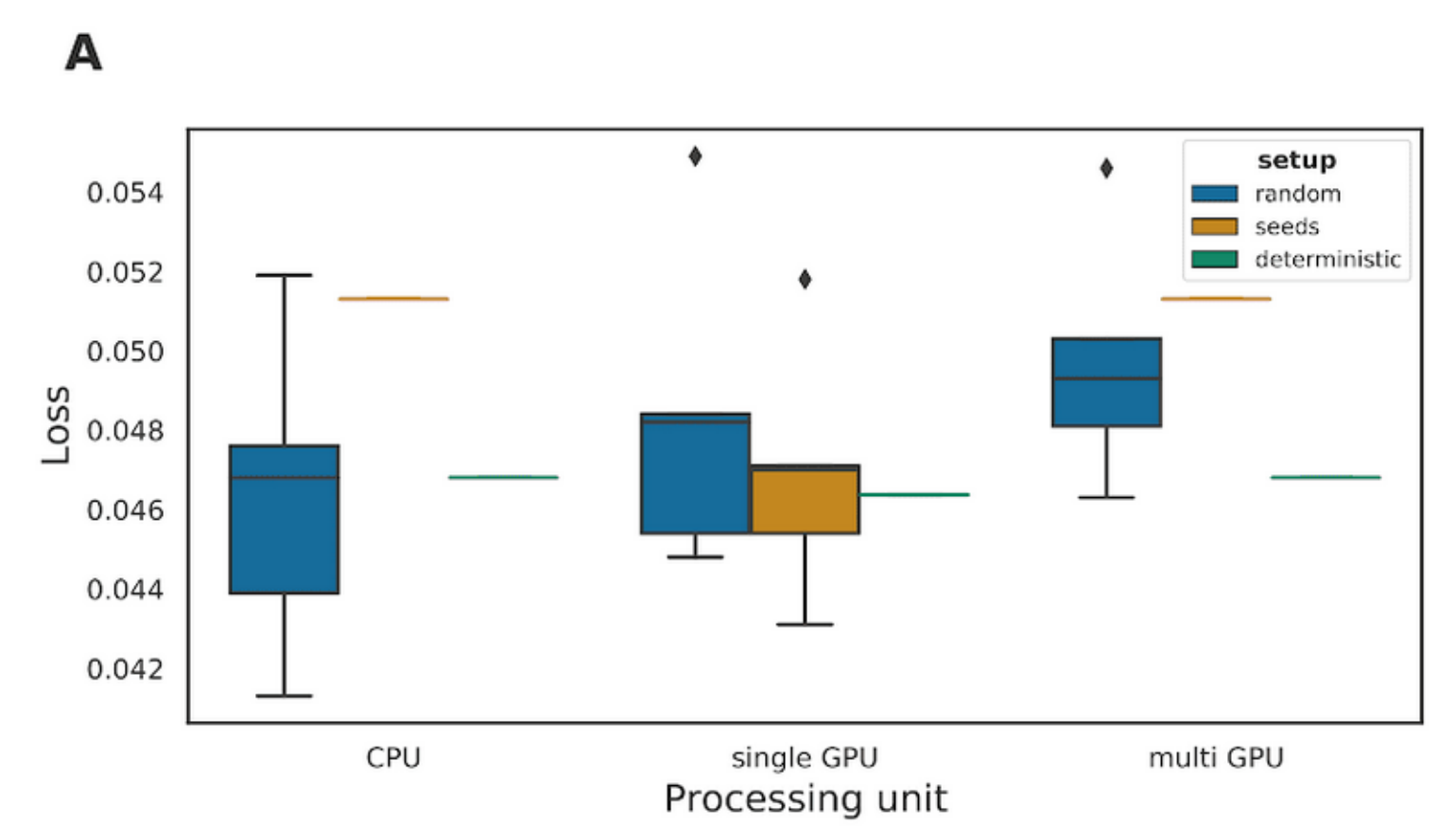


# Example Seed setting for PyTorch for deterministic evalutation

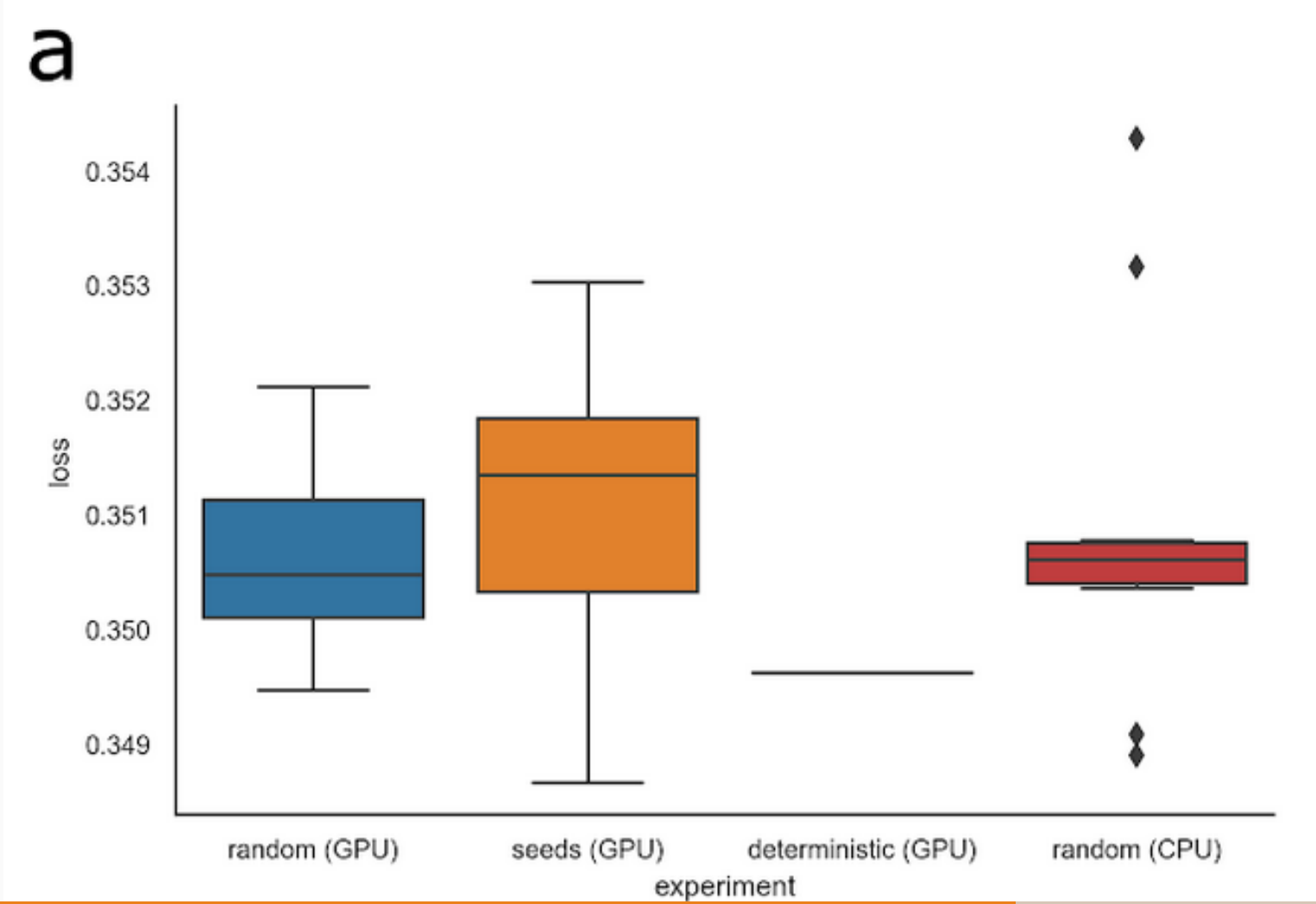
```
import numpy as np
import torch
import os
import random

os.environ["PYTHONHASHSEED"] = SEED
random.seed(SEED)
np.random.seed(SEED)
torch.manual_seed(SEED)
torch.backends.cudnn.deterministic = True
torch.backends.cudnn.benchmark = False
```

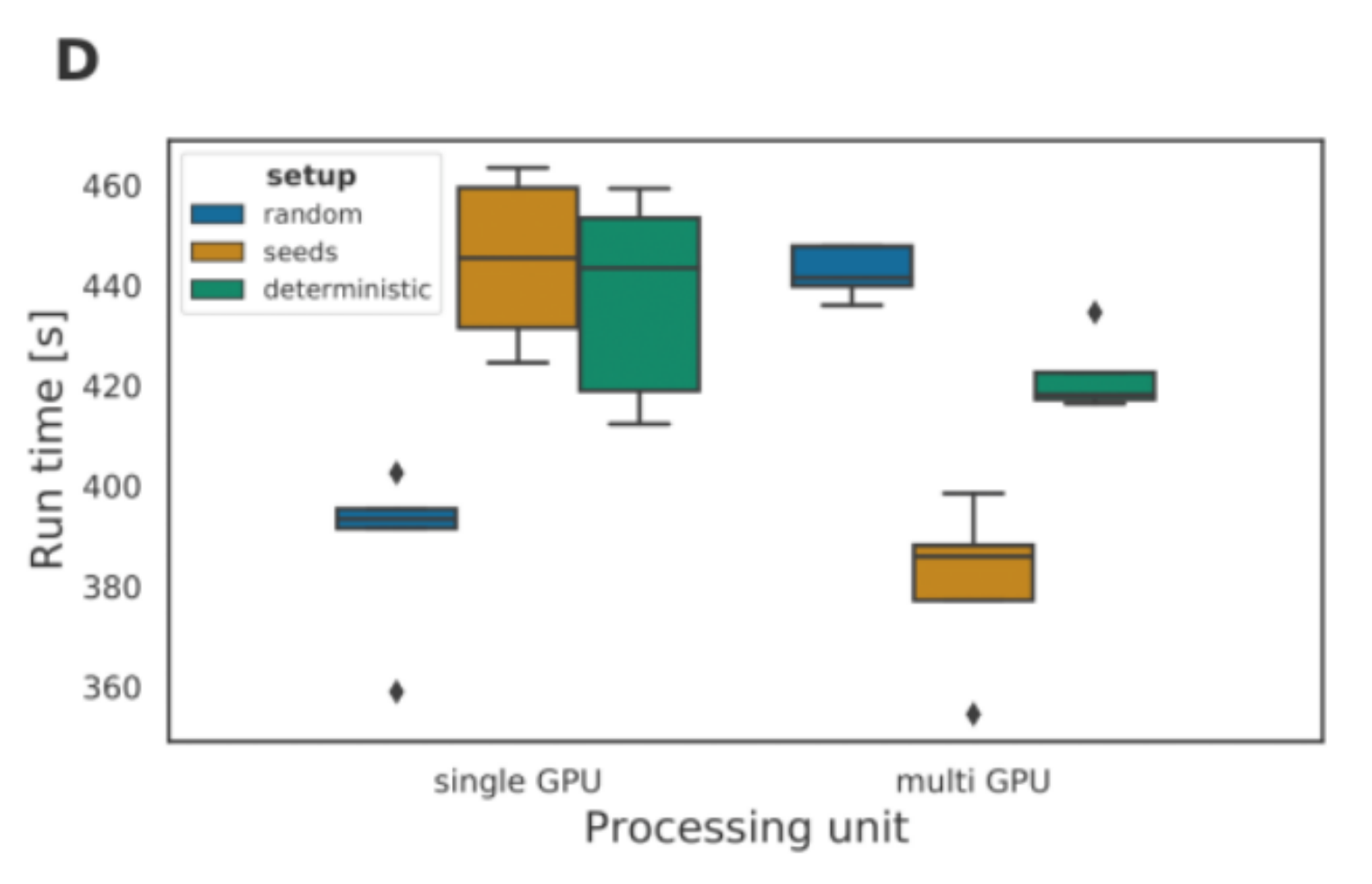
# Determinism evaluation of a convolutional neural network



# Autoencoder model for single-cell RNA-seq data and on an XGboost classification model on a liver cancer dataset



# So what's the catch?



# Tutorial

# GPU Server Gotchas

```
# gpu server specifc stuff
ml load anaconda3
conda create -n mlf-core
# conda activate mlf-core # This won't work!
source activate mlf-core
pip install mlf-core
```

# Creating a project

```
mlf-core list
cd scratch
mlf-core create
```

# Enable Conda

```
cd exploding_springfield
vi MLproject
mlflow run . -A t -A gpus=all -P gpus=2 -P acc=ddp
```

mlf-core/mlf-core#298 MNIST dataset download failes with 403 using torchvisio...

# Call to Action

# Call to Action

- Take Applied Genomics in the Summer!
- Join mlf-core
- Hackathon - March 2021 » nf-core
- #1 AI Conference | GPU Technology Conference | NVIDIA
- SciPy Conference 2021, Austin – Scientific Computing with Python
- Practice Research Computing Box Share