remote sensing

MDPI

Article

PhySR-Bio: Multi-Stage Physics-Aware Symbolic Regression for Interpretable SAR-Based Forest Biomass Retrieval

Yaxuan Xing [1,2], Huiping Lin [3] and Yinghui Zhang [1,*]

[1] School of Electronic Information Engineering, Inner Mongolia University, Hohhot 010021, China; xingyaxuanimu@163.com
[2] Key Laboratory for Information Science of Electromagnetic Waves (Ministry of Education), Fudan University, Shanghai 200433, China
[3] School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China; linhuiping@cqu.edu.cn
* Correspondence: zhangyinghui@imu.edu.cn

Highlights

What are the main findings?

- PhySR-Bio discovers an explicit AGB retrieval formula using SAR scattering priors.
- The discovered formula outperforms representative semi-empirical models and achieves accuracy close to black box machine learning baselines.

What is the implication of the main finding?

- The study provides a case in which symbolic regression is successfully used to construct an analytical SAR-based biomass retrieval formula.
- The formula suggests that AGB retrieval benefits from combining volume scattering information, double-bounce-related structural terms, and incidence angle modulation.

Abstract

Forest above-ground biomass (AGB) is essential for understanding forest carbon stocks and ecosystem dynamics. Despite the strong potential of synthetic aperture radar (SAR) for forest monitoring, retrieving AGB from SAR observations remains difficult because the relationship between SAR scattering responses and forest biomass is highly nonlinear, spatially variable, and often difficult to represent with both accuracy and interpretability. To address this issue, this study proposes PhySR-Bio, a multi-stage physics-aware symbolic regression framework for interpretable SAR-based forest biomass retrieval. The physics-aware design is achieved by incorporating SAR scattering-related variables, incidence angle information, and physically meaningful feature combinations into symbolic regression. Unlike black box regression models, PhySR-Bio derives an analytical formula in which the selected variables and their mathematical relationships can be directly inspected. On the independent test set, PhySR-Bio achieved an RMSE of 45.083 Mg/ha, an MAE of 33.997 Mg/ha, and an R of 0.527. It reduced RMSE and MAE by up to 6.35% and 8.57%, respectively, compared with representative semi-empirical formulas, and achieved a test RMSE within 2.3% of the best-performing black box machine learning model while retaining an explicit and interpretable expression.

Keywords: above-ground biomass; symbolic regression; polarimetric SAR; P-band SAR

Check for updates

Academic Editor: Eric Casella

Received: 24 April 2026
Revised: 4 June 2026
Accepted: 8 June 2026
Published: 27 October 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.

1. Introduction

Forest above-ground biomass (AGB) is a core variable for characterizing forest carbon stocks, productivity, and ecological structural status [1]. Accurate AGB estimation is directly linked to carbon accounting, forest resource inventory, ecosystem monitoring, and climate change assessment [2]. Remote sensing has become the mainstream approach for AGB estimation from regional to global scales, owing to its clear advantages over conventional field surveys in spatial coverage, temporal continuity, and cross-regional scalability [3,4].

Among various remote sensing modalities, synthetic aperture radar (SAR) has long played an important role in forest AGB retrieval because of its all-weather, all-day imaging capability and its sensitivity to forest vertical structure and woody components [4]. In particular, long-wavelength SAR is generally more suitable than shorter wavelengths for characterizing woody biomass and deeper structural layers [5–7]. Previous studies have shown that L- and P-band SAR are key frequency bands for forest biomass estimation. Sandberg et al. [8] reported that both L- and P-band polarimetric backscatter can be used for stand-scale biomass estimation, with P-band retaining stronger sensitivity in higher-biomass ranges, while Schlund et al. [9] showed that joint L- and P-band modeling can outperform single-frequency retrieval. Overall, existing studies consistently indicate the importance of L- and P-band SAR for forest AGB retrieval [5,10,11].

From a physical perspective, forest SAR backscatter is not a direct monotonic response to biomass, but rather the combined outcome of multiple scattering mechanisms, attenuation effects, and imaging geometry. This nonlinearity becomes particularly evident in high-biomass forests, where signal saturation may appear not only as a weakened or plateaued response, but also as a non-monotonic or even reversed backscatter–biomass relationship [12]. Such behavior has been attributed to propagation attenuation and compensatory changes among structural attributes, including stem size, tree density, and tree height [13,14]. In addition, studies based on L-band polarimetric UAVSAR data show that relying only on backscatter can lead to structural mixing, suggesting that additional structural and mechanism-related information is needed for stable biomass retrieval [15]. Volume scattering, double-bounce scattering, and surface scattering each encode different aspects of forest structure, and their relative contributions vary with forest type, biomass range, and environmental conditions. Consequently, forest AGB is more likely to correspond to a coupled expression of multiple mechanism-related variables than to a simple function of any single backscatter quantity. Fully polarimetric SAR is particularly valuable in this context because it provides a mechanism-related variable space through polarimetric decomposition and scattering descriptors [16–18].

Under such conditions, traditional empirical and semi-empirical models are often too restrictive to capture the complex coupling among scattering mechanisms, attenuation, and geometry, whereas black box machine learning models, although capable of achieving high predictive accuracy, often provide limited physical transparency. For forest AGB retrieval, this creates a practical dilemma: a model should not only be accurate, but should also reveal which scattering mechanisms dominate and how they jointly affect biomass estimation [19,20]. Symbolic regression provides a promising alternative because it can directly search for explicit analytical relationships in a candidate-variable space while balancing predictive error and expression complexity [21,22]. In recent years, SR has begun to show potential in remote sensing inversion problems. Existing studies have used SR to combine morphological and spectral parameters for constructing explicit analytical models for leaf area index (LAI) retrieval [23,24]. In forest canopy height estimation, SR has also been used alongside random forest (RF) as a viable modeling approach in combination with optical remote sensing data [25]. More broadly, from the perspective of Earth science modeling, a study proposed knowledge-guided deep symbolic regression for parameterization discov-

ery [26], suggesting that SR is evolving from a generic fitting tool toward an interpretable framework for relationship discovery and physically meaningful parameterization. Nevertheless, its application to fully polarimetric SAR-based forest biomass retrieval remains very limited, particularly in terms of systematically integrating scattering mechanism-related variables, interpretable analytical expressions, and deployable modeling forms.

Furthermore, when applied directly to fully polarimetric SAR biomass retrieval, SR typically suffers from large search spaces, sensitivity to initialization and hyperparameters, and a tendency toward expression bloat, which can reduce interpretability and robustness [22,27]. To address these challenges, this study proposes PhySR-Bio, a multi-stage physics-aware symbolic regression framework for interpretable SAR-based forest biomass retrieval. The proposed framework aims to reduce search ambiguity, suppress redundant expressions, and discover compact analytical models that retain clearer links to underlying scattering mechanisms. The main contributions of this study are summarized as follows:

- A physics-aware symbolic regression framework, termed PhySR-Bio, is proposed by embedding SAR scattering mechanism variables, incidence angle information, and physically meaningful feature combinations into AGB formula discovery.
- A multi-stage symbolic discovery strategy is developed to search compact analytical formulas from physical prior-guided candidate variables, reducing search ambiguity and retaining clearer links to SAR scattering mechanisms.
- The discovered formula improves upon representative semi-empirical models and approaches black box machine learning accuracy while keeping the selected variables and interaction structures interpretable.

2. Study Area and Dataset

This study uses the TomoSense dataset. It was acquired over the Kermeter forest in Eifel National Park, Germany, as part of the ESA TomoSense campaign [28]. The study area represents a temperate forest environment with heterogeneous stand structure and pronounced topographic variation, making it suitable for investigating the joint effects of forest scattering behavior and terrain-related geometric modulation.

The Kermeter forest consists of mixed deciduous and coniferous stands distributed over varying slopes and elevation gradients. Beech-dominated stands are mainly found on moist northern slopes, whereas oak and spruce stands are more common on relatively drier southern slopes. Such variations in species composition, stand structure, and terrain conditions provide a representative testbed for SAR-based forest biomass analysis.

The TomoSense campaign provides airborne P-band SAR observations specifically designed for forest tomographic studies. The P-band data were acquired by MetaSensing in July 2020 in a multi-baseline monostatic configuration, with fully polarimetric observations collected under two opposite viewing headings, i.e., north-west and south-east. According to the dataset description, the P-band system has a bandwidth of 30 MHz, an along-track resolution of about 1 m, and a vertical tomographic resolution of approximately 5–10 m over the main study area. To construct the vertical baseline distribution required for tomography, the first ten passes were flown at constant altitude, followed by additional passes acquired with progressively reduced flight altitude. Subsequent studies based on the same dataset further reported a center frequency of 435 MHz, corresponding to a wavelength of about 0.69 m and a slant-range resolution of about 5 m. In addition, the released SAR products are already coregistered, phase-calibrated, and ground-steered, which facilitates direct analysis of forest scattering characteristics.

To obtain a method that is easier to deploy in practice, this study selects a single fully polarimetric P-band scene for formula discovery and analysis. Specifically, we use one north-west-looking single-look complex (SLC) acquisition. By restricting the input to one

fully polarimetric SAR scene, the proposed method avoids dependence on multi-baseline acquisitions and tomographic reconstruction, and is therefore more consistent with the goal of developing a simple and practically deployable biomass modeling approach.

In addition to the SAR data, the TomoSense dataset provides forest AGB reference data and a slope product derived from a Digital Elevation Model (DEM). These auxiliary data are used to support the analysis of scattering features and geometry-related effects in the discovered expressions.

An overview of the dataset is shown in Figure 1. The four subfigures present the P-band pseudo-color image, the Freeman–Durden decomposition pseudo-color image, the AGB reference map, and the slope map, respectively. Together, these data provide the basis for subsequent formula discovery and interpretation.

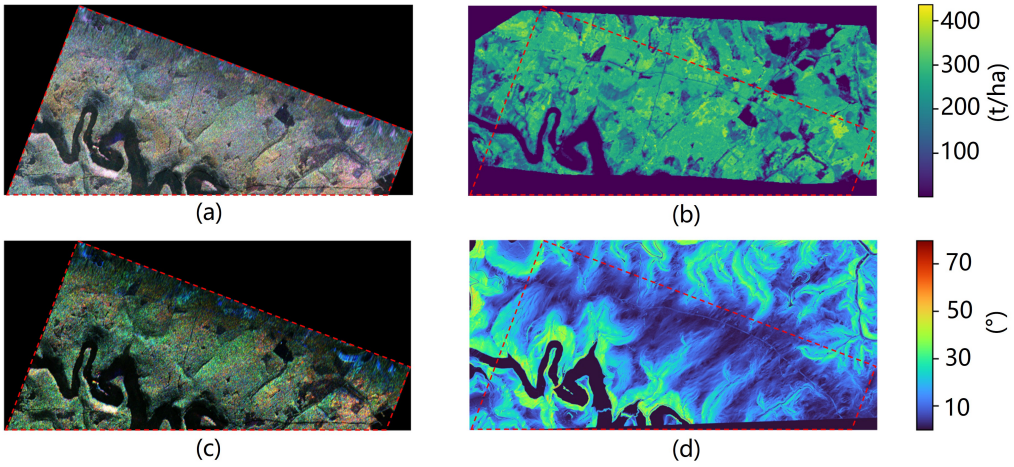


Figure 1. Multi-source data of the study area. (**a**) the P-band polarimetric SAR false-color composite, (**b**) the reference AGB map, (**c**) the Freeman decomposition false-color composite, and (**d**) the slope map. The red dashed line delineates the study area boundary.

3. Physical Priors

To construct a physics-guided symbolic modeling framework for forest AGB retrieval, prior SAR knowledge must be translated into constraints and candidate structures for symbolic search. In this study, these physical priors are derived from two complementary sources: mechanistic understanding of forest SAR scattering, which supports the construction of mechanism-related candidate variables, and existing analytical or semi-empirical biomass formulations, which provide structural references for symbolic formula design.

3.1. Scattering Mechanism Priors for Candidate Variable Construction

From the perspective of scattering mechanisms, forest SAR backscatter is essentially the result of multiple scattering mechanisms, propagation attenuation, and imaging geometry acting together. Arii et al. [29] summarized that the core scattering framework in vegetated scenes still consists of three main components: surface scattering, volume scattering, and double-bounce scattering. For forests, volume scattering mainly corresponds to the aggregated response of leaves, branches, and other randomly oriented scatterers within the canopy; double-bounce scattering is more often associated with trunk–ground interactions; and surface scattering mainly reflects the roughness and dielectric properties of the forest floor. These components correspond to different structural aspects of forests and therefore provide a direct basis for constructing mechanism-related candidate variables. The relationship between polarimetric SAR responses and biomass may also vary with environmental conditions and forest states, indicating that SAR signals should be interpreted together with scattering mechanisms and observational context [18].

The usefulness of different scattering mechanisms has also been supported by empirical observations. Chowdhury et al. [30] showed that volume and double-bounce scattering are positively correlated with growing stock volume, whereas surface scattering is negatively correlated. This pattern is physically meaningful: in sparse or low-biomass forests, canopy gaps and exposed ground can strengthen surface scattering; with forest development, canopy closure and woody structure increase volume scattering and trunk–ground double-bounce responses while reducing the relative contribution of surface scattering. Dickinson et al. [31] showed that polarimetric decomposition parameters can identify structurally similar forest composition classes across different landscapes, which also indicates that the structural information of a given forest type is often distributed across multiple polarimetric and scattering descriptors rather than being monopolized by a single component. Musthafa et al. [32] further demonstrated that inputs jointly constructed from multiple decomposed scattering powers are more favorable for building stable models than a single backscatter quantity. Scattering components derived from polarimetric decomposition have also been used directly in extended water cloud models for AGB retrieval, showing their modeling value [33]. More recent work further suggests that decomposed surface, volume, and double-bounce components carry different types of structural information, including stem density, mean stem biomass, and background or soil-related effects [34].

These studies indicate that AGB-related information is distributed across multiple scattering components rather than being represented by a single backscatter variable. Therefore, PhySR-Bio introduces volume-, double-bounce-, and surface-scattering variables, together with cross-polarized responses and incidence angle-related terms, as candidate variables for symbolic regression. In this way, the symbolic search is guided by known SAR scattering behavior while still allowing the final analytical expression to be discovered from data.

3.2. Analytical and Semi-Empirical Formula Priors for Symbolic Structure Design

Existing empirical and semi-empirical formulas provide structural priors for symbolic formula design. They express the SAR–AGB relationship through compact mathematical forms, such as logarithmic response terms, exponential saturation functions, vegetation–ground decomposition, attenuation factors, and multi-polarization combinations. These formulations do not fully represent the complexity of SAR–forest interactions, but they indicate which mathematical structures and physical components have been repeatedly used to describe biomass response. Table 1 summarizes representative analytical and semi-empirical formulas and compares whether volume scattering, double-bounce scattering, and incidence angle or geometric effects are explicitly included.

In the table, F1–F4 represent empirical regression families centered on the statistical relationship between backscatter and AGB. These formulas use logarithmic, exponential, or power–exponential functions to describe nonlinear backscatter–AGB relationships, especially saturation or asymptotic behavior in high-biomass conditions. They usually rely on cross-polarized channels or observables closely related to volume scattering.

F5–F6 correspond to the water cloud model and its variants. These models decompose the total backscatter into ground and vegetation terms, and use exponential attenuation factors to describe the suppression of ground returns by canopy extinction. In this way, they provide a mechanistic explanation for volume scattering enhancement and signal saturation. Although this class of models is more closely connected to physical processes in form, trunk–ground interaction terms are not explicitly introduced, and double-bounce scattering therefore remains implicit or absent.

Table 1. Summary of analytical and semi-empirical formulations for forest AGB retrieval.

ID	Formula	V	D	G	Ref.
F1	$\sigma^0_{\mathrm{HV}} = a_0 + a_1 \ln(\mathrm{AGB}) - a_2 \left(\ln(\mathrm{AGB})\right)^2$	T	F	F	[35]
F2	$\gamma^0 = a_0 \times e^{-a_1 \times \mathrm{AGB}} + a_2$	T	F	T	[36]
F3	$\gamma^0 = a_0\, \mathrm{AGB}^{\alpha}\left(1 - e^{-a_1 \mathrm{AGB}}\right) + a_2$	T	F	T	[37]
F4	$\gamma^0 = a_0\, \mathrm{AGB}^{\alpha}\left(1 - e^{-a_1 \mathrm{AGB}}\right) + \left(a_2\, \mathrm{AGB}^{\beta} + a_3\right)\gamma^0_s e^{-a_1 \mathrm{AGB}}$	T	F	T	[37]
F5	$\sigma^0_{\mathrm{for}} = \sigma^0_{gr} e^{-\beta \mathrm{AGB}} + \sigma^0_{\mathrm{veg}}\left(1 - e^{-\beta \mathrm{AGB}}\right)$	T	F	F	[33]
F6	$\sigma^0_{\mathrm{for}} = (1-\eta)\sigma^0_{gr} + \eta\, \sigma^0_{gr} e^{-ah} + \eta\, \sigma^0_{\mathrm{veg}}\left(1 - e^{-ah}\right)$	T	F	F	[38]
F7	$\mathrm{AGB}^{\lambda} = a_0 + a_1 \gamma^0_{\mathrm{HH}} + a_2 \gamma^0_{\mathrm{HV}} + a_3 \gamma^0_{\mathrm{VV}}$	T	T	T	[37]
F8	$\ln(\mathrm{AGB}) = a_0 + a_1 \ln(\sigma^0_{\mathrm{HH}}) + a_2 \ln(\sigma^0_{\mathrm{HV}}) + a_3 \ln(\sigma^0_{\mathrm{VV}})$	T	T	F	[39]
F9	$\sigma^0_{\mathrm{for}} = \sigma^0_{gr} e^{-\beta \mathrm{AGB}} + \sigma^0_{\mathrm{veg}}\left(1 - e^{-\beta \mathrm{AGB}}\right) + \sigma^0_{gs} e^{-\beta \mathrm{AGB}}$	T	T	F	[33]
F10	$\mathrm{AGB} = a_0 + a_1 \sigma^0_{\mathrm{HV}} - a_2 \dfrac{\sigma^0_{\mathrm{HH}}}{\sigma^0_{\mathrm{HV}}}$	T	T	F	[40]

Note: V = volume scattering; D = double-bounce scattering; G = geometric effect; T = True; F = False.

In contrast, F7–F10 are more closely related to a modeling strategy that combines mechanism-related features with regression. By introducing multi-polarization features, such as combinations or ratios of HH, HV, and VV, as well as terms associated with double-bounce scattering (e.g., HH/HV ratios, three-channel combinations, or explicit double-bounce contribution terms), these models enhance the representation of woody structural information. In addition, the "G" column in Table 1 indicates whether geometric effects are incorporated into the model. For formulations using γ^0 as input (e.g., F2–F4 and F7), the influence of local incidence angle is implicitly included through the normalized observable, and these cases are marked as "T". By contrast, models that do not use γ^0 or do not explicitly account for geometric correction are marked as "F".

Overall, existing formulas show that SAR-based biomass retrieval has commonly been represented through saturation functions, attenuation terms, vegetation–ground decomposition, and polarization-based variable combinations. These structures provide useful physical and mathematical references for symbolic search. This motivates the use of symbolic regression to retain these physical priors while allowing a more flexible analytical expression to be discovered from data.

4. Methods

To achieve interpretable SAR-based forest biomass retrieval with compact analytical formulas, this study proposes PhySR-Bio, a multi-stage physics-aware symbolic regression framework. The framework is designed to control expression bloat and reduce overfitting in unconstrained symbolic regression while preserving mechanism-related interpretability and practical deployability. As shown in Figure 2, the whole procedure consists of three stages: optimal feature selection, symbolic structure optimization, and model parameter estimation.

The first stage identifies the most effective feature triplet from the candidate feature pool. The second stage builds the analytical formula progressively from this optimal triplet. It starts from a baseline additive form, then introduces interaction-enhanced and log-stabilized forms, and finally adds a residual correction term through symbolic regression. The third stage estimates the coefficients of the selected formula. In this way, the frame-

work determines which features to use, how they should be combined, how a physically meaningful and data-informed formula can be constructed, and how its coefficients can be stably estimated.

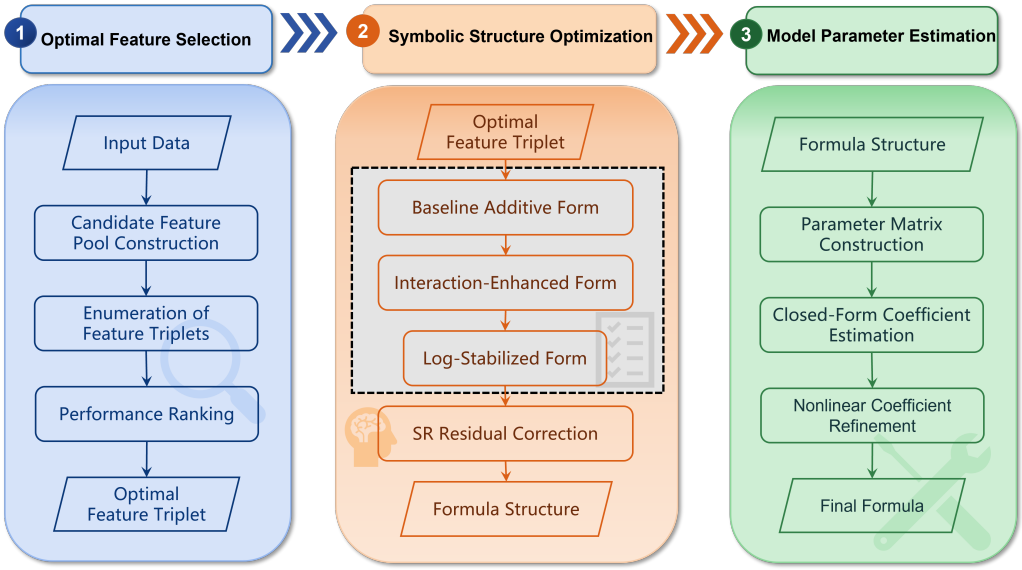


Figure 2. Overview of the proposed PhySR-Bio framework. The method follows a multi-stage modeling pipeline including optimal feature selection, symbolic structure optimization, and model parameter estimation. Within the symbolic structure optimization stage, the formula is refined in a step-wise manner through baseline additive, interaction-enhanced, and log-stabilized forms, followed by residual symbolic correction when needed, to obtain the final analytical formula.

For mathematical formulation, the problem is defined as searching for an optimal symbolic formula that maps SAR-derived variables to forest AGB. Let the training set be denoted by

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}, \tag{1}$$

where y_i is the forest AGB of the i-th sample, and $\mathbf{x}_i$ is the corresponding feature vector derived from fully polarimetric SAR observations.

According to the physical analysis in Section 2, the input variables are grouped into two categories:

1. Scattering mechanism-related variables, including polarization channel intensities, decomposition-derived volume, double-bounce, and surface scattering powers, as well as their ratio features;
2. A geometry-related variable, represented by the local incidence angle term $\cos\theta$.

The objective is to learn an explicit analytical mapping, shown as

$$\hat{y} = F(\mathbf{x}; \Theta), \tag{2}$$

where Θ denotes the coefficient set of the symbolic formula. The learning process is divided into three stages. Stage I selects the most informative feature subset. Stage II determines the formula structure under explicit complexity control. Stage III estimates the model parameters after the symbolic structure is fixed.

All candidate formulas are ranked according to both predictive accuracy and structural simplicity. The ranking score is defined as

$$\text{Score}(F) = \text{RMSE}(F) + \lambda C(F), \tag{3}$$

where RMSE is root mean square error, $C(F)$ denotes the formula complexity, and λ controls the trade-off between fitting accuracy and structural compactness.

4.1. Stage I: Optimal Feature Selection

The purpose of Stage I is to identify a compact and effective feature combination for subsequent symbolic modeling. This stage reduces the search space of symbolic regression and provides a physically meaningful initialization for formula construction.

The candidate feature pool includes the original polarization intensities $(\mathrm{HH}, \mathrm{HV}, \mathrm{VV})$, the volume, double-bounce, and surface scattering powers (P_V, P_D, P_S) derived from Freeman–Durden polarimetric decomposition, and their ratio features. The ratio features are defined as

$$r_{P_S} = \frac{P_S}{P_S + P_D + P_V + \varepsilon}, \tag{4}$$

$$r_{P_D} = \frac{P_D}{P_S + P_D + P_V + \varepsilon}, \tag{5}$$

$$r_{P_V} = \frac{P_V}{P_S + P_D + P_V + \varepsilon}, \tag{6}$$

where ε is a small positive constant to avoid division by zero. The ratio features describe the relative contribution of different scattering mechanisms and reduce the influence of overall intensity scale. In addition, the local incidence angle term $\cos\theta$ is introduced to represent geometry-related modulation. To further account for geometry-modulated scattering responses, product terms between $\cos\theta$ and scattering-related variables are also included in the candidate pool, such as $\mathrm{HV}\cos\theta$, $\mathrm{VV}\cos\theta$, $P_D\cos\theta$, $P_S\cos\theta$, and their ratio-modulated counterparts. Therefore, Stage I searches for the optimal three-feature combination from an expanded candidate pool consisting of the original variables, ratio features, the geometric term $\cos\theta$, and geometry-modulated product terms. Thus, the number of possible three-feature combinations in the expanded candidate pool is $C_{16}^{3} = 560$. For each candidate combination, a baseline additive model is first constructed, denoted as

$$\hat{y} = w_0 + \sum_{j=1}^{3} w_j x_j. \tag{7}$$

All three-feature combinations are enumerated. Their coefficients are estimated by least squares on the training set. The top-ranked combination is passed to Stage II.

The detailed ranking results of the candidate feature combinations are provided in Appendix A, Table A1. Overall, the high-ranked combinations consistently involve HV, $\cos\theta$, and one additional double-bounce-related variable, such as r_{P_D}, P_D, or $P_D\cos\theta$.

This pattern is physically consistent with the scattering analysis in Section 3. The HV channel mainly reflects canopy volume scattering responses and provides the dominant biomass-sensitive signal. The double-bounce-related variables supply complementary information associated with woody structure and trunk–ground interaction. The local incidence angle term $\cos\theta$ captures geometry-related modulation under varying observation conditions. Therefore, the ranking results indicate that effective biomass modeling should be built on the joint expression of volume scattering-dominant information, double-bounce-related structural information, and geometry-related modulation.

Based on this result, Stage II takes the top-ranked feature combination from Stage I as the input to symbolic structure optimization and progressively refines the corresponding analytical forms.

4.2. Stage II: Symbolic Structure Optimization

Stage II determines the symbolic structure of the final analytical model. This stage follows a step-wise refinement strategy, in which the formula is progressively refined from simple to more expressive forms. At each step, only a restricted family of symbolic structures is considered, so as to control formula complexity while preserving physical interpretability.

4.2.1. Step-Wise Structure Refinement

Let X_V denote the selected volume scattering-related variable, X_D denote the selected double-bounce-related variable, and $\cos\theta$ denote the local incidence angle-related geometric term. Based on the selected feature combination from Stage I, three candidate formula families are constructed.

Step 1: Baseline Additive Form

The first step considers a simple additive structure, described as

$$F_{\mathrm{A}}(\mathbf{x}) = a_0 + a_1 X_V + a_2 X_D + a_3 \cos\theta, \tag{8}$$

where X_V represents a volume scattering-related variable, and X_D represents a double-bounce scattering-related variable. This form captures the first-order contributions of the dominant volume scattering term, the complementary double-bounce-related term, and the observation-geometry-related term. It serves as the basic physical skeleton of the model.

Step 2: Interaction-Enhanced and Geometry-Modulated Form

The second step introduces geometry-dependent modulation of scattering contributions. One candidate family augments the additive model with a single modulation term, shown as

$$F_{\mathrm{B}}(\mathbf{x}) = a_0 + a_1 X_V + a_2 X_D + a_3 \cos\theta + a_4 (X_V \cdot \cos\theta), \tag{9}$$

or

$$F_{\mathrm{B}}(\mathbf{x}) = a_0 + a_1 X_V + a_2 X_D + a_3 \cos\theta + a_4 (X_D \cdot \cos\theta). \tag{10}$$

In addition, geometry-modulated variables are also allowed to directly replace their original counterparts in the candidate structure, described as

$$F_{\mathrm{B}}(\mathbf{x}) = a_0 + a_1 (X_V \cdot \cos\theta) + a_2 \cos\theta + a_3 X_D, \tag{11}$$

or

$$F_{\mathrm{B}}(\mathbf{x}) = a_0 + a_1 X_V + a_2 \cos\theta + a_3 (X_D \cdot \cos\theta). \tag{12}$$

These forms are introduced to account for cases in which geometry-modulated scattering variables are more informative than their original unmixed forms.

Step 3: Log-Stabilized Form

The third step introduces logarithmic transformation for selected variables to improve numerical stability and describe mild nonlinear relationships. The resulting log-stabilized family can be described as

$$\ln y = a_0 + \sum_{j=1}^{M} a_j \phi_j(\mathbf{x}), \tag{13}$$

where each $\phi_j(\mathbf{x})$ denotes either an original term or its logarithmic counterpart, depending on the candidate configuration. In this way, both additive and interaction-enhanced structures can be extended into a family of log-stabilized variants, including cases in which

selected scattering variables or geometry-modulated terms are represented in logarithmic form.

All candidate structures generated from Equations (8)–(13) are fitted and ranked using Equation (3). The best-performing structure is retained as the baseline symbolic formula for subsequent residual correction.

The detailed candidate structures generated in Stage II and their validation performance are listed in Appendix A Table A2. The ranking results show that the best-performing structures are still centered on HV, $\cos\theta$, and double-bounce-related variables, while the inclusion of interaction terms with $\cos\theta$ further improves the baseline expression. In particular, the best-performing structure is

$$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 r_{P_D}\cos\theta, \tag{14}$$

which is ranked most favorably under the score criterion and, at the same time, achieves the lowest validation RMSE among all candidate formulas while maintaining relatively low structural complexity.

Compared with the baseline additive forms in Step 1, the interaction-enhanced and geometry-modulated forms in Step 2 generally lead to lower test errors. This suggests that double-bounce-related information is more effective when expressed together with the local incidence angle. By contrast, the log-stabilized variants in Step 3 rank consistently lower. In many existing biomass formulas, logarithmic terms are introduced to linearize the strongly nonlinear relationship between raw backscatter observables and biomass, especially when single-channel responses exhibit saturation-like behavior. In this framework, the candidate variables have been physically reorganized through mechanism-related feature construction and explicit geometry modulation. Applying logarithmic transformation further compresses high-value variation and weakens the effective response contrast, leading to no additional predictive benefit for the current dataset.

4.2.2. Residual Symbolic Correction

After the baseline symbolic formula is selected, the remaining systematic error is modeled by a short residual correction term. Let the selected baseline formula be denoted as

$$F_{\mathrm{base}}(\mathbf{x}) = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 r_{P_D}\cos\theta. \tag{15}$$

The residual is defined as

$$r = y - F_{\mathrm{base}}(\mathbf{x}). \tag{16}$$

A constrained symbolic regression is then applied to model the residual. In this study, PySR is adopted as the symbolic regression tool. PySR is designed for scientific modeling and aims to automatically discover analytical expressions from given input variables and operator sets [41,42]. PySR performs heuristic search directly in the expression space by combining evolutionary search with local constant optimization, while jointly considering fitting error and expression complexity. As a result, its output is usually not a single formula, but a set of candidate expressions with different levels of complexity, which can then be selected according to the desired balance between interpretability and predictive accuracy.

In this study, the residual search is guided by both error reduction and complexity control. The operator set is restricted to basic arithmetic operations, exponentials, and logarithms, and additional complexity penalties are introduced to keep the correction term as compact as possible. As illustrated in Figure 3, PySR performs symbolic search on the residual through a multi-island evolutionary process involving parallel expression populations, repeated selection, crossover, mutation, and migration under explicit complexity control. To evaluate the stability and reproducibility of the residual search, PySR is independently

run 10 times under the same settings, and the resulting residual forms and validation performance are recorded.

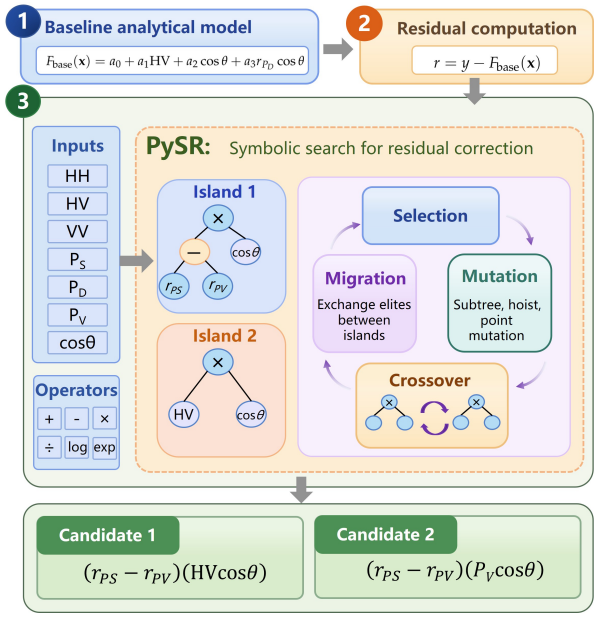


Figure 3. Workflow of the PySR-based residual symbolic search. PySR searches compact residual correction expressions using selected variables, restricted operators, and a multi-island evolutionary strategy under complexity control. The gray arrows indicate the workflow of the processing steps, whereas the purple arrows denote the iterative process within the symbolic search. Different colors are used to distinguish the main functional modules of the framework.

Although the exact symbolic expressions differ across runs, the 10 repeated searches reveal a highly consistent structural pattern. In most cases, the residual term can be summarized into two components:

1. A mechanism contrast term, usually expressed as the difference between a surface-related proportion and a canopy-volume-related proportion, such as $(r_{P_S} - r_{P_V})$;
2. An energy scale term, which controls the magnitude of the correction and is typically represented by a cosine-modulated volume scattering-related intensity, such as $HV\cos\theta$ or $P_V\cos\theta$.

Accordingly, the residual formula is written in the compact multiplicative form, shown as

$$F_{res}(\mathbf{x}) = k\,(r_{P_S} - r_{P_V})(X_V\cos\theta), \tag{17}$$

where $X_V\cos\theta$ denotes a cosine-modulated volume scattering-related term, such as $HV\cos\theta$ or $P_V\cos\theta$. This residual term has a clear physical interpretation. The factor $(r_{P_S} - r_{P_V})$ describes the relative dominance between surface-related and volume-related scattering contributions. The factor $(X_V\cos\theta)$ controls the magnitude of the correction by combining volume scattering-related information with geometry-related modulation.

The final analytical formula is written as

$$F_{final}(\mathbf{x}) = F_{base}(\mathbf{x}) + F_{res}(\mathbf{x}). \tag{18}$$

4.3. Stage III: Model Parameter Estimation

Once the symbolic structure is fixed, the final analytical formula is written as

$$F_{final}(\mathbf{x}) = a_0 + a_1 HV + a_2\cos\theta + a_3 r_{P_D}\cos\theta + k\,(r_{P_S} - r_{P_V})(X_V\cos\theta). \tag{19}$$

Stage III estimates the coefficients of this formula. This stage separates symbolic structure determination from numerical fitting, so that the final formula remains fixed while its coefficients are optimized in a stable and reproducible way. The objective function is defined as

$$\min_{\Theta} \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i(\Theta))^2, \tag{20}$$

where Θ denotes the coefficient vector of the final formula. Since the symbolic terms are fixed, the final model is linear with respect to its coefficients and can be rewritten in matrix form as

$$\hat{\mathbf{y}} = \boldsymbol{\Phi}\boldsymbol{\alpha}, \tag{21}$$

where $\boldsymbol{\alpha}$ is the coefficient vector and $\boldsymbol{\Phi}$ is the design matrix determined by the selected symbolic terms. Accordingly, the coefficient vector is written as

$$\boldsymbol{\alpha} = [a_0, a_1, a_2, a_3, k]^{\mathrm{T}}, \tag{22}$$

and the corresponding design matrix is

$$\boldsymbol{\Phi} = \left[1,\ \mathrm{HV},\ \cos\theta,\ r_{P_D}\cos\theta,\ (r_{P_S} - r_{P_V})(X_V \cos\theta)\right]. \tag{23}$$

The coefficient solution is obtained by least squares, described as

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\alpha}\|_2^2. \tag{24}$$

The detailed parameter estimates and validation performance of the candidate residual terms are provided in Appendix A, Table A3. The comparison indicates that the residual term using $P_V \cos\theta$ provides slightly better validation performance than the alternative using $\mathrm{HV}\cos\theta$. Therefore, $P_V \cos\theta$ is selected as the volume-related scaling term for residual correction.

Accordingly, the residual term in the final model is determined using $P_V \cos\theta$, and the final analytical formula is expressed as

$$F_{\mathrm{final}}(\mathbf{x}) = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 r_{P_D}\cos\theta + k(r_{P_S} - r_{P_V})(P_V \cos\theta). \tag{25}$$

This stage-wise design separates feature selection, symbolic structure determination, and coefficient estimation. It keeps the formula compact, preserves mechanism-related interpretability, and improves numerical stability for forest biomass retrieval.

5. Experimental Results and Analysis

This section presents the experimental evaluation of the proposed PhySR-Bio framework. The final output of the method is an explicit analytical formula, and its construction includes optimal feature triplet selection, symbolic structure optimization, residual term search, and parameter estimation. Accordingly, the experiments are designed to validate this data-driven formula learning process and to further assess the generalization ability, stability, and effectiveness of the resulting model under a spatially independent test setting.

5.1. Experimental Setup

The dataset used in this study consists of SAR-derived variables and their corresponding reference AGB samples. Since remote sensing data usually exhibit strong spatial continuity, neighboring samples often share similar scattering characteristics and biomass distributions. Under such conditions, random pixel-level splitting may lead to over-optimistic

performance estimates. To balance spatial independence and spatial representativeness, this study adopts a spatially blocked train–validation–test partition.

Specifically, the study area is then divided into five spatial groups along the main horizontal direction. Within each group, the rightmost strip is reserved as the test set, while the remaining valid blocks are evenly assigned to the training and validation sets. This design ensures that the test samples are spatially separated from the training and validation samples, while the test set is distributed across the whole study area rather than concentrated in only one side of the scene. All compared methods use exactly the same training, validation, and test partitions.

The training set is used for model fitting, symbolic expression search, and coefficient estimation. The validation set is used for feature selection, candidate symbolic structure ranking, residual term selection, and hyperparameter optimization. The test set is kept unseen during model development and is used only once after the final model is determined.

During preprocessing, the P-band SAR variables, local incidence angle, and biomass reference data are first read. The SAR-derived variables and incidence angle data are then spatially matched to the AGB map. The P-band SAR data have an along-track resolution of approximately 1 m, while the biomass map has a spatial resolution of approximately 10 m. To reduce local noise and potential residual misregistration effects, the matched variables are further aggregated using a 3×3 spatial window before sample extraction. Invalid values and samples outside the AGB range of 70–420 Mg/ha are removed. After these preprocessing steps, the final effective sample size is 16,202. The final train:validation:test ratio is approximately 4.5:4.5:1. The statistical characteristics of the resulting train–validation–test split are summarized in Table 2.

Table 2. Summary of AGB statistics for the train–validation–test split.

Subset	Sample Size	AGB Min	AGB Max	AGB Mean	AGB Std
Training	7491	70.099	419.825	258.244	58.980
Validation	7295	70.071	419.246	258.732	58.363
Test	1416	75.924	414.161	254.534	61.103

The input feature pool contains 16 variables, including the original polarimetric channels $(\mathrm{HH}, \mathrm{HV}, \mathrm{VV})$, decomposition-derived scattering powers (P_S, P_D, P_V), cosine-modulated intensity terms (e.g., $\mathrm{HV}\cos\theta$, $P_V\cos\theta$, and $P_D\cos\theta$, together with the corresponding related terms), scattering mechanism proportions $(r_{P_S}, r_{P_D}, r_{P_V})$, and the geometric term $\cos\theta$. Both the input variables and the output variable, AGB, are standardized using z-score normalization. The normalization parameters are calculated only from the training set and are then applied to the validation and test sets in the same experiment. Before reporting the final accuracy metrics, the predicted AGB values are transformed back to the original unit of Mg/ha.

To quantify model performance, three evaluation metrics are adopted, including RMSE, MAE, and the Pearson correlation coefficient (R). RMSE and MAE are reported in Mg/ha, where 1 Mg/ha is equivalent to 1 t/ha. These metrics are defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{26}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{27}$$

and

$$R = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}. \quad (28)$$

The experiments consist of two parts. The first part is the ablation study, which is used to analyze the contribution of the major components in PhySR-Bio, including symbolic residual correction and different residual term constructions. The second part is the comparative evaluation, which assesses the overall predictive performance of the proposed method against representative baselines.

5.2. Ablation Study

To quantify the contribution of the residual correction term, the selected baseline formula is compared with four extended models. These include three additive variants, obtained by directly adding r_{P_S}, $P_V \cos\theta$, or their combination to the baseline formula, and the final PySR-based residual correction model. This comparison is designed to distinguish the effect of simply adding physically meaningful variables from that of introducing a structurally informed symbolic correction. All models are trained on the same training set and evaluated on the same test set.

The ablation results are summarized in Table 3. The baseline model already captures the dominant variation trend of AGB, but its performance remains limited by residual systematic bias. Directly adding r_{P_S} or $P_V \cos\theta$ does not lead to consistent improvement, and the gains remain marginal even when the two terms are introduced simultaneously in a linear additive form.

Table 3. Ablation results of different residual correction designs.

Variant	Formulation	RMSE	MAE	*R*
base (Equation (15))	F_{base}	45.392	34.402	0.5244
base + r_{P_S}	$F_{\text{base}} + b_1 r_{P_S}$	45.232	34.464	0.5247
base + $P_V \cos\theta$	$F_{\text{base}} + b_2 P_V \cos\theta$	45.361	34.080	0.5256
base + $r_{P_S} + P_V \cos\theta$	$F_{\text{base}} + b_1 r_{P_S} + b_2 P_V \cos\theta$	45.190	34.182	0.5268
base + PySR	$F_{\text{base}} + k(r_{P_S} - r_{P_V})(P_V \cos\theta)$	45.083	33.997	0.5270

These results indicate that the performance gain does not mainly come from increasing the number of predictors. Instead, the way in which the additional variables are structurally coupled plays a more important role. Although r_{P_S} and $P_V \cos\theta$ are physically meaningful, their isolated or purely additive use is insufficient to characterize the residual relationship left by the baseline model.

The PySR-based residual correction achieves the best performance among all variants. Its advantage lies in introducing a compact multiplicative structure, in which $(r_{P_S} - r_{P_V})$ describes the relative balance between surface- and volume-related scattering contributions, while $P_V \cos\theta$ acts as a geometry-modulated energy scale term. This structurally informed correction is more effective than direct linear addition, showing that the residual improvement is associated with symbolic structure discovery rather than variable accumulation alone. Overall, the ablation result provides quantitative support for the physical interpretation of the residual correction term. It shows that the selected variables are not effective merely because they are added to the model, but because their PySR-derived multiplicative coupling captures a useful interaction between surface–volume scattering contrast and geometry-modulated volume scattering.

5.3. Comparative Evaluation

To further evaluate the effectiveness of the proposed formula search modeling strategy, we selected five representative baseline methods for comparison, including two semi-empirical formulas, F7 and F10, from Table 1, and three machine learning models: Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Support Vector Regression (SVR). F7 and F10 explicitly include multi-polarization or double-bounce-related information and therefore provide relevant analytical baselines for comparison. RF and GBDT are adopted as representative nonlinear ensemble methods, whereas SVR is used as a kernel-based nonlinear regression model. These methods are compared with the proposed PhySR-Bio model under the same feature setting. The hyperparameters of the three machine learning models are listed in Table 4.

Table 4. Hyperparameter settings of the machine learning baseline models.

Model	Hyperparameter Settings
RF	n_estimators = 400, max_depth = 10, min_samples_split = 10, min_samples_leaf = 4, random_state = 0, n_jobs = -1.
GBDT	HistGradientBoostingRegressor with max_depth = 3, learning_rate = 0.03, max_iter = 400, min_samples_leaf = 20, l2_regularization = 1.0, and random_state = 0.
SVR	Radial basis function kernel with $C = 8.0$, $\epsilon = 0.10$, and `gamma` = "scale".

The results of RF, GBDT, and SVR on the training and test sets are shown in Figure 4, whereas the results of F7, F10, and PhySR-Bio are shown in Figure 5. RF shows the strongest fitting ability among the compared machine learning models. On the training set, it achieves an RMSE of 38.041 Mg/ha, an MAE of 27.940 Mg/ha, and an R of 0.7897. On the test set, its RMSE, MAE, and R are 44.050 Mg/ha, 31.851 Mg/ha, and 0.5530, respectively. GBDT presents slightly weaker training performance, with a training RMSE of 43.542 Mg/ha, an MAE of 32.420 Mg/ha, and an R of 0.6742. On the test set, it achieves an RMSE of 44.116 Mg/ha, an MAE of 32.200 Mg/ha, and an R of 0.5507. These results indicate that the two tree-based ensemble models retain strong nonlinear fitting ability.

SVR shows a smoother but less accurate regression behavior. On the training set, its RMSE, MAE, and R are 46.717 Mg/ha, 31.240 Mg/ha, and 0.6136, respectively. On the test set, the corresponding values are 46.019 Mg/ha, 31.862 Mg/ha, and 0.5164. Compared with RF and GBDT, SVR shows weaker correlation and higher RMSE on the test set, although its MAE remains comparable. This suggests that the kernel-based global regression structure provides a relatively stable mapping, but its ability to capture the nonlinear SAR–AGB relationship remains limited under the present feature setting.

The two semi-empirical formulas, F7 and F10, provide useful analytical baselines. F7 achieves an RMSE of 55.722 Mg/ha, an MAE of 42.985 Mg/ha, and an R of 0.3099 on the training set. On the test set, its RMSE, MAE, and R are 48.060 Mg/ha, 36.968 Mg/ha, and 0.4215, respectively. F10 shows similar behavior, with a training RMSE of 55.818 Mg/ha, an MAE of 43.190 Mg/ha, and an R of 0.3049. On the test set, it achieves an RMSE of 48.140 Mg/ha, an MAE of 37.184 Mg/ha, and an R of 0.4177. Although these formulas include physically meaningful polarization or double-bounce-related information, their predefined structures are still too restrictive to fully capture the coupled effects of scattering mechanisms, incidence angle modulation, and nonlinear biomass response.

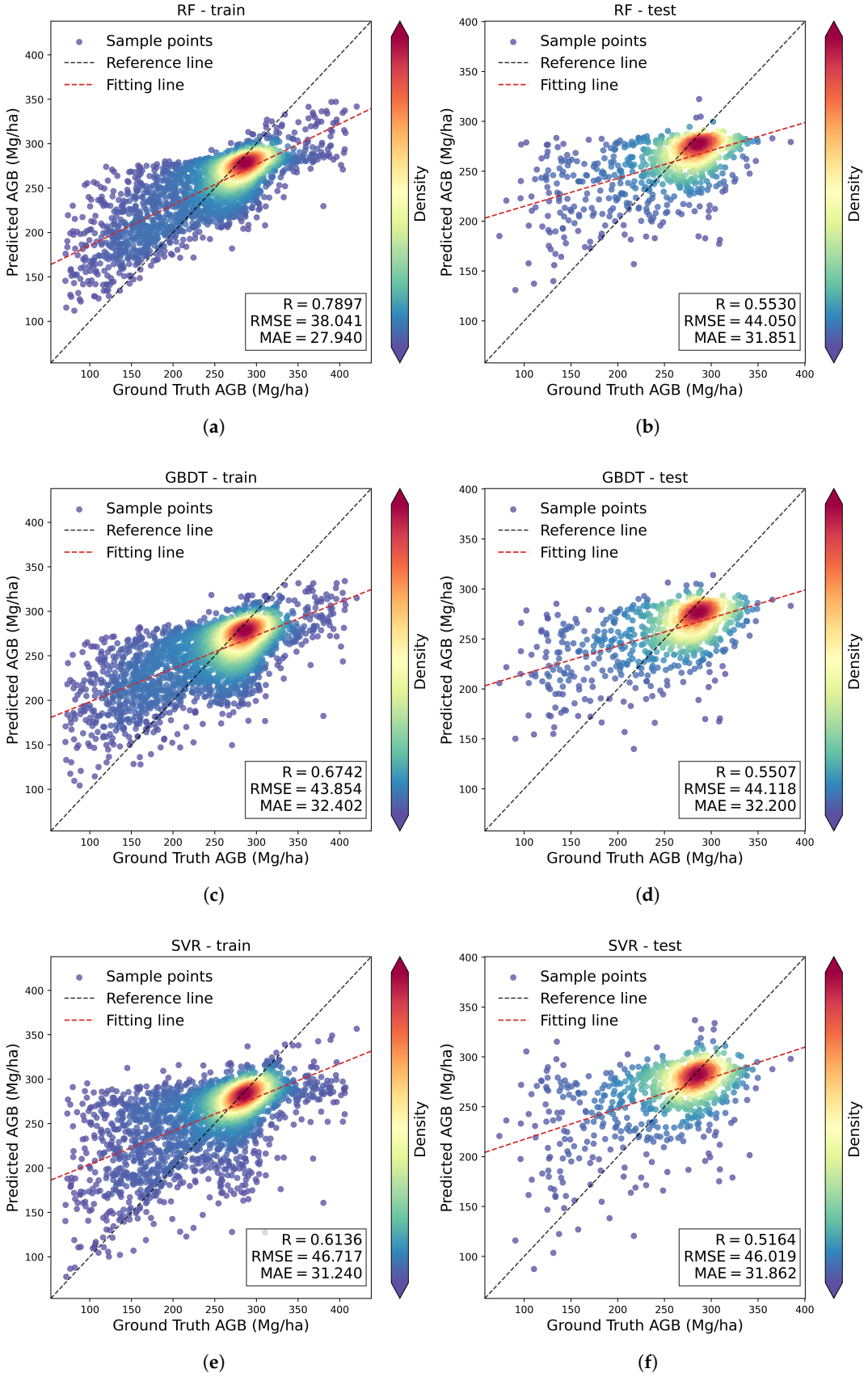


Figure 4. Scatter plots of AGB retrieval results for RF, GBDT, and SVR. (**a**) RF on the training set; (**b**) RF on the test set; (**c**) GBDT on the training set; and (**d**) GBDT on the test set; (**e**) SVR on the training set; (**f**) SVR on the test set.

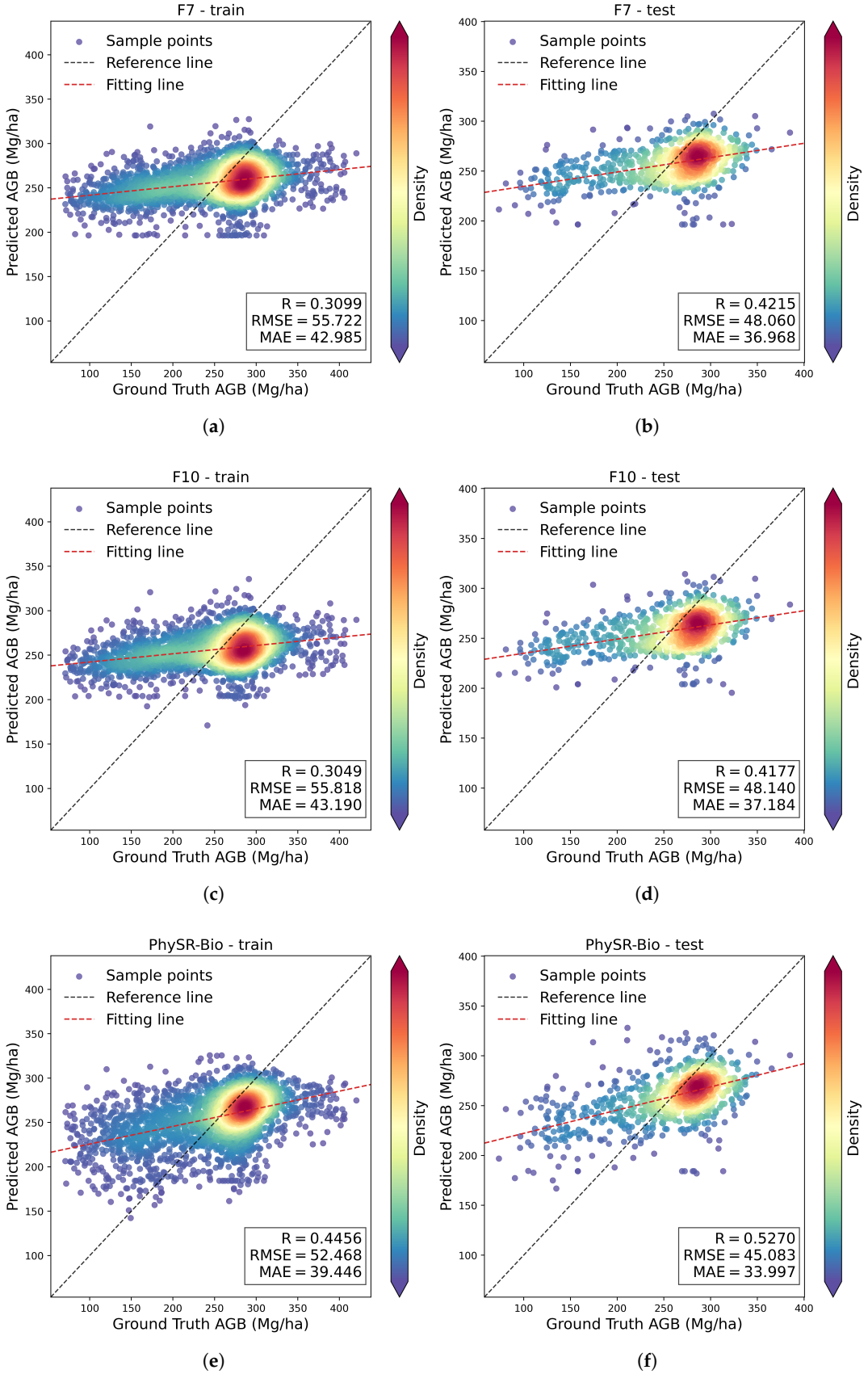


Figure 5. Scatter plots of AGB retrieval results for the semi-empirical formulas and PhySR-Bio. (**a**) F7 on the training set; (**b**) F7 on the test set; (**c**) F10 on the training set; (**d**) F10 on the test set; (**e**) PhySR-Bio on the training set; and (**f**) PhySR-Bio on the test set.

PhySR-Bio improves upon these semi-empirical analytical baselines. On the training set, it achieves an RMSE of 52.468 Mg/ha, an MAE of 39.446 Mg/ha, and an R of 0.4456. On the test set, the corresponding values are 45.083 Mg/ha, 33.997 Mg/ha, and 0.5270. Compared with F7, PhySR-Bio reduces the test RMSE from 48.060 to 45.083 Mg/ha and the MAE from 36.968 to 33.997 Mg/ha while increasing R from 0.4215 to 0.5270. Compared with F10, it reduces the test RMSE from 48.140 to 45.083 Mg/ha and the MAE from 37.184 to 33.997 Mg/ha while increasing R from 0.4177 to 0.5270. These improvements indicate that the proposed symbolic formula search can provide a more flexible and effective analytical expression than predefined semi-empirical formulas.

Compared with the black box machine learning models, PhySR-Bio does not achieve the lowest test RMSE or MAE. RF and GBDT still provide slightly better global accuracy metrics on the test set. However, the performance of PhySR-Bio remains comparable to these machine learning baselines, while offering a compact explicit formula. In contrast to RF, GBDT, and SVR, the proposed method directly reveals the selected variables, mathematical operators, and interaction structures, which improves the transparency of the retrieval model and facilitates practical deployment.

Overall, the comparative results show that predefined semi-empirical formulas may be insufficient for representing complex SAR–AGB relationships, whereas highly flexible machine learning models can achieve strong predictive accuracy but provide limited structural transparency. PhySR-Bio provides an intermediate modeling strategy: it substantially improves upon fixed analytical formulas and approaches the accuracy of machine learning baselines, while retaining an explicit and interpretable analytical form. This supports the use of physics-aware symbolic regression as a transparent formula search strategy for SAR-based AGB retrieval.

To further examine whether the global accuracy metrics mask range-dependent errors, the test samples are divided into several bins according to the reference AGB values. For each bin, the mean bias and RMSE are calculated separately for PhySR-Bio and RF, and the residual distributions are further compared using boxplots, as shown in Figure 6.

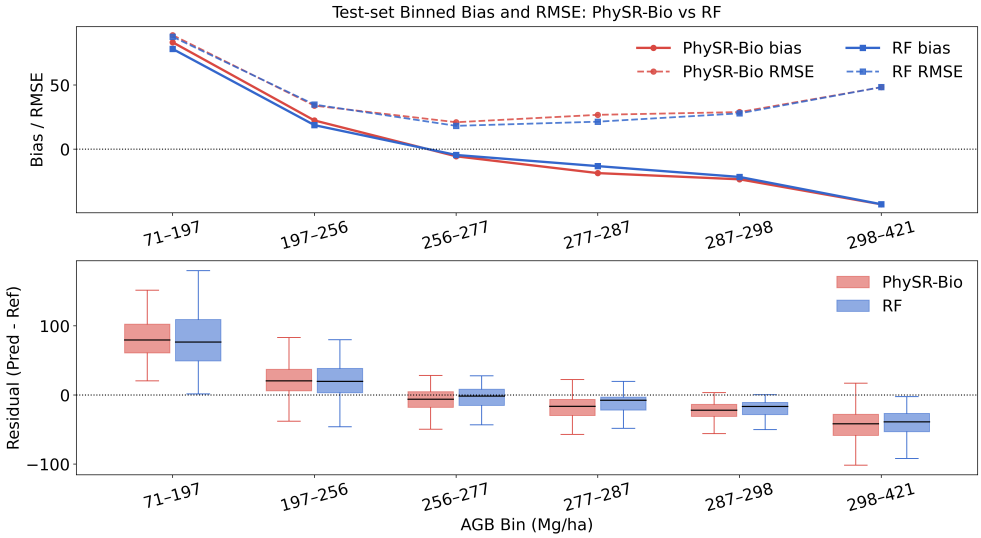


Figure 6. Binned test set error analysis for PhySR-Bio and RF. The upper panel shows the mean bias and RMSE calculated within each reference AGB bin. The lower panel shows the residual distributions in each AGB bin, where the residual is defined as the difference between the prediction and the reference AGB.

The binned results show a clear range-dependent error pattern. In the low-AGB bin, both PhySR-Bio and RF tend to overestimate AGB, as indicated by the positive mean bias

and positive residual distribution. As AGB increases, the mean bias gradually decreases and becomes negative in the medium- to high-AGB bins, indicating an underestimation tendency for higher biomass samples. This pattern suggests a compression of the predicted biomass range, where low biomass values are overestimated and high biomass values are underestimated.

This systematic bias is consistent with the known difficulty of SAR-based biomass retrieval under nonlinear response, saturation, and structural-mixing conditions. Although PhySR-Bio provides an explicit analytical formula and competitive global test accuracy, the binned analysis indicates that range-dependent bias remains, especially at the two ends of the biomass distribution.

To further examine the spatial manifestation of the range-dependent bias, several representative test regions are selected and visualized in Figure 7. The selected regions include high-AGB, low-AGB, complex-texture, and representative cases. Overall, PhySR-Bio captures the broad spatial gradient of AGB, but the estimated maps are smoother than the reference maps, indicating that fine-scale spatial variability is partly suppressed by the compact analytical formula.

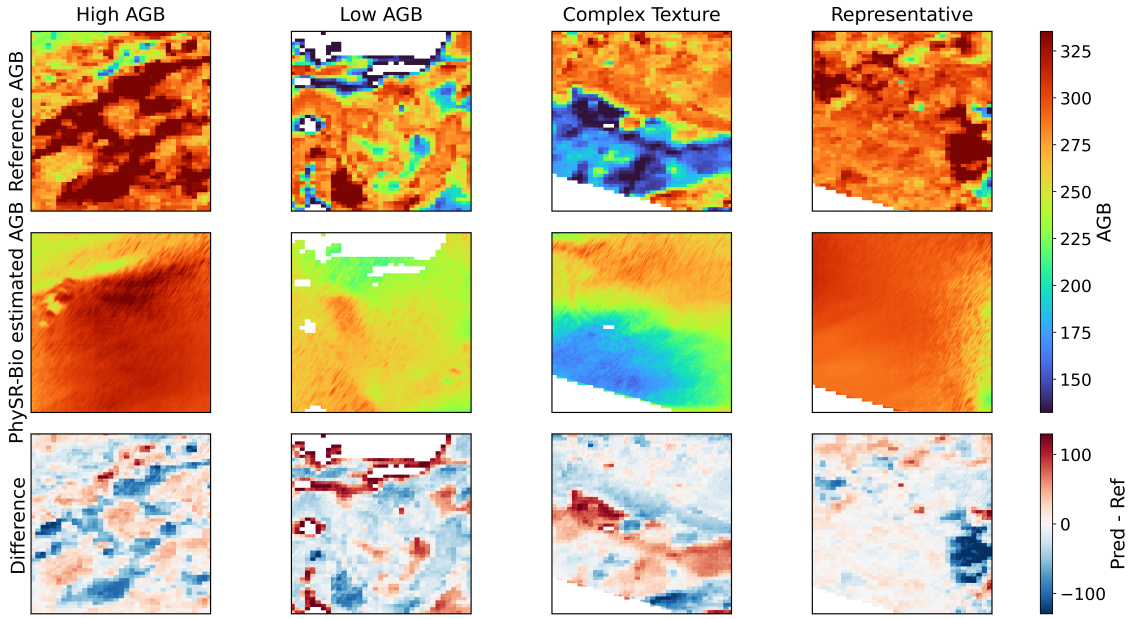


Figure 7. Examples of spatial AGB retrieval results for PhySR-Bio in representative test regions. The four columns correspond to high-AGB, low-AGB, complex-texture, and representative regions. The first row shows the reference AGB maps, the second row shows the PhySR-Bio estimated AGB maps, and the third row shows the difference between the prediction and the reference.

In the high-AGB region, the model tends to underestimate the highest biomass values, as reflected by negative residuals in several high-biomass patches. In the low-AGB region, positive residuals are more apparent, indicating that low biomass values are partly overestimated. This behavior is consistent with the binned bias analysis, where low-AGB samples show positive bias and high-AGB samples show negative bias. In the complex-texture region, the residuals are spatially heterogeneous, suggesting that abrupt local changes in biomass and scattering characteristics remain difficult to represent using a single compact formula. In the representative region, the residuals are relatively weaker and more spatially balanced, indicating that the model performs more stably in areas with moderate biomass variation. Although the proposed model provides an explicit and interpretable analytical expression, further improvement is still needed to better preserve fine-scale

spatial heterogeneity and reduce overestimation in low-AGB areas and underestimation in high-AGB areas.

6. Discussion

The experimental results show that PhySR-Bio improves upon representative semi-empirical formulas and achieves test accuracy close to black box machine learning baselines while retaining an explicit analytical expression. Therefore, the main value of the proposed method is providing a transparent formula search strategy for SAR-based AGB retrieval. In this section, the findings are discussed.

6.1. Physics-Prior-Guided Symbolic Discovery and Structural Interpretability

The physics-aware design of PhySR-Bio should be understood as physics-prior-guided symbolic discovery. Prior SAR scattering knowledge is translated into candidate variables, interaction terms, and structural constraints, so that the data-driven formula search is conducted in a physically interpretable hypothesis space rather than in an arbitrary mathematical expression space. This is the main difference between the proposed method and unconstrained symbolic regression.

The final formula shows that the physical priors introduced before symbolic search are reflected in the discovered expression. Forest SAR responses are jointly affected by volume scattering, double-bounce scattering, surface scattering, attenuation, and observation geometry [18,29]. The selected terms include HV, $\cos\theta$, $r_{P_D}\cos\theta$, and $(r_{P_S} - r_{P_V})(P_V \cos\theta)$. Among them, HV and $P_V \cos\theta$ are related to cross-polarized and volume scattering information, which has been widely reported to be important for forest biomass and structural retrieval [8,9,15]. The term $r_{P_D}\cos\theta$ reflects the coupling between double-bounce-related structure and incidence angle modulation, which is consistent with previous findings that double-bounce scattering is associated with trunk–ground interaction and woody structure [30,33,34]. The residual term $(r_{P_S} - r_{P_V})(P_V \cos\theta)$ further indicates that the contrast between surface- and volume-related scattering contributes to correcting the baseline formula. This is consistent with the interpretation that stronger surface scattering is more likely to appear in sparse or low-biomass conditions, whereas volume scattering becomes more dominant as canopy structure develops [30,34].

This structure also helps explain the map bias behavior observed in the binned error analysis. Map bias in SAR-based AGB retrieval is closely related to the non-separability between SAR inputs and AGB. Different biomass states may produce similar SAR responses because low-biomass and high-biomass conditions, surface background effects, volume scattering attenuation, double-bounce contribution, and terrain-related observation geometry are coupled in the observed signal. By including volume scattering, double-bounce scattering, surface–volume contrast, and incidence angle modulation in the symbolic search space, PhySR-Bio introduces additional physical cues for distinguishing different scattering and biomass conditions. This is one reason why it improves upon fixed semi-empirical formulas, which usually rely on more restricted predefined structures.

The interpretability of PhySR-Bio should be understood as structural interpretability rather than as strict physical parameterization. Symbolic regression remains a data-driven formula discovery method. Therefore, the physical interpretation of PhySR-Bio is based on the selected variables, operators, and interaction structures rather than on the absolute values of the fitted coefficients. This view is consistent with recent symbolic regression studies, which emphasize its role in discovering compact and inspectable mathematical relationships from data [21,22]. Similar applications in remote sensing and Earth science have also shown that symbolic regression can support explicit model discovery for biophysical parameter retrieval and process representation [23–26].

6.2. Limitations

Several limitations remain. First, the binned error analysis and spatial examples show that range-dependent bias remains. PhySR-Bio tends to overestimate low-AGB samples and underestimate high-AGB samples, indicating a compression of the predicted biomass range. This limitation does not contradict the proposed method, but is consistent with the known difficulty of SAR-based biomass retrieval. Previous studies have reported saturation, non-monotonic response, and structural compensation in high-biomass forests [12–14].

Second, the current experiment is based on a single airborne P-band SAR scene over a temperate forest site. Therefore, the discovered formula should be interpreted as a site-specific analytical approximation rather than a universally applicable biomass retrieval model. Further validation across different forest types, biomass ranges, topographic conditions, and acquisition geometries is required.

7. Conclusions

This study proposes PhySR-Bio, a physics-aware symbolic regression framework for AGB retrieval from polarimetric SAR data. By incorporating SAR scattering mechanism knowledge, incidence angle information, and semi-empirical formulation priors into a multi-stage symbolic search process, the method derives a compact analytical formula rather than relying on a fixed empirical structure or a fully black box regression model.

The experimental results show that PhySR-Bio improves upon representative semi-empirical formulas, reducing RMSE by up to 6.35%. Its test RMSE remains within 2.3% of the best-performing RF model, indicating that the explicit formula achieves accuracy close to the strongest machine learning baseline in this experiment. At the same time, the formula directly presents the selected variables, mathematical operators, and interaction terms, which makes the retrieval structure more interpretable than black box models.

The discovered expression indicates that SAR-based AGB retrieval benefits from the joint use of volume scattering information, double-bounce-related structural terms, surface–volume scattering contrast, and incidence angle modulation. These results suggest that physics-aware symbolic regression can serve as a transparent analytical modeling strategy for SAR-based biomass retrieval, especially when both prediction accuracy and model interpretability are required.

In future work, the proposed framework should be further evaluated using multi-site, multi-temporal, and multi-sensor P-band SAR datasets covering different forest types, biomass ranges, topographic conditions, and acquisition geometries. With such extended validation, the framework may provide an interpretable and computationally efficient candidate approach for large-area biomass estimation.

Author Contributions: Conceptualization, Y.X.; Methodology, Y.X.; Validation, Y.X. and H.L.; Formal analysis, Y.Z.; Resources, Y.X.; Writing—original draft, Y.X.; Writing—review and editing, H.L.; Supervision, Y.X. and Y.Z.; Project administration, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62461047; in part by the Natural Science Foundation of Inner Mongolia Autonomous Region of China under Grant 2021JQ07; in part by the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region under Grant NJYT22109; and in part by the Training Plan for Young Innovative of Grassland Talents Project in Inner Mongolia Autonomous Region under Grant Q2022003.

Data Availability Statement: The code used in this study is available at https://github.com/xingyaxuan/PhySR-Bio (accessed on 7 June 2026). The repository includes the implementation of the proposed PhySR-Bio framework and the comparison experiments with representative semi-empirical formulas and machine learning baseline models. The experimental data are from the open TomoSense dataset.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Detailed Formula Search Results

Table A1. Detailed Stage I ranking results of candidate feature combinations and baseline formula forms.

Rank	Features	Formula
1	$\mathrm{HV}, \cos\theta, r_{P_D}$	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 r_{P_D}$
2	$\mathrm{HV}, \cos\theta, P_D\cos\theta$	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 P_D\cos\theta$
3	$\mathrm{HV}, \cos\theta, P_D$	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 P_D$
4	$\mathrm{HV}, \cos\theta, r_{P_S}$	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 r_{P_S}$
5	$\mathrm{HV}, \mathrm{VV}, \cos\theta$	$y = a_0 + a_1\mathrm{HV} + a_2\mathrm{VV} + a_3\cos\theta$

Table A2. Detailed Stage II ranking results of candidate symbolic structures and their validation performance.

Rank	Step	Val-RMSE	Formula
1	2	52.855	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 r_{P_D}\cos\theta$
2	1	52.857	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 r_{P_D}$
3	2	52.932	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 r_{P_D} + a_4(r_{P_D}\cos\theta)$
4	2	52.934	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 P_D + a_4(P_D\cos\theta)$
5	2	52.935	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 P_D\cos\theta$
6	2	52.941	$y = a_0 + a_1(\mathrm{HV}\cos\theta) + a_2\cos\theta + a_3 P_D$
7	2	52.948	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 r_{P_D} + a_4(\mathrm{HV}\cos\theta)$
8	1	53.187	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 P_D$
9	2	53.266	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 P_D + a_4(\mathrm{HV}\cos\theta)$
10	2	53.326	$y = a_0 + a_1\mathrm{HV} + a_2\cos\theta + a_3 P_D + a_4(P_D\cos\theta)$
11	3	53.520	$\ln y = a_0 + a_1\ln(\mathrm{HV}) + a_2\cos\theta + a_3 r_{P_D} + a_4\ln(\mathrm{HV}\cos\theta)$
12	3	53.547	$\ln y = a_0 + a_1\ln(\mathrm{HV}) + a_2\cos\theta + a_3 r_{P_D}$
13	3	56.037	$\ln y = a_0 + a_1\ln(\mathrm{HV}) + a_2\cos\theta + a_3\ln(r_{P_D}\cos\theta)$
14	3	56.038	$\ln y = a_0 + a_1\ln(\mathrm{HV}) + a_2\cos\theta + a_3\ln(P_D) + a_4\ln(\mathrm{HV}\cos\theta)$
15	3	56.038	$\ln y = a_0 + a_1\ln(\mathrm{HV}) + a_2\cos\theta + a_3\ln(P_D) + a_4\ln(P_D\cos\theta)$
16	3	56.052	$\ln y = a_0 + a_1\ln(\mathrm{HV}\cos\theta) + a_2\cos\theta + a_3\ln(P_D)$
17	3	56.144	$\ln y = a_0 + a_1\ln(\mathrm{HV}) + a_2\cos\theta + a_3\ln(P_D)$
18	3	57.840	$\ln y = a_0 + a_1\ln(\mathrm{HV}) + a_2\cos\theta + a_3\ln(P_D\cos\theta)$
19	3	58.005	$\ln y = a_0 + a_1\ln(\mathrm{HV}) + a_2\cos\theta + a_3\ln(P_D) + a_4\ln(r_{P_D}\cos\theta)$
20	3	58.107	$\ln y = a_0 + a_1\ln(\mathrm{HV}) + a_2\cos\theta + a_3\ln(r_{P_D}) + a_4\ln(P_D\cos\theta)$

Table A3. Parameter estimates and validation performance of candidate PySR residual terms.

Residual Term	a_0	a_1	a_2	a_3	k	RMSE	MAE
$\mathrm{HV}\cos\theta$	256.847	47.217	26.754	7.586	−23.478	52.925	40.215
$P_V\cos\theta$	256.847	38.010	21.297	6.596	−15.383	52.793	40.056

References

1. Ma, T.; Zhang, C.; Ji, L.; Zuo, Z.; Beckline, M.; Hu, Y.; Li, X.; Xiao, X. Development of forest aboveground biomass estimation, its problems and future solutions: A review. *Ecol. Indic.* **2024**, *159*, 111653. [CrossRef]
2. Martes, L.; Köhl, M. Improving the contribution of forests to carbon neutrality under different policies—A case study from the Hamburg Metropolitan area. *Sustainability* **2022**, *14*, 2088. [CrossRef]

3. Hu, Y.; Sun, Z. Assessing the capacities of different remote sensors in estimating forest stock volume based on high precision sample plot positioning and random forest method. *Nat. Environ. Pollut. Technol.* **2022**, *21*, 1113–1123. [CrossRef]
4. Ye, Z.; Long, J.; Zhang, T.; Lin, B.; Lin, H. L-band synthetic aperture radar and its application for forest parameter estimation, 1972 to 2024: A review. *Plants* **2024**, *13*, 2511. [CrossRef] [PubMed]
5. Zeng, P.; Zhang, W.; Li, Y.; Shi, J.; Wang, Z. Forest total and component above-ground biomass (AGB) estimation through C- and L-band polarimetric SAR data. *Forests* **2022**, *13*, 442. [CrossRef]
6. Hamdan, O.; Aziz, H.K.; Rahman, K.A. Remotely sensed L-Band SAR data for tropical forest biomass estimation. *J. Trop. For. Sci.* **2011**, *23*, 318–327.
7. Hu, Y.; Nie, Y.; Liu, Z.; Wu, G.; Fan, W. Improving the potential of coniferous forest aboveground biomass estimation by integrating C- and L-Band SAR data with feature selection and non-parametric model. *Remote Sens.* **2023**, *15*, 4194. [CrossRef]
8. Sandberg, G.; Ulander, L.M.H.; Fransson, J.E.S.; Holmgren, J.; Le Toan, T. L- and P-band backscatter intensity for biomass retrieval in hemiboreal forest. *Remote Sens. Environ.* **2011**, *115*, 2874–2886. [CrossRef]
9. Schlund, M.; Davidson, M.W.J. Aboveground forest biomass estimation combining L- and P-band SAR acquisitions. *Remote Sens.* **2018**, *10*, 1151. [CrossRef]
10. Bouvet, A.; Mermoz, S.; Le Toan, T.; Villard, L.; Mathieu, R.; Naidoo, L.; Asner, G.P. An above-ground biomass map of African savannahs and woodlands at 25 m resolution derived from ALOS PALSAR. *Remote Sens. Environ.* **2018**, *206*, 156–173. [CrossRef]
11. Das, S.; Chatterjee, B.; Choudhury, M.R.; Dutta, S.; Mondal, B.P.; Awasthi, A. Synthetic aperture radar for a changing planet: A 25-year global synthesis in hazard assessment, urban development, and ecological applications. *Ecol. Inform.* **2025**, *92*, 103477. [CrossRef]
12. Mermoz, S.; Réjou-Méchain, M.; Villard, L.; Le Toan, T.; Rossi, V.; Gourlet-Fleury, S. Decrease of L-band SAR backscatter with biomass of dense forests. *Remote Sens. Environ.* **2015**, *159*, 307–317. [CrossRef]
13. Joshi, N.; Mitchard, E.T.A.; Brolly, M.; Schumacher, J.; Fernández-Landa, A.; Johannsen, V.K.; Marchamalo, M.; Fensholt, R. Understanding 'saturation' of radar signals over forests. *Sci. Rep.* **2017**, *7*, 3505. [CrossRef] [PubMed]
14. Persson, H.J.; Huuva, I. Polarimetric measures in biomass change prediction using ALOS-2 PALSAR-2 data. *Remote Sens.* **2024**, *16*, 953. [CrossRef]
15. Zhang, Z.; Ni, W.; Sun, G.; Huang, W.; Ranson, K.J.; Cook, B.D.; Guo, Z. Biomass retrieval from L-band polarimetric UAVSAR backscatter and PRISM stereo imagery. *Remote Sens. Environ.* **2017**, *194*, 331–346. [CrossRef]
16. Ferrazzoli, P.; Paloscia, S.; Pampaloni, P.; Schiavon, G.; Sigismondi, S.; Solimini, D. The potential of multifrequency polarimetric SAR in assessing agricultural and arboreous biomass. *IEEE Trans. Geosci. Remote Sens.* **2002**, *35*, 5–17.
17. Wiseman, G.; McNairn, H.; Homayouni, S.; Shang, J. RADARSAT-2 polarimetric SAR response to crop biomass for agricultural production monitoring. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4461–4471. [CrossRef]
18. Tanase, M.A.; Panciera, R.; Lowell, K.; Tian, S.; Hacker, J.M.; Walker, J.P. Airborne multi-temporal L-band polarimetric SAR data for biomass estimation in semi-arid forests. *Remote Sens. Environ.* **2014**, *145*, 93–104. [CrossRef]
19. Tian, L.; Wu, X.; Tao, Y.; Li, M.; Qian, C.; Liao, L.; Fu, W. Review of remote sensing-based methods for forest aboveground biomass estimation: Progress, challenges, and prospects. *Forests* **2023**, *14*, 1086. [CrossRef]
20. Zhang, Y.; Zou, Y.; Wang, Y. Remote sensing of forest above-ground biomass dynamics: A review. *Forests* **2025**, *16*, 821. [CrossRef]
21. Makke, N.; Chawla, S. Interpretable scientific discovery with symbolic regression: A review. *Artif. Intell. Rev.* **2024**, *57*, 2. [CrossRef]
22. Dong, J.; Zhong, J. Recent advances in symbolic regression. *ACM Comput. Surv.* **2025**, *57*, 1–37. [CrossRef]
23. Wu, J.; Chen, P.; Fu, S.; Chen, Q.; Pan, X. Co-inversion of island leaf area index combination morphological and spectral parameters based on UAV multi-source remote sensing data. *Ecol. Inform.* **2023**, *77*, 102190. [CrossRef]
24. Paramanik, S.; Behera, M.D.; Dash, J. Symbolic regression-based allometric model development of a mangrove forest LAI using structural variables and digital hemispherical photography. *Appl. Geogr.* **2022**, *139*, 102649. [CrossRef]
25. Ghosh, S.M.; Behera, M.D.; Paramanik, S. Canopy height estimation using sentinel series images through machine learning models in a mangrove forest. *Remote Sens.* **2020**, *12*, 1519. [CrossRef]
26. Li, Q.; Zhang, C.; Wei, Z.; Jin, X.; Shangguan, W.; Yuan, H.; Zhu, J.; Li, L.; Liu, P.; Chen, X.; et al. Advancing symbolic regression for earth science with a focus on evapotranspiration modeling. *npj Clim. Atmos. Sci.* **2024**, *7*, 321. [CrossRef]
27. Virgolin, M.; Pissis, S.P. Symbolic regression is NP-hard. *arXiv* **2022**, arXiv:2207.01018.
28. Tebaldini, S.; d'Alessandro, M.M.; Ulander, L.M.H.; Bennet, P.; Gustavsson, A.; Coccia, A.; Macedo, K.; Disney, M.; Wilkes, P.; Spors, H.-J.; et al. TomoSense: A unique 3D dataset over temperate forest combining multi-frequency mono- and bi-static tomographic SAR with terrestrial, UAV and airborne lidar, and in-situ forest census. *Remote Sens. Environ.* **2023**, *290*, 113532. [CrossRef]
29. Arii, M.; Yamada, H.; Kojima, S.; Ohki, M. Review of the comprehensive SAR approach to identify scattering mechanisms of radar backscatter from vegetated terrain. *Electronics* **2019**, *8*, 1098. [CrossRef]

30. Chowdhury, T.A.; Thiel, C.; Schmullius, C.; Stelmaszczuk-Górska, M. Polarimetric parameters for growing stock volume estimation using ALOS PALSAR L-band data over Siberian forests. *Remote Sens.* **2013**, *5*, 5725–5756. [CrossRef]
31. Dickinson, C.; Siqueira, P.; Clewley, D.; Lucas, R. Classification of forest composition using polarimetric decomposition in multiple landscapes. *Remote Sens. Environ.* **2013**, *131*, 206–214. [CrossRef]
32. Musthafa, M.; Kumar, P.; Alsulamy, S.; Khedher, K.; Avtar, R. Development of L-band fully polarimetric SAR algorithm for forest biomass retrieval using 7SD and random forest regression. *Phys. Chem. Earth* **2024**, *136*, 103688. [CrossRef]
33. Kumar, S.; Garg, R.D.; Govil, H.; Kushwaha, S.P.S. PolSAR-decomposition-based extended water cloud modeling for forest aboveground biomass estimation. *Remote Sens.* **2019**, *11*, 2287. [CrossRef]
34. Carreiras, J.M.B.; Higginbottom, T.; Godlee, J.L.; Harrison, S.; Benitez, L.; Mograbi, P.J.; Levesley, A.; Melgaço, K.; Milodowski, D.; Pickavance, G.; et al. Determinants of L-band backscatter in dry tropical ecosystems: Implications for biomass mapping. *Remote Sens. Environ.* **2026**, *334*, 115213. [CrossRef]
35. Mitchard, E.T.A.; Saatchi, S.S.; Woodhouse, I.H.; Nangendo, G.; Ribeiro, N.S.; Williams, M.; Ryan, C.M.; Lewis, S.L.; Feldpausch, T.R.; Meir, P. Using satellite radar backscatter to predict above-ground woody biomass: A consistent relationship across four different African landscapes. *Geophys. Res. Lett.* **2009**, *36*, L23401. [CrossRef]
36. Liao, Z.; He, B.; Shi, Y. Improved forest biomass estimation based on P-band repeat-pass PolInSAR data across different forest sites. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103088. [CrossRef]
37. SERVIR Global. *The SAR Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation* , 1st ed.; SERVIR Global Science Coordination Office: Huntsville, AL, USA, 2019. [CrossRef]
38. Santoro, M.; Cartus, O.; Quegan, S.; Kay, H.; Lucas, R.M.; Araza, A.; Herold, M.; Labrière, N.; Chave, J.; Rosenqvist, Å.; et al. Design and performance of the Climate Change Initiative Biomass global retrieval algorithm. *Sci. Remote Sens.* **2024**, *10*, 100169. [CrossRef]
39. Nie, Y.; Hu, Y.; Sa, R.; Fan, W. Inversion of forest above ground biomass in mountainous region based on PolSAR data after terrain correction: A case study from Saihanba, China. *Remote Sens.* **2024**, *16*, 846. [CrossRef]
40. Avtar, R.; Suzuki, R.; Takeuchi, W.; Sawada, H. PALSAR 50 m mosaic data based national level biomass estimation in Cambodia for implementation of REDD+ mechanism. *PLoS ONE* **2013**, *8*, e74807. [CrossRef] [PubMed]
41. Cranmer, M. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv* **2023**, arXiv:2305.01582.
42. Cranmer, M. *PySR: High-Performance Symbolic Regression in Python and Julia*; Astrophysics Source Code Library, Record ascl:2409.018; NASA Astrophysics Data System: Cambridge, MA, USA, 2024.

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.