MACHINE LEARNING

Prediction-powered inference

Anastasios N. Angelopoulos*†, Stephen Bates*†, Clara Fannjiang*†, Michael I. Jordan*†, Tijana Zrnic*†

Prediction-powered inference is a framework for performing valid statistical inference when an experimental dataset is supplemented with predictions from a machine-learning system. The framework yields simple algorithms for computing provably valid confidence intervals for quantities such as means, quantiles, and linear and logistic regression coefficients without making any assumptions about the machine-learning algorithm that supplies the predictions. Furthermore, more accurate predictions translate to smaller confidence intervals. Prediction-powered inference could enable researchers to draw valid and more data-efficient conclusions using machine learning. The benefits of prediction-powered inference were demonstrated with datasets from proteomics, astronomy, genomics, remote sensing, census analysis, and ecology.

Imagine a scientist has a machine-learning system that can supply accurate predictions about a phenomenon far more cheaply than any gold-standard experimental technique. The scientist may wish to use these predictions as evidence in drawing scientific conclusions. For example, accurate predictions of three-dimensional structures have been made for a vast catalog of known protein sequences (*1*, *2*) and are now being used in proteomics studies (*3*, *4*). Such machine-learning systems are increasingly common in modern scientific inquiry, in domains ranging from cancer prognosis to microclimate modeling. Predictions are not perfect, however, and this may lead to incorrect conclusions. Moreover, as predictions beget other predictions, the cumulative effect can amplify the imperfections. How can modern science leverage machine-learning predictions in a statistically principled way?

One way to use predictions is to follow the imputation approach: Proceed as if they are gold-standard measurements. Although this lets the scientist draw conclusions cheaply and quickly owing to the high-throughput nature of the machine-learning system, the conclusions may be invalid because the predictions may have biases.

Another possibility is to apply the classical approach: Ignore the machine-learning predictions and only use the available gold-standard measurements, which are typically far less abundant than predictions. The resulting discoveries will be statistically valid, but the smaller amount of data will limit the scope of possible discoveries.

This manuscript presents prediction-powered inference, a framework that achieves the best of both worlds: extracting information from the predictions of a high-throughput machine-learning system and guaranteeing statistical validity of the resulting conclusions. Prediction-powered inference provides a protocol for combining predictions, which are abundant but not always trustworthy, with gold-standard data, which are trusted but scarce, to compute confidence intervals and *P* values. The resulting confidence intervals and *P* values are statistically valid, as in the classical approach, but also leverage the information contained in the predictions, as in the imputation approach, to make the confidence intervals smaller and the *P* values more powerful.

Prediction-powered inference applies to any machine-learning system; as such, it absolves the need for case-by-case analyses dependent on the machine-learning algorithm on hand. The proposed protocol thereby could enable researchers to report on and assess the evidence for their conclusions in a fully standardized way.

Protocol for prediction-powered inference

The protocol for prediction-powered inference proceeds as follows. The scientist wishes to construct a confidence interval for a quantity θ^*, such as the mean outcome or a regression coefficient quantifying the statistical association between the outcome and a feature. Toward this goal, they have access to a small gold-standard dataset of features paired with outcomes, $(X, Y) = ((X_1, Y_1), ..., (X_n, Y_n))$, as well as the features of a large unlabeled dataset, $(X', Y') = \left((X_1', Y_1'), ..., (X_N', Y_N')\right)$, where the true outcomes $Y_1', ..., Y_N'$ are not observed. Typically, N is much larger than n . Both datasets are sampled at random from a larger population. Further, for both datasets the scientist has predictions of the outcomes made by a machine-learning algorithm based on the features, denoted $\left(\hat{Y}_1, ..., \hat{Y}_n\right)$ and $\left(\hat{Y}_1', ..., \hat{Y}_N'\right)$, respectively. The following exposition focuses on confidence intervals; however, by the standard duality between confidence intervals and *P* values, the presented tools immediately carry over to valid *P*-value constructions and hypothesis tests; see supplementary materials (SM) for details.

Prediction-powered inference uses the gold-standard dataset to quantify and correct for the errors made by the machine-learning algorithm on the unlabeled dataset, thereby enabling researchers to reliably incorporate predictions when constructing confidence intervals. The three-step protocol is outlined below and visualized in Fig. 1.

1) Estimand. The first step is to select an estimand θ^*. The estimand is the quantity the scientist is interested in knowing—for example, the mean outcome $E[Y_i]$, median outcome $\text{median}(Y_i)$, a linear regression coefficient obtained by regressing Y onto X, etc.

2) Measure of fit and rectifier. The key step is to identify the right measure of fit m_θ and rectifier Δ_θ for the selected estimand. For every candidate value of the estimand θ, the measure of fit m_θ is computed on the unlabeled dataset imputed with predictions, $(X', \hat{Y}')$ and quantifies how likely θ^* is to be equal to θ on the basis of the imputed data. The closer m_θ is to zero, the more plausible it is for θ^* to be equal to θ.

The rectifier Δ_θ is a notion of prediction error that is relevant for the estimand of interest. It is defined as the difference of the measure of fit m_θ computed on the labeled data, (X, Y), and the labeled data when the true outcomes are replaced with predicted ones, $\left(X, \hat{Y}\right)$. If the predictions are perfect, the rectifier is equal to zero.

Table 1 states the appropriate measure of fit and rectifier for common estimands of interest: the mean outcome, median outcome, q-quantile of the outcome, and linear and logistic regression coefficients when regressing Y onto X. A general recipe for deriving the right measure of fit and corresponding rectifier for a broad class of other estimands is provided in the SM.

3) Prediction-powered confidence interval. Finally, the measure of fit and rectifier are carefully combined to form a prediction-powered confidence interval for θ^*. This process is called rectifying the confidence interval. The prediction-powered confidence interval is constructed as $C^{\text{PP}} = \{\theta \text{ such that } |m_\theta + \Delta_\theta| \leq w_\theta(\alpha)\}$ and is guaranteed to contain the estimand with probability at least $1 - \alpha$. Here, $w_\theta(\alpha)$ is a constant that depends on the confidence level; it is explicitly stated in Theorem S1 in the SM.

Properties of prediction-powered inference

We proved mathematically that prediction-powered inference yields a confidence interval that contains the true value of the estimand at the desired confidence level, such as 95%. Notably, this validity is guaranteed for any machine-learning algorithm and any underlying data distribution. Similarly, the corresponding *P* values are also valid for any machine-learning algorithm and data distribution. See SM for the details of the mathematical proof

Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720, USA.
*Corresponding author. Email: angelopoulos@berkeley.edu (A.N.A); stephenbates@berkeley.edu (S.B.); clarafy@berkeley.edu (C.F.); michael_jordan@berkeley.edu (M.I.J.); tijana.zrnic@berkeley.edu (T.Z.)
†These authors contributed equally to this work.

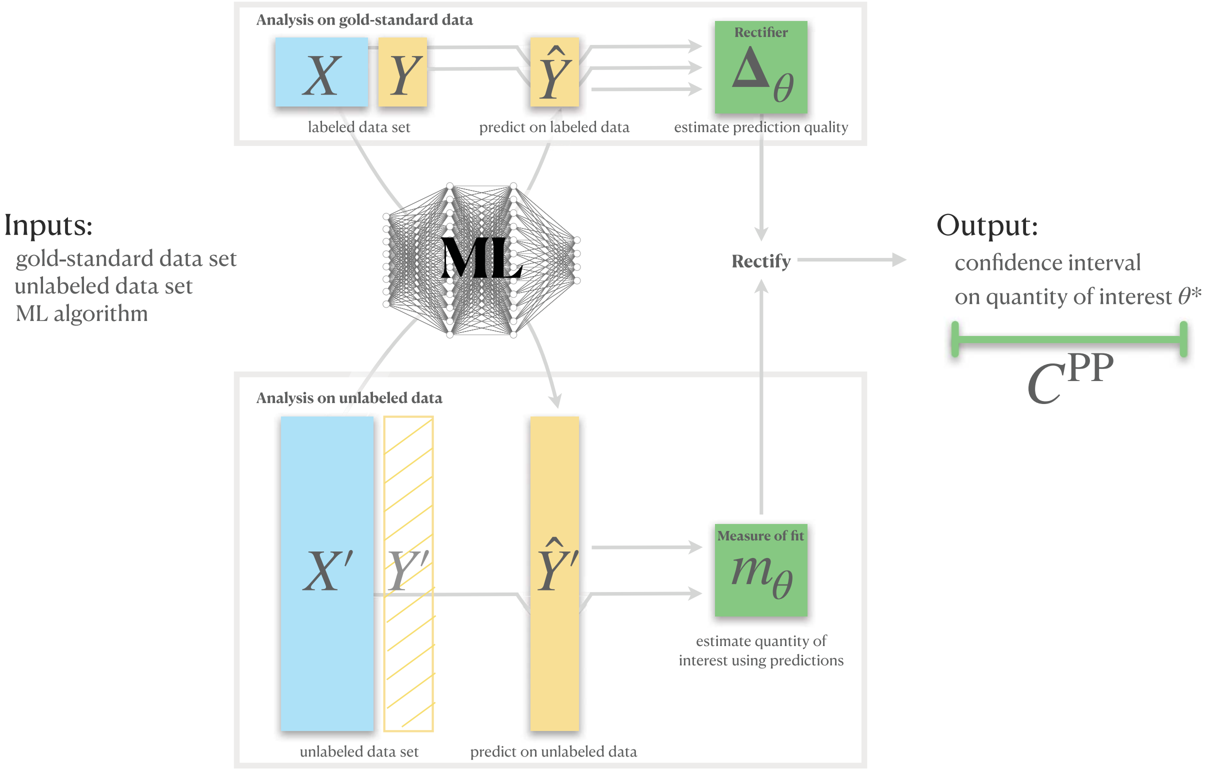

Fig. 1. Protocol for prediction-powered inference. The protocol is illustrated graphically as a block diagram. The inputs are the gold-standard dataset, the unlabeled dataset, and the machine-learning (ML) algorithm. The top block contains an analysis on gold-standard data, in which the rectifier, a measure of the prediction errors, is estimated using the labeled dataset. The bottom block contains an analysis on unlabeled data, wherein the quantity of interest is estimated using predictions. These analyses combine to form the prediction-powered confidence interval. For concrete examples of the rectifier and measure of fit, see Table 1. For a detailed theoretical exposition and more general definitions of these quantities, see SM.

of validity. A researcher relying on a deep neural network for predictions can therefore draw reliable conclusions, even though its predictions will inevitably be imperfect. Furthermore, prediction-powered inference enables more informative inferences than the classical approach, in which the researcher does not use machine-learning predictions: The confidence intervals are narrower, and the P values are more powerful. This is intuitive; prediction-powered inference carefully extracts information from the imputed data and thus has access to a larger sample size.

General applicability

Beyond quantities such as means, quantiles, and regression coefficients, the principle of prediction-powered inference can be used for constructing valid confidence intervals for any estimand that can be expressed as the minimizer of a convex objective function. This master protocol, which generalizes all the special cases instantiated in Table 1, is the core technical contribution of this work. We explained prediction-powered inference in greater generality and proved its validity in this general case in the SM. Because many important quantities can be expressed in terms of a convex-optimization problem, prediction-powered inference thus addresses many data-analysis goals beyond those explicitly demonstrated in this article.

Inference under distribution shift

Prediction-powered inference is also applicable to settings with distribution shift, i.e., the more challenging case where the unlabeled data are collected under different conditions than the gold-standard data. Two types of distribution shift are considered: label shift and covariate shift. The protocol retains the same properties as before: It is statistically valid for any machine-learning algorithm and boosts statistical power by making use of machine-learning predictions.

For covariate shift—the setting where only the feature distribution changes between the labeled and the unlabeled data—prediction-powered inference handles all estimation problems handled by the master protocol. This is done by appropriately reweighting the data; see Corollary S13 in the SM for details.

For label shift—the setting where only the label proportions change between the labeled and the unlabeled data—prediction-powered inference can be applied to estimands of the form $\theta^* = \mathrm{E}\left[\nu\left(Y_i'\right)\right]$, for a fixed function ν. For example, choosing $\nu(y) = 1\{y = k\}$ asks for inference on the proportion of instances that belong to class k. See Theorem S3 in the SM for a full description of the method.

Application of prediction-powered inference to real datasets

We demonstrated prediction-powered inference on several real tasks. In each, we computed a prediction-powered confidence interval for an estimand and compared it to intervals obtained through the classical approach and the imputation approach. In all cases, the imputation approach, which uses machine-learning predictions without accounting for prediction errors, did not contain the true value of the estimand. The widths of the two valid approaches, prediction-powered and classical, were compared as a function of the amount of labeled data used. In addition, we compared the number of labeled examples needed to reject a null hypothesis at level $1 - \alpha = 95\%$ with high probability. See (*5*) for a Python package implementing prediction-powered inference, which contains code for reproducing the experiments, and (*6*) for the data used in the experiments.

Relating protein structure and posttranslational modifications

The goal was to characterize whether various types of posttranslational modifications

Table 1. Prediction-powered inference for common statistical problems. Given a measure of fit m_θ and rectifier Δ_θ, prediction-powered inference computes a confidence interval as $C^{PP} = \{\theta \text{ such that } |m_\theta + \Delta_\theta| \le w_\theta(\alpha)\}$, where $w_\theta(\alpha)$ is a constant that depends on the error level α (see Theorem S1 in the SM). Algorithms S1 to S6 are stated in the SM. The last row ("convex minimizer") refers to a method that generalizes the methods in previous rows.

Estimand	Measure of fit m_θ	Rectifier Δ_θ	Procedure
Mean outcome	$\theta - \frac{1}{N}\sum_{i=1}^{N}\hat{Y}'_i$	$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)$	Alg. S1
Median outcome	$\frac{1}{2N}\sum_{i=1}^{N}\text{sign}\left(\theta\text{-}\hat{Y}'_i\right)$	$\frac{1}{n}\sum_{i=1}^{n}\left(1\{Y_i \le \theta\} - 1\left\{\hat{Y}_i \le \theta\right\}\right)$	Alg. S2
q-quantile of outcome	$-q + \frac{1}{N}\sum_{i=1}^{N}1\left\{\hat{Y}'_i \le \theta\right\}$	$\frac{1}{n}\sum_{i=1}^{n}\left(1\{Y_i \le \theta\} - 1\left\{\hat{Y}_i \le \theta\right\}\right)$	Alg. S3
Linear regression	$\theta - (X')^{+}\hat{Y}'$	$X^{+}(\hat{Y} - Y)$	Alg. S4
Logistic regression	$\frac{1}{N}\sum_{i=1}^{N}X'_i\left(\frac{1}{1+e^{-\theta^T X'_i}} - \hat{Y}'_i\right)$	$\frac{1}{n}\sum_{i=1}^{n}X_i\left(\hat{Y}_i - Y_i\right)$	Alg. S5
Convex minimizer	$\frac{1}{N}\sum_{i=1}^{N}\nabla L_\theta\left(X'_i, \hat{Y}'_i\right)$	$\frac{1}{n}\sum_{i=1}^{n}\left(\nabla L_\theta\left(X_i, \hat{Y}_i\right) - \nabla L_\theta(X_i, Y_i)\right)$	Alg. S6

(PTMs) occurred more frequently in intrinsically disordered regions (IDRs) of proteins (*7*). Recently, Bludau *et al.* (*3*) studied this relationship on an unprecedented proteome-wide scale by using structures predicted by AlphaFold (*1*) to predict IDRs, in contrast to previous work, which was limited to far fewer experimentally derived structures.

To quantify the association between PTMs and IDRs, the authors applied the imputation approach: They computed the odds ratio between AlphaFold-based IDR predictions and PTMs on a dataset of hundreds of thousands of protein sequence residues (*8*). Using prediction-powered inference, we could combine AlphaFold-based predictions together with gold-standard IDR labels to give a confidence interval for the true odds ratio that is statistically valid, in contrast with the interval constructed with the imputation approach, and smaller than the interval constructed using the classical approach. We used the fact that the odds ratio could be written in terms of two means and applied the recipe from the first row of Table 1; see SM for details.

We had 10,803 data points from Bludau *et al.* (*3*). For each of 100 trials, we randomly sampled n points to serve as the labeled dataset and treated the remaining $N = 10,803 - n$ points as the unlabeled dataset for which we did not observe the IDR labels. For all values of n and all three different types of PTMs that we examined, the prediction-powered confidence intervals were smaller than classical intervals; see row A in Fig. 2. Often, the classical intervals were large enough that they contained the odds ratio value of one, which means the direction of the association could not be determined from the confidence interval. However, the imputed confidence interval was far too small and significantly overestimated the true odds ratio. To reject the null hypothesis that the odds ratio is no greater than one, prediction-powered inference required $n = 316$ labeled observations, and the classical approach required $n = 799$ labeled observations; see row A in Table 2.

Galaxy classification

The goal was to determine the demographics of galaxies with spiral arms, which are correlated with star formation in the disks of low-redshift galaxies, and therefore, contribute to the understanding of star formation in the Local Universe. A large citizen science initiative called Galaxy Zoo 2 (*9*) has collected human annotations of roughly 300,000 images of galaxies from the Sloan Digital Sky Survey (*10*) with the goal of measuring these demographics. We sought to explore the use of machine learning to improve the effective sample size and decrease the requisite number of human-annotated galaxies.

We focused on estimating the fraction of galaxies with spiral arms. We had 1,364,122 labeled galaxy images from Galaxy Zoo 2, from which we simulated labeled and unlabeled datasets as follows. For each of 100 trials, we randomly sampled n points to serve as the labeled dataset and used the remaining $N = 1,364,122 - n$ points as the unlabeled dataset. We then used the first row of Table 1 to construct prediction-powered intervals. The prediction-powered confidence intervals for the mean were consistently much smaller than the classical intervals and they retained validity, and the imputation strategy failed to cover the ground truth; see Fig. 2, row B. To reject the null hypothesis that the fraction of galaxies with spiral arms is at most 0.2, prediction-powered inference required $n = 189$ labeled examples, and classical inference required $n = 449$ examples; see Table 2, row B.

Distribution of gene expression levels

Next, we constructed prediction-powered confidence intervals on quantiles that characterize how a population of promoter sequences affects gene expression. Recently, Vaishnav *et al.* (*11*) trained a state-of-the-art transformer model to predict the expression level of a particular gene induced by a promoter sequence. They used the model's predictions to study the effects of promoters—for example, by assessing how quantiles of predicted expression levels differ between different populations of promoters.

Here we focused on estimating different quantiles of gene expression levels induced by native yeast promoters. We had 61,150 labeled native yeast promoter sequences from Vaishnav *et al.* (*11*), from which we simulated labeled and unlabeled datasets as follows. For each of 100 trials, we randomly sampled n points to serve as the labeled dataset and used the remaining $N = 61,150 - n$ points as the unlabeled dataset. We then used the second and third row of Table 1 to construct prediction-powered intervals for the median, as well as the 25% and 75% quantiles, of the expression levels. The prediction-powered confidence intervals for all three quantiles were much smaller than the classical intervals for all values of n. See row C in Fig. 2 for the results for the median and fig. S6 for the other two quantiles. We also evaluated the number of labeled examples required by prediction-powered inference and classical inference, respectively, to reject the null hypothesis that the median gene expression level is at most five. Prediction-powered inference required $n = 764$ examples and classical inference required $n = 900$ examples; see row C in Table 2.

Estimating deforestation in the Amazon

The goal was to estimate the fraction of the Amazon rainforest lost between 2000 and 2015. Gold-standard deforestation labels for parcels of land are scarce, having been collected in large part through field visits, an expensive process not suited for large areas (*12*). However, machine-learning predictions of forest cover based on satellite imagery are readily available for the entire Amazon (*13*). We began with 1596 gold-standard deforestation labels for parcels of land in the Amazon. For each of 100 trials, we randomly sampled n data points to serve as the labeled dataset and used the remaining data points as the unlabeled dataset. We used the first row of Table 1 to construct the prediction-powered intervals. The imputation approach yielded a small confidence interval that failed to cover the true deforestation fraction. The classical

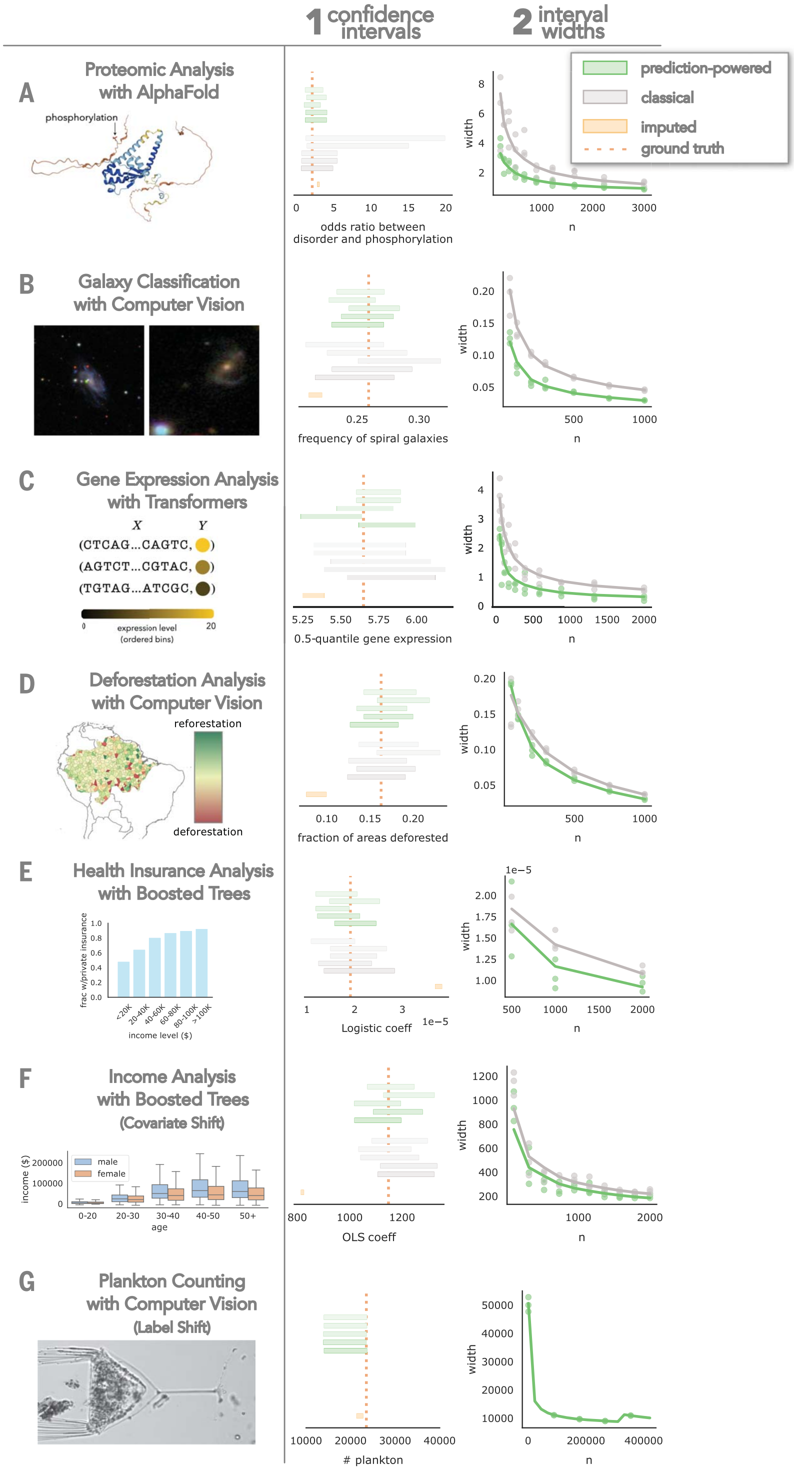

Fig. 2. Comparison of prediction-powered inference to the classical and imputation approaches on real tasks. Each row (**A** to **G**) is a different application domain. Panel 1 plots confidence intervals computed using the three approaches; for prediction-powered inference and the classical approach, intervals for five randomly chosen splits into labeled and unlabeled data are plotted. The value denoted as "ground truth" is the estimate computed on all $n + N$ data points (the true labels were available for all data points for the purpose of conducting the experiments). Panel 2 plots the average confidence interval width, as well as the width in five randomly chosen trials, for varying n, for prediction-powered inference and the classical approach; both are statistically valid solutions. The last problem setting (G) does not have a classical counterpart because the data are collected under distribution shift, hence the classical approach is not valid.

approach did cover the truth at the expense of a wider interval and, accordingly, diminished inferential power. The prediction-powered intervals were smaller than the classical intervals and retained validity; see row D in Fig. 2. We also compared the number of gold-standard deforestation labels required by prediction-powered inference and the classical approach to reject the null hypothesis that there is no deforestation. We obtain $n = 21$ labels for prediction-powered inference and $n = 35$ labels for the classical approach; see row D in Table 2.

Relationship between income and private health insurance

The goal was to investigate the quantitative effect of income on the procurement of private health insurance using US census data. Concretely, we used the Folktables interface (*14*) to download census data from California in the year 2019 (378,817 individuals). As the labeled dataset with the health insurance indicator, n census entries were randomly sampled. The remaining data were used as the unlabeled dataset. We used a gradient-boosted tree (*15*) trained on the previous year's data to predict the health insurance indicator in 2019. We constructed a prediction-powered confidence interval on the logistic regression coefficient using the fifth row of Table 1. Results in row E in Fig. 2 show that prediction-powered inference covered the ground truth, the classical interval was wider, and the imputation strategy failed to cover the ground truth. We also compared the number of gold-standard labels required by prediction-powered inference and the classical approach to reject the null hypothesis that the logistic regression coefficient is no greater than 1.5×10^{-5}. We observed a significant sample size reduction with prediction-powered inference, which required $n = 5569$ labels, whereas classical inference required $n = 6653$ labels.

Relationship between age and income in a covariate-shifted population

The goal was to investigate the relationship between age and income using US census data. The same dataset was used as in the previous experiment, but the features were age and sex, and the target was yearly income in dollars. Furthermore, a shift in the distribution of the covariates was introduced between the gold-standard and unlabeled datasets by randomly sampling the unlabeled dataset with sampling weights of 0.8 for females and 0.2 for males. We used a gradient-boosted tree (*15*) trained on the previous year's raw data to predict the income in 2019. We constructed a prediction-powered confidence interval on the ordinary least squares (OLS) regression coefficient using a covariate-shift robust version of prediction-powered inference, stated in Corollary S13 in the SM. Results in row F in Fig. 2 show that

Table 2. Number of labeled examples needed to make a discovery with prediction-powered inference and classical inference. The rows (A to F) correspond to the application domains from Fig. 2. For each application, a null hypothesis about θ^* is tested at level 95%. For details, see the SM.

Problem	Prediction-powered inference	Classical inference
A Proteomic analysis with AlphaFold	$n = 316$	$n = 799$
B Galaxy classification with computer vision	$n = 189$	$n = 449$
C Gene expression analysis with transformers	$n = 764$	$n = 900$
D Deforestation analysis with computer vision	$n = 21$	$n = 35$
E Health insurance analysis with boosted trees	$n = 5569$	$n = 6653$
F Income analysis with boosted trees	$n = 177$	$n = 282$

prediction-powered inference covered the ground truth, the classical interval was wider, and the imputation strategy failed to cover the ground truth. We also compared the number of gold-standard labels required by prediction-powered inference and the classical approach to reject the null hypothesis that the OLS regression coefficient is no greater than 800 . We observed a significant sample size reduction with prediction-powered inference, which required $n = 177$labels, whereas classical inference required $n = 282$ labels.

Counting plankton

Assessment of the increases in phytoplankton growth during springtime warming is important for the study of global biogeochemical cycling in response to climate change. We counted the number of plankton observed by the Imaging FlowCytobot (*16*, *17*), an automated, submersible flow cytometry system, at Woods Hole Oceanographic Institution in the year 2014. We had access to data from 2013, which were labeled, and we imputed the 2014 data with machine-learning predictions from a state-of-the-art ResNet fine-tuned on all data up to and including 2012. The features, X_i, are images of organic matter taken by the FlowCytobot and the labels, Y_i, are one of {detritus, plankton}, where detritus represents unspecified organic matter.

The labeled dataset consisted of 421,238 image–label pairs from 2013, and we received 329,832 labeled images from 2014. We used the data from 2014 as our unlabeled data and confirmed our results against those that were hand-labeled. The years 2013 and 2014 had a distribution shift, primarily caused by the change in the base frequency of plankton observations with respect to detritus. To apply prediction-powered inference to count the number of plankton recorded in 2014, we used the label-shift-robust technique described in Theorem S3 in the SM. The results in row G in Fig. 2 show that prediction-powered inference covered the ground truth and the imputation strategy failed to cover the ground truth.

Related work

Thematically, prediction-powered inference is most similar to the work of Wang *et al.* (*18*), who introduced a method to correct machine-learning predictions for the purpose of subsequent inference. However, this procedure is not guaranteed to provide coverage in general and requires strong assumptions about the relationship between the prediction model and the true response, whereas prediction-powered inference provides provably valid conclusions under minimal assumptions about the data-generating distribution.

There has been an increasing body of work on estimation with many unlabeled data points and few labeled data points (*19*–*27*), focusing on efficiency in semiparametric or high-dimensional regimes. Prediction-powered inference continues in this vein but focuses on the setting where the scientist has access to a good predictive model fit on separate data. This allows tackling a much wider range of estimands (e.g., minimizers of any convex objective) and gives valid inferences without assumptions about the machine-learning model. Second, prediction-powered inference goes beyond random sampling and applies to certain forms of distribution shift.

Prediction-powered inference is conceptually related to conformal prediction (*28*). Both methodologies leverage a predictive model and a labeled dataset. From this point on, however, the two methods diverge: Prediction-powered inference has additional unlabeled data and gives a confidence set that contains a population-level quantity such as the mean outcome with high probability; conformal prediction gives a confidence set for a test instance that contains the true label with high probability. Thus, the goals of prediction-powered inference and conformal prediction differ greatly from the statistical perspective. Furthermore, the mathematical tools used in the frameworks are entirely different, and neither method can be applied nontrivially to solve the objective of the other.

See SM for a further discussion of related work and the relationship of prediction-powered inference to existing baselines, as well as for empirical comparisons.

Conclusions

The past decade has witnessed rapid development and deployment of large-scale machine-learning systems across science. This surge is proceeding, however, with little statistical justification to allow these black-box systems to be used to draw scientific conclusions responsibly. Prediction-powered inference is a standardized protocol for constructing provably valid confidence intervals and P values, allowing the scientist to use the power and scale of machine-learning systems. On an array of scientific problems, we demonstrated that prediction-powered inference achieved high statistical power owing to the use of machine-learning predictions and retained statistical validity.

One question that remains open is how to handle more general forms of distribution shift. In practice, distribution shifts are often a result of a joint influence of several different forms of shift, including covariate shift and label shift and possibly others. Understanding how to handle such settings remains an important avenue for future work.

A limitation of prediction-powered inference is that it does not improve upon the classical approach when the predictions are not accurate enough or when the unlabeled dataset is not large enough compared to the gold-standard dataset. These points are demonstrated, both theoretically and empirically, in SM section "Cases Where Prediction-Powered Inference Is Underpowered." Nevertheless, given the growing number of settings with excellent predictive models and abundant unlabeled data, there is increasing potential for prediction-powered inference to benefit scientific research.

REFERENCES AND NOTES

1. J. Jumper *et al.*, *Nature* **596**, 583–589 (2021).
2. K. Tunyasuvunakool *et al.*, *Nature* **596**, 590–596 (2021).
3. I. Bludau *et al.*, *PLOS Biol.* **20**, e3001636 (2022).
4. I. Barrio-Hernandez *et al.*, Clustering predicted structures at the scale of the known protein universe. bioRxiv **2023-03** (2023).
5. A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, ppi-py: A Python package for scientific discovery using machine learning, Zenodo (2023); https://doi.org/10.5281/zenodo.8403931.
6. A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, Prediction-Powered Inference: Data Sets, Zenodo (2023); https://doi.org/10.5281/zenodo.8397451.
7. L. M. Iakoucheva *et al.*, *Nucleic Acids Res.* **32**, 1037–1049 (2004).
8. UniProt Consortium, *Nucleic Acids Res.* **43**, D204–D212 (2015).
9. K. W. Willett *et al.*, *Mon. Not. R. Astron. Soc.* **435**, 2835–2860 (2013).
10. D. G. York *et al.*, *Astron. J.* **120**, 1579–1587 (2000).
11. E. D. Vaishnav *et al.*, *Nature* **603**, 455–463 (2022).
12. E. L. Bullock, C. E. Woodcock, C. Souza Jr., P. Olofsson, *Glob. Chang. Biol.* **26**, 2956–2969 (2020).
13. J. O. Sexton *et al.*, *Int. J. Digit. Earth* **6**, 427–448 (2013).
14. F. Ding, M. Hardt, J. Miller, L. Schmidt, in *Advances in Neural Information Processing Systems* **34** (2021), pp. 6478–6490.
15. T. Chen, C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794.

16. R. J. Olson, A. Shalapyonok, H. M. Sosik, *Deep Sea Res. Part I Oceanogr. Res. Pap.* **50**, 301–315 (2003).
17. E. C. Orenstein, O. Beijbom, E. E. Peacock, H. M. Sosik, WHOI-Plankton- A large scale fine grained visual recognition benchmark dataset for plankton classification. arXiv:1510.00745 [cs.CV] (2015).
18. S. Wang, T. H. McCormick, J. T. Leek, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30266–30275 (2020).
19. M. S. Pepe, *Biometrika* **79**, 355–365 (1992).
20. J. Lafferty, L. Wasserman, in *Advances in Neural Information Processing Systems* **20** (2007), pp. 801–808.
21. A. Zhang, L. D. Brown, T. T. Cai, *Ann. Stat.* **47**, 2538–2566 (2019).
22. A. Chakrabortty, G. Dai, E. Tchetgen Tchetgen, A general framework for treatment effect estimation in semi-supervised and high dimensional settings. arXiv:2201.00468 [stat.ME] (2022).
23. A. Chakrabortty, T. Cai, *Ann. Stat.* **46**, 1541–1572 (2018).
24. Y. Zhang, J. Bradic, *Biometrika* **109**, 387–403 (2022).
25. S. Deng, Y. Ning, J. Zhao, H. Zhang, Optimal and safe estimation for high-dimensional semi-supervised learning. arXiv:2011.14185 [stat.ME] (2020).
26. D. Azriel *et al.*, *J. Am. Stat. Assoc.* **117**, 2238–2251 (2022).
27. A. Chakrabortty, G. Dai, R. J. Carroll, Semi-supervised quantile estimation: robust and efficient inference in high dimensional settings. arXiv:2201.10208 [stat.ME] (2022).
28. V. Vovk, A. Gammerman, G. Shafer, *Algorithmic Learning in a Random World* (Springer, 2005), vol. 5.

ACKNOWLEDGMENTS

We thank A. Kohli for suggesting the term “rectifier,” E. Orenstein for helpful discussions related to the WHO-I Plankton dataset, and S. Wang for help with the remote-sensing experiment. **Funding:** Office of Naval Research grant N00014-21-1-2840 (M.I.J.); National Science Foundation Graduate Research Fellowship (A.N.A., C.F.). **Author contributions:** Conceptualization: A.N.A., S.B., C.F., M.I.J., T.Z.; Methodology: A.N.A., S.B., C.F., M.I.J., T.Z.; Formal analysis: A.N.A., S.B., C.F., M.I.J., T.Z.; Funding acquisition: M.I.J.; Supervision: M.I.J.; Writing – original draft: A.N.A., S.B., C.F., M.I.J., T.Z.; Writing – review and editing: A.N.A., S.B., C.F., M.I.J., T.Z. **Competing interests:** None declared. **Data and materials availability:** The data and code are available in the accompanying Python package (*5*) and data repository (*6*). All other data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary Materials.

SUPPLEMENTARY MATERIAL

science.org/doi/10.1126/science.adi6000
Supplementary Text
Figs. S1 to S6
References (*29*–*56*)

Submitted 7 May 2023; accepted 5 October 2023
10.1126/science.adi6000