Exolife: Predicting Exoplanet Habitability to Support Astrobiological Discovery (v1.0)

Carlos Hernán Guirao

August 12, 2025

Abstract

The search for life beyond Earth requires tools capable of ranking thousands of known exoplanets by their potential to sustain habitable conditions. This paper presents *Exolife*, a machine learning pipeline designed to estimate the habitability of exoplanets using heterogeneous astrophysical data. The approach integrates self–supervised representation learning, physics–based simulations, weak supervision with positive unlabeled learning, and explicit bias and uncertainty modeling. A selection–function layer accounts for survey detection biases, ensemble climate and atmospheric–escape emulators provide physically grounded surrogate predictions with abstention capability, and monotonic constraints link the habitability likelihood to interpretable subscores. Conformal calibration ensures well–quantified predictive probabilities at decision time. The model outputs a probabilistic Habitability Likelihood (HL) and an human-interpretable Exolife Habitability Score (EHS*) to support observational prioritization. Limitations, risks, and potential enhancements are discussed to maintain transparency, robustness, and scientific value in the context of astrobiological exploration.

1 Introduction & Scope

The discovery of nearly six thousand exoplanets by 2025 has shifted the field from detection to characterization and prioritization. While Earth remains our only confirmed life–bearing planet, the increasing volume of data demands scalable methods to identify worlds that might support surface liquid water or other life–friendly conditions. *Exolife* aims to build a *robust, interpretable and generalizable* machine learning model that can estimate the probability of habitability for a given exoplanet, despite small labeled datasets, biased detection methods, and heterogeneous uncertainties. The scope encompasses the ingestion of multimission astrophysical catalogs, incorporation of physics–based simulations, representation learning on unlabeled time series and spectra, construction of a probabilistic target via weak supervision, and delivery of both a continuous habitability likelihood (HL) and a composite presentation score (EHS*). The methodology is designed to be reproducible and open, allowing integration into observational planning workflows while adhering to the principles of FAIR data.

2 Scientific Motivation

2.1 What is Exoplanet Habitability?

The concept of *habitability* refers to the ability of a planetary environment to sustain liquid water at or near the surface for geologically significant periods. Historically, assessments relied on single metrics such as membership in the stellar *Habitable Zone* (HZ) or proximity of the planet's equilibrium temperature T_{eq} to Earth's. These proxies are easy to compute but embed strong assumptions—Earth–like atmospheric composition, constant albedo and redistribution efficiency—and ignore complex interactions among climate, stellar activity and planetary geophysics. Modern approaches treat habitability as a *multidimensional, probabilistic property* that evolves as new observations and improved models become available. Hence, Exolife predicts a probability $\hat{p} = P(y = 1 \mid x)$ that a planet could sustain surface liquid water under plausible conditions, where y denotes a latent "habitable" state and x represents the measured features.

Key astrophysical factors include:

- **Orbital distance a and stellar luminosity $L_\star$:** These determine the incident stellar flux S and thereby control surface energy balance.
- **Atmospheric properties:** Composition, pressure and greenhouse effect regulate how incoming energy is absorbed and reradiated.
- **Planetary mass M_p and radius R_p:** Together they set surface gravity g and escape velocity v_{esc}, influencing atmospheric retention.
- **Stellar activity:** Ultraviolet (UV/XUV) radiation and flare rates can erode atmospheres or sterilize surfaces.

Because many of these variables are unknown or poorly constrained for exoplanets, habitability estimation must rely on a combination of direct measurements, derived indicators and simulated surrogates.

2.2 Why We Need Better Estimators

With thousands of confirmed planets and limited telescope time, prioritizing promising targets is crucial. Existing methods suffer from: (i) **static thresholds**, such as fixed HZ boundaries, which ignore feedbacks between orbital, climatic and geophysical processes; (ii) **Earth–centric bias**, assuming only Earth–like conditions are viable; (iii) **sparse parameter coverage**, since atmospheric and magnetic measurements are rare; and (iv) **poor generalization**, particularly for extreme systems or new detection methods. A data–driven but physics–informed estimator can reduce these limitations by learning from integrated datasets, incorporating physical priors through simulations, and quantifying uncertainty.

2.3 The Habitable Zone

The HZ is defined as the range of orbital distances where a rocky planet with an Earth–like atmosphere could sustain liquid water on its surface. For a star of luminosity L and effective temperature T_{eff}, the equilibrium distance d at which the planet receives a flux S_{eff} can be approximated by

$$d = \sqrt{\frac{L}{S_{\text{eff}}}}, \tag{1}$$

where S_{eff} depends on T_{eff} through empirically calibrated coefficients [1]. Conservative HZ limits use inner "recent Venus" and outer "early Mars" boundaries, while optimistic limits extend to allow less certain climate feedbacks. Despite its utility, this classification assumes Earth–like atmospheric chemistry, neglects other heat sources (tidal or geothermal), and treats habitability as binary; hence, it serves as only one component of Exolife's label model.

2.4 Metrics for Habitability Estimation

Several indices have been proposed to quantify habitability from limited data: (i) the *Earth Similarity Index* (ESI) compares radius, density, surface temperature and escape velocity to Earth's; (ii) the *Planetary Habitability Index* (PHI) incorporates potential chemical energy and substrate availability; (iii) equilibrium temperature T_{eq} estimates the effective surface temperature assuming a grey, uniform atmosphere; and (iv) machine–learned scores trained on known systems. While informative, each metric embeds assumptions and may fail to generalize. Exolife fuses multiple indicators via a probabilistic label model to produce a continuous habitability likelihood accompanied by interpretable sub–scores.

3 Project Objectives

Exolife seeks to deliver a machine–learning framework that can be integrated into astrobiological workflows. The specific objectives are:

(1) **Integrated Data Construction:** Compile and harmonize data from multiple astronomical catalogues—*NASA Exoplanet Archive, PHL Exoplanet Catalog, Gaia, SWEET–Cat*—alongside simulated planetary populations and derived features. Cross–identification of stars and planets uses primary keys (e.g., `gaia_source_id` and host/planet identifiers) with proper motion corrections, ensuring consistent joining across missions.

(2) **Comprehensive Feature Space:** Include a broad set of stellar, orbital and planetary parameters (Section 6) as well as derived indicators such as stellar flux, equilibrium temperature, and tidal heating. This allows models to capture non–trivial interactions that affect habitability.

(3) **Probabilistic Habitability Estimation:** Move beyond binary classification by producing a probability in the unit interval, with well–calibrated uncertainty bounds,

reflecting both measurement noise and model uncertainty. This supports risk–aware decision making.

(4) **Robustness to Incomplete and Biased Data:** Develop methods to handle missing values, measurement errors, and selection biases inherent in exoplanet detection (e.g., transit surveys favor short–period planets). The pipeline must remain functional when key parameters are absent and should report increased uncertainty or abstain when predictions are unsupported.

(5) **Scientific Interpretability:** Provide interpretable outputs such as feature attributions, parameter sensitivity analyses and physically meaningful sub–scores. Domain experts must be able to audit how the model arrives at high or low habitability assessments.

(6) **Reproducibility and Open Science:** Release the methodology, code and processed datasets under an open–access license, with containerized environments to ensure results can be replicated. All releases include model cards documenting intended use and limitations.

(7) **Operational Relevance:** Design the system for direct integration into observational planning. In addition to HL, Exolife outputs a composite EHS* and flags such as "conservative" or "optimistic" that align with mission criteria for follow–up observations.

By achieving these objectives, Exolife addresses critical gaps in existing habitability estimation methods and positions itself as a living system that evolves with incoming data and improved physics.

4 Problem Approach

Estimating exoplanet habitability is a *probabilistic inference challenge* where the target (true habitability) is unobserved and proxy labels are noisy. Exolife outputs two kinds of results:

– **Habitability Likelihood (HL):** A continuous probability $\hat{p} \in [0, 1]$ representing the model's belief that a planet could sustain surface liquid water. HL is trained using weak supervision and positive–unlabeled learning on curated soft labels.

– **Exolife Habitability Score (EHS*):** A human–readable composite score obtained by aggregating five interpretable sub–scores via a logit–space weighted mean. EHS* is not a training target; rather, it is calibrated to HL post–training and used for communication and triage.

The sub–scores capture distinct physical aspects of habitability:

SLWP – Surface Liquid Water Probability: Probability that climate simulations yield liquid water, derived from climate emulators trained on general circulation models (GCMs).

ARP – Atmospheric Retention Probability: Likelihood of retaining a volatiles atmosphere, based on escape velocity v_{esc}, surface gravity g and ultraviolet irradiation history.

HZP – Habitable Zone Proximity: A tapered score reflecting how close the planet is to the HZ boundaries defined by Kopparapu et al. [1]. Planets deep inside the HZ score near one, while those far outside score near zero.

SAP – Stellar Activity Penalty: An attenuation factor derived from the integrated XUV dose and flare rate; high stellar activity lowers habitability prospects.

TSI – Tidal/Spin Stability Index: A metric combining tidal locking timescale, tidal heating and obliquity; extremely high or low values may destabilize climate.

These sub–scores provide interpretable insights while serving as auxiliary heads during training. To avoid proxy feedback loops, EHS* is computed only after training using a monotonic logit–space combiner:

$$\mathrm{EHS}^* = \sigma\left(\frac{\sum_{j\in\mathcal{A}} w_j \,\mathrm{logit}(s_j)}{\sum_{j\in\mathcal{A}} w_j}\right), \tag{2}$$

where $\mathcal{A}$ indexes the available sub–scores s_j, w_j are positive weights (currently fixed but potentially learnable under monotonic constraints), $\mathrm{logit}(x) = \log\frac{x}{1-x}$, and σ is the logistic function. Equation (2) prevents multiplicative collapse when any single s_j is low and ensures the aggregate increases monotonically with each sub–score. EHS* is then calibrated to HL via isotonic regression on a validation set to align the presentation layer with the predictive model.

4.1 Challenges

Several factors complicate exoplanet habitability modeling:

– **Small labeled sample:** Only $\sim$ 6,000 confirmed planets have measured or inferred parameters, and even fewer reside in the temperate, rocky regime. The majority of Kepler and TESS candidates lack confirmation.

– **No true labels:** We have no direct observations of life or surface conditions on exoplanets. Habitability must be inferred from proxies such as HZ membership, equilibrium temperature, atmospheric escape models and climate simulations.

– **Selection bias:** Transit and radial–velocity surveys preferentially detect short–period, large–radius planets around certain stellar types. This leads to covariate shift between the training distribution and the population of interest (e.g., long–period temperate planets).

– **Sparse features and missing data:** Many planets lack measured masses, radii or stellar activity indicators. Missingness itself carries information—detected planets are those whose properties allow detection—and must be encoded.

– **Heterogeneous uncertainties:** Different instruments and catalogs report measurements with varying precision. Uncertainty must be propagated into derived quantities and model predictions.

To overcome these challenges, Exolife integrates multi–source data, physics–guided simulations, weak supervision, selection weighting and active learning.

5 Data Sources

The data strategy supports the generation of HL targets, training of sub–score heads and computation of EHS*. We categorize sources into observational catalogues, unlabeled corpora, physics–based simulations and derived features.

5.1 Observational Catalogs

Core features are drawn from:

– **NASA Exoplanet Archive:** Canonical planetary and stellar parameters (e.g., P, R_p, M_p) and detection metadata.

– **PHL Exoplanet Catalog:** Habitability–oriented aggregates such as the ESI and preliminary classes; these inform the label model but are down–weighted to mitigate circularity.

– **Gaia Mission Data:** High–precision parallaxes and photometry yield stellar luminosities $L_\star$ and radii $R_\star$ with quantified uncertainties.

– **SWEET–Cat:** Spectroscopic stellar parameters (effective temperature T_{eff}, metallicity [Fe/H]) for accurate host characterization.

– **Mission vetting catalogs:** The Kepler Object of Interest (KOI) list, TESS Objects of Interest (TOI) and ExoFOP provide ephemerides, contamination flags and high–resolution imaging follow–up.

Cross–matching uses primary keys `gaia_source_id` for stars and (host name, planet letter) for exoplanets. Proper motion corrections align positional data across missions, and conflicting measurements are resolved via deterministic precedence (Gaia → SWEET–Cat → mission catalogs → archive). For each numeric parameter, we store the posterior mean and standard deviation (and covariances when available), enabling downstream uncertainty propagation.

5.2 Large Unlabeled Corpora

For self–supervised representation learning, we leverage unlabeled data:

– **Unconfirmed planets and candidates (KOIs, TOIs):** Treated as unlabeled examples for pretraining.

– **Light curves (Kepler, K2, TESS):** Time–series of brightness variations encode stellar rotation, activity and planetary transits.

– **Radial velocity (RV) series:** Spectroscopic Doppler shifts provide information on mass, eccentricity and stellar activity.

– **Stellar spectra:** High–resolution spectra constrain $T_{\rm eff}$, metallicity and activity proxies such as the chromospheric index $\log R'_{\rm HK}$.

– **Large stellar catalogs:** Additional photometric surveys (e.g., LAMOST) improve coverage and support cross–modal alignment.

These corpora train modality–specific encoders that produce astrophysical embeddings used in HL and sub–score models.

5.3 Physics–Based Simulations & Surrogate Models

Simulations fill under–represented regimes and provide supervised targets for sub–scores:

– **Population synthesis:** Planet occurrence models calibrated to observed distributions across radius R_p, orbital period P and stellar $T_{\rm eff}$ generate synthetic systems, especially long–period and moderate–eccentricity planets.

– **Climate emulators:** Surrogate models trained on 1D/3D general circulation model (GCM) grids predict surface temperature, presence of liquid water, ice line latitude and day–night contrast given inputs such as incident flux S, albedo A, surface gravity g, atmospheric composition, eccentricity e and rotation rate. Following our improvements, we train *ensembles* of emulators and report mean predictions together with disagreement metrics. When ensemble members disagree beyond a threshold, the model abstains and records a high uncertainty for SLWP and related targets.

– **Atmospheric escape models:** Energy–limited and photo–chemical surrogates estimate mass–loss timescales conditioned on UV/XUV histories, stellar wind pressure and planet size. We similarly adopt ensemble surrogates and propagate their disagreement into ARP.

– **Domain randomization:** To avoid simulator overfitting, nuisance parameters like albedo, greenhouse factor and redistribution efficiency are sampled from astrophysically plausible priors. This ensures the emulator training covers a broad parameter space.

These simulations are used both to pretrain encoders (auxiliary tasks) and to produce weak targets (SLWP, ARP, TSI) for the habitability label model.

5.4 Derived and Engineered Features

From observational and simulated data we compute features such as:

– **Stellar flux** $S/S_\oplus = (L_\star/L_\odot)/a^2$, where $L_\odot$ is the Sun's luminosity. This enters both HL and HZP.

– **Equilibrium temperature** $T_{\rm eq}$, estimated by

$$T_{\rm eq} = \left(\frac{(1-A)S}{4\sigma\varepsilon} \right)^{1/4}, \tag{3}$$

where A is the Bond albedo, σ is the Stefan–Boltzmann constant and ε is the heat redistribution efficiency. We propagate uncertainties in S and A through Monte Carlo samples.

– **Surface gravity** $g = GM_p/R_p^2$ and **escape velocity** $v_{\rm esc} = \sqrt{2GM_p/R_p}$, with G the gravitational constant. These inform ARP and HL features.

– **Tidal locking and heating proxies:** Using orbital elements, we estimate the time to tidal locking and tidal heating rate $\propto e^2/a^6$; high values penalize TSI.

– **Activity proxies:** Flare rates and integrated XUV dose, derived from light curves and spectra, drive SAP.

– **Data quality indicators:** Signal–to–noise ratio (S/N), completeness and missingness flags help estimate prediction uncertainty.

5.5 Label Model Inputs

The HL target is constructed by fusing multiple noisy sources $h_i(x) \in [0, 1]$ representing HZ proximity, equilibrium temperature windows, atmospheric retention likelihood, stellar activity penalties, climate emulator outputs and catalog–derived indices (e.g., ESI). The generative label model described in Section 7.5 combines these sources into a probabilistic soft label $\tilde{y}$.

6 Key Astrophysical Parameters

Habitability arises from interactions among stellar, orbital and planetary properties. Exolife's feature space includes direct measurements, derived quantities and their uncertainties.

6.1 Stellar Parameters

The host star defines the radiation environment and long–term stability of the habitable zone. Relevant variables include:

– **Luminosity** $L_\star$**:** The total energy output determines the incident flux and sets the scale of the HZ via Eq. (1).

– **Effective temperature** $T_{\rm eff}$**:** Governs the spectral energy distribution and enters the HZ coefficients [1].

– **Radius** $R_\star$ **and mass** $M_\star$**:** Affect the star's evolution, rotation and gravitational influence on the system.

– **Metallicity [Fe/H]:** Correlates with planet occurrence rates and composition; higher metallicity hosts tend to produce more massive and volatile–rich planets.

– **Stellar activity indicators:** Rotation period, flare frequency, chromospheric activity (e.g., $\log R'_{\mathrm{HK}}$) and UV/XUV flux; high activity can erode atmospheres or sterilize surfaces.

– **Age and multiplicity:** Stellar age influences cumulative radiation dose and HZ evolution, while multiplicity (binary or higher) can destabilize planetary orbits.

6.2 Orbital Parameters

Orbital geometry and dynamics determine climate variability and tidal effects:

– **Semi–major axis a:** Primary determinant of stellar flux S and equilibrium temperature T_{eq}.

– **Eccentricity e:** High eccentricities induce seasonal variations, affect tidal heating and can make habitability intermittent.

– **Inclination i:** Relevant for transit probability and potential climate asymmetries in highly inclined systems.

– **Orbital period P:** Linked to a by Kepler's third law; influences tidal locking timescale.

– **Obliquity and precession cycles:** Axial tilt and long–term oscillations modulate climate patterns (Milankovitch cycles).

6.3 Planetary Parameters

Intrinsic planetary properties shape surface conditions and atmospheric retention:

– **Mass M_p and radius R_p:** Combined to estimate bulk density $\rho_p = 3M_p/(4\pi R_p^3)$ and composition class (rocky, icy or gaseous).

– **Surface gravity g and escape velocity v_{esc}:** Determine the ability to retain an atmosphere and the efficiency of thermal escape processes.

– **Rotation rate:** Influences atmospheric circulation and the generation of a magnetic dynamo.

– **Magnetic field indicators:** Although rarely measured directly, proxies such as rotation period and core composition inform magnetic shielding.

6.4 Derived Habitability Metrics

To condense multiple inputs into physically meaningful indicators, we compute:

– **Stellar flux S:** As above.

– **Equilibrium temperature** $T_{\rm eq}$: given by Eq. (3) with Monte Carlo propagation of uncertainties.

– **Habitable zone classification:** Conservative and optimistic boundaries computed using the polynomial fits of Kopparapu et al. [1].

– **Tidal locking likelihood and heating rate:** Derived from orbital period, semi–major axis and eccentricity.

– **Earth Similarity Index (ESI):** Combines radius, density, temperature and escape velocity; used cautiously in the label model.

6.5 Measurement Quality & Availability

Each parameter is accompanied by an uncertainty estimate and a missingness flag. We propagate measurement uncertainties into derived features via Monte Carlo sampling—drawing $N \in [1, 5\,000]$ samples from the reported posterior for $(L_\star, T_{\rm eff}, R_p, M_p, a, e)$ and computing derived quantities for each sample. Summary statistics (mean, standard deviation, quantiles) are cached for efficient training. Missing values are not globally imputed; instead, we encode binary indicators and perform multiple imputation within cross–validation folds when necessary. This approach preserves uncertainty and mitigates leakage.

7 Planned Methodology

The Exolife pipeline comprises several stages designed to learn a robust, uncertainty–aware habitability estimator under scarce, biased and noisy labels. Figure **??** (not shown) illustrates the overall flow.

7.1 Pipeline Overview

The high–level steps are:

(a) **Data harmonization & uncertainty propagation:** Curate and cross–match catalogs; standardize units; propagate measurement posteriors via Monte Carlo sampling; derive features; encode missingness.

(b) **Self–supervised representation learning:** Train modality–specific encoders on unlabeled light curves, RV series, spectra and tabular data to learn astrophysical embeddings without habitability labels (Section 7.3).

(c) **Physics–guided simulations & surrogate modeling:** Augment sparse regimes using population synthesis and train emulators for climate, atmospheric escape and tidal effects, using ensembles with abstention signals (Section 7.4).

(d) **Weak supervision via probabilistic label model:** Fuse multiple noisy heuristics into soft labels $\tilde{y}$ using a generative model with latent true label y, source accuracies α_i and correlations C_{ij}; incorporate a PU prior $\pi = P(y = 1)$ (Section 7.5).

(e) **Fine–tuning on confirmed exoplanets:** Train the habitability head (HL), ranking head and sub–score heads on confirmed planets using soft labels and selection–aware losses; then unfreeze top layers of encoders for joint end–to–end optimization.

(f) **Validation & interpretability:** Employ temporal, system–wise and population splits; compute calibration curves, SHAP values and independent conditional expectation plots; perform ablations to assess the contribution of each component (Section 7.10).

(g) **Active learning loop:** Identify high–value follow–up measurements (e.g., RV mass for 1–1.5 $R_\oplus$ candidates) that maximize expected information gain; incorporate expert feedback.

(h) **Operational aggregation:** Compute EHS* from sub–scores using Eq. (2); calibrate EHS* against HL; produce flags (conservative, optimistic) aligned with mission criteria.

7.2 Data Ingestion, Harmonization & Uncertainty Propagation

Cross–identification and uncertainty storage were described in Section 5. Here we note that for each system we draw a modest number of Monte Carlo samples ($N = 1{,}000$ by default) from the posterior distributions of key parameters to compute derived features (e.g., S, $T_{\rm eq}$, g, $v_{\rm esc}$, HZ distances, tidal proxies). These samples enable the estimation of predictive means, variances and percentiles. Missingness is encoded as binary flags; imputation is performed only within cross–validation folds to avoid leakage.

7.3 Representation Learning (Self–Supervised Pretraining)

The goal is to learn modality–specific embeddings that capture stellar variability, activity and system context without requiring habitability labels. We use the following approaches:

– **Light curves:** A temporal transformer or 1D convolutional neural network (CNN) with multi–scale dilations encodes normalized flux sequences. Objectives include masked reconstruction (similar to masked autoencoders) and contrastive instance discrimination with physics–preserving augmentations: time warping within ±5–10%, random windowing, sector dropout and additive Gaussian or red noise. Auxiliary forecasting heads predict rotation period and flare counts.

– **Radial velocity series:** A gap–aware transformer or temporal CNN separates Keplerian and stellar activity signals. The encoder is trained using masked prediction and contrastive forecasting, with an auxiliary head to predict activity indicators.

– **Stellar spectra:** A 1D CNN or conformer operates on normalized wavelength grids. Objectives include masked token modeling and across–epoch instance discrimination (spectra of the same star at different times are positive pairs).

– **Tabular catalogs:** A feature transformer (FT–Transformer) or TabTransformer processes static features. We employ denoising autoencoding with uncertainty–aware noise (sampling within reported posteriors) and masked–feature modeling.

– **Cross–modal alignment:** We align embeddings across modalities by maximizing a CLIP–style contrastive objective:

$$\mathcal{L}_{\mathrm{XMod}} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp(\langle z_k^{(\mathrm{LC})}, z_k^{(\mathrm{Spec})}\rangle/\tau)}{\sum_{j=1}^{B}\exp(\langle z_k^{(\mathrm{LC})}, z_j^{(\mathrm{Spec})}\rangle/\tau)}, \tag{4}$$

where $z^{(\mathrm{LC})}$ and $z^{(\mathrm{Spec})}$ denote normalized embeddings of the same star from light–curve and spectral encoders, B is the batch size and τ is a temperature. The objective is applied to all modality pairs using star identity as supervision.

– **Domain–adversarial debiasing:** To suppress mission and detection–method signals that could leak into the habitability prediction, we append a gradient–reversal head that predicts the detection method d from the embeddings and subtract this loss: $\mathcal{L}_{\mathrm{selfsup}} - \lambda\mathbb{E}[\log p(d \mid z)]$. Increasing λ penalizes detection–method information, encouraging the representations to be survey–agnostic.

7.4 Physics–Guided Simulations and Surrogate Models

Simulations supplement observations by generating synthetic data and providing surrogate outputs for sub–scores. Our improvements emphasize *ensemble* emulators and *abstention* when models disagree.

Population Synthesis. We draw synthetic planetary systems from occurrence rates calibrated to $(R_p, P, T_{\mathrm{eff}})$ distributions, sampling moderate–eccentricity orbits and long periods beyond current detection thresholds. These populations fill the training set with under–represented cases and help avoid extrapolation.

Climate Emulators. We train ensembles of neural surrogates on gridded outputs from GCMs. Inputs include incident flux S, albedo A, surface gravity g, atmospheric composition, eccentricity e and rotation rate; outputs include surface temperature, presence of liquid water and day–night contrast. Ensemble disagreement is quantified via predictive variance or pairwise KL divergence. When disagreement exceeds a threshold, the emulator abstains (SLWP is set to missing), and this missingness propagates into HL. This ensemble approach reduces simulator bias and provides an epistemic uncertainty signal.

Atmospheric Escape. We similarly train ensemble surrogates for energy–limited and photo–chemical escape processes conditioned on UV/XUV histories and planetary parameters. Disagreement yields an abstention on ARP.

Domain Randomization. We sample nuisance parameters (albedo, greenhouse factor, heat redistribution) from realistic priors during emulator training. This mitigates overfitting to specific climate assumptions and encourages generalization.

All surrogate outputs carry provenance metadata (simulation version, input sample index) enabling auditability and targeted improvement.

7.5 Weak Supervision: Probabilistic Label Model

Direct labels for habitability are unavailable. Instead, we construct soft labels $\tilde{y} \in [0, 1]$ by fusing m noisy sources $h_i(x)$ that represent HZ status, temperate T_{eq} ranges, atmospheric escape thresholds, climate emulator outputs, PHL indices and activity penalties. We adopt a generative label model akin to Snorkel [7], with latent true label $y \in \{0, 1\}$ and parameters capturing source accuracies α_i and pairwise correlations C_{ij}:

$$\max_{\alpha, C} \sum_{n} \log \sum_{y \in \{0,1\}} p(y) \prod_{i=1}^{m} p(h_i(x_n) \mid y; \alpha_i)\, p(h_{i<j}(x_n) \mid y; C_{ij}). \tag{5}$$

We estimate parameters via expectation–maximization (EM) and marginalize over y to obtain $p(y = 1 \mid h_1, \ldots, h_m)$. A key improvement is the introduction of *structure priors* and jackknife ablations: we encode prior beliefs (e.g., the correlation between HZ and T_{eq} heuristics should be high) and systematically remove each heuristic to ensure that HL does not collapse to a single proxy. Furthermore, we estimate the positive class prior $\pi = P(y = 1)$ using anchor–point methods [8], and incorporate it via PU learning; the nnPU risk estimator then debiases training on soft labels.

Finally, we calibrate the resulting soft labels against a high–confidence subset (e.g., temperate rocky planets with well–measured parameters) using isotonic regression or Platt scaling. This reduces bias in the label model and aligns it with known physically plausible cases.

7.6 Model Architectures and Objectives

The Exolife estimator combines strong tabular baselines with multimodal encoders. Key components include:

Tabular Baselines. We train gradient–boosted tree ensembles (LightGBM, CatBoost, XGBoost) on the Monte Carlo–derived summary statistics. To ensure physically plausible monotonicity, we impose shape constraints (e.g., HL is non–decreasing in SLWP, ARP and HZP, and non–increasing in SAP). Loss functions such as quantile and Poisson/Huber losses handle heavy–tailed distributions and outliers. Ensembles provide fast, interpretable baselines and serve as a control in ablation studies.

Multimodal Encoder–Fusion. We stack the pre–trained encoders from Section 7.3 with a tabular encoder (FT–Transformer). A gated attention or cross–attentional pooling mechanism aggregates modality embeddings and Monte Carlo summary statistics into a unified representation z. The unified representation feeds multiple heads:

– *Habitability likelihood head:* Outputs logits for HL; trained on soft labels with a heteroscedastic likelihood that models aleatoric noise. The positive–unlabeled risk estimator accounts for class prior π.

– *Ranking head:* Optimizes a listwise ranking objective (e.g., differentiable NDCG) to align predictions with expert–defined priorities for telescope scheduling.

– *Sub–score heads (SLWP, ARP, HZP, SAP, TSI):* Each head is trained against its corresponding surrogate or heuristic target. Monotonic constraints are imposed across heads and HL, as noted above, using differentiable inequality penalties.

– *Auxiliary physics heads:* Regress emulator outputs (e.g., surface temperature) and astrophysical properties (e.g., rotation period) to encourage physically meaningful representations. These heads yield inductive bias and help disentangle astrophysical factors.

Missingness in sub–score inputs leads to abstentions in those heads; EHS* reweights available terms accordingly.

7.7 Bias, Selection Effects & Missing Data

Detection methods (transit, radial velocity) and mission sensitivities distort the training distribution relative to the population of interest. To address selection bias, we introduce an explicit **selection–function layer**. For each planet, we estimate its detectability $p_{\text{train}}(x)$ by fitting a propensity model on observed versus synthetic populations (or between older and newer catalog releases). The inverse propensity weight $w(x) = p_{\text{target}}(x)/p_{\text{train}}(x)$ down–weights over–represented systems during training and enters the loss function:

$$\mathcal{L}_{\text{weighted}} = \sum_n w(x_n)\, \ell(\hat{p}_n, \tilde{y}_n). \tag{6}$$

This weighting, combined with domain–adversarial training and worst–group risk minimization, mitigates covariate shift. We report performance across strata (stellar types, detection methods) and monitor worst–group metrics to ensure fairness. Missing data are handled as discussed previously, with imputation inside folds and explicit missingness indicators; models are trained to operate on subsets of features via random feature dropout, improving robustness.

7.8 Uncertainty Quantification & Calibration

Following the decomposition of uncertainty into *aleatoric* (data noise) and *epistemic* (model uncertainty) components, we implement:

– **Aleatoric uncertainty:** The HL head outputs both a mean logit and variance; we assume a Gaussian distribution on the latent logit. This captures heteroscedastic noise due to measurement error.

– **Epistemic uncertainty:** We train deep ensembles (multiple random seeds), apply Monte Carlo dropout, or use SWAG (Stochastic Weight Averaging Gaussian) to approximate posterior uncertainty. For tree baselines, we use bootstrap aggregation. Ensemble variance provides a measure of model confidence.

– **Probability calibration:** Temperature scaling, isotonic regression and Dirichlet calibration are applied to map raw logits to calibrated probabilities. We measure the

calibration error via Expected Calibration Error (ECE) and Negative Log–Likelihood (NLL). EHS* is calibrated to HL via isotonic regression.

– **Conformal risk control:** To produce decision–theoretic thresholds with guarantees, we layer conformal prediction on top of the calibrated probabilities. Split–conformal calibration, optionally stratified by stellar type (Mondrian conformal), yields prediction sets or probability thresholds that achieve a user–specified miscoverage rate. This improvement ensures that planets flagged as promising meet a pre–defined risk tolerance.

7.9 Training Protocol & Hyperparameter Optimization

Training proceeds in three phases:

Phase A (Pretraining): Train modality encoders with self–supervised objectives. Early stopping is based on proxy tasks such as rotation period retrieval. The domain–adversarial penalty coefficient λ is swept over a grid to balance survey–agnosticity and representation quality.

Phase B (Label Modeling): Fit the generative label model (Eq. (5)) and estimate the PU prior π using anchor–point methods. Calibrate soft labels against high–confidence subsets via isotonic or Platt scaling.

Phase C (Fine–Tuning): Freeze encoders and train HL, ranking and sub–score heads using the calibrated soft labels and selection weighting. Then unfreeze the top layers of encoders with a small learning rate and perform joint fine–tuning. Multi–task loss weights are annealed over training. Optimization uses AdamW with cosine decay, gradient clipping and mixed precision. Hyperparameters (learning rates, dropout rates, fusion sizes, loss weights) are tuned via Bayesian optimization.

Regularization techniques include stochastic depth, label smoothing on soft labels and Mixout to stabilize partial fine–tuning.

7.10 Evaluation Protocol & Metrics

To ensure robustness and generalization, we employ a rigorous evaluation scheme:

– **Data splits:** We use temporal splits (training on earlier catalog snapshots and testing on later discoveries), system–wise splits (all planets from the same host in one fold) and population holdouts (e.g., late–M dwarfs, long–period terrestrials). Detection–method holdouts evaluate how well the model generalizes from transit to radial–velocity discoveries and vice versa.

– **Primary metrics:** Probabilistic metrics include NLL, Brier score, ECE and AUROC/PR–AUC (with PU adjustments) for HL. Ranking metrics include NDCG@k and Precision@k for telescope scheduling lists, as well as Kendall's τ correlation with expert rankings.

– **Robustness tests:** We report worst–group performance across stellar types and detection methods, perform covariate–shift stress tests by synthetically altering detection probability distributions, and compute out–of–distribution (OOD) detection AUROC using energy–based or Mahalanobis scores on the fusion embeddings.

– **Statistical confidence:** For all metrics we compute bootstrap confidence intervals (10,000 resamples) and conduct paired permutation tests to assess significance between models. For AUROC comparisons we use DeLong's test.

– **Ablations:** We systematically remove components—pretraining, simulations, the label model, selection weighting, multimodal fusion—and measure performance degradation. We also drop each sub–score head to test the sensitivity of EHS*.

– **Interpretability:** We compute global SHAP values across folds, SHAP interaction values (e.g., flux $\times$ R_p), Individual Conditional Expectation (ICE) curves for key features ($T_{\rm eq}$, HZ distance), per–planet sub–score attributions and case studies of high and low HL predictions. We quantify the consistency of SHAP attributions across model seeds.

7.11 Out–of–Distribution Monitoring & Drift Detection

Energy–based or Mahalanobis scores on the fused embeddings flag predictions with low support in the training data. Population–level drift is measured via population stability index (PSI) or Kullback–Leibler divergence on key features between training and new catalog snapshots. Significant drift triggers retraining or recalibration. Conformal prediction thresholds adapt to distributional shifts, providing formal guarantees on the miscoverage rate.

7.12 Active Learning and Experiment Design

Active learning closes the loop between model predictions and new observations. Acquisition functions of the form $\alpha(x) = \mathrm{UQ}(x) \times \mathrm{Utility}(x)$ select planets with high model uncertainty and high scientific value (e.g., being in the temperate rocky regime). The system recommends the type of follow–up measurement that most reduces posterior entropy—RV mass measurements for planets with known radii, UV/XUV monitoring for active M–dwarfs, high–resolution imaging for diluted transits. Human experts vet the recommendations, and their feedback is incorporated as pairwise ranking constraints in retraining.

7.13 Reproducibility, Governance & Model Cards

All data and model releases are versioned with immutable snapshots and documented seeds. Tests include schema and physical–sanity unit tests, pipeline integration tests and continuous integration (CI) for reproducibility. A model card accompanies each release, documenting intended use, limitations (proxy labels, selection bias), calibration plots, population–wise performance, OOD policy and recommended follow–ups for uncertain predictions. This transparency is essential for scientific trust.

8 Evaluation and Validation

Section 7.10 described the metrics and splits used for model evaluation. In this section we note that the evaluation protocol intentionally simulates deployment conditions: models are trained on data available at a given time and tested on later discoveries. This temporal holdout prevents leakage of information from future discoveries and assesses how well the pipeline generalizes to the continually expanding exoplanet catalog. System–wise and population splits prevent contamination across planets from the same host or across rare classes. The worst–group and selection–weighted metrics ensure that high performance is not driven solely by over–represented populations.

In addition, we validate the reliability of the HL and EHS* outputs via calibration curves (reliability diagrams). We compute the correlation and Kendall's τ between HL and EHS* to verify that the presentation layer reflects the learned habitability likelihood. Conformal prediction sets with a miscoverage rate of $\alpha = 0.1$ produce calibrated confidence intervals for each prediction. For interpretability, we ensure that SHAP attributions for key features are stable across seeds and that their global importance aligns with astrophysical expectations (e.g., increased flux lowers habitability after a threshold).

9 Project Roadmap

The development of Exolife follows a phased roadmap:

(i) **Data Harmonization and Pretraining (Months 1–6):** Cross–match catalogues; implement Monte Carlo uncertainty propagation; train self–supervised encoders. Develop and test domain–adversarial debiasing.

(ii) **Simulation & Label Model (Months 4–8):** Construct population synthesis; train ensemble climate and escape emulators; fit and calibrate the label model; estimate the PU prior.

(iii) **Fine–Tuning & Validation (Months 7–12):** Train HL, ranking and sub–score heads; implement selection–function weighting; perform hyperparameter optimization; conduct evaluation on temporal and population splits; refine monotonic constraints.

(iv) **Deployment and Active Learning (Months 12+):** Integrate the model into observation planning tools; roll out EHS* scoring; implement active learning for follow–up measurements; begin iterative retraining with new data.

Each phase includes thorough documentation and release of code and models. Continuous evaluation guides go/no–go decisions based on calibration (ECE ≤ 0.05), ranking improvement (ΔNDCG@20 ≥ 0.10 over baseline) and OOD detection performance (AUROC ≥ 0.85).

10 Conclusion

Exolife provides a cohesive and scientifically grounded model for estimating the habitability potential of exoplanets in the face of scarce, biased and noisy data. Its central output, the Habitability Likelihood (HL), is a probabilistic, calibrated score trained via physics–guided, weakly supervised and PU–aware methods. Five interpretable sub–scores (SLWP, ARP, HZP, SAP, TSI) accompany HL, allowing domain experts to understand the contributing factors. The Exolife Habitability Score (EHS*) aggregates sub–scores in logit space to provide a human–readable triage metric without influencing training. A comprehensive data strategy, encompassing harmonized catalogues, unlabeled corpora and physics–based simulations, ensures broad coverage. Explicit handling of selection bias, missingness, measurement noise and model calibration yields robust predictions. Evaluation protocols simulate real–world deployment via temporal and population holdouts, and active learning facilitates iterative improvement.

The improvements discussed herein further strengthen Exolife. Introducing an explicit selection–function layer addresses detectability bias; ensemble emulators with abstention reduce simulator–to–real bias and provide uncertainty estimates; monotonic constraints enforce physical plausibility across HL and sub–scores; conformal calibration offers rigorous decision–time guarantees; and generative label modeling with structure priors prevents collapse to single heuristics. These enhancements enhance trustworthiness and reliability, paving the way for Exolife to serve as a decision–support system for future astrobiological discoveries.

11 Future Refinements

While the current plan is robust, several refinements can enhance performance and adaptability:

- **Quantitative go/no–go thresholds:** Define explicit acceptance criteria for calibration (e.g., ECE ≤ 0.05), ranking improvement (ΔNDCG@20 ≥ 0.10) and OOD detection (AUROC ≥ 0.85) to guide model release.
- **Cost–weighted ranking:** Integrate observing cost and science–value weights into the ranking head's loss function, aligning prioritization with mission scheduling constraints.
- **Expanded physics models:** Incorporate coupling between tides, atmosphere and magnetic fields for more realistic TSI and ARP estimates. Use multi–simulator ensembles to reduce bias from any single climate or escape model.
- **Dynamic label model updating:** Automate label model retraining with each data refresh, incorporating new observational constraints and updated simulation outputs. This prevents drift in the soft labels.
- **Enhanced provenance and auditability:** Ship sub–score outputs with full input provenance, measurement uncertainties and abstention flags, enabling domain experts to audit the decision path.

– **Adaptive EHS* weighting:** Explore learning weights w_j in Eq. (2) within cross–validation folds, subject to monotonicity constraints, to better reflect the relative importance of sub–scores.

– **Human–in–the–loop feedback:** Incorporate expert feedback on top–ranked candidates into the active learning loop, refining both HL calibration and ranking performance.

– **Uncertainty–aware simulation targeting:** Use HL uncertainty maps to guide further simulation runs, focusing computational effort on under–explored regions of parameter space where emulator errors are highest.

Addressing these refinements will strengthen Exolife's role as a decision–support tool for exoplanet science, enhance adaptability to new data and models and maintain a balance between prediction performance, interpretability and scientific plausibility.

References

[1] Kopparapu, R. K., et al. (2013). Habitable zones around main–sequence stars: new estimates. *The Astrophysical Journal*, **765**(2), 131.

[2] Schulze–Makuch, D., et al. (2011). A two–tiered approach to assessing the habitability of exoplanets. *Astrobiology*, **11**(10), 1041–1052.

[3] Heller, R., & Armstrong, J. (2014). Superhabitable worlds. *Astrobiology*, **14**(1), 50–66.

[4] Gaia Collaboration (2023). Gaia data release 3: summary of the content and survey properties. *Astronomy & Astrophysics*, **674**, A1.

[5] Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[7] Ratner, A., et al. (2020). Snorkel: rapid training data creation with weak supervision. *VLDB Journal*, **29**, 709–730.

[8] Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 213–220).

[9] Luger, R., et al. (2016). starry: analytic occultation light curves. *The Astronomical Journal*, **152**(4), 100.

[10] Foreman–Mackey, D., et al. (2013). emcee: the MCMC hammer. *Publications of the Astronomical Society of the Pacific*, **125**(925), 306–312.