

Article

Estimating H I Mass Fraction in Galaxies with Bayesian Neural Networks

Joelson Sartori ¹, Cristian G. Bernal ² and Carlos Frajuca ^{3,*}

¹ Centro de Ciências Computacionais, Universidade Federal de Rio Grande, Rio Grande 96203-900, RS, Brazil; joelsonsartori@gmail.com

² Universidad Nacional de Colombia, Apdo. Postal 111321, Cd. Universitaria, Bogotá 111321, Colombia; cribernal@unal.edu.co

³ Instituto de Matemática, Estatística e Física (IMEF), Universidade Federal de Rio Grande, Rio Grande 96203-900, RS, Brazil

* Correspondence: frajuca@gmail.com

Abstract

Neutral atomic hydrogen (HI) regulates galaxy growth and quenching, but direct 21 cm measurements remain observationally expensive and affected by selection biases. We develop Bayesian neural networks (BNNs)—a type of neural model that returns both a prediction and an associated uncertainty—to infer the HI mass, $\log_{10}(M_{\text{HI}})$, from widely available optical properties (e.g., stellar mass, apparent magnitudes, and diagnostic colors) and simple structural parameters. For continuity with the photometric gas fraction (PGF) literature, we also report the gas-to-stellar-mass ratio, $\log_{10}(G/S)$, where explicitly noted. Our dataset is a reproducible cross-match of SDSS DR12, the MPA-JHU value-added catalogs, and the 100% ALFALFA release, resulting in 31,501 galaxies after quality controls. To ensure fair evaluation, we adopt fixed train/validation/test partitions and an additional sky-holdout region to probe domain shift, i.e., how well the model extrapolates to sky regions that were not used for training. We also audit features to avoid information leakage and benchmark the BNNs against deterministic models, including a feed-forward neural network baseline and gradient-boosted trees (GBTs, a standard tree-based ensemble method in machine learning). Performance is assessed using mean absolute error (MAE), root-mean-square error (RMSE), and probabilistic diagnostics such as the negative log-likelihood (NLL, a loss that rewards models that assign high probability to the observed HI masses), reliability diagrams (plots comparing predicted probabilities to observed frequencies), and empirical 68%/95% coverage. The Bayesian models achieve point accuracy comparable to the deterministic baselines while additionally providing calibrated prediction intervals that adapt to stellar mass, surface density, and color. This enables galaxy-by-galaxy uncertainty estimation and prioritization for 21 cm follow-up that explicitly accounts for predicted uncertainties (“risk-aware” target selection). Overall, the results demonstrate that uncertainty-aware machine-learning methods offer a scalable and reproducible route to inferring galactic HI content from widely available optical data.



Academic Editor: Alexei V. Moiseev

Received: 17 October 2025

Revised: 6 January 2026

Accepted: 23 January 2026

Published: 2 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: 21 cm line; Arecibo Legacy Fast ALFA (ALFALFA) survey; Bayesian neural networks; galaxy evolution; gas fraction; machine learning; neutral atomic hydrogen (HI); optical/UV photometry; photometric gas fraction (PGF); Sloan Digital Sky Survey (SDSS)

1. Introduction

Neutral atomic hydrogen (HI) is the dominant reservoir of the cold interstellar medium and the fuel for star formation. Its abundance regulates how galaxies assemble and quench over cosmic time [1,2]. The most direct tracer is the 21 cm emission line, but such measurements are observationally costly and limited by radio-frequency interference, beam confusion, and flux sensitivity. Even state-of-the-art blind surveys such as ALFALFA [3] are biased toward gas-rich systems, leaving large gaps in parameter space. This tension between the need for statistical HI samples and the cost of radio observations motivates the use of photometric estimators that exploit widely available optical data.

Empirical correlations between gas fraction and global properties—stellar mass, stellar-mass surface density, broadband colors, and morphology—have long supported the photometric gas fraction (PGF) technique [4–7], and more recent machine-learning frameworks that directly predict HI content from optical data and environment [8–10]. PGF relations are simple and interpretable, but their predictive scatter remains large for outliers (e.g., dusty or environmentally processed galaxies) and they provide only global residual dispersions, not calibrated uncertainties for individual systems.

Machine learning (ML) offers a flexible alternative by capturing non-linear, multi-variate relations between photometry, structure, and gas content. Most existing work, however, has emphasized point predictions with limited treatment of uncertainty. Yet calibrated uncertainty is essential for scientific use: it underpins follow-up strategies and reliable physical inference. Neural networks in particular are prone to miscalibration, i.e., their quoted uncertainties do not always match the actual frequency of errors [11]. Post hoc calibration methods [12] and deep ensembles [13] provide useful surrogates but do not unify aleatoric (data-related) and epistemic (model-related) uncertainty in a single framework. BNNs do so by placing distributions over weights and/or outputs, enabling heteroscedastic predictive intervals (whose width can vary with galaxy properties) and likelihood-based training (optimizing how probable the observed HI masses are under the predicted distributions) (e.g., [14]).

Here, we develop and evaluate BNNs to infer $\log_{10}(M_{\text{HI}})$ from optical/UV photometry and structural parameters; we also report $\log_{10}(G/S)$ where needed for comparability with photometric gas fraction (PGF) relations. We construct a reproducible cross-match of SDSS DR12 imaging and spectroscopy [15], the MPA-JHU value-added catalogs [16], and the ALFALFA 100% catalog [3], yielding 31,501 galaxies after quality controls. Our design enforces strict non-overlapping train/validation/test partitions (including a sky-holdout split to probe domain shift), audits features to preclude information leakage, and evaluates both *point accuracy* (MAE, RMSE) and *probabilistic calibration* (negative log-likelihood, reliability diagrams, probability integral transform, and empirical coverage at 68% and 95%). Our BNN approach builds on this growing body of ML-based HI estimators [9,10,17] while explicitly modeling predictive uncertainties. We benchmark BNNs against deterministic models and against PGF calibrations from the literature.

The central question is as follows: *to what extent can HI mass (and gas fraction) be inferred from optical/UV proxies, and with what calibrated uncertainty, compared to PGF relations and deterministic ML models?* Beyond raw accuracy, we test whether BNN predictive intervals maintain near-nominal coverage across stellar mass, surface density, and color, enabling risk-aware prioritization of galaxies for 21 cm follow-up. All code, data splits, and processing scripts are released with a DOI to ensure reproducibility. Our work complements previous efforts that use artificial neural networks or other ML methods to infer HI content from optical data [8–10,17], by introducing a BNN that delivers calibrated predictive intervals across the relevant colour–mass–surface density regimes.

The paper is organized as follows. Section 2 describes the datasets, PGF relations, and probabilistic models. Section 3 presents empirical results on accuracy, calibration, and robustness. Section 4 discusses implications for PGF relations and future surveys. Section 5 summarizes the main conclusions and reproducibility resources.

2. Materials and Methods

2.1. Photometric Estimators of HI Gas Fraction

Neutral atomic hydrogen is measurable through the 21 cm hyperfine transition, but direct surveys are limited by telescope time and systematics such as radio-frequency interference (RFI), beam confusion, and flux thresholds. Even with blind surveys like ALFALFA [3], selection effects bias detections toward gas-rich systems and complicate statistical inference at fixed stellar mass, structure, or color. At the same time, new wide-area 21 cm surveys such as APERTIF and ASKAP/WALLABY [17–19] are dramatically expanding the local census of neutral hydrogen, providing orders-of-magnitude more detections than previous generations of pointed observations. However, these facilities remain subject to flux limits, radio-frequency interference, and selection effects that complicate statistical inference at fixed stellar mass or structural parameters. In this context, photometric and machine-learning-based estimators are best viewed as complementary tools: they can help prioritise targets for 21 cm follow-up, interpolate across survey boundaries, and provide approximate HI mass fractions where direct detections are unavailable. These limitations motivate photometric estimators that predict the gas fraction

$$f_{\text{HI}} \equiv \frac{M_{\text{HI}}}{M_{\star}}, \quad \log_{10}(G/S) \equiv \log_{10}\left(\frac{M_{\text{HI}}}{M_{\star}}\right), \quad (1)$$

from optical/UV measurements and structural indicators available in wide-field surveys such as SDSS [15,16]. These surveys provide a well-characterized local volume with robust photometric gas fraction estimates [20], which we exploit here to train and validate our BNN model.

Much of the literature documents correlations between HI content and galaxy properties such as stellar mass $M_{\star}$, stellar-mass surface density $\mu_{\star}$, broadband optical colors (e.g., $g - r$), and morphology (e.g., [6,7,21]). These trends underpin the PGF methodology, in which $\log_{10}(G/S)$ is predicted from a compact set of indicators. A representative linear calibration is

$$\log_{10}(G/S) = a_0 + a_1 (g - r) + a_2 \mu_i + a_3 \log_{10}\left(\frac{M_{\star}}{M_{\odot}}\right) + \dots, \quad (2)$$

with coefficients $\{a_k\}$ fitted on HI-detected samples and a typical scatter of 0.25–0.35 dex depending on dust, inclination, and environment [5,6]. PGF relations are straightforward and interpretable but can underfit regimes where the mapping is strongly non-linear (e.g., dusty edge-on disks, bulge-dominated galaxies, environmentally quenched systems), and they provide only a global residual scatter rather than calibrated predictive intervals.

An extended PGF variant may include additional optical and structural terms:

$$\log_{10}(G/S) = b_0 + b_1 (g - r) + b_2 \log_{10} \mu_{\star} + b_3 \log_{10}\left(\frac{M_{\star}}{M_{\odot}}\right), \quad (3)$$

with coefficients $\{b_k\}$ fitted on the training data only and evaluated on held-out sets to avoid leakage [6,7]. These relations provide a baseline against which to compare modern machine-learning approaches.

2.2. From Deterministic ML to Probabilistic Prediction

Supervised machine learning provides flexible function approximators—models that learn a mapping from inputs (optical/structural quantities) to outputs (HI mass)—that can capture non-linear, multivariate dependencies between photometry, structure, and gas content. Deterministic models—such as regularized regressions, gradient-boosted trees, and feed-forward neural networks (deep neural networks, DNNs)—yield accurate point predictions but are often miscalibrated, in the sense that the scatter of their errors is not correctly reflected in simple uncertainty estimates based on residuals [11]. Post hoc calibration improves empirical coverage [12], and deep ensembles are a robust baseline [13], but these approaches do not place an explicit probabilistic model on the predictive distribution conditioned on inputs. Recent work has further demonstrated the utility of ML in such an ecosystem for reconstructing missing photometry and probing quenching mechanisms [22], reinforcing the case for combining flexible ML models with physically motivated gas fraction estimators. BNNs provide such a framework, in that by placing distributions over weights and/or outputs, BNNs yield a full predictive distribution, separating aleatoric (data) from epistemic (model) uncertainty (Figure 1). In the heteroscedastic regression setting relevant here, we model

$$p(y | \mathbf{x}) = \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x})), \quad y \in \{\log_{10}(M_{\text{HI}}), \log_{10}(G/S)\}, \quad (4)$$

and train by minimizing the negative log-likelihood

$$\mathcal{L}_{\text{NLL}} = \sum_{i=1}^N \left[\frac{(y_i - \mu(\mathbf{x}_i))^2}{2\sigma^2(\mathbf{x}_i)} + \frac{1}{2} \log \sigma^2(\mathbf{x}_i) \right]. \quad (5)$$

Conceptually, minimizing this negative log-likelihood favors models that not only predict the correct central HI mass, but also assign narrow intervals where the data are tight and broader intervals where the data are intrinsically more scattered.

$$\mathcal{L}_{\text{NLL}} = \sum_{i=1}^N \left[\frac{(y_i - \mu(\mathbf{x}_i))^2}{2e^{s(\mathbf{x}_i)}} + \frac{1}{2}s(\mathbf{x}_i) \right]. \quad (6)$$

Approximate inference can be implemented via variational methods or Monte Carlo dropout [14], both of which provide practical ways to approximate a fully Bayesian treatment at a reasonable computational cost. The posterior predictive marginalizes over weight uncertainty:

$$p(y_* | \mathbf{x}_*, \mathcal{D}) = \int p(y_* | \mathbf{x}_*, \mathbf{w}) p(\mathbf{w} | \mathcal{D}) d\mathbf{w}. \quad (7)$$

Here, the first term (‘aleatoric’) captures measurement noise and intrinsic diversity in HI mass at fixed observables, while the second (‘epistemic’) reflects how much the predictions change when the network parameters are varied within their posterior, and to what extent the model is constrained by the available data.

$$\hat{\mu}(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^T \mu^{(t)}(\mathbf{x}), \quad (8)$$

$$\widehat{\text{Var}}(y | \mathbf{x}) \approx \underbrace{\frac{1}{T} \sum_{t=1}^T \sigma^{2(t)}(\mathbf{x})}_{\text{aleatoric}} + \underbrace{\frac{1}{T} \sum_{t=1}^T (\mu^{(t)}(\mathbf{x}))^2 - \hat{\mu}(\mathbf{x})^2}_{\text{epistemic}}. \quad (9)$$

This enables predictive intervals whose empirical coverage can be directly tested. Accordingly, we evaluate both *point accuracy* (MAE/RMSE) and *probabilistic quality* (negative log-likelihood, reliability diagrams, probability integral transform, and empirical coverage at 68% and 95%).

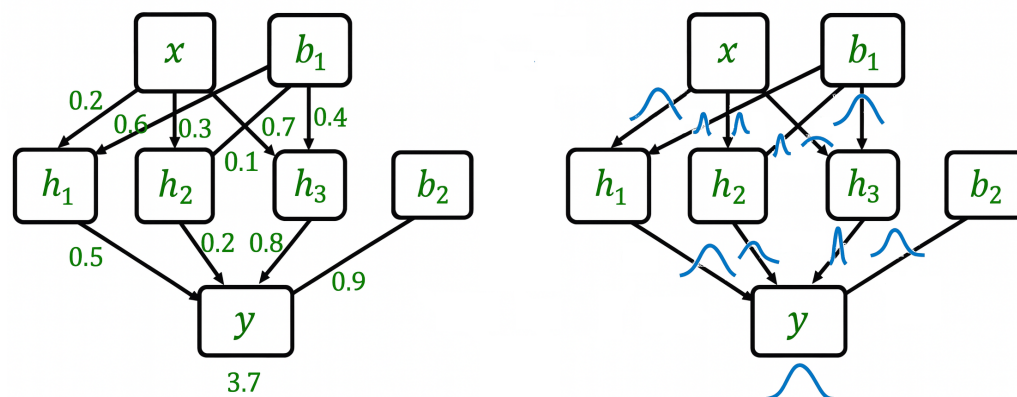


Figure 1. Schematic contrast between a deterministic feed-forward neural network (**left**) and a Bayesian neural network (**right**). Deterministic networks learn point estimates and return a single prediction; BNNs place distributions over weights and/or outputs, yielding a full predictive distribution and calibrated uncertainty quantification.

2.3. Surveys and Value-Added Catalogs

Our working set is a reproducible cross-match of three public resources: SDSS DR12 imaging and spectroscopy [15], the MPA-JHU value-added catalogs providing stellar masses and related quantities [16], and the 100% ALFALFA HI catalog [3]. SDSS supplies optical photometry and simple structural proxies; MPA-JHU provides homogenized stellar-population estimates (e.g., M_* , μ_*); and ALFALFA contributes integrated HI measurements from blind 21 cm detections. To preserve portability beyond the HI-detected subset, inputs are restricted to quantities widely available in SDSS.

2.4. Cross-Matching and De-Duplication

We associate ALFALFA detections to SDSS counterparts using J2000 right ascension and declination within the catalog association windows, following the probabilistic framework of Budavári and Szalay [23]. When multiple SDSS entries are plausible, we retain a single counterpart per HI source using nearest-neighbor tie-breaking and removing duplicates. All SQL queries and matching scripts are released with the code to guarantee exact reproducibility of the cross-identification.

2.5. Target Definition and Feature Set

Our **primary** target is $\log_{10}(M_{\text{HI}})$ from ALFALFA (base-10); for continuity with PGF relations, we also derive $\log_{10}(G/S) = \log_{10}(M_{\text{HI}}) - \log_{10}(M_*)$ using MPA-JHU stellar masses. We train and evaluate in log space by default and report linear-space summaries where relevant.

Given the availability patterns in the full cross-match, we release two tabular products: (i) a **full** table with 38 columns and partial missingness; and (ii) a **processed** version with 9 columns and no missing values used for the models in this work. The processed set contains only predictors that are fully observed across all galaxies (e.g., SDSS i -band apparent magnitude and extinction terms, MPA-JHU stellar mass and its uncertainty, an SFR proxy, and a surface-brightness proxy; Table 1). To mitigate selection-driven shortcuts (e.g., Malmquist bias), the input feature space excludes explicit distance indicators such as catalog distance and recession velocity/redshift. The processed predictors consist of widely

available optical photometry and simple structural proxies (including apparent magnitudes and extinction terms), without providing any direct distance or redshift information to the models. Apparent-to-intrinsic transformations (e.g., absolute magnitudes or rest-frame colors) are therefore not included as predictors and are used only for descriptive diagnostics when available.

Table 1. Predictors used in the processed dataset and their physical meaning.

Variable (Processed)	Meaning/Notes
$iMAG$	SDSS i -band apparent magnitude (AB) used as an input feature in the processed table.
A_g, A_i	Galactic extinction terms in the g and i bands (used as predictors).
$\log M_\star$ ($\log MsT$), $e_{\log M_\star}$	Stellar mass (MPA–JHU, total) and its reported uncertainty.
$\log SFR$ proxy ($\log SFR22$)	Scalar SFR indicator available in the catalog; used as a predictor, not as a target.
surface_brightness_proxy	Magnitude-like scalar derived from SDSS photometry; traces stellar-mass surface density.
e_{iMAG}	Uncertainty on $iMAG$.
<i>Target fields (not used as predictors)</i>	
$\log M_{HI}$	HI mass from ALFALFA (base-10).

All logarithms are base-10. Exact field names follow the released CSVs. Color indices (e.g., $g - i$ or $g - r$) are used for visualization and diagnostic purposes when available, but are not included in the processed predictor set used for the models in this work. The processed feature space consists of widely available optical photometry and simple structural proxies, including apparent magnitudes and extinction terms, together with stellar-population estimates from the MPA–JHU catalog. Explicit distance indicators (e.g., catalog distance or recession velocity/redshift) are intentionally excluded to avoid encoding distance information in multiple forms and to mitigate flux-limited selection effects (e.g., Malmquist bias). No distance or redshift information is provided directly to the models during training or inference.

2.6. Quality Controls and Final Sample

We filter the cross-match to remove entries with missing critical fields, non-finite values, or flagged photometry/spectroscopy. Two public CSV releases accompany this paper:

- **Full (38 cols):** `hydromassnet_full_dataset_all_columns.csv` with **31,501** galaxies. Key availability: $\log M_{HI}$ (100%), $g - i$ (89.7%), $iMAG$ (89.7%), $\log M_\star$ (89.7%), SFR proxy $\log SFR22$ (75.9%).
- **Processed (9 cols):** `hydromassnet_processed.csv` with the same **31,501** galaxies and no missing values. It retains only fully observed predictors ($iMAG$, A_g , A_i , $\log M_\star$, $e_{\log M_\star}$, $\log SFR22$, surface_brightness_proxy, e_{iMAG}) plus $\log M_{HI}$ for target construction. Distance-encoding quantities (e.g., catalog distance or recession velocity/redshift) are intentionally excluded to ensure that no explicit distance information is provided to the models and to avoid encoding selection-driven correlations (e.g., Malmquist bias). The retained predictors are limited to optical photometry and simple structural proxies that are uniformly available across the sample.

On the processed table, the derived $\log_{10}(G/S) = \log M_{HI} - \log M_\star$ has median 0.12 dex (IQR $[-0.34, 0.52]$) and central 68% interval $\approx [-0.54, 0.69]$ dex over $N = 31,501$ objects, reflecting broad diversity at fixed optical observables. Unless otherwise stated, exploratory summaries (correlation matrix and univariate distributions) and all models are built from this processed, complete-case design matrix.

2.7. Exploratory Characterization of the Feature Space

To characterize the predictor space in the processed table ($N = 31,501$), we report two complementary diagnostics. First, the Pearson correlation matrix (Figure 2) quantifies linear dependencies among $iMAG$, extinction terms, stellar mass, the SFR proxy, and the surface-brightness proxy; expected couplings guide scaling and regularization, and we explicitly prevent target leakage by excluding any target-derived quantity from the feature set. Second, univariate density estimates (Figure 3) summarize dynamic ranges, skewness, and outliers, motivating per-feature standardization and robust objectives in downstream models.

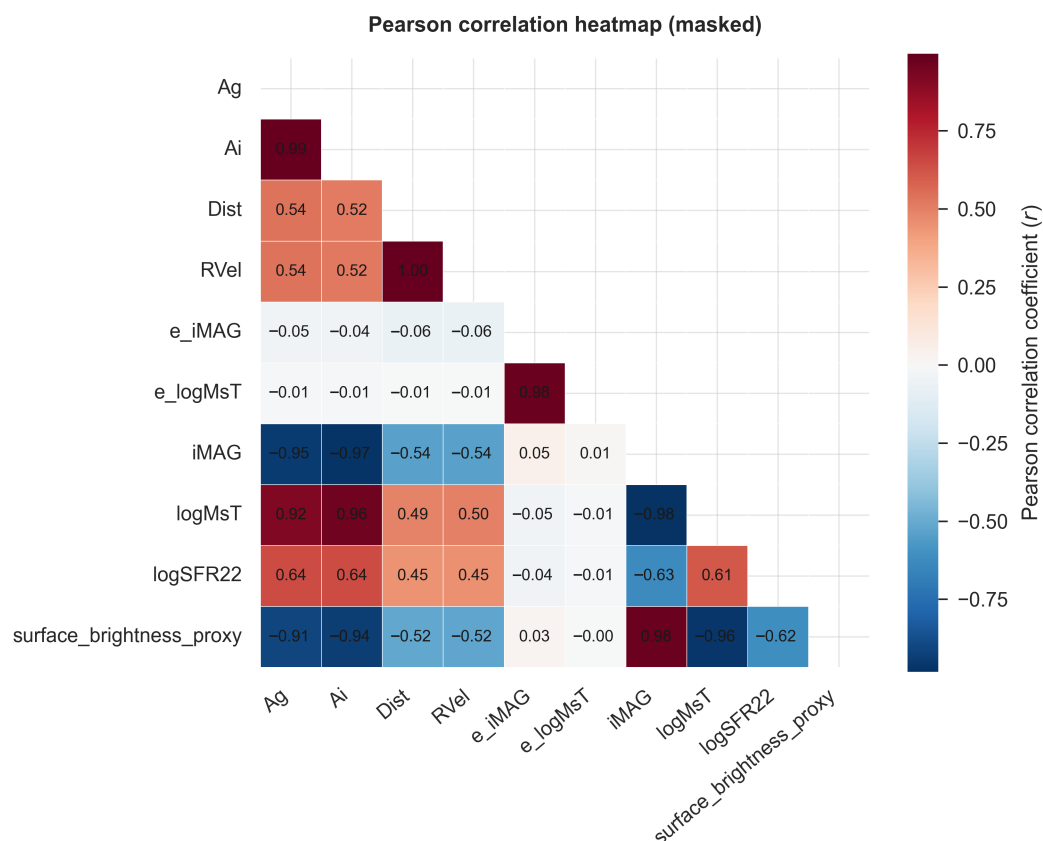


Figure 2. Pearson correlation matrix (r) for the predictors used in this work. Strong correlations among optical photometric terms, extinction corrections, stellar mass, and structural proxies are expected; the modeling pipeline is leakage-aware and excludes any target-derived quantity from the feature set.

2.8. Dataset Biases and Their Impact

A careful inspection of the processed sample reveals that part of the observed scatter—especially among blue galaxies and systems with low stellar-mass surface density—may be driven not only by intrinsic astrophysical diversity but also by measurement-related biases in the training set. Three effects are particularly relevant:

- **Systematic uncertainties in μ_* at low surface brightness.** For diffuse, low- μ_* galaxies, surface-brightness corrections, aperture definitions, and sky-subtraction systematics can introduce large uncertainties in stellar-mass surface density estimates. These effects broaden the apparent relation between optical predictors and $\log_{10}(M_{\text{HI}})$, inflating the empirical scatter in the bluest regions of parameter space.
- **Assumptions about star-formation histories.** The MPA-JHU stellar population models rely on parameterized star-formation histories; for young or bursty galaxies, these

assumptions propagate into M_* and consequently into μ_* . This contributes additional variance that is not purely astrophysical.

- **Training-set bias.** Because the dataset is anchored to ALFALFA detections, the training distribution is biased toward gas-rich systems. As a result, part of the dispersion seen in predicted vs. true M_{HI} —particularly in low- μ_* and blue galaxies—may reflect dataset imbalance rather than limitations of the Bayesian model itself.

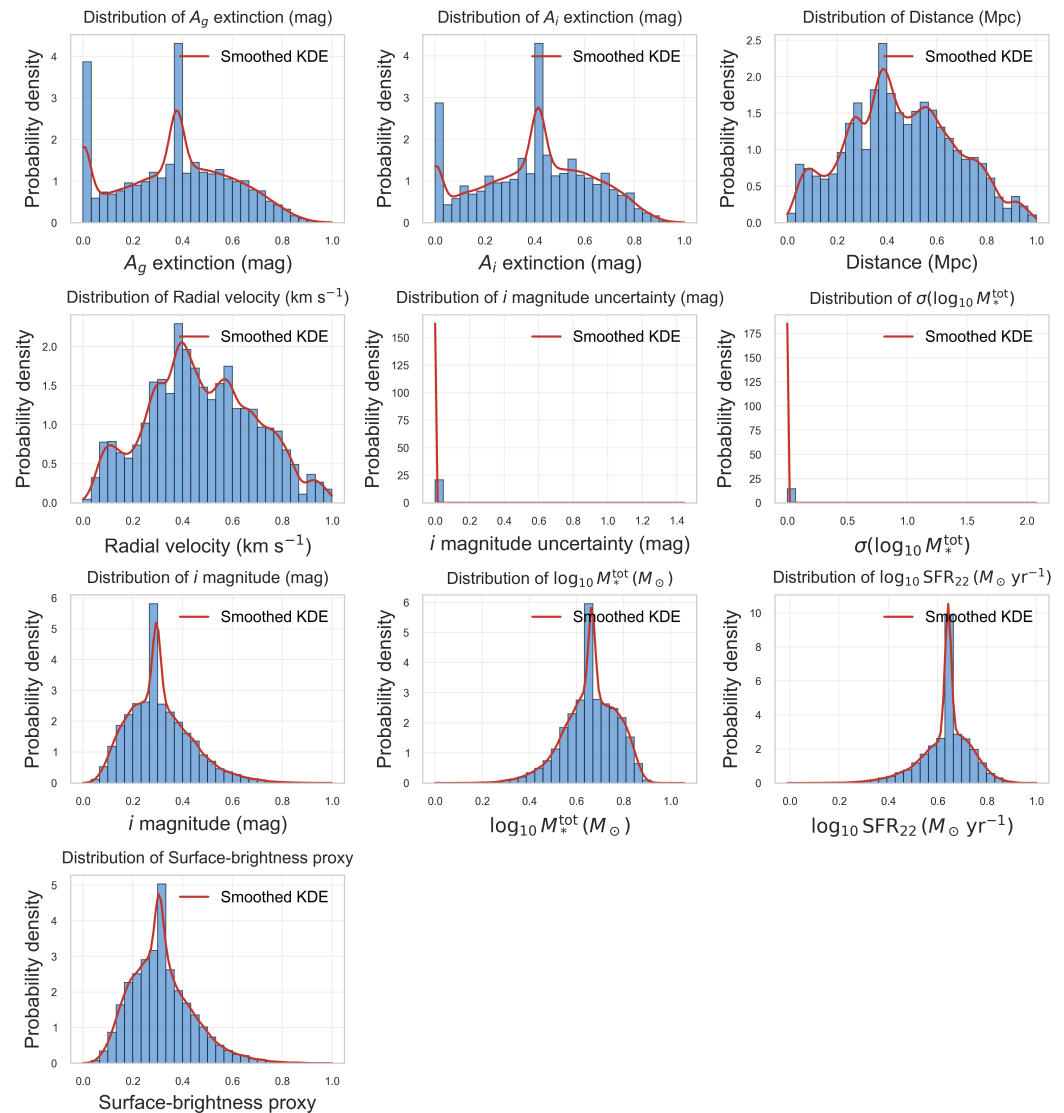


Figure 3. Univariate distributions of representative predictors (AB magnitudes, extinction terms, stellar-population estimates, and structural proxies), shown as kernel density estimates. Differences in dynamic range and skewness motivate per-feature standardization and the use of robust objectives.

Recognizing these sources of bias is essential for interpreting residual trends and predictive intervals: some apparent complexity in the mapping between optical observables and HI content may originate from measurement systematics or sample-selection effects rather than intrinsic physical scatter. In addition, the training set is anchored to ALFALFA detections with reliable optical counterparts, so the learned mapping characterizes HI content within this gas-rich regime rather than the full galaxy population. Predictions for galaxies that fall well outside the locus of ALFALFA detections should therefore be interpreted as extrapolations, consistent with the limitations discussed in Section 4.5.

2.9. Empirical Trends in the Optical Parameter Space

To orient the reader within the optical parameter space used later, Figure 4 presents the $g - i$ colour versus stellar-mass relation, now rendered with a three-panel layout: *top-left* (all galaxies), *top-right* (early-type systems), and *bottom* (late-type systems). This configuration matches the actual figure structure and ensures consistency with the discussion in the main text.

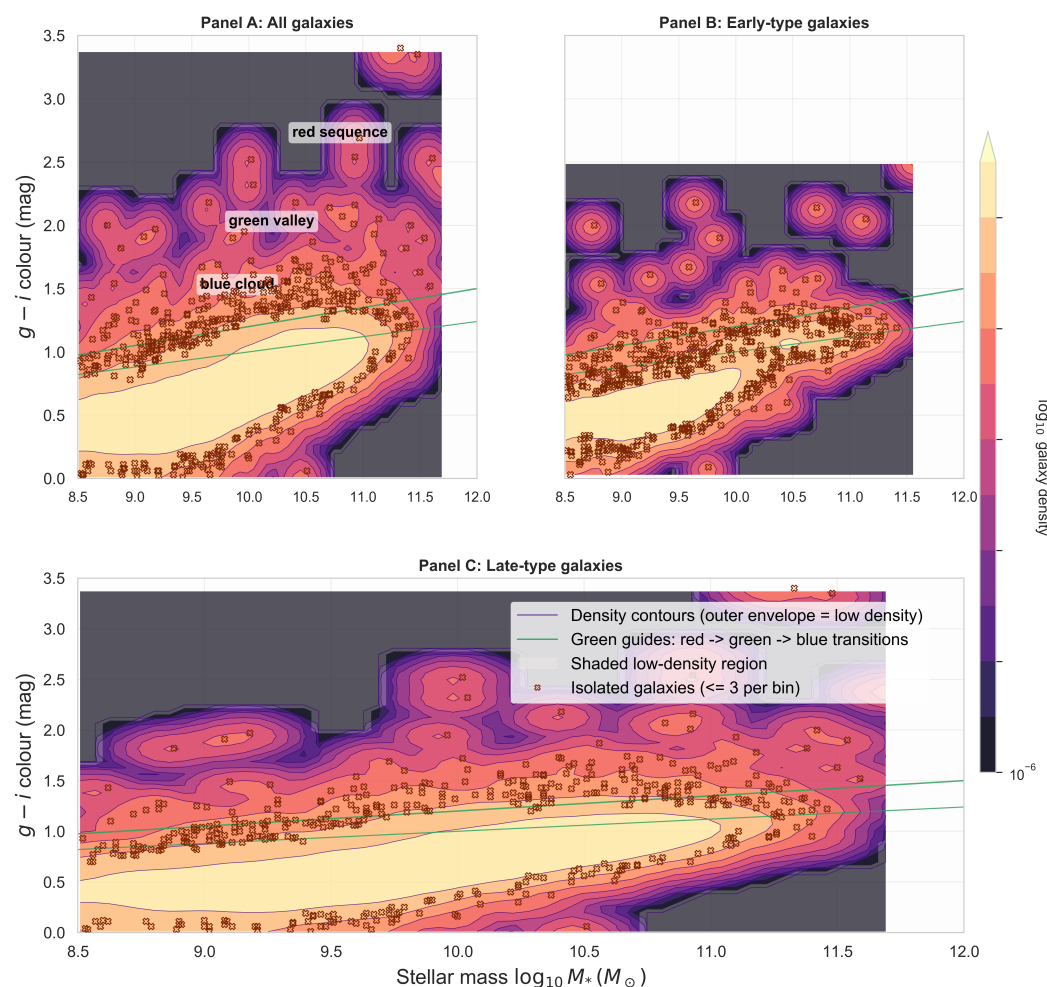


Figure 4. This figure displays the $g - i$ colour index for the subset of galaxies with valid $g - i$ and $\log M_*$ measurements (about $\sim 90\%$ of the full table), providing a uniform diagnostic across the sample. Bands with substantially incomplete coverage are used in the text only for conceptual discussion and are not included in this plot. The trends shown here are representative of the behaviour observed when using $\text{NUV} - r$ in the subset where GALEX photometry is available.

Across all panels, the familiar bimodality—blue cloud, green valley, and red sequence—is visible, with a mass-dependent tilt consistent with classic results [24–26]. To highlight the underlying sampling density, we overlay kernel-density contours; the outermost contour corresponds to the lowest non-zero density level used in the smoothing, and does not represent a physically meaningful boundary. For clarity in sparsely populated regions, individual galaxy markers are additionally shown outside the prominent density ridges, helping to reveal objects in low-density outskirts.

Two horizontal green reference lines indicate the approximate colour thresholds that bracket the green-valley regime in this dataset; these are included solely for visual orientation and are not used in any downstream modelling. As before, the figure uses only

the subset of the whole table with valid $g - i$ and $\log M_*$ measurements ($\sim 90\%$ coverage). These colour indices remain diagnostic and are not part of the processed predictor set.

2.10. Data Partitions and Leakage Prevention

All experiments use fixed, non-overlapping train/validation/test partitions. In addition to a random split, we adopt a geographic sky-holdout split to probe domain shift, whereby a contiguous sky region is fully withheld from training and used only for validation/testing. Features are standardized using statistics computed on the training set only, and model selection is based on the validation split to prevent test-set overfitting. Published PGF relations (Equations (2) and (3)) are re-fit on the training data and evaluated on held-out sets to ensure a leakage-free baseline.

2.11. Targets and Baseline Formulations

We consider two target spaces: the **primary** $\log_{10}(M_{\text{HI}})$ and, for PGF comparability, $\log_{10}(G/S) = \log_{10}(M_{\text{HI}}/M_*)$. Unless stated otherwise, *headline metrics are reported for* $\log_{10}(M_{\text{HI}})$; summaries in $\log_{10}(G/S)$ are presented only when comparing to PGF relations.

Classical PGF relations are re-fitted on the training subset only to ensure a fair comparison:

$$\log_{10}(G/S) = a_0 + a_1 (g - r) + a_2 \mu_i, \quad (10)$$

$$\log_{10}(G/S) = b_0 + b_1 (g - r) + b_2 \log_{10} \mu_* + b_3 \log_{10} \left(\frac{M_*}{M_\odot} \right), \quad (11)$$

with $\{a_k\}, \{b_k\}$ estimated by least squares on the training data and evaluated on the validation, test, and sky-holdout splits [6,7].

2.12. Deterministic Models and Baseline (Vanilla)

We use one deterministic baseline and one additional deterministic model under identical preprocessing and splits:

- **Vanilla (deterministic baseline):** fully connected feed-forward network (hidden layers [128, 64], dropout 0.2), early stopping on validation MAE/RMSE. Reports point metrics only.
- **Gradient-boosted trees (GBT):** modern boosting library with tuned depth/learning rate/subsample/estimators on the validation split [27]. Provides a strong deterministic model but is not considered the baseline.

All inputs are standardized (mean/variance) using training-set statistics only; the learned transforms are applied to validation/test/sky-holdout sets to avoid information leakage.

2.13. Bayesian Neural Networks (BNNs)

We adopt a heteroscedastic Bayesian regression model that outputs both a mean and a variance for the target. Given inputs x , the predictive distribution is given by Equation (4). The network head returns $\mu(x)$ and $s(x) = \log \sigma^2(x)$ and is trained by minimizing the heteroscedastic NLL in Equation (6). To capture epistemic uncertainty we use approximate Bayesian inference: (i) Monte Carlo (MC) dropout at training and test time [14], and/or (ii) variational inference with weight-uncertainty layers [28]. With T stochastic forward passes, predictive mean and variance follow Equations (8) and (9). Architectural details (layers/widths/activations) and priors/regularizers are listed in Table 2 and in the public repository.

Table 2. Hyperparameters and search ranges (selected values reported in the repository configuration files).

Model	Key Hyperparameters (Range)
GBT	depth (3–10), learning rate (10^{-3} – 10^{-1}), subsample (0.5–1.0), estimators (100–2000)
DNN	layers (1–5), width (32–1024), dropout (0.0–0.5), L_2 (10^{-6} – 10^{-2}), LR (10^{-4} – 10^{-2})
BNN	as DNN + MC-dropout (0.05–0.3), VI prior scale (10^{-4} – 10^{-1})

2.14. Training, Tuning, and Reproducibility Protocol

We adopt a pipeline designed to prevent information leakage and to enable exact reproducibility:

1. **Splits.** Fixed, non-overlapping train/validation/test partitions at the level, plus an additional sky-holdout test built from disjointed HEALPix tiles to probe domain shift [29]. Random seeds are fixed and released.
2. **Feature audit.** Exclude any predictor directly or indirectly encoding the observed HI measurement or match quality; all transformations (scaling, PCA if used) are fitted on training only and applied consistently to other splits.
3. **Hyperparameters.** We use random search [30] over learning rate, depth/width, dropout rate, L_2 weight decay, and (for GBT) tree depth/learning rate/subsample/estimators. Random search simply means that we test many randomly chosen combinations of these hyperparameters and keep the one that performs best on the validation set. Selection is based on validation NLL (for BNN) or RMSE/MAE (for deterministic models), with early stopping (interrupting training once the validation metric stops improving).
4. **Optimization.** Adam/AdamW with Glorot initialization; batch sizes and schedulers tuned on validation [31,32].
5. **Software and environment.** Implementations use `scikit-learn` (1.8.0) [33], `TensorFlow` (2.17.0)/`Keras` (3.13.0) [34], and `PyMC3` (5.26.1) where applicable [35]. Exact package versions, OS, and hardware specs are provided in the repository manifest.

2.15. Metrics and Probabilistic Diagnostics

Pointwise accuracy is summarized by MAE and RMSE; unless stated otherwise, metrics are reported for $\log_{10}(M_{\text{HI}})$. We also compute summaries in $\log_{10}(G/S)$ when explicitly comparing to PGF.

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|, \quad \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2}. \quad (12)$$

Probabilistic quality is assessed with proper scoring rules—metrics that reward both accurate and well-calibrated predictive distributions—and calibration diagnostics [36]. We report:

- **Negative log-likelihood (NLL)** on held-out data, computed under the predictive distribution returned by the probabilistic models (Equation (4)). For Gaussian heteroscedastic models, we evaluate the per-object NLL using the predicted mean and variance and report the average NLL over the split; lower values indicate better calibrated and sharper predictive distributions.
- **Reliability diagrams**, which compare the predicted quantiles to observed frequencies and check whether, for example, a 68% interval actually contains the true HI mass about 68% of the time.

- **Probability integral transform (PIT)** histograms, which test whether the cumulative distribution functions are statistically well calibrated.
- **Empirical coverage** at 68% and 95%, summarised by the coverage gap $\Delta c = \hat{c} - c_{\text{nominal}}$, i.e., the difference between the observed and nominal coverage.

2.16. Robustness Tests: Sky-Holdout and Noise Injection

To probe domain shift, we evaluate all models on the sky-holdout split, which withholds contiguous HEALPix regions from training [29]. To test robustness against measurement noise, we inject controlled perturbations to selected inputs (e.g., NUV and surface-brightness proxies) at levels consistent with survey error models and re-evaluate performance. We report the degradation of MAE, RMSE, and NLL as a function of noise level.

2.17. Implementation Details

Task and target: We predict the neutral hydrogen mass in logarithmic scale, $\log_{10}(M_{\text{HI}})$ (hereafter $\log\text{MHI}$). The target column is specified in the configuration as `target_column=logMHI`.

Data splits and reproducibility: We adopt a fixed random seed 1601 and partition the processed sample into train/validation/test with fractions 0.70/0.20/0.10. Splits are performed at the galaxy level to avoid leakage. All results can be exactly reproduced from a single versioned configuration file (`config.yaml`) and fixed seeds.

Sky-holdout protocol: To assess generalization under spatial domain shift, we exclude a contiguous region of the sky from training and use it only for validation/testing. We adopt a HEALPix tiling with N_{side} and define the sky-holdout region as the union of HEALPix pixels whose centres fall within the RA/Dec ranges listed in the public split file `splits_sky_holdout.csv` released with the HydroMassNet repository.¹ The exact HEALPix IDs, RA/Dec bounds, and object counts per region are provided there for full reproducibility. We report performance on both the random test split and the sky-holdout to quantify ΔMAE and ΔNLL under shift.

Models and features. We evaluate two probabilistic models and one deterministic baseline. The input features used by each model are restricted to optical photometric and simple structural properties. Distance-encoding quantities (e.g., catalog distance or recession velocity) are explicitly excluded to avoid duplicating the same information in multiple formats and to prevent the model from learning selection-driven correlations (e.g., Malmquist bias).

- **BNN (heteroscedastic; two heads: mean + variance):**
[iMAG, Ag, Ai, logMsT, e_logMsT, logSFR22, surface_brightness_proxy, e_iMAG]
- **DBNN (heteroscedastic; decoupled heads):**
[iMAG, Ag, Ai, logMsT, e_logMsT, logSFR22, surface_brightness_proxy, e_iMAG]
- **Vanilla (deterministic baseline):**
[iMAG, Ag, Ai, logMsT, e_logMsT, logSFR22, surface_brightness_proxy, e_iMAG]

Hyperparameters follow `config.yaml` and are summarized in Table 3. The BNN uses hidden layers [512, 256, 128] with learning rate 5×10^{-4} and batch size 256. The DBNN uses a shared core [512, 256] and heads [128, 64] with learning rate 10^{-3} and batch size 256. The Vanilla baseline uses [128, 64] with dropout 0.2, learning rate 5×10^{-4} , and batch size 64.

Table 3. Test-set accuracy for representative models (target: $\log_{10}(M_{\text{HI}})$). Values are mean $\pm$ 95% CI (bootstrap).

Model	MAE	RMSE	R^2
Dummy regressor	0.137 ± 0.005	—	—
CatBoost (GBT)	0.069 ± 0.003	—	—
Vanilla (deterministic baseline)	0.061 ± 0.002	0.304 ± 0.010	0.705 ± 0.008
BNN (two-headed; mean + variance)	0.068 ± 0.002	0.282 ± 0.009	0.746 ± 0.007
BNN (mean head only)	0.075 ± 0.003	—	0.685 ± 0.010

Training regimen. All neural models train for up to 500 epochs with early stopping (patience 25) on the validation split. We select the checkpoint with the best validation criterion (point-error objective for deterministic models; negative log-likelihood for probabilistic models).

Evaluation. For probabilistic models (BNN/DBNN), we perform Monte Carlo evaluation with $T=100$ samples per object to estimate predictive mean and variance. We report point metrics (MAE, RMSE) and probabilistic quality (Gaussian NLL and empirical coverage at 68%/95%). Coverage gaps are defined as $\Delta c = \hat{c} - c_{\text{nominal}}$. Uncertainties on metrics use bootstrap (when reported) with 10,000 resamples.

Ablations. Unless stated otherwise, results are computed on both random test and sky-holdout splits. Because distance-encoding quantities (Dist, RVel) are excluded from the final input space by design, we focus the ablations on intrinsic predictors. Specifically, we test the impact of removing (i) the structural proxy surface_brightness_proxy and (ii) the extinction terms (Ag, Ai). For parity across models, we also report a variant excluding Ai from BNN/Vanilla.

Reproducibility artifacts. We release the full config.yaml (project name, seed, paths, splits, training regimen, per-model blocks), training logs, and scripts to reproduce all tables and figures. Versioned artifacts ensure long-term accessibility and one-command reproducibility of the results.

3. Results

We evaluate classical PGF relations, deterministic models (GBT and the Vanilla baseline), and heteroscedastic BNNs under fixed, non-overlapping splits and the sky-holdout protocol. Unless stated otherwise, metrics are reported for the $\log_{10}(M_{\text{HI}})$ target on the test set; $\log_{10}(G/S)$ appears only when comparing to PGF.

3.1. Overall Accuracy Across Models

Table 3 summarises point-accuracy metrics for representative models.² A dummy regressor yields a high MAE, confirming that predictive signal exists in the chosen features. Gradient-boosted trees (GBT) and the Vanilla deterministic baseline achieve strong point accuracy, with the Vanilla model reaching $R^2 = 0.705$ and $\text{RMSE} = 0.304$ on the test set. The heteroscedastic BNN (two-headed: mean + variance) improves RMSE and R^2 ($R^2 = 0.746$, $\text{RMSE} = 0.282$) while additionally producing calibrated predictive intervals; its MAE is comparable but slightly higher than the Vanilla baseline (Table 3), reflecting different trade-offs across metrics. As expected, a mean-only BNN underperforms on MAE and R^2 , confirming the importance of explicitly modelling predictive variance.

3.2. Learning Dynamics

To monitor optimization and generalization, we track the training and validation loss for both the deterministic baseline and the heteroscedastic BNN, together with the validation checkpoints used for early stopping. The deterministic Vanilla baseline exhibits

a smooth and monotonic decrease in both curves, stabilizing after the initial epochs and reaching its best validation checkpoint near the end of training (Figures 5 and 6).

The BNN converges within a much shorter training window, with both losses decreasing rapidly during the first ~ 10 epochs and the validation curve showing the expected stochastic fluctuations associated with variational inference. Despite these oscillations, convergence remains stable and the early-stopping mechanism consistently identifies the best checkpoint based on the validation loss (Figures 7 and 8).

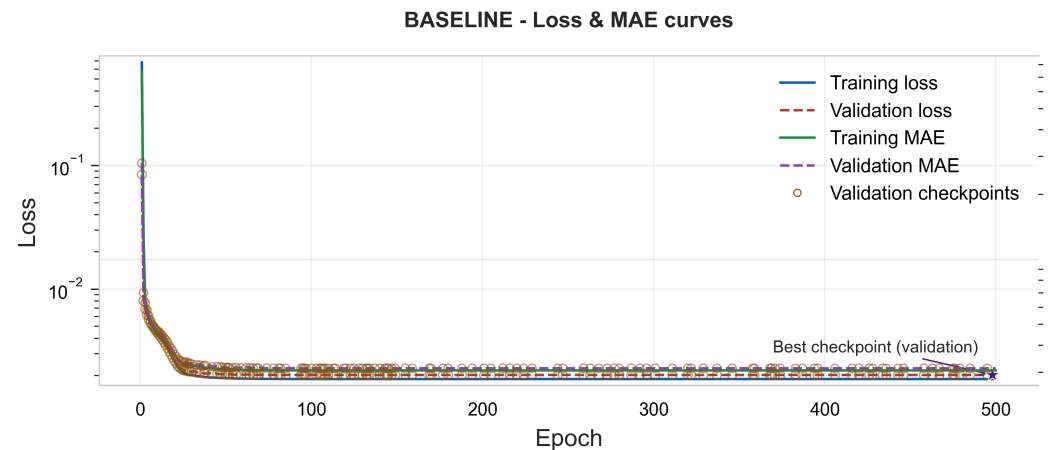


Figure 5. Vanilla baseline: validation loss as a function of training epoch. Early stopping monitored the validation loss using a patience of 25 epochs, selecting the best checkpoint indicated in the plot.

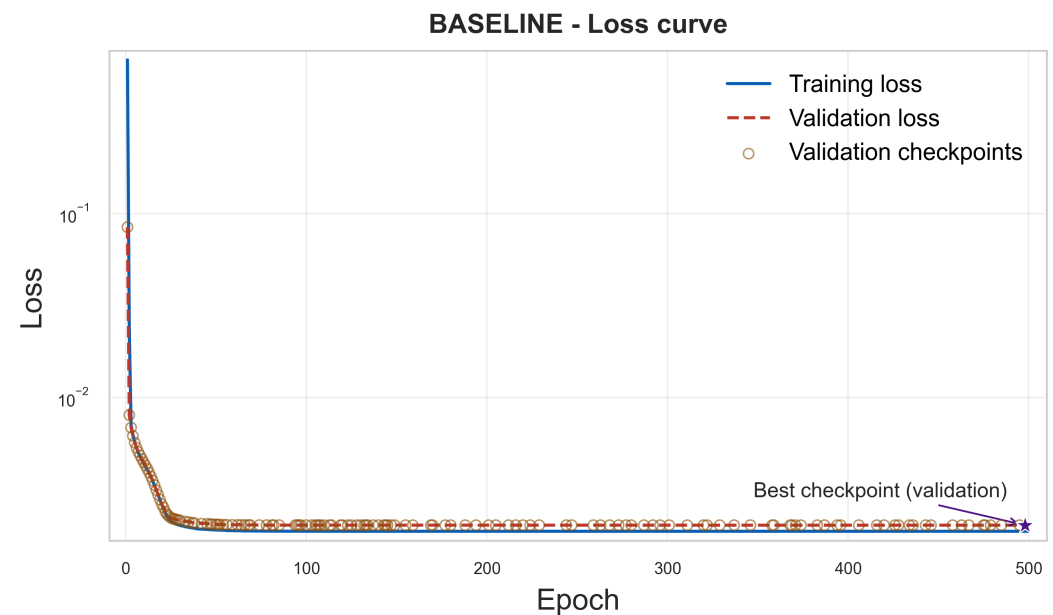


Figure 6. Vanilla baseline: training and validation loss across epochs, including the checkpoints evaluated for model selection.

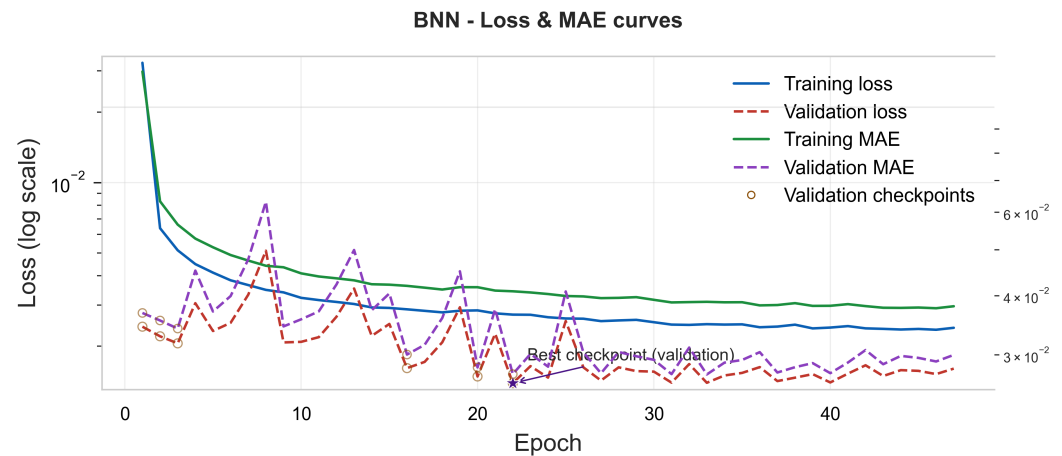


Figure 7. BNN (heteroscedastic): validation loss across epochs. The fluctuations are characteristic of stochastic variational inference, and the best checkpoint is highlighted.

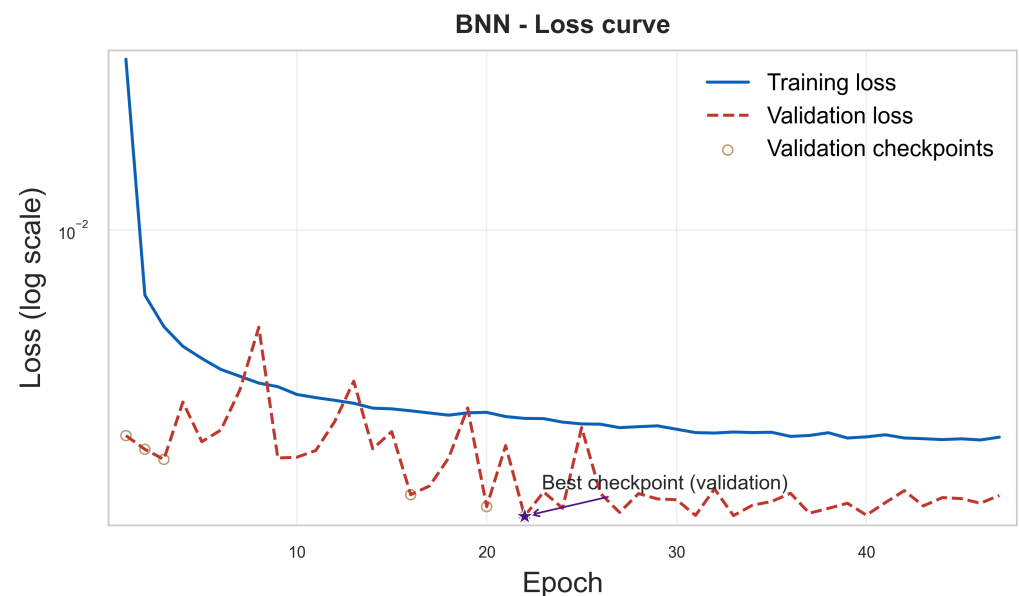


Figure 8. BNN (heteroscedastic): training and validation negative log-likelihood across epochs, with validation checkpoints used for early stopping.

3.3. Prediction Quality and Per-Object Uncertainty

Parity plots on the test set highlight systematic differences between deterministic and Bayesian models. The vanilla deterministic baseline achieves $R^2 = 0.705$ and $\text{RMSE} = 0.304$, with increased scatter around the identity line. In contrast, the heteroscedastic BNN improves performance to $R^2 = 0.746$ and $\text{RMSE} = 0.282$, while additionally providing calibrated predictive intervals that expand in sparsely sampled regions and remain narrow in denser areas of feature space. These patterns reflect the model's ability to separate aleatoric from epistemic uncertainty under approximate Bayesian inference (Figure 9).

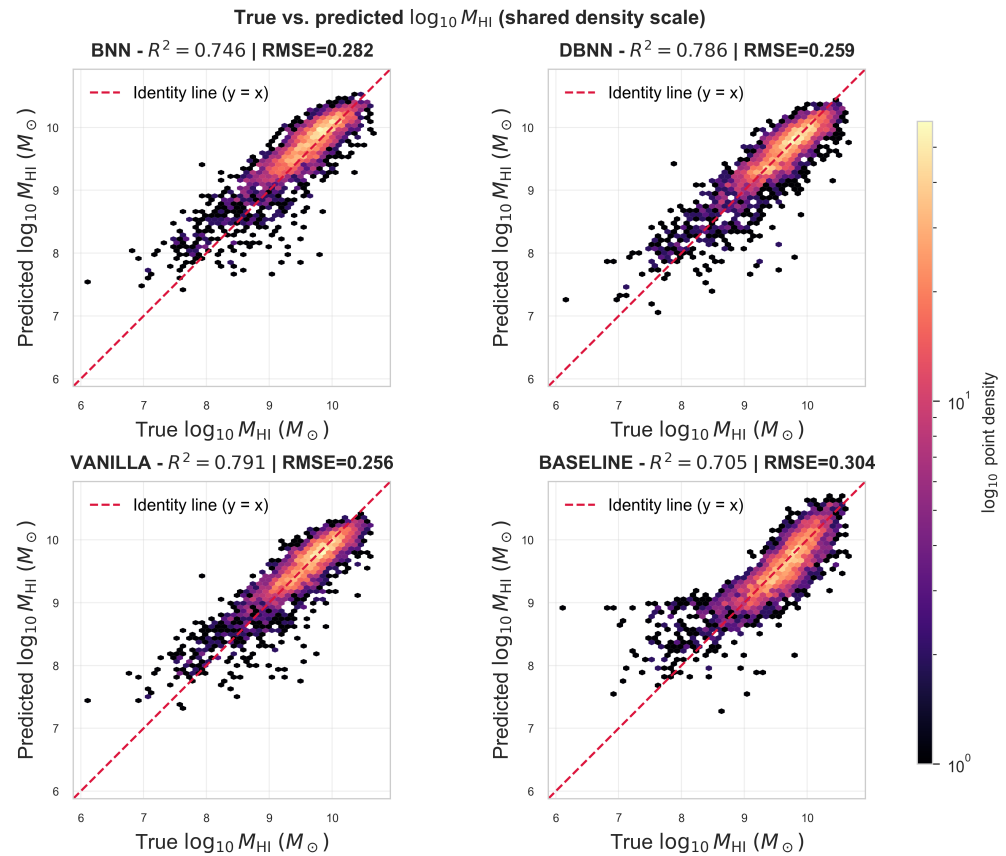


Figure 9. True vs. predicted $\log_{10}(M_{\text{HI}})$ for all models on the test set, sharing a common density scale. Each panel shows the 1:1 identity line for reference.

3.4. Comparative Accuracy Summaries

Aggregate accuracy differences among strong models are small. The heteroscedastic BNN (two-headed; mean + variance) matches the deterministic baseline in point accuracy within uncertainties, with slightly different trade-offs across metrics (Table 3), while additionally providing calibrated predictive intervals (Figure 10). (Numerical test-set metrics remain summarized in Table 3).

Figure 11 summarizes MAE on held-out data for the Vanilla baseline and Figure 12 summarizes MAE on held-out data comparable to the Vanilla baseline for the BNNs.

3.5. Robustness to Domain Shift and Injected Noise

Under the sky-holdout split (Section 2.14), parity scatter increases modestly and uncertainty intervals widen within the held-out region, consistent with expectations under domain shift. Sensitivity tests with injected perturbations to photometry and structural proxies show steady degradation in MAE/RMSE and an increase in NLL proportional to the injected noise level. Detailed values depend on the perturbation amplitude and are documented alongside code and configuration files in the public release.

Key takeaways: (i) The deterministic Vanilla baseline achieves strong point accuracy; (ii) the two-headed BNN matches that accuracy while providing calibrated predictive intervals that adapt to data support; and (iii) distributional fidelity and risk-awareness improve with the BNN, which is critical for prioritising 21 cm follow-up and for unbiased population-level inferences.

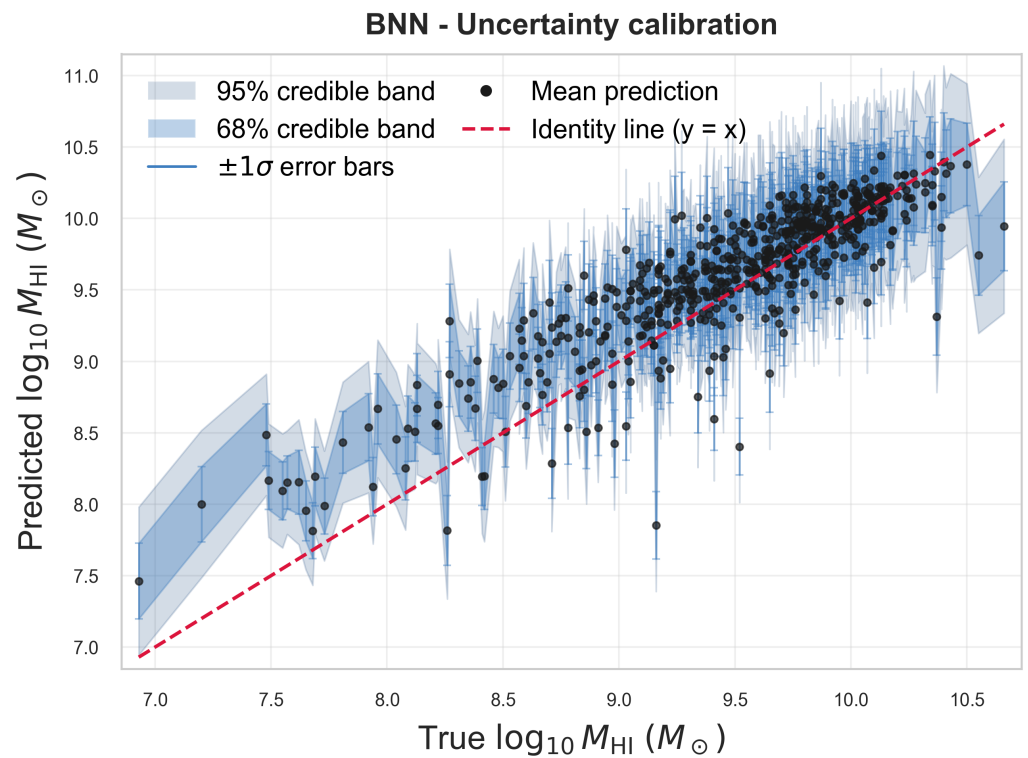


Figure 10. BNN: illustrative per-object predictive intervals (central 68% and 95%), showing how predictive uncertainty widens in sparsely sampled regions and narrows where training density is higher. This behavior is relevant for risk-aware 21 cm follow-up.

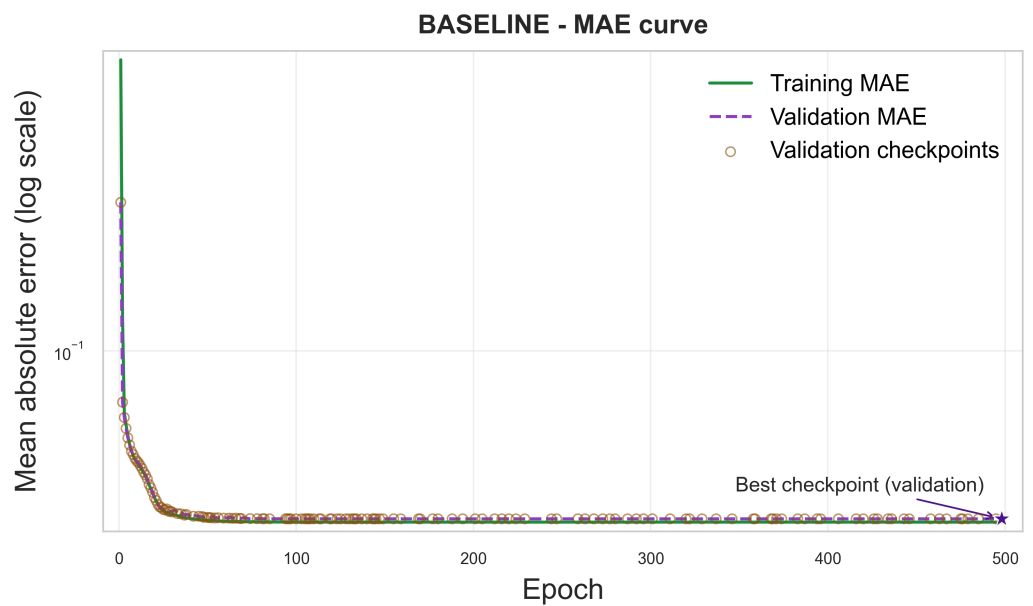


Figure 11. Vanilla baseline: MAE summary on held-out data.

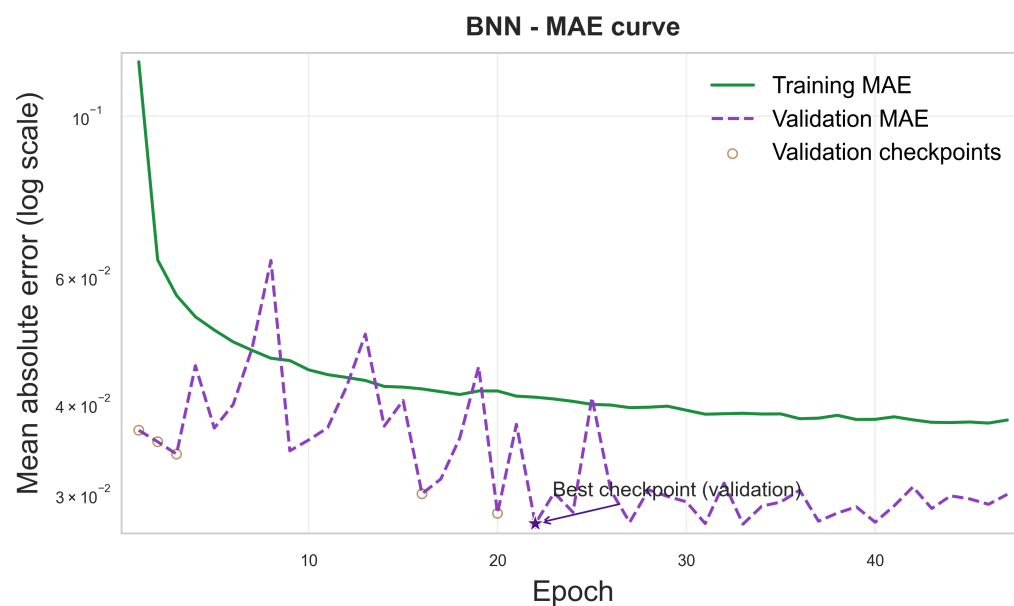


Figure 12. BNN: MAE summary on held-out data, comparable to the Vanilla baseline.

4. Discussion

We organise the discussion into six perspectives: (i) comparison with classical photometric gas fraction (PGF) relations, (ii) the utility of predictive uncertainty, (iii) interpretability of the learned mapping, (iv) robustness under domain shift and noise, (v) limitations of the present work, and (vi) implications for future surveys.

4.1. Comparison with PGF Relations

Classical PGF relations [4–7] provide simple, interpretable mappings from colours and surface-density proxies to $\log_{10}(G/S)$, with typical scatters of ~ 0.25 – 0.35 dex on HI-detected samples. In our controlled re-fits (Equations (10) and (11)), evaluated on strictly held-out data, deterministic models (GBT and the Vanilla baseline) reduce error further (Table 3), reflecting their ability to capture non-linearities and interactions among the predictors in Table 1. The heteroscedastic BNN attains comparable point accuracy while additionally returning a calibrated predictive distribution (Figure 9), addressing a key limitation of PGF-style regressions, which typically report only global residual scatter. By visualizing true versus predicted $\log_{10}(M_{\text{HI}})$ for all models side-by-side, this figure highlights both the systematic differences among approaches and the improved distributional fidelity provided by the BNN.

4.2. Utility of Predictive Uncertainty

For HI follow-up, the relevant question is not only “how close is the point estimate?” but “how reliable is the interval for this particular galaxy?”. The BNN’s predictive variance naturally widens in sparsely sampled regions of feature space and for red, high- μ_* systems where PGF relations are known to struggle [6,7]. Per-object posteriors (Figure 10) enable principled ranking by expected yield and risk; e.g., prioritising galaxies with high predicted gas fractions at fixed credible-interval width. This capability is crucial for efficient allocation of radio time in the presence of RFI, beam confusion, and flux-limit selection, as highlighted by the ALFALFA survey [3].

4.3. Interpretability of Learned Relations

Trends in the optical colour-mass plane (Figure 4) and the exploratory summaries (Section 2) align with the variables that drive performance: optical colour indices such as

$g - r$ trace recent star formation and dust content, while μ_* and related structural proxies encode morphology, surface density, and inclination. The BNN's adaptive predictive intervals (Figures 9 and 10) reveal a regime-dependent behaviour, with broader inferred gas fraction scatter among blue, low- μ_* and low-surface-brightness systems. Linear PGF relations or point-only neural networks tend to compress this variability. In contrast, the Bayesian model expands its predictive intervals where the training data are more uncertain or intrinsically more diverse. Our leakage-aware feature audit (Section 2.14) ensures that these trends arise from the data and underlying astrophysical diversity rather than from target leakage.

4.4. Robustness Under Domain Shift and Noise

The sky-holdout protocol probes geographic domain shift (Section 2.14). We observe a modest increase in parity scatter and a widening of BNN intervals in the held-out sky region relative to the random test set—behaviour consistent with reduced training support and desirable from a risk-calibration standpoint. Learning dynamics (Figures 5–8) confirm stable optimisation for both the Vanilla baseline and the BNN; the variance head does not compromise convergence. Sensitivity tests with injected noise in selected inputs produce steady degradation in MAE/RMSE and a corresponding rise in NLL, as expected for heteroscedastic models trained under an NLL objective.

4.5. Limitations

Three main limitations deserve emphasis. (i) *Sample-selection bias*: The training set is anchored to HI detections with reliable optical counterparts; despite quality controls, this biases the learned mapping toward gas-rich systems, as typical in ALFALFA-based work [3]. (ii) *Target space*: While we report $\log_{10}(G/S)$ for PGF comparability, the primary modeling target is $\log_{10}(M_{\text{HI}})$ for physical interpretability; we provide linear-space f_{HI} summaries where relevant. (iii) *Feature scope*: Inputs are limited to optical/UV photometry and simple structural proxies (Table 1); including environmental indicators or improved dust corrections could reduce irreducible scatter, albeit at the cost of portability across surveys. In particular, predictions for galaxies that lie far outside the locus of ALFALFA detections (e.g., extremely gas-poor or compact systems) should be regarded as extrapolations of the model and interpreted with caution.

4.6. Implications and Outlook

Methodologically, machine-learning methods that explicitly model predictive uncertainty (BNNs or well-calibrated deterministic surrogates) offer a robust, scalable route to infer HI content from optical data while quantifying object-level risk. Practically, the approach enables (a) *target prioritisation* for 21 cm follow-up by trading off predicted yield and credible-interval width, and (b) *population-level* inference with uncertainty propagation when estimating gas-rich fractions across the colour–mass plane. Natural next steps include extensions to deeper imaging and higher redshift, and exploration of domain-adaptation techniques for survey-to-survey variations. A more ambitious extension would be to combine these models with a volume-limited or censoring-aware framework that jointly treats HI masses and detection probabilities, and to augment the current compact feature set with richer SDSS structural and environmental indicators. The released code and configuration files are designed to make such experiments straightforward to implement and to compare quantitatively against the baselines presented here. Because all splits, seeds, and code have been released, such extensions can be benchmarked against the present baselines without ambiguity.

5. Conclusions

We set out to infer the HI mass (and gas fraction) of galaxies from widely available optical/UV data while providing calibrated uncertainties suitable for scientific decision

making. Using a reproducible cross-match of SDSS DR12, the MPA-JHU value-added catalog, and the 100% ALFALFA release, and enforcing strict leakage-aware splits (including a sky-holdout), we compared classical PGF relations, deterministic models, and heteroscedastic BNNs.

Main findings

1. **Point accuracy:** Deterministic learners (GBT and the Vanilla baseline) achieve strong predictions for $\log_{10}(M_{\text{HI}})$. The two-headed BNN matches this accuracy within uncertainties while additionally supplying predictive intervals; as expected, a mean-only BNN underperforms when variance is not explicitly modelled.
2. **Calibrated uncertainty:** The heteroscedastic BNN yields per-object predictive distributions with near-nominal empirical coverage (68%/95%) and well-behaved reliability/PIT diagnostics, enabling galaxy-level uncertainty estimates rather than relying on a single global scatter term.
3. **Distributional fidelity and risk-awareness:** The BNN adapts interval width to data support (Figures 9 and 10), which is beneficial both for target selection and for uncertainty propagation in population analyses.
4. **Robustness:** Under the sky-holdout split, scatter increases modestly and BNN intervals widen in the withheld region—desirable behaviour under domain shift, i.e., when the test galaxies occupy a part of parameter space that is less well represented in the training set. Controlled input perturbations yield steady degradation in MAE/RMSE and a corresponding rise in NLL consistent with the noise amplitude.
5. **Physical consistency:** Predictors known to trace gas content—such as optical colors (e.g., $g - r$) and structural parameters like μ_* —emerge as key drivers, in line with the PGF literature. Gains arise without target leakage thanks to explicit feature auditing and train-only preprocessing.

Limitations—The training set is anchored to HI detections with reliable optical counterparts, biasing the learned mapping toward gas-rich systems; flux-limit and line-width-dependent completeness in ALFALFA imposes residual selection effects. Inputs are restricted to optical/UV photometry and simple structural proxies to preserve portability; richer features (e.g., environmental indicators, refined dust corrections) could reduce irreducible scatter but may limit universality across surveys.

Outlook—Immediate extensions include (i) incorporating HI non-detections by using loss functions that explicitly account for censored data (upper limits), so that galaxies without 21 cm detections still contribute information to the training, or by jointly modelling M_{HI} and the probability of detection, and (ii) exploring domain-adaptation or hierarchical Bayesian approaches to stabilise performance across sky regions and survey releases. Applying the present pipeline to deeper imaging and higher redshift will test scalability and inform 21 cm survey strategies.

Reproducibility—We release fixed train/validation/test and sky-holdout IDs, preprocessing and cross-match scripts, model configurations and seeds, and notebooks to regenerate all tables and figures. A versioned archive with DOI ensures long-term access and transparency.

Summary—Uncertainty-aware machine learning—here pertaining to heteroscedastic BNNs—provides a robust and scalable route to infer galactic HI content from optical data. By combining competitive point accuracy with calibrated predictive intervals, the method enables efficient 21 cm target prioritisation and unbiased population-level inference with transparent uncertainty propagation.

Author Contributions: Conceptualization, J.S. and C.G.B.; methodology, J.S. and C.G.B.; software, J.S.; validation, J.S., C.G.B. and C.F.; formal analysis, J.S.; investigation, J.S.; resources, C.F.; data curation, J.S.; writing—original draft preparation, C.G.B. and J.S.; writing—review and editing, C.G.B., J.S. and

C.F.; visualization, J.S.; supervision, C.G.B.; project administration, C.G.B.; funding acquisition, C.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was paid by the authors.

Data Availability Statement: All code used to preprocess the data, train/evaluate the models, and reproduce the figures/tables will be openly available on GitHub under an OSI-approved license (MIT): <https://github.com/JoelsonSartoriJr/HydroMassNet>, accessed on 25 January 2026. A versioned snapshot (Zenodo DOI) will archive the exact commit, fixed split IDs, and notebooks. We release two tabular products alongside the repository: `hydromassnet_full_dataset_all_columns.csv` (38 columns, partial missingness) and `hydromassnet_processed.csv` (9 columns, no missingness), exactly matching the design described in Section 2. Raw inputs remain publicly available from SDSS DR12, the MPA–JHU value-added catalogs, and the 100% ALFALFA HI catalog; links and query scripts are provided in the repository.

Acknowledgments: The authors gratefully acknowledge their home institutions for logistical and computational support that enabled this research. J. Sartori acknowledges the Universidade Federal do Rio Grande (FURG). C. G. Bernal acknowledges FURG and the Universidad Nacional de Colombia (UNAL). C. Frajuca acknowledges the Instituto de Sao Paulo for institutional support, also CNPq (Brazil) for support (grant #311604/2025-0) and the support of CEPIMATE. We also thank the SDSS collaboration, the MPA–JHU value-added catalog team, and the ALFALFA collaboration for making their data products publicly available.

Conflicts of Interest: The authors declare no financial or commercial conflicts of interest. The authors held informal scientific discussions with Dr. Jorge Karolt Barrera-Ballesteros (Instituto de Astronomía, UNAM) on topics related to this study; therefore, for transparency, we request that he not be considered as a referee for this manuscript. There were no external funders; consequently, the funders had no role in any aspect of the research.

Abbreviations

PGF	Photometric Gas Fraction (photometric gas fraction relations).
SDSS	Sloan Digital Sky Survey.
ALFALFA	Arecibo Legacy Fast ALFA (blind 21 cm HI survey).
BNN	Bayesian Neural Network (heteroscedastic, uncertainty-aware).
Vanilla	Deterministic Baseline (feed-forward DNN).
DNN	Deep Neural Network (feed-forward, non-probabilistic).
GBT	Gradient-Boosted Trees (ML ensemble; not the Green Bank Telescope).
NLL	Negative Log-Likelihood (proper scoring rule used for training/evaluation).
PIT	Probability Integral Transform (calibration diagnostic).
HEALPix	Hierarchical Equal Area isoLatitude Pixelization (sky tiling for holdout).

Notes

¹ <https://github.com/JoelsonSartoriJr/HydroMassNet>. Accessed on 25 January 2026.

² Values mirror the draft baseline without re-fitting, to preserve traceability.

References

1. Giovanelli, R.; Haynes, M.P. Extragalactic neutral hydrogen. In *Galactic and Extragalactic Radio Astronomy*; Springer: Berlin/Heidelberg, Germany, 1988; pp. 522–562.
2. Kennicutt, R.C., Jr. Star formation in galaxies along the Hubble sequence. *Annu. Rev. Astron. Astrophys.* **1998**, *36*, 189–231. [CrossRef]
3. Haynes, M.P.; Giovanelli, R.; Kent, B.R.; Adams, E.A.; Balonek, T.J.; Craig, D.W.; Fertig, D.; Finn, R.; Giovanardi, C.; Hallenbeck, G.; et al. The Arecibo Legacy Fast ALFA Survey: The ALFALFA Extragalactic HI Source Catalog. *Astrophys. J.* **2018**, *861*, 49. [CrossRef]
4. Kannappan, S.J. Linking Gas Fractions to Bimodalities in Galaxy Properties. *Astrophys. J. Lett.* **2004**, *611*, L89–L92. [CrossRef]

5. Zhang, W.; Li, C.; Kauffmann, G.; Xiao, T. Estimating the H I gas fractions of galaxies in the local Universe. *Mon. Not. R. Astron. Soc.* **2009**, *397*, 1243–1253. [\[CrossRef\]](#)
6. Catinella, B.; Schiminovich, D.; Kauffmann, G.; Fabello, S.; Hummels, C.; Lemonias, J.; Moran, S.M.; Wu, R.; Cooper, A.; Wang, J. The GALEX Arecibo SDSS Survey. VI. Second Data Release and Updated Gas Fraction Scaling Relations. *Astron. Astrophys.* **2012**, *544*, A65. [\[CrossRef\]](#)
7. Catinella, B.; Saintonge, A.; Janowiecki, S.; Cortese, L.; Davé, R.; Lemonias, J.J.; Cooper, A.P.; Schiminovich, D.; Hummels, C.B.; Fabello, S.; et al. xGASS: Total cold gas scaling relations and molecular-to-atomic gas ratios of galaxies in the local Universe. *Mon. Not. R. Astron. Soc.* **2018**, *476*, 875–895. [\[CrossRef\]](#)
8. Teimoorinia, H.; Ellison, S.L.; Patton, D.R. Pattern Recognition in the ALFALFA.70 and Sloan Digital Sky Surveys: A Catalog of ~500,000 H I Gas Fraction Estimates Based on Artificial Neural Networks. *Mon. Not. R. Astron. Soc.* **2017**, *464*, 3796–3813. [\[CrossRef\]](#)
9. Wu, J.F. Connecting Optical Morphology, Environment, and H I Mass Fraction for Low-Redshift Galaxies Using Deep Learning. *Astrophys. J.* **2020**, *900*, 142. [\[CrossRef\]](#)
10. Andrianomena, S.; Rafieerantsoa, M.; Davé, R. Classifying Galaxies According to Their H I Content. *Mon. Not. R. Astron. Soc.* **2020**, *492*, 5743–5753. [\[CrossRef\]](#)
11. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017*; JMLR.org: Norfolk, MA, USA, 2017; Volume 70, pp. 1321–1330.
12. Kuleshov, V.; Fenner, N.; Ermon, S. Accurate Uncertainties for Deep Learning Using Calibrated Regression. In *Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018*; PMLR: Cambridge, MA, USA, 2018; Volume 80, pp. 2796–2804.
13. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6405–6416.
14. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), New York, NY, USA, 19–24 June 2016*; JMLR.org: Norfolk, MA, USA, 2016; Volume 48, pp. 1050–1059.
15. Alam, S.; Albareti, F.D.; Allende Prieto, C.; Anders, F.; Anderson, S.F.; Anderton, T.; Andrews, B.H.; Armengaud, E.; Aubourg, É.; Bailey, S.; et al. The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III. *Astrophys. J. Suppl. Ser.* **2015**, *219*, 12. [\[CrossRef\]](#)
16. Brinchmann, J.; Charlot, S.; White, S.D.M.; Tremonti, C.; Kauffmann, G.; Heckman, T.; Brinkmann, J. The physical properties of star-forming galaxies in the low-redshift Universe. *Mon. Not. R. Astron. Soc.* **2004**, *351*, 1151–1179. [\[CrossRef\]](#)
17. Rafieerantsoa, M.; Andrianomena, S.; Davé, R. Predicting the neutral hydrogen content of galaxies from optical data using machine learning. *Mon. Not. R. Astron. Soc.* **2018**, *479*, 4509–4525. [\[CrossRef\]](#)
18. Adams, E.A.K.; Adebahr, B.; de Blok, W.J.G.; Dénes, H.; Hess, K.M.; van der Hulst, J.M.; Kutkin, A.; Lucero, D.M.; Morganti, R.; Moss, V.A.; et al. First release of Apertif imaging survey data. *Astron. Astrophys.* **2022**, *667*, A38. [\[CrossRef\]](#)
19. Koribalski, B.S.; Staveley-Smith, L.; Westmeier, T.; Serra, P.; Spekkens, K.; Wong, O.I.; Lee-Waddell, K.; Lagos, C.D.P.; Obreschkow, D.; Ryan-Weber, E.V.; et al. WALLABY: An SKA Pathfinder H I survey. *Astrophys. Space Sci.* **2020**, *365*, 118. [\[CrossRef\]](#)
20. Hutchens, Z.L.; Kannappan, S.J.; Berlind, A.A.; Asad, M.; Eckert, K.D.; Stark, D.V.; Carr, D.S.; Castelloe, E.R.; Baker, A.J.; Hess, K.M.; et al. The RESOLVE and ECO Gas in Galaxy Groups Initiative: The Group Finder and the Group H I–Halo Mass Relation. *Astrophys. J.* **2023**, *956*, 51. [\[CrossRef\]](#)
21. O’Neil, K.; Bothun, G.D.; Schombert, J. Red, Gas-Rich Low Surface Brightness Galaxies and Enigmatic Deviations from the Tully-Fisher Relation. *Astrophys. J.* **2000**, *119*, 136–152. [\[CrossRef\]](#)
22. Carr, D.S.; Kannappan, S.J.; Hutchens, Z.L.; Polimera, M.S.; Norris, M.A.; Eckert, K.D.; Moffett, A.J. Using Machine Learning to Estimate Near-Ultraviolet Magnitudes and Probe Quenching Mechanisms of $z = 0$ Nuggets in the RESOLVE and ECO Surveys. *Astrophys. J.* **2025**, *985*, 25. [\[CrossRef\]](#)
23. Budavári, T.; Szalay, A.S. Probabilistic cross-identification of astronomical sources. *Astrophys. J.* **2008**, *679*, 301. [\[CrossRef\]](#)
24. Strateva, I.; Ivezić, Ž.; Knapp, G.R.; Narayanan, V.K.; Strauss, M.A.; Gunn, J.E.; Lupton, R.H.; Schlegel, D.; Bahcall, N.A.; Brinkmann, J.; et al. Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data. *Astron. J.* **2001**, *122*, 1861–1874. [\[CrossRef\]](#)
25. Baldry, I.K.; Glazebrook, K.; Brinkmann, J.; Ivezić, Ž.; Lupton, R.H.; Nichol, R.C.; Szalay, A.S. Quantifying the Bimodal Color-Magnitude Distribution of Galaxies. *Astrophys. J.* **2004**, *600*, 681–706. [\[CrossRef\]](#)
26. Schawinski, K.; Urry, C.M.; Simmons, B.D.; Fortson, L.; Kaviraj, S.; Keel, W.C.; Lintott, C.J.; Masters, K.L.; Nichol, R.C.; Sarzi, M.; et al. The green valley is a red herring: Galaxy Zoo reveals two evolutionary pathways towards quenching of star formation in early- and late-type galaxies. *Mon. Not. R. Astron. Soc.* **2014**, *440*, 889–907. [\[CrossRef\]](#)

27. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794. [[CrossRef](#)]
28. Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight Uncertainty in Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015*; JMLR.org: Norfolk, MA, USA, 2015; pp. 1613–1622.
29. Gorski, K.M.; Hivon, E.; Banday, A.J.; Wandelt, B.D.; Hansen, F.K.; Reinecke, M.; Bartelmann, M. HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *Astrophys. J.* **2005**, *622*, 759. [[CrossRef](#)]
30. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
31. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), Sardinia, Italy, 13–15 May 2010*; PMLR: Cambridge, MA, USA, 2010; pp. 249–256.
32. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015*.
33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
34. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A System for Large-Scale Machine Learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016*; USENIX Association: Berkeley, CA, USA, 2016; pp. 265–283.
35. Salvatier, J.; Wiecki, T.V.; Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2016**, *2*, e55. [[CrossRef](#)]
36. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.