Bioverse: a simulation framework to assess the statistical power of future biosignature surveys

ALEX BIXEL 1,2 and DÁNIEL APAI 1,2,3

1 Department of Astronomy/Steward Observatory, The University of Arizona, 933 N. Cherry Avenue, Tucson, AZ 85721, USA
2 Earths in Other Solar Systems Team, NASA Nexus for Exoplanet System Science
3 Lunar and Planetary Laboratory, The University of Arizona, 1629 E. University Blvd, AZ 85721, USA

(Accepted January 22, 2021)
Submitted to AJ

ABSTRACT

Next-generation space observatories will conduct the first systematic surveys of terrestrial exoplanet atmospheres and search for evidence of life beyond Earth. While in-depth observations of the nearest habitable worlds may yield enticing results, there are fundamental questions about planetary habitability and evolution which can only be answered through population-level studies of dozens to hundreds of terrestrial planets. To determine the requirements for next-generation observatories to address these questions, we have developed Bioverse. Bioverse combines existing knowledge of exoplanet statistics with a survey simulation and hypothesis testing framework to determine whether proposed space-based direct imaging and transit spectroscopy surveys will be capable of detecting various hypothetical statistical relationships between the properties of terrestrial exoplanets. Following a description of the code, we apply Bioverse to determine whether an ambitious direct imaging or transit survey would be able to determine the extent of the circumstellar habitable zone and study the evolution of Earth-like planets. Given recent evidence that Earth-sized habitable zone planets are likely much rarer than previously believed (Pascucci et al. 2019), we find that space missions with large search volumes will be necessary to study the population of terrestrial and habitable worlds. Moving forward, Bioverse provides a methodology for performing trade studies of future observatory concepts to maximize their ability to address population-level questions, including and beyond the specific examples explored here.

1. INTRODUCTION

The field of exoplanet science stands at an exciting turning point. In the past, most exoplanet surveys aimed only to constrain bulk properties - such as size, period, and mass. Moving forward, several groups are developing concepts for space telescopes which would enable the atmospheric characterization of temperate terrestrial planets. Such concepts include the Large UV/Optical/Infrared Surveyor (LUVOIR, The LUVOIR Team 2019), the Habitable Exoplanet Observatory (HabEx, The HabEx Team 2019), the Origins Space Telescope (Origins Space Telescope Study Team 2019), the Nautilus Space Observatory (Apai et al. 2019), the Large Interferometer for Exoplanets (LIFE, Quanz et al. 2018), and the Mid-Infrared Exoplanet Climate Explorer (MIRECLE, Staguhn et al. 2019). By looking for biosignatures in the atmospheres of temperate Earth-sized planets, these observatories would conduct the first systematic search for life beyond the Solar System.

Next-generation observatories will be able to study some of the closest terrestrial exoplanets in unprecedented detail, but this is only the start of their scientific capability: observatories which can study tens to hundreds of terrestrial planets will allow for the first statistical constraints on the atmospheric, geological, and biological properties of terrestrial planets. Some recent works have explored statistical trends and patterns which may only be evident at the population level. For example, habitable zone models predict patterns in atmospheric CO$_2$ and H$_2$O abundance (Bean et al. 2017; Lehmer et al. 2020) as well as color and albedo across a range of stellar insolations (Checlair et al. 2019). Venus analogs may have larger apparent radii than their temperate siblings due to their thick, post-runaway greenhouse atmospheres (Turbet et al. 2019). Earth’s geological record suggests a possible relationship between the ages and oxygen content of Earth-like planets, assuming their atmospheres evolve similarly to Earth’s (Bixel & Apai 2020a), and with a large enough sample size of potentially habitable planets, next-generation surveys...
could place the first constraints on the frequency of life in the universe (Checlair et al. 2020). An understanding of population-level trends will provide context for the interpretation of possible biosignatures on individual worlds and could illuminate their potential false positive (i.e. non-biological) sources (Apai et al. 2017; Meadows et al. 2018). To avoid statistical false positive scenarios, efforts must also be made to understand which distinct mechanisms could produce the same apparent trends. For example, an increase in cloud deck altitude with insolation could masquerade as a signature of atmospheric erosion in a sample of transiting exoplanets (Lustig-Yaeger et al. 2019).

Recent research has identified key outstanding questions about terrestrial exoplanets, their planetary systems, and the processes which shape them for which future observatories might provide insights (see the SAG 15 report for an overview of such several questions in the context of direct imaging missions, Apai et al. 2017). For example: what are the processes which shape their atmospheric loss (e.g., Zahnle & Catling 2017)? Is the habitable zone wide (e.g., Kasting et al. 1993; Koppa-rapu et al. 2014) or narrow (e.g., Hart 1979)? What is the relationship between planet size and tectonic activity (e.g., Valencia et al. 2007; Dorn et al. 2018)? Are habitable planets equally common around stars of different mass and activity levels (e.g., Shields et al. 2016)? Which, if any, of these questions could be answered with a next-generation observatory will depend on its technical design and observing strategy. One important metric is the number of terrestrial habitable zone planets which it could realistically detect, but only a subset of these will be habitable, and even inhabited worlds may vary substantially from Earth in their atmospheric composition and evolutionary history. Furthermore, deep spectroscopic characterization of individual planets will be time-consuming, so strategic choices must be made as to which planets to characterize and at what wavelengths. For these reasons, analyses based solely on the detection yield predictions of future space mission concepts will provide an optimistic assessment of their statistical power.

To enable meaningful statistical hypotheses which can be tested by future observatories, we have developed Bioverse. Bioverse estimates the statistical power of next-generation exoplanet surveys to detect and study population-level trends by simulating the underlying planet population, survey limitations, observing biases, and statistical analyses which a future observer would perform on a large set of observations of terrestrial planets. After the following brief description of the code structure, we describe its three main components in Sections 3 through 5. In Sections 6 and 7, we use Bioverse to determine the requirements for next-generation surveys to test the habitable zone concept and study the evolution of Earth-like planets.

2. CODE OUTLINE

Bioverse consists of three components, outlined in Figure 1. The first component generates planetary systems with bulk properties (e.g., size and period) drawn from statistical distributions, then applies theoretical models or parametric relationships to generate secondary properties of interest (e.g., atmospheric composition). The second component is a survey simulator which conducts observations of the simulated exoplanet population in direct imaging or transit spectroscopy mode. The survey simulator first determines which planets could be characterized within a finite allotted observing time, then generates a simulated data set representative of the telescope and instrument capabilities. The third component is a Bayesian framework which uses simulated datasets to test statistical hypotheses and estimate model parameters. By iterating through these components, we can use Bioverse to determine the statistical power of a proposed observatory to test different hypotheses.

Bioverse is written in Python and designed for flexibility, so that different statistical assumptions and testable hypotheses can be implemented in the future. The specific set of assumptions which Bioverse is currently based on are listed in Table 1. Given the large number of parameters involved in Bioverse, we provide a table of abbreviations and symbols used in the text in Appendix A.

3. PLANET GENERATION

The first component of Bioverse creates simulated planetary systems around host stars in the solar neighborhood with a period and radius distribution informed by Kepler statistics. Other planet properties (such as mass and geometric albedo) are derived from empirical relationships or best-guess prior distributions. Finally, the simulated planet properties reflect the effects of hypothetical population-level trends which could be uncovered by a future survey of terrestrial planets.

3.1. Stellar properties

We begin by considering which stars in the solar neighborhood would be targeted by future biosignature surveys. Our strategy for simulating stellar systems is

---

1 A current version of the code can be found on GitHub, while the version used in this paper is archived on Zenodo (Bixel & Apai 2021).
A high-level outline of the Bioverse code. In this paper, we apply Bioverse to assess the detectability of two hypothetical population-level trends (green) with next-generation survey telescopes. These relationships are injected into the simulated planet population by the first module, then tested as statistical hypotheses by the third module.

Direct imaging surveys will primarily target the habitable zones of higher-mass (FGK) stars within the nearest 30 pc, the majority of which have already been cataloged by space-based astrometry missions. Not all of these will be equally valid targets, due to the combined effects of distance and background noise sources, such as zodiacal dust (Stark et al. 2019). Sophisticated simulations for the LUVOIR mission concept (The LUVOIR Team 2019) have produced an optimized list of targets whose habitable zones could feasibly be probed for Earth-like planets. In imaging mode, we use an optimized stellar target list for the 15-meter LUVOIR-A concept as the basis for simulating nearby planetary systems (C. Stark, private correspondence).

A survey of transiting habitable zone planets would be most sensitive to planets around low-mass (K and M) stars, as their habitable zone planets are more likely to transit, transit more frequently, and produce a deeper relative transit depth. However, the census of low-mass stars is not complete out to ~ 100 pc. Therefore, in transit mode, all stellar masses are randomly drawn from a present-day stellar mass function (Chabrier 2003) and distribute them isotropically in space. We do not include any known stars or transiting planets in the transit mode sample; as most nearby transiting planets remain undiscovered, this would have little effect on the overall statistical distribution of host star properties.

In both imaging and transit modes, we relate the stellar mass ($M_*$) to its radius, luminosity, and effective temperature ($R_*, L_*, T_*$) by interpolating a list of these properties as a function of spectral type (Pecaut & Mamajek 2013). Each star is assigned an age drawn uniformly from 0–10 Gyr, reflecting the (to first order) constant star formation rate in the Milky Way for the past
Table 1. Summary of statistical assumptions and modeling choices in Bioverse, with associated references.

| Topic                                      | Assumptions                                                                                                                                  | References                                                                 |
|--------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| Host star distribution and properties      | (Imaging mode) LUVOIR-A optimized target catalog                                                                                               | The LUVOIR Team (2019) and C. Stark (private correspondence)                    |
|                                            | (Transit mode) Stellar mass function                                                                                                          | Chabrier (2003)                                                               |
|                                            | Main sequence mass-radius-luminosity relations                                                                                               | Pecaut & Mamajek (2013)                                                       |
| Planet occurrence rates                    | SAG13 occurrence rates, with modifications:                                                                                                  | Pascucci et al. (2019); Neil & Rogers (2020)                                  |
|                                            | - $\eta_\oplus \approx 7.5\%$ for G stars (down from $\approx 24\%$)                                                                         | Mulders et al. (2015a,b)                                                      |
|                                            | - More planets around lower-mass stars                                                                                                         |                                                                               |
| Exo-Earth candidates                       | approximately Earth-sized ($0.85^{0.25} < R < 1.4R_\oplus$) within the circumstellar habitable zone                                          | various (see Section 3.4)                                                     |
|                                            | K14                                                                                                                                           |                                                                               |
| Observatory templates                      | (Imaging mode) 15-meter LUVOIR-A observatory                                                                                                  | The LUVOIR Team (2019)                                                       |
|                                            | (Transit mode) 50-meter equivalent Nautilus Space Observatory                                                                              | Apai et al. (2019a)                                                          |
| Target prioritization                      | Finite observing time with overheads                                                                                                          |                                                                               |
|                                            | Observe in order of required time                                                                                                            |                                                                               |
|                                            | Prioritize targets to reduce survey biases                                                                                                    |                                                                               |
| Measurement noise                          | Photon-noise limited observations with characteristic wavelength $\lambda_{\text{eff}}$                                                       |                                                                               |
|                                            | Required exposure time scales with distance, stellar brightness, and signal strength                                                          |                                                                               |
| Model comparison                           | Compare alternative to null hypothesis through Bayesian evidence $Z$                                                                           |                                                                               |
|                                            | Significant evidence when $\Delta(Z) > 3$                                                                                                       |                                                                               |
|                                            | (where applicable) Frequentist comparison tests (e.g., t-tests)                                                                               |                                                                               |

3.2. Period and radius occurrence rates

Kepler has provided excellent insights into the frequency of planets as a function of period and size for a wide range of host stars. However, these statistics are only complete to periods $\lesssim 100$ days, and as such do not reach the habitable zone of Sun-like stars. As a result, estimates of $\eta_\oplus$ (the average number of habitable zone Earth-sized planets per star) have so far been based on extrapolation and are therefore model-dependent. NASA’s Exoplanet Program Analysis Group 13 (hereafter SAG13) to consolidate the results of several studies of Kepler occurrence rates into a single set of estimates for community use, resulting in an oft-cited value of $\eta_\oplus \approx 24\%$ for G stars. Here, and elsewhere in this paper, the value of $\eta_\oplus$ uses the habitable zone model of Kopparapu et al. (2014) (hereafter K14: 0.95 – 1.67 AU for an Earth twin). We use the SAG-13 consensus occurrence rate power laws as the basis for determining the number, radii ($R_p$), periods ($P$), semi-major axes ($a$), and insolations ($S$) of planets.

10 Gyr (e.g., Snaith et al. 2015; Fantin et al. 2019; Mor et al. 2019).

2 see this URL as well as Kopparapu et al. (2018)
in each system. However, the SAG13 metastudy was based largely on studies published before 2017, many of which did not assess planet occurrence as a function of stellar mass. We make the following two modifications to the SAG13 rates to reflect recent work.

First, we unilaterally decrease the number of planets per star by a factor of 3.2, such that $\eta_\oplus \approx 7.5\%$ for G stars. This is in response to the findings of Pascucci et al. (2019) that Earth-sized planets are more common at shorter orbital periods ($P \lesssim 25$ d) than in the habitable zone, which they ascribe to the effects of photoevaporation. Specifically, they argue that a large fraction of Earth-sized planets on close-in orbits are the evaporated cores of ice giants - planets which maintain their envelopes (and are therefore not Earth-like) if they form in the habitable zone. In another analysis, Neil & Rogers (2020) find evidence for two distinct populations of rocky planets, and as a result fewer Earth-sized planets in the habitable zone, for which they suggest a similar explanation. The chosen value of 7.5% is in the mid-range of values estimated by Pascucci et al. (2019) when they exclude the planets most affected by photoevaporation.

Second, we modulate the occurrence rates as a function of spectral type following Mulders et al. (2015a), who find that rocky planets are more common around lower-mass stars and tend to occupy shorter orbits. Specifically, we gradually increase the number of planets for stars less massive than the Sun and decrease their semi-major axes by interpolating between the scaling factors provided by Mulders et al. (2015a)$^3$ (normalized to 1 for the typical Kepler host star). Later, Mulders et al. (2015b) found evidence that the number of rocky planets around the typical Kepler M dwarf (M0 – M5) was $\sim 3.5 \times$ as high as for G dwarfs, so we further increase the number of planets around M dwarfs to reflect this result. Finally, since Kepler was not sensitive to late M dwarfs, we assume the number of planets per star to plateau for these stars (which we believe to be a conservative extrapolation given the general trend).

We note that more recent studies of M dwarf planet occurrence rates reaffirm the finding that lower-mass stars have more Earth-sized planets, including estimates from Kepler data (e.g., Hardegree-Ullman et al. 2019; Hsu et al. 2020) and radial velocity detections (Tuomi et al. 2019).

The net impact of these two decisions on the number of Earth-sized planets per star, as a function of stellar mass, is shown in Figure 2. Our estimate of $\eta_\oplus$ may seem pessimistic when compared to higher values used in predicting the detection yield of mission concepts (e.g., The HabEx Team 2019; The LUVOIR Team 2019; Origins Space Telescope Study Team 2019; Apai et al. 2019), but we view it to be a realistic estimate based on the most recent studies available. Our estimate is lower than those of Bryson et al. (2021), who avoid bias due to photoevaporation by excluding planets at high insolation. However, their resulting sample size is limited, and thus their confidence intervals are broad; indeed, our value of $\eta_\oplus = 7.5\%$ is within the 95% confidence interval of some of their estimates.$^4$

In general, all existing estimates of $\eta_\oplus$ for G stars - including our own - are based on extrapolation and are therefore uncertain. For example, existing data cannot rule out an increase in terrestrial planet occurrence rates at orbital periods beyond $\sim 100$ d, which would enhance $\eta_\oplus$. To accommodate this uncertainty, we express our results in Sections 6 and 7 in terms of either $\eta_\oplus$ or the number of planets observed (which is typically linear to

---

$^3$ see Table 1 and Figure 4 therein

$^4$ see Table 6 therein
As a result, the validity of our results is not tied to any specific value for $\eta_0$.

3.3. **Habitable zone boundaries**

The circumstellar habitable zone refers to the theoretical region around a star in which a planet can sustain liquid surface water. Many formulations of the habitable zone exist, but the most commonly cited estimates are based on Kasting et al. (1993) and subsequent papers which expanded on their methodology (Kopparapu et al. 2013, 2014). In Bioverse we use the results of K14 to calculate the inner edge ($a_{\text{inner}}$, corresponding to the runaway greenhouse limit) and outer edge ($a_{\text{outer}}$, corresponding to the maximum greenhouse limit) of the habitable zone. To account for the dependence on planetary mass, we interpolate between the three planetary masses modeled therein.

3.4. **Classification**

Following Kopparapu et al. (2018), we classify planets as “hot”, “warm”, or “cold” depending on their insolation, and “rocky”, “super-Earth”, “sub-Neptune”, “sub-Jovian”, or “Jovian” depending on their size. Approximately Earth-sized planets within the habitable zone are of particular interest, as these are the most likely planets to have liquid water and habitable surface conditions. Following recent studies of detection yield estimates for direct imaging missions (Kopparapu et al. 2018; Stark et al. 2019; The HabEx Team 2019; The LUVOIR Team 2019), we classify as “exo-Earth candidates” (hereafter EECs) any planets with radii $0.85 < R < 1.4$ and orbits within the habitable zone boundaries calculated above. The lower limit on the size of EECs is the theoretical minimum size for which a terrestrial planet can maintain an atmosphere suggested by Zahnle & Catling (2017), while the upper limit reflects the findings of several authors that planets larger than ~1.4–1.6$R_\oplus$ tend to resemble mini-Neptunes in composition more than super-Earths (e.g., Weiss & Marcy 2014; Rogers 2015; Fulton et al. 2017).

3.5. **Albedo and contrast ratio**

Imaging measurements will be able to use a planet’s brightness as a rough proxy for its size, but its brightness also depends on its geometric albedo, orbital phase, and semi-major axis. The latter two of these can feasibly be constrained by revisiting the system over several months, but it will be difficult to precisely disentangle geometric albedo and planet size. Albedo is highly sensitive to surface and atmospheric composition and will likely be highly variable for directly imaged exoplanets, so estimates of a planet’s size based on brightness alone will be highly uncertain (Guimond & Cowan 2018; Bixel & Apai 2020b; Carrión-González et al. 2020). To properly represent this source of uncertainty, we assign geometric albedos ($A_g$) to each planet ranging uniformly from 10 – 70% (approximately the range of values encountered at visible wavelengths for solar system planets, e.g., Madden & Kaltenegger 2018).

Next, we compute the planet-to-star brightness contrast ratio for each planet, modeling them as Lambertian spheres observed at quadrature phase (Traub & Oppenheimer 2010):

$$\zeta = \frac{A_g}{\pi} \left( \frac{R_p}{a} \right)^2$$

Note that the determination of a planet’s phase from imaging data is also not trivial, requiring multiple follow-up observations to establish the orbit. Nevertheless, such observations will be a likely component of any future imaging survey in order to distinguish temperate planets from their hotter and colder peers (The HabEx Team 2019; The LUVOIR Team 2019).

3.6. **Surface gravity and scale height**

To translate planet radii into masses, we use the probabilistic mass-radius relationship derived by Wolfgang et al. (2016), which separates terrestrial planets and ice giants. Given each planet’s mass and surface gravity, we then estimate the atmospheric scale height ($h$), which is important for determining the relative spectroscopic signal due to atmospheric absorption (as described in Section 4.3.2). We assign an atmospheric mean molecular weight $\mu$ to each planet based on its size. For “sub-Neptune” planets and larger, we assume $^3\text{He}$ dominated atmospheres similar to Neptune or Uranus, with $\mu = 2.5 m_\text{H}$. For “rocky” and “super-Earth” planets, we calculate the ratio of $N_2$ to $\text{CO}_2$ based on their position relative to the habitable zone as follows. For planets within $a_{\text{inner}}$, we assume $\text{CO}_2$ dominated atmospheres similar to Venus’ ($\mu = 44 m_\text{H}$). Within the habitable zone, we adopt a positive correlation between semi-major axis and $\text{CO}_2$ partial pressure, which climate models predict as a result of the carbonate-silicate negative feedback mechanism (e.g., Bean et al. 2017). Specifically, we follow the correlation derived by Lehmer et al. (2020)5, add $N_2$ as necessary to reach a minimum total pressure of 1 bar, and calculate the mean molecular weight between the two species ($28 < \mu < 44 m_\text{H}$). Finally, for planets beyond $a_{\text{outer}}$, we assume the $\text{CO}_2$ to condense, leaving behind a pure $N_2$ atmosphere ($\mu = 28 m_\text{H}$). We set the atmospheric mass, $\text{He}$.

5 We adopt the best-fit line in Figure 1 therein
temperature equal to the equilibrium temperature, assuming the Bond albedo to equal the geometric albedo. However, for EECs we assume an Earth-like atmospheric temperature due to greenhouse warming.

3.7. Inclination and transiting planets

Planets are assigned inclinations \((i)\) from an isotropic distribution (i.e. a uniform distribution in \(\cos(i)\) from \(-1\) to \(1\)). From this, and assuming circular orbits, we calculate the impact parameter on the stellar surface:

\[
b = \frac{a \cos(i)}{R_*} \quad (2)
\]

For transiting planets (with \(|b| < 1\)) we calculate the transit depth \((\delta = (R_p/R_*)^2)\) and duration:

\[
T_{\text{dur}} = \frac{R_p P}{\pi a} \sqrt{1 - b^2} \quad (3)
\]

3.8. Hypothetical population-level trends

The primary goal of our study is to understand which population-level trends may be detectable with a next-generation exoplanet survey. For example, could such a survey empirically determine the location of the habitable zone based on which planets have \(\text{H}_2\text{O}\)-rich atmospheres (Section 6), or study how oxygen evolves over time in the atmospheres of Earth-like planets (Section 7)?

To enable these inquiries, we apply hypothetical population-level trends to the simulated planet sample which will later be studied by simulated direct imaging and transit surveys. Specifically, we determine which planets have atmospheric water vapor based on their size and semi-major axis (following Equation 13), and determine which Earth-like planets have atmospheric oxygen based on their age (following Equation 15). A more detailed description of these assumed trends, and an assessment of their detectability by future biosignature surveys, can be found in Sections 6 and 7.

4. SURVEY SIMULATION

The second component of Bioverse translates the simulated planet population from the previous section into a data set representing the result of a lengthy characterization effort with a next-generation observatory. There are a few methods by which future observatories could characterize statistically-relevant samples of habitable planets, but in Bioverse we focus on space-based direct imaging and transit spectroscopy. The data sets produced by these next-generation surveys will be inherently biased by the observing approach. Most notably, an imaging survey is most efficient in targeting the habitable zones of nearby FGK stars, while a transit survey is optimized for M stars. Strategic decisions also bias the data set - for example, an imaging survey must dedicate \(\sim 4\times\) as much time to study a planet at 2 AU from its star versus an Earth twin, so studying planets near the outer edge of the habitable zone will come at a steep cost.

4.1. Survey setup

As our template for a direct imaging survey we use LUVOIR (The LUVOIR Team 2019, hereafter L19), a proposed NASA Flagship-class mission which would use an 8–15 meter segmented mirror and a multi-channel coronagraphic instrument to study terrestrial planets around nearby stars. While the details of the LUVOIR concept have been studied in-depth, our results are based only on its high-level characteristics - specifically, we adopt the 15-meter LUVOIR-A mirror diameter, coronagraphic inner (IWA) and outer (OWA) working angles and noise floor, and the host star catalog used to simulate its detection yield estimates (C. Stark, private correspondence). Our results should be generally applicable to any imaging mission with a similar mirror size and coronagraph.

As our template for a transit survey, we use the Nautilus Space Observatory concept (Apai et al. (2019a); Apai et al. (2019c)), which aims to study transiting exoplanets with the equivalent light-collecting area of a single 50-meter diameter telescope. To achieve this light-collecting power, Nautilus would employ an array of large telescopes with ultralight diffractive-refractive optical elements (Milster et al. 2020) (the launch of a single, up to 8.5m diameter telescope has recently been proposed as a NASA Probe-class mission, Apai et al. (2019b)). To generate the potential list of transiting planets, we simulate systems to a distance of 150 parsecs, as our simulated surveys tend not to observe targets beyond this distance even when they are available.

Our analyses are based on a 15-meter mirror diameter imaging survey and a 50-meter diameter (equivalent area) transit survey, because among all concepts currently under consideration by the community, these are the ones purporting to offer the largest EEC sample sizes for their respective techniques. It should be noted that a 15-meter imaging survey would also be capable of characterizing nearby transiting planets as a secondary science goal, but we do not model any dual mode surveys here.

4.2. Which planets can be detected?

After simulating a catalog of nearby planetary systems, we discard any planets which cannot be detected by a given mission architecture. In transit mode, we
exclude all non-transiting planets. In imaging mode, we exclude all planets whose maximum angular separation is less than the IWA, or whose average angular separation is greater than the OWA, or for which the planet-to-star contrast ratio \( \zeta \) is below the instrument noise floor.

The remaining planets can, in principle, be detected by the survey, but to actually detect most of them will require preliminary observations either using the same telescope architecture or a precursor survey. A dedicated imaging mission would likely be able to detect all of the EECs which it is capable of characterizing during preliminary observations (Stark et al. 2019), but the vast majority of transiting planets within the nearest \( \sim 100 \) pc remain undiscovered. Most likely, a large-aperture spectroscopic survey of hundreds of transiting planets must be preceded by a space-based all-sky survey, similar to TESS (Ricker et al. 2015) or PLATO (Rauer et al. 2014) but with sensitivity comparable to \textit{Kepler}. The cost and complexity of such a mission, though considerable, would likely be much less than that of a subsequent characterization effort requiring orders of magnitude greater light-collecting area.

### 4.3. Which planets can be characterized?

In-depth spectroscopic characterization is time-consuming, so the number of targets which can be characterized is a function of the total time budget allotted to the characterization effort \( t_{\text{total}} \). Note that \( t_{\text{total}} \) is not necessarily the same as the total survey lifetime (which might be e.g. 5–10 yr). To determine which planets can be observed within \( t_{\text{total}} \), we first determine the amount of time required to characterize each planet, including overheads, and prioritize targets based on both their required observing time and their relative importance to the survey’s goals.

#### 4.3.1. Required exposure time

To determine which planets can be characterized within the time budget \( t_{\text{total}} \), we first determine the amount of exposure time required to spectroscopically characterize a reference planet whose host star properties reflect the typical target for each survey mode. For both observing modes, the reference planet has exactly the same bulk parameters and receives the same incident flux as modern Earth. For direct imaging observations, its star is a nearby solar-type star \( (T_{\star,\text{ref}} = 5777 \ \text{K}, R_{\star,\text{ref}} = R_{\odot}, d_{\text{ref}} = 10 \ \text{pc}) \) while for transit observations it is a more distant early M dwarf \( (T_{\star,\text{ref}} = 3300 \ \text{K}, R_{\star,\text{ref}} = 0.315 R_{\odot}, d_{\text{ref}} = 50 \ \text{pc}) \). In the examples to follow, we only consider the detection or non-detection of an absorption feature associated with a species, rather than constraints on the abundance.

We use two general circulation models (GCMs) published by Komacek & Abbot (2019) to quantify the three-dimensional atmospheric abundance profiles of our reference planets. Both models are water-covered planets around a Sun-like star (imaging mode) or early M dwarf (transit mode) with the same size, mass, and insolation as Earth and 1 bar \( N_2/H_2O \) atmospheres. These models include a treatment of ice and liquid cloud cover, which is an important factor affecting the detectability of molecular features through imaging and transit observations. Notably, because the M dwarf planet is tidally-locked, convection on its dayside is more efficient, leading to strong, high-altitude cloud cover and greater stratospheric \( H_2O \) abundance (T. Komacek, private correspondence). Finally, to enable the analysis in Section 7, we inject Earth’s modern oxygen abundance \( (pO_2 = 20.7\%) \) into the model atmospheres, reducing the background \( N_2 \) pressure accordingly.

To simulate spectra for both models, we use the Planetary Spectrum Generator (hereafter PSG, Villanueva et al. 2018), which accepts three-dimensional atmospheric profiles through its GlobES module\(^6\). The directly imaged planet is observed at quadrature phase, while the transiting planet is observed with the nightside facing the observer. Both simulated spectra are shown in Figure 3 for atmospheres with and without cloud cover. Next, we use PSG to compute noise estimates for each survey architecture as a function of on-target exposure time. In imaging mode, we use the PSG template for the 15-meter LUVOIR-A observatory, including the projected throughput, spectral resolution, raw contrast, and detector noise for the visible and near-infrared imagers, as well as 4.5 zodis of background dust. In transit mode, we simulate observations for a 50-meter diameter aperture with 60% total throughput, ignoring detector and instrument noise. To determine whether a molecular feature can be detected, we simulate spectra with and without the target molecule and compute the detection signal-to-noise ratio (SNR) across the absorption band in a manner similar to Lustig-Yaeger et al. (2019)\(^7\):

\[
\text{SNR} = \sqrt{\frac{\sum_i (\Delta y_i / \sigma_{y_i})^2}{\text{SNR}}} \quad (4)
\]

where \( \Delta y_i \) is the difference between the two spectra in each spectral bin and \( \sigma_{y_i} \) is the measurement uncertainty. Finally, we compute the exposure time required to achieve a SNR = 5 detection of the feature for the

---

6 The PSG configuration files for this study can be found in the code repository

7 Equations 4–6 therein
reference planet ($t_{\text{ref}}$) in each survey mode, then scale this value to determine the exposure time required for each individual planet detected by the survey.

4.3.2. Exposure time scaling

We define $t_i$ as the amount of exposure time required to spectroscopically characterize a planet at wavelength $\lambda_{\text{eff}}$. If we assume that $t_i$ depends primarily on the number of photons collected, then we can estimate it by scaling $t_{\text{ref}}$ (as determined using PSG) as follows:

$$
\frac{t_i}{t_{\text{ref}}} = f_i \left( \frac{d_i}{d_{\text{ref}}} \right)^2 \left( \frac{R_*}{R_{*,\text{ref}}} \right)^{-2} \left( \frac{B_{s,i}(\lambda_{\text{eff}}, T_{s,i})}{B_{\odot}(\lambda_{\text{eff}}, T_{s,\text{ref}})} \right)^{-1}
$$

(5)

where $f_i$ summarizes the factors affecting the signal strength unique to each observing mode. In imaging mode, the exposure time is inversely proportional to the planet-to-star contrast ratio (assuming observations at quadrature phase):

$$
f_i^{\text{im}} = \left( \frac{\kappa_{i}}{\kappa_{\odot}} \right)^{-1}
$$

(6)

In transit mode, the transit depth signal induced by the atmosphere is (to first order) $\Delta \delta \sim (R_p/R_*)^2(h/R_p)$ (Winn 2010) and the required exposure time is inversely proportional to its square:

$$
f_i^{\text{tr}} = \left( \frac{h_i}{h_{\odot}} \right)^{-2} \left( \frac{R_{p,i}}{R_{\odot}} \right)^{-2} \left( \frac{R_{s,i}}{R_{s,\text{ref}}} \right)^4
$$

(7)

We round up $t_i$ to the next integer multiple of the planet’s transit duration, because a transit survey would likely observe complete transits to measure the baseline. Planets are considered to be invalid targets if the total number of required transit observations is greater than either the number of available transits within 10 years or $10^3$.

These scaling relations are meant to capture the main factors affecting the relative exposure time required for each target so as to provide an approximate mapping between the total amount of time dedicated to a survey and the number and distribution of targets it can observe. Ultimately, the primary metric affecting a survey’s statistical power is usually the number of EECs characterized, and we translate $t_{\text{total}}$ into the number of characterized EECs so the reader can interpret our results as a function of sample size.

4.3.3. Overheads

In imaging mode, following L19 we increment each planet’s required exposure time by 2 hr to account for slew overheads and overheads associated with wavefront control. These overheads end up being relatively insignificant except for the closest targets. In transit mode, we assume 0.5 hr of slew overheads per observation, plus a total overhead equal to the transit duration for baseline observations before and after each transit event.

4.3.4. Target prioritization

Given a limited time budget, it seems reasonable to prioritize observations of planets in order of increasing $t_i$ so as to maximize the number of planets observed. However, prioritizing targets strictly by $t_i$ will lead to a biased sample, especially in the case of transit surveys which are strongly biased towards the detection of close-in planets. To counter-act these biases, we assign a weight $w_i$ to each planet, and calculate its priority as follows:

$$
p_i = w_i/t_i
$$

(8)

The specific choice of $w_i$ depends on the hypothesis being tested and is discussed in Sections 6 and 7. To create the final simulated data set, we observe targets in order of decreasing $p_i$ until some pre-determined time limit $t_{\text{total}}$ is reached.

4.4. Comparison between survey modes

In the following sections, we use Bioverse to evaluate the statistical potential of direct imaging and transit spectroscopy surveys, but we avoid direct comparisons of their results for the following reasons. First, the technical requirements for and limitations of a direct imaging biosignature survey have been more thoroughly explored due to investments in the LUVOIR and HabEx mission concepts. As a result, our results for the transit survey are likely more optimistic. Second, we do not wish to imply that a survey’s statistical power is the only or most important dimension for comparison, as each architecture enables unique capabilities which the other does not.

For the topics discussed here, the primary difference between the two surveys is the number of EECs each can characterize. For the 15-meter imaging survey, this number is 15–20, and is volume- rather than time-limited. This estimate is consistent with that of L19 when adjusted for our updated value of $\eta_{\odot}$ ($\approx 7.5\%$ for G stars). For the 50-meter (equivalent area) transit survey, this number grows with time, with e.g. 60–70 EECs being surveyed for H$_2$O absorption or $\sim 200$ for O$_3$ absorption given $t_{\text{total}} = 2$ yr.

5. HYPOTHESIS TESTING

The third component of Bioverse assesses the information content contained within the simulated data sets
from the previous section. This assessment focuses on two primary questions: first, how likely is it that the survey would be able to detect the effects of a statistical trend injected into the simulated planet population (Section 3.8)? Second, how precisely could the survey constrain the parameters of that trend? To answer these questions, we rely on a standard Bayesian hypothesis testing approach.8

5.1. Null and alternative hypotheses

Each simulated data set can be thought of as a set of independent variables \( x \) and dependent variables \( y \). For this section (and the examples to follow), we consider \( x \) and \( y \) to each represent measurements of a single variable, but this hypothesis testing framework can extend to multivariate measurements as well. The hypothesis \( h(\theta, x) \) describes the relationship between the \( x \) and \( y \) in terms of a set of parameters \( \theta \). The simplest hypothesis is the null hypothesis, in which there is no relationship:

\[
h_{\text{null}}(\theta, x) = \theta
\]

The null hypothesis is compared to an alternative hypothesis, which proposes a relationship between \( x \) and \( y \), using a Bayesian parameter estimation and hypothesis testing approach.

5.2. Likelihood function and prior distribution

Given a hypothesis \( h \), the likelihood function takes on one of two forms. In the case where \( y \) is binary (e.g., the detection or non-detection of an atmospheric species), then \( h \) is the probability that \( y = 1 \), and the likelihood function is:

\[
\mathcal{L}(y|\tilde{\theta}) = \prod_i \left[ y_i h(\tilde{\theta}, x_i) + (1 - y_i)(1 - h(\tilde{\theta}, x_i)) \right]
\] (9)

Alternatively, if \( y \) is a continuous variable measured with normal uncertainty \( \sigma_y \), then \( h \) predicts the expectation value of \( y \), and the likelihood is described by the normal distribution:

\[
\mathcal{L}(y|\tilde{\theta}) = \prod_i \frac{1}{\sqrt{2\pi\sigma_{y,i}^2}} \exp \left( -\frac{(y_i - h(\tilde{\theta}, x_i))^2}{2\sigma_{y,i}^2} \right)
\] (10)

Note that in both example applications of Bioverse to follow, we consider a detection or non-detection as our dependent variable and use the likelihood function defined by Equation 9.

The parameter prior distribution is denoted by \( \Pi(\tilde{\theta}) \). Given limited prior information about the true values of parameters \( \tilde{\theta} \), we generally assume uniform or log-uniform distributions spanning the range of plausible values. Further justification for our choice of prior distributions can be found in the examples to follow.

5.3. Parameter estimation and Bayesian evidence

For each simulated data set, we sample the posterior distribution of the hypothesis parameters \( \tilde{\theta} \) using a Markov Chain Monte Carlo (MCMC) algorithm, implemented by \texttt{emcee} (Foreman-Mackey et al. 2013). This sampling yields measurement constraints of the parameters \( \tilde{\theta} \). We also use a nested sampling algorithm (Skilling 2006), implemented by \texttt{dynesty} (Speagle 2020), to esti-
mate the Bayesian evidence for the alternative hypothesis:

$$ Z = P(y|h) = \int \mathcal{L}(y|\theta) \Pi(\theta) d\theta $$  

(11)

To test a hypothesis, we can compare its evidence to that of the null hypothesis, finding evidence to reject the null hypothesis when:

$$ \Delta \ln(Z) = \ln(Z) - \ln(Z_{null}) > 3 $$  

(12)

We choose $\Delta \ln(Z) > 3$ as our threshold because it corresponds to the common $p < 0.05$ threshold for hypothesis testing with other frequentist tests (e.g., Student’s t-test).

It should be noted that `dynesty` also samples the parameter posterior distributions - so why use `emcee` to do this separately? In short, nested sampling is optimized to measure $Z$, while MCMC is optimized to determine the posterior distribution. While `dynesty` can quickly compute the Bayesian evidence with sufficient accuracy ($\sigma_{\ln(Z)} \lesssim 0.5$), we find it takes significantly longer to converge to the same parameter posterior distributions as `emcee`. Since we repeat each simulated survey $> 100,000$ times, we find this mixed approach to be necessary to achieve both accurate evidence and parameter estimations on a reasonable timescale.

5.4. Statistical power

Whether or not an individual simulated survey is able to reject the null hypothesis can often depend on stochastic error; one simulated survey may be able to reject the null hypothesis where another cannot. To summarize our results, we re-run each simulated survey several times under the same set of assumptions and calculate the fraction of survey realizations which achieve a positive result. This metric is also known as the statistical power, and it allows us to assess a survey’s statistical potential as a function of both survey parameters (such as total survey duration) and as-yet unknown astrophysical parameters (such as the frequency of habitable planets).

This concludes the description of the three primary components of Bioverse. In the following two sections, we will demonstrate applications of Bioverse to its stated goal of assessing the statistical power of next-generation biosignature surveys.

6. EXAMPLE 1: EMPIRICAL DETERMINATION OF THE HABITABLE ZONE BOUNDARIES

Models of the habitable zone predict that planets with oceans can only exist within a finite - and perhaps very narrow - range of insolations. An associated prediction is that terrestrial planets in the habitable zone with water-rich atmospheres are the most likely candidates for ocean-bearing worlds. These models will play an important role in the design and target prioritization of next-generation observations; for example, preliminary search strategies for future biosignature surveys often dedicate intensive follow-up to water-bearing habitable zone planets (The LUVOIR Team 2019), while delegating non-habitable zone planets to a lower priority. However, models for the habitable zone have not been tested outside of the solar system, and estimates of its location and width have varied by factors of several over the past few decades.

Could future observatories use data acquired from preliminary observations to test the “habitable zone hypothesis” i.e., the hypothesis that planets with water vapor should be more abundant within a narrow and finite range of orbital separations? Further, could these data be used to empirically determine the location and width of the habitable zone? The practical benefit of testing the habitable zone hypothesis would be to make the survey’s target prioritization strategy more efficient and to better determine which of its targeted planets are most likely to be habitable. By measuring its boundaries, observers could test the predictions of various habitable zone models, and therefore the physical mechanisms on which they rely. Finally, empirical constraints on the width of the habitable zone will be important for determining the occurrence rate of habitable worlds. Here, we use Bioverse to explore how a survey of atmospheric water vapor could be used to test the habitable zone hypothesis.

6.1. Model predictions

Climate models predict a steep decline in water vapor abundance of terrestrial planets outside of the habitable zone. Within the inner edge, an Earth-like planet may undergo a runaway greenhouse as on Venus, leaving behind only a tenuous amount of atmospheric water vapor. Beyond the outer edge, the oceans may freeze, and water vapor would not accumulate except in very low pressure atmospheres which permit its sublimation.

In Bioverse we implement these predictions as follows. We assume that a fraction $f_{\text{EEC}}^{\text{H}_2\text{O}}$ of EECs are in fact habitable, meaning they bear surface water and atmospheric water vapor. We also allow a fraction $f_{\text{non-EEC}}^{\text{H}_2\text{O}}$ of non-EECs to have atmospheric water vapor, serving as a source of noise and “false positives”
for habitable planets. Then the fraction of planets with atmospheric water vapor can be described as:

\[
f_{H_2O}^{\text{EEC}} \begin{cases} 
  f_{H_2O}^{\text{non-EEC}} & \text{if } a < a < a_{\text{outer}} \\
  0 & \text{if } R > 0.85 R_{\odot}^{0.25} \\
  f_{H_2O}^{\text{EEC}} & \text{if } a_{\text{inner}} < a < a_{\text{outer}} \\
  0 & \text{if } R < 0.85 R_{\odot}^{0.25} \\
\end{cases}
\]

where the habitable zone boundaries and planet size limits are those discussed in Sections 3.3 and 3.4.

### 6.2. Simulated survey

#### 6.2.1. Measurements

The imaging and transit surveys perform a set of measurements outlined in Table 2 to determine the size and orbital separation of each potential target. In imaging mode, the planet’s size is not determinable without prior knowledge of the geometric albedo, so an estimated size \(R_{\text{est}}\) which assumes Earth-like reflectivity is used as a proxy. In both modes, the orbital separation is converted to the “effective” semi-major axis \(a_{\text{eff}}\) for which the planet would receive the same insolation around a Sun-like star.

These preliminary measurements are used to prioritize targets as discussed in the following section. Those targets of high enough priority are spectroscopically characterized to determine whether their atmospheres contain H\(_2\)O. The final output of each simulated survey as a data set consisting of \((a_{\text{eff}}, R_{\text{H}2O})\), where H\(_2\)O = \{0, 1\} reflects the absence or presence of water absorption features in the planet’s spectrum. One example of a simulated data set is shown in Figure 4.

#### 6.2.2. Target prioritization

To test the habitable zone hypothesis we must observe planets spanning a broad range of semi-major axes, but prioritizing targets solely based on required exposure time will bias observations towards close-in planets. Furthermore, planets much smaller or larger than Earth are not likely to be habitable regardless of insolation, and therefore serve as a source of noise. The counter these effects, we weight each target according to its size and orbital separation following Figure 5. We tuned this prioritization based on trial and error to achieve the following goals:

1. Prioritize observations of more probable Earth analogs (planets receiving 50–150% of Earth’s incident flux).

2. Balance observations of widely-separated planets versus close-in planets.

3. Minimize observations of non-Earth sized planets.

In transit mode, we additionally weight each target by \((a/R_s)\) to negate the bias due to close-in planets being more likely to transit. The resulting distribution of observed planets is also shown in Figure 5.

6.2.3. Time budget

Following the procedure in Section 4.3.1, we use PSG to determine the exposure time required for a 5\(\sigma\) detection of water vapor absorption through its near-infrared absorption bands. In transit mode, we combine the SNR from the 1.4 and 1.9 \(\mu\)m features. In imaging mode, we only target the 1.4 \(\mu\)m band, as LUVOIR will be unable to observe the full near-infrared spectrum simultaneously, and the 1.9 \(\mu\)m band is harder to observe due primarily to lower stellar flux.

In imaging mode, we find \(t_{\text{ref}} = 0.9\) hr are necessary to probe the reference planet for water vapor, while in transit mode we require \(t_{\text{ref}} = 181\) hr of in-transit exposure time; the details of these calculations are shown in Table 4. Next, we scale \(t_{\text{ref}}\) according to Equations 5 through 7 to determine the amount of exposure time required to characterize each planet, then add observing overheads. We weight the targets according to Equation 8, then finally observe by order of decreasing probability until the total time budget \(t_{\text{total}}\) is reached.

The average number of EECs observed by each survey is displayed in Figures 6a and 7a as a function of either \(\eta_{\text{EEC}}\) or \(t_{\text{total}}\). While we use the number of EECs observed as our primary metric of sample size, note that most observed targets are non-habitable. Since the imaging survey yield quickly becomes volume-limited, we investigate the impact of varying \(\eta_{\text{EEC}}\) for a fixed time budget \(t_{\text{total}} = 120\) d (which is sufficient to characterize > 90% of detectable EECs). For the transit survey, we fix \(\eta_{\text{EEC}} = 7.5\%\) and investigate the impact of varying \(t_{\text{total}}\).

6.3. Habitable zone hypothesis

Now, let us approach the simulated data from the view of an observer who no prior knowledge of Equation 13 using the Bayesian hypothesis testing framework outlined in Section 5. The habitable zone hypothesis states that planets within the habitable zone are more likely to have water vapor than those outside of it:

\[
h_{\text{HZ}}^{\text{HZ}}(a_{\text{eff}}) = \begin{cases} 
  f_{\text{HZ}} & \text{if } a_{\text{inner}} < a < a_{\text{inner}} + \Delta a \\
  f_{\text{HZ}}(f_{\text{non-HZ}}/f_{\text{HZ}}) & \text{otherwise} \\
\end{cases}
\]
Figure 4. An example of a simulated direct imaging data set for Section 6. Planets are probed for the presence of atmospheric water vapor across a broad range of orbital separations. We assume the habitable zone (gray) to be marked by an abundance of water-rich atmospheres. The separation $a_{\text{eff}} = a(L_*/L_\odot)^{-1/2}$ is the solar-equivalent semi-major axis.

Table 2. Measurements made by the simulated surveys in Examples 1 and 2. Parameters marked by † are calculated from other measured values.

| Parameter | Measurement uncertainty | Description / notes |
|-----------|-------------------------|---------------------|
| **Example 1** | | |
| Imaging survey | | |
| $L_*$ | negligible | Host star luminosity |
| $\zeta$ | 15% | Planet-to-star contrast |
| $a$ | 10% | Semi-major axis |
| $a_{\text{eff}}^\dagger$ | 10% | Solar-equivalent semi-major axis |
| $R_{\text{est}}^\dagger$ | 10% | Estimated radius assuming Earth-like reflectivity |
| $\text{H}_2\text{O}$ | Detected / not detected | Presence of 1.4 $\mu$m $\text{H}_2\text{O}$ absorption |
| Transit survey | | |
| $M_*, R_*$ | 5% | Host star mass and radius |
| $P$ | negligible | Orbital period |
| $\delta$ | negligible | Baseline transit depth |
| $a_{\text{eff}}^\dagger$ | 1.7% | Solar-equivalent semi-major axis |
| $R^\dagger$ | 5% | Planet radius |
| $\text{H}_2\text{O}$ | Detected / not detected | Presence of 1.4 $\mu$m and 1.9 $\mu$m $\text{H}_2\text{O}$ absorption |
| **Example 2** | | |
| Imaging survey | | |
| $t_*$ | 10% | Age (as measured through asteroseismology) |
| $\text{O}_2$ | Detected / not detected | Presence of 0.7 $\mu$m $\text{O}_2$ absorption |
| Transit survey | | |
| $t_*$ | 30% | Age (model-based estimate) |
| $\text{O}_3$ | Detected / not detected | Presence of 0.6 $\mu$m $\text{O}_3$ absorption |
Example 1: Target prioritization and distribution

Imaging survey

| Target priority | Resulting distribution |
|------------------|------------------------|
| $R_{\text{eff}} (R_e)$ |                         |
| 0.1              | 1.0                    |
| 1.0              | 1.5                    |
| 2.0              |                         |

Transit survey

| Target priority | Resulting distribution |
|------------------|------------------------|
| $R (R_e)$ |                         |
| 0.1              | 1.0                    |
| 1.0              | 1.5                    |
| 2.0              |                         |

Figure 5. Summary of target prioritization for the simulated imaging (top) and transit (bottom) surveys in Section 6. The left panel shows the relative weight assigned to each target as a function of size and orbital separation ($w_i$ in Equation 8). The right panel shows the resulting relative distribution of targets which can be probed for the presence of water vapor within the survey duration. In the case of the imaging survey, the planet size cannot be directly measured, so the “estimated” radius (assuming Earth-like reflectivity) is used as a proxy. In the case of the transit survey, an additional weight is applied to counteract the $R_\star/a$ transit probability (not shown above).
Table 3. Parameter prior distributions for Equations 14 and 16.

| Parameter       | Description                             | Prior limits (log-uniform distribution) |
|-----------------|-----------------------------------------|------------------------------------------|
| $a_{\text{inner}}$ | Inner edge of the habitable zone        | 0.1 – 2.0 AU                             |
| $\Delta a$      | Width of the habitable zone             | 0.01 – 10 AU                             |
| $f_{\text{HZ}}$ | Fraction of habitable zone planets with $\text{H}_2\text{O}$ | 0.001 – 1                                |
| $(f_{\text{non-HZ}}/f_{\text{HZ}})$ | Fraction of non-habitable zone planets with $\text{H}_2\text{O}$ (relative to $f_{\text{HZ}}$) | 0.001 – 1                                |

Example 2

| Parameter          | Description          | Prior limits |
|--------------------|----------------------|--------------|
| $f_{\text{life}}$ | Fraction of EECs with life | 0.001 – 1    |
| $t_{1/2}$          | Oxygenation timescale of inhabited planets | 0.1 to 100 Gyr |

Table 4. The predicted signal strengths of $\text{H}_2\text{O}$ (Example 1) and $\text{O}_2$ or $\text{O}_3$ (Example 2) absorption for a representative target of each survey mode, expressed as the peak amplitude of the change in planet-to-star contrast ratio (in parts-per-trillion) or transit depth (in parts-per-million) within the absorption band. The exposure time required for a $5\sigma$ detection is determined using PSG, and scaled for each individual target according to Equation 5. In transit mode, the signals from two bands are combined to achieve the detection of $\text{H}_2\text{O}$. In imaging mode, we select the feature which requires the least exposure time to detect.

| Survey mode | Feature | Wavelength | Signal strength (without clouds) | Signal strength (with clouds) | Time required (hr) |
|-------------|---------|------------|---------------------------------|------------------------------|-------------------|
| Imaging     | $\text{H}_2\text{O}$ | 1.4 $\mu$m | 90 ppt                          | 55 ppt                       | 0.9               |
|             | $\text{O}_2$     | 0.76 $\mu$m | 90 ppt                          | 70 ppt                       | 2.6               |
| Transit     | $\text{H}_2\text{O}$ | 1.4 $\mu$m | 3.5 ppm                         | 0.5 ppm                      | 181               |
|             | $\text{O}_3$     | 1.9 $\mu$m | 5 ppm                           | 0.7 ppm                      |                   |
|             | $\text{O}_3$     | 0.6 $\mu$m | 4 ppm                           | 2 ppm                        | 74                |

This is a four parameter model with $\vec{\theta} = [a_{\text{inner}}, \Delta a, f_{\text{HZ}}, (f_{\text{non-HZ}}/f_{\text{HZ}})]$. The choice of parameters was driven by two factors: first, the width of the habitable zone ($\Delta a$) is relevant for testing “rare Earth” models in which the habitable zone is very narrow. Second, we can use simple log-uniform prior distributions for these parameters without having to filter out parameter combinations which violate the assumptions of the habitable zone hypothesis (e.g., $f_{\text{non-HZ}} > f_{\text{HZ}}$).

6.4. Prior assumptions

K14’s model for the habitable zone, which we implement in the simulated planet population, assumes a carbon-silicate feedback cycle which enhances CO$_2$ concentrations for planets further from their host stars, and spans 0.95–1.67 AU for the Sun. While this estimate has strong heritage (Kasting et al. 1993; Kopparapu et al. 2013), it has also been preceded and succeeded by more conservative or generous estimates, which we use to set the prior distribution of values considered for $a_{\text{inner}}$ and $\Delta a$.

Estimates of the inner edge range as far inward as 0.38 AU (for highly-reflective desert worlds with a minimal greenhouse effect, Zsom et al. 2013) and we allow that the inner edge could be as far out as 2 AU, in which case Earth would be an unusually cool outlier. Estimates of the habitable zone width have varied as well; a classic estimate by Hart (1979) suggests a very narrow habitable zone ($\Delta a < 0.1$ AU), which would imply that Earth-like planets are especially rare. More recent estimates have proposed mechanisms by which the habitable zone could extend as far as 2.4 AU (Ramirez & Kaltenegger 2017) or even 10 AU (Pierrehumbert & Gaidos 2011) from the Sun. Given this wide range of estimates for $a_{\text{inner}}$ and
\( \Delta a \), we assume broad prior distributions for both, shown in Table 3.

### 6.5. Results

We repeat the simulated survey and Bayesian analysis \( > 10,000 \) times over a grid of values for the astrophysical parameters in Equation 13 for both survey architectures. With each simulated survey, we use dynesty to calculate the Bayesian evidence in favor of the habitable zone hypothesis, and emcee to sample the posterior distributions of \( \sigma_{\text{inner}} \) and \( \Delta a \). The results are summarized by Figure 6 and 7 for the simulated imaging and transit surveys, respectively.

#### 6.5.1. Imaging survey

An ambitious direct imaging survey with a 15-meter telescope could confidently detect the habitable zone with a 3-month long observing campaign provided most EECs are habitable. If habitable planets are less common, however, then more EECs must be observed. The EEC yield of an imaging survey is typically volume-limited, so higher values of \( \eta_B \) would be required to test this hypothesis for more pessimistic astrophysical parameters. In the best case scenario (\( \eta_B \approx 40\% \)), a 15-meter imaging mission could perform the test if 20\% of EECs are habitable, but this value for \( \eta_B \) is likely too optimistic.

If \( \sim 80\% \) of EECs are habitable, the imaging survey would be able to measure the location of the habitable zone with sufficient accuracy to exclude some more extreme estimates of its boundaries with reasonable confidence. In particular, it would be able to place a confident lower bound on \( \Delta a \), rejecting some “rare Earth” models which predict a very narrow habitable zone (e.g., Hart 1979).

Finally, it should be noted that imaging surveys will have access to planet brightness and color information which could be incorporated into this analysis; for example, albedo and photometric color may vary predictably across the habitable zone (Checlair et al. 2019). Hypotheses which include this information could be tested with better statistical power and parameter constraints than the one examined here.

#### 6.5.2. Transit survey

The transit survey can confidently detect the habitable zone even in the case where most EECs are not habitable (\( f_{\text{H}_2\text{O}} \approx 25\% \)), provided 60–70 EECs can be probed for atmospheric water vapor during a 2-year characterization effort. Furthermore, the full survey duration may not be necessary if most EECs end up to be habitable (in which case a shorter 3–6 month survey would suffice).

The transit survey can precisely measure the inner edge of the habitable zone to within \( \pm 0.1 \) AU of its true location in most simulated surveys if most EECs are habitable, and can sometimes accomplish this even if most EECs are not habitable. The width (or outer edge) is more difficult to constrain as the transit survey only observes a handful of planets beyond the outer edge (\( \sim 10 \)). This bias has two causes: first, planets outside of the habitable zone are less likely to transit, so they are typically found around more distant stars. Second, colder planets have smaller atmospheric scale heights, and therefore weaker \( \text{H}_2\text{O} \) absorption features. Both of these effects increase the time required to characterize cold planets, making them low priority targets.

### 6.6. Discussion

#### 6.6.1. Impact of clouds

Clouds will have a major impact on the transit survey’s ability to test the habitable zone hypothesis, as they dampen the absorption signal due to tropospheric water vapor and therefore increase the number of transit observations required to detect it. As shown in Figure 7a, this means that a much smaller number of targets can be observed within a fixed time budget, and many of the most distant targets become infeasible to characterize as it would require the combination of decades’ worth of transit observations. A possible mitigating strategy would be to expand the observatory’s light-collecting area. The Nautilus Space Observatory, on which we base our transit survey results (Apai et al. 2019), would consist of 35 identically-manufactured unit telescopes. As such, the cost would scale linearly with light-collecting area, and doubling the number of telescopes would reduce by nearly half the number of transit observations required to characterize each planet.

Our cloud assumptions are based on the GCM models of Komacek & Abbot (2019), who show that tidally locked planets around M dwarfs have much higher day-side cloud covering fractions than Earth-like planets. If this bears true, it will likely prevent the characterization of such planets through transit spectroscopy by JWST (Fauchez et al. 2019; Komacek et al. 2020; Suissa et al. 2020; Pidhorodetska et al. 2020) and possibly even larger observatories. In the pessimistic case, even a 50-meter equivalent area transit survey may be unable to detect atmospheric water vapor for all but a handful of nearby exo-Earths orbiting M dwarfs, so the survey must target more distant K and G dwarfs instead. This will come at the cost of sample size, as we estimate that the increased average distance, less frequent transits, and lower transit depths for habitable zone planets around these stars.
Example 1: Results for 15-meter imaging survey
How many exo-Earth candidates are probed for H$_2$O?

Can this survey test the habitable zone hypothesis?

Can this survey determine the location of the habitable zone?
(results for six random realizations of the survey)

Figure 6. Results for the imaging survey in Section 6. (a) The number of EECs observed versus $\eta_\oplus$ (for G stars), assuming $t_{\text{total}} = 120$ d. As our baseline case, we set $\eta_\oplus = 7.5\%$. (b) The statistical power to test the habitable zone hypothesis as a function of the astrophysical parameters in Equation 13. (c) The minimum number of EECs which must be characterized to achieve 80% statistical power, with the corresponding values of $\eta_\oplus$. (d) The posterior probability that a planet with effective separation $a_{\text{eff}}$ is in the habitable zone, as estimated by six random realizations of the survey under an optimistic case (80% of EECs are habitable, left) and pessimistic case (20% of EECs are habitable, right). The true habitable zone is highlighted in green, and in both cases 1% of non-habitable planets have H$_2$O.
Example 1: Results for 50-meter (equivalent area) transit survey
How many exo-Earth candidates are probed for H$_2$O?

Can this survey test the habitable zone hypothesis?

Can this survey determine the location of the habitable zone?
(results for six random realizations of the survey)

Figure 7. Results for the transit survey in Section 6. (a) The number of EECs observed versus the observing time budget, assuming $\eta_\oplus = 7.5\%$ for G stars and cloudy atmospheres (solid). 4–10$x$ as many planets could be observed if clouds were neglected (dashed), or 1.5–3$x$ as many with clouds if assuming the higher SAG13 estimate of $\eta_\oplus = 24\%$ (dotted). As our baseline case, we set $t_{total} = 2$ yr. (b) The statistical power to test the habitable zone hypothesis as a function of the astrophysical parameters in Equation 13. (c) The minimum number of EECs which must be characterized to achieve 80% statistical power, with the necessary observing time budget $t_{total}$. (d) The posterior probability that a planet with effective separation $a_{eff}$ is in the habitable zone, as estimated by six random realizations of the survey under an optimistic case (80% of EECs are habitable, left) and pessimistic case (20% of EECs are habitable, right). The true habitable zone is highlighted in green, and in both cases 1% of non-habitable planets have H$_2$O.
will outweigh their higher stellar luminosity in terms of observing time cost.

Clouds impact imaging observations as well, although they have little effect on the results presented here. We expect cloud cover to be less prevalent for nondiably locked planets orbiting Sun-like stars, and highly-reflective clouds at low enough altitudes can have a beneficial effect on imaging observations as they amplify the absorption due to molecules in higher layers. More importantly, the imaging survey modeled here is volume-rather than time-limited, so with or without clouds we find that the survey can probe its entire EEC sample for water absorption within less than three months.

Our exposure time and sample size estimates for the transit survey are based on one possible realization of cloud conditions, but cloud cover may vary greatly across targets and observation epochs. Indeed, in the GCM models we employ, the effect of clouds on transmission spectra is sensitive to orbital period, spectral type, cloud particle size, and many other parameters (Komacek et al. 2020), suggesting that the actual distribution of cloud properties in terrestrial exoplanet atmospheres may be fairly broad. An efficient transit survey could seek to identify planets with clearer atmospheres (e.g. through scattering features in visible light) and prioritize these over cloudier targets, thereby increasing the sample size.

6.6.2. Effect of non-habitable H$_2$O-rich atmospheres

Naturally, the habitable zone hypothesis is easier to test if more habitable planets are observed, and the number of EEC characterizations required to test it is approximately proportional to the fraction of EECs which are habitable ($f_{\text{H}_{2}\text{O}}^{\text{EEC}}$). However, non-habitable planets are far more common than habitable planets, so if even a small fraction ($f_{\text{H}_{2}\text{O}}^{\text{non-EEC}}$) of these have H$_2$O, the statistical excess of H$_2$O in the habitable zone will be muted. In general, we find the statistical power to be unaffected provided that $f_{\text{H}_{2}\text{O}}^{\text{non-EEC}} \lesssim 1\%$, but the impact can be considerable if $f_{\text{H}_{2}\text{O}}^{\text{non-EEC}} \gtrsim 10\%$. This result seems sensible, as approximately 10% of the total sample are EECs, so $f_{\text{H}_{2}\text{O}}^{\text{non-EEC}} > 10\%$ would imply that H$_2$O-rich non-habitable planets are more common than habitable planets.

Our assumption in Equation 13 is that all EECs with water vapor are habitable, and the fraction of non-EECs with water vapor is mostly independent of insolation. However, if such “false positives” exist, their abundance is likely a function of insolation. For example, consider a population of non-habitable planets whose surfaces have been desiccated by a runaway greenhouse effect but which still maintain thick, H$_2$O-rich atmospheres. Such planets should be clustered near the inner edge of the habitable zone (e.g., Turbet et al. 2019), appearing as an extension of the habitable planet population to high insolations rather than as a distinct planet population. Even planets defined as EECs may actually be non-habitable (due to differences in initial volatile content, plate tectonics, outgassing rates, etc.) yet still possess water vapor, making them statistically indistinguishable from habitable EECs. Again, the effect of these false positives will likely be negligible provided they are much less common than habitable planets, but indicators of planetary (non-)habitability other than H$_2$O may be necessary to filter them out.

7. EXAMPLE 2: EVOLUTION OF EARTH-LIKE PLANETS

By characterizing a sufficiently-large sample of terrestrial worlds, a next-generation observatory could test hypotheses for how they evolve over time. One such hypothesis is that inhabited planets with oxygen-producing life, like Earth, evolve towards greater oxygen content over Gyr timescales due to long-term changes in global redox balance. As we propose in Bixel & Apai (2020a) (hereafter B20), the impact on a population level would be a positive “age-oxygen correlation”, wherein older inhabited planets are more likely to have oxygenated atmospheres.

If inhabited planets do tend to evolve towards greater oxygen content over time, then what is the typical timescale for this evolution? Earth underwent major oxygenation events at 2–2.5 Gyr of age and again at $\sim 4$ Gyr (Lyons et al. 2014), suggesting a $\sim 4$ Gyr “oxy-genation timescale” (Catling et al. 2005). These two events mark the boundaries between the Archean, Proterozoic, and Phanerozoic eras, and correspond to shifts in Earth’s redox balance where the amount of oxygen being produced by life became large enough, and/or the geological sinks for oxygen became diluted enough, that oxygen was allowed to build up in the atmosphere. However, a great diversity of planetary factors might affect redox balance, such as outgassing rates, the stellar radiation profile, biogenic oxygen flux, and the planet’s initial reducing matter inventory (Catling et al. 2005; Bixel & Apai 2020a). As a result, Earth’s oxygenation timescale could be unusually fast or slow compared to the overall population of inhabited worlds.

A next-generation biosignature survey could not only detect the proposed age-oxygen correlation, but also measure the typical timescale over which this evolution occurs. This measurement could be used to test models for the geological and biological evolution of Earth-like planets and offer insight into how Earth relates to the
rest of that population. Here, we assess the ability of direct imaging and transit surveys to study the oxygenation history of Earth-like planets. This section follows a similar methodology to our previous analysis (B20), but expands upon it by incorporating a more thorough assessment of planet occurrence rates, detection sensitivity, and survey strategy, and by studying a broader range of evolutionary timescales.

7.1. Model predictions

We assume a fraction $f_{\text{life}}$ of EECs to be inhabited by life - note that this parameter absorbs factors affecting both the planet’s habitability and the likelihood of life originating. Over time, simulated inhabited planets transition from anoxic to oxygenated atmospheres at an average rate described by a half life $t_{1/2}$. The resulting fraction of habitable planets which have oxygenated atmospheres as a function of age $t_*$ is:

$$f_{O_2}(t_*) = f_{O_3}(t_*) = f_{\text{life}} \left(1 - 0.5^{t_*/t_{1/2}}\right) \quad (15)$$

Note that we assume oxygenated atmospheres to have both O$_2$ and its photochemical byproduct O$_3$. We run simulations for $f_{\text{life}}$ ranging from 0–100% and for $t_{1/2}$ ranging from 500 Myr – 50 Gyr.

7.2. Simulated survey

7.2.1. Measurements

The measurements performed by each simulated survey are summarized in Table 7. First, we measure the age ($t_*$) of every planet’s host star with 10% precision for the imaging survey and 30% precision for the transit survey. These estimates represent the state of the art for high- and low-mass stars, respectively. For high-mass stars, asteroseismology has yielded highly precise age constraints for Kepler targets (e.g., Creevey et al. 2017; Kayhan et al. 2019; Lund et al. 2019), and will likely be able to do so for most of the O(100) stellar targets probed by an imaging mission. For low-mass stars, asteroseismology has not been successful (e.g., Rodríguez-López et al. 2015; Rodríguez et al. 2016; Berdiñas et al. 2017), and age determination currently relies on a synthesis of model-based estimates. As an example, Burgasser & Mamajek (2017) use a combination of approaches to determine the age of TRAPPIST-1 planetary system (Gillon et al. 2017) with $\sim$ 30% precision.

Next, each planet is observed to constrain the presence of oxygen. For an Earth-like planet, O$_2$ can be detected directly through its 0.77 $\mu$m absorption feature or inferred through absorption by stratospheric ozone in the Chappuis (0.40–0.65 $\mu$m) or Hartley (0.2–0.3 $\mu$m) bands. It should be noted that our calculations assume...
modern Earth O$_2$ and O$_3$ abundances, an assumption which we revisit in Section 7.6.1.

For each survey mode, we determine which of these three features would be easiest to observe across the full range of detected EECs. In imaging mode we observe O$_2$-A absorption; while the Hartley band may be easier to detect for a solar-type star, it becomes more expensive to observe for lower-mass stars, and the Chappuis band signal is too shallow. Ultimately this consideration is unimportant for the volume-limited imaging survey, and it is likely that all three features will be searched for in the atmospheres of all detected EECs. In transit mode we observe the Chappuis band, as its signal is strong in transit observations. The Hartley band is inaccessible for the vast majority of (predominantly M dwarf) transit survey targets, and the O$_2$-A feature is too shallow and narrow to detect for distant targets.

In total, the simulated surveys produce measurements of $(t_*, O_2)$ for each observed EEC, where $O_2 = \{0, 1\}$ indicates the detection or non-detection of either O$_2$-A absorption (imaging mode) or O$_3$ Chappuis band absorption (transit mode).

### 7.2.2. Target prioritization

Unlike in the previous example, we do not prioritize targets by size or insolation except that we assume all targets have previously been identified as EECs (perhaps with follow-up observations to confirm the presence of H$_2$O). This assumption is not trivial; imaging surveys cannot easily determine a planet’s size, and the true range of planet sizes and insolutions which permit habitability are not yet known. In reality, it is likely that an actual biosignature survey will probe some planets which are not habitable for reasons yet unknown to the observer, which will serve as a source of noise (i.e. by reducing $f_{\text{life}}$).

However, we do prioritize targets by age according to Figure 9, with observations of the youngest and oldest planets being preferred. This is not intended to counter any bias in the underlying sample, as there are no factors which bias the number of planets which can be characterized by our simulated surveys as a function of age. Rather, as we demonstrate in B20, a survey which prioritizes younger and older planets will be more sensitive to monotonic, age-dependent trends because of the larger contrast between those categories. While this prioritization strategy is optimal for studying the evolution of Earth-like planets, it must be balanced versus the survey’s other goals. Notably, it de-prioritizes observations of modern Earth analogs, which may be the best planets to probe if the sole goal is to maximize the chance of detecting O$_2$.

### 7.2.3. Time budget

As discussed in Section 7.2.1, we consider the detection of O$_2$-A absorption in imaging mode and O$_3$ Chappuis band absorption in transit mode. The details of the exposure time calculations are shown in Table 4. Using PSG, we determine the exposure time required for the reference target to be $t_{\text{ref}} = 2.6$ hr for imaging mode and $t_{\text{ref}} = 74$ hr for transit mode.

#### 7.3. Hypothesis and prior assumptions

Once more, we take the role of an observer interpreting the results of each simulated survey. Our hypothesis is that inhabited planets tend to evolve towards greater oxygen content over time, and can be stated in similar terms as Equation 15:

$$h(t_*) = f_{\text{life}} \left(1 - 0.5^{t_*/t_{1/2}}\right)$$

(16)

We adopt broad, log-uniform prior distributions for $f_{\text{life}}$ and $t_{1/2}$, shown in Table 3, reflecting our significant prior uncertainty as to frequency and evolutionary timescales of inhabited planets.

#### 7.4. Correlation test

In lieu of the Bayesian evidence test used in the previous example, we employ the Mann-Whitney test (Mann & Whitney 1947) to determine whether $t_*$ correlates with the presence of oxygen, as we previously have done in B20. This model-independent test is more sensitive for detecting the correlation than the Bayesian evidence-based approach, especially in the limit of small sample sizes. However, it does not allow for the estimation of $t_{1/2}$, for which we rely on MCMC sampling.

#### 7.5. Results

We assess the statistical power of each survey to test the age-oxygen correlation hypothesis, using the Mann-Whitney test to determine whether a positive correlation can be detected in each simulated data set, and emcee to sample the posterior distributions of $t_{1/2}$ and $f_{\text{life}}$. Our results are summarized in Figure 10 for the imaging survey and Figure 11 for the transit survey.

#### 7.5.1. Imaging survey

Assuming $\eta_B \approx 7.5\%$, it is unlikely (though not impossible) that a direct imaging survey will be able to detect the age-oxygen correlation with a sample of 15–20 EECs. This is generally consistent with our analysis in B20, which suggests a statistical power of 50% for a sample size of $\sim 20$ EECs only if most are inhabited. In order for an imaging survey to be reliably capable of studying the oxygen evolution of Earth-like
Example 2: Results for 15-meter imaging survey
How many exo-Earth candidates are probed for $O_2$?

Could this survey detect the age-oxygen correlation?

Figure 10. Results for the imaging survey in Section 7. (a) The number of EECs observed versus $\eta_\oplus$ (for G stars), assuming $t_{\text{total}} = 120$ d. For our baseline case, we set $\eta_\oplus = 7.5\%$. (b) The statistical power to detect the age-oxygen correlation as a function of the astrophysical parameters in Equation 15. (c) The minimum number of EECs which must be characterized to achieve 80% statistical power, with the corresponding values of $\eta_\oplus$.

7.5.2. Transit survey

By probing 100-150 EECs for ozone, the transit survey is able to detect the age-oxygen correlation with high statistical power assuming life to be somewhat common ($f_{\text{life}} \gtrsim 50\%$) and the typical oxygenation timescale to be 1–10 Gyr. If life is very common ($f_{\text{life}} \gtrsim 80\%$), high statistical power can be achieved even if the average oxygenation timescale is as short as $t_{1/2} \sim 500$ Myr or as long as $t_{1/2} \sim 20$ Gyr.

Under the case where life is very common, the transit survey could place meaningful constraints on the oxygenation timescale. As shown in Figure 11, the survey can distinguish between scenarios where global oxygenation occurs very quickly ($t_{1/2} \sim 0.5$ Gyr) or at a more Earth-like pace ($\sim 3$ Gyr), but it will be difficult to accurately measure the oxygenation timescale if it is much longer than Earth’s ($\gtrsim 10$ Gyr), since no planets of that age exist in the sample. This is due in part to the high degeneracy between $t_{1/2}$ and $f_{\text{life}}$ - that is, if only a few oxygenated planets are found, it may be because life is uncommon, or because life is common but global oxygenation is very slow and has not yet had time to occur on most inhabited worlds.

7.6. Discussion

7.6.1. Detectability of oxygen through Earth’s history

In this section we consider all oxygenated planets to have the same $O_2$ and $O_3$ abundance as modern Earth. However, during the Proterozoic era (approx. 2.2 – 0.6 Gya), Earth had a partially oxygenated atmosphere with $p_{O_2} < 1\%$ (Lyons et al. 2014). If other inhabited planets do evolve like Earth, this suggests that many of them

planets under optimistic circumstances, a sample size of $> 50$ EECs is necessary, requiring either $\eta_\oplus \gtrsim 20\%$ or a smaller inner working angle than assumed here (3.5 $\lambda/D$).
Example 2: Results for 50-meter (equivalent area) transit survey

How many exo-Earth candidates are probed for $O_3$?

Could this survey detect the age-oxygen correlation?

Could this survey determine the oxygenation timescale?

(results for six random realizations of the survey)

Figure 11. Results for the transit survey in Section 7. (a) The number of EECs observed versus the observing time budget, assuming $\eta_\oplus = 7.5\%$ for G stars and cloudy atmospheres (solid). $2\sim4\times$ as many planets could be observed if clouds were neglected (dashed), or $1.5\sim3\times$ as many with clouds if assuming the higher SAG13 estimate of $\eta_\oplus = 24\%$ (dotted). As our baseline case, we set $t_{\text{total}} = 730$ d. (b) The statistical power to detect the age-oxygen correlation as a function of the astrophysical parameters in Equation 15. (c) The minimum number of EECs which must be characterized to achieve $80\%$ statistical power, with the necessary observing time budget $t_{\text{total}}$. (d) Distribution of possible values for the oxygenation timescale as measured by six random realizations of the survey under the optimistic assumption that $80\%$ of EECs are inhabited. Results are shown for fast (0.5 Gyr, left), Earth-like (3 Gyr, center), and slow (10 Gyr, right) evolutionary scenarios, with the truth values marked by a blue line.
may have 1–3 orders of magnitude less O\textsubscript{2} than modern Earth.

In our analysis, \( t_{1/2} \) is the typical timescale require for a planet to achieve a detectable amount of O\textsubscript{2} or O\textsubscript{3}. Even if Proterozoic Earth analogs are common and their oxygen is undetectable, our results should not be affected provided that they will eventually develop richly-oxygenated atmospheres like modern Earth’s. In this case \( t_{1/2} \) corresponds to the end of the Proterozoic (\( \sim 4 \) Gyr for Earth). On the other hand, it may be that inhabited and oxygenated planets are common but very few of them ever evolve beyond \( p\text{O}_2 = 0.1–1\% \), in which case \( t_{1/2} \) corresponds to the end of the Archean (\( \sim 2 \) Gyr for Earth). In this case, a survey aiming to detect the age-oxygen correlation would need to focus on a smaller number of targets with much deeper observations, and would likely need ultraviolet sensitivity to detect the deep O\textsubscript{3} Hartley band absorption which would have been detectable throughout the Proterozoic (Reinhard et al. 2017). For transit spectroscopy, ultraviolet sensitivity will be difficult to achieve in a sample of predominantly M stars, so to detect O\textsubscript{3} at Proterozoic-like levels will require the prioritization of G and K targets instead. A LUVOIR-like direct imaging survey targeting G and K dwarfs may be capable of detecting Proterozoic-like ozone levels for individual targets, but the sample size will still be too small unless both \( \eta_\oplus \) and \( f_{\text{life}} \) are large (\( \gtrsim 30\% \)).

### 7.6.2. Abiotic oxygen sources

We only consider planets on which O\textsubscript{2} is biologically produced - as it was in Earth’s history - but others have considered scenarios through which an Earth-sized planet near or within the habitable zone could acquire detectable levels of oxygen through abiotic processes (for a review, see Meadows et al. 2018). The oxygen in these models tends to initially derive from H\textsubscript{2}O or CO\textsubscript{2} dominated atmospheres shortly after the planet’s formation and can linger in the atmosphere long enough to serve as a potential “false positive” biosignature for next-generation observatories. In B20, by assuming the fraction of planets with abiotically produced oxygen to be independent of age, we show that these false positives will have a small impact on the detectability of the age-oxygen correlation provided that they are less common than Earth-like planets with biogenic O\textsubscript{2}.

In reality, atmospheres with abiotically-produced oxygen will evolve over time. On Earth, oxygen is continually produced in large enough quantities to overcome its substantial geological sinks. On planets where oxygen is, e.g., a remnant of primordial ocean loss, it would be depleted over time. This suggests a statistical test to determine whether oxygen is a reliable biosignature: if the fraction of EECS with oxygen decreases with age, this would suggest much of the oxygen to be of a primordial, abiotic origin.

Finally, it is plausible that both populations of oxygen-rich worlds exist in comparable numbers: one with abiotically-produced oxygen which diminishes over time, and another with biologically produced atmospheres which increases over time. Whether the Earth-like age-oxygen correlation could be detected would depend on the timescales of the two processes. For example, if most planets with abiotically produced oxygen lose it before 1 Gyr, and most planets with biogenic oxygen acquire it by 10 Gyr, then it should be possible to distinguish the two populations.

### 8. SUMMARY

We have presented Bioverse, a simulation tool designed to gauge the potential of future observatories to test statistical hypotheses about the formation and evolution of planetary systems and habitable worlds. To achieve this, Bioverse leverages statistically realistic simulations of nearby planetary systems, a survey simulator designed to produce data sets representative of different observatory configurations and survey strategies, and a hypothesis testing module to assess the information content of the data. We demonstrated two applications of our code.

In the first example, we determined whether a future direct imaging (15-meter diameter) or transit spectroscopy (50-meter equivalent diameter) survey could empirically test the concept of a habitable zone as well as measure its location and width. With samples as small as 15–20 EECS, we found that both surveys will be capable of testing the habitable zone hypothesis if habitable planets are common (\( \gtrsim 50\% \) of EECS), and that they can constrain the habitable zone’s width well enough to rule out very wide (e.g., 1–10 AU) or narrow (e.g., 1–1.1 AU). A survey which can characterize 60–70 EECS for atmospheric water vapor can test the habitable zone hypothesis even if habitable planets are less common (20–40% of EECS), but would be difficult to achieve with currently-envisioned direct imaging observatories. Our estimates suggest that this would be feasible for a large aperture transit survey, but the EEC sample size is sensitive to the impact of cloud cover (and other factors not considered here, such as stellar contamination (Rackham et al. 2018)).

In the second example, we expanded upon the age-oxygen correlation proposed in B20, finding that future surveys which aim to study the oxygen evolution of Earth-like planets must expect to characterize at least
∼ 50 EECs by detecting the presence of modern Earth-like \text{O}_2 or \text{O}_3 absorption. With a sample size of 100-150 EECs – if most of them are inhabited – a survey could begin to constrain the evolutionary timescale with meaningful precision, and could determine whether the oxygenation of Earth-like planets proceeds at an Earth-like pace (2–5 Gyr timescale) or much faster (∼0.5 Gyr). The ability to detect far-UV \text{O}_3 absorption will be beneficial if Proterozoic Earth analogs are common, but may not be necessary provided they eventually evolve to a modern Earth-like state.

The statistical power of either survey to test these hypotheses depends critically on the number of EECs detected, but recent evidence suggests that existing estimates of \( \eta_\oplus \) are too high (Pascucci et al. 2019; Neil & Rogers 2020). Assuming \( \eta_\oplus = 7.5\% \) for Sun-like stars, we found that an ambitious 15-meter mirror diameter imaging survey would likely detect 15–20 EECs. Such a survey may have high statistical power for studies of terrestrial planets in general (including those outside the habitable zone), but will only be able to test the habitable zone concept if most EECs are habitable, or if tracers of habitability other than \text{H}_2\text{O} absorption are considered. Unless \( \eta_\oplus > 20\% \), an imaging survey will probably not be able to study the oxygen evolution of truly Earth-like planets, though it might still offer constraints on how common such planets are (Checlair et al. 2020).

In this paper we discussed the statistical power to test hypotheses as a function of sample size given a single measurement for each target. Bioverse can also combine multiple measurements for each planet which trace the same underlying physical conditions (such as habitability), allowing surveys to achieve greater statistical sensitivity with limited sample sizes. For example, by incorporating measurements of planetary brightness and color in addition to \text{H}_2\text{O} absorption, imaging surveys may be able to test the habitable zone concept with smaller sample sizes - provided a hypothesis exists for how these properties should vary with orbital separation (e.g., Checlair et al. 2019). Similarly, if clouds make the detection of \text{H}_2\text{O} difficult for a transit survey, then stratospheric \text{O}_3 may offer an alternative tracer of planetary habitability (provided \text{O}_2 is predominantly produced by life).

With Bioverse, we aim to enable future space-based exoplanet surveys to test hypotheses including and beyond the examples explored here, and to emphasize the importance of population-level studies for next-generation exoplanet surveys. While target-by-target analyses of the closest planets will be valuable, population-level studies will reveal fundamental truths about the laws governing non-habitable, habitable, and inhabited worlds.

ACKNOWLEDGMENTS

We thank Chris Stark for providing a possible realization of the LUVOIR-A target list and Tad Komacek for providing the GCM model outputs used in this work. A.B. acknowledges support from the NASA Earth and Space Science Fellowship Program under grant No. 80NSSC17K0470. The results reported herein benefited from collaborations and/or information exchange within NASA’s Nexus for Exoplanet System Science (NExSS) research coordination network sponsored by NASA’s Science Mission Directorate. This research has made use of NASA’s Astrophysics Data System.

Software: dynesty (Speagle 2020), emcee (Foreman-Mackey et al. 2013), Matplotlib (Hunter 2007), NumPy (Oliphant 2006), PSG (Villanueva et al. 2018), SciPy (Virtanen et al. 2020)

REFERENCES

Apai, D., Milster, T. D., Kim, D. W., et al. 2019, The Astronomical Journal, 158, 83, doi: 10.3847/1538-3881/ab2631

Apai, D., Milster, T. D., Kim, D. W., et al. 2019a, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 11116, Astronomical Optics: Design, Manufacture, and Test of Space and Ground Systems II, 1111608, doi: 10.1117/12.2529428

Apai, D., Cowan, N., Kopparapu, R., et al. 2017, arXiv e-prints, arXiv:1708.02821. https://arxiv.org/abs/1708.02821

Apai, D., Bixel, A., Rackham, B. V., et al. 2019b, in Bulletin of the American Astronomical Society, Vol. 51, 141

Bean, J. L., Abbot, D. S., & Kempton, E. M.-R. 2017, The Astrophysical Journal, 841, L24, doi: 10.3847/2041-8213/aa738a

Berdiñas, Z. M., Rodríguez-López, C., Amado, P. J., et al. 2017, MNRAS, 469, 4268, doi: 10.1093/mnras/stx1140

Bixel, A., & Apai, D. 2020a, arXiv e-prints, arXiv:2005.01587. https://arxiv.org/abs/2005.01587

—. 2020b, AJ, 159, 3, doi: 10.3847/1538-3881/ab5222
## APPENDIX

### A. LIST OF SYMBOLS

Table 5. A list of common abbreviations and symbols used in this paper.

| Symbol | Description |
|--------|-------------|
| **Abbreviations** |  |
| EEC    | “exo-Earth candidate” (or “potentially habitable planet”); planets in the radius range $0.8(S/S_\odot)^{0.23} < R_p < 1.4 R_\odot$ |
| LUVOIR | Large UV/Optical/Infrared Surveyor ([The LUVOIR Team 2019](#)) |
| SAG13  | NASA’s Exoplanet Program Analysis Group Science Analysis Group 13 |
| PSG    | NASA/GSFC Planetary Spectrum Generator ([Villanueva et al. 2018](#)) |
| IWA, OWA | Inner, outer working angles of a coronagraphic instrument |
| MCMC   | Markov Chain Monte Carlo |
| **Stellar properties** |  |
| $d$    | Distance to star |
| $M_*, R_*, L_*$ | Mass, radius, and luminosity |
| $T_*$  | Effective temperature |
| $t_*$  | Age of star and planetary system |
| $a_{inner}, a_{outer}$ | Inner and outer edge of the star’s habitable zone |
| **Planet properties** |  |
| $M_p, R_p, g_p$ | Mass, radius, and surface gravity |
| $h$    | Atmospheric scale height |
| $P$    | Orbital period |
| $a$    | Semi-major axis |
| $a_{eff}$ | Semi-major axis scaled by the stellar luminosity, $a_{eff} = a(L_*/L_\odot)^{-0.5}$ |
| $\cos(i)$ | (Cosine of) orbital inclination |
| $b$    | Transit impact parameter, assuming a circular orbit |
| $\delta$ | Planet transit depth, $\delta = (R_p/R_*)^2$ |
| $\Delta\delta$ | Approximate transit depth induced by planet’s atmosphere, $\Delta\delta \sim 2(h/R_p)$ |
| $\zeta$ | Planet-to-star contrast ratio |
| $R_{est}$ | Estimated planet radius assuming Earth-like reflectivity (direct imaging only), $R_{est}/R_\odot = (\zeta/\zeta_\oplus)^{0.5}(a/1 \text{ AU})$ |
| **Simulated survey** |  |
| $D_{tel}$ | Telescope diameter or effective diameter (based on total light-collecting area) |
| $\lambda_{eff}$ | Effective wavelength of a spectroscopic measurement |
| $R_{*, ref}, T_{*, ref}$ | Radius and effective temperature of the reference star; $(R_{*, ref}, T_{*, ref}) = (5777 \text{ K}, 1 R_\odot)$ for the imaging survey, $(3000 \text{ K}, 0.15 R_\odot)$ for the transit survey |

Table 5 continued
Table 5 (continued)

| Symbol | Description |
|--------|-------------|
| $t_i$  | Amount of time required to characterize the $i$-th planet in a sample |
| $t_{ref}$ | Amount of time required to characterize an Earth twin orbiting the reference star with $a_{eff} = 1$ AU |
| $t_{total}$ | Time budget allocated to characterizing planets for a specific spectral feature (may overlap with observations at other wavelengths) |
| $\zeta_\oplus$ | Contrast ratio of the Earth with respect to the Sun, $\zeta_\oplus \approx 10^{-10}$ |
| $p_i$, $w_i$ | Observing priority and relative weight assigned to each planet, where $p_i = w_i/t_i$ |

**Hypothesis testing**

$x$, $y$

Independent and dependent variables in the simulated data sets

$h(\vec{\theta}, x)$

Alternative hypothesis describing the relationship between $x$ and $y$, to be compared to the null hypothesis

$\vec{\theta}$

Set of parameters which define $h$

$L(y|\vec{\theta})$

Likelihood function, described by Equation 9 or 10

$\Pi(\vec{\theta})$

Prior probability distribution of $\vec{\theta}$, described for each example in Table 3

$\mathcal{Z}$

Bayesian evidence in favor of the null or alternative hypothesis, computed by nested sampling

**Habitable zone hypothesis**

$a_{inner}$, $a_{outer}$

Inner and outer edges of the habitable zone in $a_{eff}$ space (i.e. for a Sun-like star)

$f_{\text{EEC}}^{H_2O}$

Fraction of EECs with atmospheric water vapor (assumed habitable)

$f_{\text{non-EEC}}^{H_2O}$

Fraction of non-EECs with atmospheric water vapor

**Age-oxygen correlation**

$f_{\text{life}}$

Fraction of EECs inhabited by life (regardless of $O_2$ content)

$t_{1/2}$

Oxygenation timescale; the time required for 50% of inhabited planets to undergo global oxygenation