NBER WORKING PAPER SERIES

# ASSET PRICING WITH PANEL TREE UNDER GLOBAL SPLIT CRITERIA

Lin William Cong
Guanhao Feng
Jingyu He
Xin He

Working Paper 30805
http://www.nber.org/papers/w30805

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2022

The authors acknowledge financial support from INQUIRE Europe and the Hong Kong Research Grants Council (Project No. CityU 11507022). Will Cong is also an advisor to BlackRock, Inc. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Asset Pricing with Panel Tree Under Global Split Criteria
Lin William Cong, Guanhao Feng, Jingyu He, and Xin He
NBER Working Paper No. 30805
December 2022
JEL No. C1,G11,G12

**ABSTRACT**

We develop a new class of tree-based models (P-Tree) for analyzing (unbalanced) panel data utilizing global (instead of local) split criteria that incorporate economic guidance to guard against overfitting while preserving interpretability. We grow a P-Tree top-down to split the cross section of asset returns to construct stochastic discount factor and test assets, generalizing sequential security sorting and visualizing (asymmetric) nonlinear interactions among firm characteristics and macroeconomic states. Data-driven P-Tree models reveal that idiosyncratic volatility and earnings-to-price ratio interact to drive cross-sectional return variations in U.S. equities; market volatility and inflation constitute the most critical regime-switching that asymmetrically interact with characteristics. P-Trees outperform most known observable and latent factor models in pricing individual stocks and test portfolios, while delivering transparent trading strategies and risk-adjusted investment outcomes (e.g., out-of-sample annualized Sharp ratios of about 3 and monthly alpha around 0.8%).

Lin William Cong
SC Johnson College of Business
Cornell University
Sage Hall
Ithaca, NY 14853
and NBER
will.cong@cornell.edu

Guanhao Feng
City University of Hong Kong
83 Tat Chee Avenue
Kowloon Tong
Hong Kong, China
gavin.feng@cityu.edu.hk

Jingyu He
City University of Hong Kong
83 Tat Chee Avenue
Kowloon Tong
Hong Kong, China
jingyuhe@cityu.edu.hk

Xin He
College of Finance and Statistics
Hunan University
Changsha, China
xinh@hnu.edu.cn

A P-Tree Github Package is available at https://github.com/Quantactix/TreeFactor

# 1 Introduction

Tree-based models are exceptional in predictive performance when dealing with data of high dimensionality, nonlinearity, asymmetric variable interactions, low signal-to-noise ratios, and small sample environments (e.g., Sorensen et al., 2000; Rossi and Timmermann, 2015; Gu et al., 2020; Bali et al., 2022). However, they are developed for pure prediction tasks, and are neither guided by economic principles nor tailored for financial panel data. Meanwhile, despite the popularity of factor models (e.g., Fama and French, 2015; Hou et al., 2021), researchers recognize the need to apply machine learning (ML) models to understand the time-series co-movement and cross-sectional variation of asset returns using high-dimensional asset characteristics from financial big data (e.g., Kelly et al., 2019; Kozak et al., 2020; Lettau and Pelger, 2020b). Nonetheless, many ML methods appear black boxes without straightforward formulation and interpretation. When taken off-the-shelf, they do not incorporate economic restrictions such as no arbitrage.

Therefore, we develop a novel unified P-Tree framework ("P" for "panel") for analyzing (unbalanced) panel data, inheriting the advantages of tree models while incorporating interpretability under economics-guided global split criteria.[1] When applied to asset pricing, P-Tree extends the scope of tree models beyond pure prediction to recovering the stochastic discount factor (SDF) from a panel of individual asset returns. Specifically, P-Tree splits the cross section top-down to generate leaf basis portfolios, achieving supervised dimension reduction and generalizing sequential security sorting for constructing characteristics-managed basis assets. We derive SDF from leaf portfolios and grow P-Tree following global split criteria to explain the cross-sectional returns or maximize the Sharpe ratio of investment portfolios.

While other applications of P-Tree exist, the asset pricing application effectively illustrates our methodological innovations. First, ML methods, including tree-based meth-

[1] We develop and post a P-Tree package in `R`, `TreeFactor`, for general explorations by other researchers (`https://github.com/Quantactix/TreeFactor`).

ods, assume observations are independently and identically distributed (*i.i.d.*). P-Tree admits the panel structure of data, incorporates vectorized leaf parameters for multi-period observations, and performs supervised dimension reduction for the unbalanced panel. Specifically, the P-Tree nodes are time series of portfolios for (value-weighted) averages of individual asset returns. This innovation allows P-Tree to overcome the high-dimensional sorting challenge raised in Cochrane (2011) and to capture potentially asymmetric interactions of multiple characteristics, which are desirable in empirical asset pricing.

Second, tree-based models for prediction, such as CART (Classification and Regression Tree), grow recursively and optimize split rules at each node without considering other sibling nodes. This "myopic" strategy focuses on local optimization for computational ease but usually leads to overfitting because it operates on fewer observations in each node as the tree grows. In contrast, P-Tree is designed to utilize data from the entire cross section to guard against overfitting when iteratively generating basis portfolios and latent factors. In our asset pricing application, P-Tree recovers the SDF by minimizing pricing errors or maximizing the Sharp ratio (theoretically equivalent), both global criteria entailing all sample observations. P-Tree thus customizes a nonlinear ML method to combine economic principles and statistical properties, while achieving transparency and interpretability of non-ensemble trees.[2]

Empirically, we apply P-Tree to individual stock returns in the U.S. from 1981 to 2020.[3] The asset pricing and investment P-Tree factor models outperform most alternatives, including the well-known observable factor models (e.g., Fama and French, 2015; Hou et al., 2021) and latent factor models (e.g., Kelly et al., 2019; Lettau and Pelger, 2020b), in terms of pricing and investment performance. In particular, P-Tree models show excellent pricing performance in explaining individual stock and test portfolio returns. Investment P-Tree models offer annualized Sharpe ratios close to 3 and significant monthly

[2]When boosted to generate multiple factors, P-Trees can be viewed as an alternative to PCA approaches but with greater interpretability and (potentially asymmetric) nonlinear interactions.

[3]P-Tree can be applied beyond equities. For example, the online appendix reports its implementation on U.S. corporate bonds.

alphas of about 0.80% in the test sample. Moreover, the multi-characteristics sequentially-sorted P-Tree leaf basis portfolios better span the efficient frontier and serve as challenging test assets for alternative factor models.

Our study of the out-of-bag variable importance validates the transparent and interpretable P-Tree models, because single P-Tree and P-Forest (ensemble) reveal the same small set of characteristics (e.g., idiosyncratic volatility and earnings-to-price ratio) as drivers of the cross-sectional return variation through nonlinear interactions. P-Tree can also split the panel of return data on the time series using macroeconomic variables: market volatility and inflation are crucial macro predictors for considering regime-switching. Idiosyncratic volatility plays a crucial role when splitting the cross section for either high- or low-market-volatility periods, whereas market equity is more important during low-inflation periods. Moreover, P-Tree displays all the split rules in sequence, which helps researchers understand the interactions among firm characteristics and between macroeconomic variables and firm characteristics, and potentially enhance trading strategies and resurrect anomalies.[4] Overall, P-Tree offers a new framework for panel data analysis that incorporates economic guidance through global split criteria and can be tailored for asset pricing applications and beyond.

**Literature.** Machine learning has gained adoption in finance in recent years, has been shown to be powerful in predicting individual asset returns with a high dimension of asset characteristics (e.g., Freyberger et al., 2020; Gu et al., 2020; Bianchi et al., 2021; Bali et al., 2022). While recent studies focus on deep learning or reinforcement learning(Chen et al., 2022; Cong et al., 2019; Feng et al., 2022), P-Tree adds a more interpretable class of Tree-based models that can be customized for asset pricing.[5] Though financial economists have

[4]In particular, P-Tree allows a long-short factor's long and short leg portfolios to interact with different characteristics, which loads the portfolio on different leaf basis portfolios. This is in contrast to the traditional treatment of a long-short portfolio as a single asset, complementing the pioneering work of Jarrow et al. (2021) to model the two legs of anomaly portfolios separately.

[5]Technically speaking, our paper also contributes to AI and data science by introducing a new form of economically guided, self-supervised learning (SSL) algorithm. Existing SSL train large models without

used the regression tree or ensemble methods such as boosted trees and random forests (e.g., Rossi and Timmermann, 2015; Gu et al., 2020), most of them apply off-the-shelf tree models for prediction. One recent exception by Creal and Kim (2021) identifies variables that best explain assets' betas by splitting currency-return observations via Bayesian additive regression trees (BART). Instead of risk premia, we focus on the panel structure of returns and interpretability, which are crucial for understanding asymmetric nonlinear interactions and guiding portfolio constructions.

Our study contributes to the growing literature on latent factor models related to statistical factor models, such as principal component analysis (PCA). Latent factor models in asset pricing start with the arbitrage pricing theory of Ross (1976), and the empirical test in Roll and Ross (1980). For recent work, the projected PCA of Kim et al. (2021), and the instrumental PCA of Kelly et al. (2019) and Kelly et al. (2022) use firm characteristics as instruments to model the time-varying factor loadings and estimate principal components. Lettau and Pelger (2020a,b) develop the risk premia PCA and provide a regularized estimation for risk premia. Recent studies, such as the auto-encoder (Gu et al., 2021), generative adversarial network (Chen et al., 2022), and characteristics-sorted factor approximation (Feng et al., 2021) have also developed nonlinear deep neural networks for latent factor modeling. P-Tree belongs to nonlinear latent factor models but is the only one providing a graphical representation for variable nonlinearity and asymmetric interactions, something PCA methods or deep learning ill-afford.

More generally, for recovering the pricing kernel, Kozak et al. (2020) use a shrinkage estimator on the SDF coefficients for characteristics-based factors. Bryzgalova et al. (2020) estimate the regularized SDF on a given set of characteristics-managed portfolios with a pruning algorithm with a global criterion. P-Tree differs in growing a tree top-down

labeled data through pretext tasks and data-driven supervisory signals (e.g., Chen et al., 2020), and is fast emerging in AI and data science with various applications. in image, audio, and language processing and recognition (e.g., OpenAI's GPT-3, Google's BERT Model, and Facebook(Meta)'s wav2vec). We join Cong et al. (2022) as the earliest studies to utilize economic principles (e.g., no-arbitrage under a factor return structure) for iteratively generating supervisory signals (e.g., based on pricing performance) for clustering observations.

rather than bottom-up (pruning), and generating test portfolios endogenously.[6] Our approach exploits a global iterative greedy algorithm to effectively search in a much larger space of tree structures than the conventional sequential sorting to uncover nonlinear signals and predictor interactions. Cong et al. (2022) is another application of the panel tree structure for clustering individual assets and estimating heterogeneous factor models.

We also join the emerging studies about imposing economic restrictions on estimating and evaluating ML or statistical factor models (e.g., Gagliardini et al., 2016; Feng et al., 2021; Chen et al., 2022; Avramov et al., 2022). For evaluating risk factors, Feng et al. (2020) and Bryzgalova et al. (2022) estimate the risk price for high dimensional factors by variable selection cross-sectional regressions. Rossi and Timmermann (2015) adopt the boosted regression to estimate the conditional covariance for ICAPM. DeMiguel et al. (2020) offer the economic rationale for why many characteristics are needed in portfolio investment, by considering transaction costs. While no-arbitrage has been applied to linear parametric models, we adopt it into flexible ML methods—a challenging task because of the additional SDF modeling.

Finally, this paper relates to the construction of basis assets and the general asset pricing literature. P-Tree contributes the first economic-supervised solution of clustering individual stocks and generating leaf-basis assets. Ahn et al. (2009) generate basis assets by unsupervised clustering of assets according to the correlation structure of returns. Zhu et al. (2020) develop a group-wise interpretable basis selection to estimate a new adaptive multi-factor model. The literature, including regularized linear models (Kozak et al., 2020; Bryzgalova et al., 2020) and the PCA approaches (Kozak et al., 2018; Haddad et al., 2020; Lettau and Pelger, 2020b), largely takes pre-specified or characteristics-managed portfolios as given. P-Tree generates these characteristics-managed portfolios, complementing the adversarial networks approach in Chen et al. (2022), among others.

[6]Enumerating all possible tree configurations to prune from is an NP-hard problem. It is not computationally feasible unless one manually specifies a small set of variables and shallow depth for initial trees.

# 2 P-Tree Factor Models for Asset Pricing

## 2.1 Tree-based Models and P-Tree Innovations

Before detailing P-Trees' growth, we formally describe CART, the building block of the tree-based method, introduce P-Tree models, and explain how they innovate.

**CART and asset pricing interpretation.** Decision trees partition the space of predictors into rectangles by a sequence of splits and provide local conditional responses.[7] Let $\mathbf{z}_i = (z_{i,1}, \cdots, z_{i,K})$ denote a vector of $K$ predictors for the $i$-th observation. A $j$-th split rule of the tree is $\tilde{\mathbf{c}}^{(j)} = (z_{\cdot,k}, c_j)$, which splits the data sample along the $k$-th predictor $z_{\cdot,k}$ according to greater or smaller than the value $c_j$[8]. A tree consists of $J$ splits in total that partition the predictor space into $J+1$ regions (leaf nodes), which we denote by $\{\mathcal{R}_j\}_{j=1}^{J+1}$. Importantly, CART assigns a constant leaf parameter $\mu_j$ to each leaf node $\mathcal{R}_j$ for prediction tasks. The regression tree $\mathcal{T}$, with parameters $\Theta_J = \{\{\tilde{\mathbf{c}}^{(j)}\}_{j=1}^{J}, \{\mu_j\}_{j=1}^{J+1}\}$, can be expressed as

$$\mathcal{T}(\mathbf{z}_i \mid \Theta_J) = \sum_{j=1}^{J+1} \mu_j \mathbf{I}(\mathbf{z}_i \in \mathcal{R}_j). \tag{1}$$

The indicator function $\mathbf{I}(\mathbf{z}_i \in \mathcal{R}_j)$ takes value $1$ for one and only one leaf node and $0$ for all others. CART takes a simple average of training data in each leaf node to estimate the leaf parameter. The averaging within a node and through ensembles guard CART models against overfitting.

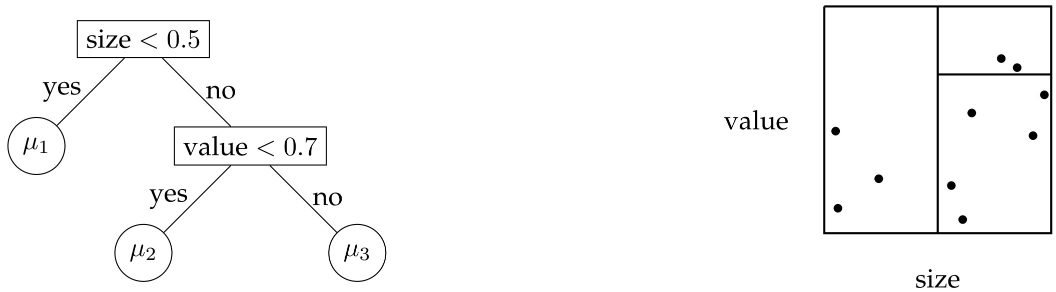
To predict a new observation, CART finds the leaf node and reports the associated leaf parameter as the prediction. Figure 1 illustrates how CART uses firm characteristics (with the corresponding partition of the characteristics space and their constant leaf parameters) to fit stock returns. Growing a tree entails finding the best split rules sequen-

[7]CART(Breiman et al., 1984) is the most influential binary decision tree model in statistics, and is the building block for ensemble methods. Well-known ensemble tree models include random forest (Breiman, 2001), boosting trees (Freund and Schapire, 1997), or Bayesian approaches such as Bayesian additive regression trees (Chipman et al., 2010) and the recent XBART (He et al., 2019; He and Hahn, 2021).

[8]A multiway-split tree can always be represented as a tree with multiple binary splits.

Figure 1: Example: A Decision Tree For Prediction

Left: A decision tree with two splits, three leaf nodes, and three leaf parameters. Right: Corresponding partition plot for the sample of predictor space on value and size.




tially. CART grows by recursively partitioning the characteristics of space. From a parent node, the best split is determined by all split rule candidates according to pre-specified split criterion, and then the data sample is divided into two regions for left and right child nodes. The growing process repeats until some pre-specified stopping conditions are met, such as the total number of leaves, maximal tree depth, or a minimal number of observations in each leaf. When splitting a particular node, the recursive algorithm only considers observations in that local node without looking at other nodes.

Although CART is designed purely for prediction, we interpret it for single-period asset pricing as follows. Let $r_{i,t}$ denote the return of asset $i$ at one same time period $t$. The split criterion typically used is the sum of squared errors:

$$\mathcal{L}(\tilde{\mathbf{c}}) = \sum_{i \in \text{left node}} (r_{i,t} - \bar{r}_{\text{L}})^2 + \sum_{i \in \text{right node}} (r_{i,t} - \bar{r}_{\text{R}})^2, \tag{2}$$

where $\bar{r}_{\text{L}} = \frac{1}{\#_{\text{left node}}} \sum_{i \in \text{left node}} r_{i,t}$ and $\bar{r}_{\text{R}} = \frac{1}{\#_{\text{right node}}} \sum_{i \in \text{right node}} r_{i,t}$ are average excess returns in the left or right leaf nodes, respectively. If we continue to grow the tree in Figure 1, the bottom nodes (leaf nodes) are associated with constant leaf parameters. In other words, assets in the same node share the same constant leaf parameter, the expected return estimated. This one-period tree structure is typically used to capture the cross-sectional difference of average excess returns for individual stocks.

**CART vs. P-Tree.** Applying the off-the-shelf tree models in ML for asset pricing instead of pure prediction has at least two serious limitations. First, CART requires same-period return observations. Like most ML methods, CART assumes *i.i.d.* observations but ignores the panel structure of asset pricing data. When one fits CART to multiple data periods, the leaf parameters lose the pricing kernel interpretation, calling for new methods adapted to panel data for cross-sectional asset pricing. An additional advantage of adapting tree models to panel data involves portfolio optimization on *leaf basis portfolios* — a vector of portfolio returns over multiple periods. We, therefore, specify the leaf parameters of P-Trees to be vectors instead of scalars.

Second, when splitting a node, CART only considers observations within that node without using information from other sibling nodes. This recursive statistical algorithm leads to overfitting idiosyncratic noises from observations in that node when growing a single tree and does not incorporate any economic guidance. P-Tree grows iteratively instead, imposing a *global* split criterion defined on all leaf nodes, which explicitly moderates the in-sample overfitting. The algorithm searches for all split rule candidates in all current leaf nodes to find the optimal split that improves the pre-defined performance metric. This metric, reflected in a global split criterion, is crucial to asset pricing studies when the goal is to construct a factor model that prices all individual stock returns or delivers the highest Sharpe ratio for investment.

## 2.2 Growing a P-Tree

A P-Tree splits the universe of unbalanced individual assets, over $T$ time periods with $N_t$ assets in each time period and $N$ assets in total, into non-overlapping leaf nodes according to values of ranked firm characteristics.[9] Leaf basis portfolios are thus created

[9] The P-Tree framework is flexible enough to accommodate a simultaneous sorting scheme (i.e., Fama-French ME - B/M 5×5 equity portfolios). One needs to split along ordered duplet or multiplet of characteristics, that is, linear partitions, not parallel to any single input variable. P-Tree can also split according to macro variables in the time series (see Section 4.4).

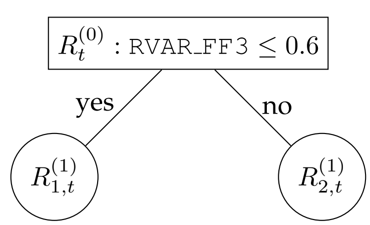
as value-weighted portfolios for stocks within each leaf node. The number of basis portfolios increases one at a time when the tree splits a parent node into two child nodes. After the $j$-th split, P-Tree generates $j+1$ leaf basis portfolios. Therefore, it reduces dimension from thousands of individual stocks to $j+1$ leaf basis portfolios. This idea of characteristics-managed leaf basis portfolios is similar to commonly used characteristics-sorted portfolios but allows for the asymmetric interactions of multiple characteristics.

Here is the growth procedure: Let $\mathbf{R}_t^{(j)}$ denote the return vector of the leaf basis portfolios after the $j$-th iteration of the tree. After the $j$-th splits, the tree has $j+1$ basis portfolios, denoted by $\mathbf{R}_t^{(j)} = \left[R_{1,t}^{(j)}, R_{2,t}^{(j)}, \cdots, R_{j+1,t}^{(j)}\right]$. Here, $R_{k,t}^{(j)}$ represents a vector of returns of $T$ periods for the $k$-th leaf basis portfolio. We denote $f_t^{(j)}$ as the P-Tree factor (latent factor generated by P-Tree) generated after the $j$-th splits and $\beta^{(j)}(\mathbf{z}_{i,t-1}) = b_0^{(j)} + \mathbf{b}^{(j)\intercal}\mathbf{z}_{i,t-1}$ as the conditional factor loadings driven by firm characteristics.

The portfolio at the root node ($\mathbf{R}_t^{(0)}$, before the first split) corresponds to the market factor: a single-leaf basis portfolio of all assets. We gradually grow the tree with additional splits by updating $\left\{\mathbf{R}_t^{(j)}, f_t^{(j)}, \beta^{(j)}(\cdot)\right\}$ iteratively. First, leaf basis portfolios $\mathbf{R}_t^{(j)}$ are expanded when the tree splits and grows one more leaf. Second, the P-Tree factor, $f_t^{(j)}$, is re-estimated using the expanded leaf basis portfolios. Finally, factor loadings $\beta^{(j)}(\cdot)$ are re-estimated for the updated P-Tree factor for individual asset returns.

Figure 2: Illustration of the First Split

To search for the optimal characteristic and cutpoint value, we consider one split rule candidate, RVAR_FF3$\leq 0.6$, for calculating the split criterion.




**First split.** Before the first split, the entire cross section lies in the root node, and the corresponding value-weighted leaf basis portfolio $\mathbf{R}_t^{(0)}$ is the market factor. Firm characteristics are normalized cross-sectionally to $[-1, 1]$ uniformly within each time period. For all characteristics and macroeconomic variables, we consider various split threshold candidates. For example, we use quintile split rule candidates at -0.6, -0.2, 0.2, and 0.6.

Consider a split rule candidate $\tilde{\mathbf{c}}_{k,m} = (z_{\cdot,k}, c_m)$, which partitions the root node to the left and right child nodes according to whether the $k$-th characteristics is smaller than value $c_m$ or not. Figure 2 illustrates a candidate for the first split. Stock-return observations in each potential child leaf form a leaf basis portfolio, denoted by $R_{1,t}^{(1)}$ and $R_{2,t}^{(1)}$ respectively. The P-Tree factor is estimated as the mean-variance efficient (MVE) portfolio of all leaf basis portfolios,[10]

$$f_t^{(1)} = \mathbf{w}^{(1)\intercal}\mathbf{R}_t^{(1)}, \quad \mathbf{w}^{(1)} \propto \mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1, \tag{4}$$

where $\mathbf{\Sigma}_1$ and $\boldsymbol{\mu}_1$ are the covariance matrix and average excess returns for leaf basis portfolios $\mathbf{R}_t^{(1)} = \left[R_{1,t}^{(1)}, R_{2,t}^{(1)}\right]$ after the first split.[11]

To gauge the quality of the split rule candidate, we measure a customized split criterion based on the loss of aggregate pricing errors,

$$\mathcal{L}(\tilde{\mathbf{c}}_{k,m}) = \sum_{t=1}^{T}\sum_{i=1}^{N_t}\left(r_{i,t} - \beta^{(1)}(\mathbf{z}_{i,t-1})f_t^{(1)}(\tilde{\mathbf{c}}_{k,m})\right)^2, \tag{5}$$

where $f_t^{(1)}$ is determined by the split rule $\tilde{\mathbf{c}}_{k,m}$, and $\beta^{(1)}(\mathbf{z}_{i,t-1}) = b_0^{(1)} + \mathbf{b}^{(1)\intercal}\mathbf{z}_{i,t-1}$ is the conditional beta driven by characteristics $\mathbf{z}_{i,t-1}$. Note that (5) is a customized split criterion

[10]For a robust estimation of the efficient portfolio weights, we include two small shrinkage parameters, $\gamma_\Sigma = 10^{-4}$ and $\gamma_\mu = 10^{-4}$ and the estimation is

$$\mathbf{w}^{(j)} = \left(\text{Cov}(\mathbf{R}_t^{(j)}) + \gamma_\Sigma \mathbf{I}_{k+1}\right)^{-1}\left(\text{E}(\mathbf{R}_t^{(j)}) + \gamma_\mu \mathbf{1}\right), \tag{3}$$

where $\mathbf{I}_{k+1}$ is the identity matrix and $\mathbf{1}$ is a vector of ones. The shrinkage parameters help stabilize the portfolio weight estimation and avoid over-leveraging. Larger shrinkage parameters imply a larger regularization. These regularized portfolio optimization studies are also addressed in Bryzgalova et al. (2020).

[11]The absolute sum of portfolio weights is normalized to two, which is equivalent to a 50% margin constraint. According to Regulation T of the Federal Reserve Board, a minimum of 50% of the security's current market value is required for the margin account.

with an embedded factor model:

$$\mathcal{T}(\mathbf{z}_{i,t-1}|\Theta) = \beta^{(1)}(\mathbf{z}_{i,t-1}) f_t^{(1)},$$

where $\Theta$ is a collection of split rule and leaf regression coefficients $\beta^{(1)}$ for the tree. The leaf parameter is a function conditional on latent factor and dynamic betas. The betas are estimated by a pooled regression model for all individual stock returns regardless of the leaf node membership. Thus the split criterion is "globally" involving all stock-return observations. In the P-Tree framework, the choice of split criteria is flexible. We demonstrate the aggregate pricing errors as an example in the current subsection, and discuss asset pricing or investment criterion further in Section 2.3.

A different split rule candidate $\tilde{\mathbf{c}}_{k,m}$ produces different partitions of the data, thus creating different leaf basis portfolios, corresponding P-Tree factors, and valuation of split criteria in (5) eventually. Therefore, we loop over all split rule candidates, and choose the one minimizing the split criteria as the first split rule.

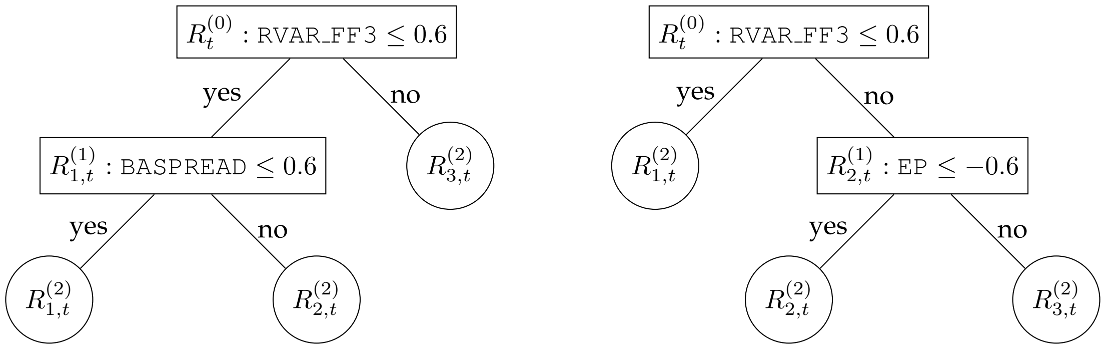
**Second split.** The second split can happen at either the root's left or right child node. We evaluate the split criteria for all split rule candidates for *both* leaf nodes, and pick the one that minimizes the split criteria as in (5). Figure 3 depicts the tree of the candidates for the second split. In either case, one leaf node splits, becomes an internal node, and creates two new leaf nodes. The P-Tree factor is now constructed based on *all three* leaf basis portfolios:

$$f_t^{(2)} = \mathbf{w}^{(2)\intercal}\mathbf{R}_t^{(2)}, \quad \mathbf{w}^{(2)} \propto \mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2, \tag{6}$$

where $\mathbf{\Sigma}_2$ and $\boldsymbol{\mu}_2$ are the covariance matrix and average excess returns for leaf basis portfolios $\mathbf{R}_t^{(2)} = \left[R_{1,t}^{(2)}, R_{2,t}^{(2)}, R_{3,t}^{(2)}\right]$ after the second split. The construction of three basis portfolios depends on which node the candidate splits, as shown in Figure 3.

Figure 3: Illustration of the Second Split

These two figures display two candidates for the second split. Either the left or right node splits, and the tree model's second ($j$-th) iteration has three ($j+1$) leaf basis portfolios to generate the latent factor.




(a) Splitting node $R_{1,t}^{(1)}$ at BASPREAD.

(b) Splitting node $R_{2,t}^{(1)}$ at EP.

The updated P-Tree factor is plugged into the split criteria,

$$\mathcal{L}(\tilde{\mathbf{c}}_{k,m}) = \sum_{t=1}^{T}\sum_{i=1}^{N_t}\left(r_{i,t} - \beta^{(2)}(\mathbf{z}_{i,t-1})f_t^{(2)}(\tilde{\mathbf{c}}_{k,m})\right)^2, \tag{7}$$

where $\beta^{(2)}(\mathbf{z}_{i,t-1}) = b_0^{(2)} + \mathbf{b}^{(2)\intercal}\mathbf{z}_{i,t-1}$. Notably, it is still defined globally on the entire cross section. Furthermore, we explore all candidates of all possible nodes to find the one with the largest global performance improvement, rather than focusing on a specific leaf node without looking at sibling nodes, as is the case in CART. This global criterion still uses a greedy algorithm, but is less myopic and helps prevent overfitting.

All subsequent splits are determined similarly. Algorithm 1 summarizes the entire tree-growing procedure.One natural turning parameter of the tree-growing process is the number of leaves. We consider a tree with $J+1 = 20$ leaf nodes in the baseline specification.[12] Once the tree growing process terminates, then $f_t^{(J)}$ and $\beta^{(J)}(\cdot)$ are final outputs. We discuss how one can create multi-factor P-Trees or incorporate exogenously given pricing factors in Section 2.4.

[12]Our data consist of 3,000 to 7,000 stocks per month. We set the condition of the minimal leaf size to be 10. Leaf nodes that cannot satisfy minimal leaf size are not further split. We demostrate model robustness in the Internet Appendix I for trees with different number of leaves (e.g., 15 an 25).

## 2.3 Incorporating SDF Objectives into Global Split Criteria

A fundamental theorem in asset pricing is the duality between the mean-variance efficient (MVE) portfolio and the stochastic discount factor (SDF). The minimal variance of the SDF equals the maximal squared Sharpe ratio of the MVE portfolio (Hansen and Jagannathan, 1991). Correspondingly, explaining the cross section and maximizing the Sharpe ratio are direct and indirect objectives of constructing the SDF. The global split criteria of the P-Tree can incorporate either asset pricing objective.

Let us first illustrate the duality of asset pricing no-arbitrage restriction and the maximal Sharpe ratio of investment for the SDF. We start with the minimum-variance SDF in the span of $N$ individual asset excess returns, $\mathbf{r}_t = [r_{1,t}, ..., r_{N,t}]^\intercal$,

$$m_{t+1} = 1 - \mathbf{w}_t^\intercal(\mathbf{r}_{t+1} - \boldsymbol{\mu}_t), \tag{8}$$

where $\boldsymbol{\mu}_t = E_t[\mathbf{r}_{t+1}]$ represents the conditional expectation of excess returns. Plugging in the linear SDF in (8) into the no-arbitrage restriction $E_t[m_{t+1}\mathbf{r}_{t+1}] = 0$ yields the solution to SDF loading $\mathbf{w}_t$,

$$\mathbf{w}_t = \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t, \tag{9}$$

where $\boldsymbol{\Sigma}_t = \text{Cov}_t(\mathbf{r}_{t+1})$ is the conditional covariance matrix of excess returns. Plugging it into (8) yields the conditional variance of SDF,

$$\text{Var}_t(m_{t+1}) = \boldsymbol{\mu}_t^\intercal\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t, \tag{10}$$

which equals the maximal conditional squared Sharpe ratio of the tangency portfolio. Therefore, the tangency portfolio weights equal the SDF parameter in (9). A solution of portfolio weights to maximize the squared Sharpe ratio also satisfies the no-arbitrage restriction for the SDF. However, the estimation of $\Sigma_t$ and $\mu_t$ for a large number of individual assets is unfeasible (Cochrane, 2014). One motivation of the P-Tree is to reduce the dimension of the individual assets and to create a small number of leaf basis portfolios, which can be considered with two types of split criteria. The below text refers P-Tree models with asset pricing or investment criteria P-Tree or Investment P-Tree, respectively.

**Asset pricing criterion.** The P-Tree algorithm in Section 2.2 illustrates the asset pricing criterion under no arbitrage. The direct objective of constructing an SDF is to explain the cross section of returns and minimize the loss of aggregate pricing errors. Thus the tree growing algorithm searches for split rule candidate that minimizes

$$\mathcal{L}_{\mathrm{A}}(\tilde{\mathbf{c}}_{k,m}) = \sum_{t=1}^{T}\sum_{i=1}^{N_t}\Big(r_{i,t} - \beta_{i,t-1}(\mathbf{z}_{i,t-1})f_t\Big)^2, \tag{11}$$

where $r_{i,t}$ is the excess return for stock $i$ at period $t$, $f_t$ is the $P \times 1$ traded factor return, and $\beta_{i,t-1}$ is the factor loading for stock $i$. With this criterion, P-Tree solves a joint problem for constructing the latent factors and estimating betas.

The realized pricing errors are the return residuals from the traded factor model. P-Tree generates latent traded factors to reduce the aggregate realized pricing errors. If we adopt this criterion for growing the P-Tree, it minimizes the aggregate realized pricing errors and thus follows the *no-arbitrage restriction*. Chen et al. (2022) and Feng et al. (2021) share the same modeling objective by constructing latent factors via deep learning. Including a benchmark factor (e.g., market factor) in $f_t$ is also possible and discussed in Section 2.4. For U.S. equities, we use P-Tree to generate latent factors to complement the benchmark factor model to reduce pricing errors.

**Investment criterion.** A traded SDF is supposed to be the tangency portfolio with the maximal Sharpe ratio. Therefore, an indirect objective of recovering the SDF and growing the P-Tree is to maximize (minimize) the (negative) squared Sharpe ratio following (10):

$$\mathcal{L}_{\mathrm{I}}(\tilde{\mathbf{c}}_{k,m}) = -\boldsymbol{\mu}_{\mathbf{F}}'\boldsymbol{\Sigma}_{\mathbf{F}}^{-1}\boldsymbol{\mu}_{\mathbf{F}}, \tag{12}$$

where $\mathbf{F} = f_t$ is the latent factor generated by the P-Tree, with mean and covariance $\boldsymbol{\mu}_F$ and $\boldsymbol{\Sigma}_{\mathbf{F}}$ respectively. Considering this alternative global criterion implies a joint problem of constructing the latent factors and estimating efficient portfolio weights. The portfolio optimization also requires a special design to accommodate the panel structure of asset returns, which is not available for standard ML methods. Including benchmark factors (e.g., market factor) in $\mathbf{F}$ is also feasible (see Section 2.4). Thus, $\mathbf{F} = [\mathbf{f}_{\text{benchmark},t}, f_t]$. The indirect construction of the latent factor relies on the squared Sharpe ratio improvement over benchmark factors (Barillas and Shanken, 2017).

## 2.4 Boosted P-Trees for Multiple Factors

P-Tree models offer practical ways to build a multi-factor model additionally. To do so, one can use boosting, an ML technique to combine an ensemble of weak learners to form a strong learner (Freund and Schapire, 1997). It grows a list of additive trees sequentially, each time generating a latent factor to augment factors generated by all previous trees. We propose the boosting split procedures as follows:

1. The first factor $f_{1,t}$ is generated by the standard P-Tree as discussed in Section 2.2. Then, we keep $f_{1,t}$ and its corresponding factor loadings $\beta_1(\cdot)$.

2. The second factor is generated to augment the first one. The tree-growing steps are the same, except we use the boosting split criteria:

$$\mathcal{L}(\tilde{\mathbf{c}}_{k,m}) = \sum_{t=1}^{T}\sum_{i=1}^{N_t}\left(r_{i,t} - \beta_1(\mathbf{z}_{i,t-1})f_{1,t} - \beta_2(\mathbf{z}_{i,t-1})f_{2,t}\right)^2.$$

   Note that $\beta_1(\cdot)$ is fixed from step 1, then $r_{i,t} - \beta_1(\mathbf{z}_{i,t-1})f_{1,t}$ represents the unexplained pricing error of the previous factor model. The second factor is generated to augment the first one. Then, we keep $f_{2,t}$ and its corresponding factor loadings $\beta_2(\cdot)$.

3. The third factor $\mathbf{f}_{3,t}$ is generated similarly to augment the first two factors, whereas

factor loadings for the first two factors are fixed at previous estimations.

$$\mathcal{L}(\tilde{\mathbf{c}}_{k,m}) = \sum_{t=1}^{T}\sum_{i=1}^{N_t}\Big(r_{i,t} - \beta_1(\mathbf{z}_{i,t-1})f_{1,t} - \beta_2(\mathbf{z}_{i,t-1})f_{2,t} - \beta_3(\mathbf{z}_{i,t-1})f_{3,t}\Big)^2.$$

4. Repeat the above process $K$ times to generate $K$ factors $\mathbf{f}_t = [f_{1,t}, \cdots, f_{K,t}]$ and obtain corresponding factor loadings $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_K]$.

The boosted P-Tree factors can augment any existing factor model. Suppose we fit the P-Tree model with $K = 3$ factors. The three factors can be listed in decreasing order of importance as $[f_{1,t}, f_{2,t}, f_{3,t}]$. Generating additional P-Tree factors is similar to generating additional principal components and each subsequent factor to help reduce the remaining pricing errors. For example, a three-factor boosted P-Tree model follows:

$$\mathcal{T}(\mathbf{z}_{i,t-1}|\Theta) = \beta_1(\mathbf{z}_{i,t-1})f_{1,t} + \beta_2(\mathbf{z}_{i,t-1})f_{2,t} + \beta_3(\mathbf{z}_{i,t-1})f_{3,t}.$$

**P-Tree on benchmark factors.** Another natural boosting application is creating an augmented factor model on benchmark factor models. Feng et al. (2021) start with a benchmark model such as CAPM or Fama-French factor models and add latent factors generated by deep learning. P-Tree factors can also be generated to augment a benchmark model $\mathbf{f}_{\text{benchmark},t}$ to explain the "unexplained" information from the benchmark factors.

$$\mathcal{L}(\tilde{\mathbf{c}}^{(j)}) = \sum_{t=1}^{T}\sum_{i=1}^{N_t}\Big(r_{i,t} - \boldsymbol{\beta}_0(\mathbf{z}_{i,t-1})^\intercal\mathbf{f}_{\text{benchmark},t} - \beta_1(\mathbf{z}_{i,t-1})f_{1,t}\Big)^2.$$

For the investment criterion, it simply expands $\mathbf{F} = [\mathbf{f}_{\text{benchmark},t}, f_t]$ and calculates its negative squared Sharpe ratio as in (12). Boosted P-Tree factors can be added similarly. Given the central role of the market factor, which is not explained by ranked characteristics, we use the market-adjusted P-Tree for our baseline empirical results.[13]

[13]Complementing the market-adjusted P-Tree (Figure 5), we present a plain-vanilla P-Tree in Figure A.1.

**Boosted Investment P-Trees.** The boosting procedures are straightforward under the investment criterion. Additional P-Tree factors are generated to maximize the Sharpe ratio for the tangency portfolio on existing factors. However, the original boosting idea for unexplained residuals might not apply to this sequential boosted investment strategy, because every additional P-Tree factor updates the weights for the tangency portfolio. The main application of this investment P-Tree is to perform supervised dimension reduction from the unbalanced panel of individual stock returns to a few P-Tree factors for improving the existing investment portfolio. Feng et al. (2022) construct latent factors generated by deep learning to complement or hedge benchmark assets, such as market factors for corporate bonds or equities. Our P-Tree framework is similarly flexible for generating additional factors to complement and hedge benchmark factors.

# 3 An Empirical Implementation of P-Tree on U.S. Equities

The P-Tree framework is easily implementable on the panel of U.S. equity data or corporate bond data. The baseline P-Tree model trained on U.S. public equity data runs about 1.5 hours on a server with an Intel Xeon Gold 6230 CPU, for a training data set with 61 characteristics and 1.3 million observations. We describe the details next and report another implementation on the U.S. corporate bond data in Internet Appendix III.

## 3.1 Data

**Equity data and Characteristics.** The standard filters (e.g., same as in Fama-French factor construction) are applied to the universe of U.S. equities: (1) include only stocks listed on NYSE, AMEX, or NASDAQ for more than one year; (2) use those observations for firms with a CRSP share code of 10 or 11; and (3) exclude stocks with negative book equity or lag market equity. We use 61 firm characteristics with monthly observations for each stock, covering six major categories: momentum, value, investment, profitability, frictions (or size), and intangibles. Table A.1 lists these input variables. We standardize the characteristics cross-sectionally in the range $[-1, 1]$.[14]

The sample period ranges from January 1981 to December 2020. We use the first 20 years for training and the latter 20 years for testing. The average and median monthly numbers of stock observations are 5,265 and 4,925 in the training sample and 4,110 and 3,837 in the testing sample, respectively. For individual stock returns in the training sample, we perform a cross-sectional winsorization on 1% and 99% to reduce the outlier effects. We perform no winsorization for the observations in the test sample.

**Macro predictors.** We use ten macro predictors for consideration of time-series splits. Table A.2 summarizes the macro predictors, including macroeconomic variables, bond market predictors, and aggregate characteristics for S&P 500. Macro predictor data are uniformly standardized by the historical percentiles of the rolling window of the past 10 years. For example, inflation greater than 0.7 implies the current inflation level is higher than 70% of observations of the past 10 years. This rolling window data standardization is useful when comparing the predictor level to detect macroeconomic regime switches.

## 3.2 Splitting the Cross Section

We use the market-adjusted asset pricing criterion as our baseline empirical specification, and report complementary results using the investment criterion. Figure 5 plots the market-adjusted P-Tree. In each leaf node, S# is the order of splits, and N# is the node index. The numbers printed in the leaves are the median number of observations in the monthly updated leaf basis portfolios. The split rules are the selected splitting characteristics and cross-sectional quintile cutpoints $[-0.6, -0.2, 0.2, 0.6]$.[15] Before the first split,

[14]For example, market equity values in December 2020 are uniformly standardized into $[-1, 1]$. The firm with the lowest value is -1, and the highest value is 1. All others are distributed uniformly in between. Missing values are imputed as 0, implying the firm is neutral in the security sorting on the characteristic.

[15]This follows conventional univariate security sorting, which uses quintiles to maintain sufficient observations in each sorted portfolio. Our R package allows for finer grid searches and flexible specifications.

the tree grows from the root node (N1), whose leaf basis portfolio represents the value-weighted market factor.

**Growth, splits, and stop.** Selecting tuning parameters to balance the model fitness and complexity is crucial in ML. We need to determine when to stop the P-Tree growth. To maintain interpretability and avoid overfitting, we advocate a natural tuning parameter, the number of leaves, to limit the tree size.[16] We report the results for the case of 20 final leaves. Our findings are robust to other tree sizes, such as 15 and 25.

Figure 5 displays a market-adjusted P-Tree, which stops growing after 19 splits.The data-driven P-Tree first splits along the idiosyncratic volatility over Fama-French three factors (`RVAR_FF3`) at 0.6 (80% quantile), which is related to characteristics including size and liquidity. After this split, 80% of the stocks go to the left leaf (labeled N2), and 20% go to the right (N3). The second split is on the earnings-to-price ratio (`EP`) at -0.6 of the right leaf (N3), whereas the third split is on the volatility of turnover (`STD_TURN`) under the high-`RVAR_FF3` and low-`EP` node (N6).

**Partition, clusters, and leaf basis portfolios.** P-Tree is helpful for clustering similar assets under the economic-guided split criterion by their underlying characteristics. Figure 5 perfectly illustrates the (asymmetric) characteristics interactions for splitting the cross section. By jointly defining the partition corresponding to the leaf node, P-Tree learns the interaction of characteristics appearing in the same path. For instance, `BASPREAD` is useful for low-`RVAR_FF3` stocks, such that liquidity might be a better indicator for clustering low-volatility stocks. However, for high-`RVAR_FF3` stocks, `EP` is useful because value might be a better indicator for clustering high-volatility stocks.

The partition plot in Figure 7 is an alternative way to visualize these clustering patterns. We report the monthly average excess returns and annualized Sharpe ratios for leaf

[16]Other conditions for controlling tree growth can be easily added to the framework. For example, one can require the minimal leaf portfolio size to be 10, although it turns out to never bind in our empirical exercise. As Figure 5 shows, the median portfolio size is well above 40 when the growth stops.

basis portfolios. These performance gaps show the usefulness of splitting the cross section via the nonlinear interaction of characteristics. The partition plot perfectly illustrates the nonlinear and asymmetric interaction between EP and RVAR_FF3. When used to split the cross section, the EP only matters for high idiosyncratic volatility (RVAR_FF3) stocks. In the training sample, the low-RVAR_FF3 partition portfolio has a 60% annualized Sharpe ratio, but the high-RVAR_FF3 low-EP partition portfolio has a -87% value.

## 3.3 Random P-Forest for Variable Importance

Random forest is another prominent ensemble approach that fits multiple trees to reduce prediction errors.[17] We deploy this strategy to measure the usefulness of characteristics for the P-Tree growth. We refer to our approach as a random P-Forest. First, we bootstrap the data on the time-series dimension with replacement. We preserve the complete cross section for the selected periods to exploit the low serial correlations of returns, which is a crucial assumption of random forest. Second, we randomly draw 20 characteristics out of 61 for each subsample. Third, we independently grow a P-Tree on each bootstrap sample and repeat the procedure 1,000 times. Any characteristic is considered about 330 times out of 1,000 subsamples for fitting the P-Forest.

**Quantifying importance.** Our empirical exercises focus on the out-of-bag (OOB) variable importance and define two measurements. The first measurement of variable importance counts the frequency of one variable being selected by splitting. Intuitively, the more often a characteristic is selected as a split rule candidate, the more important it is. For each bootstrap sample, a subset of characteristics is randomly drawn to grow the tree, and only a fraction of them are actually selected as split rules. We count the number of times a particular $l$-th characteristic $z_l$ used in the first $J$ splits and the total number of

[17]The idea of random forest is bagging, where sampling data from empirical distribution approximates sampling from the true underlying distribution, thus enabling quantification of the estimation uncertainty and variable importance.

appearances in bootstrap subsamples. We define the first measurement of importance as

$$\text{Selection Probability}(z_l) = \frac{\#\left(z_l \text{ is selected at first } J \text{ splits}\right)}{\#\left(z_l \text{ appears in the bootstrap subsamples}\right)}. \tag{13}$$

The second measurement aims to capture a feature's "treatment effect". For the 1,000 bootstrap subsamples, we know a characteristic is randomly included in a subsample by following a Bernoulli distribution. Even if one characteristic is included in a subsample, it is not necessarily selected as a split rule. The with- and without-sampling scheme for a particular characteristic creates the treatment effect evaluation and the significance evaluation of its importance. We compute the loss function reduction for the asset pricing criterion in (11) or the investment criterion in (12) as the characteristic importance:

$$\text{Char. Importance}(z_l) = \left[\frac{E\left(\text{loss function} \mid \text{with } z_l\right)}{E\left(\text{loss function} \mid \text{without } z_l\right)} - 1\right] \times 100. \tag{14}$$

Panel A of Table 3 summarizes the selection frequency in (13) of being the top splitting characteristics. We find idiosyncratic volatility (`RVAR_FF3`) has a 40% chance to be the first split characteristic once included in the bootstrap sample. Other high selections are highly correlated, including IVOL of CAPM (`RVAR_CAPM`), return volatility (`SVAR`), and market equity (`ME`). The second split characteristic in Figure 5 is the earnings-to-price ratio (`EP`), for which we also find a high selection probability, as well as its similar value measure, the cash flow-to-price ratio (`CFP`). The two metrics reveal similar important characteristics for growing the P-Tree. These characteristics are essential for capturing the asymmetric nonlinear interactions and do not overlap with the important ML return predictors (Gu et al., 2020), which only focuses on return prediction.

Figure 8 summarizes the variable importance in a market-adjusted random P-Forest. According to (14), a negative value implies a reduction in model performance when we excludethis characteristic. We conduct a two-sample $t$-test using the bootstrap samples and plot those significantly useful characteristics in a deeper color. Only seven out of 61 characteristics show statistical significance. The in-sample results identify the significant characteristics such as return and idiosyncratic volatility (SVAR, RVAR_FF3, RVAR_CAPM), analysts' earning revisions (RE), bid-ask spread (BASPREAD), corporate investment (CINVEST), and dollar trading volume (DOLVOL). The return and idiosyncratic volatility characteristics also show consistently significant asset pricing improvement for the test sample in the bottom plot of Figure 8, so do RE, BASPREAD, and CINVEST.

**Interpretable non-ensemble P-Tree without overfitting.** Visualizing characteristics of asymmetric nonlinear interactions gives P-Tree models excellent interpretability. A nonlinear tree structure and the interactions of characteristics are displayed when the decision tree continues splitting further. However, these patterns gleaned from the single P-Tree are only helpful if the P-Tree model can avoid overfitting. We use the aforementioned characteristic importance metrics to examine if a P-Tree splits on similar characteristics without overfitting. Specifically, we compare important characteristics from the random P-Forest with those revealed in our single P-Tree model. If they are similar, the P-Tree model with global split criteria behaves similarly to the less interpretable random P-Forest that typically achieves better out-of-sample performance.

The results are indeed consistent. Those important characteristics from random P-Forest, namely, idiosyncratic volatility (RVAR_FF3) and the earnings-to-price (EP), are also the characteristics used in the first two splits of the P-Tree (Figure 5). We deduce that P-Tree models behave similarly to random P-Forest and are not subject to overfitting. Yet, the nonlinearities and interactions revealed in P-Tree models provide helpful information that is impossible to glean directly from ensemble models.

# 4 Performance and Further Applications

We now report the performance of P-Tree models for pricing individual stocks and test portfolios and constructing investment strategies. We then discuss how P-Tree helps uncover nonlinear patterns in the cross-section of asset returns and enhance trading strategies, all while maintaining interpretability. Finally, we illustrate how P-Tree augments tree models by allowing both time-series and cross-sectional splits and incorporating interactions of macroeconomic variables (and with asset-specific characteristics).

## 4.1 Asset Pricing Performance

**Performance metrics.** We follow the literature (e.g., Feng et al., 2021) to include multiple performance metrics. Total $R^2$ and cross-sectional $R^2$ evaluate economic asset pricing performance for variation in time-series and cross-sectional dimensions. Specifically,

$$\text{Total } R^2 = 1 - \frac{\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(r_{i,t} - \widehat{r}_{i,t}\right)^2}{\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} r_{i,t}^2}, \tag{15}$$

where $\widehat{r}_{i,t} = \widehat{\beta}(\mathbf{z}_{i,t-1})f_t$. Total $R^2$ represents the fraction of realized return variation explained by the model-implied contemporaneous returns.

$$\text{Stock CS } R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}(r_{i,t} - \widehat{r}_{i,t})\right)^2}{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} r_{i,t}\right)^2}, \tag{16}$$

where $\widehat{r}_{i,t} = \widehat{\beta}(\mathbf{z}_{i,t-1})f_t$. Cross-sectional $R^2$ for individual stocks represents the fraction of the squared unconditional mean returns described by the common factor model.

$$\text{Portfolio CS } R^2 = 1 - \frac{\sum_{i=1}^{N}\left(\bar{r}_i - \widehat{\bar{r}}_i\right)^2}{\sum_{i=1}^{N}\bar{r}_i^{\,2}}, \tag{17}$$

where $\widehat{\bar{r}}_i = \widehat{\boldsymbol{\beta}}_i^{\intercal}\widetilde{\boldsymbol{\lambda}}_f$. The risk premium estimation adopts the cross-sectional regression estimates of $\bar{r}_i$ on $\widehat{\boldsymbol{\beta}}_i$. The portfolio cross-sectional $R^2$ represents the fraction of assets' average excess returns explained by the model-implied expected returns.

**Comparing asset pricing models.** Table 1 summarizes the P-Tree performance under the asset pricing criterion. Panel A shows two P-Tree models on the market benchmark, plus one and four P-Tree factors. The additional number of P-Tree factors generated by boosting is selected by the squared Sharpe ratio test of Barillas and Shanken (2017). Panel B of Table 1 compares these results with observable factor models such as CAPM, Fama-French five-factor model (Fama and French, 2015), and the recent Q5 model of Hou et al. (2021). We also implement latent factor models from the ML finance literature, such as RP-PCA of Lettau and Pelger (2020b) and IPCA of Kelly et al. (2019). Panel C reports the standard P-Tree factor models. Panels D and E report the results of the P-Tree with time-series splits on market volatility and inflation, which we discuss in Section 4.4.

Both P-Trees with and without the market factor show consistently positive values of Total $R^2$ in pricing individual stock returns, outperforming most observable factor models both in-sample and out-of-sample. The multifactor P-Tree models show comparable Total $R^2$ to two strong latent factor models. Cross-sectional $R^2$ for individual stocks quantifies the aggregate unexplained average pricing errors by the factor model. For the cross-sectional $R^2$, the multifactor P-Tree models produce highly positive numbers for in-sample and out-of-sample analysis, which demonstrates the performance of the model-implied expected returns. These numbers are close to the strong IPCA benchmark.

Next, we investigate the P-Tree performance on pricing multiple sets of test portfolios. In this case, cross-sectional $R^2$ in (17) quantifies the aggregate unexplained pricing errors by the model-implied expected returns and is calculated using the entire sample of data. We consider Fama-French ME-B/M $5 \times 5$ and 49 industry portfolios. In addition, 20 leaf basis portfolios are generated by P-Tree in Figure 5 and P-Tree generates another 20 without the market factor in Figure A.1. Generating basis assets is another data-driven output from our P-Tree framework, which resolves the multiple characteristics challenge and follows a chosen economic-guided split criterion.

Observable factor models, including CAPM, FF5, and Q5, explain well the cross-sectional average excess returns for 25 ME-B/M and 49 industry portfolios. However, they cannot explain the leaf basis portfolios generated by P-Trees. Lewellen et al. (2010)

question that many factors are only tested on 25 ME-B/M portfolios with a strong factor structure, suggesting using more new test assets. Our P-Tree factor models in Panel A consistently price all sets of test assets. The results improve when considering the time-series split, as seen in Panels D and E. Though the one-factor P-Tree does not work well for the 20 leaf basis portfolios, the boosted four-factor P-Tree does. Among other latent factor models, IPCA stands out as competing with robust positive performances. But P-Trees offer more transparency and interpretability than other latent factor models.

## 4.2 Investment Performance

P-Tree is a flexible framework that reduces the dimension of thousands of individual stocks into a few latent factors for asset pricing or investment split criterion. Because P-Tree factors are traded portfolios, we can assess the investment performance of P-Tree factors for evaluating their in-sample and out-of-sample model fitness. Though our empirical studies generate P-Trees on the monthly characteristics, its rebalancing frequency can be lowered to quarterly or annual to reduce transaction costs.

On the one hand, though the asset pricing P-Tree is grown for fitting cross-sectional returns, it should also deliver a high Sharpe ratio because it generates an SDF. On the other hand, the investment P-Tree follows the squared Sharpe ratio split criterion and is supposed to provide excellent investment performance. Figure 6 plots investment P-Tree on the market benchmark and finds `RVAR_FF3` is again consistently the first characteristic for splitting the cross section. The growth of these two P-Trees reveals the duality of asset pricing and risk-adjusted investment performance of an SDF.

We compare the risk-adjusted investment performances of these observable or latent factors. We consider two common strategies for investing multiple factors: the $1/N$ equal-weighted portfolio and the MVE portfolio.[18] Table 2 reports the monthly average return, Jensen's alpha, and the annualized Sharpe ratio of the training and test samples.

[18]We follow (3) to calculate portfolio weights and rescale their absolute sum to 1.

The asset pricing and investment P-Tree factor models deliver exceptionally positive performances. For the in-sample analysis, the MVE strategies of the five-factor model (market plus four boosted P-Tree factors) deliver a 3.47 and 12.55 annualized Sharpe ratio, in Panels A and B. Their monthly Jensen's alphas are beyond 1%. These numbers are consistently high out of the sample. In Panels B and C, the out-of-sample MVE strategies of two investment P-Tree five-factor models offer annualized Sharpe ratios close to 3, plus significant monthly alphas of about 0.80%. If we focus on the long-only $1/N$ strategies, the numbers become lower but are still consistently positive. Overall, these P-Trees and the five-factor IPCA perform well above other models. The portfolio construction, variable importance, and interactions of characteristics are more transparent in P-Tree models.

We decompose each P-Tree factor on benchmark models to evaluate the additional investment information, alphas, in Table 4. For example, the first investment P-Tree factor `RVAR_FF3-ABR` in Panel B shows economically and statistically significant alphas against FF5 and Q5 models. A 5-factor IPCA model cannot explain this top P-Tree factor for either in-sample or out-of-sample evaluation. The efficient portfolio of four investment P-Tree factors plus the market factor still shows significant alphas everywhere. Moreover, the risk-adjusted investment performances for asset pricing P-Tree factors in Panel A are slightly lower but still consistently positive.

P-Tree factors are generated on leaf basis portfolios—the data-driven basis assets generated by P-Trees on multiple characteristics. We also compare the investment performance on different basis portfolios, such as conventional clustering ($5 \times 5$ ME-B/M and 49 industry portfolios) in Panel C, which shows lower investment performance. We can conclude these positive performances of investment P-Tree factors come from both the interaction and nonlinearity of characteristics and the data-driven way of identifying these characteristics (as opposed to ad-hoc sorting).

**Hedging benchmark Portfolios.** The generated P-Tree latent factor complements any benchmark model for investment purposes too. Note that the objective of the Sharpe ratio is not about the performance of the generated P-Tree factors, but the tangency portfolio formed together with the benchmark model. The economically-guided P-Tree factor plays two fundamental roles: (i) it should have a low or even negative correlation with the benchmark factor, providing a potentially hedging portfolio; and (ii) when the benchmark factor alone cannot span the efficient frontier, the P-Tree factor should enter the SDF if it helps to improve the multi-factor portfolio efficiency (Fama, 1996), which implies potentially missing factors from a large number of characteristics. Furthermore, the investment P-Tree factors rely on the squared Sharpe ratio improvement over the benchmark without using any test assets, sharing the same spirit with Barillas and Shanken (2017).

This hedging perspective of boosted P-Tree factors on the investment criteria differs from the multi-factor model on the asset pricing P-Tree. The statistical boosting scheme forms a strong learner by adding orthogonal weak learners, which works perfectly for risk-adjusted investment for diversification. The innovation of P-Tree is to provide a dimension reduction strategy for investing individual assets to hedge or complement benchmark assets. In addition to equity risk factors, benchmark assets can include corporate bonds, treasury bonds, or commodities.

Figure 9 shows the annualized returns of the market-adjusted investment P-Tree and market factors. We find P-Tree factor has positive returns for every year in the training sample, and only one slightly negative year in the test sample. However, the market factor is particularly positive that year and the tangency portfolio is still good. The purpose of hedging is to generate a positive return when the market is down. Over the past four decades, the market has had negative returns for ten years, yet the P-Tree factor perfectly plays the role of a hedge. Therefore, the finding of persistently high performance—an annualized Sharpe ratio of 2.78 and a monthly alpha of 1.10% for the recent two decades—for the efficient portfolio combining two factors should not be a surprise.

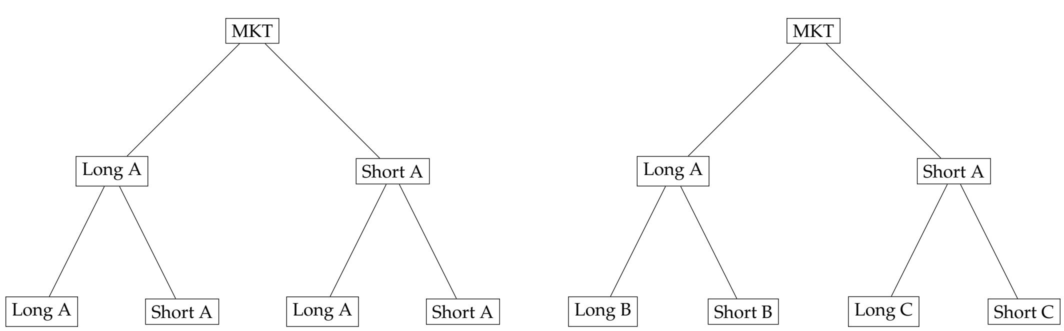
Finally, we report in Internet Appendix II how P-Tree asset pricing and investment performances are robust under economic restrictions such as the exclusion of small stocks.

## 4.3 Nonlinearity, Interactions, and Interpretability

P-Tree essentially generalizes (sequential) sorting to capture nonlinear patterns. The information can be used to understand characteristic interactions, resurrect anomalies, and enhance other trading strategies.

**Sorted Factors.** Economic theory and extant empirical literature inform us of the long and short directions for constructing the portfolios, such as small-minus-big or high-minus-low. As shown in Figure 4, two types of characteristics-sorted factors follow the original directions for long-short positions. Specification (a) is the standard univariate sorting scheme. Specification (b) is the interaction factors generated by our P-Tree models for a sequential sorting scheme. As for P-Tree specification, we consider both the market-adjusted asset pricing P-Tree and the market-adjusted investment P-Tree to generate interaction factors. For each characteristic, we train a P-Tree and restrict the first split on that characteristic, the second and third split excluding that characteristics, and the maximal depth is three. We restrict all splitting cutpoints to be the median for this example and fit these interaction factors using the training sample.

Figure 4: Specifications of Characteristics-Sorted Factors




(a) Uni-Sort 4x1 (b) P-Tree Interaction

We summarize the number of significant positive risk premia and alphas cases in Table 5, Panels A and B, at 10% and 5% significance levels. First, more than half of univariate-sort factors feature significant risk premia and alphas at the 10% level in the entire sample. Second, fewer significant factors show in the test sample than in the training sample, which might reflect the post-publication bias (McLean and Pontiff, 2016). Third, our market-adjusted asset pricing and investment P-Tree interaction factors have more significant alphas than univariate-sort factors, which reflect incremental information gains beyond the market. Last, the market-adjusted investment P-Tree interaction factors show exceptional risk premia and alphas in and out of the sample. These empirical findings confirm the advantages of P-Tree that extract asymmetric characteristics interactions beyond the conventional sorting.

**Enhancing and resurrecting anomalies through interactions.** The information on asymmetric nonlinearity and interaction provides new economic insights and can be used to construct profitable trading strategies. We analyze how these interaction factors behave using six examples and report the findings in Table 6 and Figure A.6. In short, interaction factors enhance investment performance over univariate-sort ones. For example, the univariate-sort factor for standardized unexpected earnings (`SUE`) delivers a positive risk premium. When we create an interaction factor with industry-adjusted B/M ratio (`BM_IA`) and dollar trading volume (`DOLVOL`), the performance almost doubles. `BM_IA` and `DOLVOL` are useful for the low and high `SUE` leaf portfolios, respectively. Figure A.6 shows the average excess returns for two legs during splitting and finds decreasing and increasing values in the corresponding direction. This important insight reveals the need for an asymmetric model of factor construction. Whereas Jarrow et al. (2021) point out that long and short legs have different underlying risk factors, and thus reflect different risks, our findings highlight how long and short legs can interact with different characteristics asymmetrically. Dividend yield (`DY` in Panel C) and change in profit margin (`CHPM` in Panel E) are also examples of enhancing factors through interaction.

Utilizing the information in asymmetric characteristics interactions might help the post-publication decay problem. Once we include the "correct" control or interaction, some of these factors or anomalies can be resurrected. For example, the univariate-sort factor for maximum daily returns (MAXRET) has a significant premium in the training sample but disappears in the test sample. Interacting with industry-adjusted size (ME_IA) and abnormal returns around earnings announcement (ABR) in the long and short legs, this interaction factor earns 0.67% for monthly average returns and 1.11% for alpha for the out-of-sample period. Growth in long-term debt (LGR) in Panel D and momentum 6-month (MOM6M) in Panel F are also resurrecting examples.

### 4.4 P-Tree for Panel Data: Cross Section + Time Series

Panel data of asset returns have both cross-sectional and time-series dimensions. Most of the literature focuses on explaining the cross-sectional return variation, as do the P-Tree models. Since the tree grows from top to bottom, the panel of assets is partitioned into many leaf nodes based on past characteristics. Each leaf node maintains the entire time series from period 1 to $T$ but a subset of assets. However, empirical findings show the factor choices, their risk premia, and even betas can be time-varying under different macroeconomic states. For example, a large literature examines regime changes in financial markets involving periods of high and low volatility or inflation, booms or recessions (e.g., Hamilton and Lin, 1996; Buraschi and Jiltsov, 2005).

P-Tree enables splits in the time-series dimension and provides an alternative approach to factor-beta construction and estimation with regime-switching. Moreover, the time-series information extracted from macroeconomic variables proves helpful in constructing the dynamic SDF. We consider creating time-varying factor models and estimating corresponding conditional betas under switching macroeconomic states, which can be achieved by splitting nodes (e.g., the root node) by macroeconomic variables instead of firm characteristics. Each child node maintains the same cross section of assets as the parent node, whereas their time-series samples do not overlap. This setting improves P-Tree's flexibility to capture cross-sectional and time-series variations, which performs clustering over the two-dimension panel.

The asset pricing split criterion for the first time-series split is

$$\mathcal{L}(\tilde{\mathbf{c}}) = \sum_{t\in\mathcal{R}_A}\sum_{i\in N_{A,t}} \left(r_{i,t} - \beta_A(\mathbf{z}_{i,t-1}) f_{A,t}\right)^2 + \sum_{t\in\mathcal{R}_B}\sum_{i\in N_{B,t}} \left(r_{i,t} - \beta_B(\mathbf{z}_{i,t-1}) f_{B,t}\right)^2, \tag{18}$$

where the split rule candidate $\tilde{c}$ partitions the time series of data into two non-overlapped sets of time periods $\mathcal{R}_A$ and $\mathcal{R}_B$, for example, high or low inflation states. Note that both sets of periods are not necessarily to be consecutive. Let $N_{A,t}$ denote the subsample of assets in the child node $\mathcal{R}_A$ in time period $t$, and $N_{B,t}$ proceeds similarly. Factors $f_{A,t}$ and $f_{B,t}$ are P-Tree factor returns (or market) at each time period in the left or right child node, respectively. Note that there are no overlapping periods between $\mathcal{R}_A$ and $\mathcal{R}_B$. Additionally, $\mathcal{R}_A$ and $\mathcal{R}_B$ might include multiple periods for the changing macroeconomic states. Compared with the cross-sectional split criteria in (5), the time-series split criterion is the *total* pricing loss of two time periods, with two corresponding factors $f_{A,t}$ and $f_{B,t}$. Below is the two-period model with conditional betas:

$$\mathcal{T}(\mathbf{z}_{i,t-1}|\Theta) = \beta_A(\mathbf{z}_{i,t-1}) f_{A,t}\mathbf{I}(t\in\mathcal{R}_A) + \beta_B(\mathbf{z}_{i,t-1}) f_{B,t}\mathbf{I}(t\in\mathcal{R}_B). \tag{19}$$

After searching for the optimal time-series split rule at the root, all subsequent split rules are chosen from cross-sectional characteristics only. Note any further split on either child of the root node only depends on the subsample on one side. This extension also informs how macroeconomic variables interact with firm characteristics to price assets, because they jointly define the partition of the space.

**Regime switching under market volatility or inflation.** Given the limited history of monthly returns, and for illustration, we only implement one time-series split at the root node (i.e., the first split rule of the tree). The first split determines the optimal macro predictor for regime switching. After a complete search, P-Tree splits market volatility (SVAR) at the 60% quantile of the past decade. Figure A.2 shows the time-series split and the two branches of a cross-sectional split. The left and right branches have different tree structures in high and low market volatility periods, implying the tree model adapts to different macro conditions. We also consider another important macro predictor, inflation, as the first split, chosen as the median of the past decade in Figure A.3. Idiosyncratic volatility RVAR_FF3 is selected as the first split characteristic for low-inflation periods, and market equity ME is selected for high-inflation periods.

Table 3 presents variable importance in the selection probability for the time-series split on stock variance and inflation. We find that the characteristics differ in economic regimes, consistent with the findings in Boons et al. (2020). Although those volatility characteristics are most important, the earnings-to-price (EP) is also helpful in highly volatile periods. Liquidity, the number of zero-trading days (zerotrade), is more critical during low-inflation periods than in high-inflation periods, which is consistent with the time-series split tree in Figure A.3, where zerotrade is chosen as the second split. In conclusion, stock variance and inflation are regime-switching indicators, and equity characteristics play different roles in different regimes.

Finally, the asset pricing results for the time-series split market-adjusted asset pricing P-Tree model are summarized in Panels D and E of Table 1, and their investment gains are reported in Table 2. The five-factor time-series P-Tree models have shown persistently high Total and cross-sectional $R^2$ for individual stocks. Furthermore, they provide robustly positive cross-sectional $R^2$ to different test portfolios than the standard P-Tree models and others. In the test sample, the long-only $1/N$ strategies for these five time-series P-Tree factors provide significant monthly alphas of about 0.50%. We can also consider the time-series investment P-Trees, which could deliver higher risk-adjusted investment performance for macro-guided market timing. In summary, by incorporating regime-switching and time-series splits, our P-Tree model can guide market-timing in equity trading and characteristics-based portfolio construction.

# 5 Conclusion

Panel Tree (P-Tree)—a new class of ML/AI models for panel data analysis—innovates upon standard tree-based models by relaxing the i.i.d. assumption on the observations and adopting global split criteria. When applied to asset pricing or investment, P-Tree performs self-supervised dimension reduction from an unbalanced panel of individual stock returns, generates leaf basis portfolios, and finally constructs latent factors for the pricing kernel. Meanwhile, P-Tree preserves the widely recognized advantages of tree-based models, such as transparently displaying asymmetric and nonlinear interactions while accommodating noisy data.

Our empirical study of U.S. equities shows that P-Tree models outperform most observable and latent factor models for pricing individual assets and portfolios. The generated asset pricing and investment P-Tree factors also show robust risk-adjusted investment performance. In addition, the resulting leaf basis portfolios better span the efficient frontier on the market benchmark and serve as useful assets for testing alternative factor models. The asymmetric interaction of characteristics of the P-Tree enhances factor investing performance over the univariate-sort factors. Finally, P-Tree learns regime-switching across macroeconomic states by time-series splits interacting with asset characteristics.

Overall, P-Tree models generalize characteristics-based sorting and capture potential nonlinearities, providing a useful alternative to PCA-based methods with similar empirical performance but maintaining transparency. They also apply to other asset classes, such as corporate bonds and general panel data structures beyond asset pricing.

# References


Ahn, D.-H., J. Conrad, and R. F. Dittmar (2009). Basis assets. *Review of Financial Studies 22*(12), 5133–5174.

Avramov, D., S. Cheng, and L. Metzker (2022). Machine Learning versus Economic restrictions: Evidence from stock return predictability. *Management Science, Forthcoming*.

Bai, J., T. G. Bali, and Q. Wen (2019). Common risk factors in the cross-section of corporate bond returns. *Journal of Financial Economics 131*(3), 619–642.

Bali, T. G., D. Huang, F. Jiang, and Q. Wen (2022). Different Strokes: Return Predictability Across Stocks and Bonds with Machine Learning and Big Data. Technical report, Georgetown University.

Barillas, F. and J. Shanken (2017). Which alpha? *Review of Financial Studies 30*(4), 1316–1338.

Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond risk premiums with machine learning. *Review of Financial Studies 34*(2), 1046–1089.

Boons, M., F. Duarte, F. De Roon, and M. Szymanowska (2020). Time-varying inflation risk and stock returns. *Journal of Financial Economics 136*(2), 444–470.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and regression trees*. Routledge.

Bryzgalova, S., J. Huang, and C. Julliard (2022). Bayesian solutions for the factor zoo: We just ran two quadrillion models. *Journal of Finance, Forthcoming*.

Bryzgalova, S., M. Pelger, and J. Zhu (2020). Forest through the trees: Building cross-sections of stock returns. Technical report, London Business School.

Buraschi, A. and A. Jiltsov (2005). Inflation risk premia and the expectations hypothesis. *Journal of Financial Economics 75*(2), 429–490.

Chen, L., M. Pelger, and J. Zhu (2022). Deep learning in asset pricing. *Management Science, Forthcoming*.

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR.

Chipman, H. A., E. I. George, R. E. McCulloch, et al. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics 4*(1), 266–298.

Cochrane, J. H. (2011). Presidential address: Discount rates. *Journal of Finance 66*(4), 1047–1108.

Cochrane, J. H. (2014). A mean-variance benchmark for intertemporal portfolio theory. *Journal of Finance 69*, 1–49.

Cong, L. W., G. Feng, J. He, and J. Li (2022). Uncommon Factors for Bayesian Asset Clusters. Technical report, City University of Hong Kong.

Cong, L. W., K. Tang, J. Wang, and Y. Zhang (2019). AlphaPortfolio: Direct Construction Trough Deep Reinforcement Learning and Interpretable AI. Technical report, Cornell University.

Creal, D. and J. Kim (2021). Empirical Asset Pricing with Bayesian Regression Trees. Technical report, University of Notre Dame.

DeMiguel, V., A. Martin-Utrera, F. J. Nogales, and R. Uppal (2020). A transaction-cost perspective on the multitude of firm characteristics. *Review of Financial Studies 33*(5), 2180–2222.

Fama, E. F. (1996). Multifactor portfolio efficiency and multifactor asset pricing. *Journal of Financial and Quantitative Analysis 31*(4), 441–465.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Feng, G., S. Giglio, and D. Xiu (2020). Taming the factor zoo: A test of new factors. *Journal of Finance 75*(3), 1327–1370.

Feng, G., J. He, N. Polson, and J. Xu (2021). Deep Learning of Characteristics-Sorted Factor Models. Technical report, City University of Hong Kong.

Feng, G., L. Jiang, J. Li, and Y. Song (2022). Deep Tangency Portfolios. Technical report, City University of Hong Kong.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences 55*(1), 119–139.

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies 33*(5), 2326–2377.

Gagliardini, P., E. Ossola, and O. Scaillet (2016). Time-varying risk premium in large cross-sectional equity data sets. *Econometrica 84*(3), 985–1046.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies 33*(5), 2223–2273.

Gu, S., B. Kelly, and D. Xiu (2021). Autoencoder asset pricing models. *Journal of Econometrics 222*(1), 429–450.

Haddad, V., S. Kozak, and S. Santosh (2020). Factor timing. *Review of Financial Studies 33*(5), 1980–2018.

Hamilton, J. D. and G. Lin (1996). Stock market volatility and the business cycle. *Journal of Applied Econometrics 11*(5), 573–593.

Hansen, L. and R. Jagannathan (1991). Implications of security market data for models of dynamic economies. *Journal of Political Economy 99*, 225–262.

He, J. and P. R. Hahn (2021). Stochastic tree ensembles for regularized nonlinear regression. *Journal of the American Statistical Association, Forthcoming*.

He, J., S. Yalov, and P. R. Hahn (2019). XBART: Accelerated Bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1130–1138.

Hou, K., H. Mo, C. Xue, and L. Zhang (2021). An augmented q-factor model with expected growth. *Review of Finance 25*(1), 1–41.

Hou, K., C. Xue, and L. Zhang (2020). Replicating anomalies. *Review of Financial Studies 33*(5), 2019–2133.

Jarrow, R. A., R. Murataj, M. T. Wells, and L. Zhu (2021). The low-volatility anomaly and the adaptive multi-factor model. Technical report, Cornell University.

Kelly, B. T., D. Palhares, and S. Pruitt (2022). Modeling corporate bond returns. *Journal of Finance, Forthcoming*.

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Kim, S., R. A. Korajczyk, and A. Neuhierl (2021). Arbitrage portfolios. *Review of Financial Studies 34*(6), 2813–2856.

Kozak, S., S. Nagel, and S. Santosh (2018). Interpreting factor models. *Journal of Finance 73*(3), 1183–1223.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

Lettau, M. and M. Pelger (2020a). Estimating latent asset-pricing factors. *Journal of Econometrics 218*(1), 1–31.

Lettau, M. and M. Pelger (2020b). Factors that fit the time series and cross-section of stock returns. *Review of Financial Studies 33*(5), 2274–2325.

Lewellen, J., S. Nagel, and J. Shanken (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial Economics 96*(2), 175–194.

McLean, R. D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *Journal of Finance 71*(1), 5–32.

Roll, R. and S. A. Ross (1980). An empirical investigation of the arbitrage pricing theory. *Journal of Finance 35*(5), 1073–1103.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13*(3), 341–360.

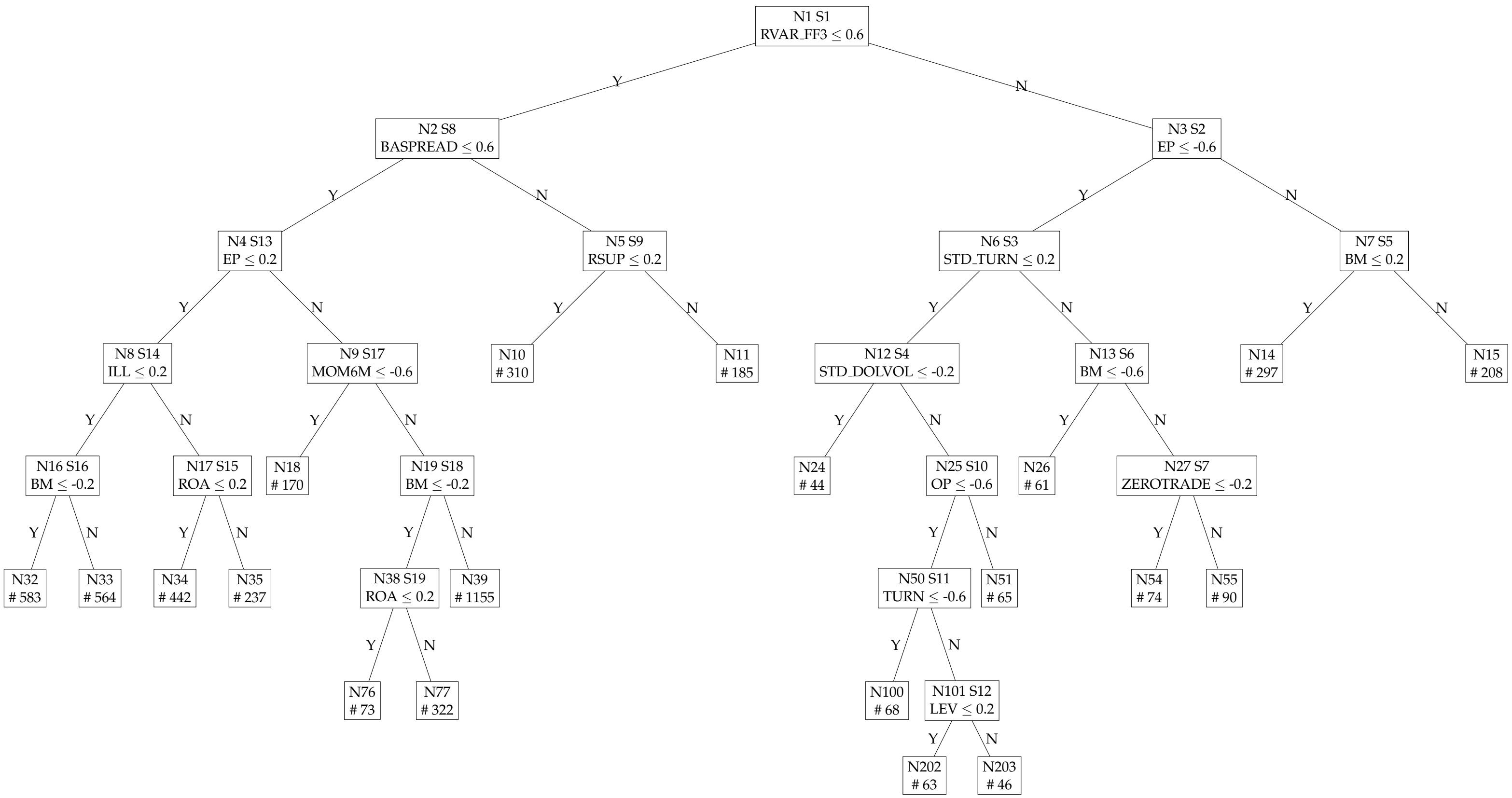
Rossi, A. G. and A. Timmermann (2015). Modeling covariance risk in Merton's ICAPM. *Review of Financial Studies 28*(5), 1428–1461.

Sorensen, E. H., K. L. Miller, and C. K. Ooi (2000). The decision tree approach to stock selection. *Journal of Portfolio Management 27*(1), 42–52.

Zhu, L., S. Basu, R. A. Jarrow, and M. T. Wells (2020). High-dimensional estimation, basis assets, and the adaptive multi-factor model. *Quarterly Journal of Finance 10*(04), 2050017.

# Figure 5: Market-Adjusted Panel Tree

This figure displays the market-adjusted asset pricing P-Tree trained from 1981 to 2000. The first factor is the market factor, and then we apply P-Tree to generate a boosted tree factor under the asset pricing criterion. We show splitting characteristics and split rule values for each parent node. The node numbers (N#) and splitting order numbers (S#) are also printed on each parent node. We have included the median monthly number of stocks in the leaf basis portfolios. The description of equity characteristics are listed in Table A.1.



# Figure 6: Market-Adjusted Investment Panel Tree

This figure displays the market-adjusted investment P-Tree trained from 1981 to 2000. Controlling the market factor, we apply P-Tree to generate a tree factor to maximize the Sharpe ratio of the SDF spanned by the market factor and the tree factor. We show splitting characteristics and split rule values for each parent node. The node numbers (N#) and splitting order numbers (S#) are also printed on each parent node. We have included the median monthly number of stocks in the leaf basis portfolios. The description of equity characteristics are listed in Table A.1.

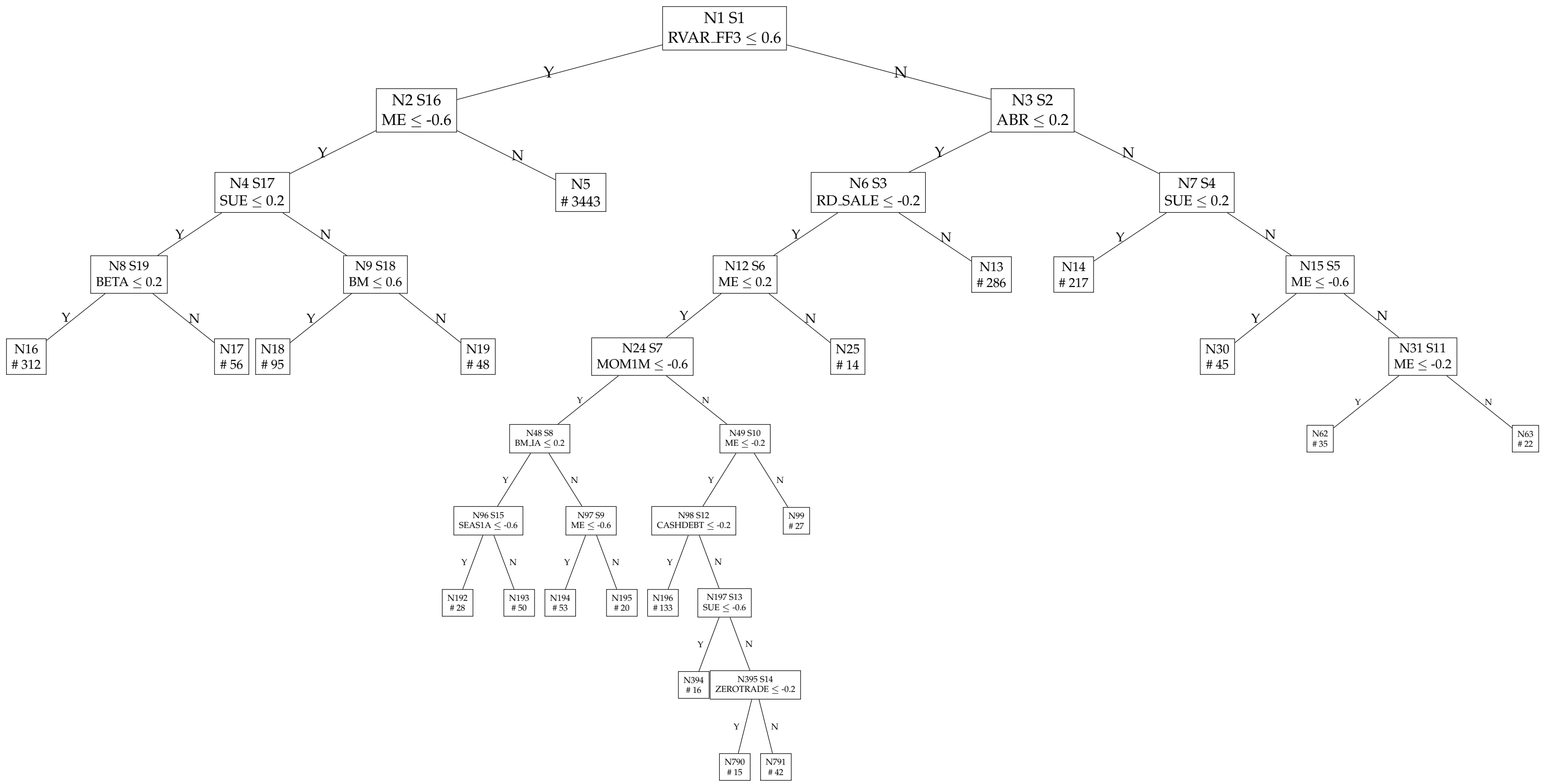
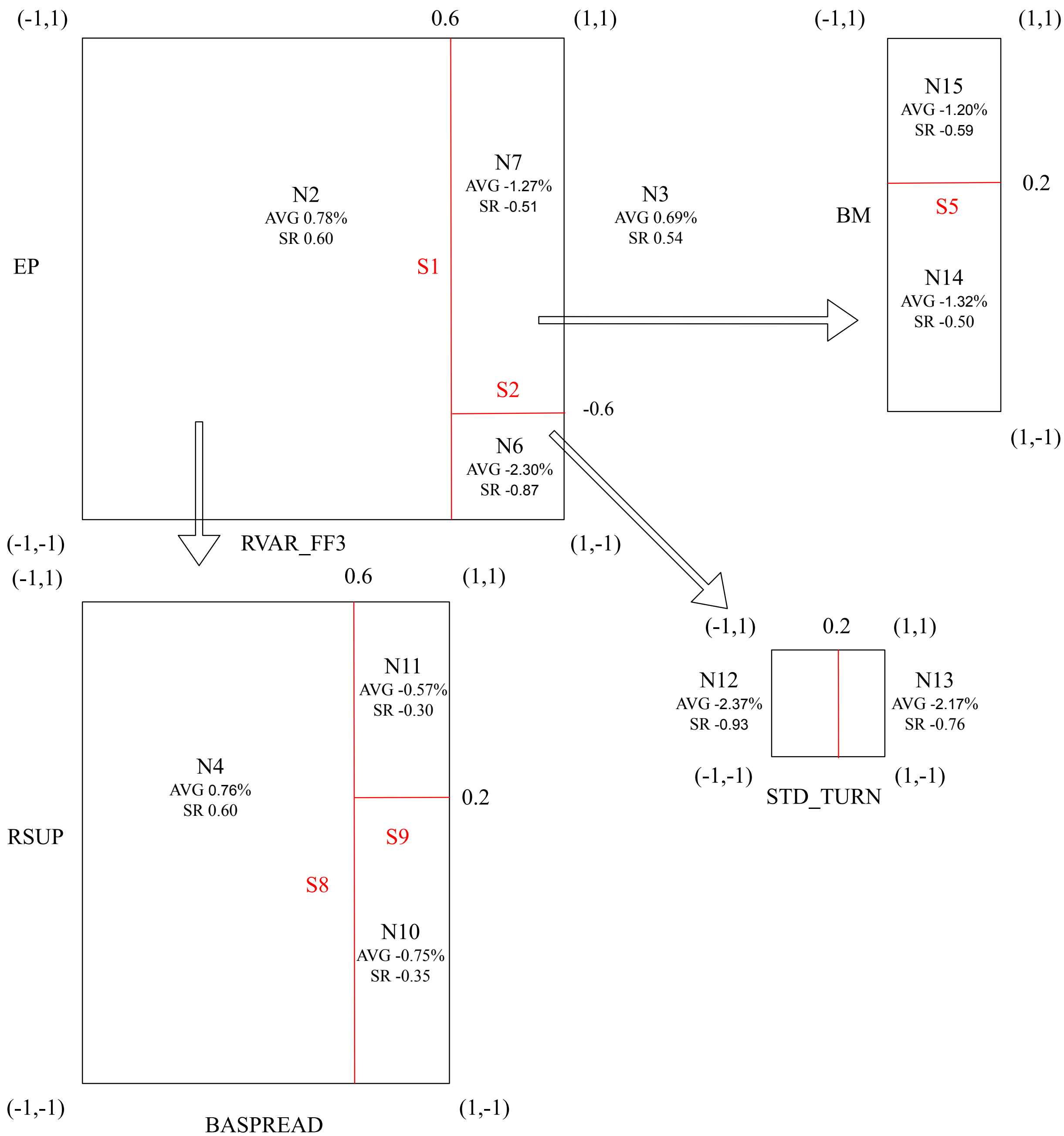


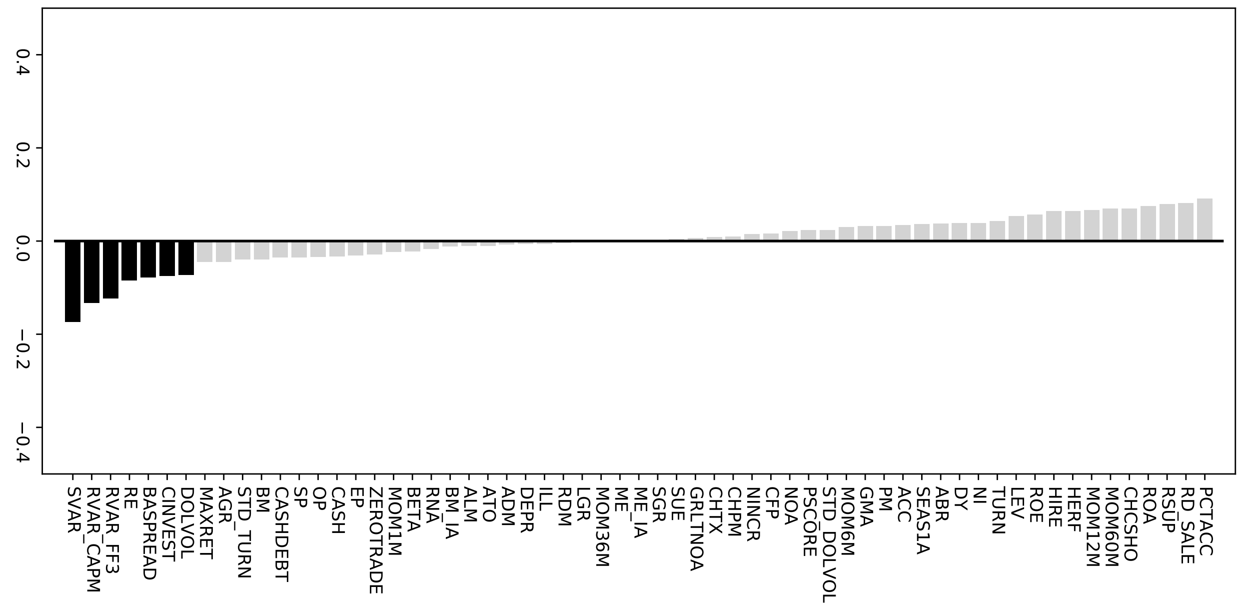
Figure 7: Partition Plots for Figure 5

This diagram illustrates the partition for the first few splits of the tree structure in Figure 5. For example, the first split (S1) is cut at 80% of `RVAR_FF3` on the entire stock universe, and the second split is cut at 20% of `EP` on the high `RVAR_FF3` partition. We also provide each partition's monthly average excess returns and annualized Sharpe ratios. The overlaid arrows show that the next split is implemented on the partitioned area from the previous one.

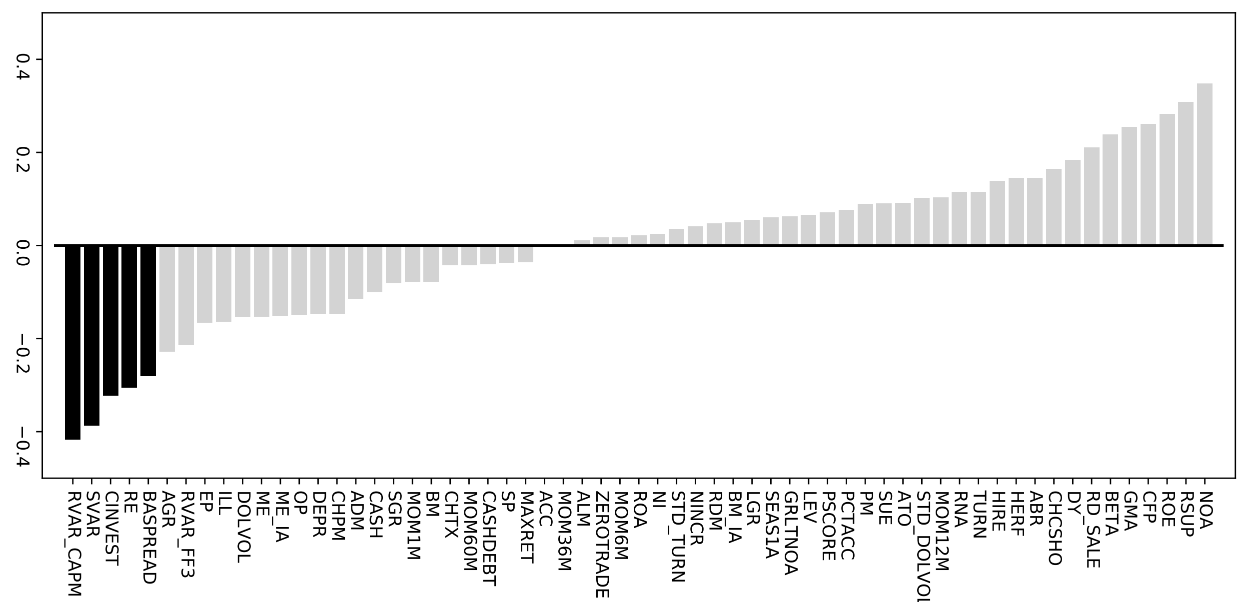


## Figure 8: Out-of-Bag Characteristics Significance

This figure reports the characteristics significance of the out-of-bag ensembles from the random P-Forest of 1,000 trees with the market benchmark. The training period is 1981-2000, and the testing period is 2001-2020. The variable importance measure is the average percentage increase of the loss function by including a characteristic in a tree model. A negative value implies including this characteristic reduces loss and is useful. The dark color bars on the left are significant characteristics at the 5% level by the two-sample t-test. The descriptions of characteristics are listed in Table A.1.



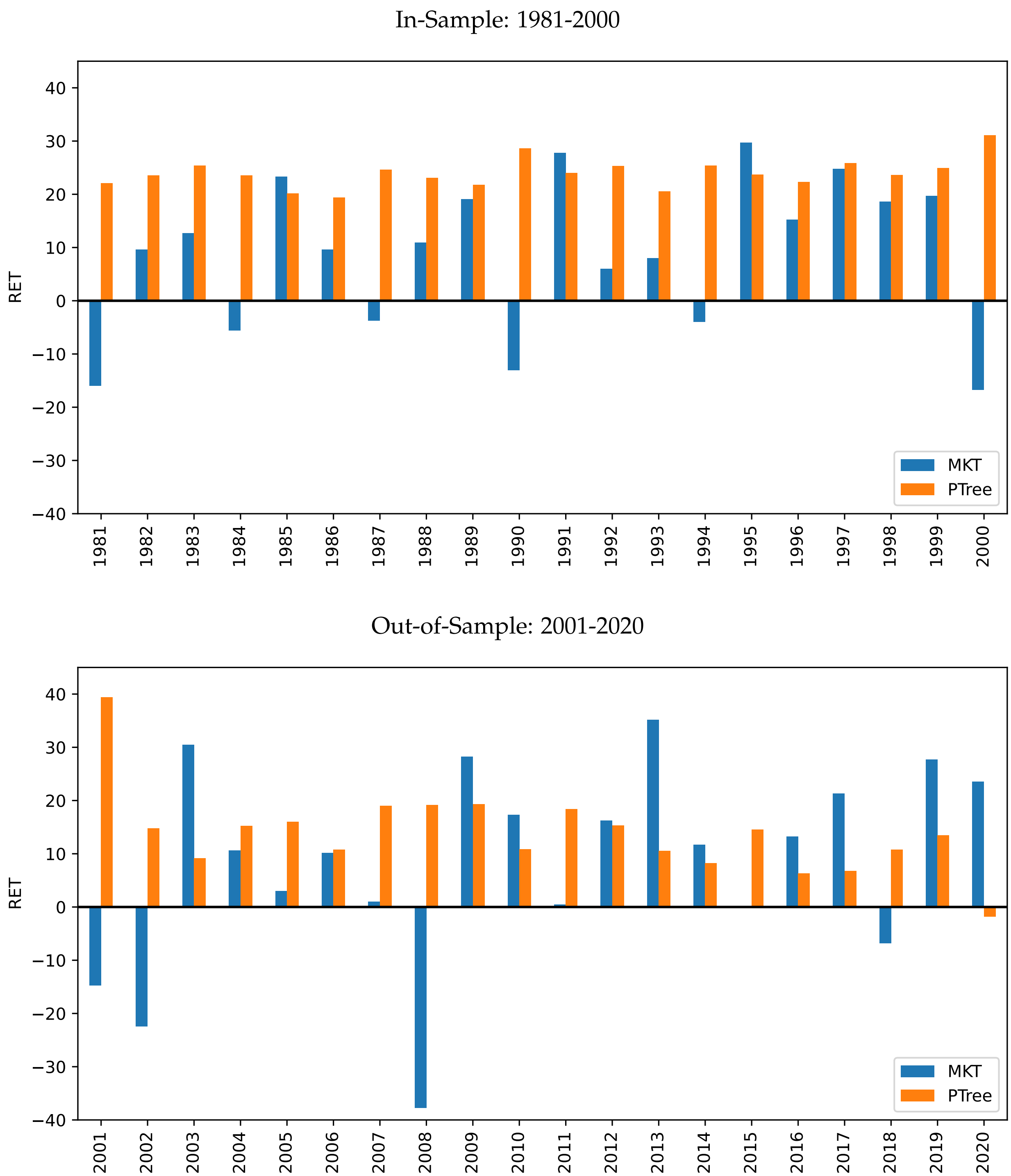

1981-2000




2001-2020

Figure 9: Annualized Returns of Market-Adjusted Investment P-Tree Factor

This figure shows the annualized returns (%) of the market-adjusted investment P-Tree factor and the market factor. The results of both in-sample (1981-2000) and out-of-sample (2001-2020) are presented.



## Table 1: Asset Pricing Performance

This table reports the asset pricing performances. "Tot" (total $R^2$ %) and "CS" (cross-sectional $R^2$ %) in equations (15) and (16) are measures for individual stock returns. The in-sample period is from 1981 to 2000, and the out-of-sample period is from 2001 to 2020. We also report cross-sectional $R^2$ % in (17), using the factor models in the rows to price the test asset portfolios in the columns. "Leaf20" indicates the 20 basis portfolios from Figure 5, and "Leaf40" indicates the 40 basis portfolios from Figure 5 and A.1. Panel A shows results for the market-adjusted asset pricing panel tree models with #factors. Specifically, "PTree2" indicates a two-factor model of the market factor and a P-Tree factor. "PTree5*" indicates a five-factor model of the market factor and four P-Tree factors. "*' indicates the optimal number of factors selected by the squared Sharpe ratio test of Barillas and Shanken (2017). Panel B provides comparisons for benchmark models introduced in section 4.1. Panel C reports the standard asset pricing P-Tree factor models. Panels D and E report the market-adjusted asset pricing P-Tree performance when applying the time-series splits on market volatility and inflation, respectively.

| | Individual Stocks | | | | Portfolios | | | |
|---|---|---|---|---|---|---|---|---|
| | In-Sample | | Out-of-Sample | | Entire Sample | | | |
| | Tot | CS | Tot | CS | FF25 | Ind49 | Leaf20 | Leaf40 |
| **Panel A: Market-Adjusted P-Tree** | | | | | | | | |
| PTree2 | 11.1 | 25.5 | 11.1 | 10.4 | 77.8 | 92.9 | 85.4 | 66.1 |
| PTree5* | 13.0 | 22.7 | 13.7 | 16.5 | 77.9 | 63.2 | 50.8 | 67.3 |
| **Panel B: Other Benchmark Models** | | | | | | | | |
| CAPM | 7.0 | 1.3 | 8.4 | 0.6 | 91.4 | 88.1 | -219.1 | -36.6 |
| FF5 | 11.0 | 13.1 | 11.3 | 5.1 | 96.1 | 78.5 | -72.7 | 22.7 |
| Q5 | 10.9 | 18.1 | 11.5 | 6.4 | 96.1 | 88.7 | 32.5 | 62.6 |
| RP-PCA5 | 12.1 | 18.3 | 13.6 | 15.0 | 69.7 | 48.6 | -66.5 | 23.2 |
| IPCA5 | 13.8 | 27.8 | 14.9 | 17.7 | 90.4 | 57.3 | 31.4 | 63.0 |
| **Panel C: P-Tree** | | | | | | | | |
| PTree1 | 9.5 | 2.7 | 10.8 | 6.6 | 89.1 | 83.3 | -373.8 | -99.5 |
| PTree4* | 12.9 | 27.2 | 13.2 | 13.0 | 78.4 | 31.8 | 15.0 | 46.7 |
| **Panel D: Time-Series Split Market-Adjusted P-Tree - Market Volatility** | | | | | | | | |
| TS-PTree2 | 11.7 | 24.7 | 11.3 | 9.3 | 73.2 | 88.5 | 83.6 | 65.9 |
| TS-PTree5* | 13.4 | 26.4 | 13.6 | 18.4 | 89.4 | 53.6 | 40.2 | 57.2 |
| **Panel E: Time-Series Split Market-Adjusted P-Tree - Inflation** | | | | | | | | |
| TS-PTree2 | 11.5 | 26.6 | 11.4 | 11.7 | 78.6 | 85.7 | 76.9 | 70.6 |
| TS-PTree5* | 13.3 | 28.6 | 13.3 | 16.1 | 91.0 | 71.2 | 80.3 | 80.0 |

## Table 2: Investing P-Tree Factors

This table reports the investment performance of the equity factor models. We report the monthly average return (%) and Jensen's $\alpha$ (%), the annualized Sharpe ratio for the factors' mean-variance efficient (MVE) and equal-weighted ($1/N$) portfolios. Panels A, E, and F use asset pricing criteria; Panels B and C are using investment criteria. The "*" indicates the optimal number of factors selected by the squared Sharpe ratio test of Barillas and Shanken (2017). For $t$-statistics *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

| | In-Sample (1981-2000) | | | | | | Out-of-Sample (2001-2020) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MVE | | | $1/N$ | | | MVE | | | $1/N$ | | |
| | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ |
| **Panel A: Market-Adjusted P-Tree** | | | | | | | | | | | | |
| PTree2 | 1.75 | 1.58 | 1.51*** | 1.31 | 1.34 | 0.86*** | 0.30 | 0.31 | 0.29 | 0.45 | 0.56 | 0.17 |
| PTree5* | 1.26 | 3.47 | 1.20*** | 1.06 | 1.69 | 0.72*** | 0.80 | 1.93 | 0.76*** | 0.70 | 1.14 | 0.42*** |
| **Panel B: Market-Adjusted Investment P-Tree** | | | | | | | | | | | | |
| PTree2 | 1.78 | 10.41 | 1.76*** | 1.27 | 1.94 | 0.90*** | 1.07 | 2.78 | 1.10*** | 0.86 | 1.36 | 0.56*** |
| PTree5* | 1.36 | 12.55 | 1.35*** | 0.78 | 1.93 | 0.56*** | 0.76 | 2.96 | 0.78*** | 0.48 | 1.34 | 0.32*** |
| **Panel C: Investment P-Tree** | | | | | | | | | | | | |
| PTree1 | 1.80 | 10.52 | 1.80*** | 1.80 | 10.52 | 1.80*** | 1.07 | 2.71 | 1.13*** | 1.07 | 2.71 | 1.13*** |
| PTree5* | 1.37 | 14.66 | 1.36*** | 0.74 | 2.96 | 0.63*** | 0.79 | 2.78 | 0.82*** | 0.44 | 1.64 | 0.35*** |
| **Panel D: Other Benchmark Models** | | | | | | | | | | | | |
| FF5 | 0.45 | 1.48 | 0.38*** | 0.38 | 1.34 | 0.33*** | 0.27 | 0.64 | 0.13* | 0.25 | 0.59 | 0.12 |
| Q5 | 0.77 | 2.78 | 0.74*** | 0.63 | 2.10 | 0.53*** | 0.34 | 1.22 | 0.34*** | 0.31 | 1.10 | 0.25*** |
| RP-PCA5 | 0.82 | 3.48 | 0.76*** | 1.07 | 1.77 | 0.75*** | 0.34 | 1.49 | 0.32*** | 0.50 | 1.00 | 0.27*** |
| IPCA5 | 1.50 | 10.37 | 1.48*** | 0.90 | 3.15 | 0.80*** | 0.97 | 4.60 | 0.98*** | 0.73 | 2.14 | 0.61*** |
| **Panel E: Time-Series Split Market-Adjusted P-Tree - Market Volatility** | | | | | | | | | | | | |
| TS-PTree2 | 1.80 | 1.67 | 1.56*** | 1.32 | 1.36 | 0.86*** | 0.60 | 0.66 | 0.57*** | 0.60 | 0.77 | 0.31*** |
| TS-PTree5* | 1.20 | 3.65 | 1.14*** | 1.03 | 1.61 | 0.69*** | 0.78 | 1.69 | 0.73*** | 0.74 | 1.14 | 0.46*** |
| **Panel F: Time-Series Split Market-Adjusted P-Tree - Inflation** | | | | | | | | | | | | |
| TS-PTree2 | 1.66 | 1.70 | 1.49*** | 1.21 | 1.34 | 0.78*** | 0.28 | 0.32 | 0.29 | 0.43 | 0.56 | 0.11 |
| TS-PTree5* | 1.37 | 4.50 | 1.32*** | 1.21 | 2.07 | 0.93*** | 0.51 | 1.24 | 0.53*** | 0.64 | 1.16 | 0.52*** |

## Table 3: Characteristics Importance by Top Splits

This table reports the most frequently selected characteristics from the random P-Forest of 1,000 trees with the market benchmark, which can be used to assess the variable importance. The "Top 1" rows only count the first split for 1,000 trees. The "Top 2" or "Top 3" rows only count the first two or three splits. The numbers reported are the selection frequency for these top characteristics selected out of the 1,000 ensembles. The description of characteristics are listed in Table A.1.

Panel A: Equity Entire Training Sample (1981-2000)

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Top1 | RVAR_FF3 | RVAR_CAPM | ME | SVAR | CFP |
| | 0.40 | 0.40 | 0.39 | 0.32 | 0.25 |
| Top2 | ME | RVAR_FF3 | RVAR_CAPM | CFP | EP |
| | 0.45 | 0.41 | 0.40 | 0.35 | 0.33 |
| Top3 | ME | RVAR_FF3 | RVAR_CAPM | CFP | EP |
| | 0.45 | 0.41 | 0.41 | 0.37 | 0.36 |

Panel B: Low Stock Variance

| | 1 | 2 | 3 |
|---|---|---|---|
| Top1 | RVAR_FF3 | RVAR_CAPM | SVAR |
| | 0.48 | 0.42 | 0.33 |
| Top2 | RVAR_FF3 | RVAR_CAPM | ME |
| | 0.48 | 0.42 | 0.34 |
| Top3 | RVAR_FF3 | RVAR_CAPM | SVAR |
| | 0.49 | 0.43 | 0.35 |

Panel C: High Stock Variance

| | 1 | 2 | 3 |
|---|---|---|---|
| Top1 | ME | RVAR_FF3 | RVAR_CAPM |
| | 0.36 | 0.34 | 0.30 |
| Top2 | ME | RVAR_FF3 | EP |
| | 0.43 | 0.36 | 0.32 |
| Top3 | ME | rvar_ff3 | EP |
| | 0.44 | RVAR_FF3 | 0.36 |

Panel D: Low Inflation

| | 1 | 2 | 3 |
|---|---|---|---|
| Top1 | RVAR_FF3 | RVAR_CAPM | SVAR |
| | 0.48 | 0.42 | 0.30 |
| Top2 | RVAR_FF3 | RVAR_CAPM | ZEROTRADE |
| | 0.48 | 0.42 | 0.34 |
| Top3 | RVAR_FF3 | RVAR_CAPM | ZEROTRADE |
| | 0.48 | 0.43 | 0.35 |

Panel E: High Inflation

| | 1 | 2 | 3 |
|---|---|---|---|
| Top1 | ME | CFP | EP |
| | 0.41 | 0.35 | 0.32 |
| Top2 | ME | CFP | RVAR_FF3 |
| | 0.46 | 0.37 | 0.37 |
| Top3 | ME | CFP | RVAR_FF3 |
| | 0.46 | 0.38 | 0.37 |

## Table 4: Factor Spanning Alpha Test

This table reports the monthly alphas (bp) and significance for the factor-spanning test. Panels A and B show results for each factor of the market-adjusted asset pricing and investment P-Tree models in Table 2. We regress each factor in the rows against a factor model in the columns. "FF5" is the five-factor model in Fama and French (2015), "Q5" is the five-factor model in Hou et al. (2021), and "IPCA5" is the five-factor model in Kelly et al. (2019). We name the P-Tree factors by the first two splitting characteristics in the tree structure for the characteristics interactions. We also show the mean-variance efficient (MVE) and $1/N$ strategies investing these five factors. Notably, "FF25" and "INF49" are the $5 \times 5$ ME-B/M and 49 industry portfolios. For $t$-statistics, *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

| | In-Sample | | | Out-of-Sample | | |
|---|---|---|---|---|---|---|
| | FF5 | Q5 | IPCA5 | FF5 | Q5 | IPCA5 |
| **Panel A: Market-Adjusted P-Tree factors** | | | | | | |
| RVAR_FF3-EP | 130*** | 101*** | 107** | 12 | 4 | 31 |
| BM_IA-Ill | 35** | 33 | -100*** | 107*** | 110*** | 34 |
| MOM12M-STD_DOLVOL | 82*** | 53*** | -24 | 25 | 22 | -95*** |
| ME-RDM | 52*** | 48*** | 109*** | 29*** | 27*** | 13 |
| MVE (4 factors + mkt) | 58*** | 45*** | -21 | 36*** | 34*** | -11 |
| 1/N (4 factors + mkt) | 60*** | 47*** | 49* | 35*** | 33*** | 1 |
| **Panel B: Market-Adjusted Investment P-Tree factors** | | | | | | |
| RVAR_FF3-ABR | 354*** | 341*** | 227*** | 215*** | 201*** | 69*** |
| BM_IA-LGR | 46*** | 58*** | 96*** | 13 | 16 | -20 |
| STD_TURN-LEV | 36*** | 32*** | 85*** | -21** | -19** | -18 |
| CFP-MOM12M | 53*** | 49*** | 90*** | 45** | 47** | 5 |
| MVE (4 factors + mkt) | 248*** | 241*** | 175*** | 147*** | 139*** | 42** |
| 1/N (4 factors + mkt) | 98*** | 96*** | 131*** | 50*** | 49*** | 12 |
| **Panel C: Other Test Assets** | | | | | | |
| MVE-FF25 | 55*** | 42*** | 27* | 19*** | 15** | 10 |
| MVE-IND49 | 13* | 20 | -14 | 10 | 8 | 28* |
| 1/N-FF25 | -8*** | -8 | 57*** | 3 | 8** | 5 |
| 1/N-IND49 | 63* | 30 | 62 | -2 | 10 | 4 |

Table 5: Uni-Sort Factors vs. Interaction Factors

This table summarizes the significant counts of long-short factors for average returns and Jensen's alphas. We count the number of significant average returns and alphas in Panels A and B at the 10% and 5% levels. The specifications include (1) 4x1 long-short portfolios, (2) market-adjusted asset pricing P-Tree interaction factors, and (3) market-adjusted investment P-Tree interaction factors.

Panel A: # of Significant Cases at 10% level

| | Uni-Sort 4x1 | | P-Tree Interaction | | Investment P-Tree Interaction | |
|---|---|---|---|---|---|---|
| | Mean | Alpha | Mean | Alpha | Mean | Alpha |
| 81-00 | 27 | 33 | 18 | 44 | 54 | 55 |
| 01-20 | 14 | 28 | 20 | 38 | 37 | 51 |
| 81-20 | 32 | 37 | 24 | 50 | 55 | 58 |

Panel B: # of Significant Cases at 5% level

| | Uni-Sort 4x1 | | P-Tree Interaction | | Investment P-Tree Interaction | |
|---|---|---|---|---|---|---|
| | Mean | Alpha | Mean | Alpha | Mean | Alpha |
| 81-00 | 20 | 31 | 10 | 34 | 52 | 54 |
| 01-20 | 6 | 22 | 14 | 28 | 35 | 49 |
| 81-20 | 24 | 32 | 21 | 46 | 54 | 58 |

## Table 6: Examples of Interaction Factors

This table demonstrates six examples in Figure A.6. We report the out-of-sample (2001-2020) monthly average returns (%) and Jensen's alpha (%) of the 4x1 long-short factors and the market-adjusted investment P-Tree interaction factors. The interaction factors are created with the train sample period from 1981 to 2000, and we have provided the corresponding interaction characteristics. For $t$-statistics, *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The descriptions of characteristics are listed in Table A.1.

| | Panel A: Stand. Unexp. Earnings | | | | Panel B: Maximum Daily Returns | | | |
|---|---|---|---|---|---|---|---|---|
| | Uni-Sort 4x1 | | +BM_IA+DOLVOL | | Uni-Sort 4x1 | | +ME_IA+ABR | |
| | Mean | Alpha | Mean | Alpha | Mean | Alpha | Mean | Alpha |
| 81-00 | 0.69*** | 0.64*** | 1.14*** | 1.26*** | 0.81** | 1.20*** | 1.08*** | 1.40*** |
| 01-20 | 0.53*** | 0.67*** | 1.35*** | 1.39*** | -0.04 | 0.47 | 0.67*** | 1.11*** |
| 81-20 | 0.61*** | 0.66*** | 1.25*** | 1.32*** | 0.39 | 0.84*** | 0.87*** | 1.26*** |

| | Panel C: Dividend Yield | | | | Panel D: Growth in Long-term Debt | | | |
|---|---|---|---|---|---|---|---|---|
| | Uni-Sort 4x1 | | +MOM12M+ME_IA | | Uni-Sort 4x1 | | +BA_IA+SUE | |
| | Mean | Alpha | Mean | Alpha | Mean | Alpha | Mean | Alpha |
| 81-00 | 0.41 | 0.87*** | 1.05*** | 1.34*** | 0.33** | 0.52*** | 0.77*** | 0.95*** |
| 01-20 | -0.07 | 0.03 | 0.47** | 0.74*** | -0.09 | -0.10 | 0.53*** | 0.56*** |
| 81-20 | 0.17 | 0.44** | 0.76*** | 1.04*** | 0.12 | 0.20* | 0.65*** | 0.75*** |

| | Panel E: Change in Profit Margin | | | | Panel F: Momentum 6-Month | | | |
|---|---|---|---|---|---|---|---|---|
| | Uni-Sort 4x1 | | +BASPREAD+ME_IA | | Uni-Sort 4x1 | | +GMA+BM_IA | |
| | Mean | Alpha | Mean | Alpha | Mean | Alpha | Mean | Alpha |
| 81-00 | 0.45*** | 0.41*** | 1.07*** | 1.20*** | 0.60* | 0.64** | 1.12*** | 1.20*** |
| 01-20 | 0.32* | 0.46*** | 0.46** | 0.65*** | 0.15 | 0.53 | 0.52* | 0.83*** |
| 81-20 | 0.38*** | 0.45*** | 0.76*** | 0.93*** | 0.37 | 0.60** | 0.82*** | 1.02*** |

# Appendices

**Algorithm** Algorithm of growing a single P-Tree

**procedure** GROWTREE(`root`)
**Input** Asset returns $\mathbf{r}_{i,t}$ and ranked characteristics $\mathbf{z}_{i,t}$
**outcome** Grow the tree from the `root` node, form leaf basis portfolios
  **for** $j$ from $1$ to $J$ **do** ▷ Loop over number of iterations
    **if** current depth $\geq d_{\text{max}}$ **then**
      **return**.
    **else**
      Search the tree, find all leaf nodes $\mathcal{N}$
      **for** each leaf node $N$ in $\mathcal{N}$ **do** ▷ Loop over all current leaf nodes
        **for** each split rule candidate $\tilde{\mathbf{c}}_{k,m,N}$ in $\mathcal{C}_N$ **do**
          Partition data temporally in $N$ according to $\tilde{\mathbf{c}}_{k,m,N}$.
          **if** Either left or right child of $N$ does not satisfy minimal leaf size **then**
            $\mathcal{L}(\tilde{\mathbf{c}}_{k,m,N}) = \infty$.
          **else**
            Calculate leaf basis portfolios.
            Estimate SDF using *all* leaf basis portfolios as in (4).
            Calculate the split criteria $\mathcal{L}(\tilde{\mathbf{c}}_{k,m,N})$ in (5).
          **end if**
        **end for**
      **end for**
      Find the best leaf node and split rule that minimizes split criteria

$$\tilde{\mathbf{c}}_j = \underset{N \in \mathcal{N}, \tilde{\mathbf{c}}_{k,m,N} \in \mathcal{C}_N}{\arg\min} \{\mathcal{L}(\tilde{\mathbf{c}}_{k,m,N})\}$$

      Split the node selected at the $j$-th split rule of the tree $\tilde{\mathbf{c}}_j$.
    **end if**
  **end for**
  **return** $\left\{\mathbf{R}_t^{(J)}, f_t^{(J)}, \beta^{(J)}(\cdot)\right\}$
**end procedure**

Table A.1: Equity Characteristics

This table lists the description of 61 characteristics used in the empirical study.

| No. | Characteristics | Description |
|---|---|---|
| 1 | ABR | Abnormal returns around earnings announcement |
| 2 | ACC | Operating Accruals |
| 3 | ADM | Advertising Expense-to-market |
| 4 | AGR | Asset growth |
| 5 | ALM | Quarterly Asset Liquidity |
| 6 | ATO | Asset Turnover |
| 7 | BASPREAD | Bid-ask spread (3 months) |
| 8 | BETA | Beta (3 months) |
| 9 | BM | Book-to-market equity |
| 10 | BM_IA | Industry-adjusted book to market |
| 11 | CASH | Cash holdings |
| 12 | CASHDEBT | Cash to debt |
| 13 | CFP | Cashflow-to-price |
| 14 | CHCSHO | Change in shares outstanding |
| 15 | CHPM | Industry-adjusted change in profit margin |
| 16 | CHTX | Change in tax expense |
| 17 | CINVEST | Corporate investment |
| 18 | DEPR | Depreciation / PP&E |
| 19 | DOLVOL | Dollar trading volume |
| 20 | DY | Dividend yield |
| 21 | EP | Earnings-to-price |
| 22 | GMA | Gross profitability |
| 23 | GRLTNOA | Growth in long-term net operating assets |
| 24 | HERF | Industry sales concentration |
| 25 | HIRE | Employee growth rate |
| 26 | ILL | Illiquidity rolling (3 months) |
| 27 | LEV | Leverage |
| 28 | LGR | Growth in long-term debt |
| 29 | MAXRET | Maximum daily returns (3 months) |
| 30 | ME | Market equity |
| 31 | ME_IA | Industry-adjusted size |
| 32 | MOM12M | Cumulative Returns in the past (2-12) months |
| 33 | MOM1M | Previous month return |
| 34 | MOM36M | Cumulative Returns in the past (13-35) months |
| 35 | MOM60M | Cumulative Returns in the past (13-60) months |
| 36 | MOM6M | Cumulative Returns in the past (2-6) months |
| 37 | NI | Net Equity Issue |
| 38 | NINCR | Number of earnings increases |
| 39 | NOA | Net Operating Assets |
| 40 | OP | Operating profitability |

Continue: Equity Characteristics

| No. | Characteristics | Description |
|---|---|---|
| 41 | PCTACC | Percent operating accruals |
| 42 | PM | profit margin |
| 43 | PS | Performance Score |
| 44 | RD_SALE | R&D to sales |
| 45 | RDM | R&D to market |
| 46 | RE | Revisions in analysts' earnings forecasts |
| 47 | RNA | Return on Net Operating Assets |
| 48 | ROA | Return on Assets |
| 49 | ROE | Return on Equity |
| 50 | RSUP | Revenue surprise |
| 51 | RVAR_CAPM | Residual variance - CAPM (3 months) |
| 52 | RVAR_FF3 | Res. var. - Fama-French 3 factors (3 months) |
| 53 | SVAR | Return variance (3 months) |
| 54 | SEAS1A | 1-Year Seasonality |
| 55 | SGR | Sales growth |
| 56 | SP | Sales-to-price |
| 57 | STD_DOLVOL | Std of dollar trading volume (3 months) |
| 58 | STD_TURN | Std. of Share turnover (3 months) |
| 59 | SUE | Unexpected quarterly earnings |
| 60 | TURN | Shares turnover |
| 61 | ZEROTRADE | Number of zero-trading days (3 months) |

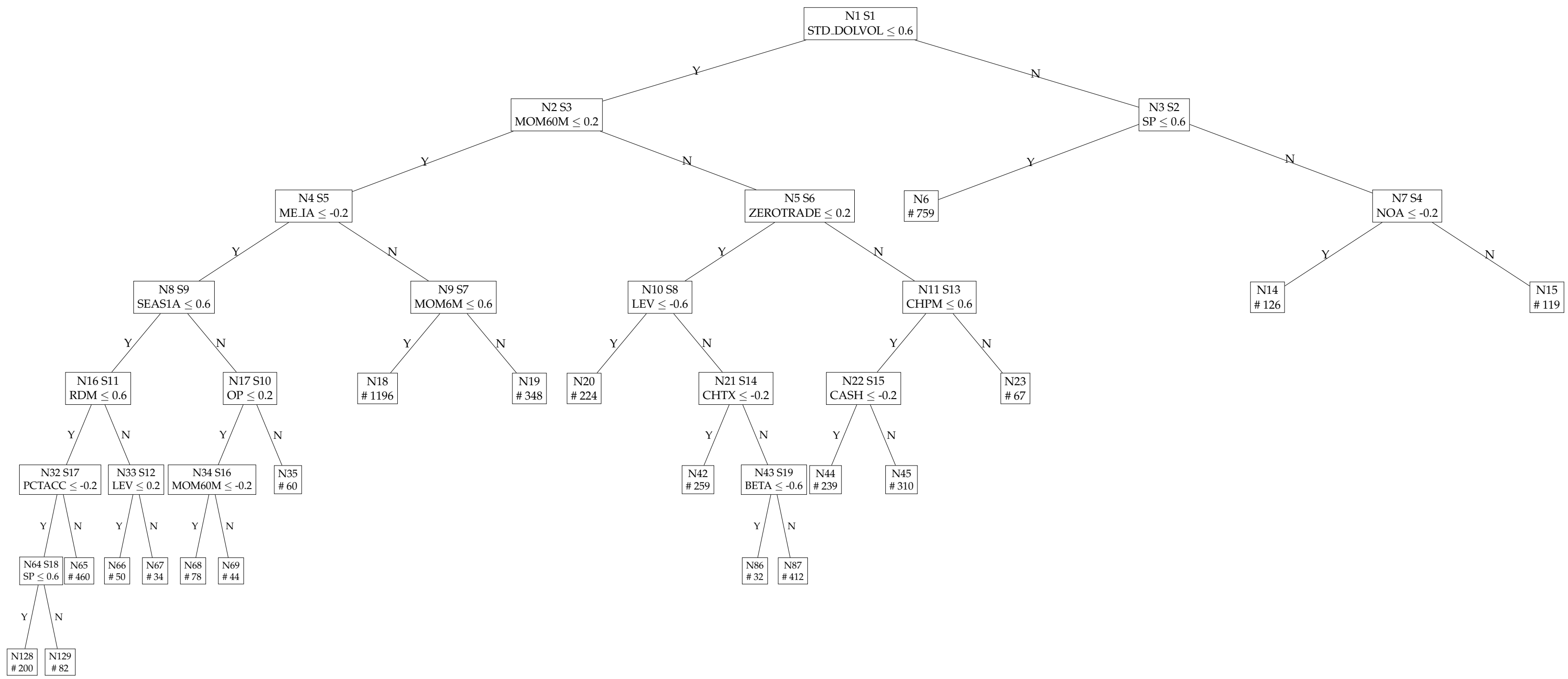
## Table A.2: Macro Predictors for Market Timing

This table lists the description of macro predictors used in the empirical study.

| No. | Variable Name | Description |
|---|---|---|
| 1 | EP | Earnings-to-price of S&P 500 |
| 2 | DY | Dividend yield of S&P 500 |
| 3 | LEV | Leverage of S&P 500 |
| 4 | NI | Net equity issuance of S&P 500 |
| 5 | SVAR | Stock Variance of S&P 500 |
| 6 | ILL | Pastor-Stambaugh illiquidity |
| 7 | INFL | Inflation |
| 8 | TBL | Three-month treasure bill rate |
| 9 | DFY | Default yield |
| 10 | TMS | Term spread |

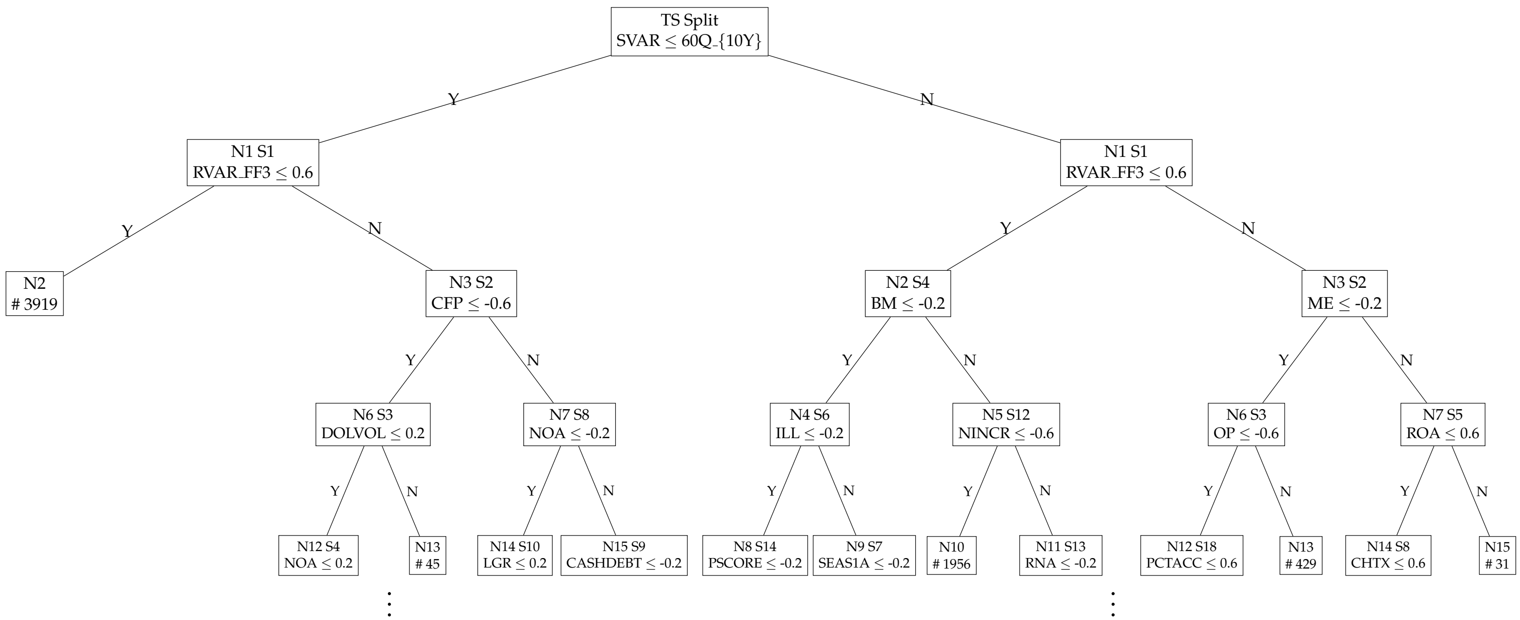
# Figure A.1: Panel Tree

This figure displays the standard asset pricing P-Tree trained from 1981 to 2000. The figure format follows Figure 5.



## Figure A.2: Panel Tree Split on High/Low Market Volatility

This figure shows the market-adjusted asset pricing panel tree by considering both cross-sectional and time-series variations. The most important macro predictor is stock market volatility `SVAR`, and the first split is implemented when the current `SVAR` level is lower than the 60% quantile of the past decade. Two tree models are provided as two child leaves for high- and low-`SVAR` periods. The figure format follows Figure 5.



# Figure A.3: Panel Tree Split on High/Low Inflation

This figure shows the market-adjusted asset pricing panel tree by considering both cross-sectional and time-series variations. One of the most commonly used macro predictors is inflation, and the first split is implemented when the current inflation level is lower than the median of the past decade. Two tree models are provided as two child leaves for high- and low-inflation periods. The figure format follows Figure 5.

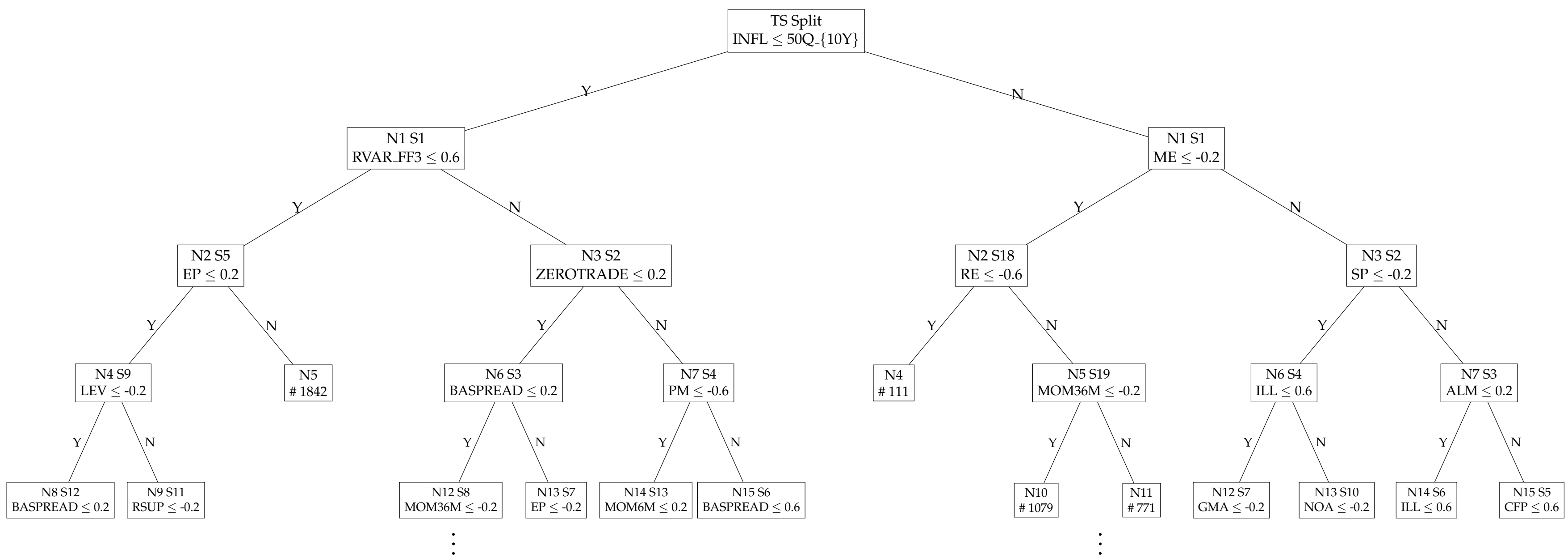
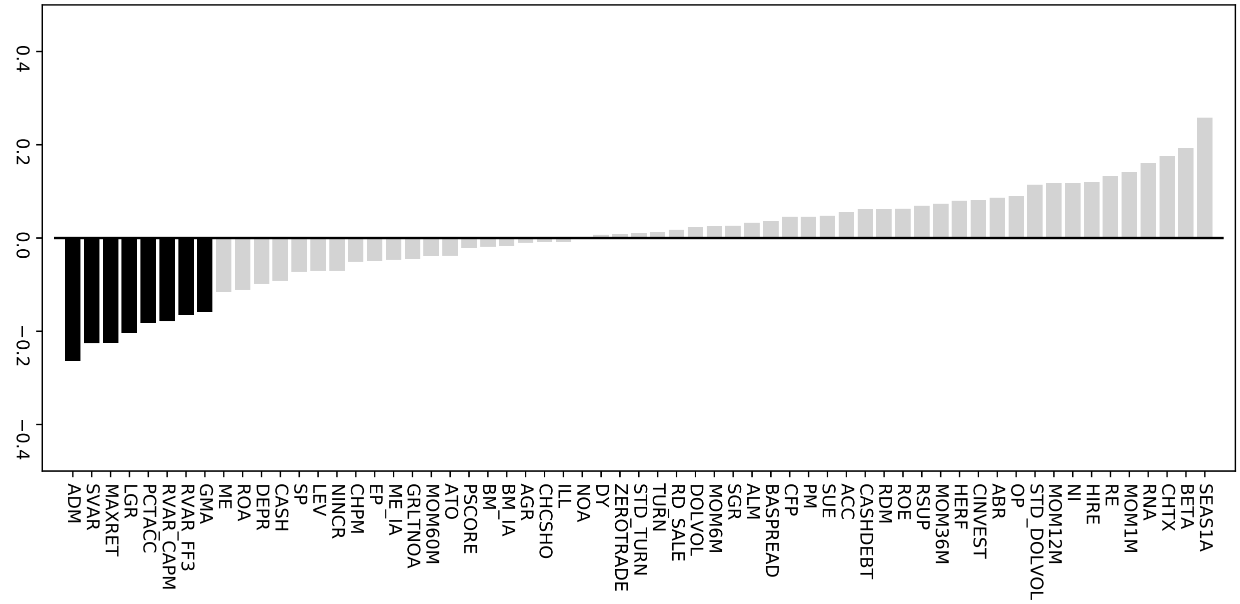
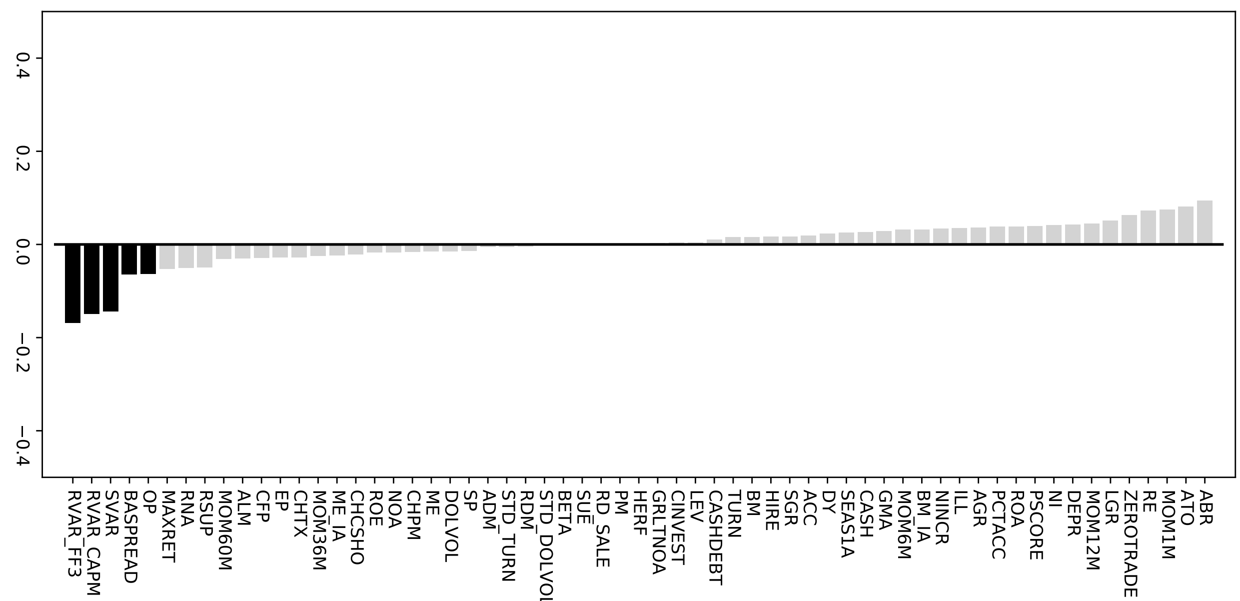


Figure A.4: Out-of-Bag Characteristics Significance: High/Low Market Volatility

This figure reports the characteristics significance of the out-of-bag ensembles from the random P-Forest of 1,000 trees with the market benchmark. We report results for high and low market volatility periods (split at 60-th quantile over the past decade) in the training sample, 1981-2000. This figure's details follow Figure 8. The left dark columns indicate significantly useful characteristics for reducing the loss function.



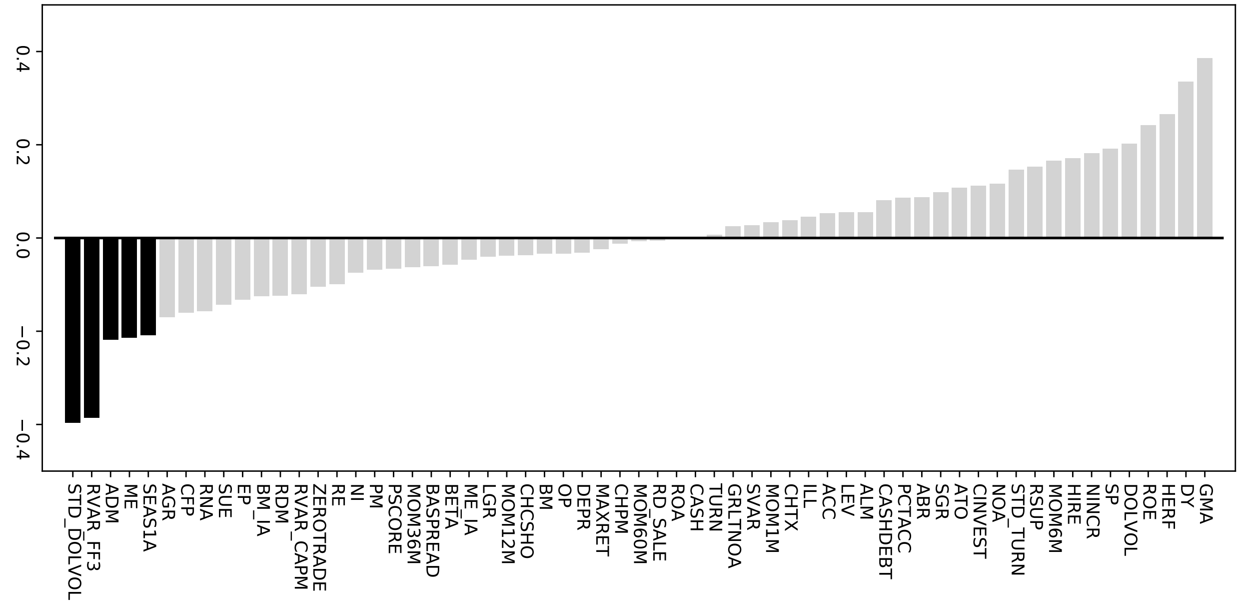

High Stock Variance



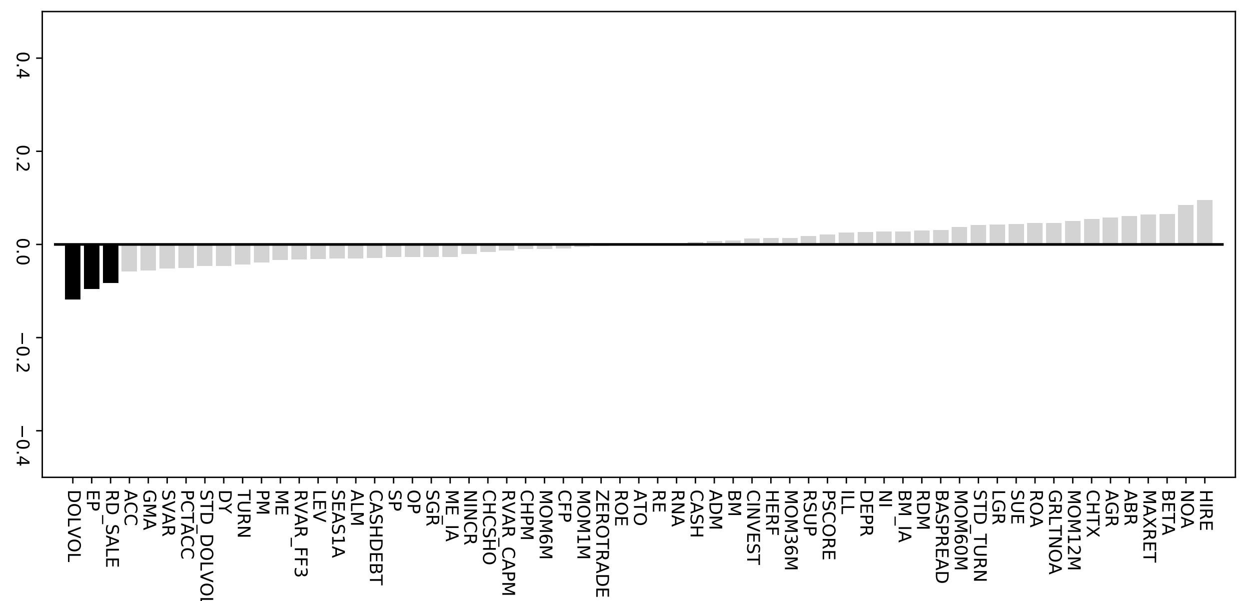

Low Stock Variance

# Figure A.5: Out-of-Bag Characteristics Significance: High/Low Inflation

This figure reports the characteristics significance of the out-of-bag ensembles from the random P-Forest of 1,000 trees with the market benchmark. We report results for high and low inflation (split at 50-th quantile over the past decade) in the training sample, 1981-2000. This figure's details follow Figure 8. The left dark columns indicate significantly useful characteristics for reducing the loss function.



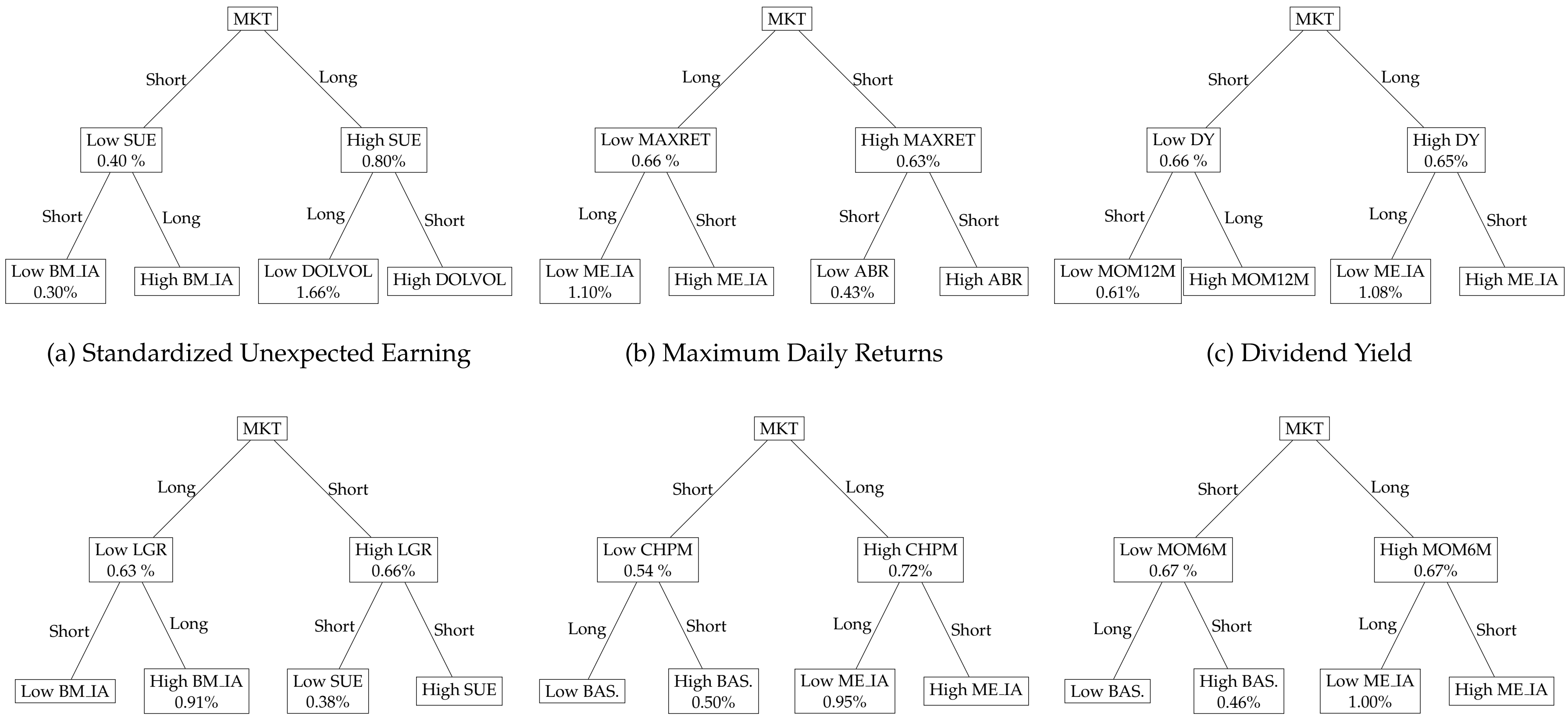

High Inflation




Low Inflation

# Figure A.6: Examples of Interaction Factors

This figure shows six examples of the market-adjusted investment P-Tree interaction factors. More details are reported in Table 6. In the second layer, we report the average excess returns of the long portfolio and short portfolio. In the third layer, the numbers are the average returns of the long-long portfolio and short-short portfolio. One can find further splitting helps create a higher return spread.




(a) Standardized Unexpected Earning

(b) Maximum Daily Returns

(c) Dividend Yield

(d) Growth in Long-term Debt

(e) Change in Profit Margin

(f) Momentum 6-Month

# Internet Appendix

## I Robustness check for Tuning Parameter *No. Leaf*

This section investigates the robustness of the P-Tree with respect to a key tuning parameter, *Number of Leaves*. The baseline empirical results are based on *Number of Leaves* equals 20, and we provide two other values, 15 and 25. We find the tree structure, asset pricing performance, and investment performance are robust for *Number of Leaves*.

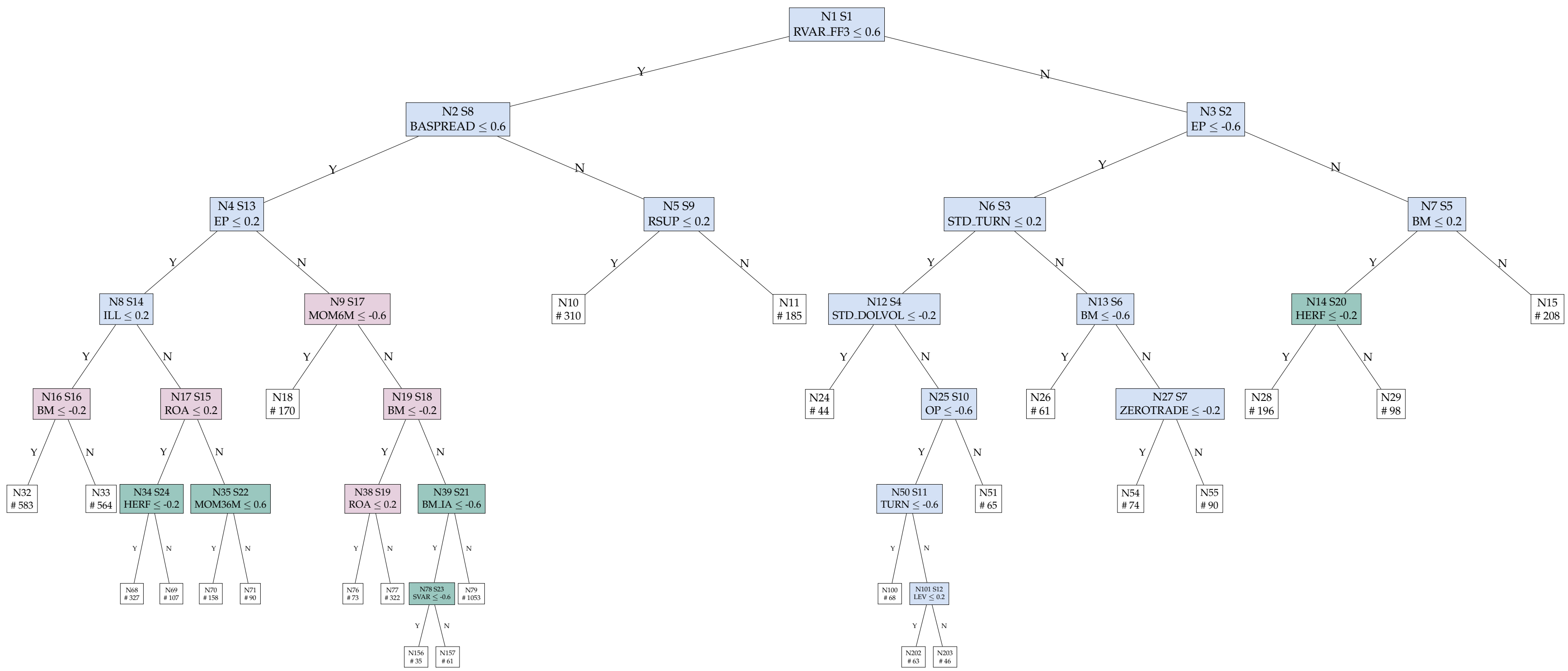
**Tree Structure** The P-Tree grows iteratively. Starting from the root node, we keep splitting and getting a sequence of P-Trees with more and more leaves. The market-adjusted P-Tree with 25 leaves is displayed in Figure I.1. We highlight the first 14 splits in blue, the 15th to 19th splits in red, and the 20th to 24th in green. The structure of the P-Tree with 25 leaves subsumes the one with 20 and 15 leaves. So, the tree structure is sequential growing and robust for the number of leaves for asset pricing P-Tree. For other specifications of P-Trees, we report the sequential structure of the standard asset pricing P-Tree in Figure I.2 and the market-adjusted investment P-Tree in Figure I.3. The deeper tree always subsumes the shallow tree.

**Asset Pricing and Investment Performance** Table I.1 shows the asset pricing performance of P-Trees with 15 and 25 leaves. Panels A and B report for the market-adjusted P-Trees. Panels C and D report for the standard P-Trees. Although the number of leaves changes, the performance measures are close to those in Panels A and C Table 1.

Table I.2 shows the asset pricing performance of P-Trees with 15 and 25 leaves. Panels A and B are for the market-adjusted P-Tree, Panels C and D are for the market-adjusted investment P-Trees, and Panels E and F are for the investment P-Trees. Consistent with Table 2, the out-of-sample MVE strategies of asset pricing P-Tree has annualized Sharpe ratio of about 1.9, and the investment P-Trees have Sharpe ratios close to 3.

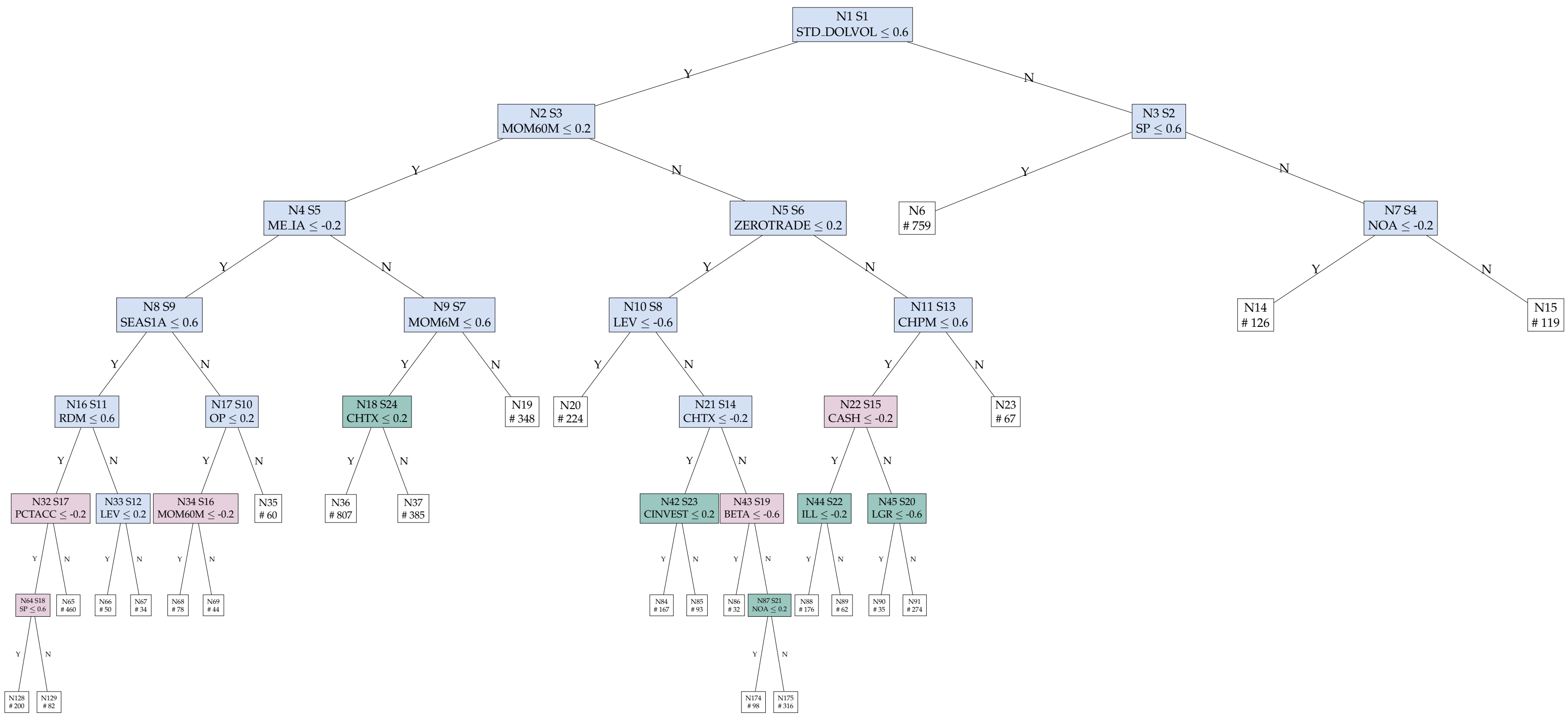
# Figure I.1: Market-Adjusted Panel Tree

This figure visualizes the market-adjusted asset pricing P-Tree. The number of leaves is 25, so there are 24 splits. We color the first 14 splits in blue, the 15th to 19th splits in red, and the 20th to 24th in green. The blue part is a P-Tree with 15 leaves (up to S14); the combination of the blue and red parts is a P-Tree with 20 leaves (up to S19); the whole diagram is a P-Tree with 25 leaves. Figure format follows Figure 5.



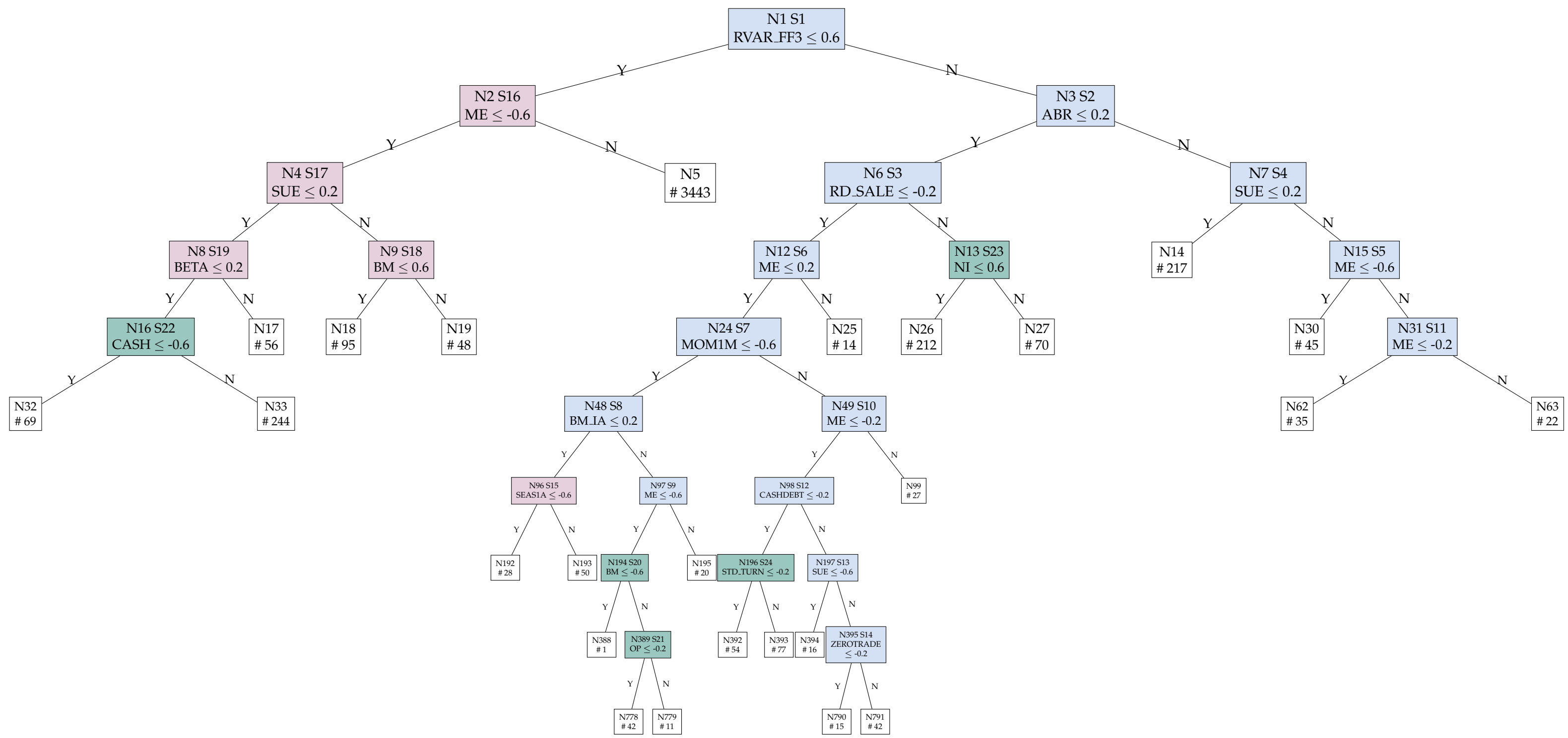
## Figure I.2: Panel Tree

This figure visualizes the asset pricing P-Tree. The number of leaves is 25. We color the first 14 splits in blue, the 15th to 19th splits in red, and the 20th to 24th in green. The blue part is a P-Tree with 15 leaves (up to S14); the combination of the blue and red parts is a P-Tree with 20 leaves (up to S19); the whole diagram is a P-Tree with 25 leaves. Figure format follows Figure A.1.



## Figure I.3: Investment Panel Tree

This figure visualizes the market-adjusted investment P-Tree. The number of leaves is 25. We color the first 14 splits in blue, the 15th to 19th splits in red, and the 20th to 24th in green. The blue part is a P-Tree with 15 leaves (up to S14); the combination of the blue and red parts is a P-Tree with 20 leaves (up to S19); the whole diagram is a P-Tree with 25 leaves. Figure format follows Figure 6.



## Table I.1: Asset Pricing Performance

Table format follows Table 1, where *Number of Leaves* equals 20. This table reports the robust asset pricing performances of P-Tree factors for the tuning parameter *Number of Leaves* equals 15 and 25. Panels A and B report for market-adjusted asset pricing P-Tree factor models, and Panels C and D report for the standard asset pricing P-Tree factor models.

| | Individual Stocks | | | | Portfolios | | | |
|---|---|---|---|---|---|---|---|---|
| | In-Sample | | Out-of-Sample | | | | | |
| | Tot | CS | Tot | CS | FF25 | Ind49 | Leaf20 | Leaf40 |
| | | | Panel A: Market-Adjusted P-Tree, No. Leaves = 15 | | | | | |
| PTree2 | 11.02 | 25.66 | 11.24 | 10.47 | 74.87 | 93.34 | 88.03 | 61.78 |
| PTree5* | 12.99 | 22.92 | 13.62 | 16.84 | 70.77 | 63.62 | 31.90 | 51.66 |
| | | | Panel B: Market-Adjusted P-Tree, No. Leaves = 25 | | | | | |
| PTree2 | 11.12 | 25.35 | 11.06 | 10.27 | 76.21 | 92.88 | 85.71 | 67.25 |
| PTree5* | 12.99 | 22.78 | 13.67 | 17.09 | 51.84 | 57.04 | 41.14 | 47.74 |
| | | | Panel C: P-Tree, No. Leaves = 15 | | | | | |
| PTree1 | 9.54 | 2.72 | 10.79 | 6.63 | 89.28 | 83.50 | -496.77 | -153.89 |
| PTree5* | 13.07 | 26.06 | 13.27 | 13.09 | 78.83 | 42.35 | 46.81 | 61.10 |
| | | | Panel D: P-Tree, No. Leaves = 25 | | | | | |
| PTree1 | 9.58 | 1.94 | 10.85 | 7.37 | 88.78 | 82.91 | -336.48 | -85.54 |
| PTree5* | 13.03 | 25.52 | 13.53 | 14.46 | 70.84 | 21.21 | 25.98 | 53.14 |

## Table I.2: Investment Performance of Factors

Table format follows Table 2, where *Number of Leaves* equals 20. This table reports the robust investment performances of P-Tree factors for the tuning parameter *Number of Leaves* equals 15 and 25. Panels A and B report for the market-adjusted asset pricing P-Tree factor models, Panels C and D report for the market-adjusted investment P-Tree factor models, and Panels E and F report for the investment P-Tree factor models.

| | In-Sample (1981-2000) | | | | | | Out-of-Sample (2001-2020) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MVE | | | $1/N$ | | | MVE | | | $1/N$ | | |
| | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ |
| **Panel A: Market-Adjusted P-Tree, No. Leaves = 15** | | | | | | | | | | | | |
| PTree2 | 2.08 | 1.59 | 1.80*** | 1.51 | 1.40 | 1.04*** | 0.41 | 0.35 | 0.45* | 0.51 | 0.60 | 0.26* |
| PTree5* | 1.47 | 3.42 | 1.40*** | 1.18 | 1.77 | 0.83*** | 0.98 | 1.95 | 0.98*** | 0.80 | 1.31 | 0.54*** |
| **Panel B: Market-Adjusted P-Tree, No. Leaves = 25** | | | | | | | | | | | | |
| PTree2 | 1.65 | 1.58 | 1.42*** | 1.25 | 1.32 | 0.81*** | 0.38 | 0.41 | 0.37* | 0.49 | 0.64 | 0.21* |
| PTree5* | 1.18 | 3.19 | 1.12*** | 1.02 | 1.69 | 0.70*** | 0.73 | 1.85 | 0.66*** | 0.69 | 1.11 | 0.40*** |
| **Panel C: Market-Adjusted Investment P-Tree, No. Leaves = 15** | | | | | | | | | | | | |
| PTree2 | 1.94 | 9.07 | 1.92*** | 1.35 | 2.05 | 0.98*** | 1.21 | 2.91 | 1.25*** | 0.93 | 1.48 | 0.64*** |
| PTree5* | 1.48 | 12.10 | 1.47*** | 0.87 | 1.99 | 0.64*** | 0.84 | 2.77 | 0.86*** | 0.46 | 0.95 | 0.24*** |
| **Panel D: Market-Adjusted Investment P-Tree, No. Leaves = 25** | | | | | | | | | | | | |
| PTree2 | 1.73 | 11.44 | 1.71*** | 1.24 | 1.91 | 0.87*** | 1.03 | 2.84 | 1.06*** | 0.84 | 1.32 | 0.54*** |
| PTree5* | 1.25 | 16.12 | 1.24*** | 0.75 | 2.11 | 0.56*** | 0.74 | 2.94 | 0.76*** | 0.50 | 1.30 | 0.33*** |
| **Panel E: Investment P-Tree, No. Leaves = 15** | | | | | | | | | | | | |
| PTree1 | 1.96 | 9.12 | 1.96*** | 1.96 | 9.12 | 1.96*** | 1.22 | 2.84 | 1.27*** | 1.22 | 2.84 | 1.27*** |
| PTree5* | 1.49 | 15.69 | 1.48*** | 1.03 | 6.31 | 0.98*** | 0.87 | 3.63 | 0.88*** | 0.61 | 3.19 | 0.60*** |
| **Panel F: Investment P-Tree, No. Leaves = 25** | | | | | | | | | | | | |
| PTree1 | 1.75 | 11.60 | 1.75*** | 1.75 | 11.60 | 1.75*** | 1.03 | 2.76 | 1.08*** | 1.03 | 2.76 | 1.08*** |
| PTree5* | 1.28 | 15.57 | 1.27*** | 0.70 | 4.22 | 0.63*** | 0.73 | 2.96 | 0.73*** | 0.45 | 2.35 | 0.40*** |

# II Robustness to Economic Restrictions

Avramov et al. (2022) highlight how many ML models are not robust to imposing various economic restrictions. Luckily, the performance of P-Trees survives these restrictions. Next, we report our findings concerning arguably the most important restriction—the exclusion of small and illiquid stocks.

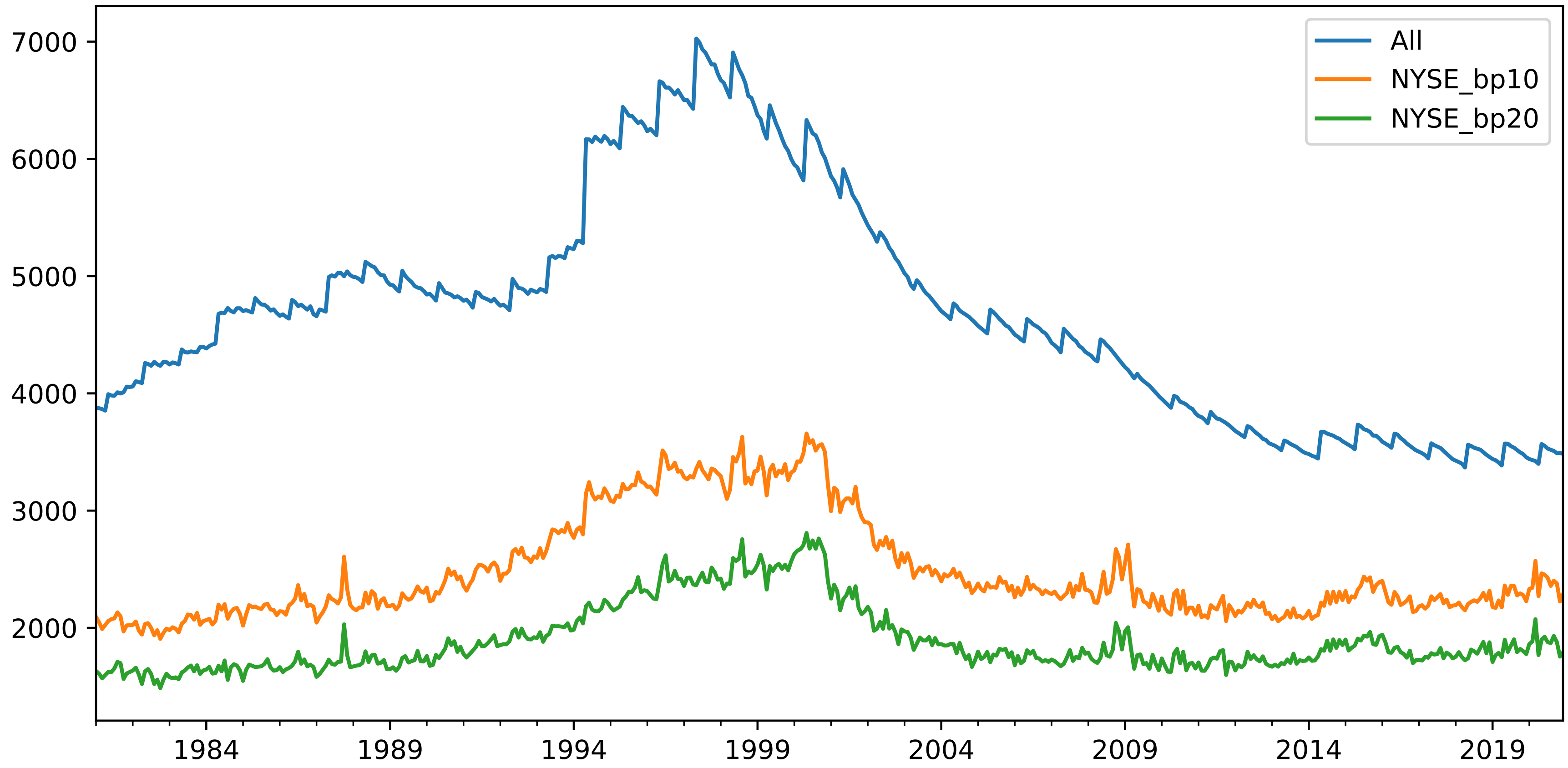
According to Hou et al. (2020), 65% of known anomalies cannot pass the test hurdle of $|t| \geq 1.96$, mainly because the original studies rely on equal-weighted returns, small stocks, and NYSE-AMEX-NASDAQ breakpoints for portfolio sorting and cross-sectional regression. In this paper, the leaf basis portfolio returns are value-weighted. Furthermore, to alleviate the concerns that P-Tree trading strategies rely on small and illiquid stocks, we revisit the empirical exercises, excluding the small stocks with NYSE breakpoints. We consider the investment pool of the stocks larger than 10% (20%) NYSE breakpoints, which removes about 47% (60%) observations in the original data. Figure I.4 reports the number of stocks and the number of stocks above 10% and 20% NYSE breakpoints in each month.

Table I.2 shows the investment performance of investment P-Tree factors, excluding the small stocks by NYSE size breakpoints of 10% and 20%. Compared to the numbers in Table 2 Panel B, the Sharpe ratio of the P-Tree five-factor model decreases from 2.96 to 1.90 and 1.46 by excluding the stocks smaller than 10% and 20% NYSE breakpoints. The Jensen's alpha decreases from 0.78% to 0.33% and 0.25% for 10% and 20% NYSE breakpoints, respectively. In Table 2 Panel C, the investment P-Tree five-factor model (without Market) has a Sharpe ratio of 2.78 and alpha of 0.82 in the test period. By contrast, the best models after excluding small stocks are with two factors, selected by the Squared Sharpe ratio test of Barillas and Shanken (2017). The Sharpe ratio decreases to 1.93 (1.32), and the alpha decreases to 0.38% (0.26%), by excluding stocks smaller than NYSE 10% (20%) breakpoints.

In summary, the investment gains are lower by excluding more small stocks. However, P-Tree factors still capture sizeable investment gains, even excluding about half of the stocks. The out-of-sample Sharpe ratios are about 1.9 (1.4) if we only invest in the top 52% (40%) largest stocks. Jensen's alphas' are also statistically significant after excluding small stocks.

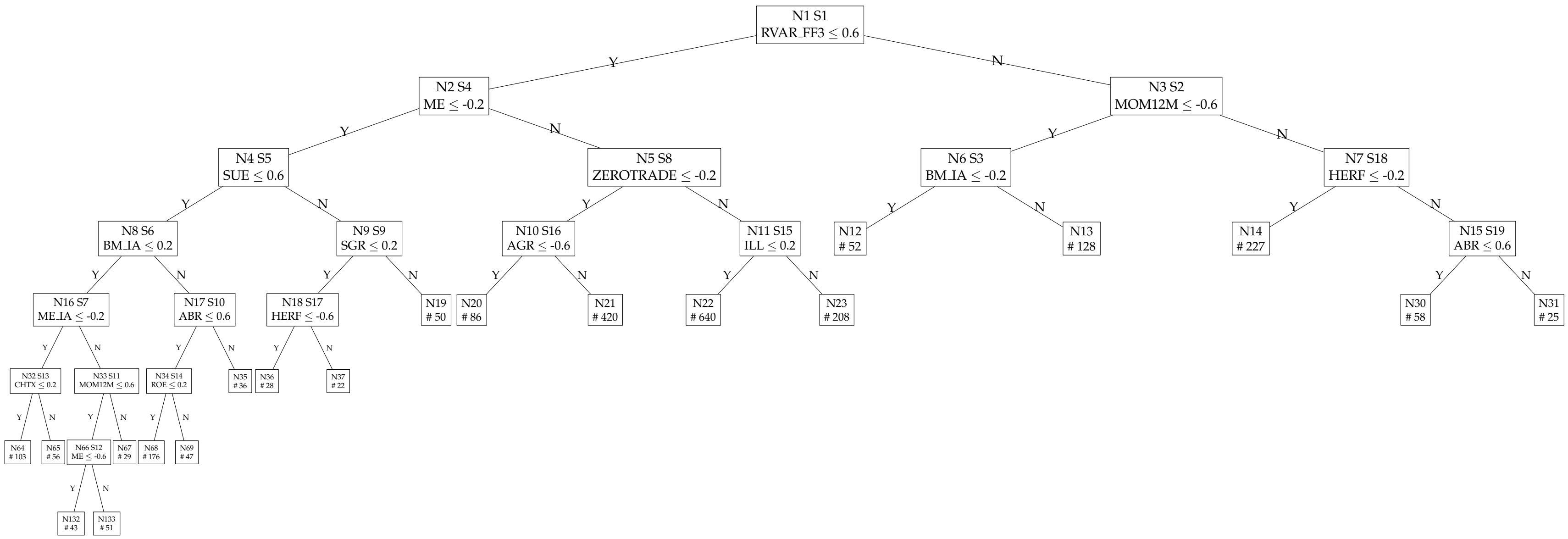
Figure I.4: Number of Stocks by Month

This figure displays the monthly number of stocks of the original sample and the filtered sample by NYSE breakpoints at 10% and 20%. The date range is 1981 to 2020.



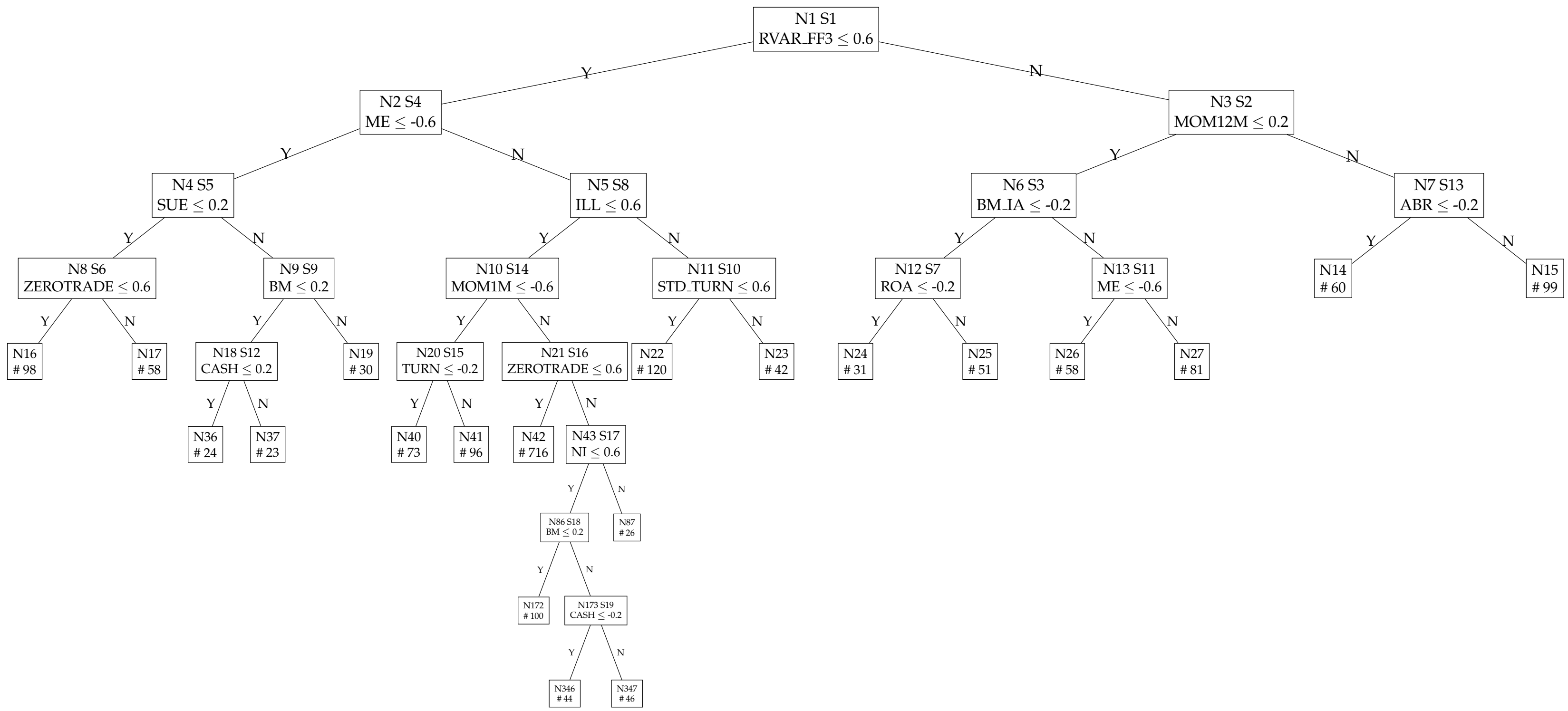
# Figure I.5: Market-Adjusted Investment Panel Tree
# Excluding small stocks, NYSE breakpoints 10%

Figure format follows Figure 6. We exclude the stocks smaller than the 10% NYSE size breakpoints.



# Figure I.6: Market-Adjusted Investment Panel Tree
# Excluding small stocks, NYSE breakpoints 20%

Figure format follows Figure 6. We exclude the stocks smaller than the 20% NYSE size breakpoints.



## Table I.3: Investment Performance of Factors

Table format follows Table 2. This table reports the investment performances of investment P-Tree factors, excluding the small stocks by NYSE size breakpoints 10% and 20%.

| | In-Sample (1981-2000) | | | | | | Out-of-Sample (2001-2020) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MVE | | | 1/N | | | MVE | | | 1/N | | |
| | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ |
| Panel A: Market-Adjusted Investment P-Tree, larger than NYSE 10% size breakpoint | | | | | | | | | | | | |
| PTree2 | 0.96 | 7.18 | 0.93*** | 0.85 | 1.31 | 0.48*** | 0.46 | 1.98 | 0.45*** | 0.55 | 0.84 | 0.23*** |
| PTree5* | 0.75 | 9.30 | 0.73*** | 0.59 | 1.14 | 0.30*** | 0.34 | 1.90 | 0.33*** | 0.40 | 0.77 | 0.15*** |
| Panel B: Market-Adjusted Investment P-Tree, larger than NYSE 20% size breakpoint | | | | | | | | | | | | |
| PTree2 | 0.87 | 5.87 | 0.84*** | 0.81 | 1.23 | 0.43*** | 0.32 | 1.41 | 0.32*** | 0.48 | 0.75 | 0.17*** |
| PTree5* | 0.67 | 7.67 | 0.65*** | 0.59 | 1.10 | 0.29*** | 0.24 | 1.46 | 0.25*** | 0.37 | 0.73 | 0.13*** |
| Panel C: Investment P-Tree, larger than NYSE 10% size breakpoint | | | | | | | | | | | | |
| PTree1 | 0.97 | 7.64 | 0.96*** | 0.97 | 7.64 | 0.96*** | 0.45 | 1.88 | 0.46*** | 0.45 | 1.88 | 0.46*** |
| PTree2* | 0.81 | 8.44 | 0.79*** | 0.64 | 6.60 | 0.62*** | 0.37 | 1.93 | 0.38*** | 0.29 | 1.74 | 0.29*** |
| Panel D: Investment P-Tree, larger than NYSE 20% size breakpoint | | | | | | | | | | | | |
| PTree1 | 0.88 | 6.24 | 0.87*** | 0.88 | 6.24 | 0.87*** | 0.31 | 1.28 | 0.33*** | 0.31 | 1.28 | 0.33*** |
| PTree2* | 0.73 | 6.81 | 0.71*** | 0.60 | 5.84 | 0.58*** | 0.25 | 1.32 | 0.26*** | 0.20 | 1.18 | 0.21*** |

# III Application of P-Tree to the U.S. Corporate Bonds

This section complements our empirical exercises by applying P-Tree to U.S. corporate bonds.

## III.1 Data

We follow Feng et al. (2022) to calculate the monthly returns and 45 characteristics as listed in Table I.4. Specifically, we obtain corporate bond issuance and rating information from FISD and the transaction data from the enhanced version of TRACE with standard filters.

Our corporate bond sample period spans from July 2004 to December 2020. We use the first ten years, from July 2004 to June 2014, for training and the remaining period, from July 2014 to December 2020, for testing. The average and median monthly numbers of corporate bond observations are 4,528 and 4,574 in the training sample and 5,337 and 5,432 in the testing sample, respectively. The individual corporate bond returns in the training sample are subjected to cross-sectional winsorization on 5% and 95% quantiles to remove outliers. Returns in the test sample are not winsorized. Similar to the equity data, bond characteristics are standardized cross-sectional in the $[-1, 1]$ range.

## III.2 Empirical Results

**Implementation** Similar to the study of U.S. stocks, in this section, we apply P-Tree to generate the SDF for U.S. corporate bonds with asset pricing and investment split criteria. The tuning parameters and shrinkage parameters are selected the same. Specifically, the number of leaves is 20, the minimal leaf size is 10, and the mean and covariance shrinkage parameters in (3) are $10^{-4}$.

Figure I.7 depicts the market-adjusted P-Tree with asset pricing criterion in the training sample. Endogenously, the P-Tree first splits along the systematic skewness of bond (`COSKEW`) at -0.6 (20% quantile) and then short-term reversal (`STR`) at -0.2 (40% quantile). Figure I.8 shows the P-Tree with asset pricing criterion, which splits on the dollar trading volume of bond (`VOLUME`) at -0.6 (20% quantile) and then duration (`DUR`) at 0.2 (60% quantile). Figure I.9 shows the market-adjusted investment P-Tree, which splits on the liquidity risk factor beta (`LRF_BETA`) at 0.6 (80% quantile) and then short-term reversal (`STR`) at 0.6 (80% quantile).

**Asset Pricing Performance** Furthermore, we generate multi-factor models by boosting P-Tree. For comparison, we include the well-known observable factors, including CAPM (corporate bond market factor), BBW4 four-factor model in Bai et al. (2019), and FF5-BOND five-factor model in Fama and French (1993), and state-of-art latent factor models, including RP-PCA with five factors in Lettau and Pelger (2020b) and IPCA with five factors in Kelly et al. (2019).

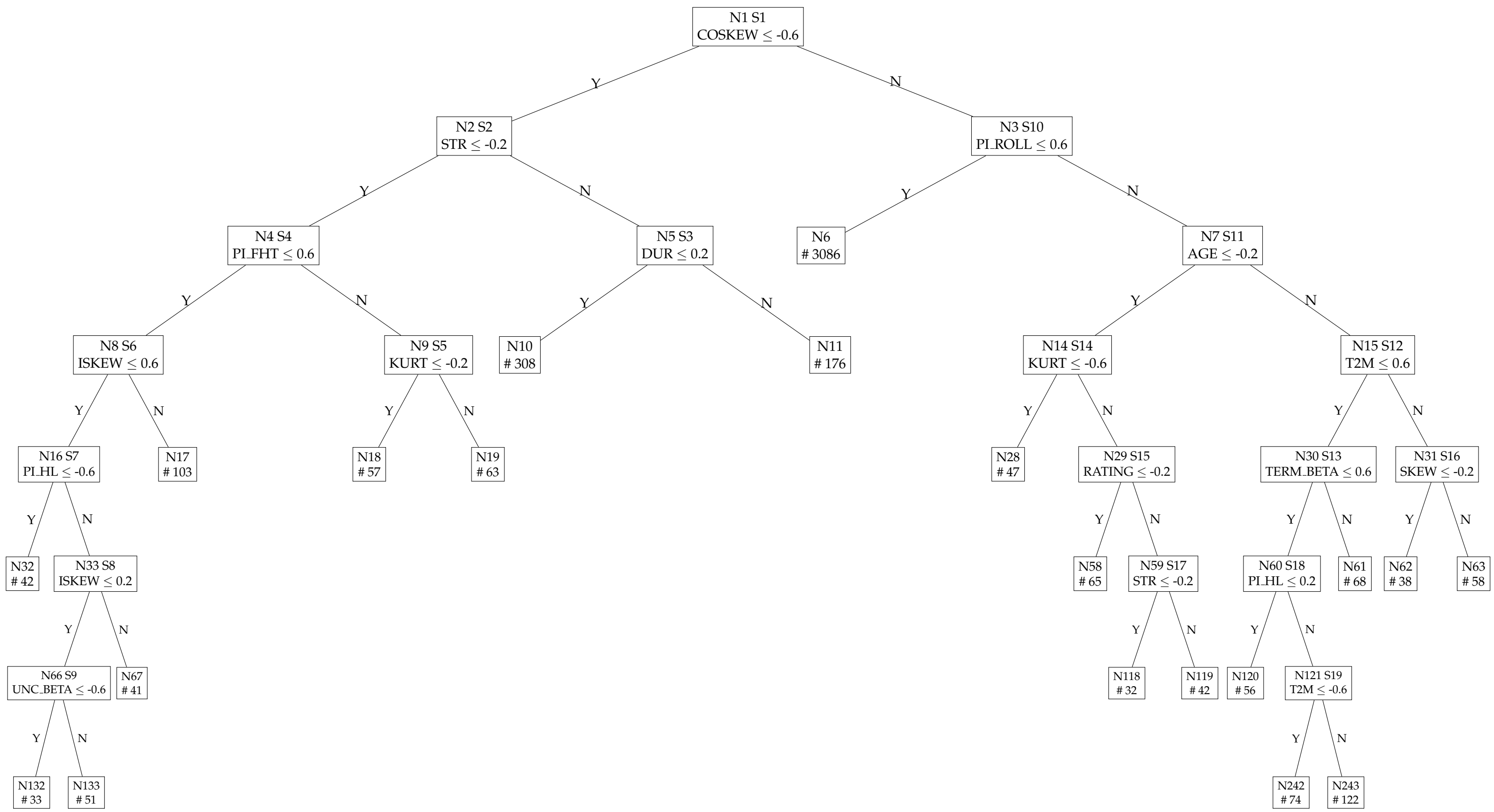
Table I.5 shows the asset pricing performance of P-Tree and alternatives. The three-factor model in Panel A (the corporate bond market factor plus two P-Tree factors) has excellent pricing performances, and is comparable to BBW4 and FF5-BOND. Excluding the market factor, the four-factor model in Panel B has higher total $R^2$ and cross-sectional $R^2$ for out-of-sample bond pricing, but lower cross-sectional $R^2$ for Rating-Duration 25 portfolios and 30 industry portfolios. The observable factor models are stronger in pricing corporate bond portfolios than the latent factor models, while the latent factor models are better in pricing individual corporate bonds. Overall, the pricing performance of P-Tree factors is comparable to that of observable and latent factors.

**Investment Performance** Table I.6 shows the investment performances of corporate bond P-Tree and other benchmark factors. In Panels A, B, C, and D, the four specifications of P-Tree factors show substantial out-of-sample risk-adjusted investment gains.

Specifically, in Panel D, the one-factor model of P-Tree under the investment criterion has 0.29% monthly expected returns and 1.40 annualized Sharpe ratio, for the out-of-sample period. Moreover, in Panel B, the four-factor model of P-Tree under asset pricing criterion has 0.18% $\alpha$ and 1.64 annualized Sharpe ratio, which performs well in pricing corporate bond returns and investment. Jensen's $\alpha$'s are significantly positive for both the MVE strategies and the 1/N strategies of the P-Tree factors. By contrast, the other benchmark factors are weaker than P-Tree factors in Sharpe ratio and $\alpha$; especially for the 1/N strategies, none of the benchmark factors has significantly positive $\alpha$. Overall, P-Tree factors are exceptionally attractive in direct investment.

## Figure I.7: Corporate Bond: Market-Adjusted Panel Tree

This figure displays the market-adjusted asset pricing P-Tree for corporate bonds, trained from July 2004 to June 2014. We show splitting characteristics and split rule values for each parent node. The node numbers (N#) and splitting order numbers (S#) are also printed on each parent node. We have included the median monthly number of bonds in the leaf basis portfolios. The figure format follows Figure 5. The description of bond characteristics are listed in Table I.4.



## Figure I.8: Corporate Bond: Panel Tree

This figure displays the asset pricing P-Tree for corporate bonds trained from July 2004 to June 2014. We show splitting characteristics and split rule values for each parent node. The node numbers (N#) and splitting order numbers (S#) are also printed on each parent node. We have included the median monthly number of bonds in the leaf basis portfolios. The figure format follows Figure 5. The description of bond characteristics are listed in Table I.4.

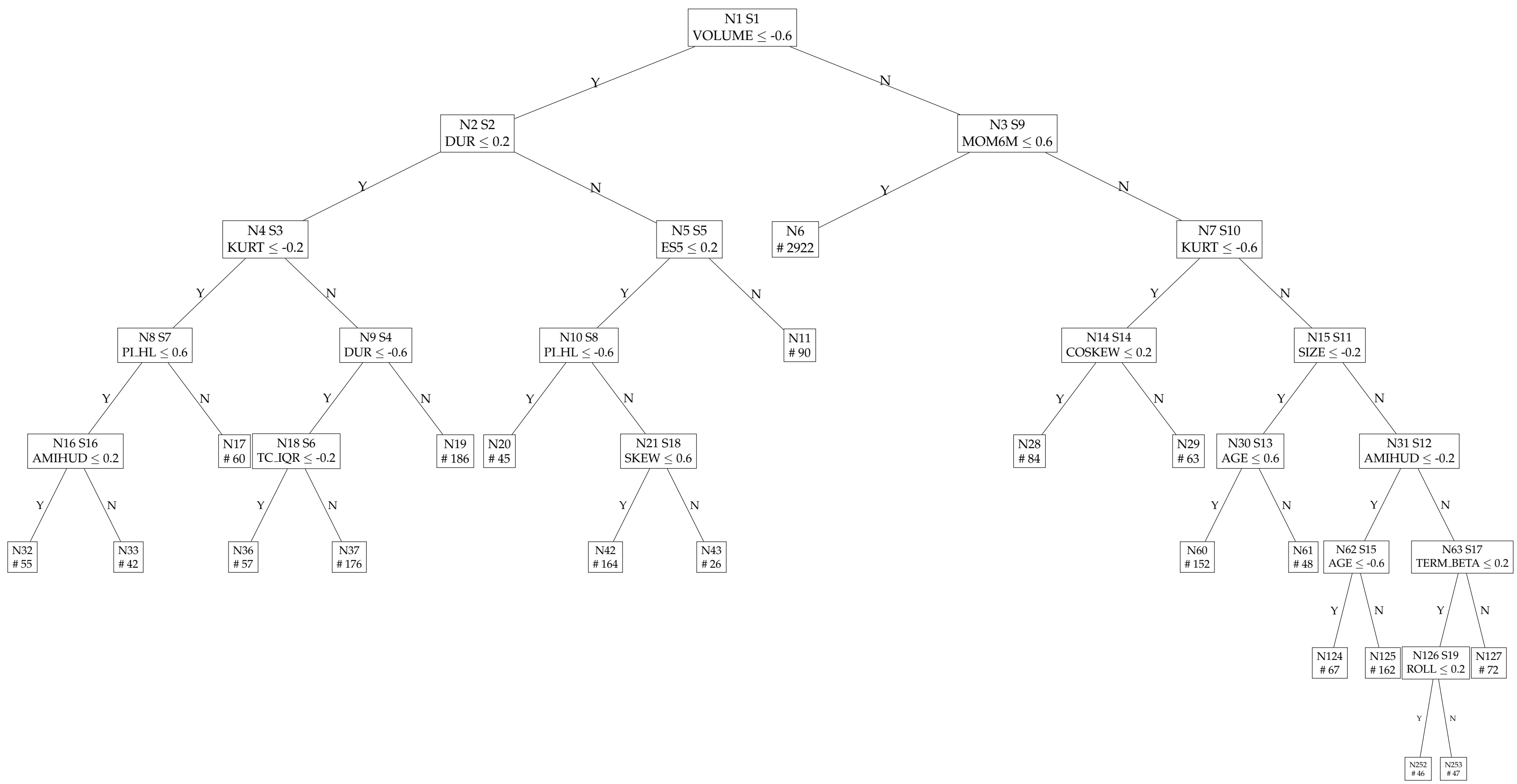


Figure I.9: Corporate Bond: Market-Adjusted Investment P-Tree

This figure displays the market-adjusted investment-guided corporate bond P-Tree with market factor trained from July 2004 to June 2014. We show splitting characteristics and split rule values for each parent node. The node numbers (N#) and splitting order numbers (S#) are also printed on each parent node. We have included the median monthly number of bonds in the leaf basis portfolios. The figure format follows Figure 5. descriptions of bond characteristics are listed in Table I.4.

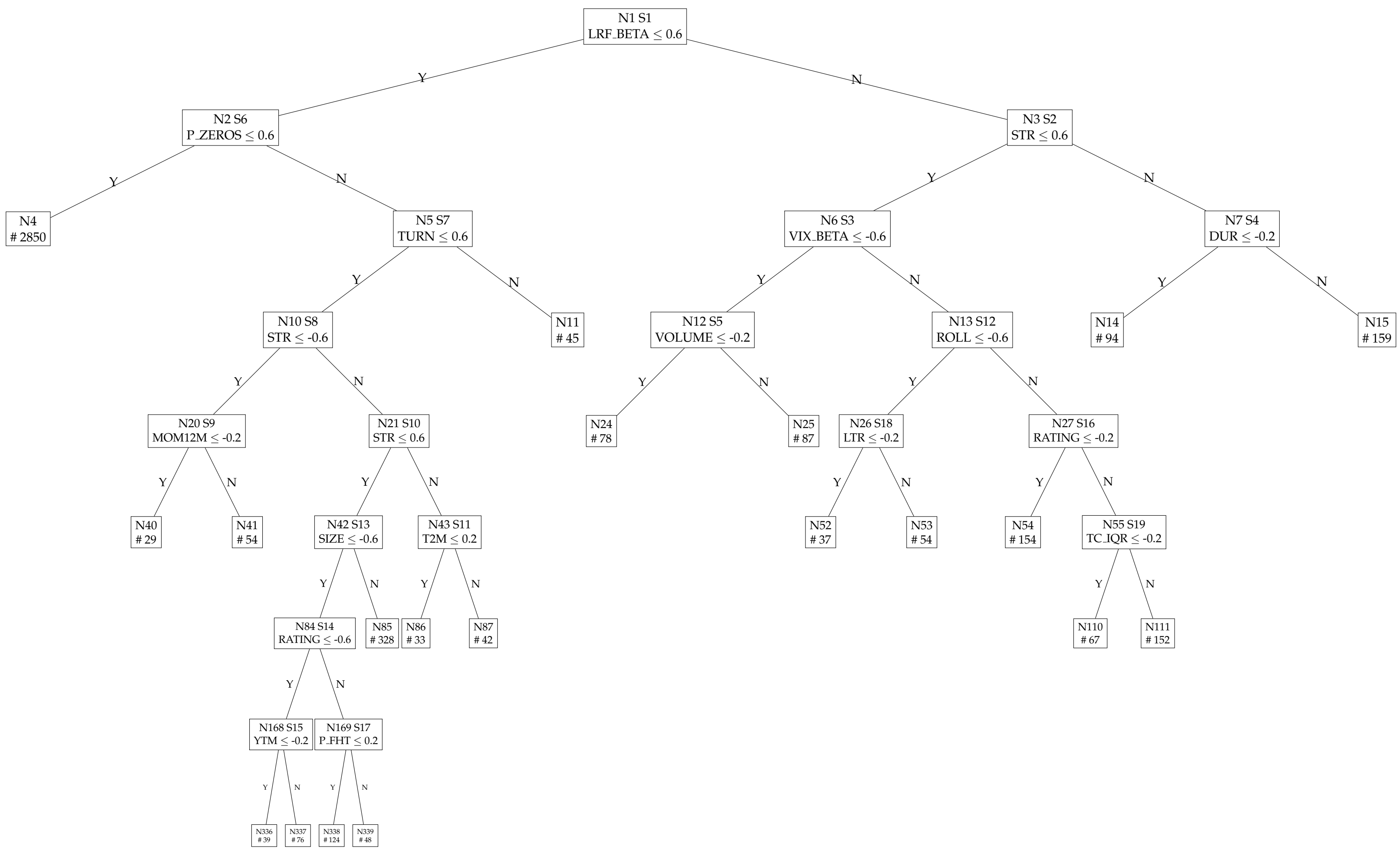


Table I.4: Corporate Bond Characteristics

This table lists the description of 45 corporate bond characteristics used in the empirical study.

| No. | Characteristics | Description |
|---|---|---|
| 1 | AGE | Time since issuance |
| 2 | LIQ_AMIHUD | Amihud liquidity |
| 3 | LIQ_BPW | Liquidity measure of Bao, Pan, and Wang (2011) |
| 4 | COSKEW | Systematic skewness |
| 5 | CRF_BETA | Credit risk beta controlling bond market factor |
| 6 | DEF_BETA | DEF (Fama and French, 1993) factor beta |
| 7 | DRF_BETA | Downside risk beta controlling bond market factor |
| 8 | DUR | Duration |
| 9 | ES10 | Expected shortfall 10% |
| 10 | ES5 | Expected shortfall 5% |
| 11 | ISKEW | Idiosyncratic skewness |
| 12 | KURT | Return Kurtosis |
| 13 | LIQ_BETA | Liquidity beta of Lin, Wang, and Wu (2011) |
| 14 | LRF_BETA | Liquidity risk beta controlling bond market factor |
| 15 | LTR | Long-term reversal from t-13 to t-48 months |
| 16 | MKT_BETA | Market beta |
| 17 | MKT_RVAR | Market residual variance |
| 18 | MOM12m | Momentum from t-7 to t-12 months |
| 19 | MOM6m | Momentum from t-2 to t-6 months |
| 20 | P_FHT | Modified illiquidity measure based on zero returns |
| 21 | BAS | Bid-Ask Spread |
| 22 | ZEROTRADE | Zero Trading Days |
| 23 | PI_FHT | An extended FHT measure based on zero returns |
| 24 | PI_HL | An extended High-low spread estimator |
| 25 | PI_ROLL | An extended Roll's measure |
| 26 | RANGE | Daily Return Range |
| 27 | RATING | Credit Ratings |
| 28 | LIQ_ROLL | Roll's liquidity |
| 29 | SIZE | Amount outstanding |
| 30 | SKEW | Return Skewness |
| 31 | STD_LIQ | Standard deviation of Amihud daily liquidity |
| 32 | STR | Short-term reversal in t-1 months |
| 33 | T2M | Time to maturity |
| 34 | TC_IQR | Interquartile range |
| 35 | TERM_BETA | TERM (Fama and French, 1993) factor beta |
| 36 | TERM_DEF_RVAR | TERM-DEF (Fama and French, 1993) residual variance |
| 37 | TRADE | Number of trades |
| 38 | TURN | Turnover |
| 39 | UNC_BETA | Macroeconomic Uncertainty Beta |
| 40 | VaR10 | Value-at-risk 10% over past 3 years |
| 41 | VaR5 | Value-at-risk 5% over past 3 years |
| 42 | VARIANCE | Return Variance |
| 43 | VIX_BETA | VIX index beta |
| 44 | VOLUME | Dollar trading volume |
| 45 | YTM | Yield-to-maturity |

Table I.5: Corporate Bond: Asset Pricing Performance

This table reports the asset pricing performances for corporate bonds. "Tot" (total $R^2$ %) and "CS" (cross-sectional $R^2$ %) in equations (15) and (16) are measures for individual bond returns. The in-sample period is from July 2004 to June 2014, and the out-of-sample period is from July 2014 to December 2020. We also report cross-sectional $R^2$ % in (17), using the factor models in the rows to price the test asset portfolios in the columns. "RD25" indicates the $5 \times 5$ portfolios sorted on rating and duration, "Ind30" indicates the 30 industry portfolios of corporate bonds, "Leaf20" indicates the 20 basis portfolios from Figure I.7, and "Leaf40" indicates the 40 basis portfolios from Figure I.7 and I.8. Panel A shows results for the market-adjusted panel tree model under asset pricing criterion with #factors. Specifically, "PTree2" indicates a two-factor model of the market factor and a P-Tree factor. "PTree3*" indicates a three-factor model of the market factor and two P-Tree factors. "*' indicates the optimal number of factors selected by the squared Sharpe ratio test of Barillas and Shanken (2017). Panel B reports P-Tree factor models under the asset pricing criterion. Panel C provides comparisons for other benchmark models, introduced in Section III.2.

| | Individual Bonds | | | | Portfolios | | | |
|---|---|---|---|---|---|---|---|---|
| | In-Sample | | Out-of-Sample | | Entire Sample | | | |
| | Tot | CS | Tot | CS | RD25 | Ind30 | Leaf20 | Leaf40 |
| **Panel A: Market-Adjusted P-Tree** | | | | | | | | |
| PTree2 | 29.53 | 6.57 | 15.84 | -12.33 | 78.37 | 93.47 | 89.23 | 88.59 |
| PTree3* | 34.73 | 22.02 | 23.36 | 14.11 | 70.78 | 95.08 | 88.46 | 91.37 |
| **Panel B: P-Tree** | | | | | | | | |
| PTree1 | 30.87 | 20.86 | 21.62 | 20.34 | 88.80 | 96.84 | 91.25 | 93.89 |
| PTree4* | 38.36 | 23.00 | 26.15 | 23.45 | 29.97 | 65.66 | 89.45 | 89.84 |
| **Panel C: Other Benchmark Models** | | | | | | | | |
| CAPM | 21.35 | 20.83 | 17.72 | 6.59 | 88.80 | 96.84 | 91.25 | 93.89 |
| BBW4 | 24.35 | 18.59 | 23.50 | 23.34 | 69.48 | 91.30 | 91.12 | 93.97 |
| FF5-BOND | 23.91 | 15.26 | 22.74 | 19.53 | 74.58 | 92.56 | 76.44 | 83.38 |
| RP-PCA5 | 38.03 | 22.95 | 28.76 | 23.86 | 57.93 | 90.44 | 89.43 | 91.79 |
| IPCA5 | 49.08 | 24.67 | 33.71 | 24.21 | -2.62 | 86.77 | 70.75 | 81.28 |

## Table I.6: Corporate Bond: Investing P-Tree Factors

This table reports the investment performance of the corporate bond factors and other benchmark factors. We report the monthly average return and Jensen's $\alpha$ (%), the annualized Sharpe ratio for the factors' mean-variance efficient (MVE) and equal-weighted ($1/N$) portfolios. The "*" indicates the optimal number of factors selected by the squared Sharpe ratio test of Barillas and Shanken (2017). For $t$-statistics *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The table format follows Table 2.

| | In-Sample (200407-201406) | | | | | | Out-of-Sample (201407-202012) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MVE | | | $1/N$ | | | MVE | | | $1/N$ | | |
| | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ | AVG | SR | $\alpha$ |
| **Panel A: Market-Adjusted P-Tree** | | | | | | | | | | | | |
| PTree2 | 0.45 | 1.44 | 0.25*** | 0.42 | 1.33 | 0.18*** | 0.26 | 1.04 | 0.10 | 0.30 | 1.06 | 0.07 |
| PTree3* | 0.47 | 2.30 | 0.33*** | 0.43 | 1.63 | 0.21*** | 0.26 | 1.24 | 0.11** | 0.30 | 1.12 | 0.07** |
| **Panel B: P-Tree** | | | | | | | | | | | | |
| PTree1 | 0.33 | 0.58 | -0.08 | 0.33 | 0.58 | -0.08 | 0.38 | 0.97 | 0.05 | 0.38 | 0.97 | 0.05 |
| PTree4* | 0.52 | 2.60 | 0.40*** | 0.48 | 1.80 | 0.26*** | 0.29 | 1.64 | 0.16*** | 0.32 | 1.34 | 0.13*** |
| **Panel C: Market-Adjusted Investment P-Tree** | | | | | | | | | | | | |
| PTree2* | 0.51 | 3.70 | 0.42*** | 0.45 | 1.92 | 0.25*** | 0.31 | 1.37 | 0.15*** | 0.34 | 1.17 | 0.09*** |
| **Panel D: Investment P-Tree** | | | | | | | | | | | | |
| PTree1* | 0.54 | 4.59 | 0.51*** | 0.54 | 4.59 | 0.51*** | 0.29 | 1.40 | 0.18*** | 0.29 | 1.40 | 0.18*** |
| **Panel E: Other Benchmark Models** | | | | | | | | | | | | |
| BBW4 | 0.58 | 1.49 | 0.37*** | 0.55 | 1.41 | 0.31*** | 0.36 | 1.03 | 0.09 | 0.33 | 0.93 | 0.06 |
| FF5-BOND | 0.52 | 0.89 | 0.21 | 0.28 | 0.63 | 0.08 | 0.71 | 1.23 | 0.34** | 0.16 | 0.35 | -0.11 |
| RP-PCA5 | 0.76 | 1.40 | 0.46*** | 0.88 | 1.12 | 0.25*** | 0.22 | 0.78 | 0.12 | 0.44 | 0.90 | 0.10 |
| IPCA5 | 0.48 | 2.36 | 0.41*** | 0.36 | 0.90 | 0.09 | 0.28 | 1.50 | 0.15*** | 0.27 | 1.00 | 0.06 |