




# DeepBench: A simulation package for physical benchmarking data

M. Voetberg <sup>1\*</sup>, Ashia Livaudais <sup>1\*</sup>, Becky Nevin <sup>1</sup>, Omari Paul <sup>1</sup>, and Brian Nord <sup>1,2,3</sup>

**1** Fermi National Accelerator Laboratory, P.O. Box 500, Batavia, IL 60510 **2** Department of Astronomy and Astrophysics, University of Chicago, 5801 S Ellis Ave, Chicago, IL 60637 **3** Kavli Institute for Cosmological Physics, University of Chicago, 5801 S Ellis Ave, Chicago, IL 60637 \* These authors contributed equally.

DOI: [10.21105/joss.06774](https://doi.org/10.21105/joss.06774)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

---

Editor: [Paul La Plante](#) 

## Reviewers:

- [@jwuphysics](#)
- [@apoudel2014](#)

Submitted: 05 March 2024

Published: 11 February 2025

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## Summary

We introduce **DeepBench**, a Python library that employs mechanistic models (i.e., analytic mathematical models) to simulate data that represent physics-related objects and systems: geometric shapes (e.g., polygon), physics objects (e.g., pendulum), and astronomical objects (e.g., elliptical galaxy). These data take the form of images (two-dimensional) or time series (one-dimensional). In contrast to natural image benchmarks and complex physics simulations, these data have simple, direct, numerical, and traceable connections between the input data and the label data. When seeking a quantifiable interpretation, this kind of data is uniquely suitable for developing, calibrating, testing, and benchmarking statistical and machine learning models. Finally, this software package includes methods to curate and store these datasets to maximize reproducibility.

## Statement of Need

There are multiple open problems and issues that are critical for the machine learning and scientific communities to address; principally: interpretability, explainability, uncertainty quantification, and inductive bias in machine learning models when they are applied to scientific data. Multiple kinds of datasets and data simulation software packages can be used for developing models and confronting these challenges. These datasets range from natural images and text to multi-dimensional data of physical processes. Indeed, multiple benchmark data and simulation software packages have been created for developing and comparing models.

However, these benchmarks are typically limited in significant ways. Natural image datasets comprising images from the real or natural world (e.g., vehicles, animals, landscapes) are widely used in the development of machine learning models. These kinds of datasets tend to be large, diverse, and carefully curated. However, they are not underpinned by or constructed upon physical principles: they cannot be generated by mathematical expressions of formal physical theory, so there is not a robust connection between the data and a quantitative theory. Therefore, these datasets have a severely limited capacity to help address many questions in machine learning models, such as uncertainty quantification.

On the other hand, complex physics simulations (e.g., cosmological  $N$ -body simulations and particle physics simulators) are accurate, detailed, and based on precise quantitative theories and models. This facilitates studies of interpretability and uncertainty quantification because there is the possibility of linking the simulated data to the input choices through each layer of calculation in the simulator. However, they are relatively small in size and number, and

they are computationally expensive to reproduce. In addition, while they are underpinned by specific physical functions, the complexity of the calculations makes them challenging as a venue through which to make connections between machine learning results and input choices. Complex physics simulations have one or more layers of mechanistic models. Mechanistic models are defined with analytic functions and equations that describe and express components of a given physical process: these are based on theory and empirical observations. In both of these scenarios, it is difficult to build interpretable models that connect raw data and labels, and it is difficult to generate new data rapidly.

The physical sciences community lacks sufficient datasets and software packages as benchmarks for the development of statistical and machine learning models. In particular, there currently does not exist simulation software packages that generates data underpinned by physical principles and that satisfies the following criteria:

- multi-domain
- multi-purpose
- fast
- reproducible
- extensible
- based on mechanistic models
- include detailed noise prescriptions.

## Related Work

There are many benchmarks—both datasets and simulation software packages—widely used for model building in machine learning, statistics, and the physical sciences. First, benchmark datasets of natural images include MNIST (Deng, 2012), CIFAR-10 (Krizhevsky, 2009), and ImageNet (Russakovsky et al., 2014). Second, there are several large astronomical observation datasets, such as the CfA Redshift Survey (Huchra et al., 1983), Sloan Digital Sky Survey (York et al., 2000), and Dark Energy Survey (Abbott et al., 2018). Third, many  $N$ -body cosmology simulation datasets serve as benchmarks, such as the Millennium (Springel, 2005), Illustris (Vogelsberger et al., 2014), EAGLE (Schaye et al., 2015), Coyote (Heitmann et al., 2010), Bolshoi (Klypin et al., 2011), CAMELS (Villaescusa-Navarro et al., 2021), and Quijote (Villaescusa-Navarro et al., 2020) projects. Fourth, there have been multiple astronomy dataset challenges that can be considered benchmarks for analysis comparison: e.g., PLAsTiCC (Hložek et al., 2023), The Great08 Challenge (Bridle et al., 2009), and the Strong Gravitational Lens Challenge (Metcalf et al., 2019). Fifth, there are multiple software packages that generate simulated data for astronomy and cosmology, such as Astropy (The Astropy Collaboration et al., 2013), GalSim (Rowe et al., 2015), lenstronomy (Birrer & Amara, 2018), deepLenstronomy (Morgan et al., 2021), CAMB (Lewis et al., 2000), pixell (Naess et al., 2021), and SOXS (ZuHone et al., 2023). Finally, particle physics projects use standard codebases for simulations, such as Geant4 (Pia et al., 2009), GENIE (Andreopoulos et al., 2015), and PYTHIA (Sjöstrand, 2020). These simulations span wide ranges in speed, code complexity, physical fidelity, and detail. Unfortunately, these datasets and software packages lack a combination of critical features, including mechanistic models, speed, and reproducibility, which are needed for more fundamental studies of statistical and machine learning models. The work in this paper is most closely related to SHAPES (Wu et al., 2016) because that work also uses collections of geometric objects with varying levels of complexity as a benchmark.

## DeepBench Software

The **DeepBench** software package simulates data for analysis tasks that require precise numerical calculations. First, the simulation models are fundamentally mechanistic: they are based on relatively simple analytic mathematical expressions, which are physically meaningful. This means that for each model, the number of input parameters that determine a simulation

output is small ( $<10$  for most models). These elements make the software package fast and the outputs interpretable: they are conceptually and mathematically relatable to the inputs. Second, **DeepBench** also includes methods to precisely prescribe noise for inputs, which are propagated to outputs. This permits studies and the development of statistical inference models that require uncertainty quantification, which is a significant challenge in modern machine learning research. Third, the software framework includes features that permit a high degree of reproducibility: e.g., random seeds at every key stage of input, a unique identification tag for each simulation run, and the tracking and storage of metadata (including input parameters) and the related outputs. Fourth, the primary user interface is a YAML configuration file, which allows the user to specify every aspect of the simulation: e.g., types of objects, numbers of objects, noise type, and number of classes. This feature—which is especially useful when building and studying complex models like deep learning neural networks—permits the user to incrementally decrease or increase the complexity of the simulation with a high level of granularity.

**DeepBench** has the following features:

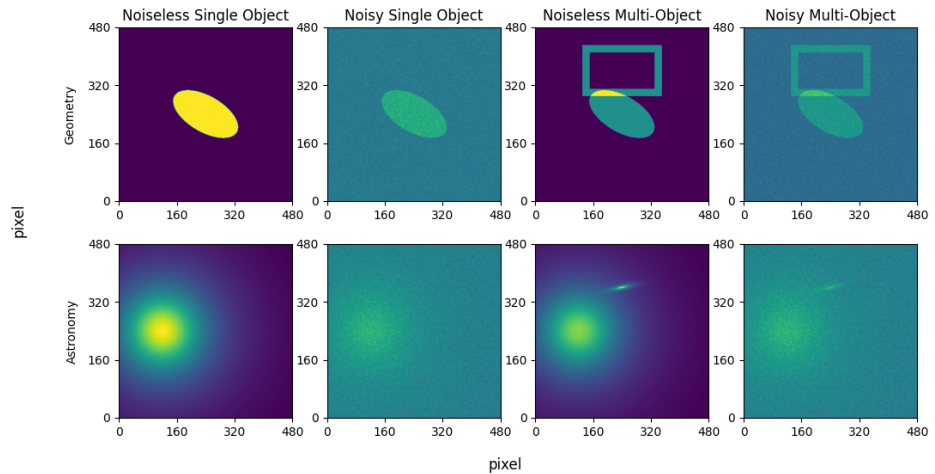
- Exact reproducibility
- Noise and error propagation
- Mechanistic modeling
- Physical sciences-based modeling
- Computational efficiency
- Simulations relevant to multiple domains
- Outputs of varying dimensions
- Readily extensible to new physics and outputs

## Primary Modules

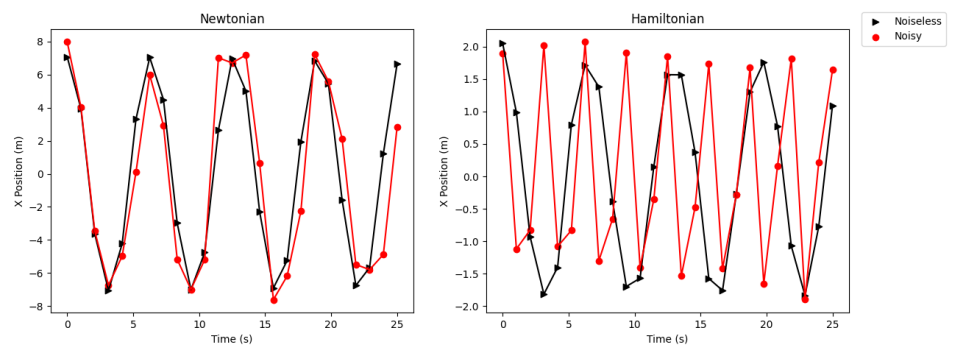
- Geometry objects: two-dimensional images generated with `matplotlib` (Hunter, 2007). The shapes include  $N$ -sided polygons, arcs, straight lines, and ellipses. They are solid, filled or unfilled two-dimensional shapes with edges of variable thickness.
- Physics objects: one-dimensional profiles for two types of implementations of pendulum dynamics: one using Newtonian physics, the other using Hamiltonian.
- Astronomy objects: two-dimensional images generated based on radial profiles of typical astronomical objects. The star object is created using the Moffat distribution provided by the `AstroPy` (The *Astropy Collaboration et al.*, 2013) library. The spiral galaxy object is created with the function used to produce a logarithmic spiral (Ringermacher & Mead, 2009). The elliptical Galaxy object is created using the Sérsic profile provided by the `AstroPy` library. Two-dimensional models are representations of astronomical objects commonly found in data sets used for galaxy morphology classification.
- Image: two-dimensional images that are combinations and/or concatenations of Geometry or Astronomy objects. The combined images are within `matplotlib` `meshgrid` objects. Sky images are composed of any combination of Astronomy objects, while geometric images comprise individual geometric shape objects.
- Collection: Provides a framework for producing module images or objects at once and storing all parameters that were included in their generation, including exact noise levels, object hyper-parameters, and non-specified defaults.

All objects also come with the option to add noise to each object. For physics objects—i.e., the pendulum—the user may add Gaussian noise to parameters: initial angle  $\theta_0$ , the pendulum length  $L$ , the gravitational acceleration  $g$ , the planet properties  $\Phi = (M/r^2)$ , and Newton's gravity constant  $G$ . Note that  $g = G * \Phi = G * M/r^2$ : all parameters in this relationship can receive noise. For astronomy and geometry Objects, which are images, the user can add Poisson or Gaussian noise to the output images. Finally, the user can regenerate the same noise using the saved random seed.

## Example Outputs



**Figure 1:** Example outputs of **DeepBench**, containing geometric and astronomy objects. Variants include a single object, a noisy single object, two objects, and two noisy objects. The geometric outputs are produced with filled ellipses and outlined rectangles, with a gaussian noise overlay for the noisy variants. The astronomy outputs feature a star and an elliptical galaxy profile with similarly applied noise.



**Figure 2:** Example physics simulations from **DeepBench**. Pendulums show noisy and noiseless variants of the Newtonian (left) and Hamiltonian (right) mathematical simulations. Both use initial conditions of an arm length of 10 meters and a starting angle of  $\pi/4$ . The noisy variants introduce uncertainty to these input parameters, along with the measurement of acceleration due to gravity.

## Acknowledgments

*M. Voetberg:* conceptualization, methodology, software, writing, project administration. *Ashia Livaudais:* conceptualization, methodology, software, writing, project administration. *Becky Nevin:* software, project administration. *Omar Paul:* software. *Brian Nord:* conceptualization, methodology, project administration, funding acquisition, supervision, writing.

We acknowledge contributions from Alex Ćiprijanović, Renee Hlozek, Craig Brechmos.

Work supported by the Fermi National Accelerator Laboratory, managed and operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy. The U.S. Government retains and the publisher, by accepting the article for publication,

acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

We acknowledge the Deep Skies Lab as a community of multi-domain experts and collaborators who have facilitated an environment of open discussion, idea-generation, and collaboration. This community was important for the development of this project.

## References

- Abbott, T. M. C., Abdalla, F. B., Allam, S., Amara, A., Annis, J., Asorey, J., Avila, S., Ballester, O., Banerji, M., Barkhouse, W., Baruah, L., Baumer, M., Bechtol, K., Becker, M. R., Benoit-Lévy, A., Bernstein, G. M., Bertin, E., Blazek, J., Bocquet, S., ... Nikutta, R. (2018). The Dark Energy Survey: Data Release 1. *The Astrophysical Journal Supplement Series*, 239(2), 18. <https://doi.org/10.3847/1538-4365/aae9f0>
- Andreopoulos, C., Barry, C., Dytman, S., Gallagher, H., Golan, T., Hatcher, R., Perdue, G., & Yarba, J. (2015). The GENIE Neutrino Monte Carlo Generator: Physics and User Manual. *arXiv e-Prints*, arXiv:1510.05494. <https://doi.org/10.48550/arXiv.1510.05494>
- Birrer, S., & Amara, A. (2018). Lenstronomy: Multi-purpose gravitational lens modelling software package. *Physics of the Dark Universe*, 22, 189–201. <https://doi.org/10.1016/j.dark.2018.11.002>
- Bridle, S., Gill, M., Heavens, A., Heymans, C., High, F. W., Hoekstra, H., Jarvis, M., Kirk, D., Kitching, T., Kneib, J.-P., Kuijken, K., Shave-Taylor, J., Lagatutta, D., Mandelbaum, R., Massey, R., Mellier, Y., Moghaddam, B., Moudden, Y., Nakajima, R., ... Erben, T. (2009). Handbook for the GREAT08 Challenge: An image analysis competition for cosmological lensing. *The Annals of Applied Statistics*, 3(1). <https://doi.org/10.1214/08-aos222>
- Deng, L. (2012). The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Processing Magazine*, 29(6), 141–142. <https://doi.org/10.1109/MSP.2012.2211477>
- Heitmann, K., White, M., Wagner, C., Habib, S., & Higdon, D. (2010). The Coyote Universe. I. Precision Determination of the Nonlinear Matter Power Spectrum. *The Astrophysical Journal*, 715, 104–121. <https://doi.org/10.1088/0004-637X/715/1/104>
- Hložek, R., Malz, A. I., Ponder, K. A., Dai, M., Narayan, G., Ishida, E. E. O., Allam, T., Jr., Bahmanyar, A., Bi, X., Biswas, R., Boone, K., Chen, S., Du, N., Erdem, A., Galbany, L., Garreta, A., Jha, S. W., Jones, D. O., Kessler, R., ... Zuo, W. (2023). Results of the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC). *The Astrophysical Journal Supplement Series*, 267(2), 25. <https://doi.org/10.3847/1538-4365/accd6a>
- Huchra, J., Davis, M., Latham, D., & Tonry, J. (1983). A survey of galaxy redshifts. IV - The data. *The Astrophysical Journal Supplement Series*, 52, 89–119. <https://doi.org/10.1086/190860>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Klypin, A. A., Trujillo-Gomez, S., & Primack, J. (2011). Dark Matter Halos in the Standard Cosmological Model: Results from the Bolshoi Simulation. *The Astrophysical Journal*, 740(2), 102. <https://doi.org/10.1088/0004-637X/740/2/102>
- Krizhevsky, A. (2009). *Learning Multiple Layers of Features from Tiny Images*. <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>
- Lewis, A., Challinor, A., & Lasenby, A. (2000). Efficient Computation of CMB anisotropies

- in closed FRW models. *The Astrophysical Journal*, 538(2), 473–476. <https://doi.org/10.1086/309179>
- Metcalf, R. B., Meneghetti, M., Avestruz, C., Bellagamba, F., Bom, C. R., Bertin, E., Cabanac, R., Courbin, F., Davies, A., re, E. D., Flamary, R., Gavazzi, R., Geiger, M., Hartley, P., Huertas-Company, M., Jackson, N., Jacobs, C., Jullo, E., Kneib, J.-P., ... Vernardos, G. (2019). The strong gravitational lens finding challenge. *Astronomy & Astrophysics*, 625, A119. <https://doi.org/10.1051/0004-6361/201832797>
- Morgan, R., Nord, B., Birrer, S., Lin, J. Y.-Y., & Poh, J. (2021). Deeplenstronomy: A dataset simulation package for strong gravitational lensing. *Journal of Open Source Software*, 6(58), 2854. <https://doi.org/10.21105/joss.02854>
- Naess, S., Madhavacheril, M., & Hasselfield, M. (2021). *Pixell: Rectangular pixel map manipulation and harmonic analysis library*. Astrophysics Source Code Library, record ascl:2102.003.
- Pia, M. G., Basaglia, T., Bell, Z. W., & Dressendorfer, P. V. (2009). Geant4 in Scientific Literature. *arXiv e-Prints*, arXiv:0912.0360. <https://doi.org/10.48550/arXiv.0912.0360>
- Ringermacher, H. I., & Mead, L. R. (2009). A new formula describing the scaffold structure of spiral galaxies. *Monthly Notices of the Royal Astronomical Society*, 397(1), 164–171. <https://doi.org/10.1111/j.1365-2966.2009.14950.x>
- Rowe, B. T. P., Jarvis, M., Mandelbaum, R., Bernstein, G. M., Bosch, J., Simet, M., Meyers, J. E., Kacprzak, T., Nakajima, R., Zuntz, J., Miyatake, H., Dietrich, J. P., Armstrong, R., Melchior, P., & Gill, M. S. S. (2015). GALSIM: The modular galaxy image simulation toolkit. *Astronomy and Computing*, 10, 121–150. <https://doi.org/10.1016/j.ascom.2015.02.002>
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge. *arXiv e-Prints*, arXiv:1409.0575. <https://doi.org/10.48550/arXiv.1409.0575>
- Schaye, J., Crain, R. A., Bower, R. G., Furlong, M., Schaller, M., Theuns, T., Dalla Vecchia, C., Frenk, C. S., McCarthy, I. G., Helly, J. C., Jenkins, A., Rosas-Guevara, Y. M., White, S. D. M., Baes, M., Booth, C. M., Camps, P., Navarro, J. F., Qu, Y., Rahmati, A., ... Trayford, J. (2015). The EAGLE project: Simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446, 521–554. <https://doi.org/10.1093/mnras/stu2058>
- Sjöstrand, T. (2020). The PYTHIA Event Generator: Past, Present and Future. *Computer Physics Communications*, 246, 106910. <https://doi.org/10.1016/j.cpc.2019.106910>
- Springel, V. (2005). The cosmological simulation code GADGET-2. *Monthly Notices of the Royal Astronomical Society*, 364, 1105–1134. <https://doi.org/10.1111/j.1365-2966.2005.09655.x>
- The Astropy Collaboration, Robitaille, Thomas P., Tollerud, Erik J., Greenfield, Perry, Droettboom, Michael, Bray, Erik, Aldcroft, Tom, Davis, Matt, Ginsburg, Adam, Price-Whelan, Adrian M., Kerzendorf, Wolfgang E., Conley, Alexander, Crighton, Neil, Barbary, Kyle, Muna, Demitri, Ferguson, Henry, Grollier, Frédéric, Parikh, Madhura M., Nair, Prasanth H., ... Streicher, Ole. (2013). Astropy: A community Python package for astronomy. *Astronomy & Astrophysics*, 558, A33. <https://doi.org/10.1051/0004-6361/201322068>
- Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., Spergel, D. N., Somerville, R. S., Dave, R., Pillepich, A., Hernquist, L., Nelson, D., Torrey, P., Narayanan, D., Li, Y., Philcox, O., Torre, V. L., Delgado, A. M., Ho, S., Hassan, S., Burkhardt, B., Wadekar, D., ... Bryan, G. L. (2021). The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. *The Astrophysical Journal*, 915(1), 71. <https://doi.org/10.3847/1538-4357/abf7ba>

- Villaescusa-Navarro, F., Hahn, C., Massara, E., Banerjee, A., Delgado, A. M., Ramanah, D. K., Charnock, T., Giusarma, E., Li, Y., Allys, E., Brochard, A., Uhlemann, C., Chiang, C.-T., He, S., Pisani, A., Obuljen, A., Feng, Y., Castorina, E., Contardo, G., ... Spergel, D. N. (2020). The Quijote Simulations. *The Astrophysical Journal Supplement Series*, 250(1), 2. <https://doi.org/10.3847/1538-4365/ab9d82>
- Vogelsberger, M., Genel, S., Springel, V., Torrey, P., Sijacki, D., Xu, D., Snyder, G., Nelson, D., & Hernquist, L. (2014). Introducing the Illustris Project: Simulating the coevolution of dark and visible matter in the Universe. *Monthly Notices of the Royal Astronomical Society*, 444(2), 1518–1547. <https://doi.org/10.1093/mnras/stu1536>
- Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & van den Hengel, A. (2016). Visual Question Answering: A Survey of Methods and Datasets. *arXiv e-Prints*, arXiv:1607.05910. <https://doi.org/10.48550/arXiv.1607.05910>
- York, D. G., Adelman, J., Anderson, J. E., Jr., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J. A., Barkhouser, R., Bastian, S., Berman, E., Boroski, W. N., Bracker, S., Briegel, C., Briggs, J. W., Brinkmann, J., Brunner, R., Burles, S., Carey, L., Carr, M. A., ... SDSS Collaboration. (2000). The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120, 1579–1587. <https://doi.org/10.1086/301513>
- ZuHone, J. A., Vikhlinin, A., Tremblay, G. R., Randall, S. W., Andrade-Santos, F., & Bourdin, H. (2023). *SOXS: Simulated Observations of X-ray Sources*. Astrophysics Source Code Library, record ascl:2301.024.