Color Classification of Extrasolar Giant Planets: Prospects and Cautions

Natasha E. Batalha1, Adam J. R. W. Smith2, Nikole K. Lewis1,3,4, Mark S. Marley5, Jonathan J. Fortney2, and Bruce Macintosh6

1 Space Telescope Science Institute, Baltimore, MD 21218, USA; batalha@stsci.edu, natasha.e.batalha@gmail.com
2 Department of Astronomy & Astrophysics, University of California Santa Cruz, Santa Cruz, CA 95064, USA
3 Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD, USA
4 Department of Astronomy and Carl Sagan Institute, Cornell University, 122 Sciences Drive, 14853, Ithaca, NY, USA
5 NASA Ames Research Center, Moffett Field, CA 94035, USA
6 Department of Physics, Stanford University, Stanford, CA 94305, USA

Received 2018 June 2; revised 2018 July 18; accepted 2018 July 22; published 2018 September 14

Abstract

Atmospheric characterization of directly imaged planets has thus far been limited to ground-based observations of young, self-luminous, Jovian planets. Near-term space- and ground-based facilities like WFIRST and ELTs will be able to directly image mature Jovian planets in reflected light, a critical step in support of future facilities that aim to directly image terrestrial planets in reflected light (e.g., HabEx, LUVOIR). These future facilities are considering the use of photometry to classify planets. Here, we investigate the intricacies of using colors to classify gas-giant planets by analyzing a grid of 9120 theoretical reflected light spectra spread across different metallicities, pressure–temperature profiles, cloud properties, and phase angles. We determine how correlated these planet parameters are with the colors in the WFIRST photometric bins and other photometric bins proposed in the literature. Then we outline under what conditions giant planet populations can be classified using several supervised multivariate classification algorithms. We find that giant planets imaged in reflected light can be classified by metallicity with an accuracy of >90% if they are a priori known to not have significant cloud coverage in the visible part of the atmosphere, and at least three filter observations are available. If the presence of clouds is not known a priori, directly imaged planets can be more accurately classified by their cloud properties, as opposed to metallicity or temperature. Furthermore, we are able to distinguish between cloudy and cloud-free populations with >90% accuracy with three filter observations. Our statistical pipeline is available on GitHub and can be extended to optimize science yield of future mission concepts.

Key words: methods: statistical – planets and satellites: atmospheres – planets and satellites: gaseous planets – techniques: spectroscopic

1. Introduction

Atmospheric characterization of directly imaged planets has thus far been limited to ground-based observations of young, self-luminous, Jovian planets (Janson et al. 2013; Konopacky et al. 2013; Kuzuhara et al. 2013; Barman et al. 2015; Macintosh et al. 2015). However, near-term space- and ground-based facilities such as the Wide-Field Infrared Survey Telescope (WFIRST; Spergel et al. 2013), and the ELTs (e.g., Kasper et al. 2010) along with future concept direct imaging missions (HabEx, Mennesson et al. 2016; LUVOIR, Bolcar et al. 2015), could enable the detection of reflected light from cooler ($T_{\text{eff}} \sim 150$–300 K) exoplanets located several au from their host star.

The WFIRST Coronograph Instrument (CGI) is currently being refocused as a technology demonstrator. As it is still in the design phase, the CGI specifications have not yet been finalized, but the current design does include both photometric and spectroscopic observing modes (Balasubramanian et al. 2016; Cady et al. 2016; Noecker et al. 2016; Seo et al. 2016; Trauger et al. 2016). Since the exact number and wavelength bandpasses for the photometric filters have not been finalized, this work (as shown in Figure 1) uses the filter set originally defined for the WFIRST CGI (Spergel et al. 2013). Although WFIRST may not carry this full filter set, the wavelengths were strategically chosen to explore key molecular features typical of Jovian or Earth-like atmospheres (~0.45–1.0 μm) (e.g., Des Marais et al. 2002; Cahoy et al. 2010; MacDonald et al. 2018), and they are still representative of the likely capabilities of missions such as HabEx and LUVOIR (Bolcar et al. 2015; Mennesson et al. 2016).

Jupiter-sized planets located a few au away from their parent stars are particularly interesting objects for direct imaging in scattered light. They are large and, if cloudy, bright, even at a few au separation, and thus offer higher contrast ratios at larger angular separations than habitable zone terrestrial planets. Such giants are not, however, Jupiter twins. The cloud decks of planets somewhat warmer than Jupiter will progressively evaporate as first the uppermost NH3 and eventually even the deep H2O clouds evaporate (Marley et al. 1999; Sudarsky et al. 2000, 2003). The nature and detailed characteristics of the uppermost cloud deck define the reflected light continuum (Cahoy et al. 2010; MacDonald et al. 2018) and the depths of absorption bands. Because the effects of clouds are so important, Sudarsky et al. (2000) suggested that cloud type should be used to classify giants, with classes I through III representing the sequence from ammonia to water to no clouds.

To explore how well such planets could be characterized by reflected light spectra of the quality which might be expected from optical wavelength coronagraphic imaging, Lupu et al. (2016) coupled a reflected light model based on Cahoy et al. (2010) and an instrument noise model based on Robinson et al. (2016) to a nested sampling algorithm in order to demonstrate the ability to retrieve cloud properties, methane abundance, gravity, and cloud-top pressure. They concluded that the
presence of clouds and methane could be determined with high confidence, with clouds being even easier to identify (from the continuum shape) than the presence or absence of methane. Nayak et al. (2017) expanded on this work to explore how poorly constrained or unknown planet radii and planet–star phase angles could impact retrievals of high fidelity planet properties. Specifically, they determined that a signal-to-noise ratio ($S/N$) of 20 was required to retrieve accurate methane abundances at all phase angles.

Most recently, Lacy et al. (2018) used an independent reflected light and instrument noise model based on the work of Madhusudhan & Burrows (2012) and Nemati et al. (2017), respectively. They determined that for the most favorable known radial velocity targets, only exposure times of hundreds of hours would enable the measurement of planet radius and methane (if the phase functions and orbital parameters were known a priori).

With regard to terrestrial planets, Feng et al. (2018) explored retrievals of atmospheric and planetary properties for Earth twins as a function of $S/N$ and spectral resolving power. As expected from previous work on Jovian systems, they concluded that WFIRST, combined with a starshade, would provide limited astrophysical information of Earth-type planets because prominent spectral features, such as water vapor, ozone and oxygen, can only be weakly detected with $S/N = 20$ data.

The consensus among all of this work, regardless of planet-type, is that it will be challenging and time intensive to attain sufficient spectroscopic $S/N$ on many targets in the era of WFIRST and future space-based coronagraphs (Traub et al. 2016). Therefore, in light of the upcoming 2020 decadal survey, it would behoove the community to determine if additional observing techniques could aid in the characterization of exoplanets or the optimization of target selection. In particular, this sentiment has motivated the idea of leveraging photometric surveys as a means of selecting scientifically interesting targets for time-intensive follow up, and even as a means of broadly classifying planet types. The dominant noise source for coronagraph or starshade spectrophotographic instruments on small or medium telescopes such as WFIRST is often detector dark current (Nemati et al. 2017). As a result, exposure times scale with the number of pixels a spatial and spectral resolution element occupies—16 or more pixels for a typical integral field spectrograph. In narrow-band photometric imaging, the signal from a single absorption feature occupies only four pixels.

Because of how stellar photometry was leveraged to create the Hertzsprung–Russell diagram, Traub (2003) reasoned that the physical characteristics of a planet and its atmosphere could be derived from its colors. This work was based on an analysis of the $0.3–1 \mu$m reflected light spectra of the solar system planets, the Moon, and Titan, all at full phase. However, Cahoy et al. (2010) pointed out that the color of planets changes with phase and thus care would have to be taken in identifying planets based solely on color. Krissansen-Totton et al. (2016) conducted a comprehensive analysis of solar system bodies and modeled exoplanets, focusing specifically on separating a habitable Earth-twin spectrum from an uninhabitable planet. Out of their sample of spectra, the Earth’s reflectance spectrum was the only one to have a “U-shape”, caused by biogenic O$_2$. Therefore, they concluded that photometry, as a preliminary characterization tool, could be valuable.

However, the brightness in any given bandpass of a planet is a complex function of several parameters. Exoplanets will likely exhibit a great diversity of atmospheric properties not fully encompassed by the solar system sample. The most dominant parameters that affect the brightness profile are (1) the abundances of absorbing gasses, which can be studied via the metallicity, (2) the cloud height and scattering properties, which largely affect the brightness in the near-infrared, (3) photochemical haze abundances and their optical properties, (4) the relative importance of scattering versus absorption, which also depends on the cloud and haze optical depths, (5) the gravity and temperature, which will impact the scale height, (6) the phase angle, which determines how important forward and back scattering is, and (7) the bandpass of the observing filter.

With such a large parameter space and so many confounding variables, it is difficult to analyze all of these parameters with computationally intensive retrievals methods (e.g., Lupu et al. 2016). Yet, in the era of next-generation direct imaging missions, it will be important to numerically assess the influences of each of these parameters on our ability to yield scientifically interesting results.

Therefore, as opposed to using retrieval algorithms, we assess the intricacies of these effects by analyzing a grid of 9120 reflected light spectra spread across metallicity, pressure–

![Table: Approximate Wavelength (µm)]

| Approximate Wavelength (µm) |
|-----------------------------|
| 0.40 | 0.46 | 0.48 | 0.54 | 0.59 | 0.62 | 0.65 | 0.73 | 0.77 | 0.78 | 0.79 | 0.83 | 0.84 | 0.86 | 0.89 | 0.91 | 0.94 | 0.99 |
| Jovian | K | CH$_4$ | CH$_4$ | CH$_4$ | Na | CH$_4$ | H$_2$O | CH$_4$ | K | CH$_4$ | CH$_4$ | H$_2$O | CH$_4$ | CH$_4$ | CH$_4$ | H$_2$O | CH$_4$ |
| Earth | O$_3$ Chappuis bands | O$_2$ | H$_2$O |
| WFIRST | Imaging | 506 | 575 | 661 | 721 | 883 | 940 |
|   | Spectroscopy | 660 | 770 | 890 |

Figure 1. Optical absorption features seen in reflected light spectra of gas-giant exoplanet models and Earth compared to WFIRST observing modes. Absorption features for Jovian systems are taken from Cahoy et al. (2010), and absorption features for Earth are taken from Krissansen-Totton et al. (2016). WFIRST observing modes are labeled by the center of their respective bandpass in units of nm. This work only focuses on the imaging modes.
temperature profiles, cloud properties, and phase angles. This large data set allows us to leverage several multivariate classification algorithms, which are computationally effective, and give numerical insights into our ability to classify giant planets by physically motivated groups using photometric observations. In doing so, we answer the following questions:

1. Is there a strong correlation between atmospheric properties and WFIRST-like optical filters?
2. If so, can those correlations be leveraged to create color–color diagrams that separate planets into any physically motivated groups?
3. If not, do any more sophisticated “machine learning” algorithms succeed in correctly classifying planets given our theoretical training set?
4. And, are there certain bandpasses (or a minimum number of bandpasses) required for successful classification?

Section 2 describes our forward modeling methodology for creating pressure–temperature profiles, chemical profiles, clouds, and reflected light spectra. Section 3 describes our statistical methodology for classifying planet types. Section 4 contains our results, and lastly, our discussion and concluding remarks are in Sections 5 and 6.

2. Modeling Approach

There are three components required for the computation of reflected light spectra: (1) the temperature–pressure \((T(P))\) profile, (2) the composition of the atmosphere as a function of altitude, and (3) the scattering properties of the clouds. We vary each of these components independently from each other, neglecting any feedback between cloud production and atmospheric temperature. Additionally, all our models are run for a single gravity planet \((25 \text{ m s}^{-2})\) that orbits a Sun-like star. In Section 5, we discuss the impact of expanding our analysis to incorporate additional gravities and stellar types.

2.1. The Temperature and Chemical Structure

Our \((T(P))\) profile is computed using the radiative-convective model initially developed by McKay et al. (1989) and modified for irradiated gas-giant planets in Marley & McKay (1999), Marley et al. (2002), Fortney et al. (2005, 2008). It has also been used in many other comparisons with data (e.g., Marley et al. 1996; Roellig et al. 2004; Mainzer et al. 2007; Saumon & Marley 2008). The model requires, as input, an internal heat flux and the incident flux from the parent star. We keep the internal heat flux equivalent to 150 K (similar to that of Jupiter Fortney et al. 2007) for all the cases considered here. For incident flux, we use semimajor axis as a proxy and explore 10 different cases, ranging from 0.5 to 5.0 au from a Sun-like star. Throughout the text we refer to planet–star distance as “distance,” and it can be assumed this also indicates a variation in temperature–pressure profile. The model then iterates to radiative-convective equilibrium using the two-stream source function technique of Toon et al. (1989).

We compute the composition of the atmosphere by using the chemical equilibrium grid most recently updated in Lodders & Fegley (2006). We consider 6 different metallicities ranging from 1 to 100× solar using the initial elemental abundances from Lodders (2003). The optically relevant gases are \(\text{H}_2\), \(\text{H}\), \(\text{CO}_2\), \(\text{CO}\), \(\text{CH}_4\), \(\text{NH}_3\), \(\text{N}_2\), \(\text{H}_2\text{S}\), \(\text{Na}\), \(\text{K}\), \(\text{TiO}\) and \(\text{VO}\). This also includes continuum opacity from \(\text{H}_2–\text{H}_2\), \(\text{H}_2–\text{He}\), \(\text{H}_2–\text{H}\), and \(\text{H}_2–\text{CH}_4\). The opacities for these gases were calculated in Freedman et al. (2008) and updated in Freedman et al. (2014).

MacDonald et al. (2018) created a similar grid of parameterized \((T(P))\) profiles by fitting to the self-consistent model used in this work. We verified that the \((T(P))\) profiles derived in this work are in agreement with the parameterized profiles from MacDonald et al. (2018).

Figure 2 shows several representative \((T(P))\) profiles for 1× and 3× solar metallicities for the full range of semimajor axes explored here. We emphasize that all our models are computed at a gravity of \(25 \text{ m s}^{-2}\) (gravity of Jupiter). This serves as an initial staring point for assessing the use of colors in classifying planets. In Section 5, we discuss how our results may be further impacted by different gravities/unconstrained masses.

2.2. The Cloud Model

The full grid, partially seen in Figure 2, contains a wide parameter space in temperature. Therefore, depending on the case, we must consider the condensation of different cloud species. On the cooler end, toward 5 au, we have the potential formation of \(\text{H}_2\text{O}\) clouds (Marley et al. 1999; Sudarsky et al. 2000; Morley et al. 2012). As we go toward hotter systems we must consider the potential condensation of several other species: \(\text{NH}_3\), \(\text{KCl}\), \(\text{ZnS}\), \(\text{Na}_2\text{S}\), \(\text{MnS}\), \(\text{Cr}\), \(\text{MgSiO}_3\) also discussed in Morley et al. (2012).

We calculate the cloud properties using a Mie scattering treatment of particle sizes calculated from a widely used cloud model Ackerman & Marley (2001). Specifically, the cloud mass and particle sizes are calculated by solving

\[
K_{zz} \frac{\partial q}{\partial z} + f_{sed} w_q = 0
\]

where \(K_{zz}\) is the vertical eddy diffusion coefficient \((\text{cm}^2 \text{s}^{-1})\), \(q_i\) is the volume mixing ratios of the total condensing species (condensate plus vapor form), \(q_e\) is the volume mixing ratio of just the condensate, and \(w\) is the convective velocity. Lastly, \(f_{sed}\) is a parameter generally used to tune the sedimentation efficiency of the atmosphere. High values of \(f_{sed}\) (i.e., >1) create thin cloud profiles with large particles, which are...
generally optically thin. Low values of \( f_{\text{sed}} \) (i.e., <1), result in the opposite—thick cloud profiles with small particles, which are generally optically thick. For reference, reflected light observations of Jupiter are well-fit with \( f_{\text{sed}} = 3 \) (Ackerman & Marley 2001; Cahoy et al. 2010). For each \( T(P) \)-metallicity point, we compute cloud profiles for \( f_{\text{sed}} = 0.01–6 \), where lowest bound \( f_{\text{sed}} \) values are motivated by observations of hot Jovian worlds (Demory et al. 2013).

Ultimately, the cloud model calculates the wavelength-dependent optical depth, the single scattering albedo and the scattering asymmetry factor, which are the last components needed for the the reflected light calculation.

### 2.3. Reflected Light: The Albedo Model

The wavelength-dependent geometric albedo is the ratio of reflected planetary flux at full phase to the flux from a perfect Lambert disk. Realistically, direct imaging observations of exoplanets will never yield full phase observations because full phase occurs too spatially near the parent star. Our albedo model was originally constructed for full phase Titan observations (McKay et al. 1989), but has since gone through substantial developments. Marley & McKay (1999) and Marley et al. (1999) extended the model to giant planets (both solar system and exoplanets). Later, Cahoy et al. (2010) added the ability to compute phase-dependent albedo spectra, crucial for the study of directly imaged planets. The radiative transfer of the model closely resembles our \( T(P) \) model (since the models both originate from McKay et al. 1989). Therefore, the radiative transfer calculations use the same Toon et al. (1989) scheme described in Section 2.3. Likewise, the opacities considered here are identical to those in the \( T(P) \) model. We feed the results of each \( T(P) \)-metallicity-cloud profile to compute albedo spectra for phases between 0° and 180° in increments of 20°. This completes the last component of the full albedo grid.

The top panel in Figure 3 shows the effect of decreasing temperature on the geometric albedo for a cloud-free solar metallicity case. As found by Sudarsky et al. (2000) and Cahoy et al. (2010), starting at high temperatures \( (a_s = 0.5 \mu \text{au}) \), the spectra of 1× solar Jovian worlds are dominated by alkali opacities and Rayleigh scattering. The clear atmosphere allows photons to penetrate deep in the atmosphere where temperatures exceed 800 K and Na and K are in relatively high abundance due to chemical equilibrium. As temperatures decrease (toward \( a_s = 2 \mu \text{au} \)), alkali opacity becomes less dominant because cooler temperatures at chemical equilibrium favor the production of \( \text{H}_2\text{O} \) and \( \text{CH}_4 \). Na and K are lost into sulfide and salt condensates (Lodders 1999; Morley et al. 2012). \( \text{H}_2\text{O} \) and \( \text{CH}_4 \) prevent photons from penetrating deep into the atmosphere and overall, results in a brighter albedo spectrum. From \( a_s = 3–5 \mu \text{au} \), there is little observable effect on the spectrum due to relatively minor changes in the \( T \) \((P) \) profile (see Figure 2). The presence or absence of Na and K lines short of 0.41 \( \mu \text{m} \) would be strong temperature indicators. However, they are outside the current proposed range of WFIRST bandpasses.

The bottom panel in Figure 3 shows the effect of increasing metallicity on the geometric albedo for a cloud-free, \( a_s = 0.5 \mu \text{au} \) case. Enhancing metallicity from solar values increases the total opacity of the atmospheres, which leads to a decrease in the atmospheric brightness (or deeper molecular bands). The biggest spectral changes occur from 1 to 10× solar, after which, the overall albedo spectra remains relatively unchanged. Ultimately this will lead to challenges in accurately constraining metallicities above 10× solar, which we discuss in Section 4.

Figure 4 shows the effect of increasing the sedimentation efficiency, \( f_{\text{sed}} \), in our cloud model. Lower values of \( f_{\text{sed}} \) lead to more vertically extended, optically thick clouds, which can be seen in Figure 4. In this 5 au, solar metallicity case, as \( f_{\text{sed}} \) increases, \( \text{H}_2\text{O} \) clouds begin to dominate the spectrum and drastically increase the brightness past ~0.6 \( \mu \text{m} \). This brightness peaks at \( f_{\text{sed}} = 1 \) then begins to decrease for \( f_{\text{sed}} < 1 \), as thick clouds decrease the Rayleigh scattering efficiency.

Lastly, Figure 5 shows the effect of our albedo model’s phase dependence, via work done by Cahoy et al. (2010). As expected the brightness of the atmosphere decreases with decreasing phase. Not pictured, is the degenerate effect between phase and unknown planet radii, not explored in this work (see Nayak et al. 2017 for a comprehensive analysis of those degeneracies).

### 3. Statistical Approach

Our full theoretical population includes: 6 metallicities ×10 \( P(T) \) profiles × 8 cloud profiles × 19 phase angles = 9120 model planetary spectra. We assume that each of these 9120 planets are observed with the six imaging filters depicted in Figure 1. We compute absolute magnitudes and colors for all possible combinations of filters.

This number of planets is of course grossly over estimating what a mission such as WFIRST (or even future mission concepts such as LUVOIR or HabEx) would provide observationally. However, we specifically aim to determine whether or not colors can be used under any scenario to classify planets, as the field of stellar astrophysics has done with stars.

Under these assumptions, our statistical problem is a classic supervised classification multivariate analysis problem (Feigelson & Babu 2012), because we assume that we know the properties of each of the 9120 planets. The ultimate goal is to determine if any modern algorithms can correctly determine physical planet properties given photometric observations in WFIRST-like filters. This is a fundamentally different approach than that of Krissansen-Totton et al. (2016). In their analysis, the goal was to determine if there were any optical filter combinations that could separate an Earth-twin from every other type of hypothetical exoplanet. Therefore, their algorithm chose filters that maximized distance in color–color space from Earth to their theoretical population.

There are dozens of open-source algorithms widely used in multivariate classification problems (Johnson & Wichern 2002; Izenman 2008; Feigelson & Babu 2012; Marsland 2015, and refs. therein). Although traditionally these algorithms are available through \( R \) packages, Python also has a widely used package for machine learning problems called scikit-learn (Pedregosa et al. 2011), which we adopt here. All of the code used in our statistical analysis is publicly available on Github\(^7\) and we footnote specific Python functions used throughout the text, when applicable.

---

\(^7\) https://github.com/natashabatalha/color-color
We first determine the effectiveness of the classification algorithms by using the $k$-fold\textsuperscript{8} cross validation\textsuperscript{9} method (Kohavi 1995). In this method, for each algorithm we: (1) randomly separated the original sample into $k$ equal groups, (2) keep one of the groups as the validation data for testing, while the other remaining groups are used as the training data, and (3) compute the mean predictive accuracy given the training and validation set. These three steps are then repeated $k$ times so that each group is used at least once as the validation data. In this analysis we also ensure that the same random seed is used to split the data into identical groups each time an algorithm is evaluated. We test our cross validation routine with values of $k$ from $k = 10$–100.

We briefly describe each of the algorithms as they relate to the problem of directly imaged planets, with further technical
1. **Linear Discriminant Analysis** (LDA) is very similar to a Principal Component Analysis (PCA) in its attempt to find the component axes that maximize the variance of the data. PCA, however, is an unsupervised algorithm and therefore ignores class labels (in our problem class labels would be information such as metallicity, temperature, cloud properties). LDA is supervised, which means it considers class labels and attempts to project the data into directions that maximize the separation between these classes. Those directions are called the linear discriminants. In Section 4.3 we will compute linear discriminants and assess whether they are successful in separating out planet populations observed with *WFIRST*-like photometric filters.

2. **K-Nearest Neighbors (KNN)** is commonly used when nonlinearities exist between classes in a data set. For example, if two different low metallicity populations appeared in distinct regions in color–color space (e.g., cloud-free versus cloudy), then a linear method would readily fail. The KNN algorithm uses the training data to compute on-the-fly calculations about the similarity between any given input and the training data. Therefore, KNN would theoretically be able to classify metallicity given varying cloud profiles, if the different cloud profiles created distinct populations of metallicities. We use the grid search method in scikit-learn to optimize the number of nearest neighbors to use in addition to testing its sensitivity to standard values, such as the Python package’s default value.

3. **Classification and Regression Trees** (CART) recursively splits the data by creating a prediction model at each partition (Breiman 2017). At each partition, the algorithm aims to minimize the inhomogeneity of the data set, called impurities. The two methods to evaluate impurities included within scikit-learn are (1) the Gini method and (2) the information gain methods. We test both to ensure insensitivity to our result. The result of a CART analysis can be represented graphically in the form of a decision tree. We tune the model using the same method described in KNN. Additionally, we take extra precaution to avoid highly complex and unintuitive decision trees that over-fit the data (discussed further in Section 4).

4. We also analyzed the results of a Multi-class Logistic Regression, and a naive Bayes classifier, and a Support Vector Machines but do not present the results because the models failed to classify planets by physical properties under all scenarios with a success rate of \(\leq 1\%\).

We test each algorithm under different observing scenarios: if two filters were available, three filters, four filters, etc., all the way to a scenario where we have all six photometric points. For each of those scenarios, we run the classifier for every single possible combination of filters. This allows us to determine what the best filter combinations are for classifying planets by metallicity or other physical properties of interest.

The metric we use to assess the success of an algorithm is the mean and standard deviation of the accuracy after the \(k\)-fold cross validation test. It should be noted that there are many other metrics to evaluate a multi-class classification algorithm (e.g., logarithmic losses and confusion matrices). We address these after our initial, top-level, analysis of all the algorithms.

## 4. Results

Before interpreting the results of the multivariate analysis, it is useful to assess the data using simpler statistical techniques in order to gain intuition for the full data set. We begin our results by presenting a correlation analysis to assess the sensitivity of the optical photometric bands to our physical planet properties of interest. Then, we present our color–color diagrams with the filter combinations expected to separate the planets into physically motivated groups (e.g., by metallicity). Lastly, we present the results of the multivariate analysis described in Section 3.

### 4.1. Sensitivity of Optical Photometric Bands to Planet Properties

Figure 6 shows the Pearson correlation matrix computed with Python’s pandas package (McKinney 2010). Correlation matrices are used to assess dependencies between variables. Therefore, it allows for the quick visualization of any dependencies between observational filters and physical planet properties. Filters where correlations exist could be helpful in separating out planet populations in color–color space. Because we are purely interested in the relative strength of correlation with respect to the physical parameters of interest, we show the absolute value, for simplicity. We emphasize that this analysis is purely used to gain intuition for the full data set.
The top panel of Figure 6 shows the strength of correlations for a cloud-free subset of the data. Metallicity exhibits strong correlations with filter combinations that include f506 or f575 minus f721 or f883. Distance/temperature exhibits the strongest correlations with filter combinations that include f721, f883, and f940. It is also sensitive to f506–f575 where the structure of the albedo spectrum is largely dictated by the presence or absence of water absorption. Phase also exhibits the strongest correlation to f506–f575. The presence of strong correlations in the cloud-free subset of the data suggests that classifying planets given WFIRST-like filters might yield fruitful results. However, the bottom panel of Figure 6 alludes to the main challenge we will encounter in classifying planets.

The bottom panel of Figure 6 shows the same Pearson correlation matrix with the full data set (cloud-free plus cloudy). With the addition of f_sed, all previous strong correlations become muted. Additionally, f_sed now holds stronger correlations than metallicity between f506 or f575 minus f721 or f883. Therefore, we should expect any behavior in color-color diagrams to be primarily driven by cloud properties and only weakly driven by metallicity and temperature/distance.

Color–Color Diagrams for Exoplanets

Moving forward, we display our color–color diagrams with the combination of three filters: f506–f575 and f575–f883 in order to illustrate the general trends of color–color diagrams for planets. We choose f883 because of its strong correlations with f_sed, and metallicity; however, we note that there are very few emergent photons at those wavelengths (see Figure 3 at ~940 nm) and the S/N is expected to be relatively low (Lacy et al. 2018). Note that our full color grid is also available for download along with a visualization application to explore additional color–color parameter space and recreate Figure 7.18

The leftmost panel in Figure 7 shows the cloud-free, zero phase, subset of our simulated grid of directly imaged planets. Temperature/distance is separated on the y-axis by f506–f575 because toward high temperatures, CH4 does not dominate the structure of the albedo spectrum. Instead, Rayleigh scattering creates a deeper slope between f506 and f575. Metallicity, on the other hand, is separated on the x-axis by f575–f883 because the total atmospheric opacity darkens the planet in the f575 bandpass at a faster rate than that at the f883 bandpass. Overall, metallicity and temperature/distance are well separated in color-space, with the exception of planets with a_t ≥ 3 au and [M/H] ≥ 30 × solar (an effect that is also seen in Figure 3 and Cahoy et al. 2010).

Adding varying phases (second from the left in Figure 7) adds an additional level of complexity. However, general trends in metallicity and temperature/distance still remain overall unchanged—majority of low metallicity systems still have largest values of f575–f883 and majority of a_t ≤ 1 au still have the lowest values of f506–f575.

Next, with the addition of all the cloud profiles (third plot from the left), the color–color plot takes on a new structure. The cloud-free sample is now degenerate with additional planets with large f_sed (thin cloud profiles). Additionally, there is a new population of planets with f506–f575 > 0. This theoretical population consists of models with f_sed values of f_sed = 0.01, and distances of a_t ≤ 1 au. Further investigation revealed these high temperature, low f_sed cases, produced vertically thick alkali clouds of Na2S and ZnS with small particles. We discuss this behavior further in Section 5.1.

Classifying Planets by Colors: Where It Works and Where It Fails

Our results in Sections 4.1 and 4.2 indicate that although strong correlations exist between optical filters and physical planet properties, those correlations are visibly diminished with a large population of planets with variable cloud profiles. Here,
we use the statistical algorithms described in Section 3 to quantitatively assess whether or not this remains true.

We first test the algorithms on a purely cloud-free sample to determine whether or not they succeed in the simplest of cases. Our results for the LDA and CART analyses are summarized in Figure 8. We only show results for the algorithms that had mean accuracy of greater than 50%. SVM had the poorest performance in the cloud-free sample, with less than 1% accuracy. KNN reached a highest mean accuracy of 41% in the case where all 6 filter combinations were available. LDA and CART both succeeded in classifying planets by metallicity with greater than 90% accuracy when a cloud-free sample was used. Additionally, they both gained the greatest accuracy with the addition of a third filter. Four filters did increase the accuracy of LDA, but not CART. The difference in performance of each algorithm depends largely on the nonlinearity of the problem (e.g., whether an increase in metallicity is always linearity correlated with a photometric point). Some classifiers, such as SVM and logistic regression, depend more heavily on linear features in the data as opposed to CART, for example. CART has the disadvantage that it is susceptible to over-fitting the data. We take care to ensure our analysis is incentive to this by scaling back the maximum levels allowed in the decision tree, discussed below.

Figure 8 also shows the optimized filter selection of each case study. For the cloud-free case, both algorithms agree that the filters at 506 and 883 nm are highest priority for planet classification. It is interesting to note that they do not agree on the third highest priority filter. LDA prefers the filter at 575 nm, while CART prefers the filter at 661 nm. However, both algorithms agree that the top four highest priority filters for planet classification in a cloud-free sample are the filters at: 506, 883, 575, and 661 nm (approximate in that order). These results intuitively agree with the results of Figure 6.

Upon introducing the full sample, we first tune/retrain the model. Figure 8 again, shows the results of the LDA and CART. SVM still performs the poorest with less than 1% accuracy. KNN again only reaches an accuracy of 40% when all six photometric points are used. LDA, which was successful before, now can only classify planets by metallicity with ∼20% accuracy. CART has the highest performance out of all of the algorithms, however still only attains a maximum success rate of 55%. This analysis suggests that classifying planets via photometric analysis of colors is incredibly difficult due to the variety of cloud compositions.

We also explore our ability to classify by metallicity if the cloud-free population is excluded from the population entirely. This is motivated by the prevalence clouds in exoplanet (e.g., Sing et al. 2016) and giant solar system planet observations (e.g., Lewis 1969; Baines et al. 1995). However, limiting ourselves to just the giant planets that have clouds, we only can attain a classification accuracy of 45%, suggesting that the prevalence of clouds inhibits our ability to accurately classify by metallicity with color observations alone.

Figure 9 graphically shows how and when the classifications scheme breaks down. The top panel shows our population of planets separated by the linear discriminants computed in the LDA for the cloud-free sample of planets at a single phase. The first linear discriminant (LD) is successful in separating all metallicities. The heat map to the right shows the confusion matrix, another metric for evaluating machine learning algorithms. The confusion matrix dictates, for each class, what fraction of planets we have correctly predicted. In this cloud-free, Phase = 90° sample we have correctly predicted nearly 100% of cases. Next we add the full sample of phases and rerun the LDA. The first LD now has trouble separating metallicities greater than 30 × solar and the second LD cannot further help distinguish between those cases. This means the predictive power of the algorithm is decreasing. With the addition of the full cloudy sample the first LD and the second LD fail to separate planets by metallicity.

Next, we repeat the analysis above in order to group planets by cloud profile ($f_{sed}$) and planet–star separation. We are unable to classify by planet-separation with higher than 47% accuracy. For cloud profile, with the full sample of planets, CART is able to classify planets by $f_{sed}$ with 83% accuracy with 3 filters. Our optimization of the CART analysis yielded a 16-level decision tree. Decreasing the maximum depth to five yields a lower accuracy of 75%, yet this is still a much higher performance than our ability to classify populations based on metallicity. Lastly, using a different criterion for distinguishing impurities (information gain versus Gini impurity) still yields an accuracy of 80% in the full sample. In order to yield this

![Figure 8](image-url) The table summarizes the results of the two most successful algorithms out of the several that were tested: linear discriminant analysis (LDA) and classification and regression trees (CART). Cells are colored by the success rate (or % accuracy) of the classifier. Each algorithm was run under different observing scenarios: observations in two filters, three filters, etc. The numbers in each cell are the optimized set of filter choices (nm) that yielded the maximum % accuracy. Main point: classification algorithms yield high accuracy for a cloud free sample, only, with at least three filter observations.
In order to verify that these results are insensitive to filter selection, we adopt the "optimized" filters from Krissansen-Totton et al. (2016). With the new set of filters our results do not change. That is, we are (1) still unable to classify our directly imaged planets by metallicity unless we specifically select a cloud-free population, (2) able to classify the full sample by clouds with an accuracy of 75% with CART, and (3) unable to classify by planet–star separation/temperature.

4.3.1. Binary Classification: Cloudy or Not?

The results depicted in Figures 8 and 9 suggest that determining accurate physical parameters such as metallicity, cloud composition, temperature, or phase, is unattainable due to the complex structures that diverse cloud species introduce to the full sample. However, photometry could still be potentially identifying targets for more time-intensive follow up. As Figure 4 shows, very cloudy cases exhibit little or no spectral features, which would not be ideal candidates for follow up. In this case, classifying giant planets based on whether they are cloudy or cloud-free would be very useful. We find that a binary supervised classification attains a maximum accuracy of 97% with three filters. Therefore, photometric observations will valuable for identifying targets for more intensive spectroscopic follow up.

5. Discussion

5.1. Population of Planets at Low \(f_{\text{sed}}\)

Figure 7 shows a unique population of theoretical planets with very bright red (883 nm) components, as compared to the blue (506 nm). Here we further explore that population. The population consists of theoretical planets with very low sedimentation efficiency \(f_{\text{sed}} = 0.01\) and high temperatures \(a_i < 1\) au. Figure 10 shows the wavelength- and vertical-dependent cloud profiles for the alkali clouds (Na2S and ZnS) that form in this region of parameter space. At \(f_{\text{sed}} = 1\), the clouds become optically thick at high pressures (1 bar) and are vertically thin (only 2 dex in pressure). However, at very low \(f_{\text{sed}} = 0.01\) the particles become small, and the clouds become vertically extended, more closely simulating the effect of hazes.
Figure 10. Vertical- and wavelength-dependent cloud profiles for Na$_2$S and ZnS at different values of $f_{sed}$. The top row shows the extinction coefficient for each cloud species, used to describe their absorbing properties. Main point: hot temperature clouds of Na$_2$S and ZnS strongly absorb short of 0.6 μm. At low $f_{sed} = 0.01$, where small particles create vertically thick clouds, this results in spectra with suppressed blue light (see the the arm in Figure 7).

Figure 10 also shows the extinction coefficient of Na$_2$S (Montaner et al. 1979; Khachatryan et al. 2009) and ZnS (Querry 1987). The extinction coefficient, also referred to as the imaginary part of the index of refraction, represent the absorbing power of each cloud species. The absorbing power of sulfur alkali species strongly suppress the blue part of the reflected light spectra, a result that has also been seen in studies of sulfur produced hazes (Gao et al. 2017).

Thus far, low values of $f_{sed}$ have only been observed in close-in hot Jupiters (Demory et al. 2013). Observations of Jupiter in reflected light suggest that widely separated exoplanets might be limited to $f_{sed} \geq 3$. Therefore, we additionally explore the feasibility of classifying a sample that only consists of $f_{sed} \geq 3$. We find that classifying by metallicity and by cloud with three imaging filters still results in a maximum accuracy of 48% and 62%, respectively. The accuracy does not increase when the sample of $f_{sed}$ is confined because small spectral changes introduced by vertically thin clouds are photometrically degenerate with changes in phase, temperature, or metallicity. See also MacDonald et al. (2018) for an exploration of how spectral morphology depends upon $f_{sed}$.

5.2. Where does the Solar System Fit in the Population?

One question not yet addressed is where the solar system fits within our giant planet population. Figure 11 shows the solar system planets in the color space chosen by Krissansen-Totton et al. (2016) to separate biogenic Earth from lifeless planets. Their filter choice is specifically sensitive to the "U-shape" created by Rayleigh scattering and ozone absorption. Disregarding our population of giant planets, the solar system planets do appear to exist in unique color spaces. The ice giants, Neptune and Uranus are brighter in the blue, and darker in the red, as compared to Jupiter, Saturn and Titan. The terrestrial planets are bright in the red, with Earth having an extra blue component from its unique "U-shape." However, when the additional population of giant planets is considered, the unique separation of these planets is washed out. Many of our giant planet cases with $f_{sed} \leq 1$ are easily degenerate in color space to Mars and Venus. While there are significantly fewer "false-positives" near Earth, a more realistic partial cloud coverage model of Earth (from Krissansen-Totton et al. 2016) is also synonymous in color space to many of our giant planet models.

Specifically, at $f_{sed} \sim 0.3$ and at $d < 1$ au: (1) temperatures are high enough so as to avoid formation of bright water clouds, (2) sedimentation efficiency is low enough to obscure alkali opacity features, and (3) sedimentation efficiency is high enough to retain the downward slope Rayleigh opacity. These three features result in the classic "Earth-like" Rayleigh slope, but of course lack a large upward slope from ozone absorption. However, when ozone is partially obscured by clouds (i.e. Earth w/partial cloud coverage in Figure 11), low resolution photometric observations could result in degeneracies. Additional intricacies could further confound this parameter space.

5.3. Additional Intricacies: Unknown Gravity, Potential Hazes, and Composition

The three parameters not explored in this work are the effect of poorly unconstrained or unknown gravities (Nayak et al. 2017; Lacy et al. 2018), potential hazes (Gao et al. 2017), and unknown chemical compositions (e.g., non-solar C/O ratio).

Increasing radii is degenerate with decreasing phase (see Figure 1 Nayak et al. 2017). Gravity alone also impacts the overall structure of the reflected light spectrum (see Figure 5 MacDonald et al. 2018). Both Nayak et al. (2017) and (Lacy et al. 2018) were able to constrain radii to relatively high fidelity with spectroscopy. This is partly because at higher (more crescent) phases, the atmospheres becomes forward scattering, and the effect of clouds become less dominant. However, photometry will be much less forgiving, even with three or four filter observations. Adding additional gravities to
our grid will ultimately decrease the accuracy of our classification algorithms.

Gao et al. (2017) investigated the effects of sulfur hazes in giant exoplanet atmospheres in reflected light and found that haze masses drastically alter a planet’s geometric albedo spectrum. These sulfur hazes mimic the behavior of our very low $f_{\text{sed}} < 0.03$ cases with sulfur-based alkali clouds (see Figure 10). Therefore, accounting for additional hazes would not drastically change the accuracy of our classification algorithm when classifying by cloud. Instead, they would be classified in the same grouping as our low $f_{\text{sed}}$ cases.

Lastly, compositions that deviate significantly from solar C/O would also change the accuracy of our classifiers if new molecular features became dominant. At the temperatures explored in this work, the chemically favorable molecule, regardless of C/O is expected to be CH$_4$ and H$_2$O (Madhusudhan 2012). The atmospheric H$_2$O abundance has not yet been measured on Jupiter, so detailed studies of this parameter are premature. Nevertheless, we would not expect varying C/O to largely change the results presented here.

5.4. Prospects for Filter Optimization Techniques

Although these additional sources of uncertainty have the potential to degrade our ability to classify planets, there are further steps we can take to improve our ability to accurately classify planets in color space. There is currently a large parameter space to explore in filter choices. The WFIRST filters have not yet been solidified. Therefore, our pipeline can be used to determine the optimal filter design in the era of WFIRST and future mission concepts.

In order to test the feasibility of using our pipeline for this type of analysis, we determined whether or not there was any 5% bandpass filter from 0.5 to 1 $\mu$m that could be added to a 575 nm photometric point to greatly increase classification accuracy. This was specifically chosen because the WFIRST science investigation teams are currently considering a 575 nm engineering filter in conjunction with just one more commissioned filter. With just two filters, we did not find any combination of 575 nm–a second filter that greatly improved the prospects of classifying our giant planet population. Regardless of the additional filter from 0.5 to 1 $\mu$m, we achieved a classification accuracy of just 25% when classifying by metallicity and just 50% when classifying by $f_{\text{sed}}$.

This result generally agrees with Figure 8, which shows that the addition of a third filter is needed to attain a relatively high classification accuracy. This simple parameter space analysis reiterates the importance of having at least three filter observations. Additionally, it is possible that decreasing the bandwidth (from 5%) and increasing the wavelength range to explore would yield a more favorable optimized filter selection. We leave a comprehensive analysis of these intricate choices and the additional uncertainties described above for a future analysis.

6. Conclusion

By analyzing a grid of 9120 theoretical gas-giant reflected light spectra, we explored our ability to classify giant planets into physically motivated populations. Specifically, we were interested in determining if the proposed WFIRST photometric bins could classify giant planets by either their temperature, metallicity, or cloud properties. In order to do so, we used several multivariate classification algorithms. Out of the six algorithms tested, only the LDA and CART yielded fruitful results. However, neither algorithm was able to yield very high accuracy when classifying by physical planet parameters.

In accordance with the questions posed in Section 1, we found that:

1. There is a strong correlation between atmospheric properties and WFIRST-like optical filters (Figure 6). However, these correlations only truly exist for a population that does not have significant cloud coverage in the visible part of the atmosphere. If a full sample of cloud-free and cloudy planets is considered, there are less strong correlations between atmospheric properties and photometric bins that can be leveraged to classify planets.

2. For giant planets, it is only possible to classify planets into physically motivated groups with greater than 90% accuracy if it is known a priori that the planet does not have significant cloud coverage in the visible. However, observations of solar system and exoplanet giant planets suggest clouds are prevalent in most planetary atmospheres.

3. Our machine learning algorithms are unable to classify planets by metallicity with greater than 55% accuracy. However, we are able to classify planets with moderate accuracy $\sim$70% by classifying by the cloud sedimentation efficiency, $f_{\text{sed}}$. Additionally, binary classification between cloudy and cloud-free populations is successful with an accuracy $>90%$.

4. We find that at least three filters are needed for any kind of classification (cloud, metallicity, etc.). We also tested our classification algorithm with another filter set proposed by Krissansen-Totton et al. (2016), but do not find more optimistic results. However, our statistical algorithm is open-source and can be used to determine optimal filters for WFIRST or future mission concepts.

These results highlight the difficulties with classifying giant extrasolar planets with photometry alone. However, photometry will still be a useful tool for planet searches and potentially identifying targets for follow up.

We thank STScI’s STARGATE team for helpful comments and Caroline Morley for critical guidance in running theoretical models. The authors acknowledge support from the NASA grant # NNG16PJ24C and grant # NNX15AJ80G.

**Software**: scikit-learn (Pedregosa et al. 2011), pandas ( McKinney 2010), bokeh (Bokeh Development Team 2018), NumPy (Walt et al. 2011), IPython (Pérez & Granger 2007), Jupyter, (Kluyver et al. 2016), SciPy (Jones et al. 2014).

**ORCID iDs**

Natasha E. Batalha https://orcid.org/0000-0003-1240-6844
Nikole K. Lewis https://orcid.org/0000-0002-8507-1304
Mark S. Marley https://orcid.org/0000-0002-5251-2943
Jonathan J. Fortney https://orcid.org/0000-0002-9843-4354
Bruce Macintosh https://orcid.org/0000-0003-1212-7538

**References**

Ackerman, A. S., & Marley, M. S. 2001, ApJ, 556, 872
Baines, K., Hammel, H., Rages, K., Romani, P., & Samuelson, R. 1995, in Neptune and Triton, ed. D. P. Cruikshank (Tucson, AZ: Univ. Arizona Press), 489
