
Cross-simulator transfer with self-supervised representations in 21cm cosmology

Yannic Pietschke* **Caroline Heneka** **Ayodele Ore**
Institut für Theoretische Physik, Universität Heidelberg
Philosophenweg 16, 69120 Heidelberg, Germany

Romain Meriot
Max Planck Institute for Radioastronomy
Auf dem Hügel 69, 53121 Bonn, Germany

Abstract

Simulation-based inference is limited by model misspecification: neural summaries trained on one simulator degrade on data from another, or from a real instrument. We study this in 21cm cosmology, where the brightness-temperature field from the Epoch of Reionization is targeted by the Square Kilometre Array (SKA). We pretrain a Vision Transformer once with a self-supervised Joint Embedding Predictive Architecture (JEPA) objective, label-free, on cheap semi-numerical simulations, freeze it, and transfer to a physically distinct hydrodynamical radiative-transfer simulator with a disjoint parameter set, under realistic instrumental noise. The frozen encoder, shown neither the target simulator, its parameters, nor any noise, matches or exceeds a supervised network retrained directly on the noisy target, and retains calibrated posteriors. A controlled ablation holding architecture and pretraining data fixed shows that the self-supervised objective, not data scale, is what enables this robustness.

1 Introduction

Simulation-based inference (SBI) [Cranmer et al., 2020] learns the map from observations to parameters directly from forward-simulated data, for models whose likelihood is intractable. Its central bottleneck is that the neural summaries involved are trained on one (or a few) forward model(s) and typically fail when faced with domain shifts and model misspecification, meaning on data from a different simulator, or on observations. Distinct codes adopt distinct parameterizations and physical approximations, and none captures the instrumental pipeline of a real measurement exactly [Cannon et al., 2022, Sooknunan et al., 2024].

A route that has recently emerged is self-supervised foundation model pretraining [e.g. Parker et al., 2024, 2025, Smith et al., 2024]: pretrain an encoder once on a low-cost, abundant dataset with a label-free objective, then reuse the frozen representation downstream with light task-specific supervision. Because the objective is not tied to any single target label set, the representation can in principle transfer to targets whose parameters, noise model and systematics were never seen during pretraining.

Whether self-supervised summaries survive such a shift has not been established for cosmological analyses. Existing demonstrations stay inside a single simulator family, testing robustness to added noise, changed resolution, or a changed parameter set while the simulator itself is held fixed [Ore et al., 2025]. The pressing question is whether such summaries transfer across simulator *families*,

*Correspondence: pietschke@thphys.uni-heidelberg.de

and thus parameterizations, and across noise regimes at once: if they do, the same frozen encoder can perform calibrated inference on data whose physics and systematics were never seen during pretraining. In this work we test exactly this model transfer. A Vision Transformer is pretrained once, label-free, on cheap semi-numerical simulations, frozen, and applied to a physically distinct hydrodynamical radiative-transfer simulator with a disjoint parameter set, first noiseless and then under realistic instrumental noise. An extended treatment of this study, with the full per-parameter and label-efficiency analysis under noise, is presented in Pietschke et al. [2026].

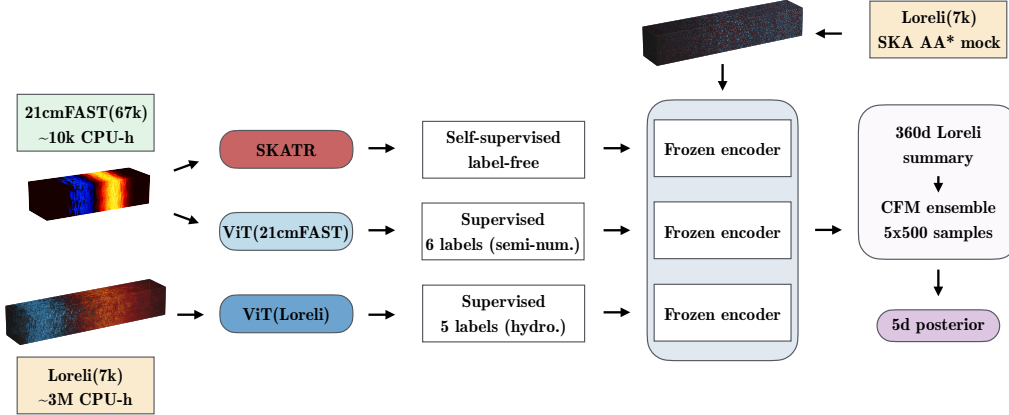


Figure 1: The pipelines compared in this work: in each case a frozen encoder maps a target simulation to a 360-d summary, which a conditional flow matching head converts into a posterior. The pipelines differ only in how the encoder is pretrained: self-supervised on the source set (SKATR), supervised on the same source set (ViT(21cmFAST)), or supervised on the budget- N target set (ViT(Loreli)). Reproduced from Pietschke et al. [2026].

- To our knowledge this is the first demonstration of cross-simulator and cross-noise transfer for a cosmological large-scale structure signal. The frozen encoder, shown neither the target simulator, its parameters, nor any noise, matches or exceeds a supervised network retrained directly on the noisy target while retaining calibrated posteriors.
- The objective, not the data scale, drives this: in a controlled ablation sharing data, architecture and downstream task, the self-supervised representation loses 4% of its mean R^2 under the noise shift against 21% for a supervised one.
- The gain is label efficiency rather than peak accuracy: an augmented supervised baseline matches at the full budget, but the gap reopens to $+0.12$ once target labels become scarce.

2 A physically defined domain shift

2.1 Source and target data

The two simulators produce 3D lightcone simulations of the 21cm brightness-temperature, a tomographic map of neutral hydrogen in the early Universe. The **source** comprises 67,325 lightcones from the semi-numerical code 21cmFAST [Murray et al., 2020, Mesinger et al., 2011, Park et al., 2019], which identifies ionized regions with an excursion-set approach, varies six parameters over flat priors and costs ~ 0.25 CPU-h per sample. The **target** comprises 7,667 lightcones from Loreli II [Meriot and Semelin, 2024, Meriot et al., 2025], generated with Licorice [Semelin et al., 2007, Baek et al., 2009, Semelin et al., 2017], in which ionization and thermal histories follow from explicit radiative transfer, varies five parameters on a grid and costs 300–500 CPU-h per sample: $\sim 180\times$ more expensive in aggregate while being $\sim 8.8\times$ smaller.

Four components of the domain shift are crossed at once: *physics*, excursion set against explicit radiative transfer; *parameterization*, disjoint parameter sets of different size; *temporal/redshift coverage*; and *instrument*, where in the second half of our experiments the target is corrupted with SKA AA* thermal noise computed with 21cmSense [Pober et al., 2013, 2014]. Noise enters the target only. Both datasets are downsampled to a common $32 \times 32 \times 576$ grid and symmetric-log

transformed; 1,150 target lightcones are held out for testing, the remaining 6,517 forming the pool from which budgets N are drawn.

2.2 Encoder, tokenization and pretraining

The encoder is a Vision Transformer (ViT) adapted to 3D inputs, following the SKA Transformer (SKATR) of Ore et al. [2025]: each lightcone is tokenized into non-overlapping $[4, 4, 12]$ patches, yielding 3072 tokens; six pre-norm blocks ($\sim 6.3\text{M}$ parameters) then produce 360-d token embeddings. A lightcone is summarized by mean-pooling its tokens into one 360-d vector. Pretraining uses a Joint Embedding Predictive Architecture (JEPA) [Assran et al., 2023], which predicts the embeddings of masked regions rather than reconstructing them in voxel space: a context encoder sees only unmasked patches, a target encoder embeds the full lightcone, and a predictor regresses the target embeddings at the masked locations under an L_1 loss. Predicting in embedding space rather than in voxel space is what we expect to provide transferability. Voxel reconstruction would force the encoder to retain every small-scale detail, including the simulator-specific texture and noise-like modes that do not survive a change of code or the addition of an instrument, whereas an embedding-space target leads to discarding what is not relevant in terms of physical predictions and rather retains structure across scales in relation to physical predictions [Assran et al., 2023, Bardes et al., 2024, Balestrieri and LeCun, 2024]. Our target encoder is a copy of the context encoder at initialization, held fixed throughout training. A fixed target cannot collapse and embeddings of a fixed randomly initialized network are known to be informative self-supervised targets [Sui et al., 2024]. Only the context encoder is retained downstream. The masking is dictated by the physics: each step draws two large (60–80% of the spatial area) and two small (10–20%) rectangular regions, each spanning the full redshift axis, so masking acts purely in the sky plane and never along the line of sight, which carries the temporal evolution of the signal. Pretraining runs for 100 epochs, takes ~ 60 GPU-hours on one Nvidia H200, and shows the encoder no target data and no parameter labels; the encoder is frozen afterwards and reused unchanged at every target budget.

2.3 Pipelines and downstream protocol

The pipelines are summarized in Fig. 1. **ViT(21cmFAST)** is the same architecture trained on the same source lightcones with a supervised L_2 objective on the six source parameters, so comparing it to **SKATR** isolates the objective from the data scale. **ViT(Loreli)** is trained supervised on the budget- N target lightcones, retrained from scratch at every budget, either without or with instrumental noise added; the source-pretrained encoders are applied unchanged, making their use on noisy data a genuine zero-shot transfer. Both supervised backbones carry a regression head on top of the pooled representation which outputs the respective set of parameters the L_2 loss is computed from. After training the head is discarded and the lightcone is summarized by mean-pooling its 3072 token embeddings into a 360-d vector, exactly as for SKATR, so all three pipelines hand summaries of identical dimension to an identical downstream head. The six source parameters therefore enter only the supervised source pretraining; all three pipelines infer the same five target hydrodynamical parameters, namely the star-formation timescale τ , the minimum star-forming halo mass $M_{\min}$, the X-ray production efficiency f_X , the hard-to-soft X-ray ratio r_H and the late-time escape fraction $f_{\text{esc,post}}$. Downstream, a conditional flow matching (CFM) head [Lipman et al., 2023, Tong et al., 2024] maps the frozen summary to a posterior over the five target parameters, with five realizations per (pipeline, budget) cell and 500 samples from each.

We assess three properties on the held-out test set: *accuracy*, the R^2 between posterior mean and truth averaged over the five parameters; *informativeness*, the mean posterior standard deviation divided by that of the labels, where 1 is a posterior as wide as the prior; and *calibration*, both joint, via the TARP coverage test [Lemos et al., 2023], and marginal, via simulation-based calibration [SBC, Talts et al., 2018]. SBC gives a consistent picture throughout and is not shown here for brevity.

3 Results

3.1 Transfer across the simulator and the noise boundary

Across the simulator boundary alone, all three pipelines recover the truth, but with markedly different posteriors (Table 1, noiseless columns). SKATR attains the best mean R^2 at 0.98 and the smallest

Pipeline	mean $R^2 \uparrow$		rel. width $\downarrow$	
	noiseless	noisy	noiseless	noisy
SKATR	0.98	0.94	0.08	0.17
ViT(21cmFAST)	0.94	0.74	0.18	0.40
ViT(Loreli)	0.96	0.91	0.13	0.22
ViT(Loreli, R+R)	0.98	—	0.07	—

Table 1: Mean R^2 and relative posterior width at the full target budget $N = 6517$. All pipelines share architecture, summary dimension and downstream head, differing only in how the encoder is obtained. ViT(Loreli) is retrained per regime, so its two columns come from different networks; the last row is the R+R-augmented in-domain baseline of Sec. 3.2.

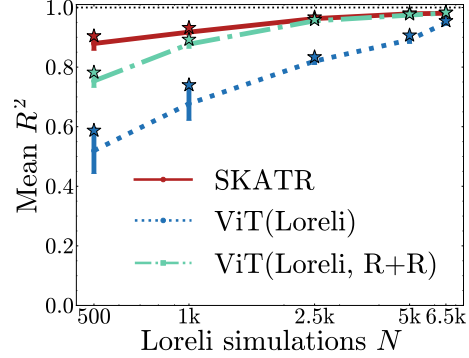


Figure 2: R^2 against the number of target simulations N , noiseless, for SKATR and the in-domain baseline with and without R+R augmentations. Error bars show the spread over the CFM ensembles; stars mark the combined ensemble. Adapted from Pietschke et al. [2026].

relative posterior width on all five parameters, against 0.96 for the in-domain baseline and 0.94 for ViT(21cmFAST). The latter is already a strong baseline: a supervised encoder pretrained on a large source dataset transfers to the target simulator without ever seeing a target label, so cross-simulator transfer is to a certain extent not restricted to the self-supervised objective. However, self-supervised pretraining adds substantial further gains against it, from 0.94 to 0.98, with more precise posteriors. The TARP coverage indicates that SKATR is marginally underconfident, while the supervised baselines carry a mix of biases and over-/under-confidence (Fig. 3a).

Adding the noise shift separates the pipelines into three regimes (Table 1, noisy columns). *Zero-shot supervised transfer is fragile*: frozen ViT(21cmFAST) drops to $R^2 = 0.74$ while more than doubling its posterior width. *The in-domain baseline maintains accuracy at the cost of calibration*: retraining ViT(Loreli) on noisy data yields $R^2 = 0.91$, only 0.03 short of SKATR, but its TARP curve falls below the diagonal (Fig. 3b), indicating posteriors that are systematically overconfident and biased. *The self-supervised representation is accurate, informative and calibrated*: SKATR reaches $R^2 = 0.94$ with the smallest posterior width on every parameter and a TARP curve close to the diagonal, having seen neither the target simulator nor the noise. Since all three pipelines share the same CFM architecture as inference head, training protocol and ensembling, these calibration differences are attributable to the summaries rather than to the density estimator. The in-domain baseline trains its encoder on the same budget- N labels the CFM is then fitted on, so an encoder that overfits produces trained summaries that lead to an underestimation of corresponding covariances, or parameter uncertainties, when presented with the test summaries, with the CFM calibrating to the training set, which is the expected route to overconfident posteriors; SKATR’s encoder never sees a target label, so no such gap arises. Per-parameter results are given in Appendix B.

The cleanest isolation of the objective comes from the two pipelines sharing pretraining data, architecture and downstream protocol: with SKATR frozen, the full-budget mean R^2 degrades as $0.98 \rightarrow 0.94$ when evaluating on noisy data, compared to $0.94 \rightarrow 0.74$ for ViT(21cmFAST), establishing the pretraining objective itself, rather than data scale, as the source of the robustness.

3.2 Label efficiency and augmented supervised baseline

Because the expensive simulator sets the cost, the relevant question is accuracy per target simulation. Noiseless at $N = 500$ (Fig. 2), $R^2 = 0.90$ for SKATR sits far above the 0.59 of ViT(Loreli), which trains its encoder from scratch and overfits where labels are scarcest. SKATR at $N = 2500$ already matches the supervised baseline trained on the full pool of 6517: $\sim 1.6 \times 10^6$ CPU-h of radiative-transfer hydrodynamics avoided, against a one-time $\sim 1.7 \times 10^4$ CPU-h source set and 60 GPU-h of pretraining. Under noise the ordering holds at every budget; the corresponding scaling curves are given in Pietschke et al. [2026].

All baselines above are trained without sample-level augmentations, a deliberate choice that isolates the pretraining objective but also leaves the supervised baseline untuned. Retraining the in-domain ViT(Loreli) with rotate-and-reflect (R+R) augmentations, the dihedral group D_4 of 90° rotations and reflections about the transverse axes, is physically natural because the signal is statistically isotropic in the sky plane. With that augmentation a fully-supervised in-domain ViT matches SKATR at the full budget on all three axes: mean R^2 of 0.98 against 0.98 and a mean relative posterior width of 0.07 against 0.08 (Table 1), and the same slight underconfidence as SKATR rather than the overconfidence it previously showed (Fig. 3c). The matching holds only in the data-rich regime, however. SKATR leads by +0.04 in mean R^2 at $N = 1000$, narrowing to parity at the full budget, but the gap widens sharply to +0.12 at $N = 500$: sky-plane augmentation does not substitute for pretraining when target labels are too few to exploit it. What distinguishes the frozen encoder is in any case independent of such tuning: the supervised pipeline commits to the target in ways SKATR avoids, retraining an encoder at every budget, locking its objective to the five target parameters, and requiring a fresh backbone once noise is added. For real cosmological observations, where both the parameterization and the noise model differ from any training simulator, this is a valuable distinction.

4 Discussion

Taken together, these results establish a self-supervised encoder as an effective foundation model for our inference problem: pretrained once on low-cost noiseless simulations and frozen, it transfers under a joint domain shift to a physically distinct simulator and to realistic instrumental noise, having seen neither. On the noisy target it exceeds a supervised model trained directly on that data, and is the only pipeline that stays calibrated.

We can only speculate about why JEPA features transfer across both the simulator and the noise shift before opening up the latent space, but the two robustness results share a plausible common origin. Predicting masked blocks in embedding space, across both smaller- and larger-scale masks, rewards an encoder that captures correlations across scales and the connections between them, encouraging a representation organized around the physics of the signal rather than any single scale; supervised regression is instead free to collapse onto whatever simulator-specific structures minimize the regression loss, which need be neither physically meaningful nor shared across simulators. Confirming this requires a latent-space analysis, asking whether the summaries cluster by physical regime rather than by simulator family. An alternative supervised baseline would train the encoder jointly with the CFM head, backpropagating through the density estimator rather than through an L_2 regression loss, and so capture uncertainty directly in pretraining. Similarly, the pretrained SKATR encoder could itself be finetuned jointly with the CFM task head, so the scope for tuning is not restricted to the supervised pipeline. We do not include it here because freezing the transfer pipelines SKATR and ViT(21cmFAST) is the premise of the comparison, and end-to-end training would change according to task objective and optimization at once. In Section 3.2 we show that a well-tuned supervised ViT can match SKATR at the full budget with respect to in-domain inference; SKATR outperforms the supervised baseline in terms of label efficiency, posterior calibration, and robustness under shift rather than peak attainable accuracy. A systematic comparison against an end-to-end supervised pipeline and the effect of finetuning is left to future work.

Acknowledgments and Disclosure of Funding

We thank Divesh Jain, Vrund Patel and M. Paola Vaccaro for useful discussions. We also thank the anonymous reviewers on OpenReview for their valuable feedback. YP’s and CH’s work is funded by the Volkswagen Foundation and supported through the Federal Ministry of Education and Research (BMBF) and the Ministry of Science, Research and the Arts of Baden-Württemberg. CH also acknowledges funding via the SciFM consortium (05D25VH3) funded by the German Federal Ministry of Research, Technology, and Space (BMFTR) in the ErUM-Data action plan. This work was supported by the DFG under Germany’s Excellence Strategy EXC 2181/1 - 3909000948 The Heidelberg STRUCTURES Excellence Cluster. The authors acknowledge support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 37/935-1 FUGG. The authors declare no competing interests.

References

- M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, June 2023.
- S. Baek, P. Di Matteo, B. Semelin, F. Combes, and Y. Revaz. The simulated 21 cm signal during the epoch of reionization: full modeling of the Ly- α pumping. *A&A*, 495(2):389–405, Feb. 2009. doi: 10.1051/0004-6361:200810757.
- R. Balestriero and Y. LeCun. How learning by reconstruction produces uninformative features for perception. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=XsDWw1Mn2p>.
- A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=QaCCuDfBk2>. Featured Certification.
- P. Cannon, D. Ward, and S. M. Schmon. Investigating the Impact of Model Misspecification in Neural Simulation-based Inference. *arXiv e-prints*, art. arXiv:2209.01845, Sept. 2022. doi: 10.48550/arXiv.2209.01845.
- K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Science*, 117(48):30055–30062, Dec. 2020. doi: 10.1073/pnas.1912789117.
- P. Lemos, A. Coogan, Y. Hezaveh, and L. Perreault-Levasseur. Sampling-Based Accuracy Testing of Posterior Estimators for General Inference. *40th International Conference on Machine Learning*, 2021:19256–19273, Jan. 2023. doi: 10.48550/arXiv.2302.03026.
- Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow Matching for Generative Modeling. In *Proceedings of ICLR*, 2023. doi: 10.48550/arXiv.2210.02747.
- R. Meriot and B. Semelin. The LORELI database: 21 cm signal inference with 3D radiative hydrodynamics simulations. *A&A*, 683:A24, Mar. 2024. doi: 10.1051/0004-6361/202347591.
- R. Meriot, B. Semelin, and D. Cornu. Comparison of Bayesian inference methods using the LORELI II database of hydro-radiative simulations of the 21-cm signal. *A&A*, 698:A80, June 2025. doi: 10.1051/0004-6361/202452901.
- A. Mesinger, S. Furlanetto, and R. Cen. 21cmfast: a fast, seminumerical simulation of the high-redshift 21-cm signal. *MNRAS*, 411(2):955–972, Feb. 2011. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2010.17731.x. URL <https://doi.org/10.1111/j.1365-2966.2010.17731.x>.
- S. G. Murray, B. Greig, A. Mesinger, J. B. Muñoz, Y. Qin, J. Park, and C. A. Watkinson. 21cmfast v3: A python-integrated c code for generating 3d realizations of the cosmic 21cm signal. *JOSS*, 5(54):2582, 2020. doi: 10.21105/joss.02582. URL <https://doi.org/10.21105/joss.02582>.
- A. Ore, C. Heneka, and T. Plehn. SKATR: A self-supervised summary transformer for SKA. *SciPost Phys.*, 18:155, 2025. doi: 10.21468/SciPostPhys.18.5.155. URL <https://scipost.org/10.21468/SciPostPhys.18.5.155>.
- J. Park, A. Mesinger, B. Greig, and N. Gillet. Inferring the astrophysics of reionization and cosmic dawn from galaxy luminosity functions and the 21-cm signal. *MNRAS*, 484(1):933–949, Jan. 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz032. URL <http://dx.doi.org/10.1093/mnras/stz032>.
- L. Parker, F. Lanusse, S. Golkar, L. Sarra, M. Cranmer, A. Bietti, M. Eickenberg, G. Krawezik, M. McCabe, R. Morel, R. Ohana, M. Pettee, B. Régald-Saint Blancard, K. Cho, S. Ho, and Polymathic AI Collaboration. AstroCLIP: a cross-modal foundation model for galaxies. *MNRAS*, 531(4):4990–5011, July 2024. doi: 10.1093/mnras/stae1450.
- L. Parker, F. Lanusse, J. Shen, O. Liu, T. Hehir, L. Sarra, L. Meyer, M. Bowles, S. Wagner-Carena, H. Qu, S. Golkar, A. Bietti, H. Bourfoune, N. Casserau, P. Cornette, K. Hirashima, G. Krawezik, R. Ohana, N. Lourie, M. McCabe, R. Morel, P. Mukhopadhyay, M. Pettee, B. Regaldo-Saint Blancard, K. Cho, M. Cranmer, and S. Ho. AION-1: Omnimodal Foundation Model for Astronomical Sciences. *arXiv e-prints*, art. arXiv:2510.17960, Oct. 2025. doi: 10.48550/arXiv.2510.17960.
- Y. Pietschke, C. Heneka, A. Ore, and R. Meriot. Cross-simulator transfer with foundation model summaries: Towards robust SKA-era reionization inference, 2026. URL <https://arxiv.org/abs/2608.26354>. Submitted to A&A, August 2026.

- J. C. Pober, A. R. Parsons, D. R. DeBoer, P. McDonald, M. McQuinn, J. E. Aguirre, Z. Ali, R. F. Bradley, T.-C. Chang, and M. F. Morales. The baryon acoustic oscillation broadband and broad-beam array: Design overview and sensitivity forecasts. *AJ*, 145(3):65, jan 2013. doi: 10.1088/0004-6256/145/3/65. URL <https://dx.doi.org/10.1088/0004-6256/145/3/65>.
- J. C. Pober, A. Liu, J. S. Dillon, J. E. Aguirre, J. D. Bowman, R. F. Bradley, C. L. Carilli, D. R. DeBoer, J. N. Hewitt, D. C. Jacobs, M. McQuinn, M. F. Morales, A. R. Parsons, M. Tegmark, and D. J. Werthimer. What next-generation 21 cm power spectrum measurements can teach us about the epoch of reionization. *ApJ*, 782(2):66, jan 2014. doi: 10.1088/0004-637X/782/2/66. URL <https://dx.doi.org/10.1088/0004-637X/782/2/66>.
- B. Semelin, F. Combes, and S. Baek. Lyman-alpha radiative transfer during the epoch of reionization: contribution to 21-cm signal fluctuations. *A&A*, 474(2):365–374, Nov. 2007. doi: 10.1051/0004-6361:20077965.
- B. Semelin, E. Eames, F. Bolgar, and M. Caillat. 21SSD: a public data base of simulated 21-cm signals from the epoch of reionization. *MNRAS*, 472(4):4508–4520, Dec. 2017. doi: 10.1093/mnras/stx2274.
- M. J. Smith, R. J. Roberts, E. Angeloudi, and M. Huertas-Company. AstroPT: Scaling Large Observation Models for Astronomy. *arXiv e-prints*, art. arXiv:2405.14930, May 2024. doi: 10.48550/arXiv.2405.14930.
- K. Sooknunan, E. Chapman, L. Conaboy, D. Mortlock, and J. Pritchard. Reproducibility of machine learning analyses of 21 cm reionization maps. *arXiv e-prints*, art. arXiv:2412.15893, Dec. 2024. doi: 10.48550/arXiv.2412.15893.
- Y. Sui, T. Wu, J. C. Cresswell, G. Wu, G. Stein, X. S. Huang, X. Zhang, and M. Volkovs. Self-supervised representation learning from random data projectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=EpYnZpDpsQ>.
- S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. *arXiv e-prints*, art. arXiv:1804.06788, Apr. 2018. doi: 10.48550/arXiv.1804.06788.
- A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. In *Proceedings of TMLR*, 2024. URL <https://openreview.net/forum?id=CD9Snc73AW>.

A Calibration diagnostics

Calibration is assessed with the TARP coverage test [Lemos et al., 2023], which estimates expected coverage from distances to random reference points and is therefore sensitive to bias as well as to posterior width. Marginal calibration was also computed via simulation-based calibration [SBC, Talts et al., 2018] and gives a consistent picture; the per-parameter rank histograms are reported in Pietschke et al. [2026]. The corresponding coverage curves are shown in Fig. 3.

B Per-parameter results

Table 2 resolves the means presented in Table 1 by parameter. SKATR attains the best R^2 on all five parameters in both the noiseless and the noisy regime, and the narrowest posterior on all five under noise, so the improvement is not driven by a subset of well-constrained parameters. The largest gains are on r_H and $f_{\text{esc,post}}$, which are also the most sparsely sampled parameters of the grid of the hydrodynamical target simulations (Loreli II). Under noise the ordering is unchanged but the spread between the noiseless and the noisy regime widens sharply for the zero-shot supervised encoder, which loses most on τ ($0.94 \rightarrow 0.66$) and $f_{\text{esc,post}}$ ($0.91 \rightarrow 0.69$).

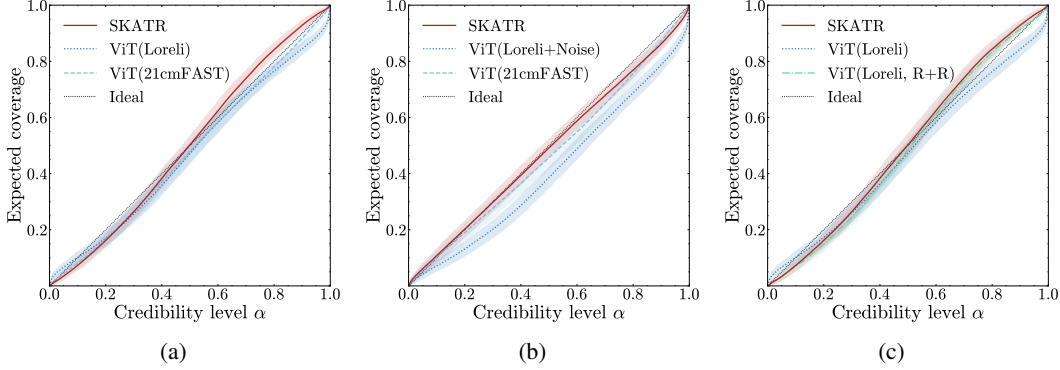


Figure 3: Joint TARP coverage at the full label budget. **(a)** Noiseless: SKATR is marginally underconfident, while the supervised baselines carry a mix of biases and over-/under-confidence. **(b)** Under SKA AA* noise: the noise-retrained in-domain baseline ViT(Loreli+Noise) falls below the diagonal, indicating overconfidence and bias, while SKATR stays close to it; ViT(21cmFAST) lies near the diagonal only because its posteriors are so broad. **(c)** Noiseless, with the R+R-augmented in-domain baseline of Sec. 3.2 added: it shows the same slight underconfidence as SKATR. Reproduced from Pietschke et al. [2026].

	noiseless			SKA AA* noise		
Parameter	ViT(Loreli)	ViT(21cmFAST)	SKATR	ViT(Loreli)	ViT(21cmFAST)	SKATR
Accuracy, R^2 ($\uparrow$)						
f_X	0.97	0.97	0.99	0.94	0.85	0.96
τ	0.95	0.94	0.99	0.88	0.66	0.92
r_H	0.92	0.89	0.97	0.89	0.70	0.92
$\log M_{\min}$	0.98	0.98	1.00	0.94	0.80	0.97
$f_{\text{esc,post}}$	0.96	0.91	0.97	0.87	0.69	0.91
mean	0.96	0.94	0.98	0.91	0.74	0.94
Relative posterior width ($\downarrow$)						
f_X	0.12	0.11	0.06	0.26	0.27	0.11
τ	0.14	0.18	0.08	0.21	0.44	0.19
r_H	0.14	0.25	0.14	0.20	0.44	0.19
$\log M_{\min}$	0.11	0.13	0.06	0.19	0.37	0.14
$f_{\text{esc,post}}$	0.12	0.20	0.09	0.22	0.45	0.20
mean	0.13	0.18	0.08	0.22	0.40	0.17

Table 2: Per-parameter accuracy and relative posterior width at the full target budget $N = 6517$, on the combined ensemble posteriors. Best value per row and regime in bold. ViT(Loreli) is retrained per regime, so its noiseless and noisy columns come from different networks; under noise it is the noise-trained in-domain baseline. Reproduced from Pietschke et al. [2026].