The Astrophysical Journal, 930:33 (13pp), 2022 May 1 https://doi.org/10.3847/1538-4357/ac610c
© 2022. The Author(s). Published by the American Astronomical Society.
OPEN ACCESS

Analytical Modeling of Exoplanet Transit Spectroscopy with Dimensional Analysis and Symbolic Regression

Konstantin T. Matchev, Katia Matcheva, and Alexander Roman
Physics Department, University of Florida, Gainesville, FL 32611, USA
Received 2021 December 23; revised 2022 March 20; accepted 2022 March 22; published 2022 May 2

Abstract

The physical characteristics and atmospheric chemical composition of newly discovered exoplanets are often inferred from their transit spectra, which are obtained from complex numerical models of radiative transfer. Alternatively, simple analytical expressions provide insightful physical intuition into the relevant atmospheric processes. The deep-learning revolution has opened the door for deriving such analytical results directly with a computer algorithm fitting to the data. As a proof of concept, we successfully demonstrate the use of symbolic regression on synthetic data for the transit radii of generic hot-Jupiter exoplanets to derive a corresponding analytical formula. As a preprocessing step, we use dimensional analysis to identify the relevant dimensionless combinations of variables and reduce the number of independent inputs, which improves the performance of the symbolic regression. The dimensional analysis also allowed us to mathematically derive and properly parameterize the most general family of degeneracies among the input atmospheric parameters that affect the characterization of an exoplanet atmosphere through transit spectroscopy.

Unified Astronomy Thesaurus concepts: Exoplanet atmospheres (487); Radiative transfer equation (1336); Hot Jupiters (753); Transmission spectroscopy (2133); Dimensionality reduction (1943); Regression (1914)

1. Introduction

Over the past two and a half decades, the number of known planets outside the solar system has soared from only a handful to several thousand,[1] and the focus of the science community has shifted from planet detection to statistical studies and characterization of the environments of these distant worlds (Madhusudhan 2019). A number of current and planned large-scale planetary surveys are based on observing planetary transits at different wavelengths. During a transit event, a planet blocks a certain fraction, $M(\lambda)$, of the original stellar flux $F_O(\lambda)$,

$$M(\lambda) \equiv \frac{F_O(\lambda) - F_T(\lambda)}{F_O(\lambda)}, \tag{1}$$

where $F_T(\lambda)$ is (the minimum of) the observed flux during transit at a given wavelength λ.

Transit spectroscopy targets the detection of commonly present atmospheric gases that have strong absorption lines in the infrared and leave a distinct imprint on the observed modulation, $M(\lambda)$, of the stellar flux (Schneider 1994; Charbonneau et al. 2000; Seager & Sasselov 2000). The theoretical basis of transit spectroscopy has been developed and discussed in a number of studies (Brown 2001; Hubbard et al. 2001; Burrows et al. 2003; Fortney 2005; Benneke & Seager 2012; de Wit & Seager 2013; Griffith 2014; Vahidinia et al. 2014; Heng & Showman 2015; Bétrémieux & Swain 2017; Heng & Kitzmann 2017; Heng 2019) and has been successfully used to extract information about the temperature, composition, and cloud opacity of the atmospheres of numerous transiting exoplanets (Fisher & Heng 2018; Cobb et al. 2019; Barstow & Heng 2020; Kitzmann et al. 2020; Blecic et al. 2022; Cubillos et al. 2022; Harrington et al. 2022; Welbanks & Madhusudhan 2021a). There are several radiative transfer models that perform detailed calculations of the absorption of the stellar flux as it is being attenuated by the gas surrounding the planet. These models typically incorporate collisionally induced absorption (CIA) provided by the main atmospheric gases, line-by-line calculation of the absorption coefficients of minor gas components, and wave-independent cloud opacity based on the gray cloud approximation. Atmospheric refraction and scattering are believed to have a higher-order effect on the observed flux and are mostly excluded from the simulations (Seager & Sasselov 2000; Brown 2001; Hubbard et al. 2001). The complexity of the underlying atmosphere can vary a great deal: from a one-dimensional, isothermal, well-mixed atmosphere (the temperature, the gas mixing ratios, and the cloud opacity are fixed to a constant value with no altitude, latitude, or longitude variations) to three-dimensional models that have a variable temperature profile and vertically resolved cloud layers. Some models even explore the day–night asymmetry in the properties of the atmosphere as the leading and the trailing hemispheres are probed at sunset or sunrise during the transit (MacDonald & Lewis 2022).

The information content of the recorded transit spectrum, $M(\lambda)$, has been the focus of discussion of several studies (Griffith 2014; Heng & Kitzmann 2017; Welbanks & Madhusudhan 2019, 2021b). It has been pointed out that the parameters and/or structure of the underlying atmosphere cannot be uniquely determined from these observations without additional independent information. Heng & Kitzmann (2017) used a simple analytical approach to derive a formula for the observed effective radius of a transiting planet as a function of the atmospheric structure and chemical composition. Despite the simplicity of their formulation, the derived analytical result

[1] The Extrasolar Planet Encyclopedia, http://exoplanet.eu.

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

nicely illustrated the limitations of transit spectroscopy to uniquely determine all atmospheric parameters.

Ultimately, the primary goal of exoplanet transit spectroscopy is the inversion of the observed spectrum in order to retrieve the parameters of the planet and its atmosphere. This process inevitably relies on a forward model that can produce large data sets of synthetic spectra, whose dependence on the underlying atmospheric parameters, however, is often obscured by the model complexity. It is therefore of great interest to have relatively simple analytical expressions as substitutes for the complicated (and slow) forward model. This not only provides valuable insights into the relevant atmospheric processes, but also helps guide the thought process during the inversion.

Recently, machine learning (ML) is increasingly used in the analysis of spectroscopic data from exoplanet transits (Márquez-Neila et al. 2018; Cobb et al. 2019; Fisher et al. 2020; Guzmán-Mesa et al. 2020; Nixon & Madhusudhan 2020; Yip et al. 2021). Most of the time, it is used to solve the inverse problem, i.e., to retrieve the parameter values given the spectroscopic observations (this type of inverse problem is commonly referred to in the ML literature as "simulation-based inference" or "likelihood-free inference"). Here we focus on applications of ML to the forward modeling itself. There are two possible approaches.

1. *Numerical approach.* Replace the slow, complex, and accurate full-blown simulation of the forward model with a fast, simple, and approximate deep-learning model (Himes et al. 2020b). The deep-learning model is typically trained on data simulated by the existing forward model. The advantages of this approach are that (i) once it is trained, the model offers significant speedup, and (ii) it opens the door to nonexperts to participate without the need to know all the specifics of the full-blown forward model. However, there are also certain disadvantages: (i) the deep-learning model is in principle a black box that hides the relevant physics (Yip et al. 2021), and (ii) the deep-learning model learns not the physics of the forward model itself, but the (finite amount of) data generated by the forward model, and this introduces additional uncertainties due to the training process (Matchev et al. 2022b).
2. *Analytical approach.* Here one asks the machine to derive an analytical formula that describes the data well (Langley 1977; Kokar 1986; Langley et al. 1987; Langley & Zytkow 1989; Zembowicz & Żytkow 1992; Todorovski & Dzeroski 1997; Bongard & Lipson 2007; Schmidt & Lipson 2009; Udrescu & Tegmark 2020). In order to succeed, the search for analytical formulas needs to (i) use the relevant (combinations of) variables, and (ii) avoid irrelevant (combinations of) variables. How to properly accomplish all of these tasks within the analytical approach is the main focus of this paper.

One of the two main goals of this paper is to demonstrate the use of symbolic regression to derive accurate analytical formulas representative of a typical forward model. For concreteness, we use the recently proposed PySR framework (Cranmer et al. 2020), which uses a genetic algorithm to perform symbolic regression.[2] Similar studies have been done in several physical domains (Battaglia et al. 2016; Chang et al. 2016; Iten et al. 2020; Udrescu & Tegmark 2020; Arechiga et al. 2021; Lemos et al. 2022), but to the best of our knowledge, not in exoplanetary science.

A typical forward model takes as inputs a relatively large number of input parameters (on the order of a dozen or more). Keeping all of them as independent degrees of freedom significantly complicates the task of symbolic regression. Fortunately, one can use dimensional analysis to restrict the relevant number of degrees of freedom. The second main goal of this paper is to demonstrate the use of dimensional analysis to identify the relevant combinations of physical parameters (the so-called Pi groups; Barenblatt 1996) that uniquely determine the observed transit spectra. In particular, we show that the initial set of seven free parameters in our example reduces to only four potentially relevant degrees of freedom. This simplification has important benefits:

1. The reduction in the relevant degrees of freedom greatly enhances the performance of the symbolic regression.
2. Our dimensional analysis correctly reproduces the parameter degeneracies already known in the literature (Heng & Kitzmann 2017; Welbanks & Madhusudhan 2019) and identifies many new ones.
3. The main result from our dimensional analysis not only agrees with previously derived analytical approximations in the literature (Heng & Kitzmann 2017), but also points to the only allowed extensions of those existing formulas that are consistent with basic physics.

The paper is organized as follows. In Section 2 we review the advantages and disadvantages of the two different approaches to the forward problem of transit spectroscopy and motivate our course of action in this paper. In Section 3 we apply dimensional analysis to the radiative transfer problem: we first derive the set of dimensionless variables describing the problem, and then in Section 4, we identify the relevant ones among them. As a byproduct, in Sections 3 and 4 we derive the most general set of degeneracies that arise among the atmospheric parameters. In Section 5 we use our parameterization in terms of dimensionless variables to fit a symbolic regression and obtain analytic expressions for the forward radiative transfer model. In the Appendix we compile a list of simple degeneracies among the atmospheric parameters.

2. Modeling the Physics of Radiative Transfer in the Atmosphere

2.1. Detailed Forward Numerical Simulations

At the heart of all observationally driven methods used to derive the properties of a planetary atmosphere is a numerical forward radiative transfer model, schematically depicted with the left blue shaded rectangle in Figure 1. The model starts with a given set, $\mathcal{S}$, of atmospheric parameters: temperature (T) and pressure (P) profiles, chemical abundances (n_j) of the present gases, mean molecular mass (m), cloud opacity ($\kappa_{\rm cl}$), specific gravity (g), and geometry (reference planet radius R_0 and stellar radius R_S). The model then calculates the atmospheric transmission of the planet and generates a synthetic spectrum of the emerging specific stellar flux, $F_T(\lambda_i)$, at several different wavelengths λ_i,

$$\mathcal{S}(T, P, n_j, m, \kappa_{\rm cl}, g, R_0, R_S) \longrightarrow F_T(\lambda_i), \tag{2}$$

[2] https://github.com/MilesCranmer/PySR

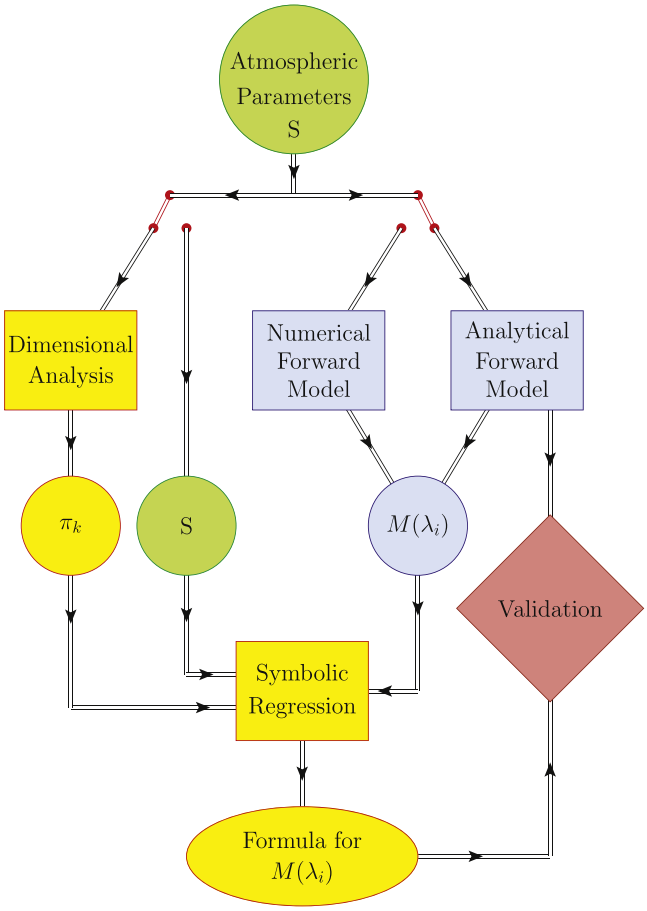

Figure 1. A simplified flowchart illustrating the traditional numerical and analytical alternatives for a forward radiative transfer model. The leftmost yellow shaded branch represents the dimensional analysis/symbolic regression approach in this paper. Rectangular shapes correspond to operations, circular/oval shapes correspond to inputs or outputs, and a diamond shape indicates a decision. The right switch toggles between a numerical or analytical forward model, while the left switch decides whether to apply dimensional analysis or not. The switches are left in their default positions corresponding to the analysis presented in this paper.

which can then be converted with the help of Equation (1) into respective modulations $M(\lambda_i)$ of the observed stellar flux, as represented with the blue shaded circle in Figure 1. For an excellent short summary and historical perspective of the published exoplanetary radiative transfer models, see Blecic et al. (2022). The forward models allow for the construction of an arbitrary complex atmosphere by implementing varying temperature profiles and chemical abundances, localized cloud layers, and can even introduce nonequilibrium processes by coupling with thermochemical models (Matcheva et al. 2005; Giles et al. 2015; Changeat et al. 2019). A more realistic description of the atmosphere, however, leads to an increased complexity of the models, which necessarily translates into increased computational cost. Moreover, due to the limited information content of the observations, the increased complexity of the forward model does not necessarily equate to a more robust characterization of the atmosphere. The reliability of the derived results and the interpretation of unresolved correlations between different atmospheric parameters are a main focus of development (Heng & Kitzmann 2017; Welbanks & Madhusudhan 2019).

2.2. Simplified Analytical Approximations

As shown in the right blue shaded branch of Figure 1, a viable alternative to the detailed computational radiative transfer models is the use of analytical expressions that provide physical insight into how the underlying atmospheric structure and composition directly impact the observed spectral flux. The analytical expressions are easy to implement and understand, they are well suited for investigating correlations between different parameters, and the model uncertainties are well understood and easily calculated. The trade-off is that the analytical expressions are based on simplifying assumptions, for example, isothermal atmosphere, gray clouds, and spherical symmetry. At the same time, these approximations seem appropriate at this early stage of exoplanet exploration because the limited information content in the observed spectrum does not allow one to probe the complexity of the atmosphere.

In this paper we proceed along the leftmost (yellow shaded) branch of Figure 1, which uses symbolic regression to derive an analytical expression for the modulation $M(\lambda_i)$ of the observed stellar flux. The inputs to the symbolic regression procedure (represented with the lower yellow shaded rectangle in Figure 1) are two different types of data:

1. *The input atmospheric variables, or combinations of them.* Instead of feeding all input variables directly into the symbolic regression, we first perform dimensional analysis in order to reduce the number of degrees of freedom by identifying the relevant dimensionless variables π_k. This not only eliminates the problem of parameter degeneracies, but also aids the performance of the symbolic regression, which scales poorly with increasing the number of input features.
2. *The modulation of the stellar flux.* As shown in Figure 1, the predicted $M(\lambda_i)$ that enters the symbolic regression can be taken either from a numerical forward model (left branch), or from its analytical counterpart (right branch). For definiteness, the analytical model we employ here uses the expression derived in Heng & Kitzmann (2017).

Note that the blue shaded branches in Figure 1 represent the paradigm of traditional computing, whereby given a numerical input, a program produces a numerical output; while the left yellow branch in Figure 1 represents the paradigm of machine learning, whereby given numerical inputs and the corresponding outputs, the program produces a method, in this case, a symbolic formula.

Finally, the result from the symbolic regression can be compared to existing analytical expressions for forward models (the red diamond in Figure 1). This validation process can go in both directions:

1. In the case when the input $M(\lambda_i)$ to the symbolic regression is taken from the right (analytical) blue shaded branch, the derived formula should be identical to the one used in the forward model. This provides an important consistency check and validation of the symbolic regression procedure, which is performed in Section 5 below.
2. Conversely, in the case when the input $M(\lambda_i)$ to the symbolic regression is taken from the left (numerical) blue shaded branch, the derived formula from symbolic regression should in principle be an improvement over the approximate analytical expressions used in the

Table 1
Notation, Names, and SI Units for the Physical Governing Variables Impacting the Transit Radius R_T

Notation	Name	SI Unit	Power
R_0	reference radius	m	α
R_S	stellar radius	m	β
k_B	Boltzmann constant	$m^2\ kg\ s^{-2}\ K^{-1}$	γ
T	temperature	K	δ
m	mean molecular mass	kg	ε
g	surface gravity	$m\ s^{-2}$	ζ
P_0	reference pressure at R_0	$m^{-1}\ kg\ s^{-2}$	η
κ	cross-section per unit mass	$m^2\ kg^{-1}$	θ

Note. The last column lists the power with which the respective variable enters the defining Equation (3).

forward analytical models. This type of exercise, however, is beyond the scope of this paper and will be undertaken in a future study.

3. Dimensional Analysis

The first step in modeling any physical phenomenon is the identification of the relevant physics variables and then finding a relation among them that describes the process of interest. For sufficiently simple systems, such a quantitative relationship can be obtained from first principles, using the known fundamental physics laws. However, for sufficiently complex phenomena such as the radiative transfer within an exoplanet atmosphere, such ab initio theory is often difficult, if not impossible, and one has to resort to alternative, typically numerical, modeling methods. For example, the recent wide spread of machine-learning methods has led to attempts to replace the forward model with a deep-learned model (or a variant thereof; Márquez-Neila et al. 2018; Zingales & Waldmann 2018; Cobb et al. 2019; Himes et al. 2020a, 2020c; Oreshenko et al. 2020; Ardevol Martinez et al. 2022). At the same time, a simple alternative to the numerical modeling approaches is the tried and true method of dimensional analysis, which relies on the simple fact that physical laws do not depend on the choice of the basic units of measurements. In this section we therefore apply dimensional analysis to our problem at hand, namely, relating the observed transit radius $R_T(\lambda)$ to the relevant variables characterizing the atmosphere of an exoplanet. We do not give a formal introduction to the method of dimensional analysis, for which we refer to the classic books on the subject (Langhaar 1951; Barenblatt 1996), instead, we use the worked-out example below to illustrate how the method works.

3.1. Choice of Variables

The first step, the choice of the so-called governing variables in the parlance of dimensional analysis, is easy—the parameters that might impact the governed variable $R_T(\lambda)$ were already mentioned in Section 2 (see Equation (2)). For convenience, they are collected in Table 1, together with their names, notation, and SI units. For consistency, we follow the notation of Heng & Kitzmann (2017). We assume an isothermal atmosphere with a mean molecular mass m and temperature T. The atmospheric structure is defined with respect to a reference pressure level P_0 at a given radius R_0, which corresponds to an optically thick atmosphere along the line of sight. The chemical composition, the gas mixing ratios, and the cloud opacity are incorporated in the gas absorption cross-section per unit mass, κ.

We note that there are two variables, the reference radius R_0 and the star radius R_S, which are commensurable with (i.e., have the same units as) the transit radius $R_T(\lambda)$. While either one of them can be chosen to provide the dimensions of $R_T(\lambda)$, for definiteness and without loss of generality, we choose R_0 because it is more directly related to the transit radius.

3.2. Identification of the Complete Set of Pi Groups

With these preliminaries, we can now write down the desired relationship in the form

$$R_T(\lambda) \sim R_0 \times [R_0^{\alpha}\, R_S^{\beta}\, k_B^{\gamma}\, T^{\delta}\, m^{\varepsilon}\, g^{\zeta}\, P_0^{\eta} (\kappa(\lambda))^{\theta}\,], \tag{3}$$

where the quantity within the square brackets is dimensionless. Dimensional homogeneity (Barenblatt 1996) implies that the powers $\alpha, \beta, \gamma, \ldots, \theta$ can be chosen to be integers. Note that eight governing variables and four SI units appear in Table 1; it is easy to check that the rank of the dimensional matrix is indeed $8-4=4$. In this case, the Buckingham π theorem (Buckingham 1914) then states that the quantity within the square brackets on the right-hand side of (3) is a function of exactly four dimensionless parameter combinations π_k with $k=1, 2, 3, 4$ (the so-called Pi groups) constructed from the governing variables (see Figure 1). To find them, we need to solve the linear equations

$$\alpha + \beta + 2\gamma + \zeta - \eta + 2\theta = 0, \tag{4a}$$
$$\gamma + \varepsilon + \eta - \theta = 0, \tag{4b}$$
$$-2\gamma - 2\zeta - 2\eta = 0, \tag{4c}$$
$$-\gamma + \delta = 0, \tag{4d}$$

which ensure the correct dimensionality of Equation (3). Eliminating δ, η, and θ from Equations 4(b)–(d), we rewrite Equation 4(a) in the form

$$\alpha + \beta + 3\gamma + 2\varepsilon + 0\zeta = 0. \tag{5}$$

This equation describes a four-dimensional flat hyperplane in the five-dimensional space spanned by the coordinates (α, β, γ, ε, ζ). Equation (5) defines this hyperplane as being orthogonal to the constant vector $\boldsymbol{\pi}_0 \equiv (1, 1, 3, 2, 0)$, therefore we can choose to parameterize points on the hyperplane as linear combinations of the following four[3] vectors that are (i) orthogonal to $\boldsymbol{\pi}_0$ and (ii) mutually orthogonal among themselves, which ensures that they are also linearly independent,

$$\begin{aligned}
\boldsymbol{\pi}_1\colon (\alpha, \beta, \gamma, \varepsilon, \zeta) &= (1, 1, -2, 2, 0)\\
\Rightarrow (\delta, \eta, \theta) &= (-2, 2, 2),\\
\boldsymbol{\pi}_2\colon (\alpha, \beta, \gamma, \varepsilon, \zeta) &= (0, 0, 0, 0, -1)\\
\Rightarrow (\delta, \eta, \theta) &= (0, 1, 1),\\
\boldsymbol{\pi}_3\colon (\alpha, \beta, \gamma, \varepsilon, \zeta) &= (1, -1, 0, 0, 0)\\
\Rightarrow (\delta, \eta, \theta) &= (0, 0, 0),\\
\boldsymbol{\pi}_4\colon (\alpha, \beta, \gamma, \varepsilon, \zeta) &= (1, 1, 0, -1, 0)\\
\Rightarrow (\delta, \eta, \theta) &= (0, 0, -1).
\end{aligned}$$

[3] Of course, any other choice of four linearly independent vectors orthogonal to $\boldsymbol{\pi}_0$ will work as well. We take advantage of this freedom below to simplify the obtained Pi groups.

Each of these vectors in turn defines a dimensionless Pi group of governing variables,

$$\boldsymbol{\pi}_1 \Longrightarrow \pi_1 = \frac{R_0\, R_S\, m^2 P_0^2\, \kappa^2}{k_B^2 T^2}, \tag{7a}$$

$$\boldsymbol{\pi}_2 \Longrightarrow \pi_2 = \frac{P_0 \kappa}{g}, \tag{7b}$$

$$\boldsymbol{\pi}_3 \Longrightarrow \pi_3 = \frac{R_0}{R_S}, \tag{7c}$$

$$\boldsymbol{\pi}_4 \Longrightarrow \pi_4 = \frac{R_0 R_S}{m\kappa}. \tag{7d}$$

Note that while the first Pi group looks complicated, it can be rewritten as

$$\frac{R_0\, R_S\, m^2 P_0^2\, \kappa^2}{k_B^2\, T^2} = \left(\frac{R_0\, mg}{k_B\, T}\right)^2 \left(\frac{P_0\, \kappa}{g}\right)^2 \frac{R_S}{R_0} = \left(\frac{R_0\, mg}{k_B\, T}\right)^2 \frac{\pi_2^2}{\pi_3}$$

and can therefore be traded for the much simpler dimensionless combination

$$\pi_1' = \frac{R_0\, mg}{k_B\, T} \equiv \frac{R_0}{H}, \tag{8}$$

where H is the pressure scale height

$$H \equiv \frac{k_B\, T}{m\, g}. \tag{9}$$

We can similarly trade π_4 in Equation 7(d) for

$$\pi_4' = \frac{R_0^2}{m\, \kappa}. \tag{10}$$

The final result from our dimensional analysis is that any meaningful expression for the transit radius must be of the form

$$\begin{aligned} R_T(\lambda) &\sim R_0 \times f(\pi_1', \pi_2(\lambda), \pi_3, \pi_4'(\lambda)) \\ &= R_0 \times f\left(\frac{R_0}{H}, \frac{P_0\, \kappa(\lambda)}{g}, \frac{R_0}{R_S}, \frac{R_0^2}{m\kappa(\lambda)}\right), \end{aligned} \tag{11}$$

where f is an unspecified function of dimensionless quantities only, i.e., of the four Pi groups listed as its arguments, plus possibly some numerical dimensionless mathematical constants like real numbers, π, e, etc. Equation (11) can be recast in terms of the modulation $M(\lambda_i) = (\pi R_T^2(\lambda))/(\pi R_S^2)$ of the observed stellar flux as follows:

$$\begin{aligned} M(\lambda) &\sim \left(\frac{R_0}{R_S}\right)^2 \times f^2(\pi_1', \pi_2(\lambda), \pi_3, \pi_4'(\lambda)) \\ &= \left(\frac{R_0}{R_S}\right)^2 \times f^2\left(\frac{R_0}{H}, \frac{P_0\, \kappa(\lambda)}{g}, \frac{R_0}{R_S}, \frac{R_0^2}{m\kappa(\lambda)}\right). \end{aligned} \tag{12}$$

Equations (11) and (12) are the main results from our dimensional analysis. At this stage, the functional dependence f remains arbitrary, and to make further progress, one has to fit these equations to data generated by a forward model (either numerical or analytical). In particular, the data will reveal (i) whether all four π variables are entering the actual relationship and (ii) the exact functional form of f. We undertake these two exercises in the next two sections. Before we do this, however, we conclude this section with a discussion of the parameter degeneracies implied by Equations (11) and (12).

Table 2
Explicit Parameterization of the Three Degrees of Freedom Degeneracy Discussed in the Text

Variable	R_0 Scaling	P_0 Scaling	κ Scaling
R_0	L_{R_0}	⋯	⋯
P_0	⋯	L_{P_0}	⋯
κ	⋯	⋯	L_κ
T	$L_{R_0}^3$	L_{P_0}	⋯
m	$L_{R_0}^2$	⋯	L_κ^{-1}
g	⋯	L_{P_0}	L_κ
R_S	L_{R_0}	⋯	⋯

Note. The three parameters were chosen to be the scaling factors for R_0, P_0, and κ, respectively.

3.3. Guaranteed Degeneracies in the Interpretation of Exoplanet Transmission Spectra

The ability of transmission spectra to uniquely constrain the atmospheric parameters of exoplanets has been an intense topic of discussion[4] since the inception of the field. Many previous studies have noted various degeneracies between different sets of atmospheric parameters, which result in the same observed transit spectra (within the error bars). The methods used in these works range from numerical to statistical to semianalytical. The results from our dimensional analysis here, Equations (11) and (12), can now help us understand from first principles the existing statements in the literature concerning the degeneracy of atmospheric retrievals. We note that the degeneracies that we discuss here are exact, and cannot be lifted by improving the experimental precision.

The main point is that there are seven variable parameters[5] in Table 1, while the physics of the transmission spectra depends on only four unique Pi groups (see Equations (11) and (12)). This implies that there exists a degeneracy among all seven atmospheric variables that (i) can be parameterized by $7 - 4 = 3$ degrees of freedom and (ii) preserves the values of all four π_k variables, and consequently, the values of $R_T(\lambda)$ and $M(\lambda)$. We choose R_0, P_0 and κ as our free-varying variables that scale as $R_0 \to L_{R_0} \times R_0$, $P_0 \to L_{P_0} \times P_0$ and $\kappa \to L_\kappa \times \kappa$, where $(L_{R_0}, L_{P_0}, L_\kappa)$ are the three scaling factors parameterizing the degeneracy. Then, in order to keep all four π_k groups constant, the remaining four atmospheric variables, T, m, g, and R_S, should scale as shown in Table 2. Explicitly,

$$R_0 \to L_{R_0} \times R_0 \tag{13a}$$

$$P_0 \to L_{P_0} \times P_0 \tag{13b}$$

$$\kappa \to L_\kappa \times \kappa \tag{13c}$$

$$T \to L_{R_0}^3 L_{P_0} \times T \tag{13d}$$

$$m \to L_{R_0}^2 L_\kappa^{-1} \times m \tag{13e}$$

$$g \to L_{P_0} L_\kappa \times g \tag{13f}$$

$$R_S \to L_{R_0} \times R_S. \tag{13g}$$

Equation (13) is the most general transformation for which the degeneracy is guaranteed in the sense that its existence does not

[4] For a recent review and a guide to the relevant literature, see Welbanks & Madhusudhan (2019) and references therein.

[5] Not counting the fundamental constant k_B, which presumably has the same value on the exoplanet (Duff 2015).

depend on the functional form of $f(\pi_1', \pi_2, \pi_3, \pi_4')$. Even broader degeneracies may arise if the function f does not explicitly depend on one or more of its π_k arguments. In analogy to classical Lagrangian mechanics, we refer to such missing variables as "cyclic" variables. In the next section, we find that this does actually occur and that π_4' is cyclic, which further enlarges the set of degeneracy transformations. For now, we conclude this section by listing the three individual degeneracies from Table 2 and Equation (13):

$$L_{R_0} \neq 1, L_{P_0} = L_\kappa = 1 \Longrightarrow R_0 \to L \times R_0, \quad T \to L^3 \times T, \quad m \to L^2 \times m, \quad R_S \to L \times R_S; \tag{14a}$$

$$L_{P_0} \neq 1, L_{R_0} = L_\kappa = 1 \Longrightarrow P_0 \to L \times P_0, \quad T \to L \times T, \quad g \to L \times g; \tag{14b}$$

$$L_\kappa \neq 1, L_{P_0} = L_{R_0} = 1 \Longrightarrow \kappa \to L \times \kappa, \quad m \to L^{-1} \times m, \quad g \to L \times g. \tag{14c}$$

The first reflects a planet radius–stellar radius–temperature–mass degeneracy, the second reveals a pressure–temperature–gravity degeneracy, and the last one implies a mass–gravity–opacity degeneracy. Of course, one can also consider arbitrary combinations of these three individual transformations, as shown in Equation (13).

Each one of the transformations (14) has a clear physics interpretation. Consider, for example, Equation 14(b). A higher temperature T would cause the atmosphere to "puff up," but this can be compensated by a corresponding increase in the specific gravity g, so that the scale height H remains constant. As a result, all three length variables R_0, R_S, and H remain the same, which in turn keeps π_1' and π_3 constant as well. Furthermore, because m and κ are unaffected by the transformation 14(b), $\pi_4' = R_0^2/(m\kappa)$ is also constant. Finally, because the pressure P_0 and g are varied at the same rate, $\pi_2 \sim P_0/g$ stays constant as well.

Note that so far, we have been considering the transit radius $R_T(\lambda)$ measured at a single value of λ, but all of our previous conclusions also apply for multiwavelength observations (Matchev et al. 2022a).

4. Identifying Relevant and Irrelevant Pi Groups

The dimensional analysis performed in the previous section identified the four relevant variables π_1', π_2, π_3 , and π_4', but did not guarantee that the function f in Equation (11) depends on all four of them. In this section we use three different approaches to investigate in more detail the dependence of $f(\pi_1', \pi_2, \pi_3, \pi_4')$ on each of its four arguments.

4.1. Comparison to Existing Analytical Approximations

We begin with an easy shortcut—simply looking up the expected functional form of $f(\pi_1', \pi_2, \pi_3, \pi_4')$ in the existing analytical approximations in the literature. For example, Heng & Kitzmann (2017) derived the following expression for the transit radius:

$$R_T(\lambda) = R_0\left\{1 + \frac{H}{R_0}[\gamma_E + E_1(\tau_0) + \ln(\tau_0)]\right\}, \tag{15}$$

where $\gamma_E = 0.577215665$ is the Euler–Mascheroni constant,

$$\tau_0(\lambda) \equiv \frac{P_0\kappa(\lambda)}{g}\sqrt{2\pi\frac{R_0}{H}} \tag{16}$$

is the optical thickness of the atmosphere along the line of sight at the reference radius R_0 , and

$$E_1(\tau_0) = \int_{\tau_0}^{\infty} \frac{e^{-t}}{t} dt \tag{17}$$

is the exponential integral of the first order with argument τ_0.

In the large τ_0 limit, the E_1 term vanishes,

$$\lim_{\tau_0 \to \infty} E_1(\tau_0) = 0,$$

and Equation (15) simplifies to

$$R_T(\lambda) = R_0\left\{1 + \frac{H}{R_0}\left[\gamma_E + \ln\left(\frac{P_0\kappa(\lambda)}{g}\sqrt{2\pi\frac{R_0}{H}}\right)\right]\right\}. \tag{18}$$

Both Equations (15) and (18) are consistent with our general result in Equation (11) derived earlier. In particular, Equation (15) allows us to identify the function f as

$$f(\pi_1', \pi_2) = 1 + \frac{1}{\pi_1'}[\gamma_E + E_1(\pi_2\sqrt{2\pi\pi_1'}) + \ln(\pi_2\sqrt{2\pi\pi_1'})], \tag{19}$$

where one should not confuse the usual constant $\pi = 3.14159265$ with the four Pi groups π_k derived in Section 3. In the large τ_0 limit, this reduces to a corresponding analog of Equation (18),

$$f(\pi_1', \pi_2) = 1 + \frac{1}{\pi_1'}[\gamma_E + \ln(\pi_2\sqrt{2\pi\pi_1'})]. \tag{20}$$

The analogous expressions for the modulation $M(\lambda)$ of the observed stellar flux are

$$M(\lambda; \pi_1', \pi_2, \pi_3) = \pi_3^2\left\{1 + \frac{1}{\pi_1'}[\gamma_E + E_1(\pi_2\sqrt{2\pi\pi_1'}) + \ln(\pi_2\sqrt{2\pi\pi_1'})]\right\}^2 \tag{21}$$

for the general case, and

$$M(\lambda; \pi_1', \pi_2, \pi_3) = \pi_3^2\left\{1 + \frac{1}{\pi_1'}[\gamma_E + \ln(\pi_2\sqrt{2\pi\pi_1'})]\right\}^2 \tag{22}$$

in the large τ_0 limit.

Note that the function f depends on only two of the identified π_k variables (π_1' and π_2), while the modulation $M(\lambda)$ depends on only three out of the four variables (π_1', π_2, and π_3). In each case, π_4' is missing and can be identified as a cyclic variable, thus expanding the previously discussed set of degeneracies in Equation (13).

In what follows, we illustrate our results with the synthetic benchmark data set of Márquez-Neila et al. (2018), which consists of 100,000 synthetic Hubble Space Telescope Wide Field Camera 3 (WFC3) spectra of transit radii of hot Jupiters observed at 13 different wavelengths λ_i in the range 0.838–1.666 μm. The data set is created by using Equation (18) and scanning the parameter space of five

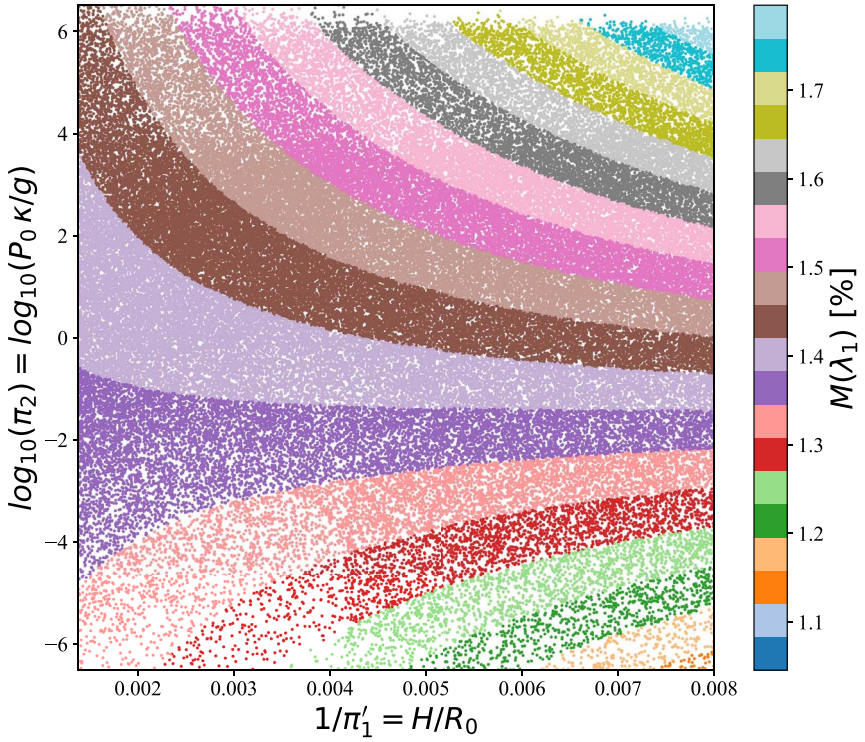

Figure 2. Scatter plot of the 100,000 points in the data set of Márquez-Neila et al. (2018) vs. the dimensionless parameters $1/\pi_1'$ and $\log_{10}(\pi_2)$, color-coded by the value of $M(\lambda_1)$ in percent.

different atmospheric parameters: temperature T, abundances of the H_2 O, NH_3, and HCN gases, and gray cloud opacity κ_{cl}.

Figure 2 shows a scatter plot of the 100,000 points in the data set of Márquez-Neila et al. (2018) versus our dimensionless parameters π_1' and π_2, where for plotting convenience, we have chosen to display $1/\pi_1'$ and $\log_{10}(\pi_2)$ on the axes. The points are color-coded by the value of $M(\lambda_1)$ in percent, where $\lambda_1 = 0.867\ \mu$m is the first wavelength in the data set. Note that R_0 and R_S were kept fixed when producing the data set, and therefore π_3 does not vary throughout it, which prevents us from using these data to illustrate any π_3 dependence.

The well-defined color bands in Figure 2 indicate that $M(\lambda_1)$ is a unique function of the two plotted variables, π_1' and π_2, as suggested by the generating formula in Equation (18). This confirms that the relevant physics information encoded in the spectroscopic measurements throughout this synthetic data set is indeed only sensitive to π_1' and π_4 (because π_3 was fixed), in agreement with Equation (22). Note that while this exercise was done here with a simple analytical forward model as an illustration of the procedure, it can easily be repeated with a full numerical forward model.

4.2. Order-of-magnitude Analysis

Dimensional analysis does provide a rule of thumb to determine whether a given governing parameter is relevant or not: "if the dimensionless parameter is either very small or very large compared to unity, it may be assumed to be not essential, and the function f can be assumed to be constant (or, in general, when there are several dimensionless parameters, a function of one fewer arguments)" (Barenblatt 1996). We determine whether this is the case in our example. Figure 3 shows distributions of (base 10 logarithms of) the dimensionless quantities π_1', π_2, and π_4' in the data set of Márquez-Neila et al. (2018). We recall that π_3 is constant throughout the data set, so adding its distribution to Figure 3 would only add a trivial delta function near the origin.

Figure 3 shows that while the distributions of π_1' and π_2 are in the proximity of $10^0 = 1$, the distribution of π_4' is shifted by almost 50 orders of magnitude. The expectations from dimensional analysis would then suggest that π_1' and π_2 are essential variables, while π_4' is a nonessential variable. This conclusion is in agreement with the discussion from the previous subsection.

4.3. Direct Tests of the Function f

While the relevancy tests in the previous two subsections relied on approximations, heuristics, or intuition from dimensional analysis, in this subsection we perform a direct quantitative test of the behavior of the function f with respect to its arguments. In particular, we change one Pi group at a time while keeping the other Pi groups constant, and then we verify whether the variation of a single Pi group leads to a change in $f(\pi_k)$ or not. We imagine that in practice, this test would be done with a full numerical forward radiative transfer model, but for the purpose of this paper, it is sufficient to illustrate the basic idea with the numerical forward model of Equation (18) that was used to generate the data set of Márquez-Neila et al. (2018).

The results from the exercise are shown in Figure 4. The bottom panels show the required variation of the input variables

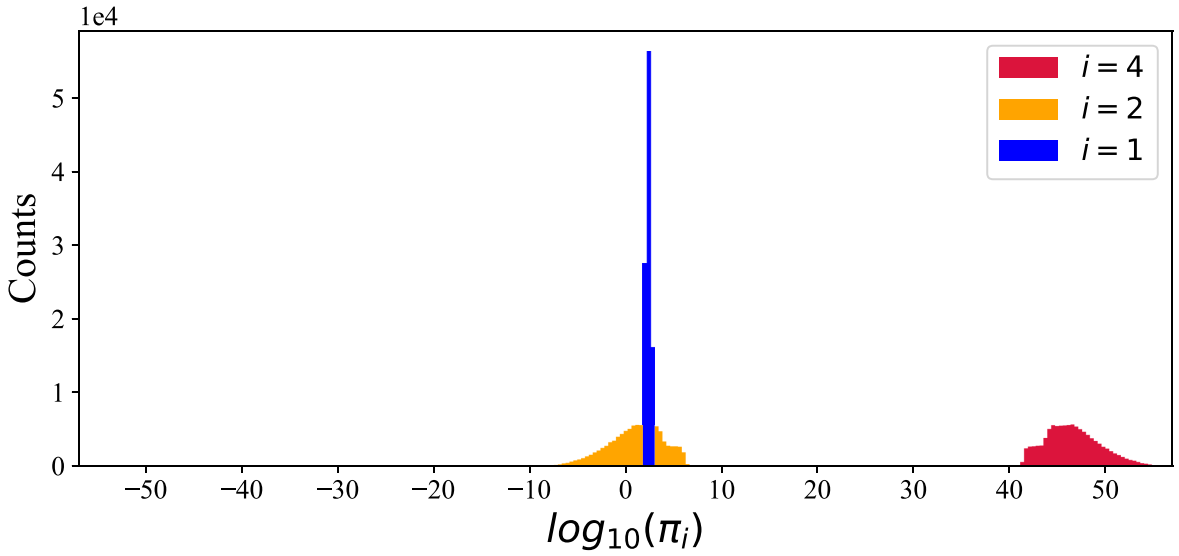

Figure 3. Distributions of the (base 10 logarithms of the) dimensionless quantities π_1', π_2, and π_4' in the data set of Márquez-Neila et al. (2018).

from Table 1, which triggers a variation in only one of the π_ks while keeping the others constant, as illustrated by the plots in the middle row. The exact scaling relations are as follows:

$$T \to T/L \Longrightarrow \pi_1^{'} \to L\pi_1^{'}, \quad \pi_2,\ \pi_3 \textit{ and } \pi_4^{'} \text{ remain constant;} \tag{23a}$$

$$T \to T/L; \quad g \to g/L \Longrightarrow \pi_2 \to L\pi_2, \quad \pi_1',\ \pi_3 \text{ and } \pi_4' \text{ remain constant;} \tag{23b}$$

$$R_S \to R_S/L \Longrightarrow \pi_3 \to L\pi_3, \quad \pi_1^{'},\ \pi_2 \text{ and } \pi_4^{'} \text{ remain constant;} \tag{23c}$$

$$T \to T/L; \quad m \to m/L \Longrightarrow \pi_4^{'} \to L\pi_4^{'}, \quad \pi_1^{'},\ \pi_2 \textit{ and } \pi_3 \text{ remain constant.} \tag{23d}$$

The top panels in Figure 4 reveal that $M(\lambda_1)$ depends on π_1', π_2, and π_3, but not on π_4'. This result is in agreement with our conclusions from Sections 4.1 and 4.2, only this time, it is made on a firm quantitative footing. Note that while we have established the dependence of $M(\lambda_1)$ on π_1', π_2, and π_3, we still do not know its exact functional form—this task is postponed for Section 5.

4.4. The Complete Set of Guaranteed and Accidental Degeneracies

The analysis in the previous subsection revealed that π_4' is a cyclic variable, and its value does not affect the observed transmission spectrum. This leads to an additional, "accidental," degeneracy, illustrated by the plots in the rightmost column in Figure 4. Using Equation 23(d), we can express this degeneracy as

$$T \to L_T \times T, \tag{24a}$$

$$m \to L_T \times m \tag{24b}$$

in terms of a new scaling parameter L_T. This degeneracy implies that if the mean molecular mass is unknown a priori, there would be difficulties in extracting the precise value of the temperature.

We recall that in Section 3.3, we already derived a family of guaranteed degeneracies, Equation (13), among the seven input atmospheric variables. For completeness, Figure 5 illustrates these earlier discussions in the same format as Figure 4. Now we can combine the results of Equations (13) and (24) to arrive at the most general family of degeneracies in $M(\lambda)$,

$$R_0 \to L_{R_0} \times R_0 \tag{25a}$$

$$P_0 \to L_{P_0} \times P_0 \tag{25b}$$

$$\kappa \to L_\kappa \times \kappa \tag{25c}$$

$$T \to L_{R_0}^3 L_{P_0} L_T \times T \tag{25d}$$

$$m \to L_{R_0}^2 L_\kappa^{-1} L_T \times m \tag{25e}$$

$$g \to L_{P_0} L_\kappa \times g \tag{25f}$$

$$R_S \to L_{R_0} \times R_S, \tag{25g}$$

which include both the degeneracies guaranteed by dimensional analysis alone and the accidental degeneracy arising due to the specific functional form of f.

Equation (25) is one of the main results of this paper. Being completely general, it should encapsulate all existing discussions of degeneracies in the literature; i.e., one should be able to reproduce any[6] previous correctly identified degeneracy as a special case of Equation (25). For illustration, in the Appendix we have compiled a list of possible degeneracies involving two or three atmospheric parameters, summarized pictorially in Figure 7. Note that according to Figure 7, the specific gravity g plays a central role in the existing degeneracies. This suggests that pinning down the value of g to high accuracy by means of independent observations would eliminate a large number of degeneracies and thus significantly decrease the uncertainties on the remaining parameters.

5. Symbolic Regression

Symbolic regression is a somewhat underused, yet very interpretable machine-learning algorithm for modeling a data

[6] Note that additional degeneracies not captured by Equation (25) may arise when one tries to disentangle the contributions of individual minor gases and/or clouds to the overall opacity of the atmosphere κ.

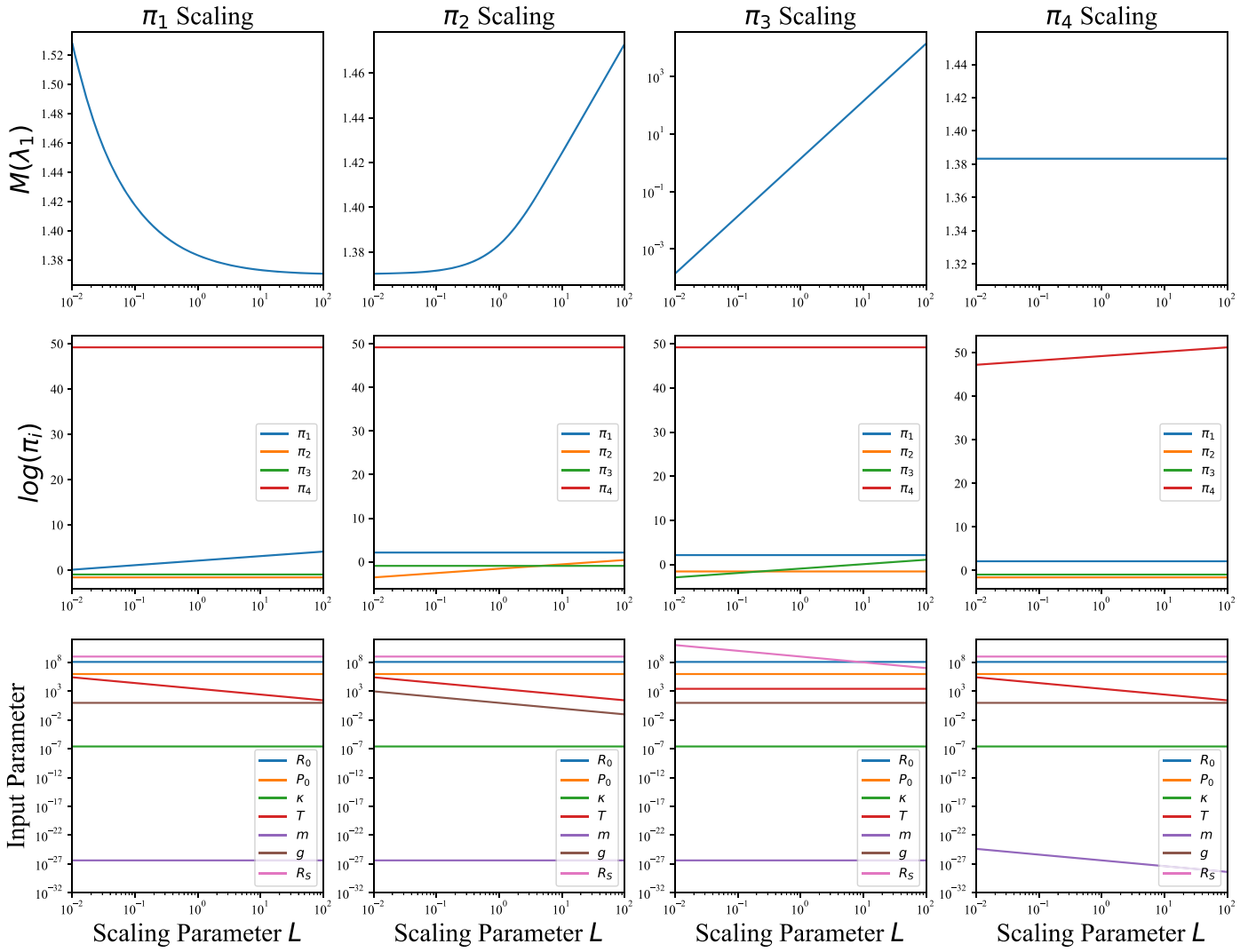

Figure 4. Testing the dependence of the modulation $M(\lambda_1)$ on each individual π_k parameter. The plots in the bottom row show the scaling of the input parameters from Table 1, which is needed to vary one of the πs but not the others, as illustrated by the plots in the middle row. The plots in the top row show the resulting variation in $M(\lambda_1)$.

set with analytic expressions by searching over a suitable space of functions. Since the numerical forward models of radiative transfer are quite complicated, the problem of finding valid and easily interpretable analytical expressions reproducing the results from the numerical models is well motivated.

In this section we fit a symbolic regression to transit spectroscopy data. For this proof-of-concept exercise, we use data generated by the analytical forward model in Equation (18) behind the benchmark data set of Márquez-Neila et al. (2018). Since the performance of symbolic regressions is known to deteriorate as the dimensionality of the data increases, we take advantage of our earlier dimensional analysis that allowed us to reduce the number of independent inputs from 7 to 4, namely π_1, π_2, π_3, and π_4.[7] To be specific, we try to learn the dimensionless function

$$f_{\rm true}(\pi_1, \pi_2, \pi_3, \pi_4) \equiv \frac{R_T(\lambda)}{R_S} - 1 = \frac{1}{\pi_1}[\gamma_{\rm E} + \ln(\pi_2\sqrt{2\pi\pi_1})]. \quad (26)$$

To do this, we make use of the recently released PySR software package (Cranmer et al. 2020). It models the data set with a graph neural network before applying symbolic regression to fit different internal parts of the learned model that operate on reduced dimension representations. We do not attempt any hyperparameter optimization and instead we use the default configuration in the PySR version 0.6.14 distribution.

We generate training data of 1000 samples as follows. We sample π_1 within its full range in the benchmark data set of Márquez-Neila et al. (2018), namely from 125 to 714. As shown in Figure 3, the values of π_2 in the data set span many orders of magnitude, thus we choose to sample π_2 only within a restricted range from $e^0 = 1$ to $e^5 = 148$. The dashed rectangle in Figure 6 shows the resulting region in the $(\pi_1, \log_{10}\pi_2)$

[7] To simplify the notation, from now on in this section, we omit the primes on π_1' and π_4'.

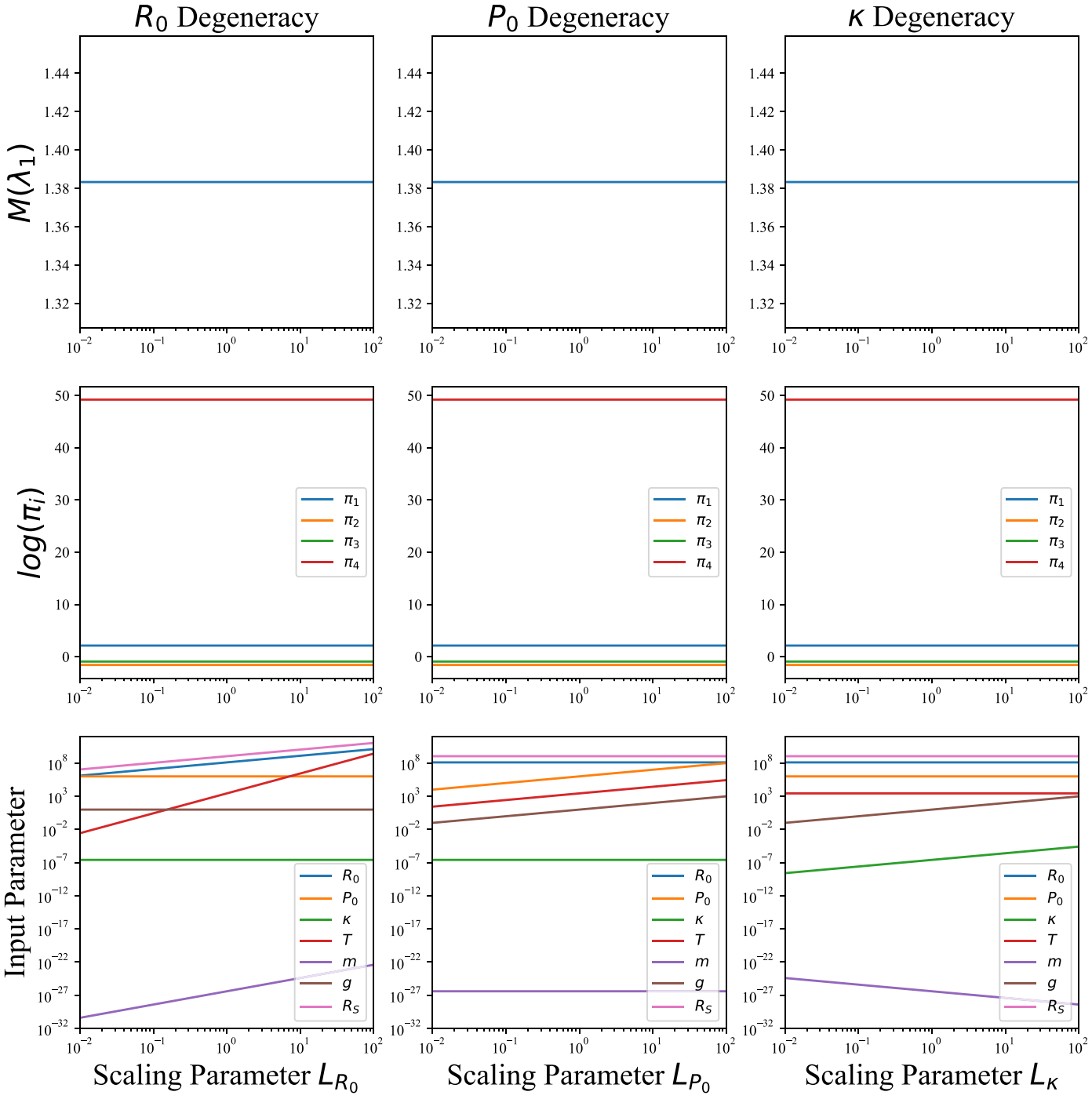

Figure 5. The same as Figure 4, but this time varying the input parameters according to Equation (14). In the lower left, lower middle, and lower right panels, the parameters are varied according to Equations 14(a)–(c), respectively. The plots in the middle and top rows confirm that these transformations are indeed true degeneracies.

parameter space where the training data were generated. In addition, we also generate values for π_3 and π_4 (sampled uniformly between -1 and 1) even though the function $f_{\rm true}$ does not explicitly depend on them—instead, we let the symbolic regression figure out on its own that these are cyclic variables. In summary, the input to the symbolic regression is a set of 1000 instances of the form

$$(\pi_1, \pi_2, \pi_3, \pi_4, f_{\rm true}).$$

The result from a PySR run is a set of functions $f_{\rm fit}^{(C)}$ of increasing complexity C (defined as the number of leaf nodes in the binary tree representing the analytical expression for $f_{\rm fit}$). The result from one typical run for the fitted functions and the

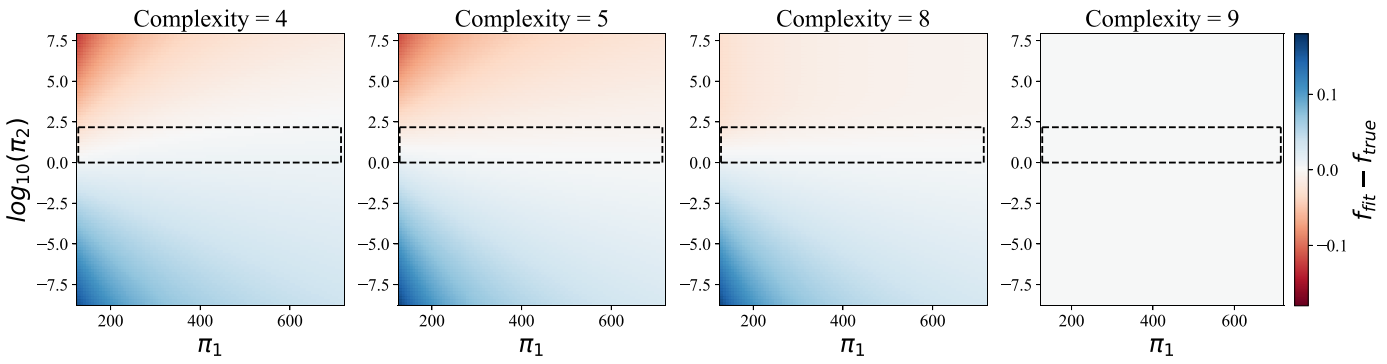

Figure 6. Heatmaps of the differences between the $f_{\rm fit}^{(C)}$ functions of Equation (27) obtained by symbolic regression and the target function $f_{\rm true}$, Equation (26). The rectangular box marked with a dashed line delineates the domain of values on which the symbolic regression was trained.

corresponding mean squared error (MSE) is

$$f_{\rm fit}^{(1)} = -0.106, \quad {\rm MSE} = 1.946031 \times 10^{-2}, \tag{27a}$$

$$f_{\rm fit}^{(4)} = \frac{0.456}{\sqrt{|\pi_1|}}, \quad {\rm MSE} = 8.314751 \times 10^{-5}, \tag{27b}$$

$$f_{\rm fit}^{(5)} = \frac{1.9413}{0.3009\ \pi_1} = \frac{6.452}{\pi_1}, \quad {\rm MSE} = 6.008056 \times 10^{-5}, \tag{27c}$$

$$f_{\rm fit}^{(8)} = \frac{\ln(|\pi_2 + 2.88\ \pi_1|)}{\pi_1}, \quad {\rm MSE} = 5.702897 \times 10^{-5}, \tag{27d}$$

$$f_{\rm fit}^{(9)} = \frac{\ln(4.4645\ |\pi_2|\ \sqrt{|\pi_1|})}{\pi_1}, \quad {\rm MSE} = 1.003324 \times 10^{-15}. \tag{27e}$$

We see that the symbolic regression was able to correctly determine that π_3 and π_4 are irrelevant variables—none of the proposed formulas involve π_3 or π_4. Of course, when we demand a low complexity like $C = 4$ or $C = 5$, the program is forced to make a choice between π_1 and π_2 to derive a good fit, because using both would exceed the set complexity threshold.

Figure 6 compares quantitatively the results from the fit, Equation (27), to the true function, Equation (26). We show heatmaps of the difference $(f_{\rm fit}^{(C)} - f_{\rm true})$ in the $(\pi_1, \log_{10}\pi_2)$ parameter space. The rightmost panel in the figure demonstrates that at complexity level 9, the symbolic regression was able to determine the target function exactly, to within the machine precision of 10^{-15}. Indeed, the original expression, Equation (26), can equivalently be rewritten as

$$f_{\rm true}(\pi_1, \pi_2, \pi_3, \pi_4) = \frac{\ln\left[(e^{\gamma_{\rm E}}\sqrt{2\pi})\pi_2\sqrt{\pi_1}\right]}{\pi_1} = \frac{\ln(4.4645\pi_2\sqrt{\pi_1})}{\pi_1}, \tag{28}$$

in perfect agreement with Equation 27(e). The result, Equation 27(e), is also favored by Occam's razor because the MSE drops significantly from complexity 8 to complexity 9. The other three panels in Figure 6 show that, as expected, at lower complexities the symbolic regression is unable to fit the data perfectly, although the fit is good within the training region identified with the dashed rectangular box.

6. Conclusions and Outlook

Observation of planetary transits at different wavelengths is a widely used technique to extract information about the structure and the composition of the atmosphere of an exoplanet. As shown in Figure 1, the methods used to constrain the atmospheric parameters can be generally divided into two groups:

1. Computational studies based on forward radiative transfer models and numerical inversion techniques, which increasingly use different novel statistical and machine-learning methods to improve the accuracy, the precision, and the speed of the performed retrievals.
2. Analytical investigations, which, provided with a set of simplifying assumptions, derive an analytical expression that allows for a better understanding of the underlying physics.

In this paper we demonstrate a novel approach to the problem (the yellow path in Figure 1) that leverages the advantages of these two approaches by harvesting the data generated by the complex and detailed forward models while preserving the ability of analytical expressions to provide physical insight into the problem.

First, we perform a formal dimensional analysis on the modulated stellar spectrum $M(\lambda)$ during planetary transit as a function of the seven input atmospheric parameters listed in Table 1:

1. We show that $M(\lambda)$ depends on only three dimensionless groups:
$$\pi_1' = \frac{R_0 mg}{k_B T}, \qquad \pi_2 = \frac{P_0\kappa}{g}, \qquad \pi_3 = \frac{R_0}{R_S}.$$
2. We mathematically demonstrate that the transit spectrum suffers from a number of degeneracies summarized in Equation (25). The simplest two- and three-level degeneracies are collected in the Appendix and are illustrated in Figure 7. The higher-level degeneracies that involve more than three parameters can be obtained from the Equation set (25). Some of the degeneracies have previously been noted in the literature, but we base our study on mathematically rigorous grounds, which allows us to find all theoretically possible degeneracies between the relevant planetary parameters.
3. We discuss methods for lifting the degeneracies by using additional observations or theoretical/model constraints

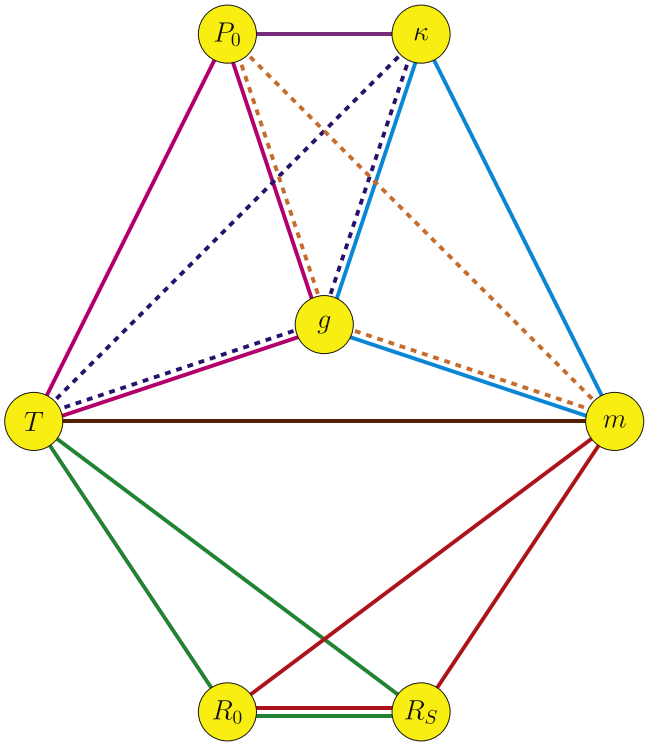

Figure 7. A pictorial representation of the two-level (line segments) and three-level (triangles) degeneracies discussed in the text. Each color corresponds to a different set of degenerate parameters, i.e., parameters, which are not uniquely fixed by the experimental observation (two triangles were drawn with dashed lines to facilitate identification).

on the stellar radius, gravity acceleration of the planet, and/or the mean molecular mass of the planet atmosphere.

4. The dimensionality reduction of the input parameter space achieved by our dimensional analysis helps greatly to optimize the performance of the symbolic regression algorithm used in the second half of the paper.

In the second part of the paper, as a proof of concept, we demonstrate the use of symbolic regression to derive analytical expressions describing the relationship between the modulated stellar spectrum $M(\lambda)$ and the input atmospheric parameters. The regression takes advantage of the dimensionality reduction accomplished by our dimensional analysis, and is performed on a synthetic data set generated with the analytical model of Heng & Kitzmann (2017), in analogy to Márquez-Neila et al. (2018). The algorithm correctly identifies both the relevant and the irrelevant variables, and provides analytic approximations of increasing complexity C, eventually reproducing the correct expression perfectly at $C = 9$ (to within the machine precision).

The analysis presented in this paper can be extended in several interesting directions.

1. Expand the dimensional analysis by including additional variables that account for further details such as temperature gradients and planet rotation, which would introduce new possible degeneracies.
2. Perform the symbolic regression on a suitable data set generated by a full forward radiative transfer model scanning all relevant parameters (π groups).
3. Explore other modern techniques for symbolic regression that are adaptable to high dimensional data.

We thank K. Kong (University of Kansas) for collaboration in the early stages of this work. This work was supported in part by the United States Department of Energy under grant No. DESC0022148.

Software: `jupyter` (Kluyver et al. 2016), `matplotlib` (Hunter 2007), `numpy` (van der Walt et al. 2011), `pysr` (Cranmer et al. 2020), `scipy` (Virtanen et al. 2020).

Appendix
List of Two-level and Three-level Degeneracies

In this appendix we list the simplest degeneracies implied by Equation (25), i.e., the degeneracies involving the fewest number of atmospheric parameters (see Figure 7 for a pictorial summary).

We start with the two-level degeneracies:

1. Pressure–opacity degeneracy (deep purple line segment in Figure 7). Heng & Kitzmann (2017) pointed out a degeneracy between the pressure P_0 and the mass-mixing ratio that is incorporated into the calculation of κ. We can reproduce this degeneracy from Equation (25) by taking $L_{R_0} = 1$ and $L_\kappa = L_T = L_{P_0}^{-1} \equiv L^{-1}$, which results in
$$P_0 \to L \times P_0, \quad \kappa \to L^{-1} \times \kappa. \tag{A1}$$
2. Temperature–mean molecular mass degeneracy (brown line segment in Figure 7). This is the accidental degeneracy shown in our previous result, Equation (24):
$$T \to L \times T, \quad m \to L \times m. \tag{A2}$$

There are also six three-level degeneracies:

1. Pressure–temperature–gravity degeneracy (solid magenta triangle in Figure 7). This is one of the guaranteed degeneracies shown in our previous result, Equation 14(b),
$$P_0 \to L \times P_0, \quad T \to L \times T, \quad g \to L \times g. \tag{A3}$$
2. Opacity–temperature–gravity degeneracy (dashed blue triangle in Figure 7). By choosing $L_\kappa = L_T \equiv L$ and $L_{R_0} = L_{P_0} = 1$ in Equation (25), we obtain
$$\kappa \to L \times \kappa, \quad T \to L \times T, \quad g \to L \times g. \tag{A4}$$
3. Pressure–mass–gravity degeneracy (dashed orange triangle in Figure 7). By choosing $L_{P_0} = L_T^{-1} \equiv L$ and $L_{R_0} = L_\kappa = 1$ in Equation (25), we obtain
$$P_0 \to L \times P_0, \quad m \to L^{-1} \times m, \quad g \to L \times g. \tag{A5}$$
4. Opacity–mass–gravity degeneracy (solid cyan triangle in Figure 7). This is another one of the guaranteed degeneracies discussed earlier in Equation 14(c),
$$\kappa \to L \times \kappa, \quad m \to L^{-1} \times m, \quad g \to L \times g. \tag{A6}$$
5. Planet radius–star radius–temperature degeneracy (solid green triangle in Figure 7). By choosing $L_{R_0} = L_T^{-1/2} \equiv L$ and $L_{P_0} = L_\kappa = 1$ in Equation (25), we obtain
$$R_0 \to L \times R_0, \quad T \to L \times T, \quad R_S \to L \times R_S. \tag{A7}$$
6. Planet radius–star radius–mass degeneracy (solid red triangle in Figure 7). By choosing $L_{R_0} = L_T^{-1/3} \equiv L$ and

$L_{P_0} = L_\kappa = 1$ in Equation (25), we obtain

$$R_0 \to L \times R_0, \quad m \to L^{-1} \times m, \quad R_S \to L \times R_S. \tag{A8}$$

These simple degeneracies can be combined together to form degeneracies involving more than three atmospheric parameters. For example, the previously discussed four-level degeneracy in Equation 14(a) can be obtained as a combination of Equations (A2), (A7), and (A8).

ORCID iDs

Konstantin T. Matchev https://orcid.org/0000-0003-4182-9096
Katia Matcheva https://orcid.org/0000-0003-3074-998X
Alexander Roman https://orcid.org/0000-0003-2719-221X

References

Ardevol Martinez, F., Min, M., Kamp, I., & Palmer, P. I. 2022, arXiv:2203.01236
Arechiga, N., Chen, F., Chen, Y.-Y., et al. 2021, arXiv:2112.04023
Barenblatt, G. I. 1996, Scaling, Self-similarity, and Intermediate Asymptotics: Dimensional Analysis and Intermediate Asymptotics (Cambridge: Cambridge Univ. Press)
Barstow, J. K., & Heng, K. 2020, SSRv, 216, 82
Battaglia, P. W., Pascanu, R., Lai, M., Rezende, D., & Kavukcuoglu, K. 2016, arXiv:1612.00222
Benneke, B., & Seager, S. 2012, ApJ, 753, 100
Bétrémieux, Y., & Swain, M. R. 2017, MNRAS, 467, 2834
Blecic, J., Harrington, J., Cubillos, P. E., et al. 2022, PSJ, 3, 82
Bongard, J., & Lipson, H. 2007, PNAS, 104, 9943
Brown, T. M. 2001, ApJ, 553, 1006
Buckingham, E. 1914, PhRv, 4, 345
Burrows, A., Sudarsky, D., & Hubbard, W. B. 2003, ApJ, 594, 545
Chang, M. B., Ullman, T., Torralba, A., & Tenenbaum, J. B. 2016, arXiv:1612.00341
Changeat, Q., Edwards, B., Waldmann, I. P., & Tinetti, G. 2019, ApJ, 886, 39
Charbonneau, D., Brown, T. M., Latham, D. W., & Mayor, M. 2000, ApJL, 529, L45
Cobb, A. D., Himes, M. D., Soboczenski, F., et al. 2019, AJ, 158, 33
Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., et al. 2020, arXiv:2006.11287
Cubillos, P. E., Harrington, J., Blecic, J., et al. 2022, PSJ, 3, 81
de Wit, J., & Seager, S. 2013, Sci, 342, 1473
Duff, M. J. 2015, ConPh, 56, 35
Fisher, C., & Heng, K. 2018, MNRAS, 481, 4698
Fisher, C., Hoeijmakers, H. J., Kitzmann, D., et al. 2020, AJ, 159, 192
Fortney, J. J. 2005, MNRAS, 364, 649
Giles, R. S., Fletcher, L. N., & Irwin, P. G. J. 2015, Icar, 257, 457
Griffith, C. A. 2014, RSPTA, 372, 20130086
Guzmán-Mesa, A., Kitzmann, D., Fisher, C., et al. 2020, AJ, 160, 15
Harrington, J., Himes, M. D., Cubillos, P. E., et al. 2022, PSJ, 3, 80
Heng, K. 2019, MNRAS, 490, 3378
Heng, K., & Kitzmann, D. 2017, MNRAS, 470, 2972
Heng, K., & Showman, A. P. 2015, AREPS, 43, 509
Himes, M. D., Cobb, A. D., Soboczenski, F., et al. 2020a, AAS Meeting Abstracts, 235, 343.01
Himes, M. D., Cobb, A. D., Wright, D. C., Scheffer, Z., & Harrington, J. 2020b, MARGE: Machine Learning Algorithm for Radiative Transfer of Generated Exoplanets, Astrophysics Source Code Library, ascl:2003.010
Himes, M. D., Harrington, J., Cobb, A. D., et al. 2020c, arXiv:2003.02430
Hubbard, W. B., Fortney, J. J., Lunine, J. I., et al. 2001, ApJ, 560, 413
Hunter, J. D. 2007, CSE, 9, 90
Iten, R., Metger, T., Wilming, H., del Rio, L., & Renner, R. 2020, PhRvL, 124, 010508
Kitzmann, D., Heng, K., Oreshenko, M., et al. 2020, ApJ, 890, 174
Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in Positioning and Power in Academic Publishing: Players, Agents and Agendas, ed. F. Loizides & B. Scmidt (Amsterdam: IOS Press), 87
Kokar, M. 1986, Mach. Learn., 1, 403
Langhaar, H. L. 1951, Dimensional Analysis and Theory of Models (New York: Wiley)
Langley, P. 1977, in Proc. of the 5th Int. Joint Conf. on Artificial Intelligence (Burlington, MA: Kaufmann)
Langley, P., Simon, H. A., & Bradshaw, G. L. 1987, in Heuristics for Empirical Discovery, ed. L. Bolc (Berlin: Springer), 21
Langley, P., & Zytkow, J. M. 1989, Artif. Intell., 40, 283
Lemos, P., Jeffrey, N., Cranmer, M., Ho, S., & Battaglia, P. 2022, arXiv:2202.02306
MacDonald, R. J., & Lewis, N. K. 2022, ApJ, 929, 20
Madhusudhan, N. 2019, ARA&A, 57, 617
Márquez-Neila, P., Fisher, C., Sznitman, R., & Heng, K. 2018, NatAs, 2, 719
Matchev, K. T., Matcheva, K., & Roman, A. 2022a, arXiv:2203.06299
Matchev, K. T., Roman, A., & Shyamsundar, P. 2022b, ScPP, 12, 104
Matcheva, K. I., Conrath, B. J., Gierasch, P. J., & Flasar, F. M. 2005, Icar, 179, 432
Nixon, M. C., & Madhusudhan, N. 2020, MNRAS, 496, 269
Oreshenko, M., Kitzmann, D., Márquez-Neila, P., et al. 2020, AJ, 159, 6
Schmidt, M., & Lipson, H. 2009, Sci, 324, 81
Schneider, J. 1994, Ap&SS, 212, 321
Seager, S., & Sasselov, D. D. 2000, ApJ, 537, 916
Todorovski, L., & Dzeroski, S. 1997, in Proc. Fourteenth Int. Conf. on Machine Learning (Burlington, MA: Morgan Kaufmann), 376
Udrescu, S.-M., & Tegmark, M. 2020, SciA, 6, eaay2631
Vahidinia, S., Cuzzi, J. N., Marley, M., & Fortney, J. 2014, ApJL, 789, L11
van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, CSE, 13, 22
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261
Welbanks, L., & Madhusudhan, N. 2019, AJ, 157, 206
Welbanks, L., & Madhusudhan, N. 2021a, ApJ, 913, 114
Welbanks, L., & Madhusudhan, N. 2021b, arXiv:2112.09125
Yip, K. H., Changeat, Q., Nikolaou, N., et al. 2021, AJ, 162, 195
Zembowicz, R., & Żytkow, J. M. 1992, in Proc. Tenth National Conf. on Artificial Intelligence, AAAI'92 (Palo Alto, CA: AAAI Press), 70
Zingales, T., & Waldmann, I. P. 2018, AJ, 156, 268