must be a function of simulation initial conditions, such as planet orbital properties. Before a search is complete, we identify regions of parameter space where more simulations are required by noting where our fitted model is uncertain (see Section 3).

To maximize the accuracy of our model, we created additional features that are physically meaningful functions of initial conditions for a given simulation. For example, for each planet, we create total and normal angular momentum proxy features of the form $\sqrt{1-e^2}$ and $\sqrt{1-e^2}\cos i$ for orbital eccentricity $e$ and inclination $i$. We created an additional 14 physically-motivated features and call this augmented feature set "Physical". We synthesized more features by transforming the Physical feature set to all monomials of degree 2 including all cross-terms. This transformation, still just a function of initial conditions, yielded a total of about 200 features per sample. We call this augmented feature set "Polynomial".

We fit our data using ordinary least squares (OLS) and ridge regression (RR) to gauge the importance of regularization. We trained each model on a training subset containing 8,000 simulations and tested on the remaining 2,000. For RR, we performed randomized 5-fold cross validation on the training set over 50 logarithmically spaced bins for $\alpha \in [10^{-10}, 10]$ to optimize the regularization parameter $\alpha$. We found $\alpha \sim 10^{-7}$ yielded optimal performance. Once fit, we evaluated both models on the training and testing set and recorded the mean squared error (MSE) and the coefficient of determination, $R^2$ (see Tables 1 and 2). All fits and hyperparameter optimization made use of [3].

## 3 Bootstrapping and Model Comparison

In order to determine the sampling distribution we used the bootstrapping method [4] with 100 realizations of our data and computed the standard deviation. Areas in parameter space with a high bootstrapped standard deviation are regions where more simulations should be run to improve the fit of the given estimator. Specifically, we computed the standard deviation using the algorithm found by [5]. We explored how the number of bootstraps influenced our standard deviation calculation and found that using an order of magnitude more realizations did not significantly impact the fit.

We found that both linear models, Ordinary Least Squares and Ridge Regression, performed comparably as seen in Tables 1 and 2 implying, at least for these models, that regularization is not important. We do notice a somewhat significant improvement in the MSE and the $R^2$ values when using the Polynomial feature set compared to the Physical feature set. We find that the bootstrapped standard deviation is roughly a factor of two larger in the Polynomial set relative to the Physical set suggesting that bootstrapping error scales with feature number. These simple linear methods show promise in being able to predict the outcomes of our complex simulations. We are therefore optimistic that more advanced methods will succeed in replacing the need to run more simulations.

Table 1: Physical feature set fit results.

| Model | Train MSE | Test MSE | Train $R^2$ | Test $R^2$ | Estimator Median Std |
|---|---|---|---|---|---|
| OLS | 0.097308 | 0.099976 | 0.575670 | 0.561866 | 0.013340 |
| RR | 0.097311 | 0.099923 | 0.575658 | 0.562102 | 0.016322 |
| RF | 0.014023 | 0.033080 | 0.941388 | 0.861286 | - |
| GP Matern | 0.050634 | 0.081951 | 0.788365 | 0.656356 | 0.279772 |
| GP RBF | 0.065208 | 0.083292 | 0.727450 | 0.650734 | 0.281678 |
| GP RQ | 0.046516 | 0.081516 | 0.805579 | 0.658183 | 0.279678 |
| XGBoost | 0.005621 | 0.028202 | 0.976505 | 0.88174 | - |

As an example, we visualized the inclinations of Proxima b and Proxima c shown in the figure xxx and colored each simulation by the bootstrapping-derived standard deviation of the fit, our proxy for