Machine-assisted classification of potential biosignatures in earth-like exoplanets using low signal-to-noise ratio transmission spectra

David S. Duque-Castaño,[1]⋆ Jorge I. Zuluaga,[1]† Lauren Flor-Torres[1]‡

[1]*SEAP/FACom, Instituto de Física - FCEN, Universidad de Antioquia, Calle 70 No. 52-21, Medellín, Colombia*

ABSTRACT

The search for atmospheric biosignatures in earth-like exoplanets is one of the most pressing challenges in observational astrobiology. Detecting biogenic gases in terrestrial planets requires high resolution and long integration times. In this work, we developed and tested a machine-learning general methodology, intended to classify transmission spectra with low Signal-to-Noise Ratio according to their potential to contain biosignatures. For that purpose, we trained a set of models capable of classifying noisy transmission spectra, as having methane, ozone, and/or water (multilabel classification), or simply as being interesting for follow-up observations (binary classification). The models were trained with $\sim 10^6$ synthetic spectra of planets similar to TRAPPIST-1 e which were generated with the package `MultiREx` especially developed for this work. The trained algorithms correctly classified test planets with transmission spectra having SNR<6 and containing methane and/or ozone at mixing ratios similar to those of modern and Proterozoic Earth. Tests on realistic synthetic spectra based on the current Earth's atmosphere show at least one of our models would classify as likely having biosignatures and using only one transit, most of the inhabited terrestrial planets observed with the JWST/NIRSpec PRISM around M-dwarfs located at distances similar or smaller than that of TRAPPIST-1 e. The implication of this result for the designing of observing programs and future surveys is enormous since machine-assisted strategies similar to those presented here could significantly optimize the usage of JWST resources for biosignature searching, while maximizing the chances of a real discovery after dedicated follow-up observations of promising candidates.

Key words: astrobiology – planets and satellites: terrestrial planets – exoplanets – techniques: spectroscopic – planets and satellites: TRAPPIST-1e – planets and satellites: atmospheres

1 INTRODUCTION

One of the strategies conceived to search for life beyond Earth involves detecting signals on exoplanets that may be closely associated with life and have a low probability of being abiotic in origin. We call these signals, biosignatures (see e.g. Schwieterman et al. 2018; Des Marais et al. 2008 and references therein). Currently, our instrumental capabilities allow us to detect biosignatures that have a global impact, especially those concentrated in the atmosphere and observable through spectroscopy (see e.g. Schwieterman et al. 2018).

However, detecting biosignatures on rocky exoplanets within the habitable zone of their stars poses significant challenges due to the low signal-to-noise ratio resulting from the smaller relative radii of the star-to-planet. M-dwarfs offer an opportunity to study the atmospheres of Earth-like planets. Although these stars may exhibit high X-ray and ultraviolet (XUV) activity in their early stages, secondary atmospheres may remain stable as they age (France et al. 2020). The fact is that M-dwarfs allow for greater transit depth compared to other types of stars (Wunderlich et al. 2019) increasing the chances of detecting and analysing exoplanetary atmospheres. Instruments such as JWST demonstrate our ability to reliably detect atmospheric species on exoplanets around M-dwarfs, such as possible H2-rich atmospheres (Madhusudhan et al. 2023; Benneke et al. 2024), and even showcase the capability to identify, as the most likely scenario, Earth-like atmospheres such as N2- or CO2-dominated atmospheres (Cadieux et al. 2024).

Recent simulations on the detectability of an Earth-like atmosphere using the James Webb Space Telescope (JWST) (Barstow & Irwin 2016; Wunderlich et al. 2019; Lin et al. 2021; Lustig-Yaeger et al. 2023) have revealed a challenging scenario. To detect robust biosignatures such as ozone (O_3), requires a large number of transits (e.g., up to 200 transits in the case of TRAPPIST-1 e; see Lin et al. 2021) to achieve statistically significant detections. Despite this challenge, detecting methane (CH_4) and water vapour (H_2O) presents a promising opportunity. Studies have demonstrated that using a reasonable number of transits, the presence of these atmospheric species, which are typically associated with a global biosphere, can be retrieved (Wunderlich et al. 2019; Lin et al. 2021; Lustig-Yaeger et al. 2023). However, it is important to note that CH_4 is a less robust biosignature compared to O_3 (for more details on the robustness of ozone as a biosignature, see subsection 4.2).

⋆ E-mail: dsantiago.duque@udea.edu.co, dasan.academico@gmail.com
ORCID: 0000-0003-3614-7904
† ORCID: 0000-0002-6140-3116
‡ ORCID: 0000-0003-4134-9615

Given the almost prohibitive cost of performing such an observational campaign, a more relevant strategy may be allocating JWST time to conduct a low signal-to-noise ratio (SNR) survey. Although this may not allow for statistically significant retrievals, it would at least enable planning for future follow-up observations of interesting targets with current and future more powerful telescopes (e.g., ELT, LUVOIR, HabEx, Roman, ARIEL).

In this paper, we propose and numerically test the methods and tools to support such an strategy. For this purpose we design and test a set of machine learning (ML) methods and tools aimed at labelling low SNR spectra that have the potential to contain interesting biosignatures. It is important to stress that the tools presented here are not designed to perform a retrieval of the abundances of chemical species, but to identify interesting candidates for follow-up observations.

The use of Machine Learning (ML) for analysing exoplanet spectra is gaining increasing interest. Multiple algorithms have been developed and trained on diverse datasets and objectives (Marquez-Neila et al. 2018; Soboczenski et al. 2018; Zingales & Waldmann 2018; Cobb et al. 2019; Nixon & Madhusudhan 2020). Primarily, these algorithms have been used to replace the retrieval process based on Bayesian methods, which is characterised by its high computational cost (Munsaket et al. 2021; Ardévol Martínez et al. 2022; Vasist et al. 2023; Ardévol Martínez et al. 2024). In Table 1 we summarise some of the strategies that have been devised for performing atmospheric composition retrieval and are available in the literature.

Additionally, the capacity of ML to support parts of the retrieval process, such as suggesting priors (Hayes et al. 2020), identifying molecules (Waldmann 2016), or using neural network parametrization of pressure-temperature profiles for better efficiency and physical consistency (Gebhard et al. 2024), or even replacing the spectrum generation process through radiative transfer, has been explored (Himes et al. 2022). Conversely, unsupervised learning techniques have also been proposed for analysing spectroscopic data and detecting anomalous chemical compositions (Forestano et al. 2023; Matchev et al. 2022a). ML has been used to assess molecule detection capabilities in various instruments of the James Webb Space Telescope (JWST), demonstrating how machine learning techniques have become robust alternatives to classical retrieval methods (Guzmán-Mesa et al. 2020).

Despite these advancements, traditional ML methods are often seen as black boxes, where the internal workings and the features driving the models are not easily interpretable, limiting the ability to achieve physical understanding. Efforts have been made to address this issue. For instance, Yip et al. (2021) developed tools to elucidate the functioning of ML methods when making predictions. Similarly, Matchev et al. (2022b) demonstrated the use of symbolic regression for characterising transit spectra, enabling a physical understanding of the problem.

Furthermore, these studies have primarily focused on gaseous planets, especially Hot Jupiters like WASP-12. There is a clear minority of studies on rocky and Earth-like planets, indicating a significant gap in the training data available for these types of planets. Nonetheless, recent research has started to consider brown dwarfs, as seen in the work of Ardévol Martínez et al. (2024) and Lueber et al. (2023). The implemented regression techniques commonly include Random Forest and various types of neural networks. Additionally, unsupervised learning techniques are particularly focused on clustering and anomaly detection.

Based on our literature review, to date, only one study has been conducted specifically on classifying potential molecules in spectra (Waldmann 2016). The training datasets contain atmospheric compositions defined by C/O ratio and metallicity, or free chemistry, which includes molecules such as H_2O, NH_3, CH_4, CO_2, and CO (see Table 1).

As pointed out before, our proposal is not to perform the full retrieval, which, even using ML techniques, requires high SNR signals. This is demonstrated by the fact that many of the previously cited works are related to giant planets. But, if we can take low SNR spectra and label interesting Earth-like planets, we can focus our search on the best candidates.

This paper is organized as follows. In section 2, we describe the ML methods used in this paper, particularly introducing the language and quantities used in classification-supervised machine learning. To test our methods, we focus on a very interesting target, namely TRAPPIST-1 e, which is described in section 3. The section 4 is dedicated to describing the most interesting species on which our numerical experiments are focused. The methods, tools, and preliminary results are detailed in section 5 and section 6. Contextualization of the results based on the required transits using the NIRSpec MIRI instrument of JWST is discussed in section 7. Our algorithms are tested using realistic spectra of Earth in section 8. The limitations of our approach and the training data are discussed in section 9. Finally, the conclusions of our numerical experiments are drawn in section 10. All the tools and data developed for this paper are publicly available, with access information provided in section 11.

2 SUPERVISED MACHINE LEARNING FOR BIOSIGNATURES DETECTION

Machine learning for classification (MLC) is an approach to building a predictive model that assigns a label or category to a data point or instance based on a set of known features (see e.g., Géron 2023). For example, we aim to label planets based on the presence of biosignatures using simple categories: O_3/No-O_3, CH_4/No-CH_4.

In MLC, a training set consists of labelled observations (e.g., a list of planets with and without biosignatures), where each observation comprises features and a classification label. The simplest form is binary classification, where an instance either belongs to a category or does not. The goal is for the classificator model to accurately predict the classification label for new, unlabelled data, enabling decisions based on the classification of such data. Our objective is for the trained algorithm to identify which planets have a specific biosignature and which do not, using a low SNR spectrum.

2.1 Random forest for biosignature classification

Many algorithms have been devised for MLC. Here, we focus on Random Forest (RF), a family of methods that is not only simple enough but also particularly well-suited to our requirements. RF demonstrates robustness to noise by maintaining high accuracy through the aggregation of predictions from multiple trees, thereby reducing the impact of noisy instances. Additionally, RF is inherently resistant to overfitting due to the Law of Large Numbers, which ensures the convergence of the model's error as more trees are added (Breiman 2001). This characteristic makes RF especially effective for complex datasets with numerous variables, each contributing marginally to the prediction, providing a robust solution for high-dimensional data classification.

In our case, RF takes signals from various spectral bands (these are the features) and creates decision trees based on the values of these specific signals. Each decision tree in the forest can use multi-

Table 1. Studies on the use of ML algorithms in exoplanet spectra. We describe the purpose and application of the algorithms used in each study. Additionally, we list the exoplanet regimes each algorithm focuses on and indicate whether they were trained or tested with C/O ratios plus metallicity (represented in the column C/O), or the most common free chemistry molecules.

Reference	ML Method(s)	Targets	C/O	H_2O	CO_2	CO	CH_4	NH_3	HCN	O_3
Gebhard et al. (2024)	Convolutional Neural Networks and Multilayer Perceptrons (Regression)	Hot Jupiters and Earth-like Planets								
Ardévol Martínez et al. (2024)	Sequential Neural Posterior Estimation with Normalizing Flows (Regression)	Wide range of planets, Brown Dwarfs	■	■	■					
Forestano et al. (2023)	LOF, 1CSVM (Unsupervised ML, Outlier detection)	Hot Jupiters		■						
Vasist et al. (2023)	Neural Posterior Estimation with Normalizing Flows (Regression)	Gas Giants	■							
Ardévol Martínez et al. (2022)	Convolutional Neural Networks (Regression)	Gas Giants	■							
Himes et al. (2022)	Convolutional Neural Networks (Regression)	Hot Jupiters	■							
Matchev et al. (2022a)	k-means, PCA and ISOMAP (Unsupervised)	Hot Jupiters		■	■		■	■		
Matchev et al. (2022b)	Symbolic Regression	Hot Jupiters	■							
Munsaket et al. (2021)	Random Forest (Supervised ML, Regression)	Hot Jupiters	■							
Yip et al. (2021)	Multilayer Perceptrons, Convolutional Neural Networks and Long Short-Term Memory Networks (Regression)	Wide range of planets		■	■	■	■			
Guzmán-Mesa et al. (2020)	Random Forest (Regression)	Warm Neptunes		■				■		
Hayes et al. (2020)	k-means, PCA (Unsupervised classification)	Jupiter-like planets								
Nixon & Madhusudhan (2020)	Random Forest (Regression)	Hot Jupiters		■	■	■	■		■	
Cobb et al. (2019)	Bayesian Neural Network (Regression)	Hot Jupiters		■	■	■	■	■	■	■
Soboczenski et al. (2018)	Convolutional Neural Networks with Monte Carlo Dropout (Regression)	Rocky terrestrial exoplanets		■	■	■				■
Marquez-Neila et al. (2018)	Random Forest (Regression)	Hot Jupiters		■	■	■	■			
Zingales & Waldmann (2018)	Generative Adversarial Networks (Regression)	Hot Jupiters	■	■			■			
Waldmann (2016)	Deep-Belief Networks (Classification)	Wide range of planets		■	■		■			

ple features simultaneously to make decisions, determining if these signals contribute to identifying a biosignature.

2.2 Metrics for classification

In classification problems, the performance of a machine learning algorithm is not only measured in terms of how many planets were correctly classified ("True" cases), i.e. the accuracy or how many were not ("False" cases), i.e. the errors. It is also required to categorize the classification results into positives and negatives, either True Positives (TP) or True Negatives (TN), which are related to accuracy; and into False Positives (FP) or False Negatives (FN), which are the error. This categorization is summarised in what is called a *confusion matrix* (Kelleher et al. 2015).

In Figure 1 we schematically represent the categorization in a confusion matrix and, more importantly, illustrate what each case means in the context of potentially biosignature-bearing exoplanets. In the diagram, we introduce the category *interesting* to distinguish planets that deserve follow-up observations or in-depth analysis. We should recall again that is the focus of this work: we do not aim at detecting biosignatures using ML but at labeling planets that are interesting or not.

Different types of errors carry specific implications for biosignature searching. For example, mistakenly classifying a planet that contains biosignatures as not interesting (FN), could mean the loss of valuable research opportunities. In contrast, erroneously considering

a planet without biosignatures as interesting (FP) could lead to inefficient resource allocation. Therefore, understanding the difference between these types of errors is essential for evaluating the accuracy of a model and for adjusting decision and exploration strategies based on the model's classification outcomes (see section 6).

Given the limitations of using accuracy alone, i.e. number of right predictions divided by the total number of predictions, especially when different errors have disproportionately large impacts, more nuanced metrics become indispensable (Kelleher et al. 2015). For instance, *recall* which is defined as

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

measures the model's efficiency in detecting all relevant instances. Ideally, a Recall = 1 means that the algorithm correctly classifies all planets actually having biosignatures (first row in Figure 1) as interesting exoplanets. In this case, the follow-up observations will be maximally successful. For this reason, we call Recall a *wasting metric*: when Recall decreases, we are excluding promising planets from future observational campaigns and wasting research opportunities.

On the other hand, we have the metric *Precision*, which is defined as

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

that assesses the accuracy of interesting predictions. This metric
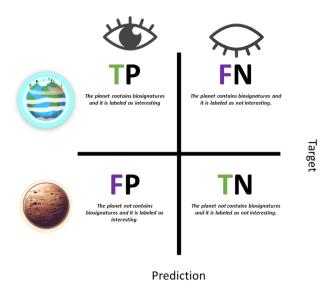
CONFUSION MATRIX

Figure 1. Schematic representation of the confusion matrix used in ML classification but for the context of biosignature searching. Instances (planets) in the first raw have biosignatures, while those in the second one are devoid of them. An algorithm labels instances (planets) in the first column as potentially having biosignatures (interesting), while the algorithm does not consider those in the second column as interesting.

is more difficult to interpret since it focuses on the classification rather than the actual presence of biosignatures. When the number of planets without biosignatures classified as interesting (FP) decreases, the Precision increases. Therefore, we call Precision a *time-saving metric*, as it helps minimize the time spent studying planets without biosignatures.

In summary, we expect that an effective algorithm applied to a given dataset would have a high value for both metrics, i.e., it will be minimally wasteful and maximally time-saving.

But there is a third metric, one quantifying another aspect of the problem: the discovery opportunity. For this purpose, we use the *F1 Score*, defined as:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

In the worst-case scenario, F1 Score = 0, the algorithm is absolutely incapable of classifying as interesting planets those that already have biosignatures. On the other extreme, the ideal case, F1 Score = 1, occurs when the algorithm has Recall = Precision = 1. Starting from this ideal case, the F1 Score decreases if the number of FN increases; i.e., this metric is very sensitive to the case when a planet with a biosignature is classified as not interesting, losing the discovery opportunity. This is the reason why we call the F1 Score a *discovery metric*.

All of the previous metric are focused on positive classifications. However, for the specific purpose of searching for biosignatures, and especially for studying the potential confusion that can arise when an algorithm is designed to detect a specific biosignature in the presence of other molecules, we should use metrics focused on the negative cases. Thus, for example, the *True Negative Rate* (TNR):

$$\text{TNR} = \frac{TN}{FP + TN} \tag{4}$$

We call this metric a *confusion metric*, in the sense that when the TNR decreases, the algorithm labels planets without biosignatures as interesting. This confusion can arise, for instance, when a biologically irrelevant molecule is "identified" by the algorithm as a biosignature. Essentially, TNR is similar to recall but focuses on the negative labels, ensuring that planets without biosignatures are correctly identified as not interesting. Therefore, TNR can also be understood as a *negative recall*.

Since understanding the definition, meaning and application of a given metric could be very confusing, we have summarized them in Table 2 with special attention on highlight their role in context of biosignature search.

2.3 Multilabel classification

Multilabel classification expands the challenges of traditional classification by allowing each instance, i.e., each planet in the sample, to be linked with multiple labels simultaneously, e.g., potentially having ozone, potentially having CO_2, etc. This increases complexity by requiring models to manage interdependencies among multiple labels.

According to Sorower (2010), conventional methods are adapted for these scenarios by transforming the problem into multiple binary classification problems or by adapting existing algorithms to process multiple labels. Among the metrics for evaluating these models is Hamming Loss,

$$\text{Hamming Loss} = \frac{1}{N \times L} \sum_{i=1}^{N} \sum_{j=1}^{L} I(y_{ij} \neq \hat{y}_{ij}), \tag{5}$$

where N is the total number of instances (i.e., planets), L is the number of labels per instance (i.e., number of molecular species), and I is the indicator function that returns 1 if the condition is true and 0 otherwise. The condition for this metric is that the jth label the algorithm associates with the ith instance, $\hat{y}_{ij}$, is not the correct one, y_{ij}. For instance, if the algorithm predicts that planet 1 has molecule 2 when it does not, then $I(y_{12} \neq \hat{y}_{12}) = 1$.

Hamming Loss measures the fraction of incorrectly predicted labels. For example, if we need to determine if a planet has three molecular species and the algorithm incorrectly assigns two of the three labels, the Hamming Loss is $0.\bar{6}$. A large Hamming Loss generally indicates worse multilabel classification performance.

On the other hand, we can use Exact Match Ratio,

$$\text{Exact Match Ratio} = \frac{1}{N} \sum_{i=1}^{N} I(Y_i = \hat{Y}_i), \tag{6}$$

where the condition $\hat{Y}_i = Y_i$ implies that the algorithm correctly assigns all possible labels to the ith instance.

It is worth mentioning that, besides evaluating Hamming Loss and Exact Match Ratio in a multilabel context, we can combine the metrics defined in previous subsections to compute averages per label, known as *macro averaging*. For instance, if we are classifying multiple molecular species in a sample of planets, we can independently calculate the F1 Score for each molecule and then average them.

In summary, evaluating the performance of a classification algorithm is not a trivial problem. Depending on which aspects of the

Table 2. Summary of metrics used in the evaluation of biosignature detection algorithms. The table includes the metric name, a nickname describing its utility, its specific use in the context of evaluating planets for biosignatures, and extreme cases illustrating the possible values and interpretations of each metric.

Metric	Nickname	Utility	Extreme cases
Recall	Wasting metric	When small we are excluding promising planets from future observational campaigns, and wasting research opportunities	• Recall = 1: all planets with biosignatures are labelled as interesting • Recall = 0: no planets with biosignatures are labelled as interesting
Precision	Time-saving metric	Helps to minimize the time spent studying planets without biosignatures	• Precision = 1: All planets labelled as interesting truly have biosignatures. • Precision = 0: Any planet is labelled as interesting even without biosignatures. **Note:** Even if the false positives are at their maximum, meaning all non-interesting planets are incorrectly labelled as interesting, the precision is not zero as long as there are true positives. This occurs because precision is calculated as the ratio of true positives to the sum of true positives and false positives. Thus, the presence of any correctly identified interesting planets ensures that precision remains above zero.
F1 Score	Discovery metric	Quantifies the discovery opportunity by balancing Precision and Recall. MAXIMIZE the opportunity to achieve a successful discovery	• F1 Score = 1: when Recall = Precision = 1, meaning the algorithm has perfect precision and Recall, successfully identifying all interesting planets without any false positives • F1 Score = 0: algorithm cannot classify any interesting planets correctly Note: When TNR = 0, F1 Score is not necessarily very low if TP is high.
TNR	Confusion metric	Helps to ensure that planets without biosignatures are not mislabelled as interesting, thereby avoiding wasted resources on unpromising candidates	• TNR = 1: all non-interesting planets are correctly identified as not interesting • TNR = 0: all non-interesting planets are incorrectly labelled as interesting
Hamming Loss	Error rate metric	It measures the fraction of incorrectly predicted molecules, indicating the algorithm's deficiency in misclassifying or omitting the presence of molecules in general.	• Hamming Loss = 0: All predicted labels (biosignatures) are correct. • Hamming Loss = 1: All predicted labels (biosignatures) are incorrect.
Exact Match Ratio	Perfect match metric	Measures how well the algorithm classifies all biosignatures for planets correctly at the same time. This metric is very demanding as it requires perfect classification for all labels. Higher values indicate better performance.	• Exact Match Ratio = 1: The algorithm correctly assigns all biosignatures to all planets. • Exact Match Ratio = 0: The algorithm fails to assign the correct set of biosignatures to any planet.

classification we want to focus different metrics can be used. In this work we will apply a combination of binary and multilabel classification metrics to assess the capabilities of our algorithms.

3 THE CASE OF TRAPPIST-1 AND ITS PLANETS

In order to test our methods and techniques, we need to select a proper investigation case that lies between an idealized illustration-purposed planet and a true target. After considering several possibilities, we found that the case of TRAPPIST-1 e fulfills the conditions for our numerical experiments. In the following paragraphs, we will present a summary of the potential that the TRAPPIST-1 system in general and TRAPPIST-1 e, in particular, have in the search for biosignatures.

The TRAPPIST-1 system has gained significant scientific attention in recent years, especially in planetary sciences and astrobiology, owing to its exceptional features. The star, with a spectral type of $M8.0 \pm 0.5$ (Gillon et al. 2016), is known for having the highest number of rocky planets discovered. This makes it an ideal candidate

for atmospheric detection, especially because, besides other stars within 40 pc and using JWST, we can achieve favorable signal-to-noise ratio levels (SNR ∼ 5.5) in a not-unrealistic observing time (200 h) (Gillon et al. 2020).

Despite being approximately 7.6 billion years old (Burgasser & Mamajek 2017) and displaying moderate magnetic activity compared to other stars of its class and age (Gillon 2024), the emission of XUV radiation remains significantly high, suggesting that it could potentially erode the atmospheres of the surrounding planets (Wheatley et al. 2017; Bourrier et al. 2017).

The orbital configuration of the TRAPPIST-1 system suggests that they may have migrated from farther regions, potentially leading to a rich accumulation of volatiles (Huang & Ormel 2022; Agol et al. 2021). This makes it an excellent laboratory to study rocky planets near ultracool dwarfs and their atmospheric retention capacity (Airapetian et al. 2020). This is important for understanding habitability around such stars, which may host most potentially habitable planets in our galaxy (Gillon 2024).

Research on atmospheric stability suggests that these planets underwent phases of runaway greenhouse effect and intense desiccation due to stellar radiation before the star reached the main sequence (Luger & Barnes 2015). Currently, planets e, d, and f reside within the conservative habitable zone, implying a conducive environment for maintaining secondary atmospheres. Models indicate the possible existence of atmospheres dominated by CO_2, such as those of Venus and Mars in the Solar System. Under more restricted conditions, their atmospheres can also be dominated by other molecular species, such as H_2O, O_2, N_2, CH_4, or NH_3 (Turbet et al. 2020). The case of Earth is one of a planet dominated by N_2.

Transmission spectroscopy using the JWST instruments has been identified as the most suitable method for detecting these atmospheres, with missions already scheduled to observe transits on each planet (Gillon et al. 2020; Turbet et al. 2020).

Regarding the composition of these exoplanets, measurements indicate they are primarily rocky, with TRAPPIST-1 e standing out for a density that makes it closest to Earth in comparative terms, though not identical in iron proportion or volatile content. These data allow for the proposition of liquid oceans based on the projection that TRAPPIST-1 e could have an iron fraction of 25% (Agol et al. 2021). The astrobiological interest in TRAPPIST-1 e is considerable, given its potential to present a wide range of atmospheric compositions, as the density suggests a high molecular mass atmosphere (Turbet et al. 2020; Lin et al. 2021), with the possibility of maintaining liquid surfaces under various atmospheric compositions and pressures (Turbet et al. 2018). This potential is enhanced by the estimation that the initial water loss, due to the runaway greenhouse effect, was significantly less compared to the inner planets of the system (Turbet et al. 2020).

Research into detecting Earth-like atmospheres in TRAPPIST-1 e using the JWST has motivated a variety of studies focused on specific aspects. On one hand, simulations of high-resolution spectra aim to replicate the spectral signature of Earth as seen from TRAPPIST-1 e, providing a crucial theoretical framework for identifying biosignatures (Kaltenegger et al. 2020; Lin & Kaltenegger 2022; Lustig-Yaeger et al. 2023). On the other hand, the capabilities of JWST instruments to detect specific biosignatures face significant challenges in detecting molecules like O_3 and N_2O, which would exceed the telescope's operational time. However, other authors highlight the viability of identifying CH_4, CO_2 and H_2O in a limited number of transits (Lin et al. 2021; Lustig-Yaeger et al. 2023). Additionally, research on the detection of stratospheric clouds concludes that, given the observational capabilities of the JWST, such detection is not yet feasible (Doshi et al. 2022).

In the following section, we enumerate and describe the biosignatures we have selected for our numerical experiments with TRAPPIST-1 e. As mentioned before, some of the biosignatures we select here have been identified as the most problematic to detect using the present observational capabilities and for that reason require novel techniques as that proposed here.

4 SELECTED BIOSIGNATURES

4.1 CH_4 and CO_2: primary and secondary biosignatures

Methane is highly common in reducing atmospheres, such as hydrogen atmospheres, and can be an anaerobic product. On Earth, methane is mainly produced by bacteria and plays a crucial role in the greenhouse effect, though it also comes from non-biological sources (Schwieterman et al. 2018). This molecule is considered a

biosignature, especially in oxidizing environments because oxidation processes with OH constantly remove it (Grenfell 2018) and its presence could only indicate a chemical disequilibrium created by a complex process such as life. For instance, Schwieterman et al. (2018) have proposed that methane would be a biosignature when O_2 or O_3 are present in an atmosphere since they are highly oxidizing molecular species. However, other molecules, such as CO_2, which also reveal an oxidizing atmospheric state, could also support the case of the case of Methane as a biosignature.

It should be noted that CH_4 can also be continuously replenished on an uninhabited planet through the reaction of CO_2 with liquid water via serpentinization, a process involving the hydration of ultramafic minerals such as olivine (Grenfell 2018; Schwieterman et al. 2018). Therefore, while methane may not be a primary biosignature in this context, its presence still indicates the existence of liquid water and, consequently, of habitable conditions. Thus, methane can be considered a secondary biosignature in this scenario. Guzmán-Marmolejo et al. (2013) suggested that N_2-CO_2 dominated atmospheres with CH_4 concentrations greater than 10 ppmv could only be produced via biological processes, providing a method to distinguish between abiotic and biotic sources of methane.

4.2 O_2 y O_3: *The crown jewel*

Molecular oxygen on Earth is primarily the result of oxygenic photosynthesis. It has been proposed that O_2 is an excellent biosignature due to its scarce abiotic production under Earth-like conditions (Léger et al. 2011). However, other potential abiotic sources, such as the photolysis of H_2O or CO_2, depend on the host star, the state of the planet's atmosphere, and internal processes that limit the reabsorption of O_2 (Schwieterman et al. 2018). Even in the case of an abiotic source of O_2, other molecules may help to disentangle its origin.

To distinguish between biotic and abiotic O_2, one can look for specific spectral fingerprints. For instance, the simultaneous presence of CO and O_2 can indicate CO_2 photolysis, while strong O_2-O_2 CIA features can signify significant hydrogen loss and oxygen-enriched atmospheres. This hydrogen loss occurs through the photolysis of H_2O, where the lighter hydrogen escapes to space, leaving behind oxygen that accumulates in the atmosphere (Schwieterman 2016). The absence of N_2 might also suggest abiotic processes due to the lack of a cold trap that retains N_2 in the upper atmosphere, allowing it to escape more easily (Wordsworth & Pierrehumbert 2014).

But detecting O_2 in exoplanetary atmospheres can be challenging. Collision-induced absorption (CIA) by O_2 molecules produces a spectral signature that, although detectable, requires a relatively close proximity to our solar system. In the case of the JWST, for instance, a maximum distance of 5 parsecs is required for achieving a successful retrieval (Fauchez et al. 2020).

On the other hand, O_3 is formed from O_2 via photochemical processes and therefore acts as a secondary biosignature, i.e. its detection suggests the presence of oxygen and hence biological activity. The concentration levels in an exoplanet atmosphere would depend heavily on the UV emission from the host star. In cool M-dwarfs, there is a Goldilocks zone for ozone concentration that varies depending on the peak of UV emission (Grenfell et al. 2014).

Ozone exhibits a strong spectral signature in the near and medium-infrared (NIR and MIR, respectively), making it more detectable than O_2 (Schwieterman et al. 2018) at least for instruments like the JWST's NIRSpec and MIRI.

The simultaneous presence of O_2 (detected indirectly with O_3) and CH_4 in an exoplanet would indicate an atmosphere in redox disequilibrium. This combination could suggest an atmosphere actively

Table 3. Parameters of planet and star (Agol et al. 2021) .

TRAPPIST-1	
Radius ($R_\odot$)	0.0898
Mass ($M_\odot$)	0.1192
T_{eff} (K)	2566

TRAPPIST-1 e	
Radius ($R_\oplus$)	0.920
Mass ($M_\oplus$)	0.692
Semi-Major Axis (au)	0.02925

regulated by life, as these molecules react quickly to produce H_2O and CO_2 (Lin et al. 2021; Schwieterman et al. 2018).

Even though many more molecular species have been proposed as biosignatures (see e.g. Grenfell 2018; Schwieterman et al. 2018 and references therein), in our numerical experiments, we will focus on CH_4 and O_3 to demonstrate the capabilities of our approach. In future works, we expect to widen the set of molecular species for studying the detectability of other biosignatures.

5 GENERATION OF TRAINING DATA

To train our algorithm, we need to generate a large sample of synthetic transmission spectra and their corresponding realizations with different levels of noise. For this purpose, we use the `TauREx 3` framework (Al-Refaie et al. 2021). Given the challenge of creating a proper set of samples with different mixing ratios of fill and biosignature gases, atmospheric temperatures, pressure, and instrumental noise, we developed an independent Python package called `MultiREx`. The package is publicly available[1]. All the scripts created for our numerical experiments and to generate the plots in this paper are available in the `GitHub` public repository of the package[2].

The synthetic transmission spectra were generated with the planetary properties of TRAPPIST-1 e, using the parameters provided by Agol et al. (2021) (see Table 3). The spectrum of TRAPPIST-1 was obtained from the Phoenix model grids (Husser et al. 2013). The atmospheric model implemented consists of 100 layers, with an isothermal temperature profile featuring three possible temperatures: 200 K, 287 K, and 400 K, with a common base pressure of 10^5 Pa ($\approx$ 1 atm) and a top pressure of 10^{-3} Pa.

Atmospheres were created with N_2 as the fill-gas, and a fixed mixing ratio of CO_2 of 10^{-2}. This apparently arbitrary concentration is a compromise between the low levels observed on present Earth and a more primitive atmosphere (see e.g. Lin et al. 2021). It should also be noticed that the higher the level of CO_2, the harder is the detection, using the traditional retrieval algorithms, of some interesting biosignatures such as ozone. However, and independently of the numerical experiments we describe in the following sections, we have ran tests using higher concentrations CO_2 and noticed that, even above this level, the effects of the molecule on our results are negligible.

In addition to the fill-gases, we include, depending on the case, three additional molecules: CH_4, O_3, and H_2O. These molecules were added such that an atmosphere could contain none, one, two, or all three of the molecules simultaneously. The mixing ratio of each molecule was selected among 10 possible values, all of them distributed log-uniformly between 10^{-10} and 10^{-1}.

[1] http://pypi.org/project/multirex
[2] https://github.com/D4san/MultiREx-public

The spectral range for the synthesis of observational data was ~0.69 μm to ~5.3 μm, corresponding to the region covered by the NIRSpec PRISM ($R = 100$), where the noise, calculated with Pandexo (Batalha et al. 2017), is relatively low.

In total, 3993 synthetic spectra were generated: [10×10×10 (all molecules included) + 3×10×10 (three combinations of 2 molecules) + 3×10 (atmospheres with a single species) + 1 (atmosphere with only fill-gases)] × 3 (values of temperature).

In the spectra synthesis, contributions due to Rayleigh scattering and molecular absorption were considered. For the opacities of CO_2, CH_4, and H_2O, we used data from the ExoMol database (Chubb et al. 2021), while the opacities of O_3 and N_2 were taken from Exo-Transmit (Lupu et al. 2014; Freedman et al. 2014, 2008).

All these molecules are present in Earth's current atmosphere and significantly contribute to the terrestrial spectrum. The combination of O_3 and CH_4 is considered a strong biosignature (see section 4). Additionally, both CH_4 and H_2O are also present in prebiotic scenarios (Lin et al. 2021).

O_2 was not included in the model because it does not exhibit a strong signal in the range of interest, and the CIA O_2-O_2 (which would be the major spectral signal of O_2) would not be detectable in the NIRSpec range with a reasonable number of transits (Fauchez et al. 2020).

The presence of clouds was also not included. Since they are located in the lower part of the atmosphere, they should not significantly affect the model's predictions (Lin et al. 2021; Barstow & Irwin 2016). Additionally, according to Doshi et al. (2022), the presence of stratospheric clouds would not affect transmission spectra, at least with the JWST instruments.

To train the algorithms, a constant Gaussian noise was added across the entire wavelength range depending on the SNR value. As a result, the 3993 synthetic spectra became many more (in some cases more than 100 times more). We trained the algorithms using spectra with six discrete SNR values: noise-less (for simplicity in notation, we will refer to this as 0), 1, 3, 6, 10, and 20. Additionally, for the training process, all spectra were normalized based on their maximum and minimum signal. For testing, however, we used a fine-grained interval of SNR.

In Figure 2, we show a representative transmission spectra, containing all interesting chemical species in concentrations similar to those of a Neoproterozoic Earth (Kaltenegger et al. 2020). We can see that many peaks can become obscured by other molecules if the mixing ratio and noise of the confounding molecule are increased.

To represent the strength of the noise in the simulated transmission spectrum in Figure 2, we include a bar over the strongest CO_2 peak, showing the magnitude of the relative error. As expected for an SNR of 3, the noise can be as one third as the peak itself.

6 BIOSIGNATURE RANDOM FOREST CLASSIFICATION

Various classification experiments utilizing RF were conducted to investigate the potential for biosignature detection:

(i) The first and simpler experiment was aimed at identifying interesting targets, namely planets which could actually have biosignatures and thus deserve further investigation. We call this experiment and the resulting model the ***Binary Classificator*** (BC).

(ii) The second experiment is more complex. It aims at classifying the planets by specific biosignatures, namely, determining if a planet could contain, for instance, ozone or not. We call this the ***Multilabel Classificator*** (MC).
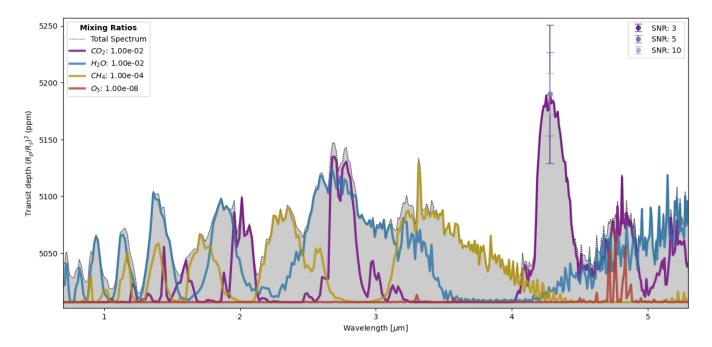
Figure 2. Absorption contributions from different molecules in a potentially inhabited TRAPPIST-1 e spectrum with surface concentrations similar to a Neoproterozoic Earth (Kaltenegger et al. 2020) and a mixing ratio of H_2O fixed at 10^{-2}. Each color represents a molecule (see legend). Error bars illustrate noise based on the SNR, using the CO_2 strongest peak as the signal value.

(iii) The third experiment seeks to train what we call specialist algorithms. These algorithms are trained on planets having a specific biosignature molecule while they are used to classify the entire synthetic spectra dataset. We call this ***Specialised Classificator*** (SC).

For the implementation and performance evaluation of the RFs, the diverse functions offered by `scikit-learn` (Pedregosa et al. 2011) were employed. This approach allowed for a comprehensive analysis of biosignatures, leveraging the robust machine learning tools available within the mentioned library.

When applying RF classification, one crucial aspect is setting the threshold for classifying a planet as positive or negative. Although a conventional threshold of 0.5 is typically used, we adjusted it to increase the Recall due to the significant need to differentiate between FP and FN. This adjustment initially led to an increase in FP, prompting further lowering of the threshold. Ultimately, we optimized the threshold to ensure that the TNR remained above 0.6 for any SNR.

6.1 Binary Classificator

As a first experiment, we generated a large set of noisy synthetic spectra encompassing all possible compositional combinations: only fill gases, only CH_4, only O_3, combinations of CH_4 and O_3, only H_2O, combinations of H_2O and CH_4, and so on. Each planet was labelled as `bio` (True or 1) if it had at least one biosignature, and `non-bio` (False or 0) otherwise.

This general experiment required the most diverse set of training and test spectra of all the experiments we performed. The distribution of planets having different combinations of biosignatures is summarised in Table 4. As we explained before, we generate synthetic spectra with 6 levels of SNR (rows in the table). The binary codes at the head of the table represent the presence (1) or absence (0) of CH_4, O_3, and H_2O, respectively. For instance, the third column, labelled 010, corresponds to planets without CH_4 (0), containing O_3

Table 4. Training data distribution for the RF classificator focused on interesting classification. A SNR of 0 correspond to a noise-less spectra.

SNR	000	100	010	001	110	101	011	111
1	20000	500	300	1000	20	20	20	5
3	20000	500	300	1000	20	20	20	5
6	20000	500	300	1000	20	20	20	5
10	10000	500	300	500	20	20	20	5
20	10000	500	300	500	20	20	20	5
0	10000	500	300	500	20	20	20	5

(1) and in the absence of H_2O (0). The number in each cell indicates how many times a noisy spectrum was generated for that type of atmosphere.

In the example before, for training the algorithm using synthetic spectra with 010 composition and having SNR of 10 (fourth row) we generate 300 noisy spectra per each value of O_3 concentration (10 in total) and per each atmospheric temperature (3 in total). This gave us a total of 9000 spectra for this composition alone. Similarly, for the case of planets having all molecules (111), we generate 5 noisy spectra per each molecule concentration ($10 \times 10 \times 10$), and per temperature value (3), ending up with 15000 spectra for this composition.

A total of 195 000 synthetic spectra was used in this experiment.

It is important to stress that the quantity and disparity of data labelled as bio or non-bio can significantly influence RF performance, especially for low SNR signals. Adjusting the training data proportion towards more negative or positive examples as necessary can enhance the model's ability to recognise the presence or absence of certain molecules under challenging noisy conditions.

Using this data, we trained a RF classificator to determine whether new spectra are `interesting` or `not-interesting`. In this context, an interesting spectrum is defined by the potential presence of CH_4 or
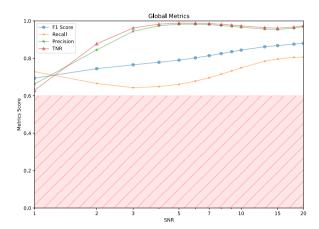
Figure 3. Global metrics of binary classification of `interesting` (with biosignatures) and `non-interesting` planets across different SNR values. The shaded area below 0.6 in the graphs indicates suboptimal classification performance.

Figure 4. Minimum mixing ratio required for detection as interesting (with Recall > 60%) across different SNR values.

O_3, while the presence of H_2O is treated as interference rather than a biosignature. The training was conducted using 80% of the data, and an initial validation was performed with the remaining 20%.

Our RF model contained 400 estimators, used an entropy criterion, and a minimum samples leaf of 3. After training, the classification thresholds were set at 0.4.

The resulting trained RF was tested using a set of 180 000 new synthetic spectra per SNR value, equally distributed between `bio` and `non-bio` categories. The `non-bio` spectra consisted of 45 000 planets with CO_2 and N_2, and 45 000 with H_2O, CO_2 and N_2. The `bio` spectra had 15 000 samples for each possible combination of biosignature molecules (6 combinations) which gives 90 000 total test spectra.

6.1.1 BC metrics

In Figure 3, we show the value of each metric defined in subsection 2.2 as a function of SNR for the trained algorithm. It is important to stress that these results are obtained when classifying spectra of all possible mixing ratios. Therefore, this metric will provide a conservative measurement of the algorithm performance. If we would focus only on the planets with the highest concentrations of the biosignatures gasses, the results will be much better.

As expected, as SNR increases, most metrics improve. Even in the worst-case scenario of SNR=1, all metrics have a value above 0.6, which means that even with the noisiest spectra, it works better than a blind guess.

Interestingly, the algorithm can classify `non-bio` as `non-interesting` planets very efficiently (TNR $\approx$ 1) even for values of SNR as low as 3. This means that the algorithm could be very good at avoiding wasting time on uninteresting planets. This is confirmed by the behaviour of the Precision metric.

An interesting behaviour is noticed in the *Recall* metric, where the algorithm seems to trade-off the improvement/worsening in TNR at the expense of Recall. In other words, for low values of SNR, the algorithm becomes more pessimistic (discarding `bio` planets as not interesting) as SNR increases, while at the same time, it avoids classifying as interesting planets that are not (high TNR).
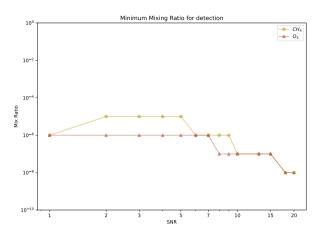
In summary, the analysis of the metrics shows that our BC algorithm performs relatively well at labeling as interesting planets that have biosignatures, but it performs much better for discarding as non interesting planets without biosignatures. This will avoid wasting valuable research time on non interesting targets.

6.1.2 BC minimum mixing ratio for detection

As a secondary analysis, for each value of the SNR, we sorted the classified synthetic spectra according to the CH_4 mixing ratio. Then, we selected the spectra having a specific CH_4 mixing ratio and calculate our binary classification metrics only for planets with this concentration.

As expected, for low values of the mixing ratio, the Recall metric was generally low. We repeated this procedure until the Recall reached a value of 0.6. The corresponding mixing ratio is called the *minimum mixing ratio for detection*. We repeated this analysis for O_3.

In Figure 4, we show the minimum mixing ratio for the detection of each biosignature as a function of SNR values. In the case of CH_4, it is observed that the algorithm, with an SNR of 1, successfully classifies as interesting atmospheres with a mixing ratio as low as 10^{-6}. For comparison, the present Earth has a mixing ratio of the order of 10^{-6}, but in the past it has reached levels as high as 10^{-3} or larger (Kaltenegger et al. 2007, 2020).

The capability of the BC algorithm for detecting low levels of CO_2 is lost when SNR > 2 and only recovers when SNR > 5. This trend is related to the decrease in Recall within this SNR range, as shown in Figure 3.

The case of O_3 is more complex. Only when the SNR reaches values larger than 10-15 (that in the case of TRAPPIST-1 e would require with JWST more than 100 transits) the minimum surface mixing ratio detection for this biosignature is of the order of that on present Earth. However, if a planet has a global biosphere with levels of CH_4 comparable to that of Earth, even if the levels of O_3 are low, our algorithm will tag the planet as interesting. In other experiments we will show that the detection capabilities for ozone can be improved by training algorithms for multilabel classification (see subsection 6.3).
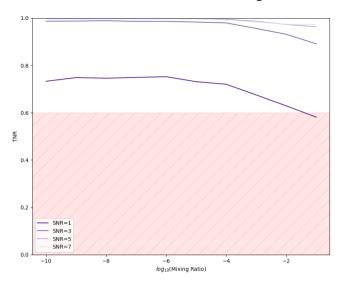
Figure 5. TNR in BC across H_2O mixing ratios evaluated at representative values of SNR. The shaded area below 0.6 in the graphs indicates suboptimal classification performance.

6.1.3 BC interference analysis

We can use similar methods as before to evaluate the performance of the algorithm in classifying `non-bio` planets with H_2O, as interesting cases. This is, to measure the tendency to misclassify planets due to the presence of water. As we observe in Figure 2, the water bands in the spectral range we are considering here overlap most of the ozone and some of the methane bands. As a result, when we have a large water mixing ratio and a low SNR, a `non-bio` planet can be tagged as interesting.

In Figure 5 we show the value of TNR (which is the best-suited metric for this case) as a function of water mixing ratio. As expected, the TNR is reduced, i.e. the algorithm is not tagging as not-interesting planets without biosignatures (TN), when the mixing ratio of water is increased. When the SNR is increased, the performance of the algorithm improves in the sense of becoming less sensitive to this water inteference.

We also analyse the performance of the algorithm with planets having two biosignatures. For that purpose, we picked samples of spectra having pairs of mixing ratios such as (CH_4, O_3) : $(10^{-10}, 10^{-10})$, $(10^{-10}, 10^{-9}), \ldots, (10^{-9}, 10^{-10}), (10^{-9}, 10^{-9}), \ldots$. For each sample, we compute the Recall. In the columns of Figure 6 we show the result as heat map of this metric as a function of mixing ratio combinations. We call this *Recall maps*.

As we observe even at low mixing ratios of one of the biosignatures, the Recall can be high if the mixing ratio of the other one is high enough. For example, at SNR = 2 (first row in Figure 6) and at a very low concentration of ozone, e.g. 10^{-9} in the first panel, the presence of abundant methane, e.g. 10^{-3} will make that the algorithm classify as interesting most planets having both biosignatures. Our interpretation of this result is that the algorithm seems to learn which are the spectral bands of each biosignature independently and it is able to know that the planet is interesting by identifying any of them even if the other is missing.

Additionally, as SNR increases, the algorithm becomes more conservative (lower values of Recall) when both mixing ratios are low

Table 5. Training data distribution for the multilabel classificator (MC). This table has the same structure as Table 4.

SNR	000	100	010	001	110	101	011	111
1	10000	250	250	250	50	50	50	5
3	10000	250	250	250	50	50	50	5
6	5000	250	250	250	50	50	50	5
10	5000	250	250	250	50	50	50	5
20	5000	250	250	250	50	50	50	5
0	10000	250	250	250	50	50	50	5

but achieves better performance (higher values of Recall) at higher mixing ratios.

We repeated the same procedure as before but now using combinations of one of the biosignatures and water. The resulting Recall maps are shown in the second and third columns of Figure 6. Although water tends to interfere in the detection of biosignatures, the fact that the frontier between low and high Recall areas in the Recall maps is almost horizontal, means that what determines the performance of the algorithm is mainly the mixing ratio of the biosignature. In other words, H_2O only affects negatively the Recall in situations where the mixing ratio of the biosignature is really low.

6.2 Multilabel Classificator

In our previous experiment we binary classified the planets as `interesting` and `not-interesting`. We aim now to improve the capacity of the method by introducing multiple labels, intended to identify individual biosignatures and other interesting molecules.

For that purpose, we specifically define three labels: `methane`, `ozone` and `water`, each of them having two possible values: 0 (the molecule is absent) and 1 (the molecule is present).

Focusing on the detection of each molecule individually is valuable because, as discussed in section 4, different biosignatures have varying degrees of robustness and implications. By independently classifying each biosignature, we can better understand the presence and relevance of each molecule, thereby improving our ability to assess the potential habitability of exoplanets.

The MC random forest differs in its hyperparameters from the BC algorithm described in subsection 6.1 by having a lower number of estimators (100) and a defined max depth of 200. The rest of the hyperparameters remain the same.

The training spectra used were of the same types as in the previous experiments, but the distribution varied. In Table 5, we show the number of training spectra according to composition and SNR for the MC dataset. From this dataset, 80% was used for training, and an initial validation was performed with the remaining 20%.

After training, the classification thresholds for CH_4, O_3, and H_2O were set at 0.49, 0.45, and 0.50, respectively.

The model was tested by generating 150 000 noisy spectra per SNR, including 30 000 spectra for the scenario without interesting molecules, and another 30 000 for the cases where all three molecules are present. For the remaining combinations, we generate 15 000 spectra per combination for a total of 90 000. This distribution ensures a balanced test dataset, i.e. for each molecule, there is an equal number of spectra with and without it.

6.2.1 MC multilabel metrics

In Figure 7, we show the algorithm's overall performance as measured with all the multilabel metrics de define in subsection 2.3.

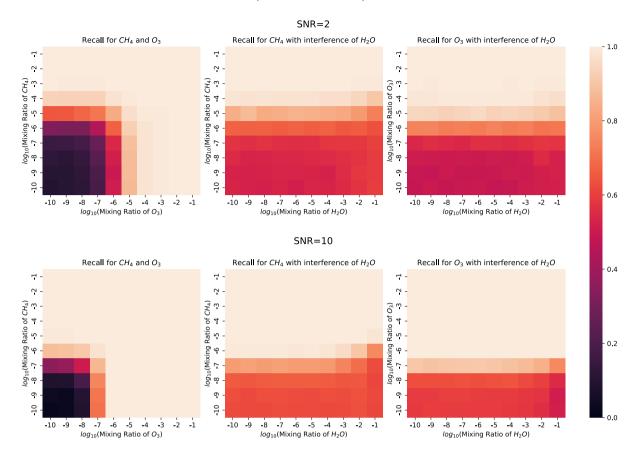The algorithm improves more rapidly in its ability to avoid false

Figure 6. Grid of Recall maps (see text) ssfor BC at two repersentative SNR levels (2 and 10). Recall was calculated for different pairs of molecules across varying mixing ratios

positives as measured by TNR and Precision when the SNR increases. This means that, on average, the model is also better at classifying planets as not having a given molecule when they actually lack it (true negatives). The Perfect Match metric performs poorly, reaching only 50% with SNR greater than 10. Although this result tends to be disappointing, this metric is so stringent that obtaining scores of the order of 0.3 is not as bad as it might seem. Finally, the global Recall seems to reach a maximum value and stops improving significantly when the SNR exceeds 7.

6.2.2 MC molecule metrics

The novel feature of MC with respect to BC is its ability to allow for a per-molecule classification. In Figure 8, we show the binary metrics for the algorithm when the planets are classified by the presence of each biosignature separately. To calculate the metric, we first choose among the training set those that have a specific value of the SNR and a given molecule, e.g. O_3. Then, we use the label, e.g. ozone (that can have two values, 1 or 0 depending on whether the planet has ozone or not, respectively) and the corresponding class assigned by the algorithm, to compute the metrics. Of course, the higher the score the algorithm achieves for a given molecule, the better it is

at identifying that molecule among others potentially present in the spectrum.

In all metrics, the model shows better performance is attained in classifying O_3 compared to CH_4 and H_2O. Both CH_4 and O_3 exhibit a recall above 60% even at low SNR. However, as the SNR increases, the Recall remains almost constant. Even though H_2O is not a biosignature, it is interesting to notice that the performance of the algorithm in terms of Recall and Precision for this molecule is lower compared to the other molecules. Specifically, H_2O generates a trade-off around an SNR of 5, where improving the TNR results in a loss of Recall. This can be interpreted by saying that improving the ability to distinguish planets without water in their atmosphere (i.e. TNR) slightly decreases the ability to detect planets having water.

Comparing all metrics, we notice that the MC algorithm is better at helping us to avoid wasting time with uninteresting candidates, this is because the TNR and Precision metrics have high values (above 80%) for SNR larger than 4.

6.2.3 MC minimum mixing ratio for detection

Using the Recall metric, we repeat the experiment in subsubsection 6.1.2 to determine the minimum mixing ratio at which the MC
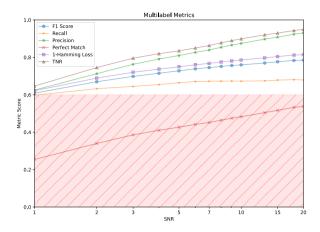
Figure 7. Multilabel classification metrics for the MC as a function of SNR. For the meaning of the shaded region, see Figure 3.

Figure 9. Minimum mixing ratio required for detection of each molecule (with Recall > 60%) with MC as a function of SNR values.
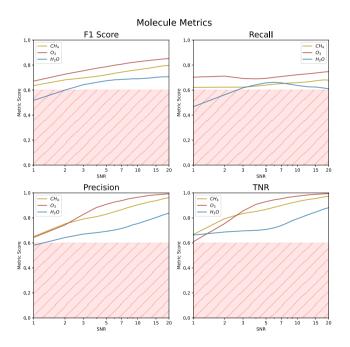

Figure 8. Molecule-specific metrics for the MC. For the meaning of the shaded region, see Figure 3.

algorithm can tag the presence of each molecule. The results are presented in Figure 9.

As expected, the minimum detectable mixing ratio improves with increasing SNR, even for the case of H_2O. However, the minimum mixing ratio for detection of water is larger than for the biosignatures. This is because in the studied spectral range, the absorption strength of water is much lower than that of the other molecules.

Two interesting comparisons between the MC and BC algorithms can be made by examining Figure 4 and Figure 9. While BC is better at detecting low levels of ozone (of the order of 10^{-7}) at relatively high SNR (greater than 15), in general, MC performs better at lower SNR value (larger or equal than 5). However, MC performs worse because of the interference of the water lines.
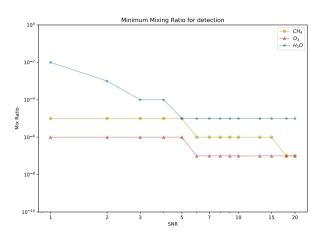
6.2.4 MC interference analysis

Now, we want to evaluate the effect that the presence of one molecule leads to misclassify a planet as if it has another molecule. For instance, at some mixing ratios, the presence of ozone makes the algorithm misclassify a planet as if it has methane although in reality, the latter molecule is not present. This misclassification interference is measured by the TNR metric. We already tested this in subsubsection 6.1.3 for the case of the BC algorithm using water as the interfering molecule.

In Figure 10, we show the result of comparing the TNR measured individually for different molecules (different panels) when the mixing ratio of interference molecules (different curves in each panel) is varied. We show the case for SNR = 3 since, in our numerical experiments, at this value of the SNR the effect of interference was the largest. However, in the numerical repository of the paper (see section 11) we have included animations showing the results for different values of the SNR.

In all cases increasing the mixing ratio of the interference molecule diminished the TNR of each biosignature molecule. In other words, the capability of the algorithm to tag as noninteresting planets that do not have a given molecule is worst when an interference molecule is present. The most significant effect is produced by ozone when classifying a planet by the presence of methane (first panel in Figure 10). At an ozone mixing ratio as low as 10^{-6} the algorithm tends to systematically classify planets without methane as if they had it. However, the TNR for O_3 (second panel in Figure 10) does not deteriorate as much at higher CH_4 concentrations. On the other hand, H_2O classification is the most affected by the presence of both biosignatures, showing confusion at various concentrations.

It is interesting to note that the presence of water creates the least interference among all molecules. Only at very high concentrations of water, interference on biosignatures can be significant. This means that this type of algorithm can tag wet planets as inhabited planets, i.e., as if they have biosignatures.

In summary, the MC algorithm faces challenges at classifying molecules individually due to the interference of other molecules. These effects are especially concerning if the mixing ratio of the interfering molecules are relatively high. Interference between ozone and methane is not as concerning since both are biosignatures, how-

TNR for Each Molecule with Interference of the Others
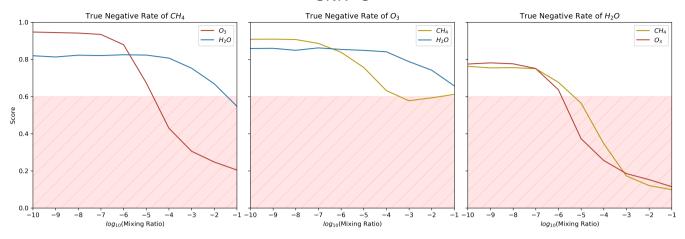### SNR=3

Figure 10. TNR in MC as a function of the mixing ratios for each interfering molecule, at a constant SNR. Each column represents the true negative rate (TNR) for the classification of: CH_4 (left), O_3 (center), and H_2O (right). For the meaning of the shaded region, see Figure 3.

ever, interference with water, which is not considered a biosignature, can be misleading. Still the mixing ratio of water required for these misclassifications are very large.

A last but not least important experiment we performed with the MC algorithm was to analysing how the presence of other molecules affects the Recall in the classification of a molecule of interest. In Figure 11, we show the Recall maps corresponding to the molecules analysed by our algorithm. This figure is analogous to Figure 6.

6.3 Specialised Classificator

To address the multilabel classification problem from another perspective, three RF binary classificators were trained, each of them specialised at detecting a specific molecule: CH_4, H_2O, or O_3. The classification of planets according to the presence of all molecules using these individual specialised algorithms will be referred to as *Specialised Classificator* (SC).

SC is considerably more scalable than previous strategies. On one hand, if we want to add new molecules and/or biosignatures to the classification effort, it only requires training another specialised RF and adding it to the already trained set of algorithms. On the other hand, it demands a lower number and complexity of training spectra as explained below.

Unlike the approach in subsection 6.2, each RF was trained using a dataset containing spectra without biosignatures (fill gases only) and spectra containing fill gases plus the molecule of interest. The numbers of training and test spectra for each molecule depending on the value of SNR are presented in Table 6.

Each RF model consisted of 400 estimators that used the entropy criterion, had a maximum depth of 200, and a minimum samples leaf of 3. The classification thresholds for CH_4, O_3, and H_2O were set at 0.40, 0.33, and 0.36, respectively.

To analyse the performance of the SC algorithm at classifying realistic spectra, namely those containing a complex mixture of molecular species (including interfering ones), we generated a spectra dataset using the same procedure described in subsection 6.2. Similarly, the same analyses and plots were generated for the SC as in the case of the

MC. This ensures consistency in the assessment of the comparative performance of the models.

6.3.1 SC multilabel metrics

In Figure 12, we show the result of calculating the multilabel classification metrics for the SC. If we compare these results with those corresponding to the MC (see Figure 7) we notice that, despite the differences between the approaches in terms of complexity of the training set and scalability, the performance is almost the same. This implies that training a set of specialised classificators, with its advantages of adding new molecules as required, produces almost the same results, at least for the test spectra we are generating here, then a more complex and CPU intensive algorithm that includes all molecules at once (the MC algorithm).

Still, in a more general case with a realistic spectrum, the MC approach will show its advantages over the SC as we will see in section 8.

6.3.2 SC molecule metrics

In Figure 13, we perform a similar comparison but for the metrics evaluated for each molecule separately. The analogous results for the MC were presented in Figure 8. Again the performance of both algorithms is comparable, at least for the biosignature molecules. This implies, again, that when classifying molecules individually, the more scalable SC algorithm could be as good as the MC one.

Interestingly, and as an example of the "hidden" behaviour of the algorithm, the TNR metric for CH_4 and H_2O (lower-right panels) tends to stagnate or worsen slightly after a certain SNR value. This behaviour indicates that as the signal becomes less noisy, the algorithm tends to confuse these molecules with others. This is likely due to those other molecules sharing many spectral peaks with the molecules of interest, leading to false positives. However, O_3 does not exhibit this behaviour, allowing it to maintain a consistently improving precision across all SNR values.
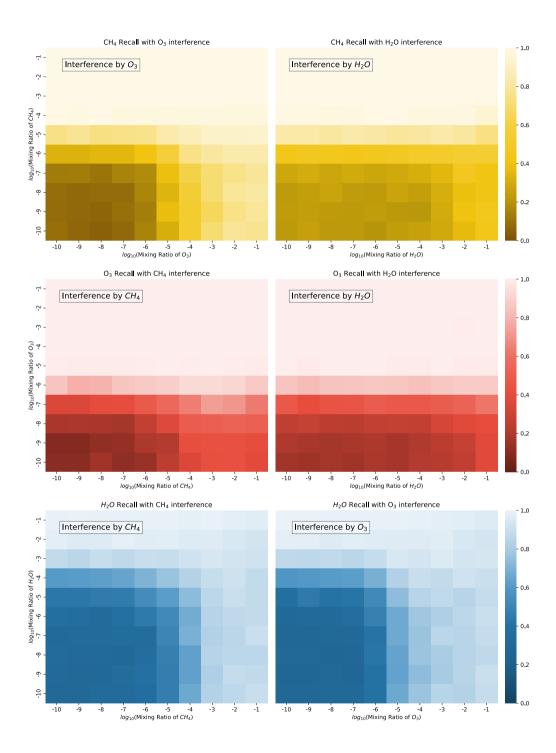
Figure 11. Grid of Recall maps for MC at SNR=3. Recall was calculated for different pairs of molecules across varying mixing ratios. Each row represents the classification of a specific molecule (CH$_4$, O$_3$, and H$_2$O from top to bottom), with interference from other molecules as indicated in the labels.

Table 6. Training dataset distribution for each Random Forest classificator specialised at detecting a specific biosignature (first column). Each column pair (labelled as ¬ mol when the molecule is absent, and mol otherwise) contains the number of spectra used at different SNRs. The last column summarise the total number of spectra used per molecule. This number is evenly distributed across the three possible temperatures.

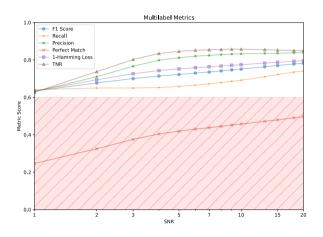| Molecule\ SNR		1		3		6		10		20		0		Total
	Label	¬ mol	mol	¬ mol	mol	¬ mol	mol	¬ mol	mol	¬ mol	mol	¬ mol	mol	
CH_4		10000	5000	10000	10000	10000	5000	10000	5000	10000	5000	10000	10000	300000
O_3		5000	2500	5000	2500	5000	5000	5000	5000	5000	5000	10000	10000	195000
H_2O		10000	5000	10000	5000	10000	5000	10000	5000	10000	5000	10000	10000	285000

Figure 12. Multilabel metrics for the SC algorithm as a function of SNR. For the meaning of the shaded region, see Figure 3 .

Figure 14. Minimum mixing ratio required for detection of each molecule (with Recall > 60%) with SC across different SNR values.
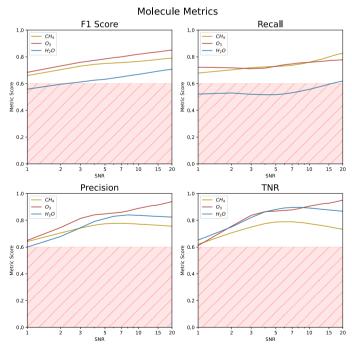

Figure 13. Molecule-specific metrics from SC. Each subplot show the performance of different metrics (F1 Score, Recall, Precision, TNR) for each molecule (CH_4, O_3, H_2O) as the SNR increases. For the meaning of the shaded region, see Figure 3.

6.3.3 SC minimum mixing ratio for detection

In Figure 14, we show the minimum mixing ratio for detection achieved by the SC algorithm for each molecule. The analogous figure for the case of the MC algorithm is Figure 9. As shown, the SC algorithm achieves lower minimum mixing ratios of CH_4 and O_3, as compared with MC at lower SNRs. For instance, in the case of methane, the SC is capable of detecting the molecule at an SNR as low as 2, while the MC requires for the same mixing ratio a minimum SNR of 6. The case of ozone is also interesting, since incredibly low values of the mixing ratio of the molecule, namely $\sim 10^{-8}$, are identifiable at SNR$\gtrsim$18 with the SC algorithm.

6.3.4 SC interference analysis

In Figure 15, we perform an analysis of the effect that interfering molecules have on the detection of each molecule for the SC algorithm. This analysis corresponds specifically to the case when the molecule of interest is absent and an interfering one can confuse the algorithm. A similar analysis was performed for the MC algorithm and presented in Figure 10. For details on the design of both figures, please refer to the explanation in subsubsection 6.2.4.

Compared to the MC case, the SC algorithm tends to be more conservative in correctly classifying the absence of all target molecules. In other words, SC tends to be less confused, especially for the case of O_3 classification (middle panel in Figure 15). The same can be said for H_2O classification (right-most panel), which, contrary to what is seen with the MC algorithm, is not confused by the presence of O_3.

It is noteworthy that, unlike the MC case, relatively high SNR levels do not reduce the confusion of these molecules at high mixing ratios. This suggests that the SC takes advantage of the noise to

Figure 15. Same as Figure 10, but for the case of SC algorithms. For an animated version click here.

avoid interference. However, this restricts the generalisation of the algorithm to other noise models, since the algorithm tends to depend on the noise structure and not on the signal itself for this particular task.

Finally, in Figure 16, we analyze the effect of interference on the Recall metric, specifically for the positive detection of each molecule. The corresponding figure for the MC algorithm is presented in Figure 11.

We notice that, in general, at low SNR, the behaviour of both algorithms is similar, at least for the biosignature molecules, ozone and methane. When the mixing ratio of these molecules is high, no interference from the other molecules is observed (upper and middle rows). At lower values of the mixing ratios, only in the case of methane, the presence of ozone can be helpful in positively identifying the molecule (upper row, left-most panel). This behaviour was also observed when using the MC algorithm.

However, for the classification of H_2O using the SC algorithm (lower row), the presence of other molecules negatively affects the Recall. This is contrary to what happens when using the MC algorithm.

7 THE SNR VALUE AND THE NUMBER OF TRANSITS

Throughout the paper, we have used SNR to quantify the precision of the transmission spectra when training and testing the algorithms. In a real-life scenario, for example, when performing observations using JWST, the level of noise of an observed spectrum will be determined by the number of transits that can be measured during an observing program or a survey campaign. In order to compare our results to those obtain in other works, we should estimate a relationship between the number of transits and the SNR of the resulting transmission spectra, at least for the scientific case we are studying here.

In Figure 17 we show the correspondence between the number of transits of TRAPPIST-1 e, hypothetically observed with the JWST NIRSpec PRISM, and the resulting SNR as computed with `Pandexo` (Batalha et al. 2017). For making this plot, we first calculate, using `TauREX`, a theoretical spectrum for a TRAPPIST-1 e twin having a N_2

atmosphere with CO_2 at a mixing ratio of 10^{-2}. Then, we synthesise the observed transmission spectrum of the planet using the observing program and instrument parameters chosen by Lin et al. (2021). We repeat the calculation assuming a number of transits between 1 and 100.

Since the noise of this instrument depends on wavelength, defining a global SNR is not trivial. For our analyisis, we are interested on the peaks of the spectrum which provides information about the presence of molecules. It is more reasonable to use the SNR of the signal at some of those peaks than at other parts of the spectrum (for instance at the Rayleigh downhill). For this reason we use as a characteristic SNR, the value corresponding to the peak of CO_2 at 4.28 μm. This peak is close enough to the maximum detectable wavelength of the instrument where the noise is also larger. We understand that, given the complexities of the instrumentation, the actual SNR value may change when different conditions are assumed.

It is not hard to verify that the SNR value in Figure 17 grows as $\sqrt{N}$, where N is the number of transits. This behaviour, although not self-evident in the figure since we have used a semi-log scale, it is the expected one since an increase in the number transits contribute to also increase the number of data points we add at a given channel in the measured flux. The main consequence of this behavior is that for increasing by a factor of 2 the value of the SNR we need almost 4 times more transits. This is one of the most important instrumental challenges we face when searching for biosignatures.

Lin et al. (2021) have calculated that for detecting methane at 2.5σ using standard retrieval procedures and in a best-case scenario (mixing ratio larger than 10^{-5}), a minimum of 10 transits (SNR>3) are required. However, for achieving a similar precision in the retrieval of ozone, as many as 200 transits will be required. This corresponds, according to our estimation, to a SNR>24.

If we are able develop, as proposed in this paper, a methodology able to select interesting targets having transmission spectra with SNR<3 for methane and SNR<24 for ozone, a significant improvement in our capabilities for searching for biosignatures could be achieved. This is precisely the ranges of SNR values where we have performed our numerical experiments.

To illustrate the capabilities of our method, we have included in Figure 17 the SNR value required to confidently classify planets

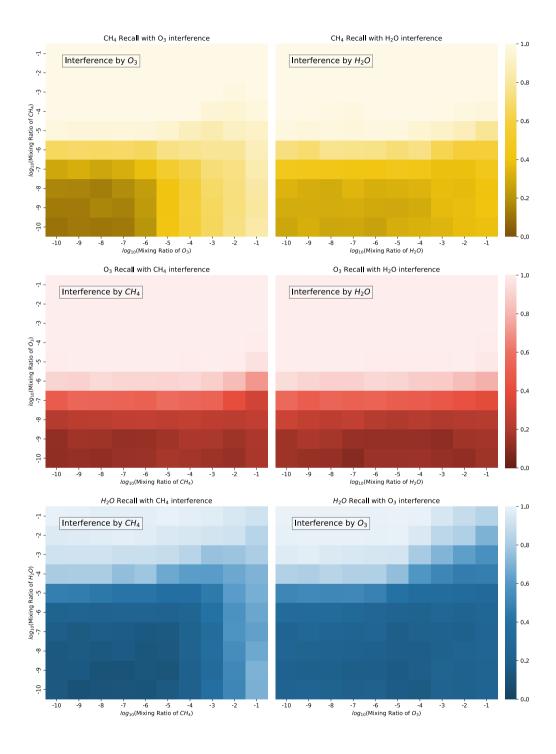# Recall for Each Molecule with Interference of Other Molecules
SNR=3

Figure 16. Grid of Recall for SC at SNR=3. Same as Figure 11. For an animate version of this figure click here.
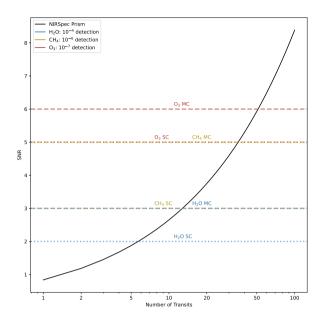
Figure 17. SNR characteristic values (as computed at 4.28 μm) of the transmission spectra of a TRAPPIST-1 e twin with a N_2-CO_2 atmosphere as a function of the number of transits observed with the JWST NIRSpec PRISM. Horizontal lines indicate the detection thresholds for H_2O (blue), CH_4 (yellow), and O_3 (red). Dashed lines represent the MC, while dotted lines represent the SC. The graph demonstrates that achieving an SNR above 10 would require an excessive number of transits, highlighting the challenges in detecting trace gases in exoplanetary atmospheres.

at different biosignature mixing ratios, using the MC and SC algorithms. We see that for classifying a planet as interesting when containing levels as low as 10^{-7} of ozone, the MC algorithm requires almost 50 transits, while the SC needs only about 30 transits. Although a direct comparison is hard, and 30 transits seems to be a lot of them, it is interesting to notice that this number of transits is almost 4-6 times fewer than those required for standard retrieval procedures. On the other hand, and in the case of methane, the number of transits required for a reliable classification using our methods is similar to those achieved with some standard retrieval procedures. However, the mixing ratio reachable by our algorithms is typically 10 times smaller.

In summary, values of the SNR between 3 and 5, where our algorithms reach acceptable performances, correspond to a number of transits between 10 and 30 which are achievable for a JWST regular observing program. With a larger number of transits ($30 - 50$), our methods can classify planets as interesting at lower mixing ratios of methane than those required for regular retrieval procedures using the same number of transits. The case of ozone is more interesting, the MC and SC algorithms may tag a planet as interesting using as much as 6 times fewer transits than those required for normal retrieval procedures.

8 CLASSIFYING REALISTIC SPECTRA

In order to test our algorithms with realistic cases, we use the high-resolution spectra of the Earth at different geologic eras calculated by Kaltenegger et al. (2020). For those spectra we synthesise observations for the JWST NIRSpec using `Pandexo`. The synthesis followed the same procedure we apply to calculate the relationship between SNR value and number of transits in section 7.

It is important to stress that the realistic spectra we use as inputs for these experiments, were originally calculated for a planet with a radius and surface gravity identical to that of the Earth. For using those spectra to study the slightly larger TRAPPIST-1 e planet, we simply rescale the line depth (expressed in kilometres) to the radius of our planet. Of course, the photochemical equilibrium composition of the Earth's atmosphere and TRAPPIST-1 e, that orbits a M-dwarf star and will have also a different gravity and surface pressure, will be different. Still, since we we are just testing the performance of our algorithm when dealing with an actual planetary spectrum, the assumption of similar atmospheres are not violating any physical law.

For a given number of transits (a given value of characteristic SNR), we synthesise 1000 observations and use the algorithms we trained in previous sections, namely the binary classificator (BC, subsection 6.1), the multilabel classificator (MC, subsection 6.2) and the specialised classificators (SC, subsection 6.3), to classify the planet as interesting or not interesting, or as having or not having specific molecules, respectively.

In Figure 18, we show the value of the Recall metric, as obtained for the set of synthetic spectra of a TRAPPIST-1 e planet having a modern and a Proterozoic Earth's atmosphere.

We observe that when dealing with more realistic spectra, namely an atmosphere with a more complex physical structure and a greater number of molecules, our binary classifier (BC) still performs well even if the number of transits is as low as the minimum of 1 transit. This promising result suggests that if we design a survey for obtaining low SNR transmission spectra, and use machine-assisted algorithms aimed at classifying the planets instead of attempting full retrieval as those proposed and exemplified in this paper, we could efficiently select planets that deserves follow-up observations. Interestingly, our results shows that even an algorithm that was trained only with synthetic and simple spectra, can correctly classify more than 60% of the planets with biosignatures in a survey sample. We speculate, that with a more complex training set and probably more competent methods, the performance of machine learning models using our methodology could be significantly enhanced.

Interestingly, the multilabel classificators (MC and SC) performs worse when dealing with realistic spectra, than the binary classificator (BC). For example, the MC classification of CH_4, achieve acceptable performances (Recall > 0.6) only at low number of transits, while the SC algorithm never reaches Recall values above the threshold, neither for the modern nor for the past Earth's atmosphere. This result does not reveal a limitation of our general strategy but the fact that, in the case of multilabel classification, not training the algorithms for dealing with a realistic, instrument-specific noise, penalize the more these strategies. We may have identified the reason of this at least in the case of the SC algorithm. As we pointed out in subsubsection 6.3.4 the specialized algorithms performance may depend on the noise structure. Therefore, to be more efficient in an actual survey, SC algorithms should be properly trained, taking into account the noise of the instrument.

Finally, it is interesting to notice that, counter intuitively, the classification of O_3 becomes harder as the number of transits increases. This fact suggests that the low levels of noise achieved in all spectral
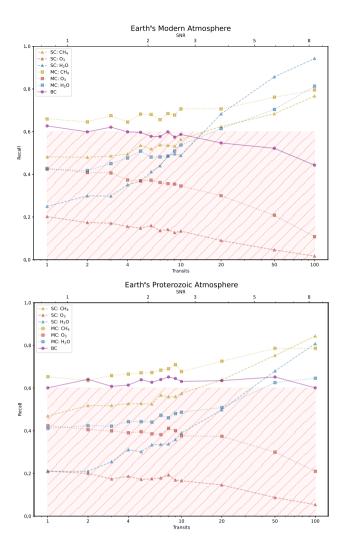
Figure 18. Recall as a function of the number of transits calculated for a set of 1000 synthetic transmission spectra of a TRAPPIST-1 e twin, generated starting with the high-resolution spectra of the Earth's modern atmosphere (upper panel) and Earth's Proterozoic atmosphere (lower panel). Dotted lines with squares and dashed lines with triangles represent the performance of MC and SC classificators for CH_4, O_3, and H_2O. The BC algorithm is shown as a continuous (purple) line.

bands achieved after many transits, may strongly interfere with the classification of this molecule. Conversely, the classification performance for H_2O and CH_4 improves significantly with the number of transits, demonstrating the capabilities of the algorithms to handle transmission spectra with many lines.

9 DISCUSSION

The validity of the results presented in this work may depend on the specific properties of the planet, its atmosphere, and the star we selected for our numerical experiments. In the following we will analyse the sensitivity of our results to those variable factors.

Changing the radius of the star or the radius of the planet will only imply a change in the signal strength and noise and hence a different value the SNR. Pressure, which essentially determines the height scale of the atmosphere, will have an analogous effect. Since our experiments explore a large range of SNR, any modification of

those parameters will fall in one of the synthetic noisy spectra we use for training and testing. In other words, a variation in SNR is equivalent to a variation in either the ratio of the radius of the planet to the radius of the star or the atmospheric pressure. Of course, in real atmospheres, all these factors may affect the equilibrium composition of the atmosphere. However, the model atmospheres we use in this work, that in some sense is a demonstrative effort, are simple enough in such a way that chemistry will not be related to these factors.

Changing the spectral type of the star or the temperature of the planetary atmosphere would not only change the SNR but also the transmission spectra in non-trivial way. This changes may actually impact the performance of the algorithms.

To test this, we performed additional numerical experiments using, first, different stellar effective temperatures in the range of M and K-dwarf stars. We observe that changing the stellar spectrum does not significantly affect the classification results.

In the second place, the effect of atmospheric temperature was implicit to our experiments. Actually, and as described, for instance, in subsection 6.1, our training and test sets already include transmission spectra with different atmospheric temperatures (in the range of 200 K to 400 K).

We observe that at lower atmospheric temperatures the performance of the algorithms improves. For instance, the minimum mixing ratio for detection, at the lowest atmospheric temperatures (200 K), may be reduced by one order of magnitude at very low values of the SNR, for several of the interesting molecules. The reason for this temperature dependence is the change in the molecular cross-section that affects the capability of the algorithm to recognize the presence of the molecules.

In summary, our models are robust against variations in stellar temperature and stellar radius, at least in the range of spectral types considered in this work (K and M dwarfs). However, our experiments reveal that planetary atmospheric temperature remains a factor that may influence the accuracy of biosignature classification.

Since the performance of the SC algorithms depends on the structure of the noise, one question that can be raised is why we do not train our algorithms using the specific noise of NIRSpec. Here, we should recall that our work was aimed to assess the feasibility of using machine learning to classify planets with biosignatures with low SNR, but not with a specific noise structure. If we have chosen a certain type of noise, the validity of our results could become noise-dependent and the generality of our conclusions will be lost. Still, the issue of studying the impact that the noise structure has on the performance of classification algorithms should be further investigated.

10 SUMMARY AND CONCLUSIONS

In this work, we explored the use of supervised machine learning algorithms to classify low SNR exoplanet transmission spectra, based on their potential for containing biosignatures. We employed Random Forest (RF) for designing and training three types of models: a binary classificator (BC) which labels a planet according to whether it is interesting or not for follow-up observations; a multilabel classificator (MC) which labels an observed spectrum according to whether it may have certain molecular species; and specialised classificators (SC) that work similar to MC by assembling multiple binary RF classificators focused on particular molecules.

For training and testing the algorithms, we generated of the order of 10^6 synthetic noisy spectra, corresponding to a planet with the same physical characteristics as TRAPPIST-1 e, but having a wide

diversity of compositions and at three atmospheric temperatures. For that purpose, we designed a package, `MultiREx`[3], that works as a wrapping for `TauREx` (Al-Refaie et al. 2021) but has the capabilities of generating, representing and manipulating large sets of synthetic noisy spectra, especially intended for machine learning purposes.

We tested our models by measuring different binary and multilabel standard metrics (Recall, TNR, F1 Score, etc.) that we interpret, in the context of biosignature searching, as time-saving metrics (minimising the time spent on following-up not interesting planets), wasting metrics (not excluding inhabited planets as non-interesting candidates), discovery metrics (maximising the opportunity to achieve an actual discovery), among others.

The BC algorithm, although lacking the accuracy for detailed molecular classification, achieves the highest values of the discovery metric. With observed spectra having values of the SNR as low as 1 or 2, which can be achieved, according to our estimations, by observing 1-5 transits with the JWST NIRSpec PRISM, the BC model labels as interesting most of the inhabited planets in a survey sample, even if they have a huge range of biosignatures mixing-ratios. Moreover, the algorithm performs incredibly well (TNR> 80%) at tagging as not interesting, planets without biosignatures, avoiding wasting time on follow-up observations.

Our results show that if we perform a survey of potentially habitable exoplanets and measure the transmission spectra using only 1 to 5 transits, we can identify most of the inhabited exoplanets in the survey for performing follow-up, time-intensive observations, only if their atmospheres have surface mixing ratios of CH_4 similar to those of present and prebiotic Earth ($O(10^{-6})$, Kaltenegger et al. 2007, 2020). This number of transits is significantly lower than those required for the detection of the same molecule using standard and non-standard retrieval procedures.

It is important to stress that the methodology proposed here is not intended for detecting biosignatures or for replacing them. In this sense, all the efforts intended at performing full retrieval are still required. Identifying a planet as interesting will only make the allocation of observing time of valuable resources such as JWST more efficient, which is an important goal in modern astronomy.

Classifying planets based on the presence of ozone will be harder than in the case of methane. With mixing ratios of the order of 10^{-7}, which are slightly lower than those of present Earth, efficient classification with our algorithms is possible for values of the SNR of 5 (SC) and 6 (MC), which corresponding to 30-50 transits. Although this number of transits seems to be large, it is important to recall that standard retrieval procedures for ozone at mixing ratios as high as 10^{-6} require more than 100 transits Lin et al. (2021). On the other hand, for planets orbiting cool M-dwarf stars, like TRAPPIST-1 e, higher mixing ratios of O_3 with lower levels of O_2 could be common if the star's UV emission promotes its production (Grenfell et al. 2014).

For future research, it would be valuable to explore the integration of other machine learning methods and increase the diversity of the synthetic spectra used for training. Additionally, including other potential biosignatures and atmospheric conditions could improve the capacity to generalise the models. These advances will contribute to more accurate and efficient biosignature detection, supporting ongoing efforts in the search for life beyond our Solar System.

[3] http://pypi.org/project/multirex

11 DATA AVAILABILITY AND REPRODUCIBILITY

All data required to replicate the results and generate the figures presented in this work, including the `Jupyter` notebooks used to train and test the models, are available in the `GitHub` public repository of the package `MultiREx` at https://github.com/D4san/MultiREx-public.

REFERENCES

Agol E., et al., 2021, The Planetary Science Journal, 2, 1
Airapetian V. S., et al., 2020, International Journal of Astrobiology, 19, 136
Al-Refaie A. F., Changeat Q., Waldmann I. P., Tinetti G., 2021, The Astrophysical Journal, 917, 37
Ardévol Martínez F., Min M., Kamp I., Palmer P. I., 2022, Astronomy & Astrophysics, 662, A108
Ardévol Martínez F., Min M., Huppenkothen D., Kamp I., Palmer P. I., 2024, Astronomy & Astrophysics, 681, L14
Barstow J. K., Irwin P. G. J., 2016, Monthly Notices of the Royal Astronomical Society: Letters, 461, L92
Batalha N. E., et al., 2017, Publications of the Astronomical Society of the Pacific, 129, 064501
Benneke B., et al., 2024, JWST Reveals CH$_4$, CO$_2$, and H$_2$O in a Metal-rich Miscible Atmosphere on a Two-Earth-Radius Exoplanet (arXiv:2403.03325)
Bourrier V., et al., 2017, The Astronomical Journal, 154, 121
Breiman L., 2001, Machine Learning, 45, 5
Burgasser A. J., Mamajek E. E., 2017, The Astrophysical Journal, 845, 110
Cadieux C., et al., 2024, Transmission Spectroscopy of the Habitable Zone Exoplanet LHS 1140 b with JWST/NIRISS (arXiv:2406.15136)
Chubb K. L., et al., 2021, Astronomy & Astrophysics, 646, A21
Cobb A. D., et al., 2019, The Astronomical Journal, 158, 33
Des Marais D. J., et al., 2008, Astrobiology, 8, 715
Doshi D., Cowan N. B., Huang Y., 2022, Monthly Notices of the Royal Astronomical Society, 515, 1982
Fauchez T. J., et al., 2020, Nature Astronomy, 4, 372
Forestano R. T., Matchev K. T., Matcheva K., Unlu E. B., 2023, The Astrophysical Journal, 958, 106
France K., et al., 2020, The Astronomical Journal, 160, 237
Freedman R. S., Marley M. S., Lodders K., 2008, The Astrophysical Journal Supplement Series, 174, 504
Freedman R. S., Lustig-Yaeger J., Fortney J. J., Lupu R. E., Marley M. S., Lodders K., 2014, The Astrophysical Journal Supplement Series, 214, 25
Gebhard T. D., Angerhausen D., Konrad B. S., Alei E., Quanz S. P., Schölkopf B., 2024, Astronomy & Astrophysics, 681, A3
Géron A., 2023, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, third edition edn. O'Reilly, Beijing Boston Farnham Sebastopol Tokyo
Gillon M., 2024, TRAPPIST-1 and Its Compact System of Temperate Rocky Planets (arxiv:2401.11815)
Gillon M., et al., 2016, Nature, 533, 221
Gillon M., et al., 2020, Bulletin of the AAS, 52
Grenfell J. L., 2018, in Deeg H. J., Belmonte J. A., eds, , Handbook of Exoplanets. Springer International Publishing, Cham, pp 1–14, doi:10.1007/978-3-319-30648-3_68-1
Grenfell J., Gebauer S., V. Paris P., Godolt M., Rauer H., 2014, Planetary and Space Science, 98, 66
Guzmán-Marmolejo A., Segura A., Escobar-Briones E., 2013, Astrobiology, 13, 550
Guzmán-Mesa A., et al., 2020, The Astronomical Journal, 160, 15
Hayes J. J. C., et al., 2020, Monthly Notices of the Royal Astronomical Society, 494, 4492
Himes M. D., et al., 2022, The Planetary Science Journal, 3, 91
Huang S., Ormel C. W., 2022, Monthly Notices of the Royal Astronomical Society, 511, 3814

Husser T.-O., Wende-von Berg S., Dreizler S., Homeier D., Reiners A., Barman T., Hauschildt P. H., 2013, Astronomy & Astrophysics, 553, A6

Kaltenegger L., Traub W. A., Jucks K. W., 2007, The Astrophysical Journal, 658, 598

Kaltenegger L., Lin Z., Madden J., 2020, The Astrophysical Journal, 892, L17

Kelleher J. D., Mac Namee B., D'Arcy A., 2015, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. The MIT Press, Cambridge, Massachusetts

Léger A., Fontecave M., Labeyrie A., Samuel B., Demangeon O., Valencia D., 2011, Astrobiology, 11, 335

Lin Z., Kaltenegger L., 2022, Monthly Notices of the Royal Astronomical Society, 516, 3167

Lin Z., MacDonald R. J., Kaltenegger L., Wilson D. J., 2021, Monthly Notices of the Royal Astronomical Society, 505, 3562

Lueber A., Kitzmann D., Fisher C. E., Bowler B. P., Burgasser A. J., Marley M., Heng K., 2023, The Astrophysical Journal, 954, 22

Luger R., Barnes R., 2015, Astrobiology, 15, 119

Lupu R. E., et al., 2014, The Astrophysical Journal, 784, 27

Lustig-Yaeger J., Meadows V. S., Crisp D., Line M. R., Robinson T. D., 2023, The Planetary Science Journal, 4, 170

Madhusudhan N., Sarkar S., Constantinou S., Holmberg M., Piette A. A. A., Moses J. I., 2023, The Astrophysical Journal Letters, 956, L13

Marquez-Neila P., Fisher C., Sznitman R., Heng K., 2018, Supervised Machine Learning for Analysing Spectra of Exoplanetary Atmospheres (arxiv:1806.03944)

Matchev K. T., Matcheva K., Roman A., 2022a, The Planetary Science Journal, 3, 205

Matchev K. T., Matcheva K., Roman A., 2022b, The Astrophysical Journal, 930, 33

Munsaket P., Awiphan S., Chainakun P., Kerins E., 2021, Journal of Physics: Conference Series, 2145, 012010

Nixon M. C., Madhusudhan N., 2020, Monthly Notices of the Royal Astronomical Society, 496, 269

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Schwieterman E. W., 2016, Thesis

Schwieterman E. W., et al., 2018, Astrobiology, 18, 663

Soboczenski F., et al., 2018, Bayesian Deep Learning for Exoplanet Atmospheric Retrieval (arxiv:1811.03390)

Sorower M. S., 2010

Turbet M., et al., 2018, Astronomy & Astrophysics, 612, A86

Turbet M., Bolmont E., Bourrier V., Demory B.-O., Leconte J., Owen J., Wolf E. T., 2020, Space Science Reviews, 216, 100

Vasist M., Rozet F., Absil O., Mollière P., Nasedkin E., Louppe G., 2023, Astronomy & Astrophysics, 672, A147

Waldmann I. P., 2016, The Astrophysical Journal, 820, 107

Wheatley P. J., Louden T., Bourrier V., Ehrenreich D., Gillon M., 2017, Monthly Notices of the Royal Astronomical Society: Letters, 465, L74

Wordsworth R., Pierrehumbert R., 2014, The Astrophysical Journal Letters

Wunderlich F., et al., 2019, Astronomy & Astrophysics, 624, A49

Yip K. H., Changeat Q., Nikolaou N., Morvan M., Edwards B., Waldmann I. P., Tinetti G., 2021, The Astronomical Journal, 162, 195

Zingales T., Waldmann I. P., 2018, The Astronomical Journal, 156, 268

This paper has been typeset from a TeX/LaTeX file prepared by the author.