NSF CSSI PI Meeting, Alexandria, VA, July 25-26, 2022

CSSI Framework: Machine Learning Materials Innovation Infrastructure

PI: Dane Morgan[1], Co-PIs: Ryan Jacobs[1], Paul Voyles[1], Michael Ferris[1], Ben Blaiszik[2]

Institutions: [1]University of Wisconsin, Madison, WI, [2]University of Chicago, Chicago, IL

Award #: 1931298

FOUNDRY

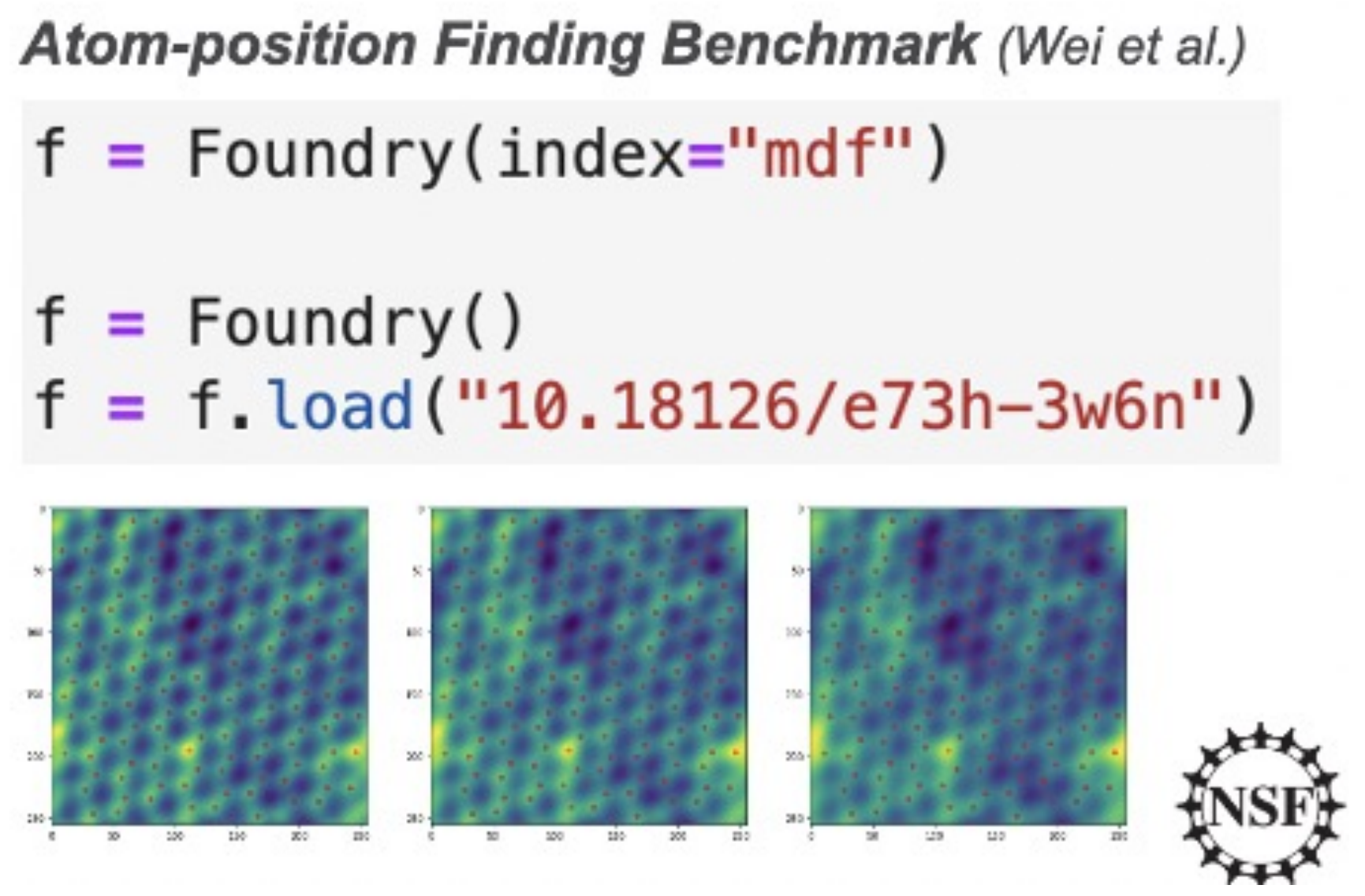
DATA, MODELS, SCIENCE

Goal: Build the *Foundry* to support rapid development of machine learning applications in MS&E[1] through (i) easy access to data, (ii) cloud-based tools for application of ML, and (iii) support for human and machine accessible and sustainable access to disseminated ML models.

Infrastructure for Model Use and Sharing

Foundry is powered by the Data and Learning Hub for Science (DLHub)[2]

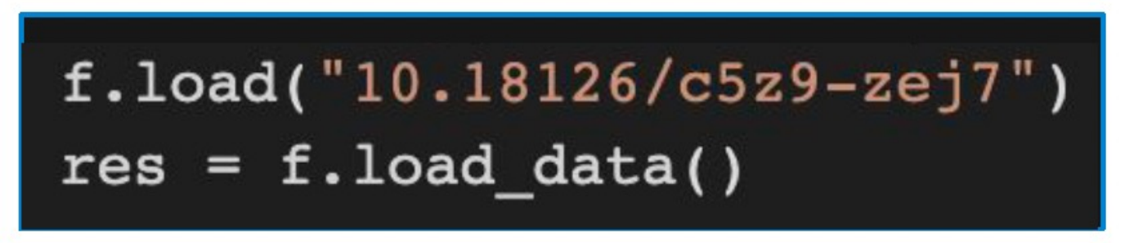
- Collect, publish, categorize models and associated code
- Operate models as a service to simplify sharing, consumption, and access
- Identify models with unique and persistent identifiers (e.g., DOI)
- Implement versioning, search, access controls etc.
- Publicly available: https://github.com/MLMI2-CSSI/foundry


Atom-position Finding Benchmark (Wei et al.)

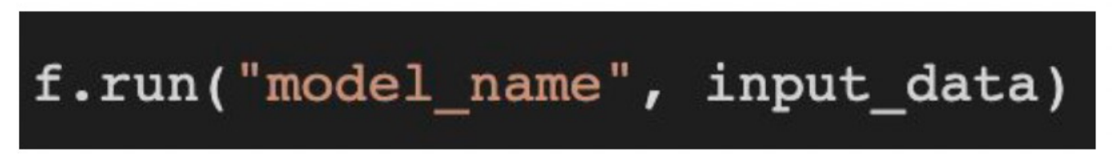
```
f = Foundry(index="mdf")

f = Foundry()
f = f.load("10.18126/e73h-3w6n")
```

Load data in 2 lines:


```
f.load("10.18126/c5z9-zej7")
res = f.load_data()
```

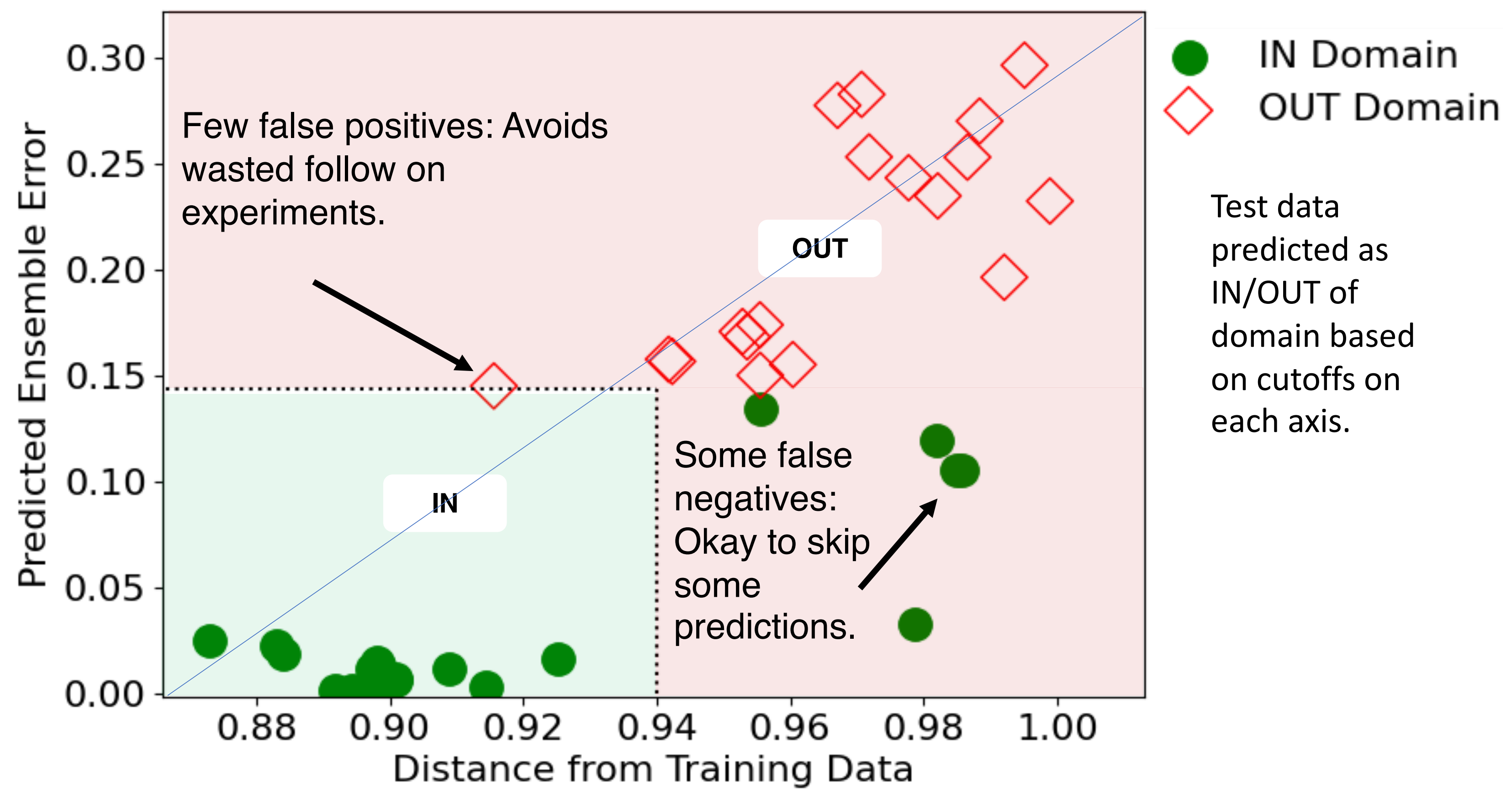
Run model in 1 line:


```
f.run("model_name", input_data)
```

Uncertainty quantification (UQ) – Error and Domain

New Methods for UQ (available now or soon in MAST-ML)

- Validated ensemble methods for error bars[5]
- New method to define and assess if data IN/OUT of domain using ensemble errors and feature space distances

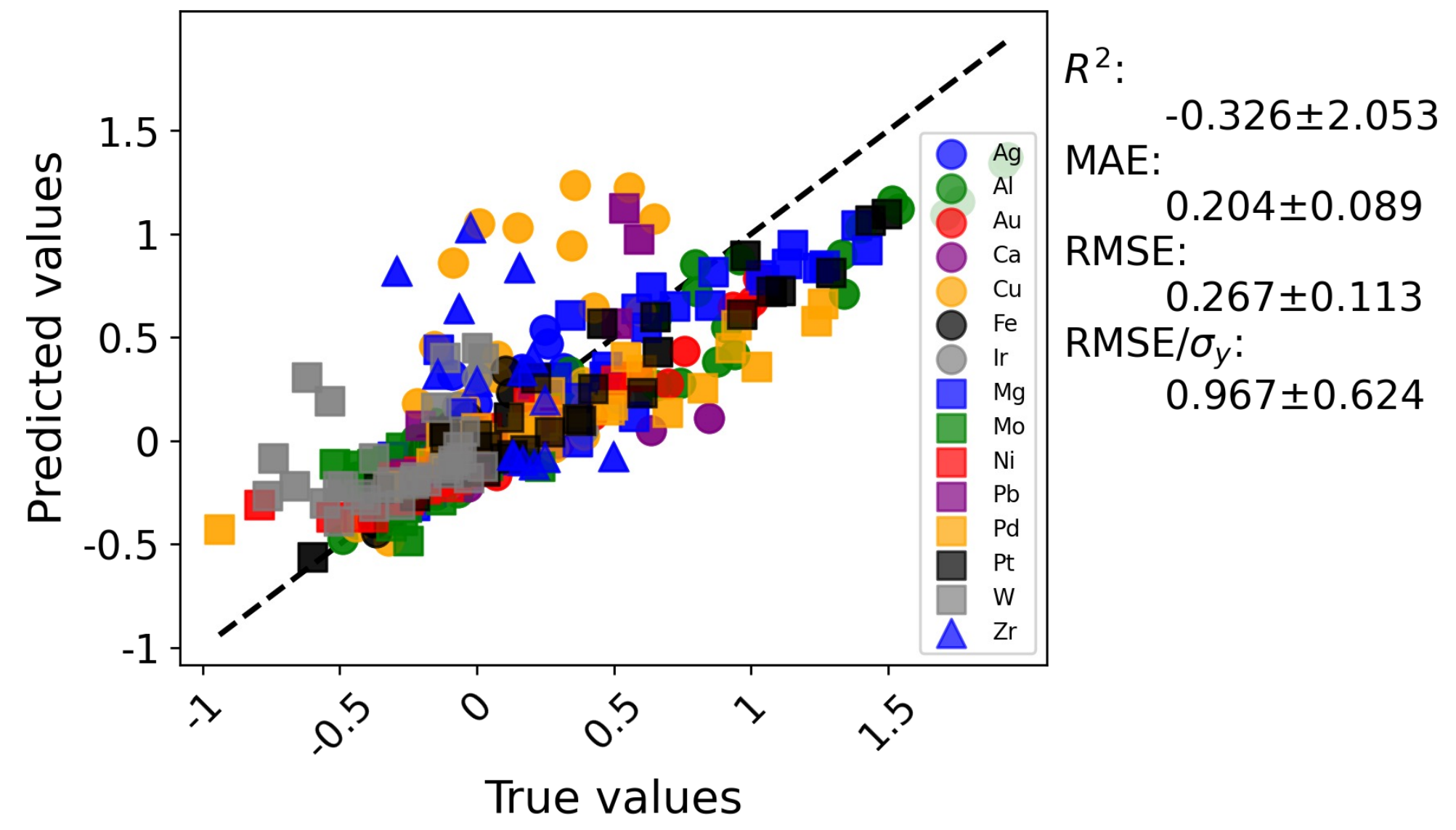


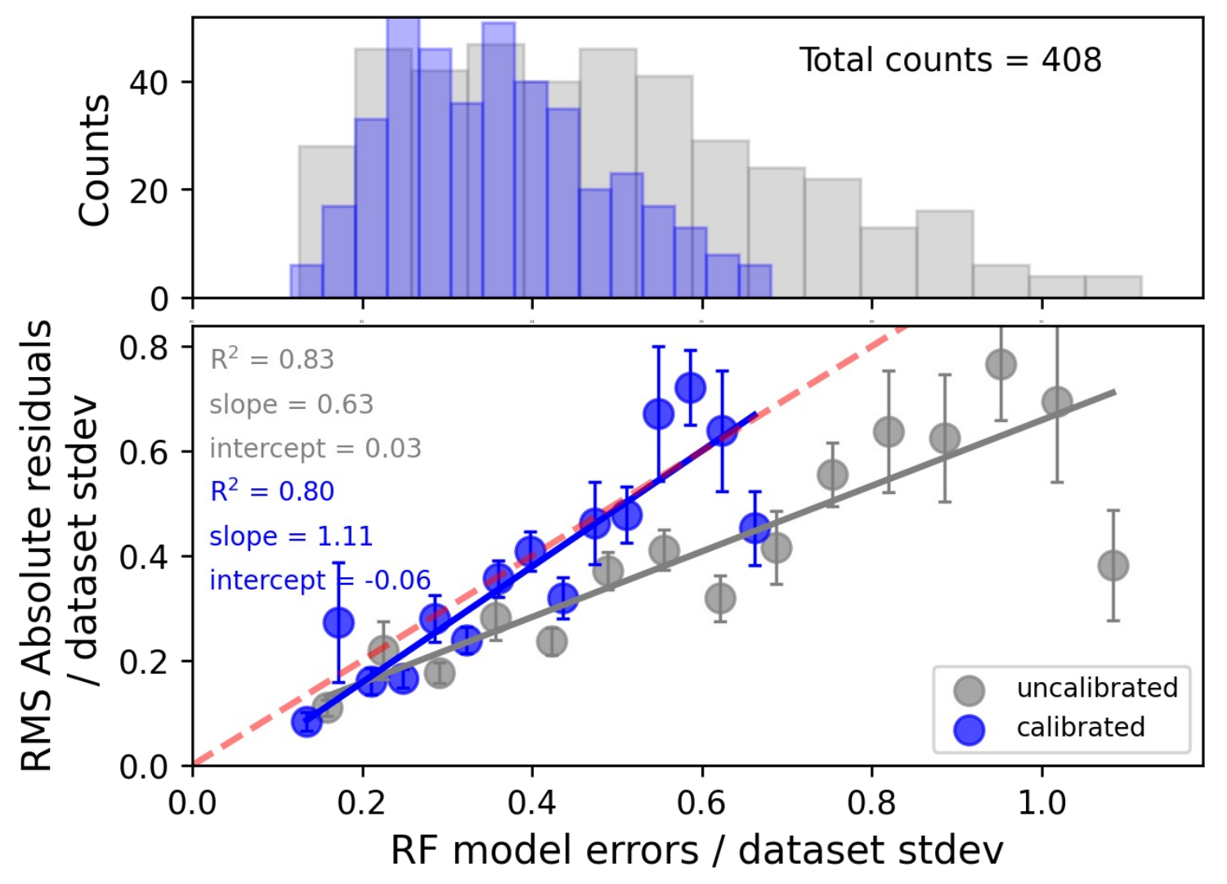
Test data predicted as IN/OUT of domain based on cutoffs on each axis.

Automated Model Generation and Analysis

Foundry leverages the Materials Simulation Toolkit for Machine Learning (MAST-ML)[3]

- Automated tools for materials informatics
- Codifies best practices for development and analysis
- Full suite of Google Colab-ready tutorial notebooks (try now by scanning QR code!)
- Publicly available: https://github.com/uw-cmg/MAST-ML

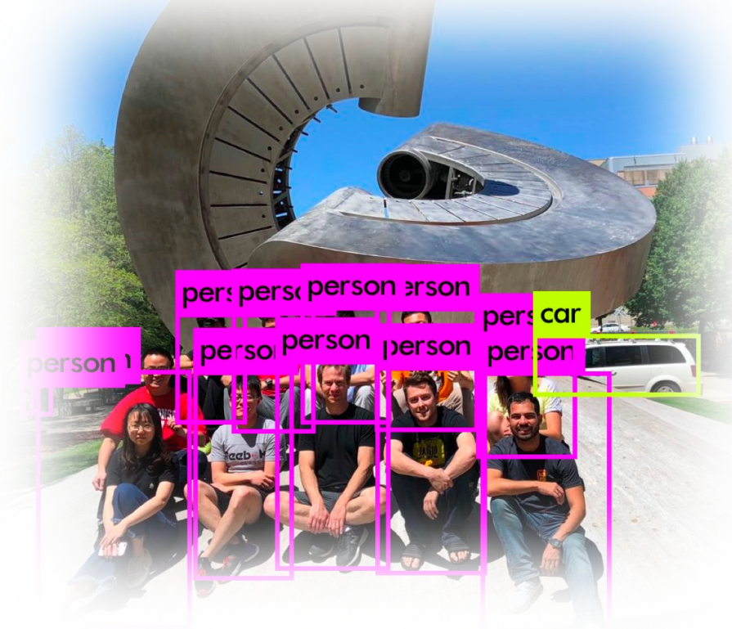




Broader Impact: The Informatics Skunkworks

A program that engages undergraduates in authentic research at the intersection of informatics and science and engineering.

Informatics Skunkworks



Vision: A Community of Practice for Undergraduate Informatics Research

- Accessible software (MAST-ML[3])
- Free online course on *ML for Engineering Research*.
- Community engagement (group activities, Slack groups, mentor-mentee teaming)

Over 350 students since 2015 have found

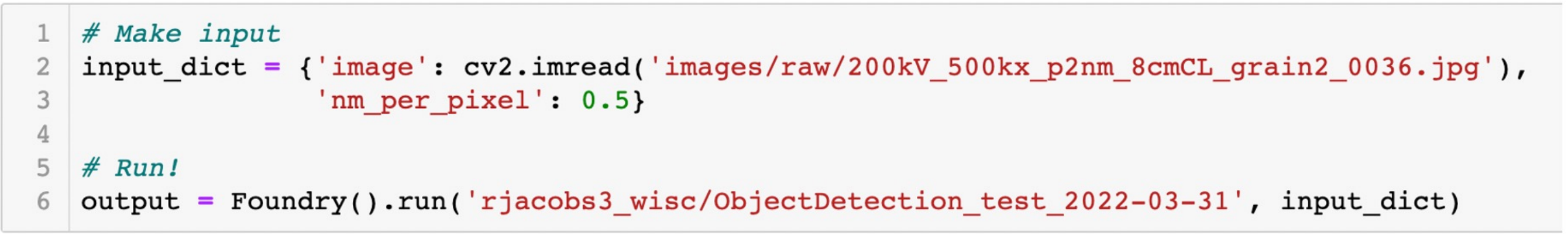
- Key skills: Research, Applied machine learning, Teamwork, Project management, Presentations
- Impactful science: 10 published papers, dozens of presentations
- Prestigious awards: >80% success rate
- Excellent Jobs/Graduate schools: MIT, Carnegie-Mellon, Google, …

https://skunkworks.engr.wisc.edu/

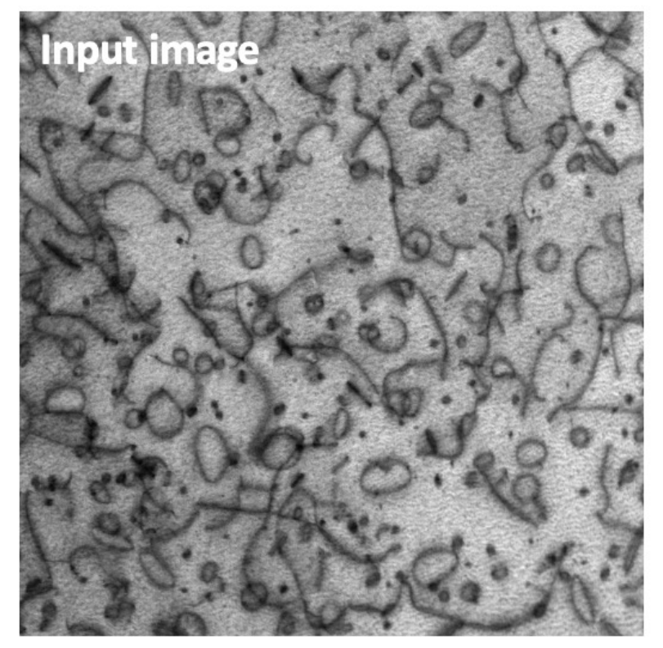
Ready-to-use models with just two lines of python

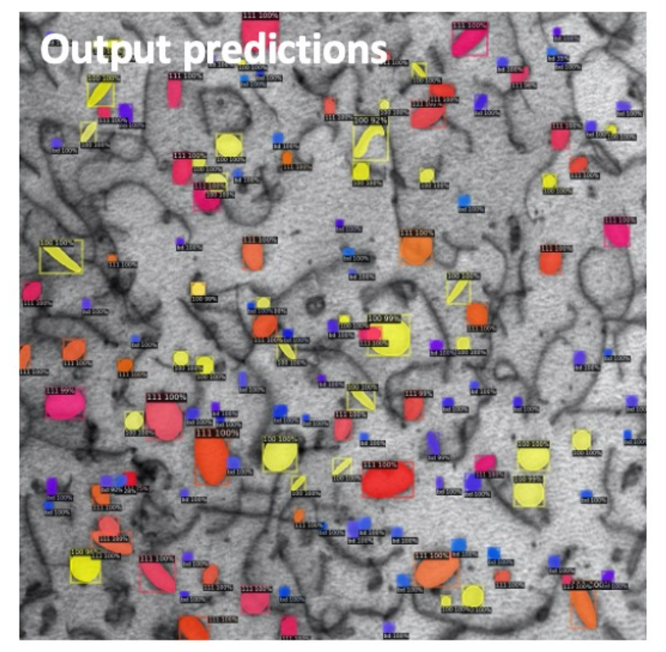
Foundry cloud-based hosted models

- Model predictions with minimal code input
- Generalized containers support wide range of model types (e.g., scikit-learn, PyTorch, Keras, etc.)
- MAST-ML enables easy model upload
- Example of deep learning object detection below[4]


```
# Make input
input_dict = {'image': cv2.imread('images/raw/200kV_500kx_p2nm_8cmCL_grain2_0036.jpg'),
              'nm_per_pixel': 0.5}

# Run!
output = Foundry().run('rjacobs3_wisc/ObjectDetection_test_2022-03-31', input_dict)
```





Impact Areas

The Foundry for Data, Models and Science powered by DLHub and MAST-ML will:

- ***Transform*** the ability of materials researchers to apply machine learning
- ***Accelerate*** material discovery and design; enable new ML research modalities, and ***Support*** users to develop and disseminate ML models in a completely open-source environment
- ***Enable*** ML models to exist in a cloud-based ecosystem, where they can be used, shared, and updated easily through intuitive APIs

References

1. Morgan, D. and Jacobs, R. *Opportunities and Challenges for Machine Learning in Materials Science.* Annual Reviews of Materials Research 50, 71-103 (2020).
2. Chard, R., Li, Z., Chard, K., Ward, L., Babuji, Y., Woodard, A., Tuecke, S., Blaiszik, B., Franklin, M., Foster, I. *DLHub: Model and Data Serving for Science.* 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS) (2019).
3. Jacobs, R., Mayeshiba, T., Afflerbach, B., Miles, L., Williams, M., Turner, M., Finkel, R., Morgan, D. *The Materials Simulation Toolkit for Machine Learning (MAST-ML): an automated open source toolkit to accelerate data-driven materials research.* Computational Materials Science 176, 109544 (2020).
4. Jacobs R., Shen, M., Liu, Y., Hao, W., Li, X., He, R., Greaves, J., Wang, D., Xie, Z., Huang, Z., Wang, C., Field, K., Morgan, D. *Performance and limitations of deep learning semantic segmentation of multiple defects in transmission electron micrographs.* Cell Reports Physical Science 3, 5 (2022).
5. Palmer, G., Du, S., Politowicz, A., Emory, J., Yang, X., Gautam, A., Gupta, G., Li, Z., Jacobs, R., Morgan, D. *Calibration after bootstrap for accurate uncertainty quantification in regression models.* npj Computational Materials 8 (2022)