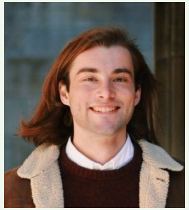
Intuitive Containerization for ML inference with Garden

Will Engler
willengler@uchicago.edu

Accelerate Team

Ari Scourtas



Owen Price Skelly

Ben Galewsky

Logan Ward

Ian Foster

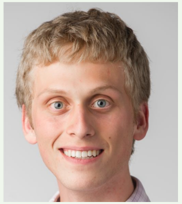
Ben Blaiszik

KJ Schmidt

Will Engler

Nick Saint
Ryan Chard
Kyle Chard
Tyler Skluzacek
Raf Vescovi
Noah Paulson
Isaac Darling
Mark Muchane

Max Tuecke
Phillip Kim
Chase Jenkins
Allison Daemicke
Jennifer Jin

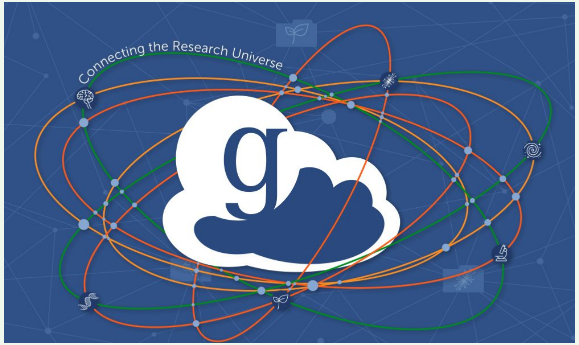


Marcus Schwarting

THE UNIVERSITY OF CHICAGO

NIST

CHiMaD



globus labs

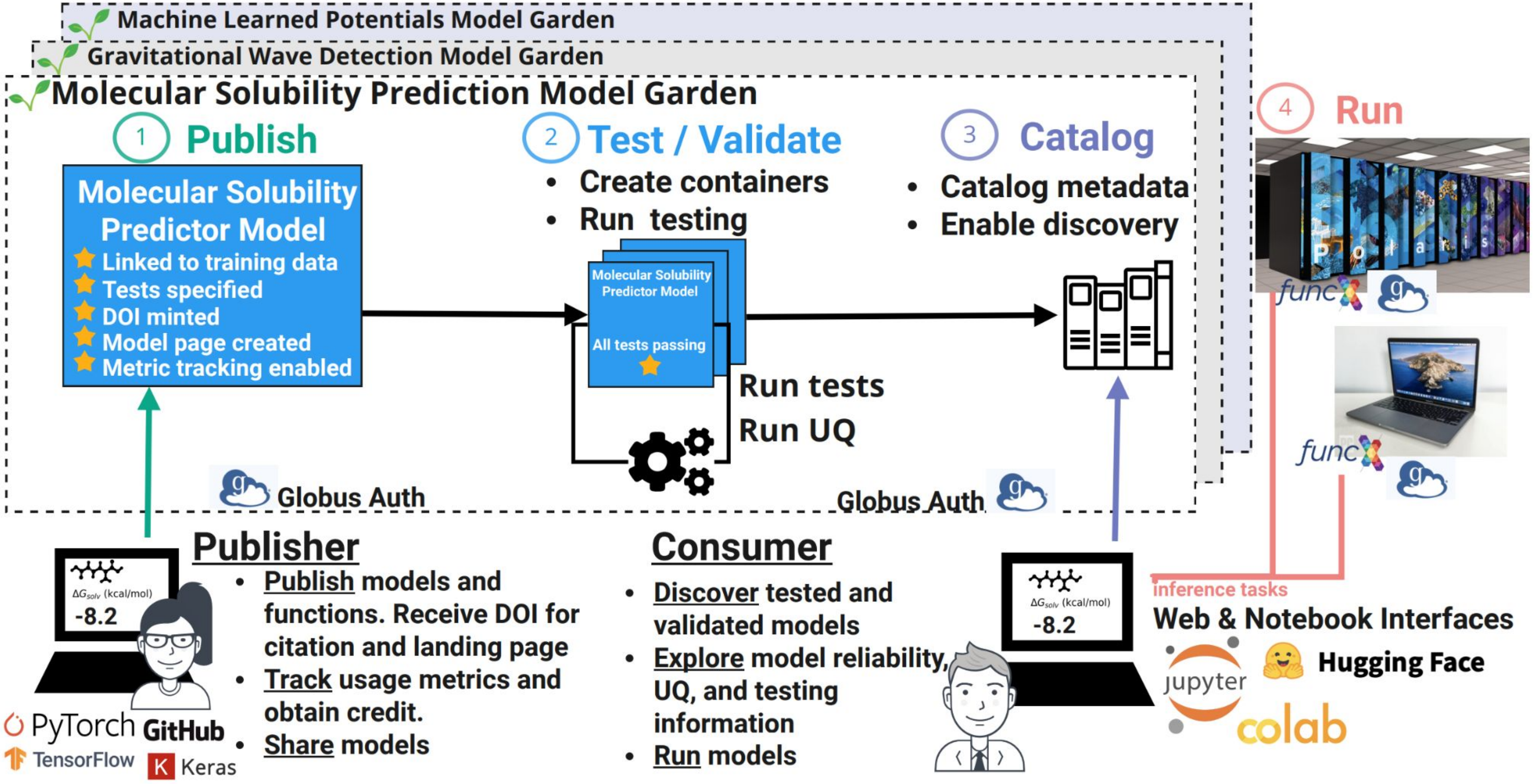
Argonne Leadership Computing Facility

The Advanced Photon Source
a U.S. Department of Energy Office of Science User Facility

3M

NSF

garden



Machine Learned Potentials Model Garden

Gravitational Wave Detection Model Garden

Molecular Solubility Prediction Model Garden

1 Publish

Molecular Solubility Predictor Model

- ⭐ Linked to training data
- ⭐ Tests specified
- ⭐ DOI minted
- ⭐ Model page created
- ⭐ Metric tracking enabled

2 Test / Validate

- Create containers
- Run testing

Molecular Solubility Predictor Model

All tests passing ⭐

Run tests

Run UQ

3 Catalog

- Catalog metadata
- Enable discovery

4 Run

Globus Auth

Globus Auth

Publisher

- Publish models and functions. Receive DOI for citation and landing page
- Track usage metrics and obtain credit.
- Share models

Consumer

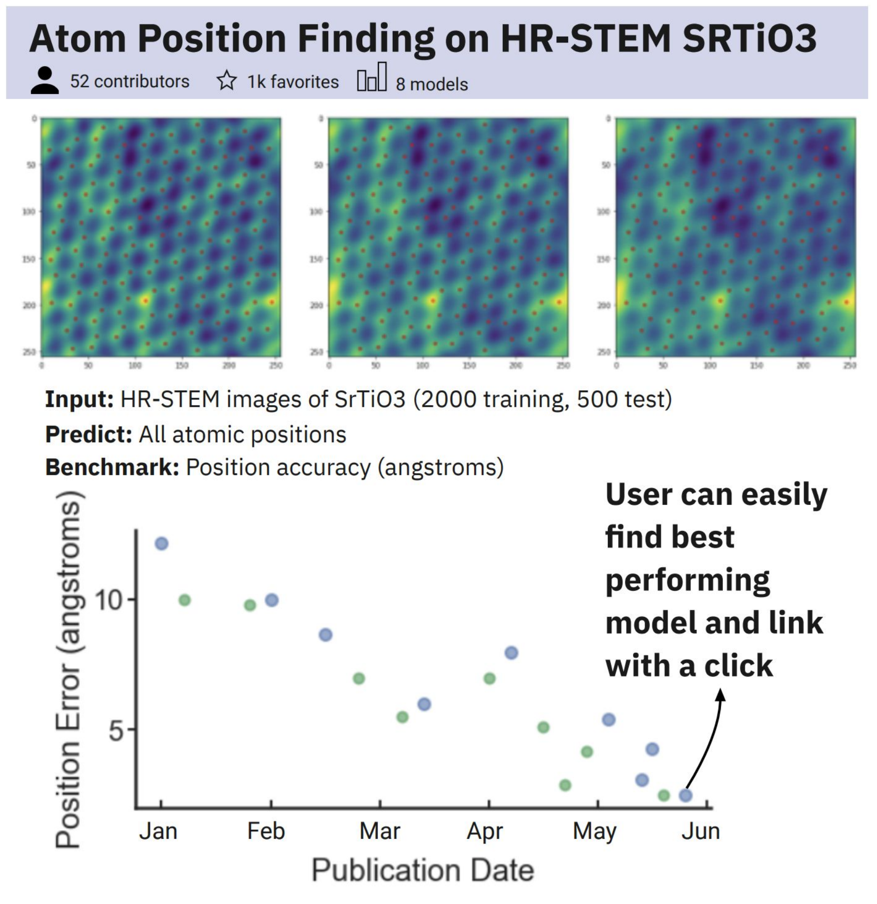
- Discover tested and validated models
- Explore model reliability, UQ, and testing information
- Run models

inference tasks

Web & Notebook Interfaces

Big Plans

- Benchmarking families of related models
- Hosting large models like (Alpha|Open)Fold and LLMs
- Tending Gardens as hubs for different subfields of scientific AI research



Atom Position Finding on HR-STEM SRTiO3

52 contributors ☆ 1k favorites 8 models

Input: HR-STEM images of SrTiO3 (2000 training, 500 test)

Predict: All atomic positions

Benchmark: Position accuracy (angstroms)

User can easily find best performing model and link with a click

Nailing The Basics First

Currently solving for Chris*

- What is Chris* trying to do?
  - Getting models ready for a paper publication
    - Models are small (generally < 100MB)
  - Needs a DOI and metadata for citations
  - Not just citable, runnable
    - Hosted inference API
    - A way to use the models in a production workload

*people who need to translate scientists' GitHub repos into runnable & citable artifacts

Both Sides

1. What does it look like for the consumer?
   a. We have a solid prototype
2. What does it look like for the publisher?
   a. We're iterating on this

Consumer’s POV

- Find a Garden that’s relevant to you
  - Maybe you searched on thegardens.ai
  - Maybe you were linked from a publication
- Try it on your own data with the Garden SDK
  - Pull in a garden by its DOI
  - Calling methods on the garden launches a Globus Compute task that runs the ML function

Publisher’s POV (Chris!)

- Lots of prospective users currently use Colab to release models with papers
    - You can’t mess up your venv
    - You have a tight feedback loop between installing libraries and testing your code

Can We Get Close To That Ease Of Use?

- garden-ai notebook create –python 3.10 –flavor torch
- garden-ai notebook publish my-notebook.ipynb

How Publishing a Notebook Works

- Start: User points to a notebook. End: They see their updated Garden online with a new Globus Compute function attached to it.
- Process
  - Spin up the base container the user specified.
  - Run the contents of the notebook in it. Side effects like library installation are fine.
  - Use dill to save the state of the notebook interpreter in a session.pkl. Save it in the container.
  - Register the container with Globus Compute.
  - Register a function with Globus Compute that uses the container. The function loads the interpreter context and calls the function the user tagged with the @garden_pipeline decorator

Thank You!

MATERIALS DATA FACILITY

https://www.materialsdatafacility.org

CHiMaD NIST

This work was performed under financial assistance award 70NANB14H012 from U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Material Design (CHiMaD). This work was performed under the following financial assistance award 70NANB19H005 from U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD).

DLHub

https://www.dlhub.org

Funding: 2018 Argonne Adv. Computing LDRD

This work was supported in part by Laboratory Directed Research and Development funding from Argonne National Laboratory under U.S. Department of Energy under Contract DE-AC02-06CH11357. We also thank the Argonne Leadership Computing Facility for access to the PetrelKube Kubernetes cluster and Amazon Web Services for providing research credits to enable rapid service prototyping. This research used resources of the Argonne Leadership Computing Facility, a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

funcX

This work was supported by the National Science Foundation under NSF Award Number: **2004894 Collaborative Research: Frameworks: funcX: A Function Execution Service for Portability and Performance**

garden

This work was supported by the National Science Foundation under NSF Award Number: 2209892 "Frameworks: Garden: A FAIR Framework for Publishing and Applying AI Models for Translational Research in Science, Engineering, Education, and Industry".

U.S. DEPARTMENT OF ENERGY

Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

globus labs

Contact: KJ Schmidt (kjschmidt@uchicago.edu)

@kj_schmidt

Thank You

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.