The Astrophysical Journal, 945:122 (24pp), 2023 March 10 https://doi.org/10.3847/1538-4357/acb25c

Revealing the Galaxy–Halo Connection through Machine Learning

Ryan Hausen[1,2], Brant E. Robertson[3], Hanjue Zhu[4], Nickolay Y. Gnedin[4,5,6], Piero Madau[3], Evan E. Schneider[7], Bruno Villasenor[3], and Nicole E. Drakos[3]
[1] Department of Physics and Astronomy, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218 USA; rhausen@ucsc.edu
[2] Department of Computer Science and Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 USA
[3] Department of Astronomy and Astrophysics, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 USA; brant@ucsc.edu
[4] Department of Astronomy and Astrophysics, University of Chicago, 5640 S. Ellis Avenue, Chicago, IL 60637 USA
[5] Fermi National Accelerator Laboratory, Batavia, IL 60510, USA
[6] Kavli Institute for Cosmological Physics, The University of Chicago, USA
[7] Department of Physics and Astronomy, University of Pittsburgh, 100 Allen Hall 3941 O'Hara Street, Pittsburgh, PA 15260 USA
Received 2022 April 21; revised 2022 December 31; accepted 2023 January 10; published 2023 March 14

Abstract

Understanding the connections between galaxy stellar mass, star formation rate, and dark matter halo mass represents a key goal of the theory of galaxy formation. Cosmological simulations that include hydrodynamics, physical treatments of star formation, feedback from supernovae, and the radiative transfer of ionizing photons can capture the processes relevant for establishing these connections. The complexity of these physics can prove difficult to disentangle and obfuscate how mass-dependent trends in the galaxy population originate. Here, we train a machine-learning method called Explainable Boosting Machines (EBMs) to infer how the stellar mass and star formation rate of nearly 6 million galaxies simulated by the Cosmic Reionization on Computers project depend on the physical properties of halo mass, the peak circular velocity of the galaxy during its formation history $v_{\rm peak}$, cosmic environment, and redshift. The resulting EBM models reveal the relative importance of these properties in setting galaxy stellar mass and star formation rate, with $v_{\rm peak}$ providing the most dominant contribution. Environmental properties provide substantial improvements for modeling the stellar mass and star formation rate in only $\lesssim$10% of the simulated galaxies. We also provide alternative formulations of EBM models that enable low-resolution simulations, which cannot track the interior structure of dark matter halos, to predict the stellar mass and star formation rate of galaxies computed by high-resolution simulations with detailed baryonic physics.

Unified Astronomy Thesaurus concepts: Galaxy formation (595); Galaxy dark matter halos (1880); Large-scale structure of the universe (902); N-body simulations (1083)

1. Introduction

Numerical simulation enables theoretical models of galaxy formation to include detailed physical models for baryonic processes. Simulations can capture the physics of cooling, supernova feedback, radiative feedback, and ionization, and the role of dynamics simultaneously while tracking the growth of cosmological structure formation (e.g., Schaye et al. 2015; Pillepich et al. 2018; Davé et al. 2019). The simulated galaxy populations that result from these models reproduce observed stellar mass sequences such as the main sequence of star-forming galaxies (Brinchmann et al. 2004; Noeske et al. 2007) or the red sequence of quiescent galaxies (Faber et al. 2007). The quest for realism in modeling these observed trends has also added substantial complexity, such that understanding which physical properties of a galaxy most influence its stellar mass and star formation rate (SFR) can prove challenging. Many theoretical frameworks to describe these relations have been developed (e.g., Wechsler & Tinker 2018), including halo occupation distribution models (e.g., Jing et al. 1998), subhalo abundance matching (Vale & Ostriker 2004; Conroy et al. 2006), and semi-analytic models (for a review, see Somerville & Davé 2015). The complex physics encoded by these models and simulations can be difficult to interpret, and the relative contribution of baryonic feedback, dark matter halo formation, and environment in setting galaxy properties remains challenging to disentangle.

This complexity extends to cosmological models of galaxy formation in the reionization epoch. To capture the distribution of sizes of ionized regions with converged simulations (Iliev et al. 2014) and the largest observed features, such as dark gaps (Zhu et al. 2021), the volume of reionization simulations should extend to a least several hundred megaparsecs. Modeling such large volumes in a single simulation while maintaining the spatial resolution needed to include the complex physics of the current state-of-the-art projects, such as Cosmic Reionization on Computers (CROC; Gnedin 2014), THESAN (Kannan et al. 2022), and Cosmic Dawn (Ocvirk et al. 2016, 2020), remains computationally infeasible. Instead, we desire an intermediate approach where large volumes are simulated and the physics of galaxy formation are implemented with an approximate model that recovers the mean trends for galaxy baryonic properties predicted by more detailed calculations. With this goal in mind, a model for reionization sources that encapsulates the results of projects like CROC in a simple module is the first necessary step for deploying lower-resolution simulations with much larger ($L \sim 500$ cMpc) simulation volumes. If the stellar mass and SFRs of ionizing sources can be predicted from their dark matter halo properties and environment, then we can account for the ionizing photons produced by these sources in large-box simulations of the reionization process without resolving the baryonic physics in detail.

This work employs a machine-learning method called Explainable Boosting Machines (EBMs; Lou et al. 2013) to infer how stellar mass $M_\star$ and SFR depend on the physical parameters θ of a host galaxy. In this work, we use the galaxy populations from the CROC simulations to provide our training and test data that populate samples in the multidimensional parameter space of $M_\star$-SFR-θ. For the additional parameters θ, we use a wide range of physical characteristics measured for galaxies in CROC, including the virial mass $M_{\rm vir}$, redshift z, environmental properties averaged on a length scale R, and the maximum peak circular velocity $v_{\rm peak}$. We can then use this approximate machine-learning-based EBM model for galaxy formation as a basis for future development to incorporate the CROC galaxy population as sources in lower-resolution, large-volume reionization simulations.

EBMs represent a form of generalized additive models (GAMs; Hastie & Tibshirani 1986) where the dependencies of a *target* quantity, such as $M_\star$ or SFR, on each physical parameter θ_i are encapsulated by feature functions of one parameter (e.g., $f^i(\theta_i)$) or interaction functions of two parameters (e.g., $f^{ij}(\theta_i, \theta_j)$). An EBM model is trained to fit these functions from a provided multidimensional data set. The predicted value of the target quantity given the parameters (e.g., $\gamma(M_\star|\boldsymbol{\theta})$) is then a sum of the functions f^i and f^{ij}. EBM models are often described as *interpretable* because the magnitudes of the functions f^i and f^{ij} directly indicate the relative importance of θ in determining the target quantity. If a given parameter θ_i is unimportant for determining the target quantity, the EBM will find $f^i \to 0$. A formal definition of the EBM is provided in Section 2.1.

Previous works have applied machine-learning models to infer connections between simulated galaxy properties. Lovell et al. (2022) use a tree-based learning method called extremely randomized trees to map baryon information to dark matter halos in the EAGLE simulations. Xu et al. (2021) train a random forest to predict the number of central and satellite galaxies in dark matter halos in the Millennium simulation. Machado Poletti Valle et al. (2021) used an XGBoost model to predict gas shapes in dark matter halos in the IllustrisTNG simulations. Bluck et al. (2022) used random forest classifiers to study quenching mechanisms in observations, semi-analytical models, and cosmological simulations. Piotrowska et al. (2022) also used random forest classifiers to examine how supermassive black hole feedback quenches central galaxies in the EAGLE, Illustris, and IllustrisTNG simulations. McGibbon & Khochfar (2022) used extremely randomized trees to predict the baryonic properties of subhalos in the IllustrisTNG simulations. Our approach complements these prior works by studying the detailed connection between the halo and environmental properties, SFR, and stellar mass in a model that can be directly implemented in future large-volume cosmological simulations with limited spatial resolution.

The paper is organized as follows. In Section 2, we review the EBM methodology, define our training data set and procedure, and introduce the evaluation metrics used to assess the performance of the model. In Section 3, we present the average contribution of each parameter to the target quantities, the best-fit feature and interaction functions, and the performance of the model in determining the distributions of stellar mass and SFR as a function of halo virial mass. We then explore in Section 4 methods for constructing *composite* EBM (CEBM) models to recover the stellar mass and SFR of simulated galaxies that only use instantaneous halo virial properties and environmental measures (i.e., excluding $v_{\rm peak}$). We discuss our results in Section 5, and summarize them and conclude in Section 6. The appendices of the paper provide detailed model results for the EBM for $M_\star$ (Appendix A), the mathematical formalism of the CEBM model (Appendix B), and detailed CEBM model results for SFR (Appendix C), and stellar mass (Appendix D).

2. Methods

To infer the connection between $M_\star$, SFR, and other physical properties of simulated galaxies, we apply EBM models to the CROC simulated galaxy catalogs. In Section 2.1, we define the EBM model. We select our model parameters and describe the simulated galaxy catalog used to train the model in Section 2.2. The training procedure is outlined in Section 2.3.

2.1. EBMs

EBM (Lou et al. 2013) models provide a fitted representation of the relationship between the target quantities y and the parameters θ. EBMs are an extension of GAMs (Hastie & Tibshirani 1986), which represent target quantities y as the sum of learned univariate functions $f^i(\theta_i)$ that depend on only one parameter θ_i. EBMs extend GAMs by including both univariate functions $f^i(\theta_i)$ and bivariate functions $f^{ij}(\theta_i, \theta_j)$ that represent dependencies on pairs of features (θ_i, θ_j) beyond the dependence of the target quantity on either feature independently. Both EBMs and GAMs are forms of regression where the feature functions f^i and f^{ij} can be quite general.

The EBM aims to encode the average dependence of a target quantity y on the parameters θ. Mathematically, an EBM can therefore be represented as

$$\gamma(y|\boldsymbol{\theta}) = \beta_y + \sum_{i=0}^{n_{\rm p}-1} f_y^i(\theta_i) + \sum_{i=0, i\neq j}^{n_{\rm p}-1} \sum_{j=0}^{n_{\rm p}-1} f_y^{ij}(\theta_i, \theta_j), \quad (1)$$

where $\gamma(y|\boldsymbol{\theta})$ is the predicted value of the target quantity y given $n_{\rm p}$ parameters $\boldsymbol{\theta} \in \mathbb{R}^n$ from the data set. We will refer to learned parameter β_y as the baseline value of the target quantity y. Though f_y^i and f_y^{ij} can be any interpretable function (e.g., linear regression, splines, etc.), Lou et al. (2012) found that gradient boosted trees (Friedman 2001) work best in practice. Using gradient boosted trees, the functions f_y^i and f_y^{ij} will be piecewise one- and two-dimensional functions, respectively. By expressing the dependence of y on θ directly through the functions f_y^i and f_y^{ij}, EBMs are interpretable and decomposable. Further, after training is complete the learned tree-based functions f_y^i and f_y^{ij} can be formulated as look-up tables for performant inference.

2.2. Simulated Galaxy Catalog Training Set

To engineer an EBM that describes the connection between simulated galaxy properties, their host dark matter halos, and features of the extrinsic environment, we turn to established observations and theoretical modeling to inform our choices for constructing a training data set.

The stellar-to-halo mass relation has been directly constrained out to redshifts $z \lesssim 0.05$ and galaxy masses $M_{\rm vir} > 10^{12} M_\odot$ using galaxy kinematics (e.g., More et al. 2009; Li et al. 2012), X-ray observations (e.g., Lin et al. 2004;

Kravtsov et al. 2018) and gravitational lensing (e.g., Mandelbaum et al. 2005; Velander et al. 2014). These constraints can be extended to higher redshifts ($z < 10$) and lower masses ($M_{\rm vir} < 10^{10}$) by including halo–galaxy connection modeling (e.g., Nelson et al. 2015; Croton et al. 2016; Rodríguez-Puebla et al. 2017; Behroozi et al. 2019; Girelli et al. 2020). Such models consistently infer that the average stellar mass of galaxies increases with halo mass.

At fixed redshift and halo mass, average galaxy masses of central galaxies differ from satellite galaxies. Halos grow through hierarchical merging, in which small halos merge to form larger halos. As subhalos merge into larger halos, tidal heating and stripping reduce the mass of the more extended dark matter halo, while the satellite galaxy mass remains largely unaffected. For this reason, galaxy mass often correlates better with halo properties at the time of accretion than the current halo mass (e.g., Conroy et al. 2006; Vale & Ostriker 2006; Moster et al. 2010; Reddick et al. 2013). In particular, SHAM models find that using the halo peak circular velocity, $v_{\rm peak}$, to assign galaxy mass and/or luminosity best reproduces observed galaxy clustering (e.g., Hearin et al. 2013; Reddick et al. 2013; Lehmann et al. 2017).

SFRs correlate tightly with galaxy masses, and increase with redshift at fixed stellar mass (e.g., Noeske et al. 2007; Stark et al. 2009; Bouwens et al. 2012). While these trends hold on average, there is a distinct bimodal distribution in the SFRs of galaxies, corresponding to star-forming and quiescent populations (e.g., Balogh et al. 2004). The observed fraction of quiescent galaxies increases as the universe evolves (e.g., Tomczak et al. 2014), with the interpretation that some mechanism turns off star formation in galaxies. Many quenching mechanisms have been proposed, including secular/mass quenching (e.g., Kauffmann et al. 2004; Contini et al. 2020) and environmental quenching (e.g., Davies et al. 2016; Trussler et al. 2020). Which of these processes dominate may vary with redshift (Kalita et al. 2021).

Overdense environments may cause environmental quenching, by providing close pairs that can suppress gas accretion (*strangulation*), removing gas through ram pressure stripping, or disrupting by interactions with other galaxies (*harassment*). Environment thereby influences SFRs, and low-mass satellite galaxies are typically the most prone to environmental quenching (e.g., Davies et al. 2019).

Given these established trends, galaxy mass and SFR may depend on redshift, halo mass, peak circular velocity, and environmental properties. We will therefore select corresponding parameters from the CROC simulated galaxy catalogs to provide our data set for training the EBM models. The CROC simulations are cosmological simulations with volumes of up 100 comoving megaparsecs and spatial resolutions in physical units approaching 100 pc. Further details of the simulations can be found in Gnedin (2014). At a range of redshifts z during the simulation, the computational grid and particle properties are written to disk. These simulation snapshots are post-processed to identify virialized galaxies, as described in Zhu et al. (2020), and the properties of the simulated galaxies are recorded in catalogs. Merger trees are used to identify the properties of simulated galaxies across redshift.

For our target quantities y, in this work we will model stellar mass $M_\star$ $[h^{-1}M_\odot]$ and SFR $[M_\odot\,{\rm yr}^{-1}]$. The parameters θ selected from the simulated catalog include both intrinsic properties of galaxies and extrinsic properties set by the large-scale environment. For intrinsic properties, we include the galaxy virial mass $M_{\rm vir}$ $[h^{-1}M_\odot]$, the redshift z at which the simulated galaxy properties were measured, and the maximum peak circular velocity $v_{\rm peak}$ $[{\rm km\,s}^{-1}]$ measured over the formation history of each galaxy. The extrinsic properties used are defined by a length scale R measured relative to each simulated galaxy. We follow convention and substitute R with a numerical value that indicates a number of comoving megaparsecs (e.g., σ_8 is the rms density fluctuations measured in spheres of radius of $R = 8$ Mpc). We compute an environmental density $\rho_1 \equiv 1 + \Delta_1$, where Δ_1 is the dimensionless matter overdensity measured within 1 Mpc. We include an environmental gas temperature T_1 [K] averaged on 1 Mpc scales. From each simulated galaxy we also find the virial mass $M_{\rm max,0.1}$ of the most massive neighboring halo within 100 kpc. We then define the mass ratio $\Upsilon_{0.1} \equiv 1 + M_{\rm max,0.1}/M_{\rm vir}$ There are many other measurable properties of simulated dark matter halo that may correlate with the stellar mass and SFRs of the simulated baryonic galaxies they contain. For instance, Lehmann et al. (2017) found that at fixed halo mass, concentration influenced the simulated stellar mass content. In choosing which simulated data to fit when training our model, we chose to fit $v_{\rm peak}$ instead of concentration because it also reflects the shape of the galaxy potential and does not share the same redshift dependence as concentration for a given halo.

Table 1
Hyperparameters Used to Train the InterpretML (Nori et al. 2019) Implementation of the EBM

EBM Training Hyperparameters	
Hyperparameter	Value
Binning $\mathcal{B}$	"uniform"
Maximum bins, univariate $\mathcal{Q}_{\rm max}$	256
Maximum bins, bivariate $\mathcal{Q}_{\rm max,2D}$	32×32
Learning rate $\mathcal{R}_l$	0.01

Note. All other model hyperparameters were set to the default values for InterpretML version 0.2.7.

The simulated galaxy catalogs include roughly 8,426,327 objects covering a wide range of halo masses, stellar masses, SFRs, redshifts, and other extrinsic properties. From the catalog of simulated galaxies, objects with an SFR < 0.001 $M_\odot\,{\rm yr}^{-1}$ were excluded owing to resolution effects artificially limiting their SFRs. After this culling, the catalog contained 5,950,357 objects that formed our data set. At this stage, we constructed the training and test data sets from our catalog using the parameter vector $\boldsymbol{\theta} = [M_{\rm vir}, z, v_{\rm peak}, \rho_1, T_1, \Upsilon_{0.1}]$ to model the target quantities $\boldsymbol{y} = [M_\star, {\rm SFR}]$. We use k-fold cross-validation (Hastie et al. 2001) with $k = 5$, such that the test/training split is 20%/80% for each k-folding.

2.3. Training Procedure

The calculations presented in this paper leverage the InterpretML (Nori et al. 2019) implementation of EBMs, using the hyperparameters in Table 1. These InterpretML hyperparameters control the number of bins in the piecewise f_y^i and f_y^{ij} functions ($\mathcal{Q}_{\rm max}$, $\mathcal{Q}_{\rm max,2D}$), the distribution of bins across the fitted domain ($\mathcal{B}$), and the learning rate of the optimization scheme ($\mathcal{R}_l$). The hyperparameters were selected by testing

various combinations of the values. We find that setting $\mathcal{B} = uniform$ improves model performance at the edges of the data distribution where there are fewer samples. Further, higher values of $\mathcal{Q}_{\max}$ and $\mathcal{Q}_{\max,2D}$ do not significantly improve performance but do affect the training runtime. The Nori et al. (2019) implementation trains an EBM in two phases. First, the univariate functions are optimized using a gradient boosting approach applied round-robin on each parameter, as detailed in Lou et al. (2012). After the univariate functions have converged, the interaction terms are computed and the bivariate functions are optimized according to the GA2M/FAST algorithms detailed in Lou et al. (2013). During training we use k-fold cross validation, and merge the training and test data sets for the final performance evaluation of the model.

We evaluate the EBM performance using the mean absolute error (MAE), a variance metric r^2, and the total outlier fraction ζ_k. These statistics provide measures of how well the EBM reproduces the mean trends in the target quantities y as a function of the features θ, the width of the distribution about the mean trends in the training data, and the tails of that distribution.

We calculate the MAE of the model applied to the simulated galaxy sample as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=0}^{N-1}|y_i - \hat{y}_i|, \tag{2}$$

where N is the number of objects, y_i is the true value of the target quantity for object i, and $\hat{y}_i$ is the predicted value from the model for object i.

We compute the $r^2 \in (-\infty, 1]$ variance metric as

$$r^2 = 1 - \frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{N-1}(y_i - \bar{y})^2}, \tag{3}$$

which provides a measure of how well the model captures the variance in the data relative to the mean $\bar{y}$, with $r^2 = 1$ reflecting a perfect reproduction of the distribution of y in the training data set. Note that the feature and interaction functions f_y^i and f_y^{ij} have a finite range, and thus not all values y_i can be represented by Equation (1) even when the input parameters θ vary about the mean trends with halo mass or environment. Hence, even for high-quality EBM models $r^2 < 1$ and we expect outliers. The ζ_k metric represents the fraction of the total data set that lies outside the range of predicted values, $\{\hat{y}\}$, as a function of one of the features θ_k. We define

$$\zeta_k = \frac{1}{N}\sum_{i=0}^{N-1} g_{k,i}(y_i, \theta_{k,i}), \tag{4}$$

where the index i runs over the total number of samples N and $g_{k,i}(y_i, \theta_{k,i})$ is a function that returns 1 if the true target quantity for object i lies outside the predicted range, i.e., $y_i \notin \{\hat{y}\}$. In practice, we compute the outlier fraction for feature $k = \log_{10} M_{\rm vir}$, and use 2D histograms of $(y_i, \theta_{k,i})$ and $(\hat{y}_i, \theta_{k,i})$ to calculate $g_{k,i}$.

In Table 2, we present the evaluation metrics for our EBM model fully trained on the simulated galaxy catalog. For the EBM model for SFR ($y = \mathrm{SFR}$), we find an MAE $\sim 0.14\ \log_{10} M_\odot\ \mathrm{yr}^{-1}$, a variance metric $r^2 \sim 0.9$, and an outlier fraction of $<3\%$. For the EBM model for stellar mass ($y = M_\star$), we report an MAE $\sim 0.19\ \log_{10} M_\odot\ \mathrm{yr}^{-1}$, a variance metric $r^2 \sim 0.88$, and an outlier fraction of $<1\%$. The good performance of the EBM models in these metrics reflects the ability of the EBMs to capture both the mean trends and full distributions of the target quantities $y = [M_\star, \mathrm{SFR}]$ in the training set given the parameters $\boldsymbol{\theta} = [M_{\rm vir}, z, \nu_{\rm peak}, \rho_1, T_1, \Upsilon_{0.1}]$. We describe the detailed model results in Section 3.

Table 2
Training Results for the EBM Using k-fold Cross-validation

EBM Training Results		
Metrics	$\gamma(\mathrm{SFR}\|\boldsymbol{\theta})$	$\gamma(M_\star\|\boldsymbol{\theta})$
r^2	0.898 ± 0.0003	0.882 ± 0.0001
ζ	0.029 ± 0.004	0.008 ± 0.0010
	$\log_{10}\mathrm{SFR}\ [M_\odot\ \mathrm{yr}^{-1}]$	$\log_{10} M_\star\ [M_\odot]$
MAE	0.144 ± 0.0001	0.189 ± 0.0001

Note. See Section 2.3 for more information on the training process. Reported are values for the variance metric r^2, the outlier fraction ζ, and the MAE. Uncertainties are computed from the variation among the k-fold trials.

3. Results

After training the EBM model to reproduce the dependence of the target quantities $M_\star$ and SFR on the parameters θ, the relationships between the target quantities and the parameters can be analyzed. Below, we provide several analyses that quantify how the target quantities relate to the parameters and illustrate the performance of the EBM for our astrophysical applications.

3.1. Average Contribution

A key advantage of using EBM models over *black box* models (e.g., neural networks) is their clear interpretability (see Section 2.1). The contribution of each parameter θ_i to the model of the target quantity y is provided by the output functions f_y^i and f_y^{ij}.

Since these functions are vectors or two-dimensional matrices with a number of elements equal to the number of bins n_b in the piecewise function (see Table 1), a summary scalar quantity for each feature function is helpful for comparing their relative importance. We can define the *average contribution* $\bar{f}_y^i$ that provides the average absolute value of f_y^i or f_y^{ij}, with the average computed over the number of bins n_b and weighted by the number of samples in each bin. Mathematically, we can write

$$\bar{f}_y^i = \frac{\sum_{j=0}^{n_b-1}|f(\theta_{i,j})|N_j}{\sum_{j=0}^{n_b-1}N_j}, \tag{5}$$

where f is the feature function being averaged (f_y^i or f_y^{ij} from Equation (1)), $\theta_{i,j}$ is the value of the parameter θ_i in the jth bin, and N_j is the number of samples in bin j. Intuitively, the average contribution $\bar{f}_y^i$ summarizes the importance of each parameter θ_i for determining the target quantities when averaged over the samples in the final, merged data set.

The average contributions of each feature (f_y^i) or combination of features (f_y^{ij}) are computed from the EBM. In each case, we rank order the features by decreasing average contribution and focus on the seven features or feature combinations with the largest average contribution. In each case, the most

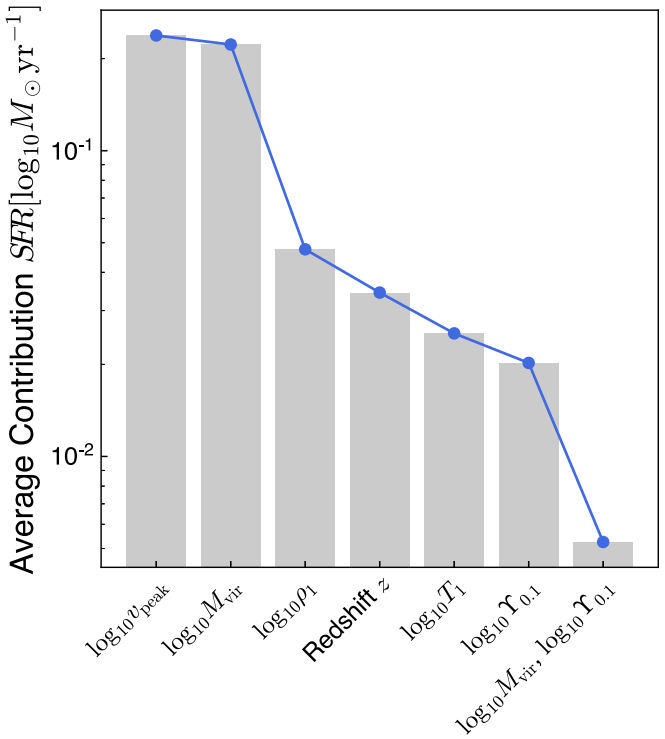


Figure 1. Top seven features with the highest average contribution in the EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ targeting the SFR. In order of decreasing importance, these features include peak circular velocity $v_{\rm peak}$, virial mass $M_{\rm vir}$, environmental density ρ_1, redshift z, environmental temperature T_1, the mass ratio of nearby halos $\Upsilon_{0.1}$, and the interaction between virial mass $M_{\rm vir}$ and $\Upsilon_{0.1}$. The average contribution is calculated using the average of the absolute value of the feature functions weighted by the number of samples in each bin (see Equation (5)).

important feature has an average contribution more than an order of magnitude larger than the seventh-ranked feature.

3.1.1. EBM Model Targeting SFR

Figure 1 shows the average contribution of the top seven features for the EBM model targeting SFR $\log_{10}\mathrm{SFR}$. In decreasing order, the seven most important features in determining SFR are maximum peak circular velocity $v_{\rm peak}$, virial mass $M_{\rm vir}$, environmental density ρ_1, redshift z, environmental temperature T_1, mass ratio of nearby halos $\Upsilon_{0.1}$, and the interaction between $M_{\rm vir}$ and $\Upsilon_{0.1}$. The numerical values for the average contributions are provided in Table 3. The baseline value of SFR is $\beta_{\log_{10}\mathrm{SFR}} = -2.1151\ [\log_{10} M_\odot\,\mathrm{yr}^{-1}]$, typical of halos with $\log_{10} M_{\rm vir} \sim 9$. The average contribution of $v_{\rm peak}$ and $M_{\rm vir}$ are quite similar, providing $\Delta \log_{10}\mathrm{SFR} > 0.2$ on average, but their interaction term is small with $\bar{f}\,(\log_{10} v_{\rm peak}, \log_{10} M_{\rm vir}) \ll 0.01$. Therefore, peak circular velocity and virial mass provide important contributions to determining the SFR, and the univariate dependence of the SFR on these properties accounts for most of their contribution. At the few-percent level, environmental density, redshift, environmental gas temperature, and the presence of nearby massive halos also contribute.

The feature functions f_y^i for each feature are plotted in Figure 2. The functions indicate that there are positive correlations between the SFR $\log_{10}\mathrm{SFR}$ and either the peak circular velocity $v_{\rm peak}$, virial mass $M_{\rm vir}$, or environmental density ρ_1. The SFR increases with increasing environmental temperature $\log_{10} T_1$, but near $T_1 \approx 10^4$ K the univariate function shows an enhancement just as hydrogen becomes mostly neutral and a deficit near the temperature at which it becomes ionized. SFR increases with decreasing redshift over the range $z \sim 5$–15, becoming more efficient after reionization.

The interaction functions f_y^{ij} learned by the EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ targeting the SFR are plotted as *heat maps* in Figure 3. Most interaction functions do not contribute significantly to the SFR, and change the SFR by $\Delta \log_{10}\mathrm{SFR} \lesssim 0.05$. However, halos with low neighboring halo mass ratios $\Upsilon_{0.1}$ and large peak circular velocity $v_{\rm peak}$ have their SFR enhanced by $\Delta \log_{10}\mathrm{SFR} \approx 0.15$. Rephrased, locally dominant halos with large peak circular velocity show enhanced star formation. Such enhancements likely owe to recent merger activity.

While Equation (1) represents a complex, multidimensional manifold that provides the SFR as a function of the parameters θ, the distributions of simulated and predicted SFR as a function of a single parameter provide a graphical summary of the EBM model performance. Figure 4 shows the simulated and predicted SFR as a function of virial mass $\log_{10} M_{\rm vir}$, and we refer to this figure as the *model summary*. Shown in this model summary are the distributions of SFR in the CROC simulated galaxy catalogs with virial mass and the SFR predicted by the EBM model $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ using the parameters θ measured for each simulated galaxy. The EBM model captures roughly 97% of the simulated distribution of SFR with virial mass. The predicted range of the EBM model then corresponds to the range of SFR values successfully recovered as shown in panel (c) of Figure 4, while panel (d) shows the range of SFR values that are not recovered by the EBM model.

Given the combined complexity of the average contribution measures, univariate feature functions, and bivariate interaction functions, in what follows we will show the summary figure for other EBM models in the main text. For completeness, the average contribution, feature function, and interaction function figures for each model will be presented in the appendices.

Table 3
Summary of the EBM Model Trained to Predict SFR

Average Contributions for the $\gamma(M_\star\|\boldsymbol{\theta})$ EBM	
Feature	Value $[\log_{10} M_\odot\,\mathrm{yr}^{-1}]$
$\beta_{\log_{10} M_\star}$	−2.1151
$\bar{f}\,(\log_{10} v_{\rm peak})$	0.2380
$\bar{f}\,(\log_{10} M_{\rm vir})$	0.2224
$\bar{f}\,(\log_{10} \rho_1)$	0.0475
$\bar{f}\,(z)$	0.0343
$\bar{f}\,(\log_{10} T_1)$	0.0252
$\bar{f}\,(\log_{10} \Upsilon_{0.1})$	0.0202
$\bar{f}\,(\log_{10} M_{\rm vir}, \log_{10} \Upsilon_{0.1})$	0.0052

Note. The first entry, $\beta_{\log_{10}\mathrm{SFR}}$, is the baseline value learned model (see Section 2.1). The next seven entries are the average contributions of the most important feature functions listed in descending order (see Equation (5)).

3.1.2. EBM Model Targeting Stellar Mass $\mathrm{M}_\star$

An EBM model $\gamma(M_\star|\boldsymbol{\theta})$ targeting stellar mass $M_\star$ using the properties θ can be constructed through simple retraining. Using the simulated galaxy catalogs from CROC, we retrain the EBM to model $M_\star$ against θ. We report the average contribution, univariate feature functions, and bivariate interaction functions for $\gamma(M_\star|\boldsymbol{\theta})$ in Appendix A. For reference, the baseline value of $M_\star$ is

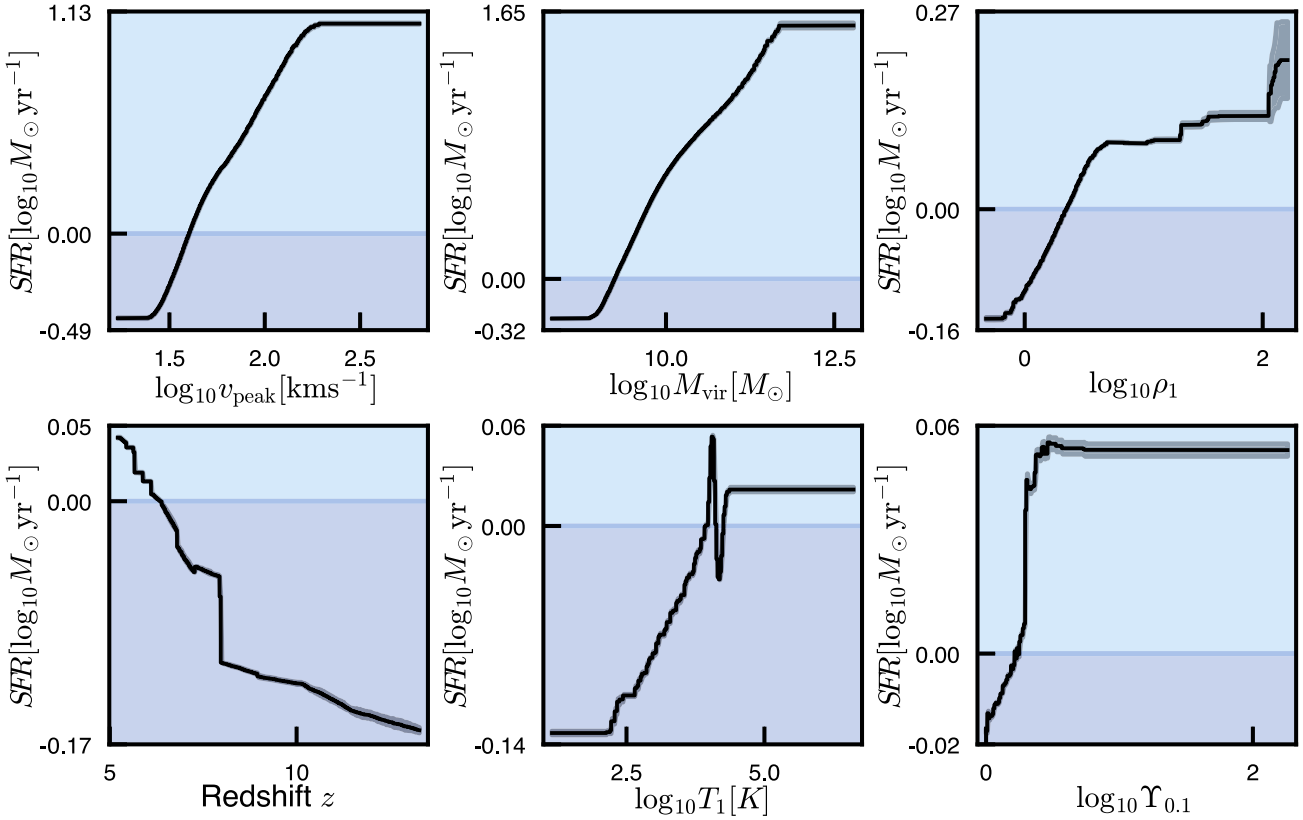


Figure 2. Learned univariate feature functions f_y^i for the EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ trained to predict the SFR. Shown (left to right) are the feature functions for peak circular velocity $v_{\rm peak}$, virial mass $M_{\rm vir}$, environmental density ρ_1, redshift z, environmental temperature T_1, and nearby halo mass ratio $\Upsilon_{0.1}$. Light blue areas indicate regions where $f_y^i > 0$ and dark blue areas indicate regions where $f_y^i < 0$. The shaded areas show the variation in f_y^i between the k-fold iterations.

$\beta_{\log_{10} M_\star} = 5.9629$ [$\log_{10} M_\odot\ {\rm yr}^{-1}$] (see Table 7 in Appendix A) typical of halos with $\log_{10} M_{\rm vir}/M_\odot \sim 9$.

Figure 5 shows the model summary for the EBM model $\gamma(M_\star|\boldsymbol{\theta})$. The EBM model provides an excellent representation of the distribution of stellar masses for the CROC simulated galaxy catalog. As the lower right panel of Figure 5 indicates, the $\gamma(M_\star|\boldsymbol{\theta})$ model results in few outliers for the CROC simulated galaxies and has an outlier fraction of $\lesssim$1%. Given the galaxy properties $\boldsymbol{\theta} = [M_{\rm vir},\ z,\ v_{\rm peak},\ \rho_1,\ T_1,\ \Upsilon_{0.1}]$, the distribution of stellar masses for CROC simulated galaxies can be recovered to 99% accuracy.

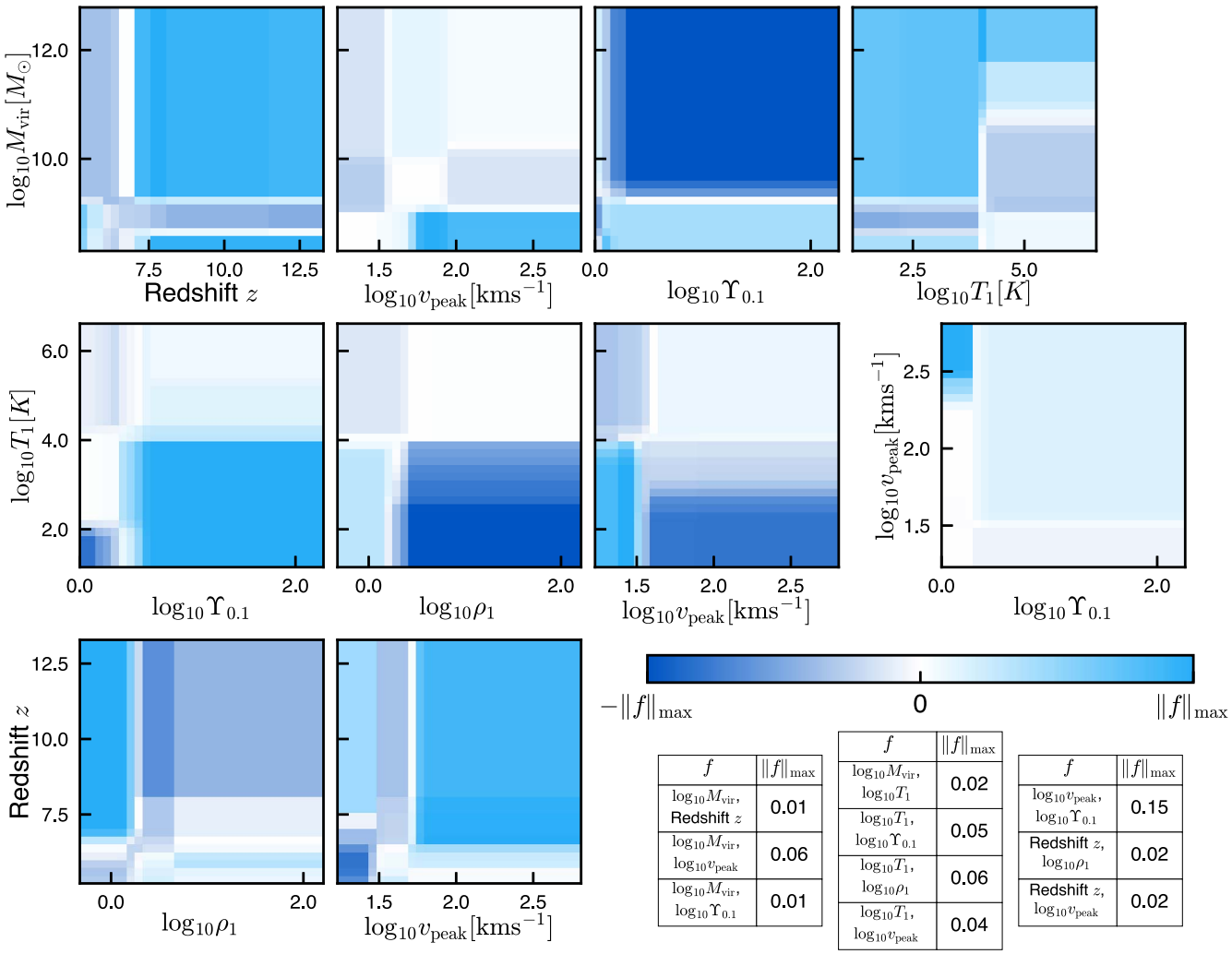
4. CEBMs for Restricted Parameter Sets

The EBM models $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ and $\gamma(M_\star|\boldsymbol{\theta})$ presented in Sections 3.1.1 and 3.1.2 are constructed using the parameter set $\boldsymbol{\theta} = [M_{\rm vir},\ z,\ v_{\rm peak},\ \rho_1,\ T_1,\ \Upsilon_{0.1}]$. Our results show that the full distribution of SFR and stellar mass in the simulated CROC galaxy catalogs can be recovered accurately with only ≈1%–3% outliers. These EBM models can therefore be applied to cosmological simulations using the parameters θ measured from simulated galaxy catalogs to recover the distribution of SFR and stellar mass computed by CROC.

The parameters θ include the peak circular velocity $v_{\rm peak}$, which requires both time-dependent tracking of formation histories for individual galaxies and high spatial resolution to capture the peak of the rotation curve for each object. As a result, as expressed above the models $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ and $\gamma(M_\star|\boldsymbol{\theta})$ cannot be applied directly to cosmological simulations with low spatial resolution or without merger trees to capture formation histories.

Instead of fitting EBM models using the full parameter set θ, consider the construction of an EBM model using the restricted parameter set $\boldsymbol{\theta}' = [M_{\rm vir},\ z,\ \rho_1,\ T_1,\ \Upsilon_{0.1}]$ that does not include $v_{\rm peak}$. The parameters $\boldsymbol{\theta}'$ can all be measured directly in cosmological simulations with sufficient resolution to capture individual galaxy-mass halos without the need to track merger trees. The EBM models $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ and $\gamma(M_\star|\boldsymbol{\theta}')$ using the restricted parameter set $\boldsymbol{\theta}'$ perform substantially less well than the models $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ and $\gamma(M_\star|\boldsymbol{\theta})$ trained on the full parameter set θ that includes $v_{\rm peak}$. With the restricted parameter set $\boldsymbol{\theta}'$, the EBM model shows 7.6% outliers when targeting SFR and 2.8% when targeting $M_\star$. Comparing with the outlier fractions reported in Table 2 for the full parameter set including $v_{\rm peak}$, the EBM model trained on the restricted data set has degraded its performance by a factor of ∼2–3. Further, the R^2 of the EBM trained on $\boldsymbol{\theta}'$ degraded by 0.068 and 0.052 for $M_\star$ and SFR, respectively. The improved performance of a single EBM model, including $v_{\rm peak}$, in predicting $M_\star$ should not be surprising. Given that dark matter is stripped from subhalos preferentially relative to their stellar mass, while $v_{\rm peak}$ reflects the maximum peak circular velocity of the subhalo's potential at a time when most of its stellar mass was in place, we expect $v_{\rm peak}$ to be informative in fitting to stellar mass.

To ameliorate the poorer performance of the EBM models trained on restricted parameter sets, we use a CEBM model. Given a target quantity y and a parameter set $\boldsymbol{\theta}'$, we fit a *base* EBM $\gamma(y|\boldsymbol{\theta}')$ in the same manner as fitting the EBMs $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ or $\gamma(M_\star|\boldsymbol{\theta})$. We construct a data set from the galaxies whose y



f	$\|f\|_{\rm max}$
$\log_{10} M_{\rm vir}$, Redshift z	0.01
$\log_{10} M_{\rm vir}$, $\log_{10} v_{\rm peak}$	0.06
$\log_{10} M_{\rm vir}$, $\log_{10} \Upsilon_{0.1}$	0.01

f	$\|f\|_{\rm max}$
$\log_{10} M_{\rm vir}$, $\log_{10} T_1$	0.02
$\log_{10} T_1$, $\log_{10} \Upsilon_{0.1}$	0.05
$\log_{10} T_1$, $\log_{10} \rho_1$	0.06
$\log_{10} T_1$, $\log_{10} v_{\rm peak}$	0.04

f	$\|f\|_{\rm max}$
$\log_{10} v_{\rm peak}$, $\log_{10} \Upsilon_{0.1}$	0.15
Redshift z, $\log_{10} \rho_1$	0.02
Redshift z, $\log_{10} v_{\rm peak}$	0.02

Figure 3. Most important learned interaction functions f_y^{ij} for the EBM model $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ targeting the SFR, as a function of their parameter pairs. Each panel shows the contribution of the bivariate interaction terms, normalized such that the color map ranges between plus or minus the maximum of the norm of each function $\|f\|_{\rm max}$. Light blue areas indicate regions of joint parameter space where the feature interactions contribute positively to the SFR, while dark blue areas indicate regions with negative contributions. The table lists $\|f\|_{\rm max}$ for the interaction functions, each with units $\log_{10} M_\odot\ \mathrm{yr}^{-1}$. In absolute terms, the largest interaction occurs for halos with large peak circular velocity $v_{\rm peak}$ and no large neighboring halos ($\Upsilon_{0.1} \approx 0$). The other interaction functions are relatively weak, and contribute changes to log SFR $\lesssim 0.05$.

values lie outside the predictions from $\gamma(y|\boldsymbol{\theta}')$, and then fit an *outlier* EBM $\delta(y|\boldsymbol{\theta}')$ to these discrepant samples. We then weight the base and outlier EBMs to construct the CEBM model $\Gamma(M_\star|\boldsymbol{\theta}')$ using a *classifier* EBM $\phi_y(\boldsymbol{\theta}')$. Instead of fitting the change in SFR or stellar mass at a given sample in θ', the classifier EBM fits the log odds that a given sample in θ' is an outlier. We then define $\phi_y(\boldsymbol{\theta}')$ to be the sigmoid of these log odds, such that $\phi_y(\boldsymbol{\theta}') \in [0, 1]$. The CEBM can then be written as

$$\Gamma(M_\star|\boldsymbol{\theta}') = [1 - \phi_y(\boldsymbol{\theta}')]\gamma(y|\boldsymbol{\theta}') + \phi_y(\boldsymbol{\theta}')\delta(y|\boldsymbol{\theta}'). \quad (6)$$

We describe the CEBM approach in more detail in Appendix B, and provide information on the CEBMs $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ and $\Gamma(M_\star|\boldsymbol{\theta}')$ in Appendices C and D.

Table 4 lists the evaluation metrics for the training of CEBM models targeting SFR and stellar mass without using $v_{\rm peak}$. The outlier fraction has improved to $\approx$5% for CEBM model $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ and to $\lesssim$2% for $\Gamma(M_\star|\boldsymbol{\theta}')$. The average parameter contributions and baseline value $\beta_{\log_{10}\mathrm{SFR}}$ from $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ are provided in Table 5 and for the CEBM targeting stellar mass in Table 6. The univariate feature functions and bivariate interaction functions for the CEBM models $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ and $\Gamma(M_\star|\boldsymbol{\theta}')$ are provided in Appendices C and D.

Figure 6 shows the model summary for the CEBM targeting SFR, and Figure 7 shows the model summary for the CEBM targeting stellar mass. As both models demonstrate, the CEBM model accurately recovers the distribution of SFR and stellar mass in the CROC simulated galaxy sample. Between the models, the outlier fraction is only $\approx$2%–5% despite using the restricted set of parameters $\boldsymbol{\theta}'$ that does not include $v_{\rm peak}$ or any time-dependent tracking of individual systems.

5. Discussion

EBM models provide a method to statistically infer relationships present in high-dimensional data. Given their statistical nature, EBM models remain ignorant of the physics that generate the connection between SFR, stellar mass, and the properties of dark matter halos that host galaxies. Nonetheless, given the results of detailed physical modeling in the form of simulated galaxy catalogs from cosmological simulations, the

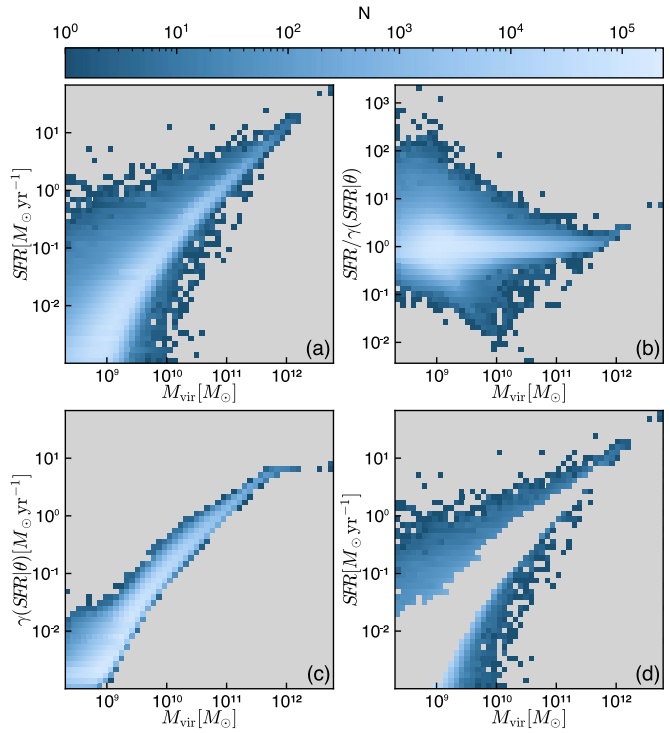


Figure 4. Summary of the EBM model $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ targeting SFR as a function of virial mass. The upper left panel shows the two-dimensional distribution of SFR with $M_{\rm vir}$ for galaxies in the CROC simulations, with the color scale showing the number of simulated galaxies at each [SFR, $M_{\rm vir}$] location. The lower left panel shows the EBM model results for the distribution of SFR with $M_{\rm vir}$, where the SFR is computed from the EBM using the parameters $\boldsymbol{\theta} = [M_{\rm vir}, v_{\rm peak}, z, \rho_1, T_1, \Upsilon_{0.1}]$. The upper right panel shows the residuals between the simulated CROC galaxy SFRs and the EBM model results. The lower right panel shows the simulated CROC galaxy SFRs that lie outside the EBM model predictions. These outliers represent $\lesssim$3% of simulated CROC galaxies.

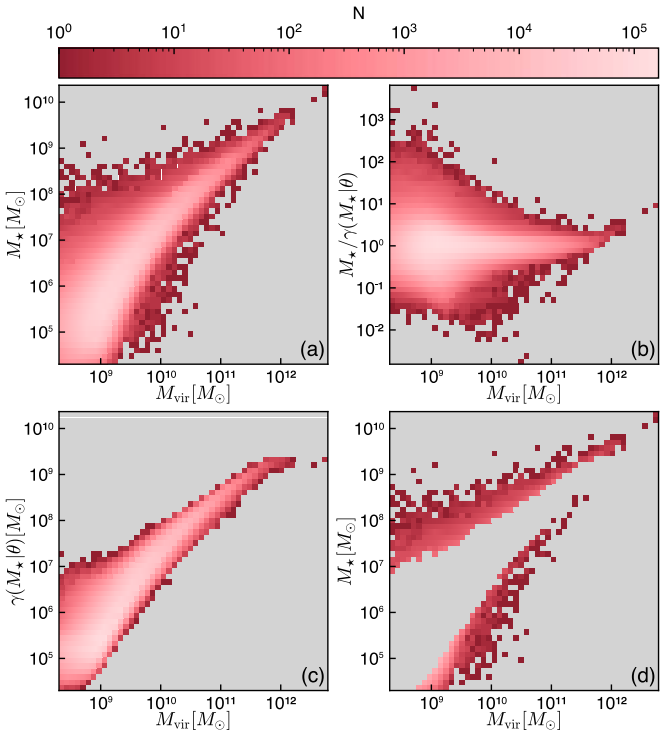


Figure 5. Summary of the EBM model $\gamma(M_\star|\boldsymbol{\theta})$ targeting stellar mass $M_\star$ as a function of virial mass. The upper left panel shows the distribution of $M_\star$ with virial mass $M_{\rm vir}$ in the CROC simulated galaxy catalogs, with the coloration indicating the number of galaxies at each [$M_\star$, $M_{\rm vir}$] location. The lower left panel shows the EBM model prediction of the stellar mass distribution with virial mass given in the input parameters $\boldsymbol{\theta} = [M_{\rm vir}, z, v_{\rm peak}, \rho_1, T_1, \Upsilon_{0.1}]$. The upper right panel shows the residuals between the simulated and predicted $M_\star$ vs. $M_{\rm vir}$ distribution, and the lower right panel shows the outliers in the simulated distribution not captured by the EBM model $\gamma(M_\star|\boldsymbol{\theta})$. The fraction of outliers is $\lesssim$1%.

EBM correctly identifies halo mass and maximum peak circular velocity as the most important halo properties for determining SFR and $M_\star$ (e.g., Figure 1). The EBM correctly infers that SFR and $M_\star$ increase with increasing halo mass or $v_{\rm peak}$, and the EBM univariate feature functions correctly identify the gas temperature at which star formation efficiency changes. To the extent that the physical connection between galaxy and halo properties are recorded in statistical relationships, the EBM models effectively recover some fraction of those relations. For instance, if we fit instead to $\mathrm{sSFR} \equiv \mathrm{SFR}/M_\star$ we find that redshift z becomes the most important feature, followed by $M_\star$ and $v_{\rm peak}$, as expected from the trends in both our simulations and observations (e.g., Feulner et al. 2005).

EBM models also provide a means to implement a *sub-grid* prescription for galaxy formation based on the properties of halos and their environments. The EBM models $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ and $\gamma(M_\star|\boldsymbol{\theta})$ capture better than 97% of the SFR and $M_\star$ distributions measured for simulated galaxies in the CROC simulations. The stellar masses and SFRs of galaxies in CROC could be accurately recovered by using only the halo and environmental parameters in $\boldsymbol{\theta} = [M_{\rm vir}, z, v_{\rm peak}, \rho_1, T_1, \Upsilon_{0.1}]$.

Using the CEBM model trained on the restricted parameter set $\boldsymbol{\theta}' = [M_{\rm vir}, z, \rho_1, T_1, \Upsilon_{0.1}]$, ≈95%–98% of the distribution of SFR and $M_\star$ of the CROC galaxies is recovered. One advantage of this parameter set is that the spatial resolution in the simulations required to compute them is less demanding than for $v_{\rm peak}$. A simulation with coarser resolution than CROC,

Table 4
Training Results for CEBM Models for SFR and $M_\star$ Using k-fold Cross-validation

CEBM Training Results		
Metrics	$\gamma(\mathrm{SFR}\|\boldsymbol{\theta})$	$\gamma(M_\star\|\boldsymbol{\theta})$
r^2	0.868 ± 0.0002	0.830 ± 0.0003
ζ	0.052 ± 0.0053	0.018 ± 0.0031
	$\log_{10}\mathrm{SFR}\ [M_\odot\mathrm{yr}^{-1}]$	$\log_{10} M_\star\ [M_\odot]$
MAE	0.165 ± 0.0001	0.233 ± 0.0002

Note. See Section 2.3 for more information on the training process. Reported are values for the variance metric r^2, the outlier fraction ζ, and the MAE. Uncertainties are computed from the variation among the k-fold trials.

such that the details of the star formation and feedback processes cannot be resolved, may still leverage the CEBM models $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ and $\Gamma(M_\star|\boldsymbol{\theta}')$ to model the SFR and stellar masses in dark matter halos. Further, the quantities $\boldsymbol{\theta}'$ used to train the CEBM models are measured at distinct redshifts such that no merger trees are required to recover accurately the CROC SFR and $M_\star$ distributions from halo and environmental properties. We note that for both the EBM and CEBM models the outlier fractions not well captured by the model are roughly percent level or less in the SFR or $M_\star$ distributions, and we expect that corresponding inaccuracies induced in, e.g., the

Table 5
Average Contribution to the CEBM Model $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ Trained to Predict SFR from the Parameter Set $\boldsymbol{\theta}'$

Overview of CEBM $\Gamma(\mathrm{SFR}\|\boldsymbol{\theta}')$	
Feature	Value $[\log_{10} M_\odot\ \mathrm{yr}^{-1}]$
$\beta_{\log_{10}\mathrm{SFR}}$	−1.7466
$\tilde{f}\,(\log_{10} M_{\mathrm{vir}})$	0.4327
$\tilde{f}\,(\log_{10} \rho_{\mathrm{l}})$	0.0625
$\tilde{f}\,(\log_{10} T_{\mathrm{l}})$	0.0327
$\tilde{f}\,(\log_{10} \Upsilon_{0.1})$	0.0215
$\tilde{f}\,(z)$	0.0190
$\tilde{f}\,(z,\ \log_{10} \rho_{\mathrm{l}})$	0.0077
$\tilde{f}\,(\log_{10} M_{\mathrm{vir}},\ \log_{10} \Upsilon_{0.1})$	0.0056

Note. The first entry, $\beta_{\log_{10}\mathrm{SFR}}$, is the learned baseline of the model. The next seven entries are the feature functions with the highest average contribution listed in descending order. The average contribution is calculated using the average of the absolute value of the base EBM function values weighted by the number of samples in each bin and the output of the classification EBM for each sample (see Appendix B.2 for more details).

Table 6
Summary of the CEBM $\gamma(M_\star|\boldsymbol{\theta}')$ Trained to Predict $M_\star$ Using the Restricted Parameter Set $\boldsymbol{\theta}'$

[1]Average Contributions for the CEBM $\gamma(M\star\|\theta')$g	
Feature	Value $[\log_{10} M_\odot]$
$\beta_{\log_{10} M\star}$	6.6995
$\tilde{f}\,(\log_{10} M_{\mathrm{vir}})$	0.5008
$\tilde{f}\,(z)$	0.0961
$\tilde{f}\,(\log_{10} \rho_1)$	0.0902
$\tilde{f}\,(\log_{10} T_1)$	0.0576
$\tilde{f}\,(\log_{10} \Upsilon_{0.1})$	0.0336
$\tilde{f}\,(z,\ \log_{10} \rho_1)$	0.0172
$\tilde{f}\,(\log_{10} M_{\mathrm{vir}},\ \log_{10} \rho_1)$	0.010

Note. The first entry, $\beta \log_{10} M\star$, is the learned baseline of the model. The next seven entries are the learned functions with the highest average contribution in descending order. The average contribution is computed via Equation (B2) (see Appendix B.2 for more details)

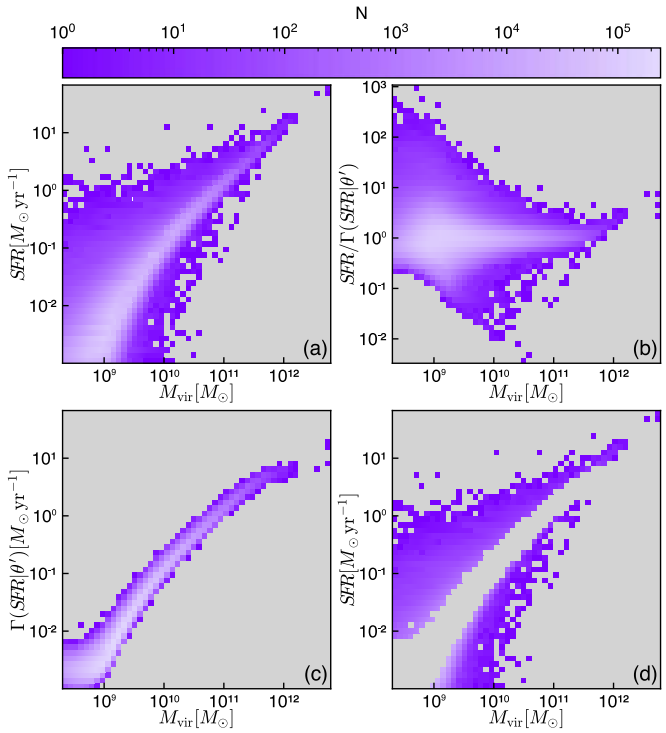


Figure 6. Summary of the CEBM model $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ targeting SFR as a function of virial mass. The upper left panel shows the two-dimensional distribution of SFR with M_{vir} for galaxies in the CROC simulations, with the color scale showing the number of simulated galaxies at each [SFR, M_{vir}] location. The lower left panel shows the CEBM model results for the distribution of SFR with M_{vir}, where the SFR is computed from the CEBM using the parameters $\boldsymbol{\theta}' = [M_{\mathrm{vir}}, z, \rho_{\mathrm{l}}, T_{\mathrm{l}}, \Upsilon_{0.1}]$. The upper right panel shows the residuals between the simulated CROC galaxy SFRs and the CEBM model results. The lower right panel shows the simulated CROC galaxy SFRs that lie outside the CEBM model predictions. These outliers represent ≈5% of simulated CROC galaxies.

ionizing photon budget or topology of reionization will be minimal.

By editing the data set and retraining, the impact of environment on the performance of the EBM models can be estimated. Relative to $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ and $\gamma(M_\star|\boldsymbol{\theta})$ that use the full data set θ, including all environmental parameters, EBM models trained only on maximum peak circular velocity v_{peak}, halo virial mass M_{vir}, and redshift z has an outlier fraction increased by only ∼1% when modeling $M_\star$ and ∼10% when modeling SFR. Further, removing v_{peak} and training only on $[M_{\mathrm{vir}}, z]$ substantially degrades the model performance, and the outlier fractions increase to ∼20% when modeling $M_\star$ and ∼40% when modeling SFR. The importance of including v_{peak} in the training data set is much larger than the importance of accounting for the environmental measures selected in this analysis.

The EBM models enable an approximate translation of the galaxy formation model from one simulation to another. Provided the parameter sets θ or $\boldsymbol{\theta}'$ can be measured in both simulations, an EBM can recover the connection between SFR, stellar masses, halo properties, and environment from the training simulation and then be used to instill those relations in a different simulation. Since the $\boldsymbol{\theta}'$ parameter set does not require very high spatial resolution to capture, the net results for SFR and stellar mass from a high-resolution simulation accurately tracking detailed baryonic physics can be translated into a simulation with resolution insufficient to capture those physics directly. In future work, we plan to transfer the CROC baryonic galaxy formation model into *Cholla* cosmological simulations (e.g., Villasenor et al. 2021, 2022) via the EBM models presented here. Such a transferred model could be used to build models of feedback from galaxy formation on resolved scales that incorporate the regulatory effects of feedback on small-scale star formation. We expect that such transferred models may have some limitations. For instance, the EBM models are deterministic and while the scatter in the properties predicted by the EBM originates from the scatter of the physical properties of the simulated galaxies, two galaxies with the same physical properties will have identical target quantities predicted by the EBM. These possible limitations of transferred models can be explored in future work.

The ability of the EBM models to recover the SFR and $M_\star$ distributions using only halo and environmental properties allows for the rapid replacement of galaxy formation models based on EBMs. Models can be trained on the simulated galaxy catalogs from a variety of expensive, high-resolution training simulations, including a wide range of physics. These EBM models can then be used interchangeably as effective galaxy formation models in the target simulations, and can also be

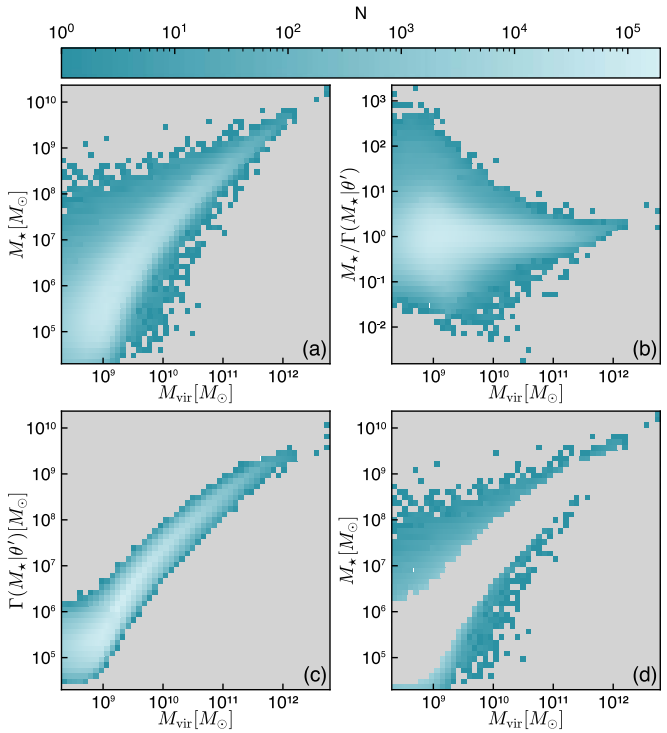


Figure 7. Summary of the CEBM model $\Gamma(M_\star|\boldsymbol{\theta}')$ targeting stellar mass $M_\star$ as a function of virial mass. The upper left panel shows the distribution of $M_\star$ with virial mass $M_{\rm vir}$ in the CROC simulated galaxy catalogs, with the coloration indicating the number of galaxies at each $[M_\star, M_{\rm vir}]$ location. The lower left panel shows the CEBM model prediction of the stellar mass distribution with virial mass given in the input parameters $\boldsymbol{\theta}' = [M_{\rm vir}, z, \rho_1, T_1, \Upsilon_{0.1}]$. The upper right panel shows the residuals between the simulated and predicted $M_\star$ vs. $M_{\rm vir}$ distribution, and the lower right panel shows the outliers in the simulated distribution not captured by the CEBM model $\Gamma(M_\star|\boldsymbol{\theta}')$. The fraction of outliers is $\lesssim$2%.

modified posteriori to allow a broad parameter search or correct the inaccuracies of the training simulation. Such an approach could reduce the sensitivity of conclusions about, e.g., the reionization process on the detailed SFR and $M_\star$ distributions as multiple EBM models for these properties could be trained and implemented in the target simulations.

Lastly, the assumptions of EBMs and EBM-based models enable a higher level of interpretability than other models, but may come at the cost of performance. In particular, when modeling data that are discrepant from the mean trend of the distribution for most or all input features, more flexible models, such as the random forest and its variants (see Section 1 for a summary of recent works) or neural networks, may perform better than EBM-based models but at the cost of interpretability. We suspect that using interpretable models like the EBM in tandem with more flexible models could offer the best of both approaches. We leave exploring the efficacy of ensembles of EBMs and more flexible models for future work.

6. Summary

A complex interplay of physical processes gives rise to the distribution of SFRs and stellar masses $M_\star$ of galaxies over cosmic time. Cosmological simulations provide powerful methods for modeling these physical processes, but the connection between SFR, $M_\star$, and other galaxy properties can be obfuscated by complexity. Leveraging machine-learning techniques, we use a variation of the GAM (Hastie & Tibshirani 1986) called EBM (Nori et al. 2019) to infer the dependence of SFR and $M_\star$ in the CROC simulations (Gnedin 2014) on dark matter halo properties, including virial mass $M_{\rm vir}$, peak maximum circular velocity $v_{\rm peak}$, redshift, environmental density, environmental gas temperature, and the mass of neighboring halos. Our findings include:

1. SFR and $M_\star$ primarily depend on $M_{\rm vir}$ and $v_{\rm peak}$, followed by redshift, environmental density, and environmental gas temperature.
2. When including $M_{\rm vir}$ and $v_{\rm peak}$ in the parameter set used to train the EBM, the model recovers better than 97% of the distribution of $M_\star$ or SFR with virial mass $M_{\rm vir}$ in the CROC simulations.
3. If the model fit excludes $v_{\rm peak}$, the fraction of outliers in the CROC data relative to the predicted model distribution increases to 7.6% for SFR and 2.8% for $M_\star$.
4. To ameliorate the degradation of the model performance when excluding $v_{\rm peak}$, we define a CEBM model comprised of a weighted sum of the base EBM model fit to the main trend of SFR and $M_\star$ with the halo properties and a second EBM model to fit the outliers not represented in the base EBM. The weighting coefficients are themselves determined by an EBM model fit.
5. The CEBM model improves the performance to ≈95%–98% accuracy in the distribution of SFR or $M_\star$ with virial mass, even when excluding $v_{\rm peak}$ measurements from the training data set.

The EBM models quantify the relative importance of halo properties like virial mass and maximum peak circular velocity for determining the stellar mass and SFR of the galaxy it hosts. Through these models, the physics of baryonic galaxy formation can be connected to the properties of dark matter halos and enable galaxy formation to be implemented as a subgrid prescription in dark matter-only simulations or hydrodynamical simulations that do not resolve the small-scale details of star formation and feedback.

This work was supported by the NASA Theoretical and Computational Astrophysics Network (TCAN) grant 80NSSC21K0271. The authors acknowledge use of the lux supercomputer at UC Santa Cruz, funded by NSF MRI grant AST 1828315. This manuscript has been co-authored by Fermi Research Alliance, LLC under contract No. DE-AC02-07CH11359 with the US Department of Energy, Office of Science, Office of High Energy Physics. CROC project relied on resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. An award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. CROC project is also part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the State of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE

Office of Science User Facility supported under contract No. DE-AC05-00OR22725. We have used resources from DOE INCITE award AST 175.

Software: Python (van Rossum 1995), NumPy (van der Walt et al. 2011), scikit-learn (Pedregosa et al. 2011), matplotlib (Hunter 2007), InterpretML (Nori et al. 2019).

Appendix A
Detailed Results for the $M_\star$ EBM

While the performance of the EBM model $\gamma(M_\star|\boldsymbol{\theta})$ targeting $M_\star$ is summarized in Figure 5, a more detailed view of the model is provided by the average contributions provided by each parameter, the univariate feature functions dependent on the parameters, and the bivariate interaction functions. The results of the model are presented below.

A.1. Average Contribution

Figure 8 shows the average contribution of the seven most important features and interactions in the EBM model $\gamma(M_\star|\boldsymbol{\theta})$. In order of decreasing importance, these features include peak circular velocity, virial mass, redshift, environmental density, environmental temperature, the mass ratio of nearby halos, and the interaction between redshift and peak circular velocity. Peak circular velocity is about 50% more important than virial mass, which in turn is roughly a factor of 2 more important than redshift. The other features and interactions contribute to stellar mass at the $\lesssim$0.1 dex level. For reference, the numerical values for the average contributions are provided in Table 7.

Table 7
Summary of the EBM Model $\gamma(M_\star|\boldsymbol{\theta})$ Trained to Predict $M_\star$ as a Function of the Full Parameter Set θ

Average Contributions for the $\gamma(M_\star|\boldsymbol{\theta})$ EBM	
Feature	Value [$\log_{10} M_\odot$]
$\beta_{\log_{10} M_\star}$	5.9629
$\bar{f}\,(\log_{10} v_{\rm peak})$	0.3284
$\bar{f}\,(\log_{10} M_{\rm vir})$	0.2123
$\bar{f}\,(z)$	0.1238
$\bar{f}\,(\log_{10} \rho_1)$	0.0722
$\bar{f}\,(\log_{10} T_1)$	0.0545
$\bar{f}\,(\log_{10} \Upsilon_{0.1})$	0.0359
$\bar{f}\,(z, \log_{10} v_{\rm peak})$	0.0135

Note. The first entry, $\beta_{\log_{10} M_\star}$, is the learned baseline value of the model (see Section 2.1). The next seven entries are the feature functions with the highest average contribution in descending order. The average contribution is calculated using the average of the absolute value of the feature functions weighted by the number of samples in each bin (see Equation (5)).

A.2. Feature Functions

The univariate functions determined by the EBM targeting stellar mass $M_\star$ are shown in Figure 9. Stellar mass increases with increasing peak circular velocity, virial mass, environmental density, and neighboring halo mass ratio. Stellar mass increases with decreasing redshift. As with SFR, the stellar mass increases with increasing environmental temperature T_1, with a sharp enhancement near the temperature where hydrogen becomes neutral and a sharp deficit near where hydrogen ionizes.

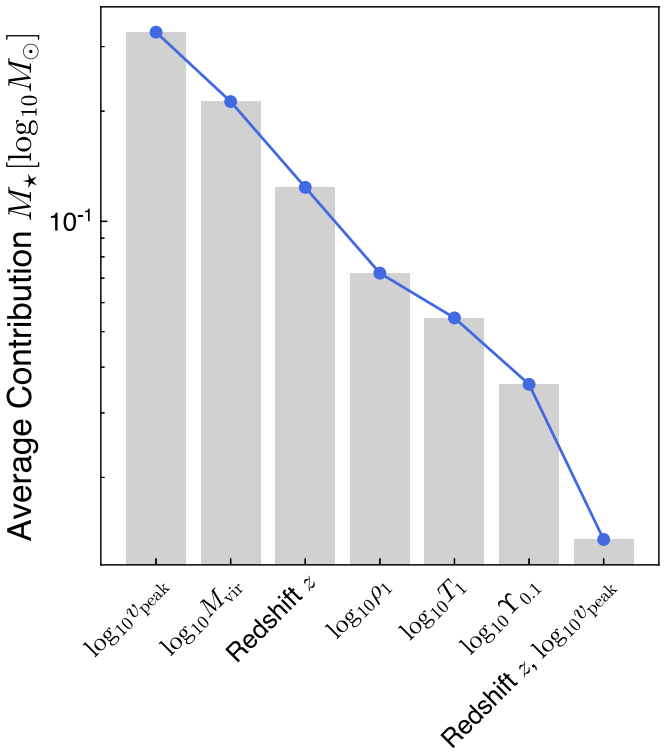


Figure 8. Features with the highest average contribution for the EBM $\gamma(M_\star|\boldsymbol{\theta})$ trained to predict $M_\star$. Average contribution is calculated using the average of the absolute value of the learned functions weighted by the number of samples in each bin (see Equation (5)). The features with the largest contribution are $v_{\rm peak}$ and $M_{\rm vir}$, followed by redshift z, environmental density ρ_1, environmental temperature T_1, and mass ratio of nearby halos $\Upsilon_{0.1}$. The interaction with the largest average contribution involves [z, $v_{\rm peak}$].

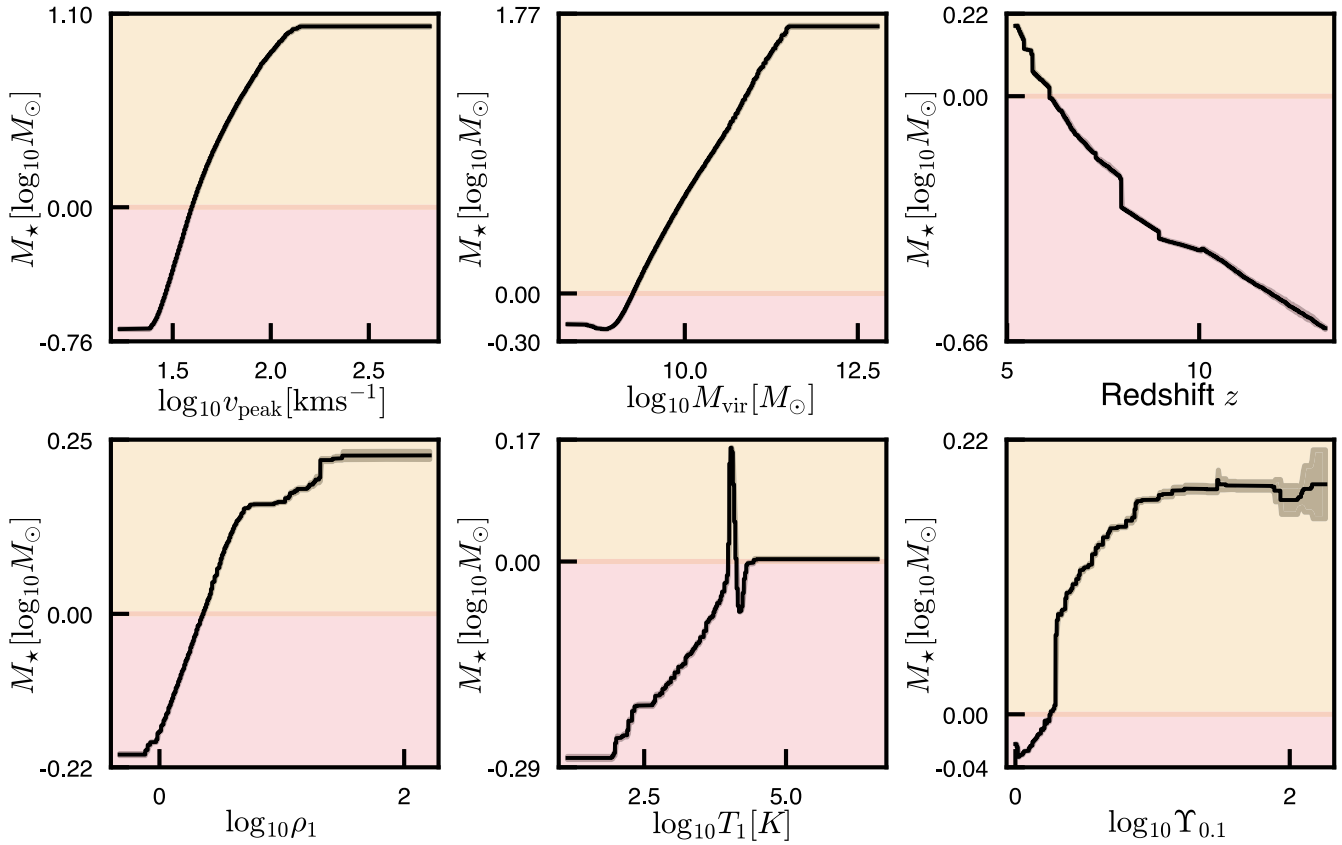


Figure 9. Learned univariate feature functions, f_y^i in Equation (1), for the EBM $\gamma(M_\star|\boldsymbol{\theta})$ trained to predict $M_\star$. Areas highlighted in orange indicate portions of the function that contribute positively to the predicted $M_\star$ and areas in red contribute negatively. Stellar mass increases with peak circular velocity and virial mass, increases with decreasing redshift, and increases with environmental density. Temperature correlates positively with stellar mass, with a strong feature near $T_1 \approx 10^4$ K where hydrogen ionizes. Stellar mass also increases with the mass ratio of neighboring halos.

A.3. Interaction Functions

The bivariate interaction functions f_y^{ij} (see Equation (1)) learned by the EBM when targeting stellar mass $M_\star$ are plotted as heat maps in Figure 10. On average most interaction functions do not contribute significantly to galaxy stellar mass, but there are regions of parameter space where the interaction functions are important. For instance, halos with low environmental temperatures and high environmental densities have suppressed stellar mass. Large virial mass halos with small neighboring halo mass ratios $\log_{10}\Upsilon_{0.1}$, indicating halos that dominate their local environment, have stellar mass enhanced by ≈0.3 dex. This effect exceeds the maximum univariate contribution of $\log_{10}\Upsilon_{0.1}$ alone. The deficit of stellar mass at environmental temperatures where hydrogen is becoming ionized is increased at high redshifts.

Appendix B
CEBM

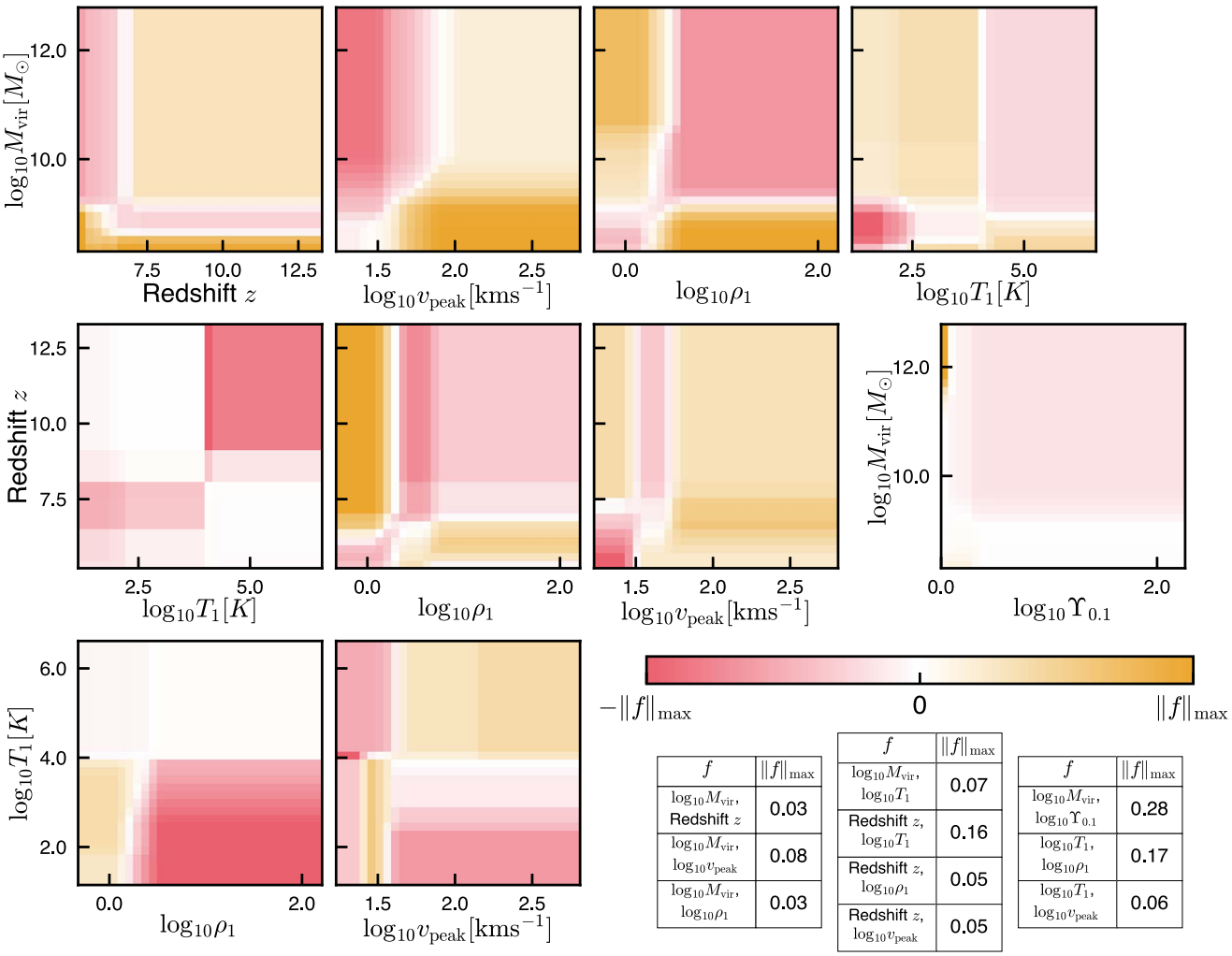
The CEBM models we present consist of a *base* EBM model trained to recover the main trend $\gamma(y|\boldsymbol{\theta}')$ of the targeted property y with the input parameters $\boldsymbol{\theta}'$, an *outlier* EBM model that captures the outlying values of y not recovered by $\gamma(y|\boldsymbol{\theta}')$, and a classification EBM model $\phi_y(\boldsymbol{\theta}')$ that interpolates between them (see Section 4). Given a CEBM, we wish to construct analogs of the average contribution, feature functions, and interaction functions determined for a single EBM. We define these quantities for the CEBM function in Sections B.2 and B.3 below.

B.1. CEBM Feature and Interaction Functions

The feature functions of a single EBM are univariate and indicate directly how the expectation value of the targeted quantity depends on each parameter $\theta_i \in \boldsymbol{\theta}$. With a CEBM comprised of a weighted sum of two base EBMs, we define the analog of the feature function to be the weighted sum of the base EBM feature functions. We can write that

$$\tilde{f}_y^i = \frac{1}{N}\sum_{j=0}^{N} \|\phi(\boldsymbol{\theta}_j) \odot \boldsymbol{f}_y^i(\boldsymbol{\theta}_j)\|_1, \tag{B1}$$

where $\odot$ is the Hadamard or element-wise product operation and the sum is over the number of samples N. The quantity $\boldsymbol{f}_y^i$ is the



f	$\|f\|_{\rm max}$
$\log_{10}M_{\rm vir}$, Redshift z	0.03
$\log_{10}M_{\rm vir}$, $\log_{10}v_{\rm peak}$	0.08
$\log_{10}M_{\rm vir}$, $\log_{10}\rho_1$	0.03

f	$\|f\|_{\rm max}$
$\log_{10}M_{\rm vir}$, $\log_{10}T_1$	0.07
Redshift z, $\log_{10}T_1$	0.16
Redshift z, $\log_{10}\rho_1$	0.05
Redshift z, $\log_{10}v_{\rm peak}$	0.05

f	$\|f\|_{\rm max}$
$\log_{10}M_{\rm vir}$, $\log_{10}\Upsilon_{0.1}$	0.28
$\log_{10}T_1$, $\log_{10}\rho_1$	0.17
$\log_{10}T_1$, $\log_{10}v_{\rm peak}$	0.06

Figure 10. Learned bivariate interaction functions f_y^{ij} for the EBM $\gamma(M_\star|\boldsymbol{\theta})$ trained to predict $M_\star$. Areas highlighted in orange indicate portions of the functions that contribute positively to the predicted $M_\star$ while areas in red contribute negatively. Halos with large environmental temperatures T_1 at high redshift z show enhanced stellar mass. The stellar masses of halos with low environmental temperature $T_1 < 10^4$ K correlate with environmental density, increasing with increasing ρ_1. Massive halos with no comparable large neighboring halos ($\Upsilon_{0.1} \approx 0$) also show enhanced stellar mass.

vector of the individual EBM feature functions f_y^i. While the base EBM feature functions are individually univariate, by weighting the sum of these feature functions with the classifier EBM the resulting feature function analog in Equation (B1) is *not* univariate.

The interaction functions $\tilde{f}_y^{ij}$ are defined as in Equation (B1) but with the vector of the individual EBM interaction functions $\boldsymbol{f}_y^{ij}$ substituted for $\boldsymbol{f}_y^i$. While the interaction functions for a single EBM are bivariate, the CEBM interaction functions are *not* bivariate.

B.2. CEBM Average Contribution

The average contribution of each feature in a CEBM can be defined in a manner analogous to the average contribution computed for a single EBM (Equation (5)). The CEBM average contribution can be written as

$$\bar{f}_y^i = \frac{\sum_{j=0}^{n_b-1} \tilde{f}\,(\theta_{i,j}) N_j}{\sum_{j=0}^{n_b-1} N_j}, \tag{B2}$$

where $\tilde{f}$ is either the CEBM feature function $\tilde{f}_y^i$ or the CEBM interaction function $\tilde{f}_y^{ij}$. Equation (B2) characterizes how important the parameter θ_i is for modeling the target quantity y.

B.3. Visualizing CEBM Feature and Interaction Functions

The feature and interaction functions $\tilde{f}_y^i$ and $\tilde{f}_y^{ij}$ are not univariate or bivariate by design, which allows them to model

the outlier distribution about the base EBM model $\gamma(y|\boldsymbol{\theta}')$. To visualize the feature and interaction functions for CEBM models in a manner similar to the univariate feature and bivariate interaction functions for single EBM, we can average the values of $\boldsymbol{f}_y^i$ and $\boldsymbol{f}_y^{ij}$. For the feature function averaged over N samples, consider n_b bins along the θ_i direction, with central values $\theta_{i,b}$ and bin widths $\Delta\theta_{i,b}$. The bin-averaged CEBM feature and interaction functions are then

$$f_y^{i,b} = \frac{1}{N}\sum_{j=0}^{N-1} \alpha(\theta_{i,b}, \Delta\theta_{i,b}, \theta_{j,i})\phi(\boldsymbol{\theta}_j) \odot \boldsymbol{f}(\boldsymbol{\theta}_j), \tag{B3}$$

where $\theta_{j,i}$ is the ith parameter of the jth sample θ_j and the function $\alpha(\theta_{i,b}, \Delta\theta_{i,b}, \theta_{j,i}) = 1$ if $\theta_{i,b} - \Delta\theta_{i,b}/2 \leqslant \theta_{j,i} \leqslant \theta_{i,b} + \Delta\theta_{i,b}/2$, and $\alpha = 0$ otherwise. The quantity $\boldsymbol{f}$ is either the vector of EBM feature functions $\boldsymbol{f}_y^i$ or the EBM interaction functions $\boldsymbol{f}_y^{ij}$. Equation (B3) calculates the mean of the $\boldsymbol{f}$ values in each of the n_b bins, and can be modified to calculate its standard deviation.

Appendix C
Composite EBM Model for SFR

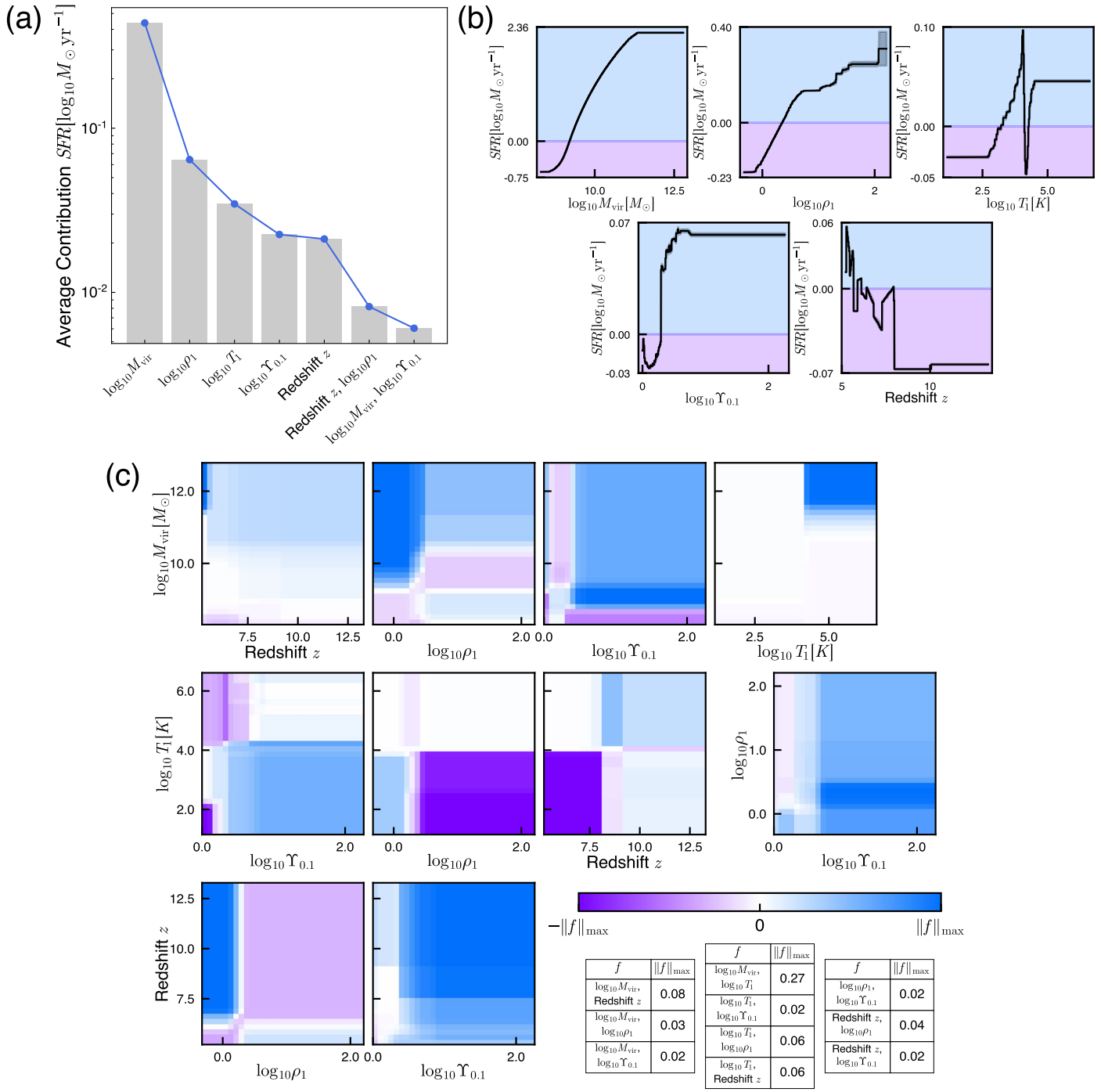
The CEBM model $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ for the SFR consists of a base EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$, a residual EBM $\delta(\mathrm{SFR}|\boldsymbol{\theta}')$ that attempts to capture the outlying values of SFR not recovered by $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$, and the classifier EBM $\phi_{\mathrm{SFR}}(\boldsymbol{\theta}')$. For each of these individual EBMs that form the CEBM model, we plot the average contribution, feature functions, and interaction functions.

Figure 11 shows the average contribution, feature functions, and interaction functions for the EBM model $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ that forms the base of the CEBM model. The differences between $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ and $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ reflect the additional information provided by the maximum peak circular velocity v_{peak}. Without access to v_{peak}, the base EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ upweights $\bar{f}(M_{\mathrm{vir}})$ such that its importance roughly equals the combined importance of M_{vir} and v_{peak} in determining $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$. The average contribution of ρ_1, T_1, z, $\Upsilon_{0.1}$, and $(M_{\mathrm{vir}}, \Upsilon_{0.1})$ is similar between the models. The additional interaction term in the top seven average contributions is (z, ρ_1), with a percent-level contribution to SFR relative to M_{vir}. The feature functions for $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ have shapes similar to the feature functions for $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$, but their minimum and maximum contributions to SFR are adjusted to account for the missing v_{peak} contribution. The feature function $\bar{f}(z)$ is noisier overall. For the interaction functions, the largest contributors now involve M_{vir} rather than the missing parameter v_{peak}, and the set of available functions is substantially different than with $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$.

Figure 12 shows the average contribution, feature functions, and interaction functions for the outlier EBM $\delta(\mathrm{SFR}|\boldsymbol{\theta}')$ fit to the deviant samples not captured by the base EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$. The outlier EBM receives the highest contribution from virial mass, with an average contribution more than an order of magnitude larger than the next most important feature ρ_1. The redshift z and environmental temperature T_1 have comparable importance to ρ_1. The remaining features provide only percent-level contributions relative to M_{vir}.

Figure 13 shows the average contribution, feature functions, and interaction functions for the classifier EBM $\delta(\mathrm{SFR}|\boldsymbol{\theta}')$ that interpolates between the base and outlier EBMs when calculating the CEBM model. For the classifier EBM, the most important features are ρ_1, M_{vir}, and $\Upsilon_{0.1}$. Redshift z has middling importance, followed by T_1, $[\rho_1, \Upsilon_{0.1}]$, and $[M_{\mathrm{vir}}, \rho_1]$. The feature functions show strong dependencies on ρ_1, M_{vir}, $\Upsilon_{0.1}$, z, and T_1. The largest interaction functions involve the environmental temperature T_1, redshift z, and virial mass M_{vir}.

By weighting the base and outlier EBM models with the classifier EBM, we construct the CEBM for SFR as $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}') \equiv [1 - \phi_{\mathrm{SFR}}(\boldsymbol{\theta}')]\gamma(\mathrm{SFR}|\boldsymbol{\theta}') + \phi_{\mathrm{SFR}}(\boldsymbol{\theta}')\delta(\mathrm{SFR}|\boldsymbol{\theta}')$. Figure 14 shows the average contribution, feature functions, and interaction functions for the SFR CEBM. The most important feature is M_{vir}, which dominates by a factor of $\sim$4–10 over environmental density ρ_1, environmental temperature T_1, $\Upsilon_{0.1}$, and redshift z. The interaction terms are roughly percent-level effects relative to M_{vir}. The feature functions show a strongly increasing SFR with M_{vir}, and enhanced SFR with environmental density ρ_1. The temperature dependence shows the feature at $\log_{10} T_1 \approx 4$ seen with the EBM model $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$. The interaction functions provide only important contributions over very limited areas of parameter space, with the most important adjustments occurring at low redshift and large virial mass, or for large temperatures and virial masses. For reference, the model summary Figure 6 illustrates the overall performance of the model.

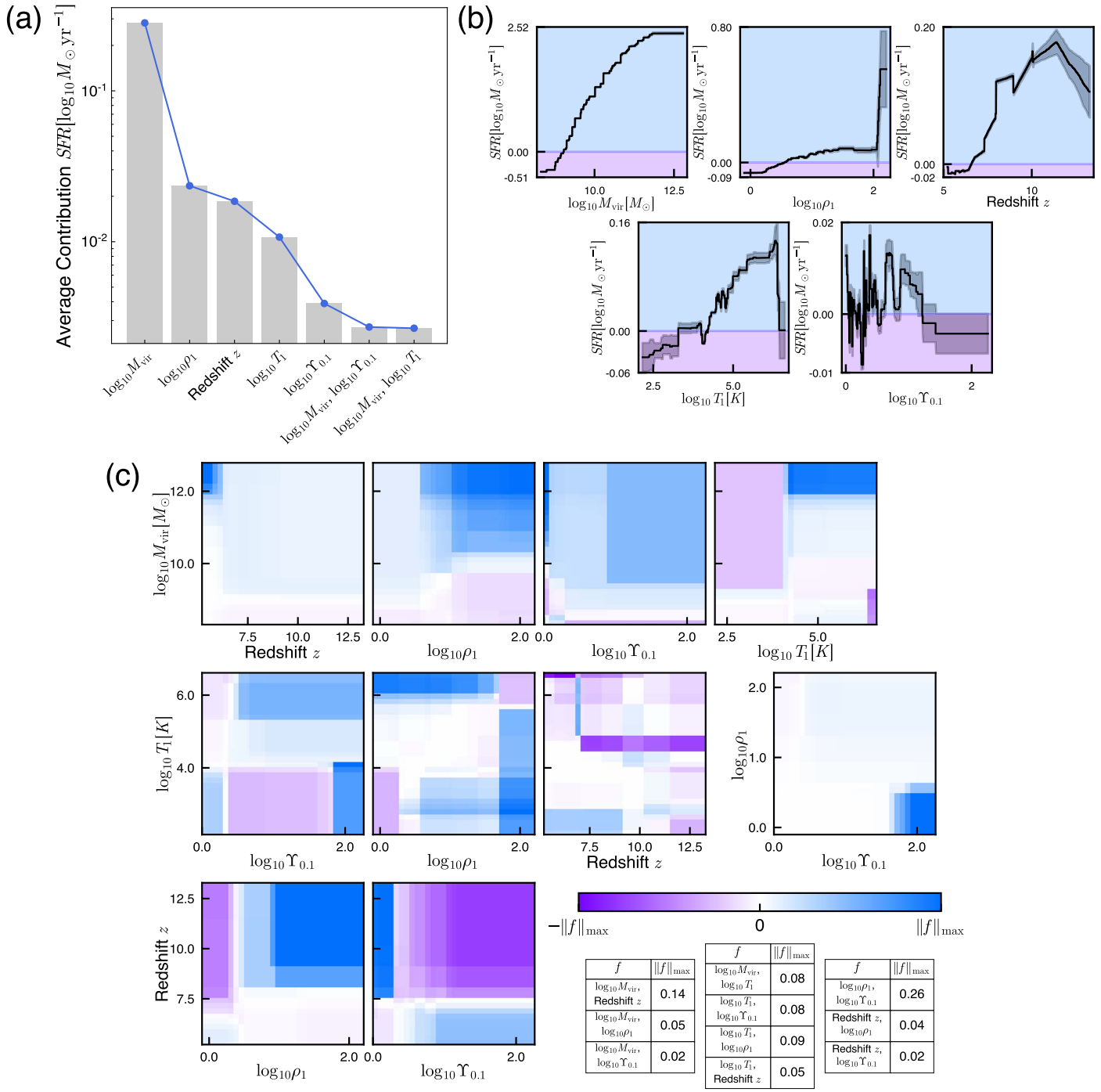


f	$\|f\|_{\max}$
$\log_{10} M_{\rm vir}$, Redshift z	0.08
$\log_{10} M_{\rm vir}$, $\log_{10}\rho_1$	0.03
$\log_{10} M_{\rm vir}$, $\log_{10}\Upsilon_{0.1}$	0.02

f	$\|f\|_{\max}$
$\log_{10} M_{\rm vir}$, $\log_{10} T_1$	0.27
$\log_{10} T_1$, $\log_{10}\Upsilon_{0.1}$	0.02
$\log_{10} T_1$, $\log_{10}\rho_1$	0.06
$\log_{10} T_1$, Redshift z	0.06

f	$\|f\|_{\max}$
$\log_{10}\rho_1$, $\log_{10}\Upsilon_{0.1}$	0.02
Redshift z, $\log_{10}\rho_1$	0.04
Redshift z, $\log_{10}\Upsilon_{0.1}$	0.02

Figure 11. Details for the base EBM model $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ component of the CEBM $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ trained to predict SFR. Panel (a) displays the average contribution of features. The dominant feature is virial mass $M_{\rm vir}$, with an average contribution to $\log_{10}$ SFR roughly 8–10 × larger than environmental density ρ_1 and temperature T_1. Compared with the average contributions to the EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta})$ (see Figure 1), $M_{\rm vir}$ subsumes the contribution provided by the missing $v_{\rm peak}$ parameter. Panel (b) shows the feature functions contributing to the base EBM model. The SFR increases with $M_{\rm vir}$, which provides the dominant contribution. A secondary contribution comes from environmental density ρ_1. Environmental temperature T_1 has a minor contribution, but shows the same feature at $T_1 \approx 10^4$ K where hydrogen ionizes. The mass ratio of nearby halos $\Upsilon_{0.1}$ and redshift z provide minor contributions. Panel (c) presents the interaction functions for the base EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$. Each panel shows the contribution of the bivariate interaction terms, normalized such that the color map ranges between plus or minus the maximum of the norm of each function $\|f\|_{\max}$. Purple indicates negative contributions and blue indicates positive contributions. The table lists $\|f\|_{\max}$ for the interaction functions, each with units $\log_{10} M_\odot\,\mathrm{yr}^{-1}$. In absolute terms, the largest interaction occurs for large virial mass $M_{\rm vir}$ and environmental temperature T_1. SFR is partially reduced for low environmental temperature T_1 and either low environmental density ρ_1 or redshift z.

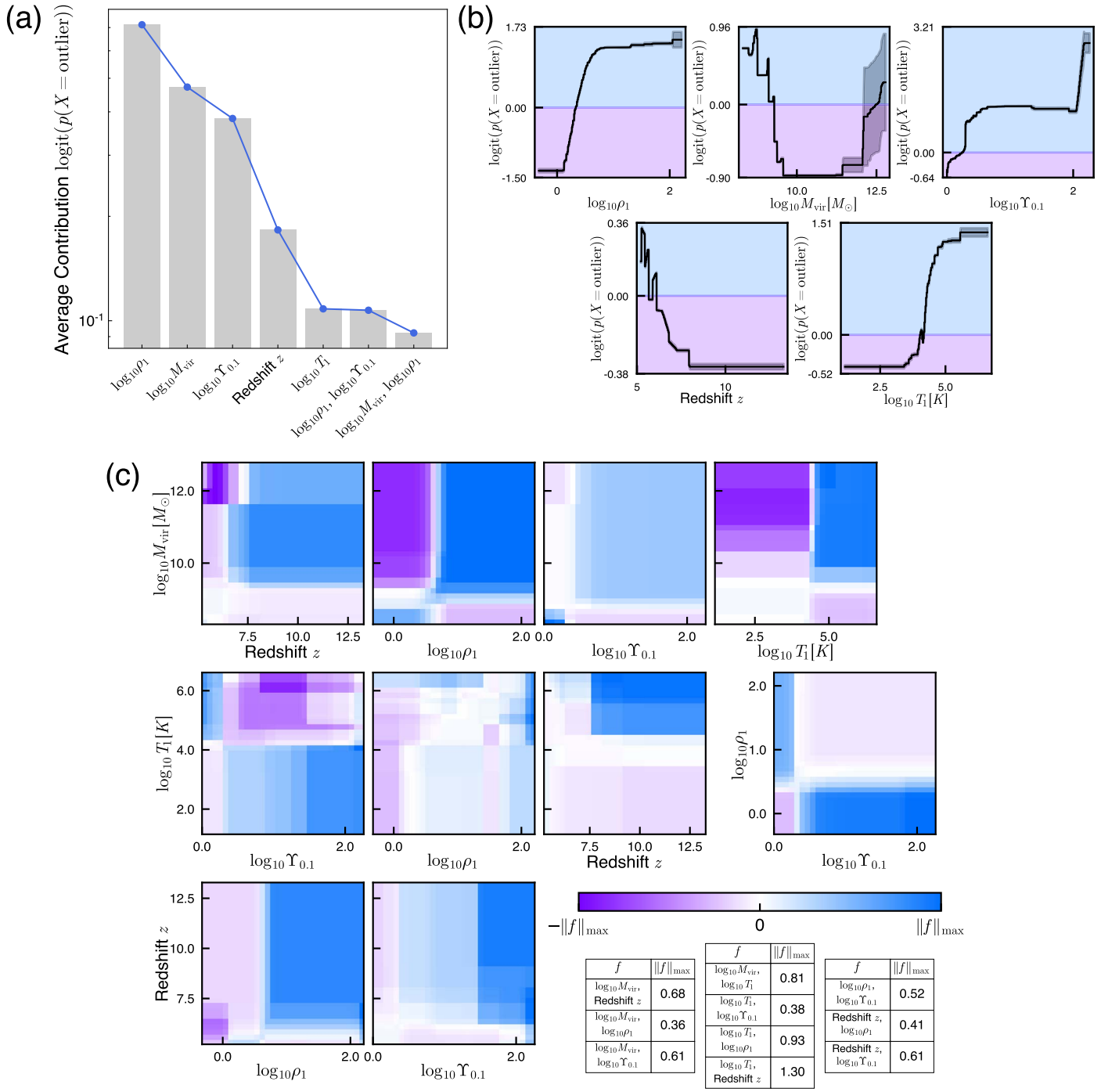


f	$\|f\|_{\max}$
$\log_{10} M_{\rm vir}$, Redshift z	0.14
$\log_{10} M_{\rm vir}$, $\log_{10}\rho_1$	0.05
$\log_{10} M_{\rm vir}$, $\log_{10}\Upsilon_{0.1}$	0.02

f	$\|f\|_{\max}$
$\log_{10} M_{\rm vir}$, $\log_{10} T_1$	0.08
$\log_{10} T_1$, $\log_{10}\Upsilon_{0.1}$	0.08
$\log_{10} T_1$, $\log_{10}\rho_1$	0.09
$\log_{10} T_1$, Redshift z	0.05

f	$\|f\|_{\max}$
$\log_{10}\rho_1$, $\log_{10}\Upsilon_{0.1}$	0.26
Redshift z, $\log_{10}\rho_1$	0.04
Redshift z, $\log_{10}\Upsilon_{0.1}$	0.02

Figure 12. Details for the outlier EBM model $\delta(\mathrm{SFR}|\boldsymbol{\theta}')$ component of the CEBM $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ trained to predict SFR. Panel (a) displays the average contribution of features. As with the base EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$, the feature with the largest average contribution is virial mass $M_{\rm vir}$, with roughly $\gtrsim$10 $\times$ larger contribution to $\log_{10}$SFR than environmental density ρ_1, redshift z, or temperature T_1. The average contributions of $\Upsilon_{0.1}$ or interactions are small. Panel (b) shows the feature functions for the outlier EBM $\delta(\mathrm{SFR}|\boldsymbol{\theta}')$. The feature function for virial mass $M_{\rm vir}$ has the largest contribution to $\delta(\mathrm{SFR}|\boldsymbol{\theta}')$, similar to the virial mass dependence of the base EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$(see panel (b) of Figure 11). The SFR of outliers increases with increasing environmental density ρ_1, with a large enhancement at very large ρ_1. Unlike the base EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$, SFR for the outliers increases with increasing redshift. The feature function for the nearby halo mass ratio $\Upsilon_{0.1}$ is weak and noisy. Panel (c) presents the interaction functions for the outlier EBM $\delta(\mathrm{SFR}|\boldsymbol{\theta}')$. Each panel shows the contribution of the bivariate interaction terms, normalized such that the color map ranges between plus or minus the maximum of the norm of each function $\|f\|_{\max}$. Purple indicates negative contributions and blue indicates positive contributions. The table lists $\|f\|_{\max}$ for the interaction functions, each with units $\log_{10} M_\odot\ \mathrm{yr}^{-1}$. For outliers, the SFR increases at low environmental density ρ_1 and large neighboring halo mass ratios $\Upsilon_{0.1}$, suggesting dynamical interactions increase SFR in low-density environments. The other interaction functions are relatively weak.

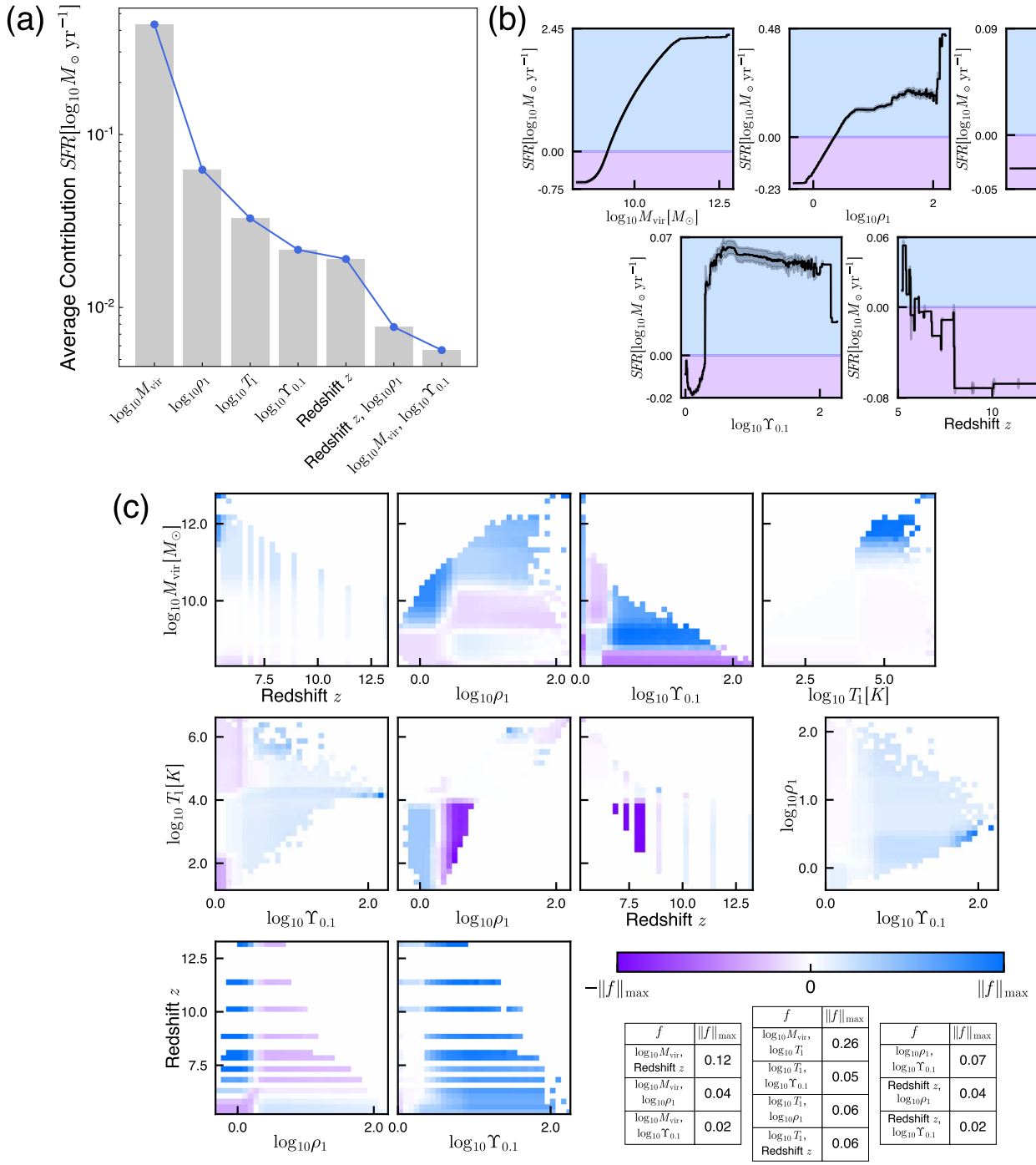


f	$\|f\|_{\max}$
$\log_{10}M_{\rm vir}$, Redshift z	0.68
$\log_{10}M_{\rm vir}$, $\log_{10}\rho_1$	0.36
$\log_{10}M_{\rm vir}$, $\log_{10}\Upsilon_{0.1}$	0.61

f	$\|f\|_{\max}$
$\log_{10}M_{\rm vir}$, $\log_{10}T_1$	0.81
$\log_{10}T_1$, $\log_{10}\Upsilon_{0.1}$	0.38
$\log_{10}T_1$, $\log_{10}\rho_1$	0.93
$\log_{10}T_1$, Redshift z	1.30

f	$\|f\|_{\max}$
$\log_{10}\rho_1$, $\log_{10}\Upsilon_{0.1}$	0.52
Redshift z, $\log_{10}\rho_1$	0.41
Redshift z, $\log_{10}\Upsilon_{0.1}$	0.61

Figure 13. Details for the classification EBM model $\phi_{\rm SFR}(\boldsymbol{\theta}')$ that interpolates between the base EBM $\gamma(\text{SFR}|\boldsymbol{\theta}')$ and the outlier EBM $\delta(\text{SFR}|\boldsymbol{\theta}')$ for creating the CEBM $\Gamma(\text{SFR}|\boldsymbol{\theta}')$. Panel (a) displays the average contribution of features to the classification EBM model $\phi_{\rm SFR}(\boldsymbol{\theta}')$. The most important features for determining whether a galaxy is an outlier in the SFR distribution are environmental density ρ_1, virial mass $M_{\rm vir}$, and nearby halo mass ratio $\Upsilon_{0.1}$. The average contributions are unit-free, and represent changes to the log odds of a galaxy being an outlier in the SFR distribution. Panel (b) shows the feature functions contributing to the classifier EBM $\phi_{\rm SFR}(\boldsymbol{\theta}')$. These feature functions represent the change in log odds that a given galaxy will be an outlier in SFR. Outliers tend to occur at high environmental density ρ_1 or very low or high virial masses $M_{\rm vir}$. Galaxies with massive neighbors, reflected by $\Upsilon_{0.1}$, or high environmental temperature T_1, are also more likely to be outliers. The lowest redshift galaxies in the data set are additionally likely be outliers in SFR. Panel (c) presents the interaction functions for the classifier EBM $\phi_{\rm SFR}(\boldsymbol{\theta}')$. Each panel shows the contributions of the interaction terms, normalized such that the color map ranges between plus or minus the maximum of the norm of each function $\|f\|_{\max}$. Purple indicates negative log odds and blue indicates positive log odds that a given galaxy is an outlier in SFR. The table lists $\|f\|_{\max}$ for the interaction functions, listed as the corresponding change in log odds. Galaxies with large environmental temperature T_1 and with high environmental density ρ_1, at high redshift z, or with large virial mass $M_{\rm vir}$ are more likely to be outliers. Massive galaxies in high environmental densities or at high redshift also are more likely outliers in SFR.



f	$\|f\|_{\max}$
$\log_{10} M_{\rm vir}$, Redshift z	0.12
$\log_{10} M_{\rm vir}$, $\log_{10}\rho_1$	0.04
$\log_{10} M_{\rm vir}$, $\log_{10}\Upsilon_{0.1}$	0.02

f	$\|f\|_{\max}$
$\log_{10} M_{\rm vir}$, $\log_{10} T_1$	0.26
$\log_{10} T_1$, $\log_{10}\Upsilon_{0.1}$	0.05
$\log_{10} T_1$, $\log_{10}\rho_1$	0.06
$\log_{10} T_1$, Redshift z	0.06

f	$\|f\|_{\max}$
$\log_{10}\rho_1$, $\log_{10}\Upsilon_{0.1}$	0.07
Redshift z, $\log_{10}\rho_1$	0.04
Redshift z, $\log_{10}\Upsilon_{0.1}$	0.02

Figure 14. Details for the CEBM model $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ trained to predict SFR. Panel (a) displays the average contribution of features to the CEBM. Virial mass $M_{\rm vir}$ provides the largest average contribution to the SFR. The environmental density ρ_1 provides a $\sim$6 $\times$ smaller average contribution. The environmental temperature T_1, nearby galaxy-mass ratio $\Upsilon_{0.1}$, and redshift z provide a relative contribution roughly 10 $\times$ smaller than $M_{\rm vir}$. Panel (b) shows the feature functions contributing to the CEBM $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$. The SFR increases with $M_{\rm vir}$, which provides the largest contribution. A secondary contribution comes from environmental density ρ_1. Environmental temperature T_1 has a minor contribution, but shows the familiar feature at $T_1 \approx 10^4$ K where hydrogen ionizes. The mass ratio of nearby halos $\Upsilon_{0.1}$ and redshift z provide minor contributions. As expected, the CEBM feature functions are similar to the base EBM feature functions that represent the parameter dependence of SFR for most galaxies in the data set (see panel (b) of Figure 11). Panel (c) presents the interaction functions for the CEBM $\Gamma(\mathrm{SFR}|\boldsymbol{\theta}')$. Each panel shows the contribution of the interaction terms, normalized such that the color map ranges between plus or minus the maximum of the norm of each function $\|f\|_{\max}$. Purple indicates negative contributions and blue indicates positive contributions. The table lists $\|f\|_{\max}$ for the interaction functions, each with units $\log_{10} M_\odot\ \mathrm{yr}^{-1}$. As for the interaction functions for the base EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$, the largest interaction occurs for large virial mass $M_{\rm vir}$ and large environmental temperature T_1 or low redshift z.

Appendix D
CEBM Model for Stellar Mass

The CEBM model $\Gamma(M_\star|\boldsymbol{\theta}')$ for stellar mass is comprised of a base EBM $\gamma(M_\star|\boldsymbol{\theta}')$, an outlier EBM that attempts to model the $M_\star$ of samples not recovered by $\gamma(M_\star|\boldsymbol{\theta}')$, and the classifier EBM function $\phi_{M_\star}(\boldsymbol{\theta}')$ that interpolates between them. The average contribution, feature functions, and interaction functions from these component EBM models are presented below.

Figures 15 shows the average contribution, feature functions, and interaction functions for the base EBM model $\gamma(M_\star|\boldsymbol{\theta}')$. By removing $v_{\rm peak}$ from the data set used to train the EBM, the base EBM model for the $M_\star$ CEBM replaces the dependence on $v_{\rm peak}$ with an additional dependence on $M_{\rm vir}$. The relative ordering and importance of redshift z, environmental density ρ_1, environmental temperature T_1, and $\Upsilon_{0.1}$ are approximately maintained. For the feature functions, the results shown for $\gamma(M_\star|\boldsymbol{\theta}')$ in panel (b) of Figure 15 can be compared with the results for $\gamma(M_\star|\boldsymbol{\theta})$ shown in Figure 8. As reflected by average contributions, the amplitude of the feature function $\tilde{f}(M_{\rm vir})$ increases to account for the removal of $v_{\rm peak}$. The feature functions for z, ρ_1, T_1, and $\Upsilon_{0.1}$ are modified and remain similar in shape to those computed for the EBM $\gamma(M_\star|\boldsymbol{\theta})$. The interaction functions shared between $\gamma(M_\star|\boldsymbol{\theta}')$ and $\gamma(M_\star|\boldsymbol{\theta})$ are similar. There is an increase in $M_\star$ contribution for large [$M_{\rm vir}$, T_1] and a decrease in the amplitude of [$M_{\rm vir}$, $\Upsilon_{0.1}$].

Figure 16 shows the average contribution, feature functions, and interaction functions for the outlier EBM model $\delta(M_\star|\boldsymbol{\theta}')$. The average contribution is dominated by $M_{\rm vir}$, with the contributions from all other single parameters lower by a factor of $\approx$10 with the order of importance maintained relative to $\gamma(M_\star|\boldsymbol{\theta}')$. For the feature functions, the redshift dependence changes dramatically and now increases with increasing redshift. The feature function for environmental density $\tilde{f}(\rho_1)$ becomes much weaker over a wide range of ρ_1, but increases dramatically at high ρ_1. Relative to the $\gamma(M_\star|\boldsymbol{\theta}')$ feature functions, the feature function $\tilde{f}(\Upsilon_{0.1})$ for $\delta(M_\star|\boldsymbol{\theta}')$ is weak and noisy. The interaction functions show increased contributions at large [z, ρ_1], and for low T_1 and large ρ_1.

Figure 17 shows the average contribution, feature functions, and interaction functions for the classifier EBM $\phi_{M_\star}(\boldsymbol{\theta}')$. For each of these properties, we note that in determining $\phi_{M_\star}(\boldsymbol{\theta}')$ a sigmoid function σ is applied to the sum of β, f_y^i, and f_y^{ij} that model the log odds that a galaxy is an outlier in stellar mass. In determining $M_\star$, the features with the largest average contribution are environmental density ρ_1, redshift z, $\Upsilon_{0.1}$, and virial mass $M_{\rm vir}$. The interaction terms with the largest contribution are (z, ρ_1) and (ρ_1, T_1). Clearly, environmental density plays an important role in determining whether a given simulated galaxy is an outlier relative to the base EBM $\gamma(M_\star|\boldsymbol{\theta}')$. The feature functions show that galaxies with large environmental densities ρ_1, at low redshift z, or with a large neighboring galaxy (expressed by $\Upsilon_{0.1}$) have an enhanced probability of being outliers relative to $\gamma(M_\star|\boldsymbol{\theta}')$. Galaxies at both high and low $M_{\rm vir}$ or large environmental temperature T_1 are also more likely to be outliers.

We construct the stellar mass CEBM with the sum $\Gamma(M_\star|\boldsymbol{\theta}') \equiv [1 - \phi_{M_\star}(\boldsymbol{\theta}')]\gamma(M_\star|\boldsymbol{\theta}') + \phi_{M_\star}(\boldsymbol{\theta}')\delta(M_\star|\boldsymbol{\theta}')$. Figure 18 shows the average contribution, feature functions, and interaction functions for $\Gamma(M_\star|\boldsymbol{\theta}')$. The feature with the largest average contribution is $M_{\rm vir}$, with redshift z, environmental density ρ_1, environmental temperature T_1, and the mass ratio of nearby galaxies $\Upsilon_{0.1}$ having a lower average contribution by a factor of $\sim$5–10. Relative to $M_{\rm vir}$, the interactions [z, ρ_1] and [$M_{\rm vir}$, ρ_1] contribute at level of a few percent. The $M_\star$ CEBM feature function $\tilde{f}(M_{\rm vir})$ has increased in amplitude relative to the $M_\star$ EBM feature function $f(M_{\rm vir})$, subsuming some of the dependence on the missing $v_{\rm peak}$ feature. The remaining feature functions for $\Gamma(M_\star|\boldsymbol{\theta}')$ are similar in shape and amplitude to those for $\gamma(M_\star|\boldsymbol{\theta})$, although the contribution at large T_1 and ρ_1 is increased and the dependence on redshift z is decreased. The interaction functions are similar between $\Gamma(M_\star|\boldsymbol{\theta}')$ and $\gamma(M_\star|\boldsymbol{\theta}')$, although there is a larger enhancement of $M_\star$ for large [$M_{\rm vir}$, T_1] and a smaller enhancement for large $M_{\rm vir}$ and small $\Upsilon_{0.1}$ for the CEBM $\Gamma(M_\star|\boldsymbol{\theta}')$. For reference, the model summary Figure 7 illustrates the overall performance of the model.

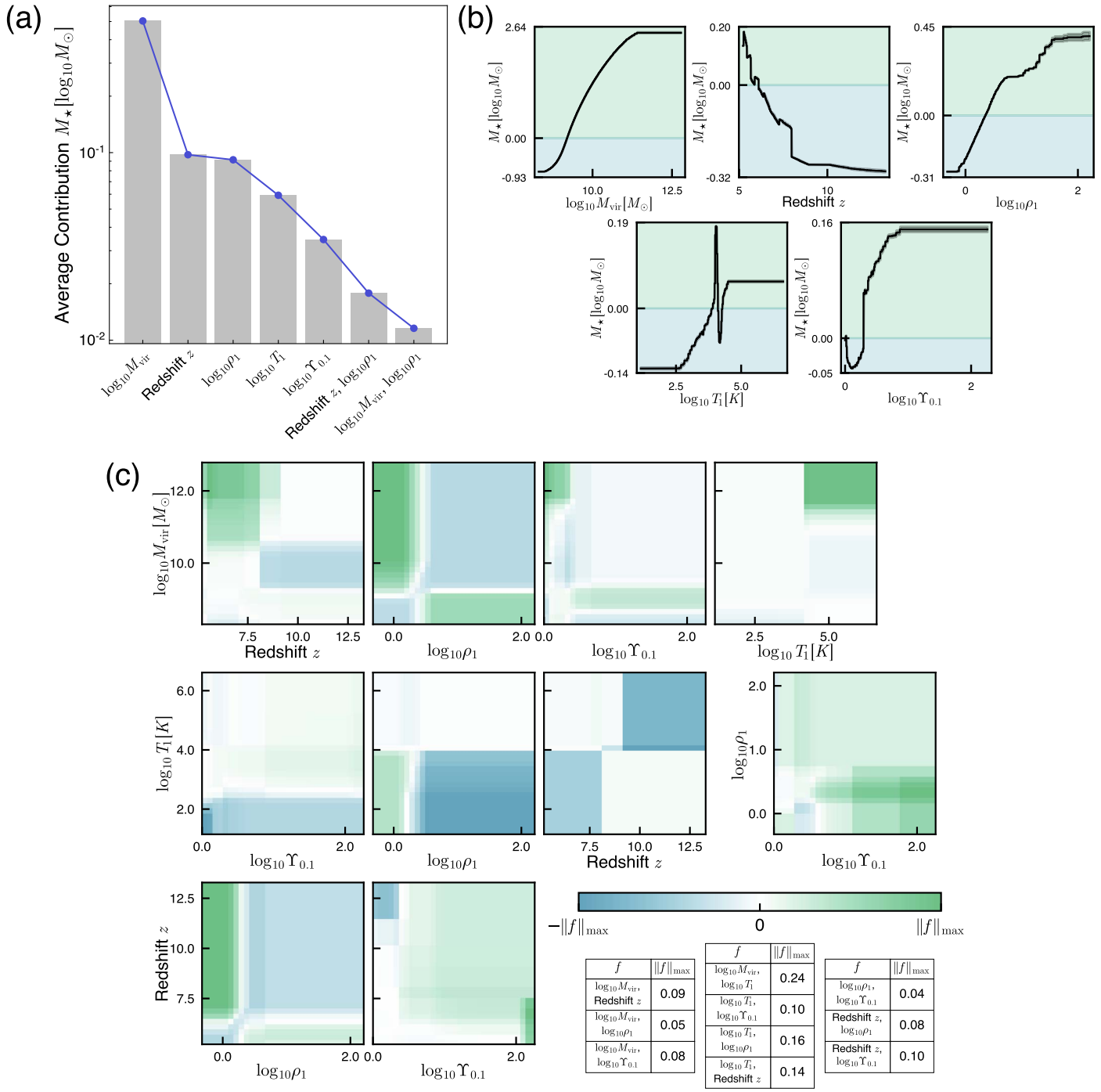
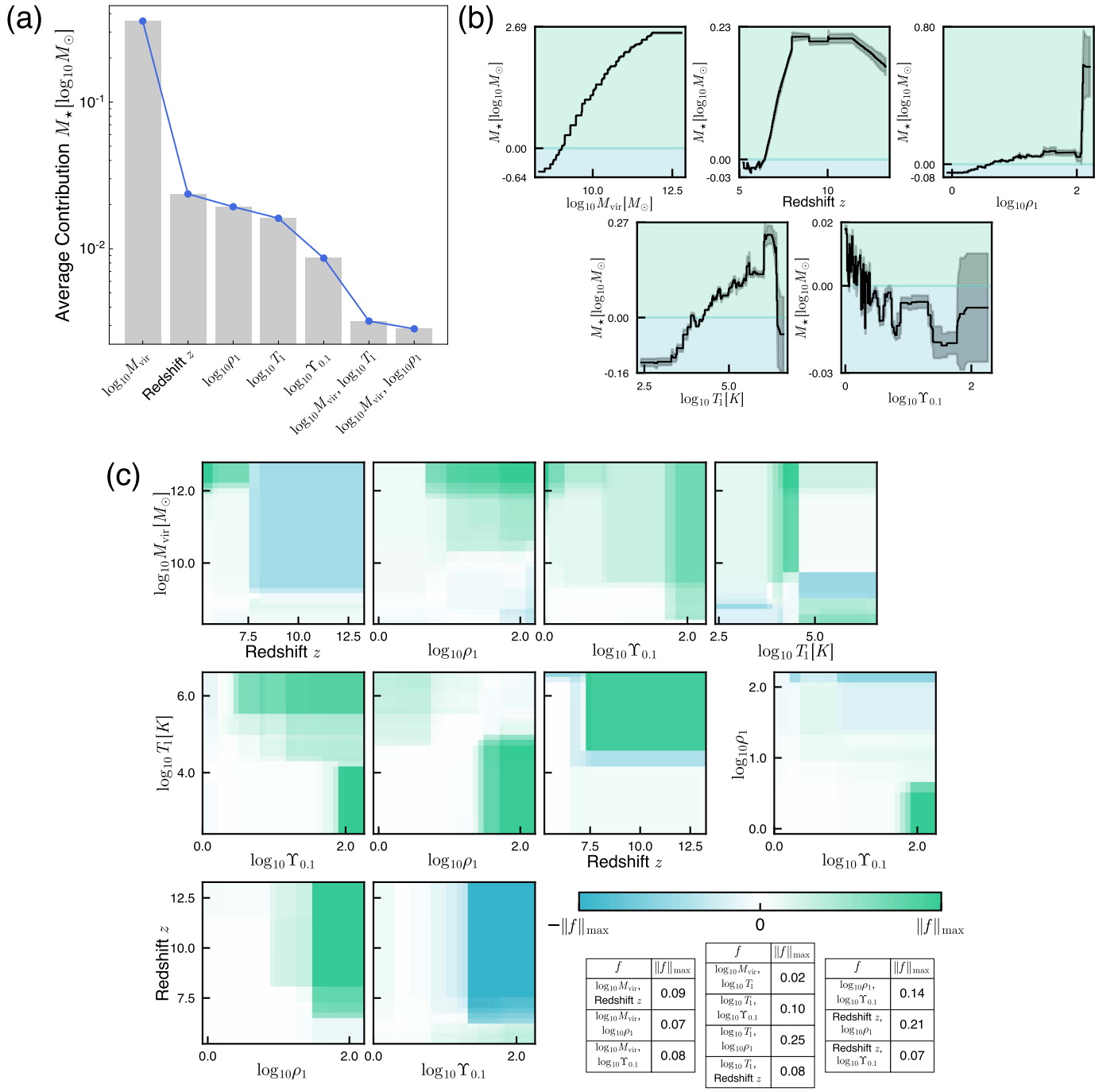


Figure 15. Details for the base EBM model $\gamma(M_\star|\boldsymbol{\theta}')$ component of the CEBM $\Gamma(M_\star|\boldsymbol{\theta}')$ trained to predict stellar mass $M_\star$. Panel (a) displays the average contribution of features to the base EBM model $\gamma(M_\star|\boldsymbol{\theta}')$. The feature with the highest average contribution is virial mass $M_{\rm vir}$, with an average contribution to $\log_{10} M_\star$ roughly $5\times$ larger than that from redshift z or environmental density ρ_1. The environmental temperature T_1 and nearby halo mass ratio $\Upsilon_{0.1}$ provide smaller average contributions, and interactions between features are yet smaller. Panel (b) shows the feature functions contributing to the base EBM model $\gamma(M_\star|\boldsymbol{\theta}')$. The stellar mass increases with $M_{\rm vir}$, which provides the largest contribution. Secondary contributions come from redshift z, which increases $M_\star$ at later times, and the positive correlate environmental density ρ_1. comes from environmental density ρ_1. Environmental temperature T_1 has a small contribution, and shows the familiar feature at $T_1 \approx 10^4$ K where hydrogen ionizes. The mass ratio of nearby halos $\Upsilon_{0.1}$ provides a minor contribution. Panel (c) presents the interaction functions for the base EBM $\gamma(M_\star|\boldsymbol{\theta}')$. Each panel shows the contribution of the bivariate interaction terms, normalized such that the color map ranges between plus or minus the maximum of the norm of each function $\|f\|_{\rm max}$. Teal indicates negative contributions and green indicates positive contributions. The table lists $\|f\|_{\rm max}$ for the interaction functions, each with units $\log_{10} M_\odot$. In absolute terms, the largest interaction occurs for large virial mass $M_{\rm vir}$ and environmental temperature T_1 (same as for the base EBM $\gamma(\mathrm{SFR}|\boldsymbol{\theta}')$ modeling SFR, see panel (c) of Figure 11). Stellar mass is partially reduced for low environmental temperature T_1 and either high environmental density ρ_1 or high redshift z.

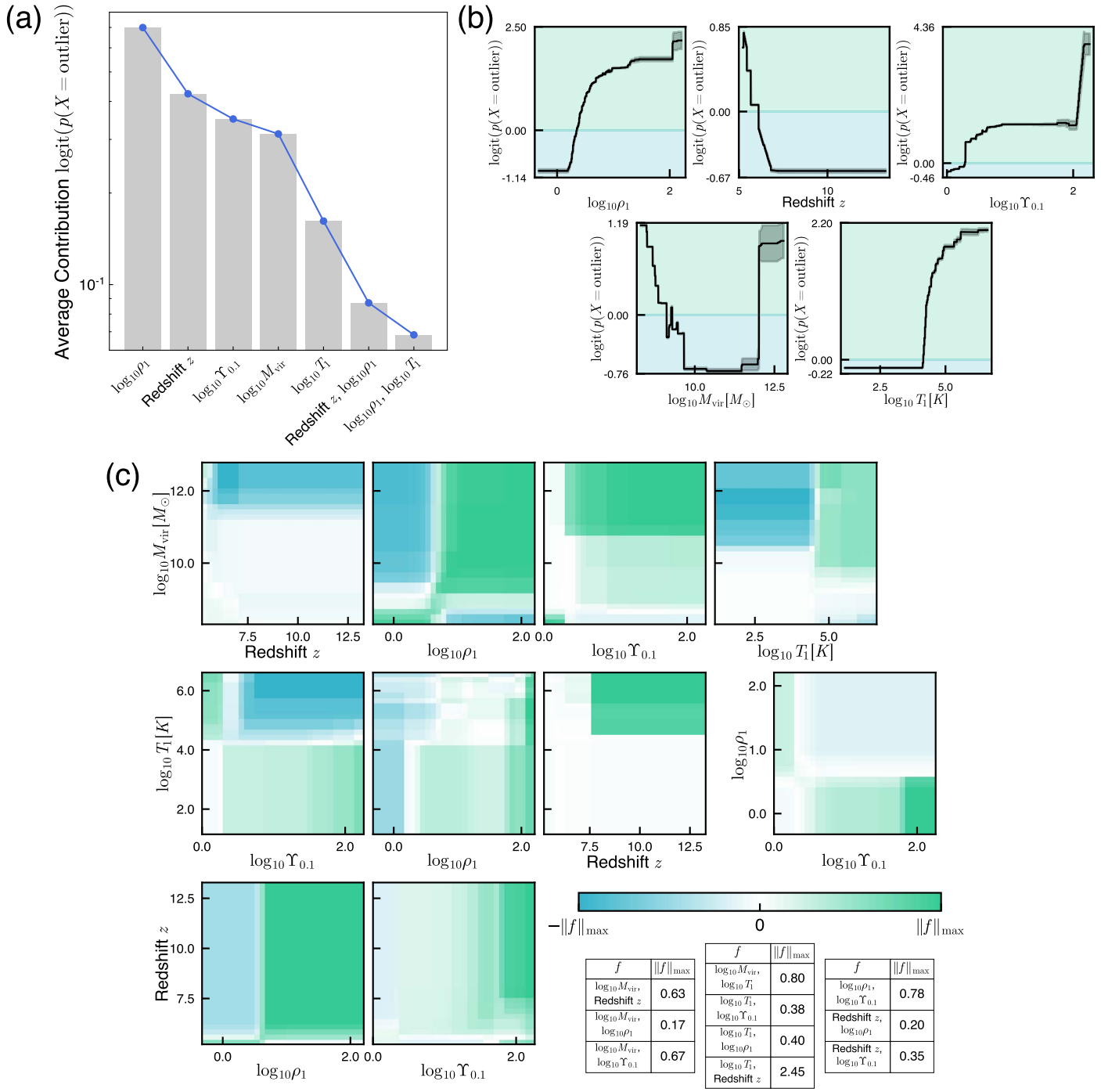


f	$\|f\|_{\rm max}$
$\log_{10} M_{\rm vir}$, Redshift z	0.09
$\log_{10} M_{\rm vir}$, $\log_{10}\rho_1$	0.07
$\log_{10} M_{\rm vir}$, $\log_{10}\Upsilon_{0.1}$	0.08

f	$\|f\|_{\rm max}$
$\log_{10} M_{\rm vir}$, $\log_{10} T_1$	0.02
$\log_{10} T_1$, $\log_{10}\Upsilon_{0.1}$	0.10
$\log_{10} T_1$, $\log_{10}\rho_1$	0.25
$\log_{10} T_1$, Redshift z	0.08

f	$\|f\|_{\rm max}$
$\log_{10}\rho_1$, $\log_{10}\Upsilon_{0.1}$	0.14
Redshift z, $\log_{10}\rho_1$	0.21
Redshift z, $\log_{10}\Upsilon_{0.1}$	0.07

Figure 16. Details for the outlier EBM model $\delta(M_\star|\boldsymbol{\theta}')$ component of the CEBM $\Gamma(M_\star|\boldsymbol{\theta}')$ trained to predict $M_\star$. Panel (a) displays the average contribution of features to the outlier EBM model $\delta(M_\star|\boldsymbol{\theta}')$. As with the base EBM $\gamma(M_\star|\boldsymbol{\theta}')$, the feature with the largest average contribution is virial mass $M_{\rm vir}$, with roughly $\gtrsim$10 $\times$ larger contribution to $\log_{10} M_\star$ than redshift z, environmental density ρ_1, or temperature T_1. The average contributions of $\Upsilon_{0.1}$ or interactions are small. Panel (b) shows the feature functions for the outlier EBM $\delta(M_\star|\boldsymbol{\theta}')$. The feature function for virial mass $M_{\rm vir}$ has the largest contribution to $\delta(\mathrm{SFR}|\boldsymbol{\theta}')$, similar to the virial mass dependence of the base EBM $\gamma(M_\star|\boldsymbol{\theta}')$ (see panel (b) of Figure 15). The stellar mass of outliers increases with increasing environmental density ρ_1, with a large enhancement at very large ρ_1. Unlike the base EBM $\gamma(M_\star|\boldsymbol{\theta}')$, the stellar mass of the outliers increases with increasing redshift. The feature function for the nearby halo mass ratio $\Upsilon_{0.1}$ is weak and noisy. Panel (c) presents the interaction functions for the outlier EBM $\delta(M_\star|\boldsymbol{\theta}')$. Each panel shows the contribution of the interaction terms, normalized such that the color map ranges between plus or minus the maximum of the norm of each function $\|f\|_{\rm max}$. Teal indicates negative contributions and green indicates positive contributions. The table lists $\|f\|_{\rm max}$ for the interaction functions, each with units $\log_{10} M_\odot$. For outliers, stellar mass increases at high environmental density ρ_1 with low environmental temperature T_1 or high redshift z.

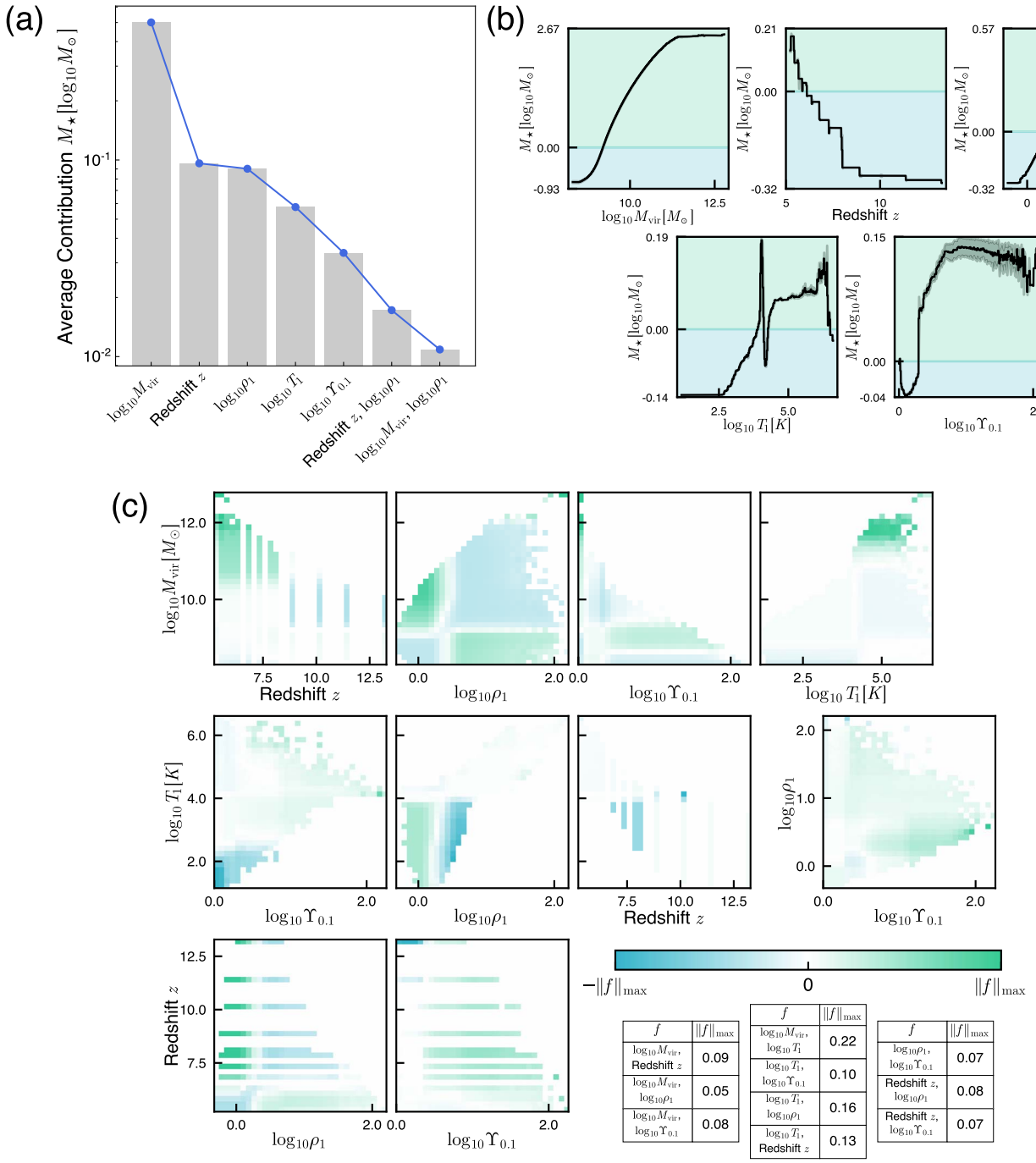


f	$\|f\|_{\max}$
$\log_{10}M_{\text{vir}}$, Redshift z	0.63
$\log_{10}M_{\text{vir}}$, $\log_{10}\rho_1$	0.17
$\log_{10}M_{\text{vir}}$, $\log_{10}\Upsilon_{0.1}$	0.67

f	$\|f\|_{\max}$
$\log_{10}M_{\text{vir}}$, $\log_{10}T_1$	0.80
$\log_{10}T_1$, $\log_{10}\Upsilon_{0.1}$	0.38
$\log_{10}T_1$, $\log_{10}\rho_1$	0.40
$\log_{10}T_1$, Redshift z	2.45

f	$\|f\|_{\max}$
$\log_{10}\rho_1$, $\log_{10}\Upsilon_{0.1}$	0.78
Redshift z, $\log_{10}\rho_1$	0.20
Redshift z, $\log_{10}\Upsilon_{0.1}$	0.35

Figure 17. Details for the classification EBM model $\phi_{M_\star}(\boldsymbol{\theta}')$ that interpolates between the base EBM $\gamma(M_\star|\boldsymbol{\theta}')$ and the outlier EBM $\delta(M_\star|\boldsymbol{\theta}')$ for creating the CEBM $\Gamma(M_\star|\boldsymbol{\theta}')$. Panel (a) displays the average contribution of features to the classification EBM model $\phi_{M_\star}(\boldsymbol{\theta}')$. The most important features for determining whether a galaxy is an outlier in the stellar mass distribution are environmental density ρ_1, redshift z, nearby halo mass ratio $\Upsilon_{0.1}$, and virial mass M_{vir}. The average contributions are unit-free, and represent changes to the log odds of a galaxy being an outlier in the stellar mass distribution. Panel (b) shows the feature functions contributing to the classifier EBM $\phi_{M_\star}(\boldsymbol{\theta}')$. These feature functions represent the change in log odds that a given galaxy will be an outlier in $M_\star$. Outliers tend to occur at high environmental density ρ_1 or very low or high virial masses M_{vir}. Galaxies with massive neighbors, reflected by $\Upsilon_{0.1}$, or high environmental temperature T_1 are also more likely to be outliers. The lowest redshift galaxies in the data set are additionally likely to be outliers in stellar mass. These trends are similar to the feature functions for the classifier EBM $\phi_{\text{SFR}}(\boldsymbol{\theta}')$ (see panel (b) of Figure 13). Panel (c) presents the interaction functions for the classifier EBM $\phi_{M_\star}(\boldsymbol{\theta}')$. Each panel shows the contributions of the interaction terms, normalized such that the color map ranges between plus or minus the maximum of the norm of each function $\|f\|_{\max}$. Teal indicates negative log odds and green indicates positive log odds that a given galaxy is an outlier in stellar mass. The table lists $\|f\|_{\max}$ for the interaction functions, listed as the corresponding change in log odds. Galaxies with large environmental temperature T_1 and at high redshift z are more likely to be outliers. Massive galaxies at high environmental density ρ_1 or with large nearby halos (large $\Upsilon_{0.1}$) also tend to be outliers. Galaxies at low environmental density but with large nearby halos also have an increased likelihood of being outliers in stellar mass.



f	$\|\|f\|\|_{\rm max}$
$\log_{10} M_{\rm vir}$, Redshift z	0.09
$\log_{10} M_{\rm vir}$, $\log_{10}\rho_1$	0.05
$\log_{10} M_{\rm vir}$, $\log_{10}\Upsilon_{0.1}$	0.08

f	$\|\|f\|\|_{\rm max}$
$\log_{10} M_{\rm vir}$, $\log_{10} T_1$	0.22
$\log_{10} T_1$, $\log_{10}\Upsilon_{0.1}$	0.10
$\log_{10} T_1$, $\log_{10}\rho_1$	0.16
$\log_{10} T_1$, Redshift z	0.13

f	$\|\|f\|\|_{\rm max}$
$\log_{10}\rho_1$, $\log_{10}\Upsilon_{0.1}$	0.07
Redshift z, $\log_{10}\rho_1$	0.08
Redshift z, $\log_{10}\Upsilon_{0.1}$	0.07

Figure 18. Details for the CEBM model $\Gamma(M_\star|\boldsymbol{\theta}')$ trained to predict stellar mass $M_\star$. Panel (a) displays the average contribution of features to the CEBM model $\Gamma(M_\star|\boldsymbol{\theta}')$. Virial mass $M_{\rm vir}$ provides the largest average contribution to the stellar mass. The environmental density ρ_1 and redshift z provide $\sim$5$\times$ smaller average contributions. The environmental temperature T_1 and nearby galaxy-mass ratio $\Upsilon_{0.1}$ provide a relative contribution to stellar mass roughly 10$\times$ smaller than $M_{\rm vir}$. Panel (b) shows the feature functions contributing to the CEBM $\Gamma(M_\star|\boldsymbol{\theta}')$. The stellar mass increases with $M_{\rm vir}$, which provides the largest contribution. Secondary contributions come from environmental density ρ_1 and redshift z. Environmental temperature T_1 has a smaller contribution, but shows the familiar feature at $T_1 \approx 10^4$K where hydrogen ionizes. The mass ratio of nearby halos $\Upsilon_{0.1}$ provides a minor contribution. As expected, the CEBM feature functions resemble the base EBM feature functions that represent the parameter dependence of stellar mass for most galaxies in the data set (see panel (b) of Figure 15). Panel (c) presents the interaction functions for the CEBM $\Gamma(M_\star|\boldsymbol{\theta}')$. Each panel shows the contribution of the interaction terms, normalized such that the color map ranges between plus or minus the maximum of the norm of each function $\|f\|_{\rm max}$. Teal indicates negative contributions and green indicates positive contributions. The table lists $\|f\|_{\rm max}$ for the interaction functions, each with units $\log_{10} M_\odot$. As for the interaction functions for the base EBM $\gamma(M_\star|\boldsymbol{\theta}')$, the largest interaction occurs for large virial mass $M_{\rm vir}$ and large environmental temperature T_1. Stellar mass is partially reduced for low environmental temperature T_1 and high environmental density ρ_1. These trends are similar to those for the base EBM $\gamma(M_\star|\boldsymbol{\theta}')$ modeling stellar mass (see panel (c) of Figure 15).

ORCID iDs

Ryan Hausen https://orcid.org/0000-0002-8543-761X
Brant E. Robertson https://orcid.org/0000-0002-4271-0364
Hanjue Zhu https://orcid.org/0000-0003-0861-0922
Nickolay Y. Gnedin https://orcid.org/0000-0001-5925-4580
Piero Madau https://orcid.org/0000-0002-6336-3293
Evan E. Schneider https://orcid.org/0000-0001-9735-7484
Bruno Villasenor https://orcid.org/0000-0002-7460-8129
Nicole E. Drakos https://orcid.org/0000-0003-4761-2197

References

Balogh, M. L., Baldry, I. K., Nichol, R., et al. 2004, ApJL, 615, L101
Behroozi, P., Wechsler, R. H., Hearin, A. P., & Conroy, C. 2019, MNRAS, 488, 3143
Bluck, A. F. L., Maiolino, R., Brownson, S., et al. 2022, A&A, 659, A160
Bouwens, R. J., Illingworth, G. D., Oesch, P. A., et al. 2012, ApJ, 754, 83
Brinchmann, J., Charlot, S., White, S. D. M., et al. 2004, MNRAS, 351, 1151
Conroy, C., Wechsler, R. H., & Kravtsov, A. V. 2006, ApJ, 647, 201
Contini, E., Gu, Q., Ge, X., et al. 2020, ApJ, 889, 156
Croton, D. J., Stevens, A. R. H., Tonini, C., et al. 2016, ApJS, 222, 22
Davé, R., Anglés-Alcázar, D., Narayanan, D., et al. 2019, MNRAS, 486, 2827
Davies, L. J. M., Robotham, A. S. G., Driver, S. P., et al. 2016, MNRAS, 455, 4013
Davies, L. J. M., Robotham, A. S. G., Lagos, C. d. P., et al. 2019, MNRAS, 483, 5444
Faber, S. M., Willmer, C. N. A., Wolf, C., et al. 2007, ApJ, 665, 265
Feulner, G., Gabasch, A., Salvato, M., et al. 2005, ApJL, 633, L9
Friedman, J. H. 2001, AnSta, 29, 1189
Girelli, G., Pozzetti, L., Bolzonella, M., et al. 2020, A&A, 634, A135
Gnedin, N. Y. 2014, ApJ, 793, 29
Hastie, T., Friedman, J., & Tibshirani, R. 2001, The Elements of Statistical Learning (New York: Springer)
Hastie, T., & Tibshirani, R. 1986, StaSc, 1, 297, http://www.jstor.org/stable/2245459
Hearin, A. P., Zentner, A. R., Berlind, A. A., & Newman, J. A. 2013, MNRAS, 433, 659
Hunter, J. D. 2007, CSE, 9, 90
Iliev, I. T., Mellema, G., Ahn, K., et al. 2014, MNRAS, 439, 725
Jing, Y. P., Mo, H. J., & Börner, G. 1998, ApJ, 494, 1
Kalita, B. S., Daddi, E., D'Eugenio, C., et al. 2021, ApJL, 917, L17
Kannan, R., Garaldi, E., Smith, A., et al. 2022, MNRAS, 511, 4005
Kauffmann, G., White, S. D. M., Heckman, T. M., et al. 2004, MNRAS, 353, 713
Kravtsov, A. V., Vikhlinin, A. A., & Meshcheryakov, A. V. 2018, AstL, 44, 8
Lehmann, B. V., Mao, Y.-Y., Becker, M. R., Skillman, S. W., & Wechsler, R. H. 2017, ApJ, 834, 37
Li, C., Jing, Y. P., Mao, S., et al. 2012, ApJ, 758, 50
Lin, Y.-T., Mohr, J. J., & Stanford, S. A. 2004, ApJ, 610, 745
Lou, Y., Caruana, R., Gehrke, J., et al. 2012, in Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD '12 (New York: ACM), 150
Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. 2013, in Proc. of the 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD '13 (New York: ACM), 623
Lovell, C. C., Wilkins, S. M., Thomas, P. A., et al. 2022, MNRAS, 509, 5046
Machado Poletti Valle, L. F., Avestruz, C., Barnes, D. J., et al. 2021, MNRAS, 507, 1468
Mandelbaum, R., Tasitsiomi, A., Seljak, U., Kravtsov, A. V., & Wechsler, R. H. 2005, MNRAS, 362, 1451
McGibbon, R. J., & Khochfar, S. 2022, MNRAS, 513, 5423
More, S., van den Bosch, F. C., Cacciato, M., et al. 2009, MNRAS, 392, 801
Moster, B. P., Somerville, R. S., Maulbetsch, C., et al. 2010, ApJ, 710, 903
Nelson, D., Pillepich, A., Genel, S., et al. 2015, A&C, 13, 12
Noeske, K. G., Weiner, B. J., Faber, S. M., et al. 2007, ApJL, 660, L43
Nori, H., Jenkins, S., Koch, P., & Caruana, R. 2019, arXiv:1909.09223
Ocvirk, P., Aubert, D., Sorce, J. G., et al. 2020, MNRAS, 496, 4087
Ocvirk, P., Gillet, N., Shapiro, P. R., et al. 2016, MNRAS, 463, 1462
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Pillepich, A., Springel, V., Nelson, D., et al. 2018, MNRAS, 473, 4077
Piotrowska, J. M., Bluck, A. F. L., Maiolino, R., & Peng, Y. 2022, MNRAS, 512, 1052
Reddick, R. M., Wechsler, R. H., Tinker, J. L., & Behroozi, P. S. 2013, ApJ, 771, 30
Rodríguez-Puebla, A., Primack, J. R., Avila-Reese, V., & Faber, S. M. 2017, MNRAS, 470, 651
Schaye, J., Crain, R. A., Bower, R. G., et al. 2015, MNRAS, 446, 521
Somerville, R. S., & Davé, R. 2015, ARA&A, 53, 51
Stark, D. P., Ellis, R. S., Bunker, A., et al. 2009, ApJ, 697, 1493
Tomczak, A. R., Quadri, R. F., Tran, K.-V.-H., et al. 2014, ApJ, 783, 85
Trussler, J., Maiolino, R., Maraston, C., et al. 2020, MNRAS, 491, 5406
Vale, A., & Ostriker, J. P. 2004, MNRAS, 353, 189
Vale, A., & Ostriker, J. P. 2006, MNRAS, 371, 1173
van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, CSE, 13, 22
van Rossum, G. 1995, Python Reference Manual, https://www.narcis.nl/publication/RecordID/oai%3Acwi.nl%3A5008
Velander, M., van Uitert, E., Hoekstra, H., et al. 2014, MNRAS, 437, 2111
Villasenor, B., Robertson, B., Madau, P., & Schneider, E. 2021, ApJ, 912, 138
Villasenor, B., Robertson, B., Madau, P., & Schneider, E. 2022, ApJ, 933, 26
Wechsler, R. H., & Tinker, J. L. 2018, ARA&A, 56, 435
Xu, X., Kumar, S., Zehavi, I., & Contreras, S. 2021, MNRAS, 507, 4879
Zhu, H., Avestruz, C., & Gnedin, N. Y. 2020, ApJ, 899, 137
Zhu, Y., Becker, G. D., Bosman, S. E. I., et al. 2021, ApJ, 923, 223