Beyond Point Estimates: Benchmarking Uncertainty Quantification Methods on the AION-1 Astronomical Foundation Model

Karla Tame-Narvaez
Scientific Computing Division
Fermi National Accelerator Laboratory
Batavia, IL 60510
karla@fnal.gov

Aleksandra Ćiprijanović
Scientific Computing Division
Fermi National Accelerator Laboratory
Batavia, IL 60510
Department of Astronomy and Astrophysics
University of Chicago
Chicago, IL 60637
NSF and Simons SkAI Institute
Chicago, IL 60637
aleksand@fnal.gov

Shubhendu Trivedi
Fermi National Accelerator Laboratory
Now at Google DeepMind
shubhendu@csail.mit.edu

Abstract

Foundation models for astronomical surveys offer powerful learned representations that can be transferred to downstream regression tasks such as galaxy property estimation. However, point predictions alone are insufficient for scientific inference; reliable uncertainty quantification (UQ) is essential. We compare seven UQ methods on galaxy property regression using frozen AION-1 foundation-model embeddings, predicting redshift, stellar mass, stellar-population age, gas-phase metallicity, and specific star-formation rate, from Legacy Survey photometry/imaging and DESI spectra, with PROVABGS-derived labels. Distribution-free conformal methods achieve marginal coverage within $\sim$1 pp of the nominal 90% across all properties, while non-conformal baselines (Deep Ensembles, MC Dropout) fail to calibrate reliably. Among conformal approaches, Conformalized Quantile Regression (CQR) delivers the best coverage in the bin with the poorest model predictions. More importantly, only the Locally Valid and Discriminative (LVD) framework—particularly when operating on AION-1 embeddings—also provides finite-sample *local validity*, producing intervals that adapt to each galaxy's local prediction difficulty rather than relying on marginal guarantees alone. These results establish conformal prediction, and LVD in particular, as the preferred UQ framework for uncertainty-aware inference on foundation-model embeddings in astrophysics.

1 Introduction

The estimation of physical properties of galaxies from photometric imaging data is a cornerstone of observational cosmology and galaxy evolution studies. Properties such as photometric and spectroscopic redshift, stellar mass, stellar and gas metallicity, stellar-population age, and specific star-formation rate encode the physical processes governing galaxy formation, from star formation

Conference on Physics and AI at Stanford University (PAI 2026)

and chemical enrichment to environmental quenching and the galaxy–halo connection. In recent years, deep learning approaches have demonstrated strong performance in inferring galaxy properties from heterogeneous and high-dimensional data [6, 11, 2, 16]. Among these, *foundation models* [13, 14]—large neural networks pre-trained on broad, unlabeled datasets via self-supervised objectives—have emerged as a particularly promising paradigm. By learning generic, transferable representations of the input data, foundation models decouple the costly pre-training phase from lightweight, task-specific fine-tuning, classification or regression heads, thereby enabling rapid deployment across heterogeneous downstream tasks.

The AION-1 model [14] exemplifies this strategy in the astronomical domain. Pre-trained on multi-band stellar and galaxy data from the Legacy Survey [5], Hyper Suprime-Cam (HSC) [1], Sloan Digital Sky Survey (SDSS) [18], the Dark Energy Spectroscopic Instrument (DESI) [4], and Gaia [7], AION-1 produces high-dimensional embeddings which integrates complex information from heterogeneous imaging, spectroscopic, and scalar data. Training is done using a two-stage architecture: modality-specific tokenization followed by transformer-based masked modeling of cross-modal token sequences. AION-1 achieves strong results on a broad suite of downstream tasks, such as galaxy and stellar property estimation, galaxy morphology classification, similarity search, galaxy image segmentation, and spectral super-resolution.

In this work we will focus on downstream regression of galaxy properties from AION-1 embeddings. However, point estimates presented in AION-1 paper are insufficient for scientific inference. Robust *uncertainty quantification* (UQ) is crucial for propagating measurement and modeling errors, identifying out-of-distribution inputs and distribution shift, and ensuring that both individual predictions and population-level inferences remain reliable and scientifically robust. Despite its importance, UQ in deep learning remains only partially understood, and for large, black-box foundation models—with their complex, high-dimensional embeddings—it is especially challenging and largely unexplored in astrophysics and science in general. In this paper, we conduct a direct comparison of seven UQ methods—LDV and LVD-MAD [12], VanillaSplit [17], MADSplit [10, 3], CQR [15], Deep Ensembles [9], MC Dropout [8]—applied to the same AION-1 backbone and the same downstream regression task as in Table 1 from [14]. We measure the marginal, tail, and worst-predicted coverage, as well as efficiency (mean prediction interval width) for each UQ method, and discuss problems and potential pitfalls related to some of these commonly used UQ methods.

2 Data

We use galaxy samples from three AION-1 surveys: photometry ($\{g, r, i, z, y\}$ fluxes) and $\{g, r, z\}$ imaging from the DESI Legacy Imaging Surveys, 3600–9800 Å DESI spectra, and PROVABGS SED-fitting posteriors providing labels for redshift, stellar mass, age, metallicity, and specific star-formation rate[1]. We leverage tokens from the pre-trained AION-1 model, but train a task-specific regressor using these tokens and PROVABGS labels. The dataset consists of 50,000 galaxies (out of 120,000 total) for training and 11,362 for testing, ensuring no overlap with AION-1 pre-training via object IDs. For downstream UQ tasks, the training set is further split 70%/30% into train/calibration. Final dataset sizes are 34,844 objects for training, 14,934 for calibration and 11,362 for testing.

Each dataset is tokenized following the modality-specific procedures described in the AION-1 framework. The resulting token sequences are concatenated and passed through the frozen AION-1 encoder, which produces a contextualized 768-dimensional embedding vector for each galaxy via attentive pooling. This configuration corresponds to the photometry + imaging + spectra (Ph+Im+Sp) input mode of AION-1, which was shown to yield the highest R^2 across all five targets [2]. The encoder weights remain frozen throughout the study; only the downstream regression head and UQ-specific components are optimized. This design ensures that all UQ methods operate on an identical input representation and that the comparison isolates the effect of the UQ strategy itself.

[1] We note that these labels are themselves derived quantities from SED-fitting rather than direct observables, and therefore carry their own modeling uncertainties; our UQ evaluation measures calibration with respect to these derived labels, which in some cases might differ from true physical values.

[2] Throughout this test study we use the AION-B variant (300 M parameters), the smallest member of the AION-1 model family—which also includes AION-L (800 M) and AION-XL (3 B).

3 Regression Network Architecture

For each galaxy we regress five physical properties: redshift z, stellar mass $\log M_\star$, mass-weighted stellar-population age $t_{\rm age}$, mass-weighted gas-phase metallicity $\log Z$, and specific star-formation rate $\log \mathrm{sSFR}$. The regression architecture used in this study is a three-layer multilayer perceptron (MLP), which maps the 768-dimensional AION-1 embeddings to a scalar prediction for each target property. The network consists of two hidden layers of 256 units each with Gaussian Error Linear Unit (GELU) activations, followed by a single linear output neuron. The second hidden-layer activations $\boldsymbol{h}_2 \in \mathbb{R}^{256}$ are returned as intermediate embeddings for use in downstream UQ methods. A separate MLP is trained independently for each of the five target properties. Each model is optimized using Adam with a learning rate of 10^{-3} and weight decay of 10^{-5}, paired with a cosine annealing learning rate schedule over $T_{\max} = 100$ cycles, for a total of 500 epochs minimizing the MSE loss. After training, the model is evaluated on the held-out test set and the coefficient of determination R^2 is recorded for each target property as a point-estimate baseline. Our model produces similar R^2 values as in AION-1 paper, namely $\{z, \log M_\star, t_{\rm age}, \log Z, \log \mathrm{sSFR}\} = \{0.98, 0.89, 0.85, 0.63, 0.86\}$. The code used in this work will be shared on Github, after paper is accepted.

4 UQ Methods

We compare seven UQ methods and it's variants. The *Locally Valid and Discriminative* (LVD) framework [12] learns a kernel in a reduced feature space to estimate conditional nonconformity quantiles, yielding prediction intervals with approximate *per-input* validity. We test two embedding spaces: **LCNet**, where the kernel operates on the 256-dimensional MLP hidden-layer activations $\boldsymbol{h}_2$, and **LCAion**, where it operates on the raw 768-dimensional AION-1 embeddings. Formally, local validity requires that the prediction interval covers the true value with probability at least $1 - \alpha$ within any local neighborhood of the input, not just on average across all inputs; this is strictly stronger than marginal validity, which only guarantees the nominal coverage rate when averaged over the entire test distribution. **LVD-MAD** extends LVD by normalizing nonconformity scores with a learned residual-magnitude predictor, producing scale-adaptive intervals; we again have two variants: **LCMADNet** (kernel and residual predictor on $\boldsymbol{h}_2$) and **LCMADAion** (kernel and residual on AION-1 embeddings). For the MAD variants, the residual predictor is trained on the training ser and calibration is performed on the held-out calibration set, preventing data leakage. For all LVD variants, the kernel is a metric-learned projection to 5 dimensions (500 iterations, learning rate 10^{-3}, early stopping after 50 non-improving steps, with L2 normalisation); for the MAD variants, the residual magnitude predictor is a two-layer MLP (hidden size 32 for **LCMADNet**, 64 for **LCMADAion**) with ReLU activations and a Softplus output, trained for 500 steps with Adam at 10^{-3}. The runtime for all four LVD variants is approximately 30min in a 40 GB GPU A-100.

VanillaSplit [17] is standard split conformal prediction with a single global nonconformity quantile, producing constant-width intervals with marginal coverage guarantees. **MADSplit** [10, 3] is its MAD-normalized variant, scaling scores by a learned residual predictor trained on the training set to adapt interval width to local prediction difficulty. **CQR** (Conformalized Quantile Regression) [15] trains separate networks for the $\alpha/2$ and $1-\alpha/2$ conditional quantiles on training set, then applies a conformal correction on the calibration set to restore finite-sample marginal coverage.

The two non-conformal baselines are **DE** (Deep Ensembles) [9], which trains $M{=}5$ independently initialized MLPs (with MSE loss) and constructs Gaussian intervals from the inter-model predictions of the mean and standard deviation, and **MC Dropout** [8], which injects dropout ($p{=}0.1$) into the pre-trained regressor at inference time and estimates predictive mean and variance from $T{=}100$ stochastic forward passes. Both capture only *epistemic* uncertainty—DE through inter-model disagreement, MC Dropout through dropout-induced variance—with no learned aleatoric term.

Conformal methods provide distribution-free, finite-sample coverage guarantees, making them well suited to astrophysical settings with heterogeneous data, complex noise, and simulation–data mismatch. They are increasingly used for reliable uncertainty intervals in tasks such as redshift estimation and stellar population inference. In contrast, non-conformal approaches (e.g., Deep Ensembles and MC Dropout) offer richer predictive uncertainty representations and integrate naturally with inference pipelines, but rely on model assumptions and lack formal coverage guarantees, often leading to miscalibration under distribution shift. While conformal methods ensure marginal validity, they may suffer from conditional miscalibration; methods such as MADSplit, CQR, and LVD address

this through input-dependent or locally adaptive calibration, improving reliability in high-dimensional regimes.

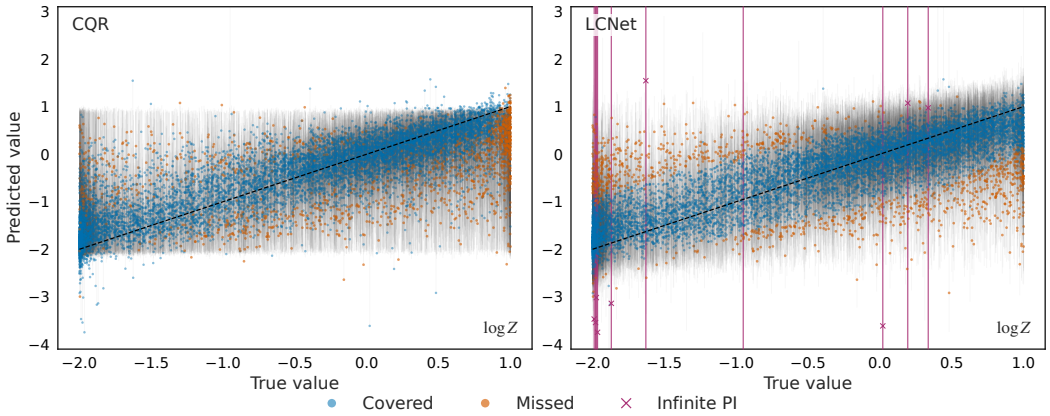

Figure 1: Predicted vs. True $\log Z$ with 90% prediction intervals for CQR (left) and LCNet (right). Both achieve similar marginal coverage ($\sim 90\%$), but CQR produces more uniform intervals, while LCNet adapts interval widths to local prediction difficulty. A small number (10 cases) of LCNet intervals are unbounded (which is actually a desirable behaviour, as it can be used to discover outliers and anomalies).

5 Results and Conclusion

We evaluate coverage at three complementary levels: over the full test set (marginal coverage), in the 10% upper and lower tails (tail coverage), and on the 10% worst-predicted points (worst-predicted bin coverage). As expected, all conformal methods—including the LVD variants, VanillaSplit, MADSplit, and CQR—achieve similar $\sim$90% marginal coverage (Table 1), consistent with their distribution-free guarantees.

In the tails, LVD variants remain competitive, with LCMADAion and LCAion generally achieving the most consistent performance, indicating that operating in the richer AION embedding space enhances robustness near the edges or in low-density regions. This advantage likely reflects structure in the $\mathbb{R}^{768}$ AION embedding that the downstream MLP discards during compression to $\mathbb{R}^{256}$, but which LVD's kernel can still exploit for local calibration. In contrast, the non-conformal baselines fail to calibrate properly: DE systematically undercovers (12–38%) with overly narrow intervals, while MC Dropout exhibits strong overcoverage (57–99%) with intervals that are 2–3$\times$ wider.

Clear differences emerge in more challenging regimes. In the worst-predicted bin, CQR attains the highest coverage (44–66%), followed by the LVD variants, where LCMADAion (22–42%) and LCMADNet (9–40%) consistently outperform their non-MAD counterparts, LCAion (12–33%) and LCNet (10–20%). This demonstrates the importance of scale-adaptive normalization in capturing heteroscedasticity and improving local calibration. Across all targets, LVD methods maintain non-trivial coverage, while VanillaSplit collapses to $\sim$0% in the hardest regions and MADSplit, although better (7-41%), shows less consistency.

Among well-calibrated methods, the LVD variants achieve the best efficiency, producing the tightest intervals while preserving coverage. Although CQR shows stronger performance in marginal and worst-bin metrics, it guarantees only marginal validity. In contrast, LVD provides approximate conditional coverage through local, kernel-based calibration, ensuring reliability at the level of individual predictions (see Fig. 1). This makes LVD—particularly the MAD and AION variants—a more principled choice when per-input validity and robustness in complex, high-dimensional regimes are required. All experiments use the AION-B (300M parameters) variant; whether the LVD-on-embedding advantage persists at the AION-L (800M parameters) and AION-XL (3B parameters) scales is left to a follow-up study. We note that the coverage figures in Table 1 do not include uncertainty estimates; bootstrap or repeated-calibration-split error bars will be reported in a more thorough follow-up study, to better understand the stability of each UQ method.

Table 1: Marginal coverage, tail coverage, worst-predicted bin coverage, and efficiency (mean prediction interval width) for each UQ method at nominal level $1 - \alpha = 0.90$.

	LCNet	*LCAion*	*LCMADNet*	*LCMADAion*	*VanillaSplit*	*MADSplit*	*CQR*	*DE*	*MC Dropout*
Marginal Coverage (%)									
z	89.64	91.42	90.07	91.43	89.48	89.49	90.04	38.14	99.18
$\log M_\star$	88.76	90.09	88.82	89.71	89.54	89.86	89.78	11.68	99.78
$t_{\rm age}$	90.68	90.55	92.55	91.11	90.07	90.83	90.77	21.62	97.48
$\log Z$	90.51	89.06	88.37	87.91	89.10	88.43	88.92	18.93	56.72
$\log \mathrm{sSFR}$	89.42	91.16	89.42	89.59	89.61	89.59	90.26	22.53	99.38
Tail Coverage (%)									
z	84.33	87.43	83.85	87.70	75.62	77.99	89.83	34.60	98.37
$\log M_\star$	79.89	79.34	79.75	83.41	75.31	83.14	83.10	9.860	99.16
$t_{\rm age}$	88.32	87.29	91.18	87.60	86.40	89.44	87.06	29.09	94.94
$\log Z$	86.34	84.05	88.11	85.51	83.23	86.49	75.88	13.82	53.96
$\log \mathrm{sSFR}$	79.62	83.65	79.49	84.86	78.52	84.15	83.93	22.27	97.71
Worst-Predicted Bin Coverage (%)									
z	18.75	31.77	20.95	34.16	0.000	7.390	65.76	0.000	91.81
$\log M_\star$	10.04	14.69	9.510	39.03	0.000	38.47	48.86	0.000	97.80
$t_{\rm age}$	20.19	16.95	40.48	22.42	0.700	27.82	43.93	0.000	74.82
$\log Z$	17.27	12.46	37.71	28.36	0.000	26.58	52.73	0.000	6.16
$\log \mathrm{sSFR}$	11.88	32.97	12.59	41.78	0.000	41.20	48.42	0.000	93.84
Efficiency									
z	0.039	0.041	0.041	0.041	0.040	0.040	0.054	0.016	0.109
$\log M_\star$	0.563	0.575	0.563	0.554	0.594	0.558	0.667	0.060	2.173
$t_{\rm age}$	0.142	0.141	0.154	0.142	0.143	0.141	0.153	0.022	0.228
$\log Z$	1.695	1.681	1.523	1.574	1.791	1.617	1.571	0.217	0.757
$\log \mathrm{sSFR}$	0.705	0.708	0.702	0.673	0.725	0.677	0.711	0.125	2.055

Acknowledgments and Disclosure of Funding

This work was produced by Fermi Forward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. The United States Government retains, and the publisher, by accepting the work for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this work, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (`http://energy.gov/downloads/doe-public-access-plan`). The work of K.T. and A.C. is supported by DOE Grant DE-SCL0000092 (HEP AmSC IDA Pilot: AI Universe).

References

[1] Hiroaki Aihara, Nobuo Arimoto, Robert Armstrong, et al. The Hyper Suprime-Cam SSP survey: Overview and survey design. *Publications of the Astronomical Society of Japan*, 70(SP1):S4, 2018.

[2] Ashraf Ayubinia, Jong-Hak Woo, Fatemeh Hafezianzadeh, Taehwan Kim, and Changseok Kim. Prediction of Star Formation Rates Using an Artificial Neural Network. , 980(2):177, February 2025.

[3] Anthony Bellotti. Constructing normalized nonconformity measures based on maximizing predictive efficiency. In Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Giovanni Cherubin, editors, *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 41–54. PMLR, 09–11 Sep 2020.

[4] DESI Collaboration, Amir Aghamousa, Jessica Aguilar, Steve Ahlen, Shadab Alam, Lori E. Allen, Carlos Allende Prieto, James Annis, Stephen Bailey, et al. The DESI Experiment Part I: Science,Targeting, and Survey Design. *arXiv e-prints*, page arXiv:1611.00036, October 2016.

[5] Arjun Dey, David J. Schlegel, Dustin Lang, Robert Blum, Kaylan Burleigh, Xiaohui Fan, Joseph R. Findlay, Doug Finkbeiner, et al. Overview of the DESI Legacy Imaging Surveys. , 157(5):168, May 2019.

[6] Mario Gai, Mario Bove, Giovanni Bonetta, Davide Zago, and Rossella Cancelliere. Simultaneous derivation of galaxy physical properties with multimodal deep learning. , 532(2):1391–1401, August 2024.

[7] Gaia Collaboration, T. Prusti, J. H. J. de Bruijne, et al. The Gaia mission. *Astronomy & Astrophysics*, 595:A1, 2016.

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

[9] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[10] Jing Lei, Max Grazier G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry A. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094 – 1111, 2016.

[11] Ping Li, Li-Li Wang, Guang-Jun Yang, Jia-Bao Feng, and Yan-Ke Tang. Estimation of age and metallicity for galaxies based on multi-modal deep learning. , 698:A222, June 2025.

[12] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Locally valid and discriminative prediction intervals for deep learning models. In *Neural Information Processing Systems*, 2021.

[13] Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Rudy Morel, Ruben Ohana, Mariel Pettee, Bruno Régaldo-Saint Blancard, Kyunghyun Cho, Shirley Ho, and Polymathic AI Collaboration. AstroCLIP: a cross-modal foundation model for galaxies. , 531(4):4990–5011, July 2024.

[14] Liam Parker, Francois Lanusse, Jeff Shen, Ollie Liu, Tom Hehir, Leopoldo Sarra, Lucas Meyer, Micah Bowles, et al. AION-1: Omnimodal Foundation Model for Astronomical Sciences. *arXiv e-prints*, page arXiv:2510.17960, October 2025.

[15] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

[16] Jing Rou Puah and Sasa Arsovski. A Deep Multimodal Multi–Head Neural Network for Joint Estimation of Stellar Age, Lifetime, and Evolutionary Stage. *arXiv e-prints*, page arXiv:2511.18477, November 2025.

[17] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[18] Donald G. York, J. Adelman, John E. Anderson, et al. The Sloan Digital Sky Survey: Technical summary. *The Astronomical Journal*, 120(3):1579–1587, 2000.