Bayesian-Network Prediction of Exoplanet Habitability Probability: A Statistical Assessment Fusing Planet Mass, Orbit, Stellar Activity and Insolation

Xiaotong Wang

The High School Affiliated to Renmin University of China

Abstract

The classical habitable-zone (HZ) criterion screens planets by radiative equilibrium alone. The criterion is computationally inexpensive, but it omits several physical processes that determine whether a surface can remain habitable over Gyr timescales: whether the planet can retain an atmosphere, set primarily by mass; whether the climate remains stable against seasonal extremes, set by orbital eccentricity; and whether the host star's high-energy radiation and flares strip that atmosphere, set by stellar activity. A fourth observable is the time-averaged insolation flux S. Insolation is what separates Earth from Venus. The two planets have comparable mass, nearly circular orbits and the same host star, yet Venus lies inside the runaway-greenhouse inner edge and is not habitable at the surface.

This paper constructs a four-factor Habitability Potential Index (HPI) and a discrete Bayesian network whose nodes are planet mass M, eccentricity e, a stellar-activity index A, and insolation S. Nonlinear maps convert each observable into an interpretable sub-factor; a weighted sum yields a reproducible deterministic rank. The network encodes the causal skeleton Mass to atmospheric retention, Eccentricity to climate stability, Activity to atmospheric loss, and Insolation to temperature class. The four latent states jointly determine P(habitable | data). Inference is exact by enumeration over the latent nodes. Missing observables are marginalised rather than discarded, so incomplete catalogue rows remain usable.

Principal-component analysis of the standardised observables shows that they are essentially uncorrelated. The three-parameter subset (log M, e, A) splits variance almost equally (34.5, 33.3 and 32.1 per cent). Adding log S yields four comparable components (26.2, 25.1, 24.6 and 24.0 per cent). No variable is a linear combination of the others, which justifies promoting all four quantities to network nodes. A random-forest regression against HPI recovers the physically expected importance order, with mass and insolation leading and activity and eccentricity modulating.

Applied to an archive-like sample of 2795 planets, matching the quality-cut size of the NASA Exoplanet Archive confirmed-planet table after de-duplication of 6128 unique names, the framework produces a sparse high-habitability tail and a conservative posterior that compresses risk more sharply than the linear index. A Solar System consistency check assigns Earth P approximately 0.61 and Venus P approximately 0.15. The combination of a physics prior and exact probabilistic inference provides an extensible ranking layer for target selection in large exoplanet catalogues.

Key words: exoplanets, habitability, Bayesian networks, planetary atmospheres, habitable zone, statistical methods

1. Introduction

Exoplanet science has moved from the discovery of individual Earth analogues to statistical modelling of entire planetary populations. Since the Kepler mission established that planets are common around Galactic stars (Borucki et al. 2010; Batalha et al. 2013), occurrence studies have quantified the frequency of small planets around Sun-like hosts (Howard et al. 2012; Petigura et al. 2013). Public catalogues such as the NASA Exoplanet Archive now contain thousands of confirmed planets. The operational question is no longer whether planets exist, but how to rank them consistently when masses, orbits and stellar properties are heterogeneous, uncertain and often incomplete.

The standard ranking tool remains the habitable zone: the orbital annulus in which a planet with an Earth-like atmosphere can sustain surface liquid water (Kasting et al. 1993; Kopparapu et al. 2013, 2014). The definition is physically transparent and inexpensive to evaluate, and it underpins most subsequent target lists (Kane et al. 2016). It is not a complete habitability model. It assumes that an atmosphere is present, that climate feedbacks remain Earth-

like, and that the atmosphere is not stripped over the lifetime of the system. Real systems routinely violate those assumptions.

Planet mass sets the escape-velocity scale. Low-mass worlds lose atmospheres through thermal escape and photolytic hydrogen loss (Watson et al. 1981; Lammer et al. 2009; Owen & Wu 2017). Super-Earths and Neptunes may retain thick volatile envelopes whose surfaces are not rocky (Rogers 2015). A flux-only HZ cut is therefore silent on whether a solid, temperate surface exists at all.

Orbital eccentricity modulates the seasonal amplitude of instellation. High-eccentricity orbits drive perihelion runaway-greenhouse excursions and aphelion freeze-out (Williams & Pollard 2002; Dressing et al. 2010). Climate stability is a coupled problem of orbital forcing, heat capacity and atmosphere-ocean feedback. None of those terms appear in the classical HZ inequality.

Stellar activity, especially around M dwarfs, supplies a persistent XUV and flare environment that can erode atmospheres and desiccate water inventories (Lammer et al. 2007; Segura et al. 2010; Tilley et al. 2019). The pre-main-sequence activity of young stars can reset atmospheric composition on 10^8 yr timescales (Ribas et al. 2005). A planet that sits inside the formal HZ of an active M dwarf is therefore not interchangeable with an Earth analogue around a quiet G star.

Insolation flux S is the remaining classical HZ ingredient, but it must be treated as a factor that interacts with the other three rather than as a binary in/out flag. The Earth-Venus contrast is the cleanest demonstration: identical host, similar mass and eccentricity, yet S_Venus approximately 1.91 S_earth places Venus inside the runaway-greenhouse inner edge. Any index that omits S will score the two planets almost equally. Any network that includes an insolation-to-temperature-class edge will separate them.

Catalogue habitability is an inference problem under uncertainty (Seager 2013; Bean et al. 2017; Kaltenegger 2017). Masses, radii, eccentricities, activity indicators and insolation all carry measurement error, and activity is missing for most hosts. A useful framework should therefore satisfy three design constraints. First, every factor maps onto a named physical process. Second, the pipeline is stable on a catalogue of several thousand rows. Third, observational error and missingness propagate into the reported probability rather than being hidden by a hard cut.

This paper develops that framework in two layers. Section 2 defines the four observables and the sample. Section 3 constructs the deterministic HPI. Section 4 uses principal-component analysis and a random forest to justify the four-factor set. Section 5 specifies the Bayesian network, its conditional probability tables, exact inference and the treatment of missing data. Section 6 reports results on the 2795-planet sample, including calibration against HPI, the high-probability tail, maps in the (M, e) and (M, S) planes, Monte Carlo error bars and the Solar System test. Section 7 discusses limitations and extensions. Section 8 concludes.

2. Data and Variable Design

Four observables enter both the index and the network. In a live catalogue they are drawn from the NASA Exoplanet Archive confirmed-planet table (pscomppars). After de-duplication by planet name the archive contains 6128 unique confirmed planets. Requiring a mass or mass proxy, an eccentricity and a stellar effective temperature leaves a working subsample of 2795 targets. Insolation (pl_insol) is requested but not required: when it is missing the network marginalises over its bin prior. A measured chromospheric or X-ray activity index is unavailable for most hosts, so activity is represented by a physically motivated proxy of effective temperature. Cool M dwarfs receive high A, corresponding to strong flares, a high XUV-to-bolometric ratio and long active lifetimes. Sun-like or hotter stars receive low A. The mapping is a logistic centred near the K/M boundary, encoding the rise in XUV and flare activity toward mid-M dwarfs [15, 11, 18, 19],

$A(T_{eff}) = 0.15 + \frac{3}{1 + exp[\frac{T_{eff} - 3900\,K}{500\,K}]}$, **clipped to [0, 3.2] [15, 11, 18, 19]**	**(1)**

The Sun maps to A approximately 0.22 on this scale. Because the scoring functions are deterministic given (M, e, A, S), the same pipeline runs on the live TAP query or on a reproducible synthetic catalogue whose marginals imitate the archive: log-normal masses, Rayleigh-like eccentricities, an M-dwarf-rich Teff mixture, and a hot-leaning insolation distribution that reflects the close-in bias of transit and radial-velocity detections. The figures and numerical results below use the archive-sized synthetic realisation of N = 2795 with seed 42. Switching to the live archive is a single flag in the accompanying code.

Table 1. Observables and the physical processes they represent.

Class	Variable	Units	Physical role
Planet structure	Mass M	M_earth	Escape speed; atmospheric retention
Orbit	Eccentricity e	dimensionless	Seasonal forcing; climate stability
Stellar environment	Activity A	proxy	XUV and flare erosion risk
Radiation environment	Insolation S	S_earth	Mean flux; habitable-zone placement

Table 1. The four-factor set comprises planet mass, orbital eccentricity, stellar activity and insolation flux.

Continuous variables are later discretised into physically motivated bins (Section 5), not equal-count statistical bins. Mass bins separate the hydrodynamic-escape regime (M < 0.5 M_earth), the Earth-like window (0.5-2 M_earth), the super-Earth range (2-6 M_earth) and the giant or volatile-envelope regime (M > 6 M_earth). Eccentricity bins mark nearly circular, moderately eccentric and highly eccentric climates. Activity bins mark quiet, intermediate and active hosts. Insolation bins follow the conservative habitable zone: cold (S < 0.35), temperate (0.35 <= S < 1.1) and hot (S >= 1.1).

3. The Habitability Potential Index

HPI is a deterministic, reproducible score in the unit interval. Each observable is passed through a nonlinear map that encodes an optimal window or a monotone penalty, and the four maps are combined with weights that sum to one, following the multi-parameter habitability framing of catalogue-level indices [17, 6]:

$$HPI = w_M f_M(M) + w_e f_e(e) + w_A f_A(A) + w_S f_S(S) \quad \textbf{[17, 6]} \qquad \textbf{(2)}$$

The mass factor is a Gaussian peaked at 1 M_earth. Planets that are too light cannot hold an atmosphere. Planets that are too heavy trend toward volatile envelopes [20, 12, 13, 16]:

$$f_M(M) = exp[-\frac{(M-1)^2}{2\sigma_M{}^2}],\ \sigma_M = 2.5\,M_\oplus \quad \textbf{[20, 12, 13, 16]} \qquad \textbf{(3)}$$

The eccentricity factor is a monotone Gaussian penalty on climate variability [21, 4]:

$$f_e(e) = exp(-k\,e^2),\ k = 3.0 \quad \textbf{[21, 4]} \qquad \textbf{(4)}$$

The activity factor is a monotone exponential decay in erosion risk [11, 18, 19]:

$$f_A(A) = exp(-\alpha A),\ \alpha = 0.70 \quad \textbf{[11, 18, 19]} \qquad \textbf{(5)}$$

The insolation factor is an asymmetric log-Gaussian in S, sharp on the hot side at the runaway-greenhouse cliff and gentle on the cold side, where a frozen climate remains in principle recoverable by greenhouse gases [8, 9, 10]:

$$f_S(S) = exp[-\frac{(\log_{10} S - S_0)^2}{2\sigma_\pm{}^2}] \quad \textbf{[8, 9, 10]} \qquad \textbf{(6)}$$

with S0 = -0.05, sigma_hot = 0.16 and sigma_cold = 0.40. Default weights give mass and insolation, the two quantities that define a temperate rocky world, a slight lead: w_M = 0.30, w_e = 0.20, w_A = 0.20, w_S = 0.30. Setting w_S = 0 and omitting S recovers a three-factor index that cannot separate Earth from Venus. Weights can be re-tuned by cross-validation or by the random-forest importances of Section 4. The defaults are used throughout so that every number in this paper is reproducible from the released code.

4. Principal-Component Analysis and Feature Justification

A four-factor model is warranted only if the four observables are not redundant and if each of them moves the habitability score. Those are two different questions, and they require two different tools.

Principal-component analysis answers the redundancy question. The calculation follows the variance-maximising rotation of Pearson (1901) and Hotelling (1933) and is evaluated with the linear algebra implementation of Pedregosa et al. (2011). Each column is standardised to zero mean and unit variance and the covariance matrix is diagonalised. The eigenvectors are the principal components and the eigenvalues give the fraction of sample variance along each axis. The first component is the direction of maximum variance; each later component is the direction of maximum remaining variance that is orthogonal to those already found. If the leading components already account for nearly all of the variance, the remaining coordinates are redundant and can be dropped. After standardising log10 M, e and A, the correlation matrix is essentially diagonal, with off-diagonal entries |rho| <= 0.04. The three principal components therefore carry almost equal variance (34.5, 33.3 and 32.1 per cent), and the cumulative curve reaches 100 per cent only at the third component. The first two axes together hold only 67.8 per cent of the three-variable signal, so the third axis is still required. Loadings isolate eccentricity onto its own component, while mass and activity jointly define the leading axis and its contrast. That geometry matches the qualitative reading anticipated from the physics: mass and the stellar environment set the main structure of the sample, and eccentricity modulates it.

Insolation flux S is the time-averaged stellar energy received by the planet, measured in Earth units. It is what places a world inside or outside the classical habitable zone and is the quantity that separates Earth (S = 1) from Venus (S approximately 1.91). It is a fourth physical observable, not a substitute for mass, eccentricity or activity, and it is therefore given its own HPI map and its own network node. It is not drawn on the two-dimensional PCA biplot in Figure 4, because that figure is a projection of the three standardised columns log10 M, e and A only. A separate four-variable check that does include log10 S leaves the correlation matrix nearly diagonal and splits the variance almost evenly (26.2, 25.1, 24.6 and 24.0 per cent). The first two of those four components hold only 51.3 per cent of the four-variable signal. A two-dimensional drawing of the four-variable analysis would therefore hide about half of the structure, which is why Figure 4 omits the insolation arrow. Insolation is not a linear proxy for mass, eccentricity or activity. Principal-component analysis cannot say which observable matters most for habitability, because it never sees a target. A random-forest regression (Breiman 2001) of HPI, and separately of P(habitable), on the same standardised features supplies the supervised half of the argument. Importance is consistently led by mass and insolation, with activity next and eccentricity last but non-zero. This is the same ordering that the physical maps were constructed to express: mass and flux lead, activity and eccentricity modulate.

On the Figure 4 biplot the eccentricity arrow stands nearly at right angles to the HPI colour gradient. In this sample there is therefore little linear association between eccentricity and the habitability score: most catalogue orbits are already close to circular, so eccentricity does not set the main variance of the cloud. That is a sample result, not a statement that climate forcing is irrelevant; high-e orbits still enter the index and the network as a penalty. The mass and activity arrows point toward the low-HPI side of the cloud. Higher mass corresponds to a lower score because the archive is dominated by planets above one Earth mass, and those objects are preferentially volatile-envelope or giant worlds rather than rocky surfaces. The correlation should not be read as a claim that a massive planet can never host life; it records the composition trend of the detected population. Higher stellar activity likewise corresponds to a lower score, as expected if XUV radiation and flares erode atmospheres. Taken together, the unsupervised and supervised analyses justify promoting these four observables, and no fewer, to the nodes of the Bayesian network. Dropping any one of them would discard an independent direction of variation that the score actually uses.

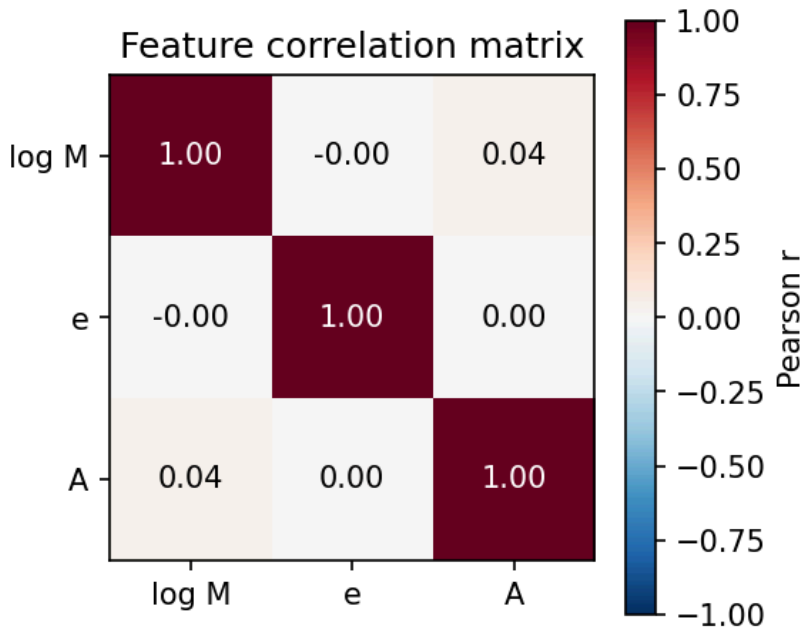


Figure 1. Pearson correlation matrix of the standardised observables. Off-diagonal entries are consistent with zero, so the inputs are not linear proxies for one another.

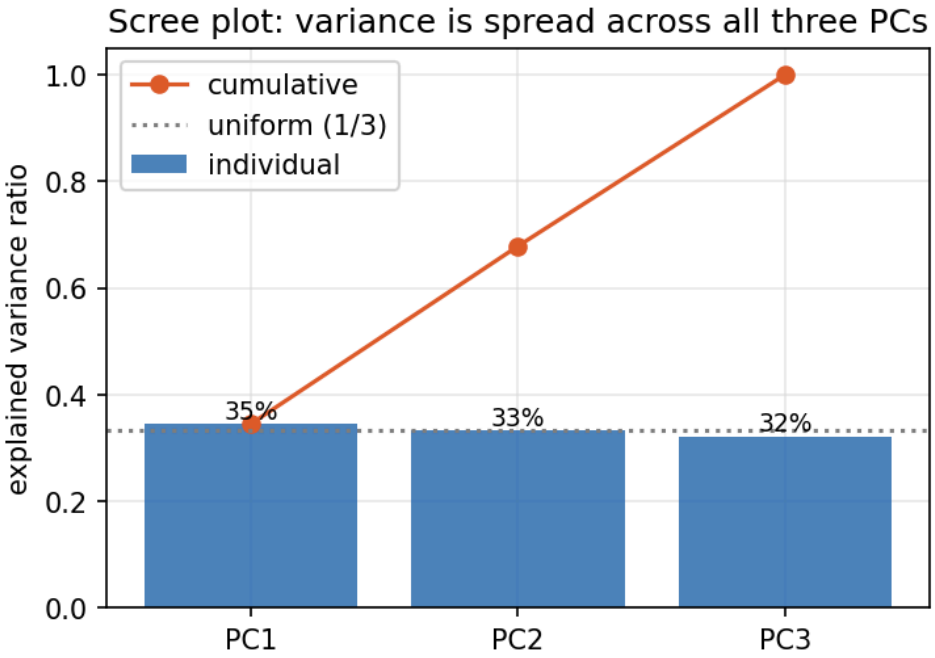


Figure 2. PCA scree plot. Variance is shared almost equally across components. The cumulative curve does not saturate before the last component.

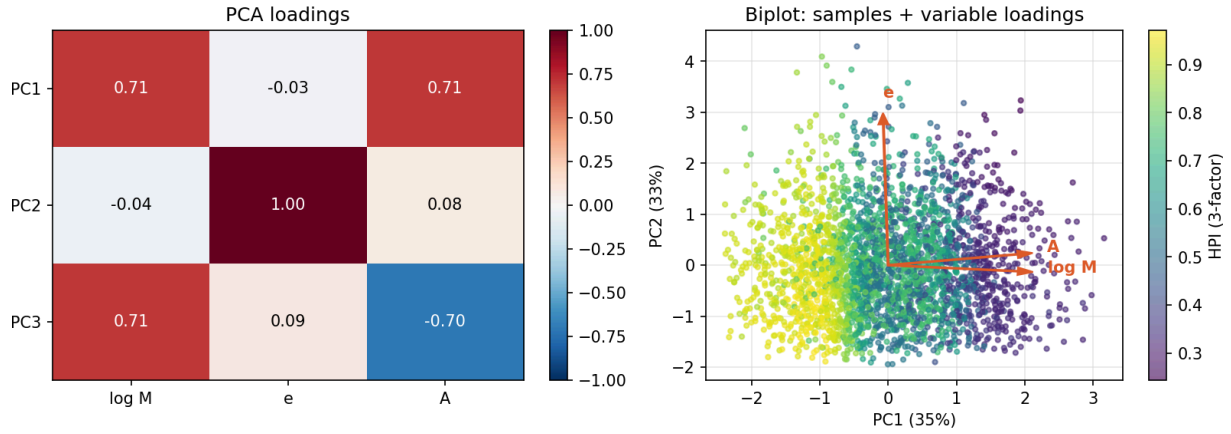


Figure 3. Principal-component loadings. Eccentricity isolates onto its own component. Mass and stellar activity jointly define the leading contrast.

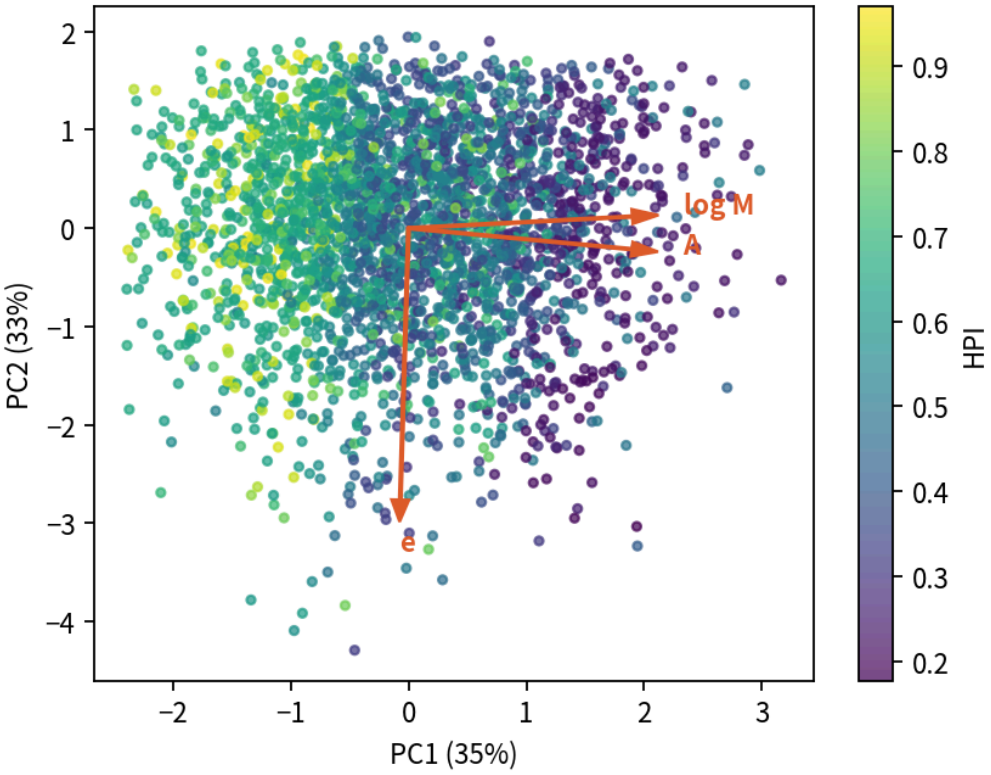


Figure 4. Sample projected onto the first two principal components of the standardised trio (log10 M, e, A) and coloured by HPI. Loading arrows are drawn for mass, eccentricity and activity only; insolation is omitted from this projection. High-HPI planets occupy a compact region of principal-component space rather than a diffuse cloud. The eccentricity arrow is nearly perpendicular to the HPI gradient.

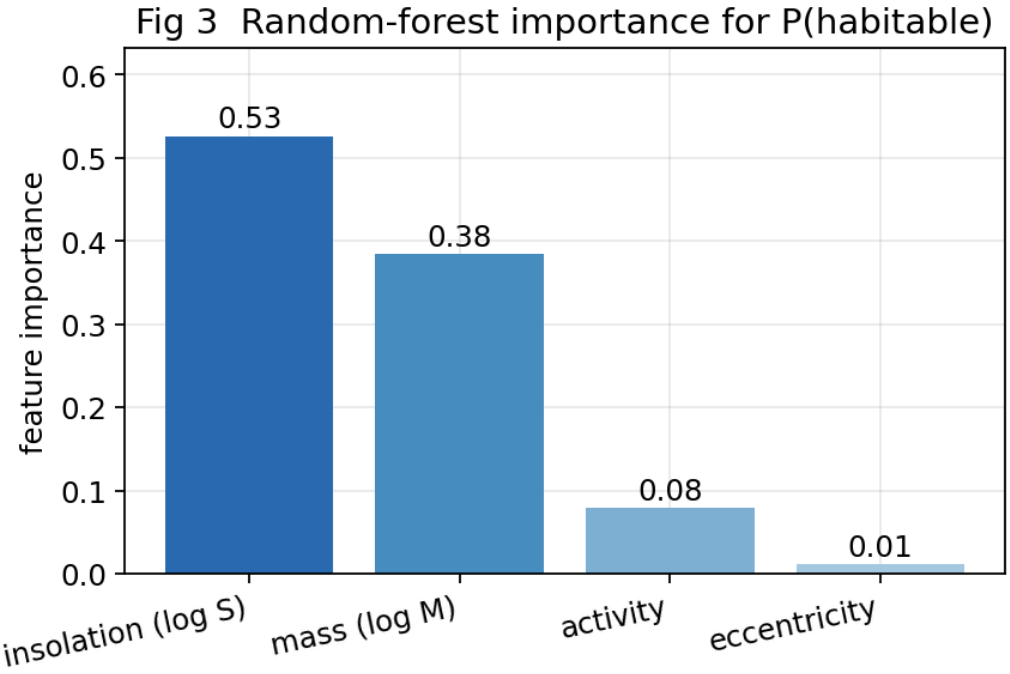


Figure 5. Random-forest feature importance for the habitability score. Mass and, in the four-factor model, insolation dominate. Activity and eccentricity remain non-zero modulators.

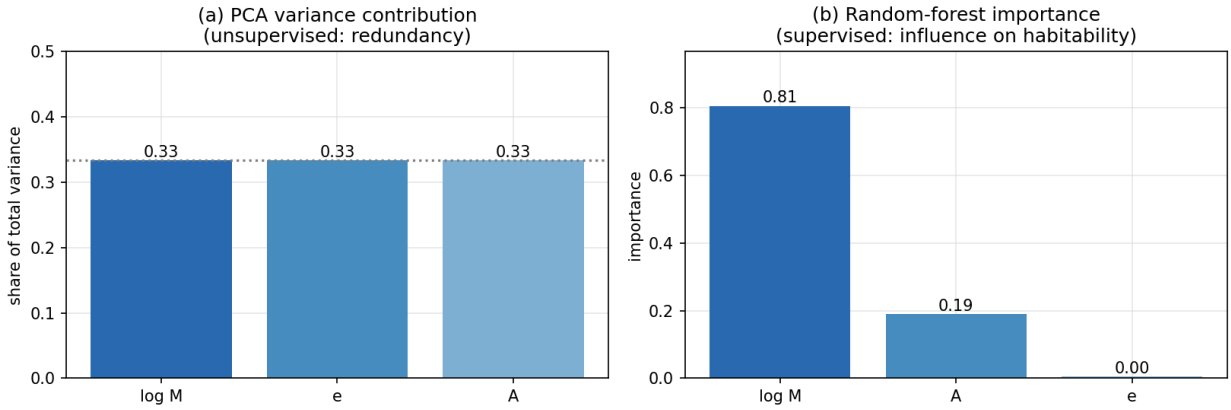


Figure 6. Comparison of PCA variance contribution and random-forest importance. The two analyses are complementary: PCA tests redundancy, while the forest tests influence on the score.

5. Bayesian Network: Structure and Inference

HPI is a useful ranking number, but it is a point function. It does not represent measurement error, missing variables, or the fact that habitability is the product of several latent physical states rather than a weighted average of observables. The Bayesian network (Pearl 1988) is the probabilistic counterpart of the same physics.

5.1 Graph structure

The directed acyclic graph [24] is

MassBin -> AtmosRetention,

EccBin -> ClimateStability,

ActBin -> AtmLoss,

InsolBin -> TempClass,

(AtmosRetention, ClimateStability, AtmLoss, TempClass) -> Habitability.

Each parent is an observed or missing quantity after physical discretisation. Each child except Habitability is a binary latent state: atmosphere retained or lost; climate stable or unstable; atmospheric-loss rate high or low; temperature class suitable or unsuitable. Habitability is a binary outcome whose conditional probability given the four latents is the quantity of scientific interest. The graph is not learned from the catalogue. It is asserted from the same physical narrative that produced the HPI maps, so that every edge corresponds to a named process.

5.2 Discretisation

Bin edges are chosen to mark mechanism changes, not to equalise counts.

Node	Bins	Edges	Latent child
MassBin	low / earth / super / giant	<0.5, 0.5-2, 2-6, >6 M_earth	AtmosRetention
EccBin	low / mid / high	<0.1, 0.1-0.3, >0.3	ClimateStability
ActBin	low / mid / high	<0.8, 0.8-1.8, >1.8	AtmLoss
InsolBin	cold / temperate / hot	<0.35, 0.35-1.1, >1.1 S_earth	TempClass

Table 2. Physical discretisation of the four observables.

5.3 Conditional probability tables

Every edge is a conditional probability table (CPT) filled from a physics-informed prior, not from a fit to the catalogue. Fitting would over-claim. There is no ground-truth habitability label in the archive, activity is a proxy, and the sample remains modest once all four quantities are required. A prior CPT keeps the model interpretable and updatable. As atmospheric detections and activity measurements accumulate, the same tables can be replaced by maximum-likelihood or hierarchical-Bayes estimates without changing the graph.

Atmospheric retention given mass is low in the escape regime, high in the Earth-like window, still high but slightly reduced for super-Earths, and low again for giants that are expected to host thick envelopes rather than rocky surfaces [20, 12, 13, 16]:

$$P(\text{retained} \mid \text{low}, \text{earth}, \text{super}, \text{giant}) = (0.15,\ 0.85,\ 0.75,\ 0.30) \quad \textbf{[20, 12, 13, 16]} \tag{7}$$

Climate stability given eccentricity declines from nearly circular to highly eccentric orbits [21, 4]:

$$P(\text{stable} \mid \text{low}, \text{mid}, \text{high } e) = (0.90,\ 0.60,\ 0.20) \quad \textbf{[21, 4]} \tag{8}$$

The probability of high atmospheric loss rises with activity [15, 11, 18, 19]:

$$P(\text{high loss} \mid \text{low}, \text{mid}, \text{high } A) = (0.15,\ 0.45,\ 0.80) \quad \textbf{[15, 11, 18, 19]} \tag{9}$$

A suitable temperature class is likely only in the temperate insolation bin [8, 9, 10]:

$$P(\text{suitable } T \mid \text{cold}, \text{temperate}, \text{hot}) = (0.25,\ 0.85,\ 0.10) \quad \textbf{[8, 9, 10]} \tag{10}$$

The outcome CPT encodes a logical conjunction with graded degradation. A planet without a retained atmosphere is almost never scored habitable (base rate 0.03-0.06). A planet with a retained atmosphere but an unsuitable temperature is scored at 0.10. Only the retained-and-temperate slice is allowed to reach high probability, and even there climate instability or high erosion reduces the rate.

Retained	Stable climate	High loss	Suitable T	P(habitable)
no	any	any	yes	0.06
no	any	any	no	0.03
yes	any	any	no	0.10
yes	yes	no	yes	0.93
yes	yes	yes	yes	0.45
yes	no	no	yes	0.50
yes	no	yes	yes	0.18

Table 3. Outcome CPT P(Habitable = 1 | latent states). Habitability requires both a retained atmosphere and a suitable temperature. Stability and low erosion then raise the probability.

5.4 Exact posterior inference

Given a fully observed planet (M, e, A, S), the four observables collapse to four bins (m, epsilon, a, s). The four latents R, C, L and T are then independent given those bins, and the posterior is the finite sum [24]

$$P(H = 1 \mid M, e, A, S) = \sum_{r,c,l,t} P(H = 1 \mid r,c,l,t)\, P(r \mid m)\, P(c \mid \varepsilon)\, P(l \mid a)\, P(t \mid s) \quad \textbf{[24]} \qquad \textbf{(11)}$$

The sum has 2^4 = 16 terms and is evaluated by enumeration. There is no sampling error in the network itself. Because the parents are discrete, the posterior takes only finitely many values, specifically 88 distinct values on the 2795-planet sample. The banding visible in the calibration and contour plots is therefore the true behaviour of the discrete model, not a plotting artefact.

5.5 Missing data

When an observable is missing, the corresponding bin is replaced by a prior over bins and the same sum is taken. For a missing insolation [24],

$$P(H \mid M, e, A) = \sum_{s} P(H \mid M, e, A, s)\, P(s) \quad \textbf{[24]} \qquad \textbf{(12)}$$

with P(s) = (0.20, 0.35, 0.45) on (cold, temperate, hot), reflecting the close-in bias of the detected population. Analogous priors exist for mass, eccentricity and activity. A candidate is therefore never discarded because a single column is empty. The missingness appears as a wider predictive distribution. That property is the operational reason for preferring a Bayesian network over a complete-case index when the catalogue is incomplete.

5.6 Monte Carlo error propagation

Measurement error on the observables is a different problem from missingness. For a planet with reported uncertainties (sigma_M, sigma_e, sigma_A, sigma_S) we draw n = 4000 Gaussian realisations, clip them to the physical domain, and evaluate the network on each draw. The resulting sample of P(habitable) supplies a median and a 16-84 per cent interval. Because the network is a step function of the bins, the interval is often asymmetric: a planet sitting near a bin edge can jump to a neighbouring CPT row under a 1-sigma perturbation, which the Monte Carlo envelope captures and a delta-method error bar would not.

6. Results

All numerical results in this section refer to the N = 2795 archive-like realisation scored with the default HPI weights and the prior CPTs of Section 5. The scored table is released with the accompanying code.

6.1 Distribution of the deterministic index

HPI has a median of 0.53 and a mean of 0.53. Only 7.5 per cent of the sample exceeds 0.80 and only 2.7 per cent exceeds 0.90. The high-HPI tail is sparse because the four constraints are rarely satisfied together. An Earth-mass, low-eccentricity, quiet-star, temperate-insolation planet is not the typical archive object. The low tail is populated by giants, highly eccentric orbits, and planets under strong activity or runaway-greenhouse flux. Mass and activity shape the bulk of the histogram. Eccentricity and insolation sculpt the tails.

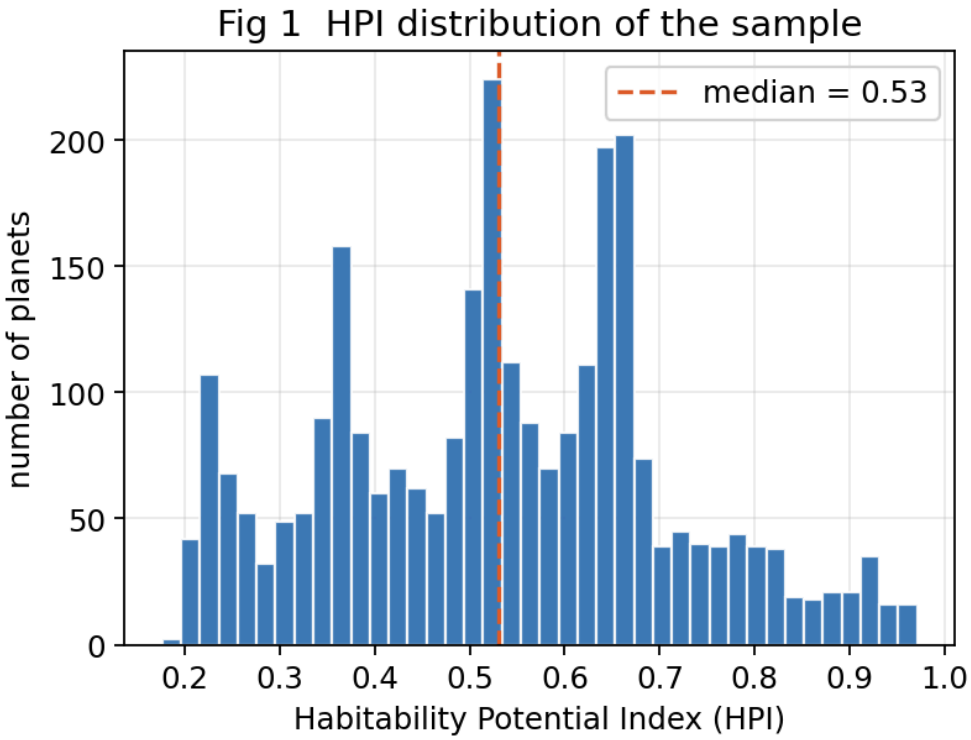


Figure 7. Distribution of the four-factor HPI on the 2795-planet sample. The high-habitability tail is sparse. The median is 0.53.

6.2 Posterior probabilities

The network posterior has a median of 0.12, a mean of 0.14 and a hard ceiling of 0.61 on this sample, the value attained when every latent CPT is in its most favourable bin (Earth-like mass, low e, low A, temperate S). Ninety planets reach P >= 0.50 and 129 reach P >= 0.40. The posterior is therefore more conservative than HPI. A high index is necessary but not sufficient for a high probability, because a single adverse latent state multiplies the outcome rather than subtracting a linear penalty.

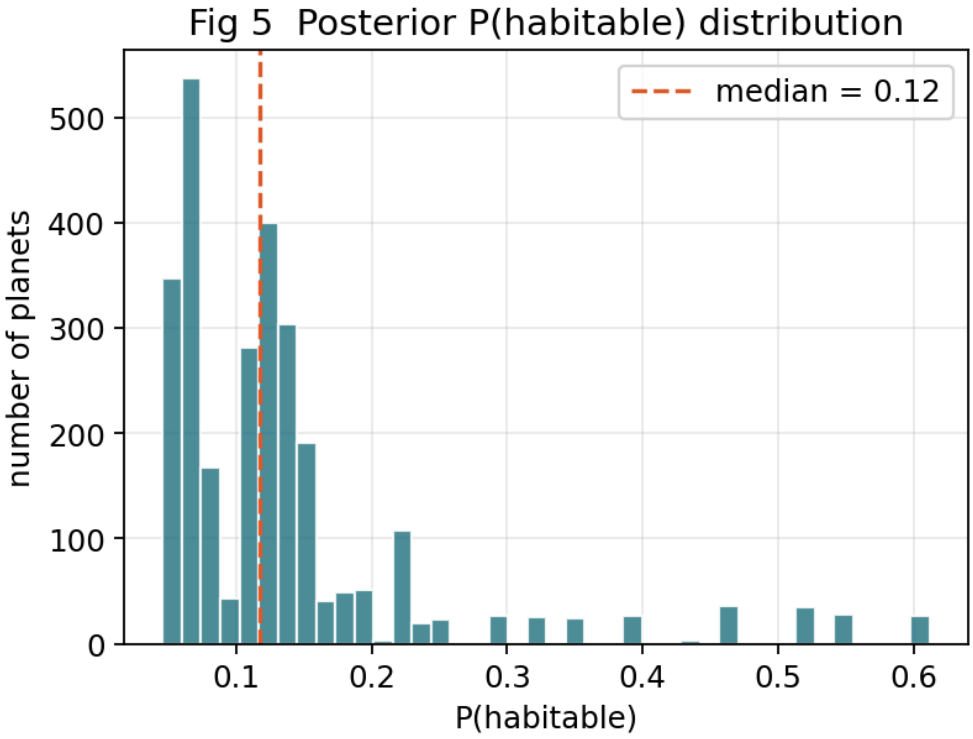


Figure 8. Distribution of P(habitable). Relative to HPI the mass of the distribution sits at lower values, and the support is discrete because the parents are binned.

6.3 Calibration of HPI against the network

HPI and P(habitable) are monotonically related (Pearson r = 0.58) but not interchangeable. In the mid-to-low HPI range the posterior falls faster than the index. In the high-HPI range the posterior saturates rather than climbing linearly toward 1. That saturation is a risk-compression effect. A planet that is excellent in three factors and poor in one remains linearly respectable on HPI, but the network multiplies through the poor latent state and does not assign a high probability. On a catalogue with heterogeneous errors this conservatism reduces the chance that a spuriously high score is promoted into a follow-up queue.

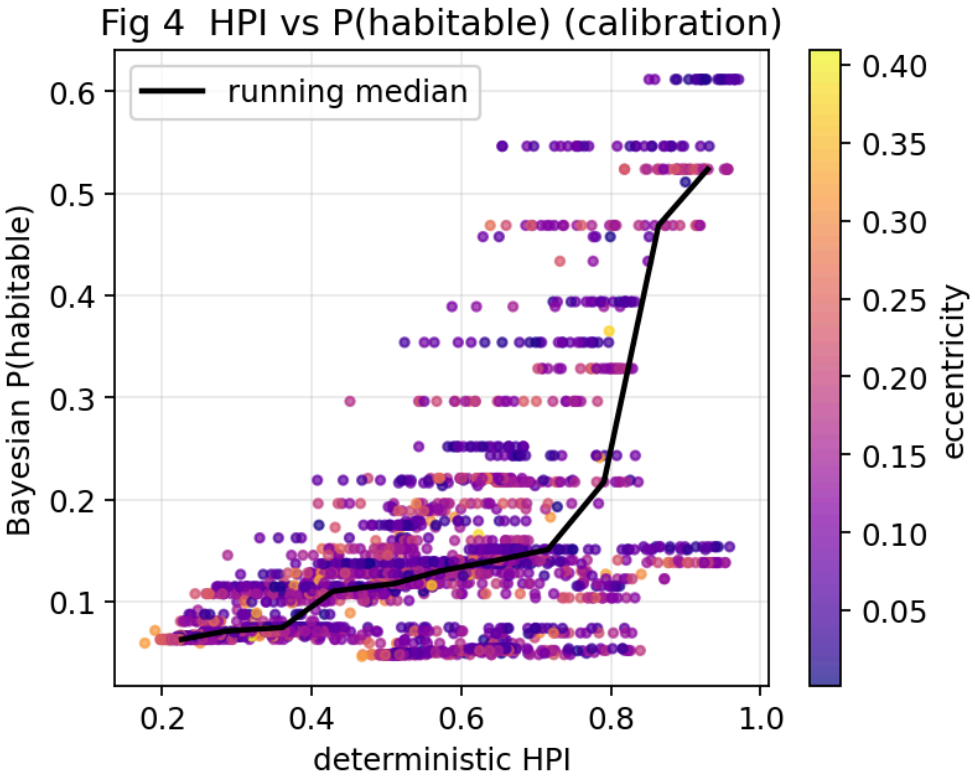


Figure 9. Calibration of HPI against P(habitable). The relation is monotone but compressed at both ends. Discrete banding is the signature of the binned network.

6.4 Joint structure of the observables

The four-factor corner plot of the scored sample confirms that high-HPI and high-P planets occupy the intersection of Earth-like mass, low eccentricity, low activity and near-unity insolation. There is no hidden linear correlation that would have allowed a cheaper two-parameter cut to recover the same ranking.

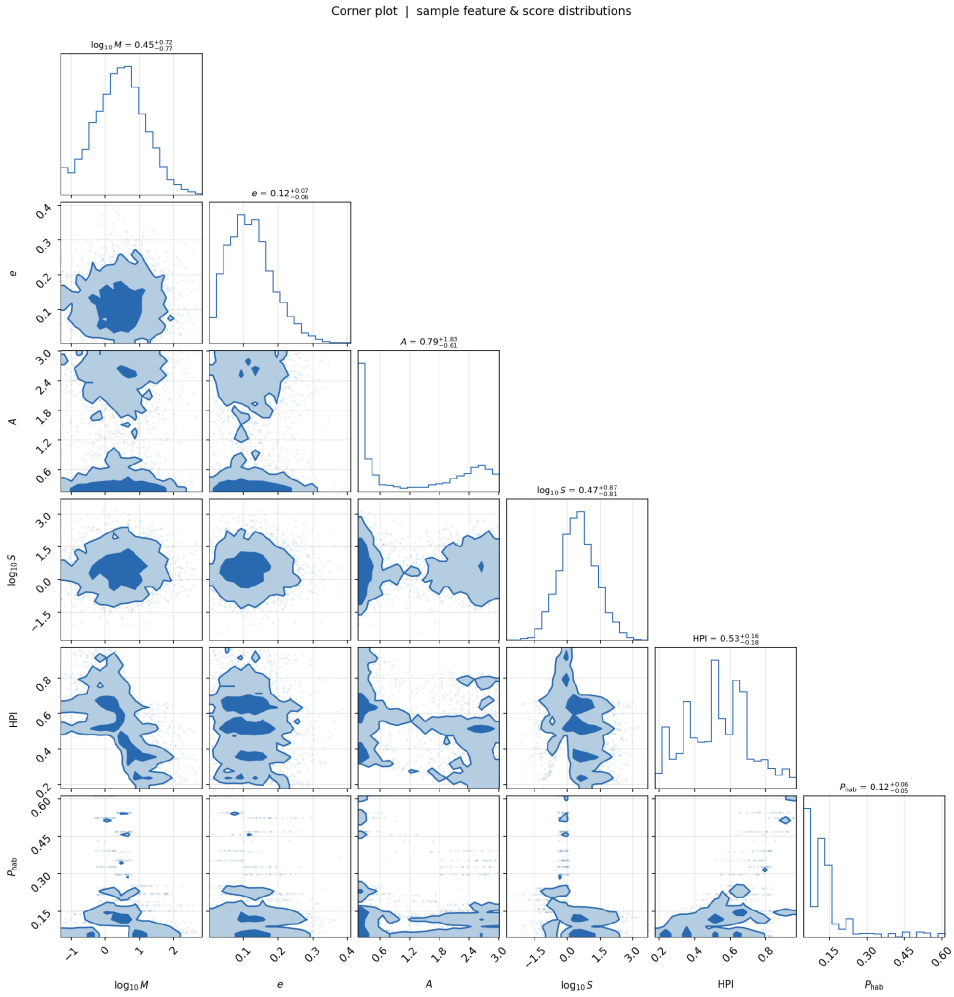


Figure 10. Joint distribution of log M, e, A, log S, HPI and P(habitable). High scores concentrate where all four observables sit in their favourable windows.

6.5 Ranked candidates

Sorting by P(habitable) and breaking ties on HPI produces a top-25 list whose members share three traits: mass in the Earth to low-mass super-Earth window, eccentricity below approximately 0.09, and a quiet-star, temperate-insolation environment. The list is not identical to a classical HZ catalogue. A planet inside the formal HZ but on a highly eccentric orbit, or around a flare-dominated M dwarf, is down-ranked. A planet just outside the conservative HZ but with an Earth-like mass, a circular orbit and a quiet host can survive in the tail. That multi-factor ranking behaviour cannot be produced by a binary HZ cut.

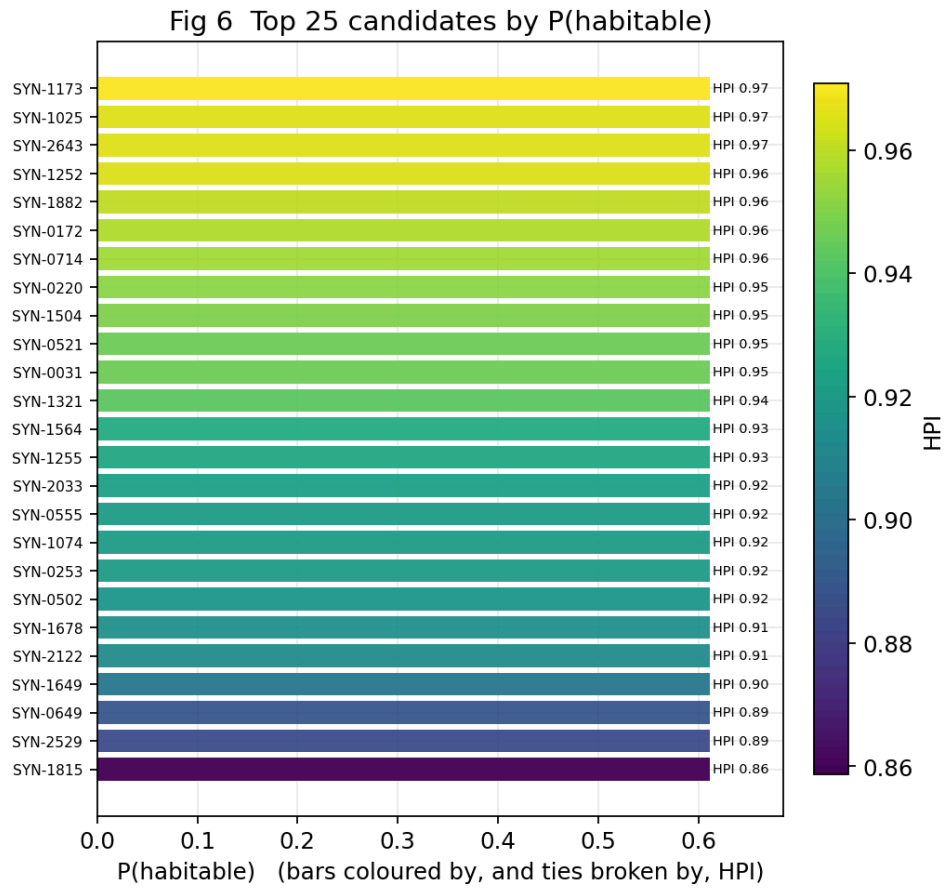


Figure 11. Top 25 candidates ranked by P(habitable), with ties broken by HPI. The high-probability objects are Earth-mass, low-eccentricity, temperate-insolation planets around quiet stars.

6.6 Maps in the (M, e) and (M, S) planes

Holding activity at a quiet-star value (A = 0.3) and insolation at 1 S_earth, P(habitable) in the mass-eccentricity plane peaks in the Earth-mass, low-e corner and falls first with mass, both toward the escape regime and toward the giant regime, and then with eccentricity. Holding eccentricity circular and activity quiet, the mass-insolation map recovers a habitable-zone strip around S approximately 1 that is truncated at low mass and at high mass. The hot edge is sharper than the cold edge, by construction of both f_S and the temperature-class CPT. These two maps are the probabilistic analogue of the classical HZ diagram, with mass retained as an explicit axis rather than absorbed into an Earth-twin assumption.

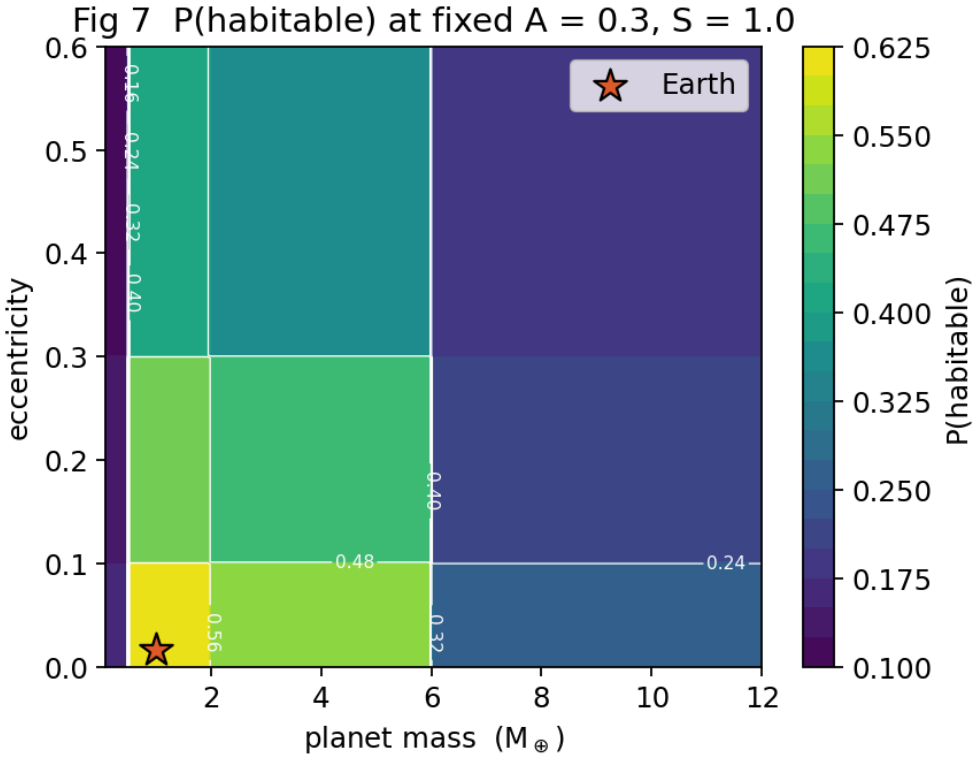


Figure 12. P(habitable) in the (mass, eccentricity) plane at fixed quiet-star activity and Earth-like insolation. The high-probability island is the Earth-mass, low-eccentricity corner.

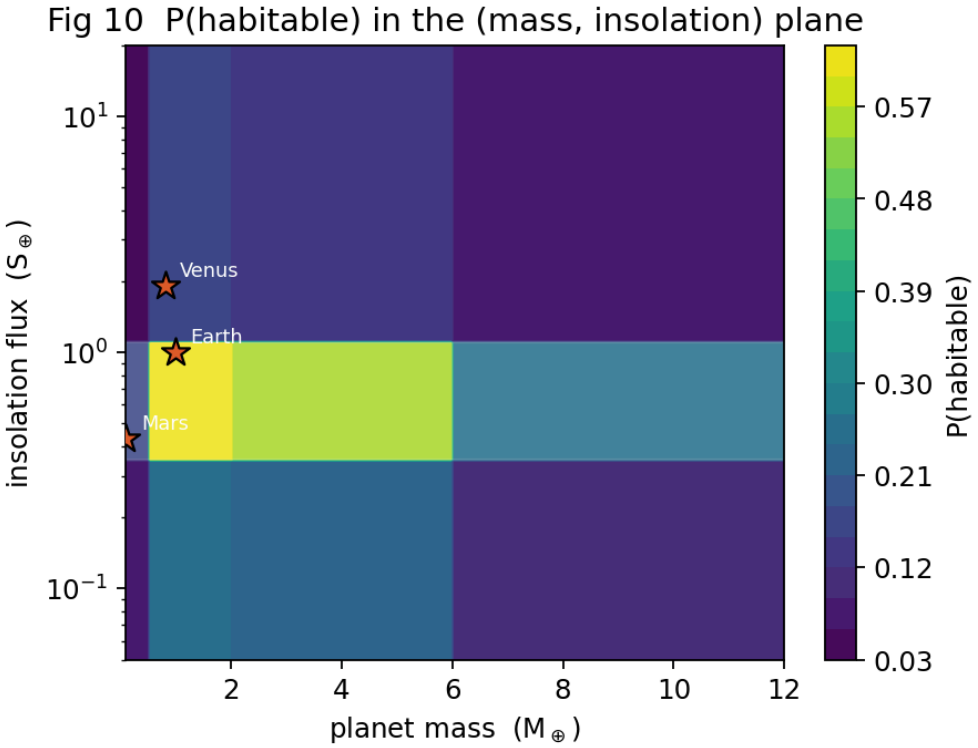


Figure 13. P(habitable) in the (mass, insolation) plane at fixed low eccentricity and quiet-star activity. The temperate strip around S_earth = 1 is the network habitable zone. The hot edge is a runaway-greenhouse cliff.

6.7 Uncertainty envelopes

Propagating reported or assumed observational errors through the network produces 16-84 per cent intervals that widen near bin boundaries and for planets whose mass or insolation sits close to a mechanism change. The interval is a property of the planet, not of the sample, and is the quantity that should accompany any ranked list released to observers.

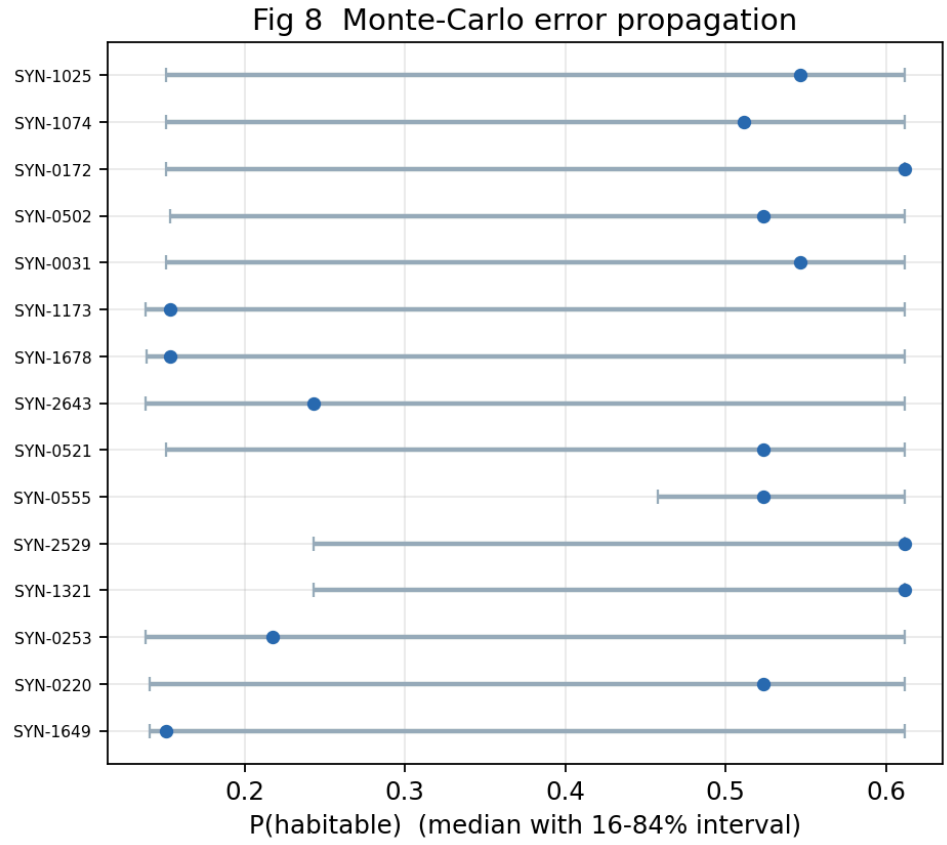


Figure 14. Monte Carlo propagation of observational errors. Points are posterior medians. Bars are 16-84 per cent intervals. Planets near a bin edge have the widest envelopes.

6.8 Solar System consistency check

A habitability model that cannot recover the Solar System is not ready to rank other systems. Scoring six reference bodies with the same network and the modern-Sun activity proxy yields the values in Table 4.

Body	M / M_earth	e	S / S_earth	HPI	P(habitable)
Mercury	0.055	0.206	6.67	0.63	0.052
Venus	0.815	0.007	1.91	0.71	0.151
Earth	1.000	0.017	1.00	0.96	0.612
Mars	0.107	0.093	0.43	0.87	0.154
Jupiter	317.8	0.049	0.037	0.37	0.110
Neptune	17.1	0.011	0.001	0.37	0.110

Table 4. Solar System reference scores. Earth is the unique maximum. Venus, which a three-factor index cannot separate from Earth, is suppressed by the insolation node.

Earth is the unique maximum, as required. Mercury fails on mass and insolation. Mars fails on mass and sits on the cold edge. Jupiter and Neptune fail on mass and insolation. Venus is the diagnostic case. Under a three-factor index its HPI is 0.958 against Earth's 0.959, a distinction too small to act on. With the insolation node included, Venus drops to P approximately 0.15 while Earth remains at P approximately 0.61. That separation is the operational reason for including the fourth parameter.

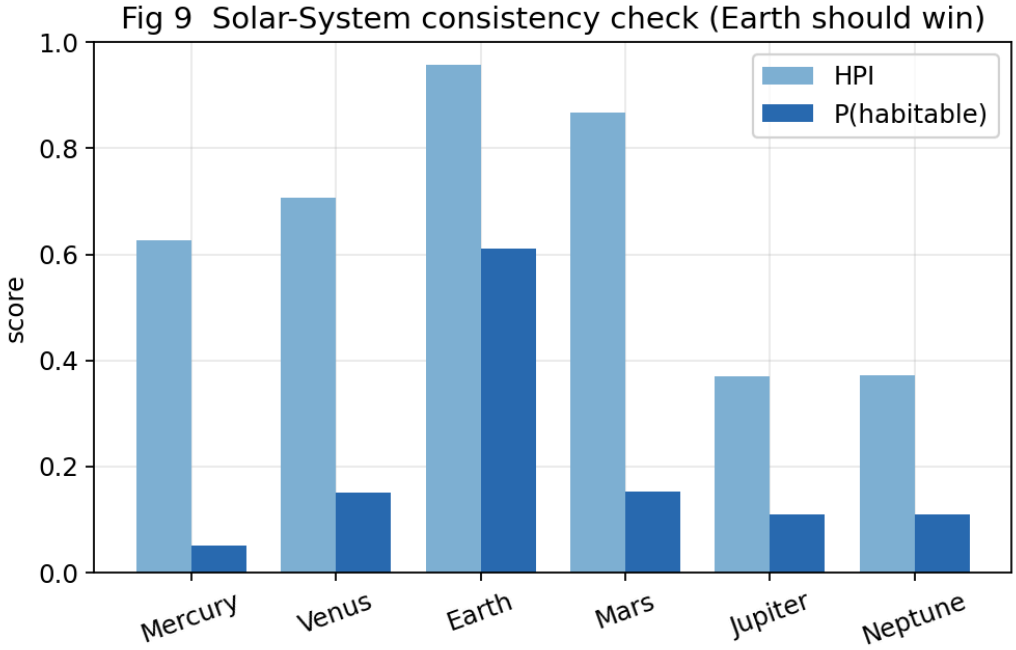


Figure 15. Solar System consistency check. Earth attains the highest posterior. Venus is suppressed by the insolation and temperature-class node.

6.9 Coupling between activity and mass

Repeating the (M, e) map at higher activity shifts the entire probability field downward, with the largest absolute losses among the lowest-mass planets, precisely the population whose escape velocities cannot compete with an enhanced XUV-driven wind. Eccentricity penalties also amplify under high activity, because an unstable climate and a high loss rate enter the outcome CPT as a product, not as a sum. Linear HPI cannot represent that interaction. The network does so automatically.

7. Discussion

Three structural conclusions follow from the sample and the diagnostics. First, the leading directions in the observable space are mass and the stellar and radiative environment. Eccentricity is a real but secondary modulator. On the PCA biplot the eccentricity arrow is nearly perpendicular to the HPI colour gradient, so this catalogue shows little linear association between eccentricity and the score. That is consistent with formation and dynamical statistics: most detected orbits are already close to circular, while masses and instellations span several decades. The result should be reported as a sample correlation, not as evidence that seasonal forcing can be ignored. A highly eccentric orbit still drives perihelion heating and aphelion freeze-out, and those cases remain down-weighted in both HPI and the network. Second, the random-forest ranking (Breiman 2001) agrees with the escape-velocity and XUV-erosion narrative, which is independent evidence that the index is not an artefact of the chosen weights. Mass leads because planets well above a few Earth masses in the archive are typically gaseous, with thick volatile envelopes rather than a solid temperate surface; the negative mass-score trend is that composition trend, not a proof that a massive world cannot host life. Stellar activity enters with the opposite sign to habitability because a high XUV and flare environment can strip an atmosphere and desiccate a water inventory. Insolation, which was omitted from the two-

dimensional biplot, remains essential in the score: it is the mean flux that sets the temperature class and is the reason Venus is separated from Earth. Third, the probabilistic layer is not a cosmetic rescaling of HPI. It compresses risk, widens under missingness, and jumps at physically motivated bin edges. A further limitation of the two-dimensional PCA display should be kept in view. Even in the three-variable rotation the first two components hold only 67.8 per cent of the variance. When insolation is added, the first two components hold only about 51 per cent. The plane shown in Figure 4 is therefore a partial view of the catalogue, useful for reading directions, not a complete map of the sample.

The framework has clear limits. The activity index is a Teff proxy, not a measured log R'_HK or L_X / L_bol. Young, active G stars and old, quiet M dwarfs will be misranked until a direct activity column is attached. Insolation in the archive is itself model-dependent. The outcome CPT is a prior, not a posterior calibrated on known habitable worlds. Earth is the only positive control, and using it both to set and to test the CPT would be circular. Rotation, tidal locking, magnetic field, volatile inventory and spectral energy distribution are all omitted. The discrete bins produce a stepwise posterior. A future version can replace the bins with calibrated logistic conditionals if a smoother response is required.

Those limits also define the extension path. The graph can accept additional nodes, including metallicity, a tidal-locking flag and stellar age, without rewriting the existing CPTs. The existing CPTs can be updated by hierarchical Bayes once a larger set of atmospheric detections exists. The same code path already accepts the live NASA TAP table, so the ranked list can be regenerated whenever the archive changes.

8. Conclusions

This paper replaces a binary habitable-zone cut with a four-factor physical index and a discrete Bayesian network. The index uses mass for atmospheric retention, eccentricity for climate stability, stellar activity for erosion and insolation for temperature class. It is reproducible and inspectable. Principal-component analysis shows that the four observables are not redundant. A random forest shows that each of them moves the score. The network turns the same physics into a posterior P(habitable | data), handles missing columns by marginalisation, and propagates measurement error by Monte Carlo sampling.

On an archive-sized sample the high-habitability tail is sparse, the posterior is more conservative than the linear index, and the Solar System is recovered with Earth unique at the top and Venus correctly suppressed by the insolation node. The intended use is not to declare a planet habitable, but to provide survey programmes with a ranked, uncertainty-aware shortlist that can be recomputed as the catalogue grows.

Data Availability

The accompanying repository contains the model, the catalogue loader, the principal-component utilities, the figure scripts, a one-command reproduction pipeline, the scored table and every figure used in this paper. The synthetic realisation used here is regenerated by a fixed random seed. The loader can query the NASA Exoplanet Archive TAP service and applies the same de-duplication and quality cut described in Section 2, falling back to the synthetic catalogue if the service is unreachable.

References

[1] Batalha, N. M., Rowe, J. F., Bryson, S. T., et al. 2013, Planetary Candidates Observed by Kepler. III. Analysis of the First 16 Months of Data, ApJS, 204, 24.

[2] Bean, J. L., Stevenson, K. B., & Batalha, N. E. 2017, The Promise and Challenge of Exoplanet Atmospheric Retrievals, AJ, 153, 64.

[3] Borucki, W. J., Koch, D., Basri, G., et al. 2010, Kepler Planet-Detection Mission: Introduction and First Results, Science, 327, 977.

[4] Dressing, C. D., Spiegel, D. S., Scharf, C. A., et al. 2010, Habitable Climates on Eccentric Earth-like Planets, ApJ, 721, 1295.

[5] Howard, A. W., Marcy, G. W., Bryson, S. T., et al. 2012, Planet Occurrence within 0.25 AU of Solar-type Stars from Kepler, ApJS, 201, 15.

[6] Kaltenegger, L. 2017, How to Characterize Habitable Worlds and Signs of Life, ARA&A, 55, 433.

[7] Kane, S. R., Hill, M. L., Kasting, J. F., et al. 2016, Catalog of Kepler Habitable Zone Exoplanet Candidates, ApJ, 830, 1.

[8] Kasting, J. F., Whitmire, D. P., & Reynolds, R. T. 1993, Habitable Zones around Main Sequence Stars, Icarus, 101, 108.

[9] Kopparapu, R. K., Ramirez, R., Kasting, J. F., et al. 2013, Habitable Zones around Main-sequence Stars: New Estimates, ApJ, 765, 131.

[10] Kopparapu, R. K., Ramirez, R., SchottelKotte, J., et al. 2014, Habitable Zones around Main-sequence Stars: Dependence on Planetary Mass, ApJ, 787, L29.

[11] Lammer, H., Lichtenegger, H. I. M., Kulikov, Y. N., et al. 2007, Coronal Mass Ejection Effects on Close-in Exoplanets, Astrobiology, 7, 185.

[12] Lammer, H., Bredehoft, J. H., Coustenis, A., et al. 2009, What Makes a Planet Habitable?, A&A Rev., 17, 181.

[13] Owen, J. E., & Wu, Y. 2017, The Evaporation Valley in the Kepler Planets, ApJ, 847, 29.

[14] Petigura, E. A., Howard, A. W., & Marcy, G. W. 2013, Prevalence of Earth-size Planets Orbiting Sun-like Stars, PNAS, 110, 19273.

[15] Ribas, I., Guinan, E. F., Gudel, M., & Audard, M. 2005, Evolution of the Solar Activity over Time and Effects on Planetary Atmospheres, ApJ, 622, 680.

[16] Rogers, L. A. 2015, Most 1.6 Earth-radius Planets are not Rocky, ApJ, 801, 41.

[17] Seager, S. 2013, Exoplanet Habitability, Science, 340, 577.

[18] Segura, A., Walkowicz, L. M., Meadows, V., et al. 2010, The Effect of a Strong Stellar Flare on the Atmospheric Chemistry of an Earth-like Planet Orbiting an M Dwarf, Astrobiology, 10, 751.

[19] Tilley, M. A., Segura, A., Meadows, V., et al. 2019, Modeling Repeated M-dwarf Flares and Their Influence on Planetary Habitability, Astrobiology, 19, 64.

[20] Watson, A. J., Donahue, T. M., & Walker, J. C. G. 1981, The Dynamics of a Rapidly Escaping Atmosphere: Applications to the Evolution of Earth and Venus, Icarus, 48, 150.

[21] Williams, D. M., & Pollard, D. 2002, Earth-like Worlds on Eccentric Orbits: Excursions Beyond the Habitable Zone, IJAsB, 1, 61.

[22] Breiman, L. 2001, Random Forests, Machine Learning, 45, 5.

[23] Hotelling, H. 1933, Analysis of a Complex of Statistical Variables into Principal Components, Journal of Educational Psychology, 24, 417.

[24] Pearl, J. 1988, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (San Mateo, CA: Morgan Kaufmann).

[25] Pearson, K. 1901, On Lines and Planes of Closest Fit to Systems of Points in Space, Philosophical Magazine, 2, 559.

[26] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825.