Recovering Missing Firm Characteristics with Attention-based Machine Learning [*]

Heiner Beckmeyer[†] Timo Wiedemann[‡]

This version: November 8, 2022

Abstract

Firm characteristics are often missing. Our study is devoted to their recovery, drawing on the informational content of other – observed – characteristics, their past observations, and information from the cross-section of other firms. We adapt state-of-the-art advances from natural language processing to the case of financial data and train a large-scale machine learning model in a self-supervised environment. Using the uncovered latent structure governing characteristics, we show that our model beats competing methods, both empirically and in simulated data. Based on the completed dataset, we show that average returns to many characteristic-sorted long-short portfolios are likely lower than previously thought.

Keywords: Machine Learning, Matrix Completion, Missing Data, Attention, Big Data, Risk Factors

JEL classification: G10, G12, G14, C14, C55

[*] We want to thank Turan G. Bali, Bernard Dumas, Andrea Barbon, Florian Weigert, Jonathan Lewellen, Neil Pearson, Julian Thimme, Simon Rottke, Dmitriy Muravyev, Ilias Filippou, Andrew Chen, Yuri Tserlukevich, Christian Schlag, Johannes Kriebel, Jörn Debener and Nicole Branger, as well as participants at the 5th Asset Pricing Conference by LTI@UniTo, the annual meeting of the Northern Finance Association 2022, Frontiers of Factor Investing Conference 2022, International Conference on Economic Modeling and Data Science 2022, VfS annual meeting 2022 on big data in economics, the annual meeting of the International Finance and Banking Society 2022, the Econometric Society European Summer Meeting 2022, Hofstra Financial Regulations & Technology Conference 2022, the 2022 Economics of Financial Technology Conference in Edinburgh, and the annual meeting of the German Statistical Society 2022 for valuable feedback and discussions, as well as seminar participants at Goethe University and the University of Munster.

[†] School of Business and Economics, Finance Center Münster, University of Münster, Universitätsstr. 14-16, 48143 Münster, Germany. E-mail: heiner.beckmeyer@wiwi.uni-muenster.de

[‡] School of Business and Economics, Finance Center Münster, University of Münster, Universitätsstr. 14-16, 48143 Münster, Germany. E-mail: timo.wiedemann@wiwi.uni-muenster.de

Recovering Missing Firm Characteristics with Attention-based Machine Learning

This version: November 8, 2022

Abstract

Firm characteristics are often missing. Our study is devoted to their recovery, drawing on the informational content of other – observed – characteristics, their past observations, and information from the cross-section of other firms. We adapt state-of-the-art advances from natural language processing to the case of financial data and train a large-scale machine learning model in a self-supervised environment. Using the uncovered latent structure governing characteristics, we show that our model beats competing methods, both empirically and in simulated data. Based on the completed dataset, we show that average returns to many characteristic-sorted long-short portfolios are likely lower than previously thought.

1. Introduction

A large amount of economic research uses the combined database by the Center for Research in Security Prices (CRSP) and Compustat for firm-level information. While it is certainly the "gold standard of stock market databases",[1] the provided data is far from complete. Table 1 showcases the severity of the issue of missing firm characteristics using a large panel of 143 firm-level characteristics from the dataset provided by Jensen, Kelly, and Pedersen (2021). Among the twelve characteristic groups, formed to capture similar aspects of a company, for example about the company's quality, return momentum or size, the median characteristic is missing for 12%–22% of firm×month observations. Characteristics with the highest degree of missingness are missing between 19% and 56% of their entries over the sample period from 1972–2020. Even those characteristics that are most often available still lack a considerable chunk of information. Depending on the characteristic's group, they may be missing for as few as 1% (for the company's market capitalization, as part of group *Size*), but also as many as 16% of all firm×month observations for *Accruals*.

Statistical tests in asset pricing typically require long time-series and large cross-sections. However, due to missing information, the panel of stock characteristics is limited both in the time series and the cross section of stocks. By removing or naively imputing missing entries of characteristics, a researcher throws away valuable information, for example about the temporal properties of risk factors based on these characteristics. At the same time, she implicitly assumes that the results obtained from the subsample of stocks with available characteristics generalize to those firm×month observations with missing information. For example, factor premia estimated from characteristic-sorted portfolios may not be representative of the full universe of stocks, if the target characteristic is

[1] https://www.crsp.org/files/Booth_Magazine_Winter_2011_CRSP_Index_Feature.pdf.

Table 1: Summary of Main Findings.

The table highlights our main findings. First, it shows how often characteristics clustered into twelve themes, designed to capture similar aspects of a firm (Jensen et al., 2021), are missing. Second, it gives our model's accuracy in reconstructing the characteristics, which we measure as the expected deviation from the true value, when characteristics are cross-sectionally discretized into percentiles (see Section 3, lower values are better). Finally, we show how using the completed dataset changes annualized returns of high-minus-low characteristic-sorted portfolios (ΔHmL). Along the median (Q_{50}), we also report results for the 10th (Q_{10}) and 90th (Q_{90}) percentile measured across characteristics within a theme.

	Q_{10}	Q_{50}	Q_{90}	Q_{10}	Q_{50}	Q_{90}	Q_{10}	Q_{50}	Q_{90}	Q_{10}	Q_{50}	Q_{90}
	Accruals$_{N=5}$			Debt issuance$_{N=6}$			Investment$_{N=21}$			Leverage$_{N=11}$		
Miss.	0.16	0.16	0.19	0.11	0.20	0.32	0.13	0.20	0.31	0.08	0.19	0.56
Acc.	3.42	4.06	4.85	3.15	4.00	6.60	2.69	3.62	5.39	1.37	2.79	5.05
ΔHmL	−4.85	−2.64	−0.95	−2.87	−2.02	−0.38	−3.93	−2.12	−0.11	−1.01	−0.25	−0.00
	Low risk$_{N=21}$			Momentum$_{N=7}$			Profit growth$_{N=12}$			Profitability$_{N=12}$		
Miss.	0.12	0.17	0.36	0.12	0.17	0.19	0.14	0.22	0.32	0.10	0.16	0.28
Acc.	2.16	3.61	7.89	2.31	3.34	8.56	2.35	4.94	7.35	1.90	2.56	4.48
ΔHmL	−5.82	−0.65	1.37	−3.13	−1.05	−0.27	−1.58	−0.80	−0.01	−4.19	−0.75	0.63
	Quality$_{N=17}$			Seasonality$_{N=7}$			Size$_{N=5}$			Skewness$_{N=6}$		
Miss.	0.08	0.12	0.33	0.14	0.20	0.35	0.01	0.12	0.41	0.14	0.17	0.19
Acc.	1.54	1.80	7.54	2.81	6.97	18.11	1.30	1.90	2.14	5.16	8.52	11.02
ΔHmL	−3.99	−0.58	0.48	−1.46	−0.37	0.85	−3.77	0.37	1.33	−1.00	−0.66	0.69

more often available for large firms.[2] Table 1 showcases that our proposed method to recover missing firm characteristics is highly accurate, on average predicting the correct cross-sectional percentile with a deviation of at most five percentiles. At the same time, correctly accounting for missing firm characteristics has a real effect on characteristic-sorted portfolio returns in empirical asset pricing. For most characteristics, the average high-minus-low return spread decreases using a dataset for which missing entries have been completed by our machine learning algorithm. Taking characteristics that proxy for a company's *Investment* behavior as an example, we find that post-completion average value-weighted factor returns decrease by −3.93% to −0.11% per year.

Our study is devoted to recovering missing firm characteristics, drawing on the informa-

[2] We show in Figure 4 that small firms are on average missing much more information than their larger counterparts. At the same time, missingness restricts the available historic sample, see Figure 3.

tional content of other – observed – characteristics, past observations of characteristics, and information from the cross-section of other firms. We adapt state-of-the-art advances from the field of natural language processing to the case of financial data and train a large-scale machine learning model in a self-supervised environment. We use the uncovered latent structure governing firm characteristics to recover missing entries and show that our model comfortably beats competing methods, both empirically and in simulated data. We furthermore quantify the model's uncertainty in its predictions and stress the importance of considering missing information in firm panels by showing that average returns to many characteristic-sorted long-short portfolio are likely lower than previously thought.

1.1. Our Findings

Masked language models randomly flag a certain fraction of words in an input sentence for reconstruction. The model consequently learns the context in which words are placed in a sentence. We apply this idea to the case of missing firm characteristics and carefully customize the model to match the task of recovering missing financial data. By asking the model to reconstruct a certain set of masked characteristics, we force it to extract a suitable context of information about other characteristics, their historical evolution, and information from other firms, which uncovers the latent structure governing firm characteristics. Our main building block is the attention mechanism used in the so-called "Transformer" architecture popularized by Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin (2017). Attention computes the similarity between a target search query and internally-updated keys to a database. The resulting attention matrix provides a direct mapping between a target characteristic and historical, as well as cross-sectional information.

We apply our model to a large dataset from Jensen et al. (2021), which provides information about 143 firm characteristics for the years of 1962 through 2020. To assure that our model learns to recover missing firm characteristics by uncovering the latent structure that governs them, we train it using data of the most recent 15 years, for which the information set is most complete. We use firm characteristics, discretized to cross-sectional percentiles, as input to the model. The discretization has several advantages: i) it allows the model to explicitly deal with and learn from missing entries, ii) it deals with outliers and reduces potential noise in the input data and iii) it produces a probability distribution across the characteristic's percentiles, which we can use to gauge the model's uncertainty associated with each prediction.

In a simulation study we show that the proposed model setup can extract information for various different processes governing the evolution of firm characteristics in a single model and outperforms simple ad-hoc methods imputing a characteristic's last available value or the cross-sectional mean. The model accurately predicts masked entries for autoregressive and cross-sectionally dependent characteristics, as well as for characteristics driven by a combination of both processes. Furthermore, we can show that the model accurately recovers the temporal information patterns of autoregressive processes.

Our main metric to assess the model's accuracy for a large panel of characteristics and firms is the *expected percentile deviation (EPD)*, which measures the average absolute deviation from the true percentile. The expected percentile deviation amounts to 3.63 in the training and 4.67 in the testing sample. Separately considering accounting- and market-based, as well as hybrid characteristics, which draw on both types of information, we find that reconstructing the latter is easiest, but that the model performs well on each type of characteristic. We find that the model's accuracy is robust over time, as it is to the degree of information provided for a target firm×month observation, measured by the number of missing characteristics.

Zooming in on how well we can reconstruct individual characteristics, we find near perfect reconstruction for `age`, `market_equity`, or intermediate momentum `ret_12_7`, among many others. The characteristics that the model struggles the most with are those relying on daily data, as the model operates at the monthly frequency and is never fed intra-month information. We can further show that the model's estimations are unbiased.

What's the merit of setting up a model that simultaneously recovers missing entries for all 143 characteristics? Given prior knowledge about how a target characteristic may evolve, researchers and investors alike may use bespoke approaches to impute missing values. We consider a wide range of competing approaches that harness the informational content of a characteristic's own past, different slices of the cross-section of firms – for example of firms within the same industry or of similar size – as well as themes of characteristics that represent similar aspects of a firm, for example its past debt issuance, overall accounting quality, or return momentum. Across all twelve characteristic themes that we consider, our model produces the lowest EPD, showcasing that the interdependencies between input characteristics and their evolution through time are highly complex, requiring a flexible modeling approach to recover missing entries. We also show that imputing the cross-sectional mean as a simple benchmark (Green, Hand, and Zhang, 2017) yields poor results. In fact, across the twelve themes, our model explains between 55% and 98% of the variation of reconstructed firm characteristics not already explained by the mean-imputation.

The proposed model architecture puts us in the unique position of being able to quantify the uncertainty attached with each prediction. This is in contrast to contemporaneous research (e.g. Bryzgalova, Lerner, Lettau, and Pelger, 2022), which rely on point estimates, as opposed to a probability distribution across percentiles. To quantify this uncertainty, we use the insight that an uninformed guess produces equal probabilities for all percentiles and use the Kolmogorov-Smirnov test statistic to show that the estimated probability distribution significantly differs from a uniform assignment for 93.8% of missing firm characteristics.

In an application to empirical asset pricing, we use the completed dataset and investigate how filling the gaps influences average returns of characteristic-sorted long-short (L-S) portfolios. Incorporating the additional information pushes the returns of most L-S portfolios towards zero. Across all 143 characteristics, the average change in the absolute return spread amounts to a significant -1.26%, with a much larger impact on many characteristics. Furthermore, we show that the returns of both the long and short leg decrease on average, but that the reduction in either leg of the long-short portfolio are uncorrelated across characteristics. We rule out that these results arise mechanically by sorting stocks based on past information or wrongfully allocating stocks to the long or short portfolio after imputation. At the same time, we can confirm that most factors survive the scrutiny of this approach, adding to recent evidence by Jensen et al. (2021) that most findings in financial research are indeed reproducible and carry over to unseen data.

Our model's architecture produces interpretable outputs and importantly is not a black box.[3] The rigorous use of the attention mechanism allows us to track the internal flow of information: which input is required to reconstruct a target firm characteristic? Using groups of accounting-, market-based and hybrid characteristics as the highest level of abstraction, we show that the model primarily draws on information about characteristics of the same group, but also greatly benefits from the inclusion of all other characteristics. Clustering firm characteristics by their informational content confirms these results. Together, this provides an intuitive justification for jointly modeling the evolution of the

[3] Explainable AI has recently garnered a lot of attention. See for example Lundberg and Lee (2017) for a great attempt at interpretation. Attention is a way to keep the model interpretable internally, see for example Lim, Arık, Loeff, and Pfister (2021) and Arik and Pfister (2019).

143 firm characteristics when recovering missing information.

Recent evidence from the literature on empirical asset pricing has showcased that past firm characteristics contain valuable information about future stock returns (Baba Yara, Boons, and Tamoni, 2020; Keloharju, Linnainmaa, and Nyberg, 2021). Investigating how important past information is when recovering missing firm characteristics, we can show that it is mostly the evolution of firm characteristics *within the last year* that is used in the model's predictions. This is in line with efficient financial markets and adequate accounting and reporting standards. Further elaborating on the importance of different parts of the information set, we find that restricting our proposed model architecture to the inclusion of only temporal or cross-characteristic information deteriorates how accurate firm characteristics can be reconstructed. Still, these restricted machine learning models manage to outperform the bespoke approaches outlined earlier. Together, this highlights the importance of jointly incorporating information about a target characteristic's past, other characteristics of the firm in question, as well as information about the cross-section of other firms.

We provide the completed dataset, including the recovered percentiles and estimates for the raw firm characteristics, as well as the estimated probability distribution across percentiles for future research.[4]

1.2. Related Literature

Our paper contributes to the literature on dealing with missing information in financial and accounting data. The issue is pervasive: Abrevaya and Donald (2017) hand-collected data from four leading economic journals over a three-year window and claim that about 40 % of all published articles had to deal with missing data and roughly 70 % of those

[4] The imputed firm characteristics can be downloaded from the first author's website.

simply dropped missing observations. This ad-hoc approach, also preferred by influential studies such as Fama and French (1993) and Kelly, Pruitt, and Su (2019), not only vastly reduces the sample size, but potentially results in biased inference, if results obtained from the sample with no missing data do not generalize to the remainder. Smaller firms provide less complete information – a direct violation of this "missing-at-random" assumption.

Another prominent way of dealing with missing data is to impute the cross-sectional mean, which dates back to Wilks (1932). The studies by Green et al. (2017), Kozak, Nagel, and Santosh (2020a), Gu, Kelly, and Xiu (2020), Chen and Zimmermann (2020), and Gu, Kelly, and Xiu (2021) are recent examples using this approach on the merged CRSP-Compustat database. Bali, Beckmeyer, Moerke, and Weigert (2021b) and Bali, Goyal, Huang, Jiang, and Wen (2021a) also use this approach on joint stock-option and stock-bond datasets, respectively. Afifi and Elashoff (1966) argue that imputing the mean yields unbiased estimates if and only if the data follows a multivariate normal distribution and the data is missing at random. Financial and accounting data likely violates both assumptions, requiring the use of novel methods more apt at dealing with the issue of missing firm characteristics.

A contemporaneous attempt at leveraging the informational content of missing firm characteristics for the use in asset pricing studies is provided by Freyberger, Höppner, Neuhierl, and Weber (2021). The authors propose an adjusted generalized method of moments framework to find suitable estimates for missing characteristics. They base their estimation on a pre-selected set of 18 characteristics, which are required to be observable at all times, and impute missing characteristics by assuming that they hold explanatory power for stock returns. As a fully cross-sectional approach, their method disregards information about past observations of firm characteristics – as we will later see, an important aspect for recovering missing firm characteristics. The objective we pursue in this study is different from that of Freyberger et al. (2021). Instead of imputing missing characteristics that best help explain stock returns, we completely abstract from the asset pricing perspective and impute the most probable unconditional values for missing entries. In an application using the completed dataset, we can show that missing characteristics indeed have an impact on characteristic-sorted long-short portfolio returns.

The two studies most closely related are by Cahan, Bai, and Ng (2022) and Bryzgalova et al. (2022). Both studies rely on extracting a latent factor structure on panel data to impute missing entries. Bryzgalova et al. (2022) have a similar objective as our study, in that the authors, after examining cross-sectional and temporal missingness patterns in a panel of 45 firm characteristics, propose an imputation method using latent factors estimated from observable characteristics. The authors thereby assume that a low number of latent factors adequately explains the evolution of observed *and missing* firm characteristics.

We deviate from Bryzgalova et al. (2022) in several important aspects: first, our model requires no dimensionality reduction, i.e., we do not impose that each characteristic is driven by the same set of latent factors. Instead, we allow the model to flexibly decide for each firm×month observation, which observed characteristics are most informative to impute missing values. This also facilitates the interpretation of the predictions, as we are able to trace the model's internal information flow. Second, we provide a full out-of-sample test spanning 37 years, showing that the estimated structure between observable characteristics and missing entries is stable over time. Third, we carefully account for imbalances in the dataset, which arise because some characteristics are missing more often than others. Not accounting for these imbalances will implicitly overweight information from characteristics with a higher availability. Estimating latent factors on observed characteristics will push these factors towards being more informative about characteristics that are already most often available. Keep in mind that the objective of our study is to impute *missing* characteristics. During the training phase, we therefore make sure that the model is asked to reconstruct each characteristic the same number of times. Fourth, we introduce nonlinearities and interaction effects between characteristics, of which current asset pricing research stresses the importance in explaining stock returns (Gu et al., 2020; Chen and Zimmermann, 2020; Kozak, Nagel, and Santosh, 2020b). Fifth, we obtain a probability distribution across percentiles of a characteristic, which puts us in the unique position to directly quantify the uncertainty associated with each imputation. Finally and most importantly, the differences in the modeling approach outlined above lead to a better imputation performance on the 143 characteristics we consider, which more than triples the 45 studied by Bryzgalova et al. (2022).

2. Machine Learning for Missing Characteristics

Our model architecture builds on recent advances from the computer science literature, and applies state-of-the-art ideas from natural language, sequence, and image processing to the question of how to deal with missing economic data. Specifically, we follow the insights of BERT, proposed by Devlin, Chang, Lee, and Toutanova (2018), which has grown to be one of the most renowned language models and is now an integral part of Google's search engine. BERT learns how words relate to one another in a self-supervised fashion. By randomly masking words of an input sentence, BERT is required to come up with a probabilistic assessment of how to reconstruct the masked words given the remaining sentence as a context. In an analogous fashion, we extend BERT to predict missing firm characteristics by leveraging the information content of observed characteristics, past observations of characteristics, and information from the cross-section of other firms. Characteristics are correlated across stocks, within a target stock, and over time. To see this, consider the case of Apple. Given that Apple operates in the technology sector, it's profitability will co-move with that of Microsoft. At the same time,

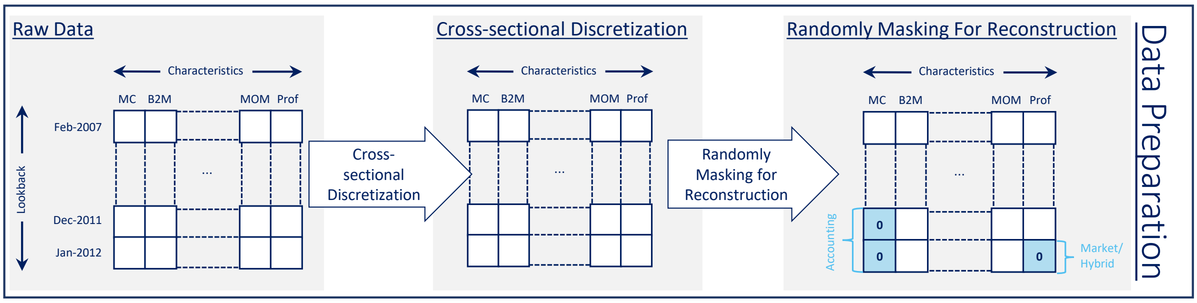

Fig. 1. Model Setup

The figure shows the first part of our model setup, the data preparation. The raw data for a target stock, for example Apple, consists of a $T \times F$ matrix, with $T = 60$ months of historical information about $F = 143$ characteristics (left panel). We first discretize the characteristics cross-sectionally into percentiles (middle panel) and then randomly mask 20% of Apple's characteristics at time t for reconstruction, denoted by the special class "0" in the right panel.

Apple's valuation ratios, such as book-to-market and price-to-earnings, will also co-move together. This correlation structure is dynamic and potentially highly complex and non-linear. Our model is able to make use of this vast array of information in an interpretable fashion, providing academics and practitioners alike unique insights into commonalities in the space of firm characteristics and a simple way to deal with missing data.

2.1. Data Preparation

Figure 1 walks the reader through the first step of our model, which deals with preparing the input dataset to a usable form.

Raw Data. The example in the Figure uses information about Apple to reconstruct masked target characteristics in January of 2012, using a total lookback window of $T = 60$ months to draw information from. The input matrix of size $T \times F$ holds information about the $F = 143$ characteristics of Apple between February 2007 and January 2012. When estimating the model we apply the steps detailed herein for all stocks in the sample.

Cross-sectional Discretization. To simplify the task of reconstructing missing firm characteristics, while at the same time retaining a high-degree of expressiveness and flexibility, we employ a simple transformation to the input characteristics: instead of considering rank-standardized firm characteristics as a real number between $[-1, +1]$ (Gu et al., 2020), we cross-sectionally discretize (middle panel) each characteristic into percentiles, yielding a total of 100 classes per characteristic.[5] First and foremost increasing the robustness of the model fit and helping with possible overfitting by dealing with outliers and reducing noise in the input characteristics, this approach has the added benefit of providing a natural way for dealing with missing data: we add an additional class to each characteristic, which captures the information of a *missing* input. We thereby allow the circumstance that a certain characteristic is missing to provide potentially valuable information to the model about why it is missing. Expressing characteristics as a real number in contrast provides no direct way to denote a missing entry. Should we assign the cross-sectional mean of "0" to missing entries? This will bias the model towards imputing the cross-sectional mean – an undesirable property as the tails of the distribution are more often than not the main point of interest. Another advantage of this discretization step is that the model produces a probability distribution across a characteristic's percentiles. This allows us to quantify the modeling uncertainty associated with each predictions.

Masking for Reconstruction. In the next step, features masked for reconstruction (right panel) are assigned class "0". Predicting their percentiles using the information available about other firm characteristics will be the objective of our model. We randomly mask 20% of the *available* firm characteristics at time t, shown in Figure 1 as the blue squares. This approach is known in natural language processing as *masked language modeling* and follows Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, and

[5] This approach is commonly used in gradient-boosted trees, such as XGBoost, which have recently enjoyed considerable interest by financial economists. In a novel paper.

Stoyanov (2019). We apply it to the case of recovering *ex-ante missing* entries. During this self-supervised stage, we have full knowledge of the true percentiles of masked characteristics. In a subsequent out-of-sample evaluation phase, we use the estimated latent structure to impute characteristics that were missing to begin with.

Note at this point that we repeat the masking step for a high number of iterations, 400 in our case. In each iteration, we mask a different set of characteristics per firm$\times$month observation. Therefore, each combination of masked characteristics is equally likely. The model extracts a suitable context from the available information, where the availability is changing from iteration to iteration. Thereby, the model learns to suitably impute missing entries of firm characteristics by flexibly accommodating to the available information, regardless of the dependence structure underlying missing information. Importantly, whereas we randomly mask 20% of the available characteristics for reconstruction, this does not require that entries of firm characteristics are missing at random. In contrast, the iterative learning algorithm makes the model robust to fluctuations in patterns of missing information, which potentially vary over time and are stock-specific.

The model is flexible enough to understand the quarterly release cycle of accounting variables. We therefore mask not only the values of these variables in month-t, but also the two preceding months ($t-1$ and $t-2$). This step assures that the model has no forward-looking information. During the estimation phase, we also assure that the model reconstructs each characteristic an equal number of times by scaling the 20% with the overall percentage of missing entries per characteristic in the training sample. Table B2 highlights that the degree of missingness varies substantially across characteristics. Without this modeling choice, the model would invariably learn to reconstruct well those characteristics that are most often available, simply because it sees them as examples more often. As we later want to use the learned latent representation of how firm characteristics relate to one another to recover ex-ante missing entries, we assure that the

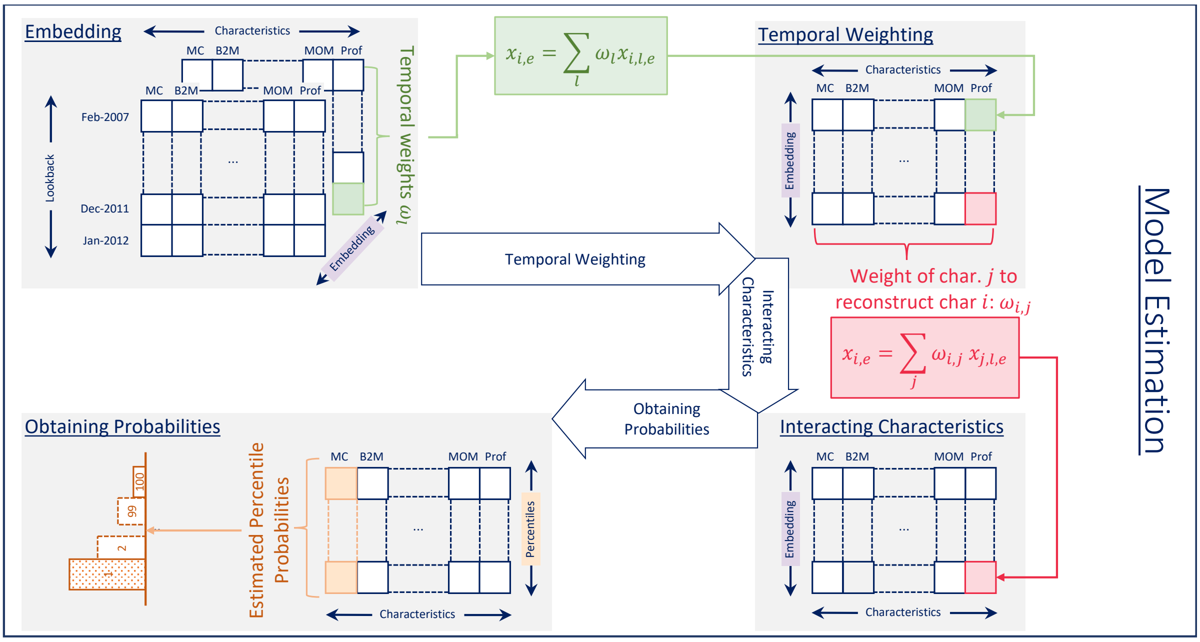

Fig. 2. Model Setup

The figure shows the second part of our model setup, dealing with the model estimation. The prepared data from the steps outlined in Figure 1 is fed through an embedding stage, which represents the 100 percentiles of each characteristic as a $D = 64$-dimensional vector, which is learned across stocks. We next weight the information from past time steps through temporal attention (top right) and then interact input characteristics (bottom right), allowing the model to use information from all available characteristics in the reconstruction of every target characteristic. In a final step, we obtain a probability distribution across the 100 percentiles for each masked characteristic and use the percentile with the highest predicted probability as the model's estimate. We compare this with the true – observed – percentile and update the model weights through stochastic gradient descent.

model learns to do so for all characteristics, regardless of how often they are missing in the training step.

2.2. Model Estimation

Figure 2 provides the second part of our model setup, dealing with the actual estimation and choice of architecture. We extend the methodology proposed in Devlin et al. (2018) to deal a) with financial data, and b) missing observations.

Embedding. First, we feed the discretized and masked characteristics through an embedding. The embedding enables the model to leverage information about other stocks in the universe, by pushing dissimilar stocks across the distribution of a target characteristic away from each other, while keeping the ones that are similar close to one another in vector space. It does so by expressing the percentiles of each firm characteristic as $D = 64$-dimensional numerical vectors, which are shared across all stocks at time t. This "lookup table" relates a percentile of a characteristic to a numerical representation, which is learned from the input data. While embeddings are useful in many ways, they are especially so in the context of recovering missing entries of firm characteristics to accommodate differences in how the distributional properties of observed characteristics relate to missing entries. For example, Fama and French (1993) highlight that stocks with high and low market capitalization have vastly different risk profiles. Analogously, large and small stocks potentially differ in how to process information needed to recover missing characteristics. The embedding accommodates these differences. The use of embeddings is standard in machine learning to deal with complex datasets (Huang, Khetan, Cvitkovic, and Karnin, 2020; Somepalli, Goldblum, Schwarzschild, Bruss, and Goldstein, 2021; Lim et al., 2021; Gorishniy, Rubachev, Khrulkov, and Babenko, 2021).

Temporal Weighting. How does the historical evolution of firm characteristics help us to reconstruct today's values? To find an optimal weighting of information from past time steps, we heavily rely on the so-called attention mechanism, a machine learning technique that allows the model to dynamically focus on the most important parts of the input data, while fading out the rest. We also apply attention in the next step when interacting information from the set of 143 input firm characteristics. The rigorous use of attention in machine learning as a standalone technique was proposed by Vaswani et al. (2017) and gave rise to the "Transformer" model type, which are by now the backbone of many state-of-the-art models in natural language and sequence processing.

Attention computes how similar a tensor of search queries $\mathbf{Q}$ is to a tensor of keys $\mathbf{K}$. Both $\mathbf{Q}$ and $\mathbf{K}$ are linear transformations of the same $\mathbf{x}$, which is the intermediate output from the previous model step, $\mathbf{W^{Q/K}x}$, where matrices $\mathbf{W^{Q}}$ and $\mathbf{W^{K}}$ are learned in the estimation process. The reliance on the same inputs for both queries and keys gives rise to the name "self-attention". Using the resulting attention (comparison) matrix $A(\mathbf{Q},\mathbf{K})$ as weights, we compute an optimally-weighted combination of the values in tensor $\mathbf{V}$, which again is a linear transform of input $\mathbf{x}$, $\mathbf{V} = \mathbf{W^{V}x}$. Each entry of $\mathbf{V}$ is associated with a certain entry of keys in $\mathbf{K}$, analogous to how SQL lookups work. Different from SQL lookups, however, which require that each query has a matching key in the database, attention is a probabilistic lookup, such that the algorithm retrieves the *most probable* keys for the query. In economics and finance this allows to answer questions such as "how important is information about Apple's book-to-market ratio from one year ago to reconstruct today's entries?" or "how important is information about Apple's price-to-earnings ratio to recover its book-to-market ratio?"

Mathematically, express attention as,

$$\mathrm{A}(\mathbf{Q},\mathbf{K}) = \mathit{Norm}\left(\frac{\mathbf{QK}^{\mathrm{T}}}{\sqrt{N^{\mathrm{A}}}}\right), \tag{1}$$

where N^{A} denotes the number of units to attend to, so either $T = 60$ in the temporal weighting step, or $F = 143$ when interacting characteristics. The resulting attention matrix per firm-month observation is thus of size $(N^{\mathrm{A}} \times N^{\mathrm{A}})$. *Norm* is a normalization function, which scales the attention matrix to row-wise sum to 1, with values between 0 and 1, thereby mapping from $\mathbb{R}^d$ to probability space Δ^d.[6] We consider normalization

[6] Such that $\Delta^d := \{\mathbf{p} \in \mathbb{R}^d : \mathbf{p} \geq 0, \|\mathbf{p}\|_1 = 1\}$.

functions of the α-Entmax family (Peters, Niculae, and Martins, 2019):

$$\alpha\text{-entmax}(\mathbf{x}) = \underset{\mathbf{p}\in\Delta^d}{\operatorname{argmax}}\, \mathbf{p}^\top \mathbf{z} + \mathrm{H}^T_\alpha(\mathbf{p}), \qquad \text{with} \tag{2}$$

$$\mathrm{H}^T_\alpha(\mathbf{p}) = \begin{cases} \frac{1}{\alpha(\alpha-1)} \sum_j \left(p_j - p_j^\alpha\right), & \alpha \neq 1. \\ -\sum_j p_j \log p_j, & \alpha = 1. \end{cases} \tag{3}$$

We consider three different normalization functions, with varying degrees of imposed sparsity in the attention matrices. $\alpha = 1$ yields the common Softmax function, with no sparsity imposed (i.e. $p_j > 0 \ \forall\ j$). Martins and Astudillo (2016) introduce Sparsemax ($\alpha = 2$), which aggressively pushes small weights towards zero. This is familiar to how Lasso-regressions push the smallest coefficient to zero. To model moderate sparsity in the attention matrices, we also consider $\alpha = 1.5$, which we refer to as Entmax. We have no prior on the degree of sparsity in the latent structure governing the evolution of firm characteristics. We therefore let the data decide on the optimal degree of sparsity in both the temporal and feature attention matrices, by tuning hyperparameter α.[7]

To increase the learning capacity, multiple attention heads – each with its own attention matrix and therefore flexibility to focus on different parts of the input data – are commonly employed. We opt for a total of $N^{\text{heads}} = 8$ temporal and feature attention heads per processing unit. Changing this has only a marginal impact on the outcome. We follow Lim et al. (2021) and use *interpretable multi-head attention (IMHA)* throughout the paper. It averages the attention matrices of each attention head before multiplying

[7] Results for this exercise are shown in Table IA2.1.

it with a single learned value matrix $\mathbf{V}$:

$$\text{IMHA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{H}\mathbf{W}_{\mathbf{H}}, \qquad \text{where} \tag{4}$$

$$\mathbf{H} = \left\{ \frac{1}{N^{\text{heads}}} \sum_{h=1}^{N^{\text{heads}}} \text{A}\left(\mathbf{Q}\mathbf{W}_{\mathbf{Q}}^{h}, \mathbf{K}\mathbf{W}_{\mathbf{K}}^{h}\right) \right\} \mathbf{V}\mathbf{W}_{\mathbf{V}}. \tag{5}$$

Here, matrices $\mathbf{W}_l \in \mathbb{R}^{D\times\left(D/N^{\text{heads}}\right)}$ with $l \in [\mathbf{Q},\mathbf{K}]$ are head-specific weights for keys and queries, and $\mathbf{W}_{\mathbf{V}} \in \mathbb{R}^{D\times D}$ are the weights for values $\mathbf{V}$, which are shared across the heads. The weight-sharing for $\mathbf{V}$ allows us to directly interpret the attention weights in terms of how important each characteristic and historic time step is in reconstructing today's characteristics.

Now that we have introduced the attention mechanism, we continue walking through the model setup. We stress the equivalence between the attention mechanism and a weighted sum of a possibly nonlinear function of inputs $\mathbf{x}$. Let $x_{i,l,e}$ denote the internal representation of Apple's ith characteristic before temporal aggregation, measured at time $t-l$. Subscript e denotes the eth value in the vector along the embedding dimension of size $D = 64$, which we have described above. We leverage past information about Apple's characteristics by computing a weighted sum of all $x_{i,\cdot,e}$ over lookback months $l \in [0,..,59]$, where weights ω_l are learned from the data itself, yielding intermediate model output $y_{i,e}$:[8]

$$y_{i,e} = \sum_{l=0}^{T=59} \omega_l\left(\mathbf{x}\right) \cdot f\left(x_{i,l,e}\right), \tag{6}$$

where function $f(\cdot)$ follows Eq. (1). After the temporal weighting, we apply additional nonlinear processing steps, which capture higher-order dependencies. These steps are further described in Appendix IA1.

[8] Where the lookback dimension is deflated.

Interacting Characteristics. Other firm characteristics hold valuable information about missing entries, which we incorporate into the model by employing the attention mechanism across the feature dimension. The model consequently learns to extract the right share of information from each characteristic $j \in F$ to reconstruct the percentile of target characteristic i.

Let $x_{i,e}$ denote Apple's ith characteristic after the temporal aggregation. The feature attention now computes weights $\nu_{i,j} \forall (i,j)$, expressing how much of characteristic j's information is required to reconstruct an entry of characteristic i. As for the temporal weights ω_l from the previous step, weights $\nu_{i,j}$ depend on the intermediate model input $\mathbf{x}$ and are therefore both stock- and time-specific:

$$y_{i,e} = \sum_{j=1}^{F=143} \nu_{i,j} \left(\mathbf{x}\right) \cdot f\left(x_{j,e}\right) \tag{7}$$

After these three steps the model has combined information about other stocks (embeddings), the past evolution of Apple's own characteristics (temporal weighting), as well as other firm characteristics of Apple (interacting characteristics).

Obtaining Probabilities. The final modeling steps takes the interacted and processed characteristics and creates a probability distribution across the 100 percentiles for each of the *masked* input characteristics. To do so, we first extend the 64-dimensional embedding dimension to span the 100 percentiles:

$$\mathbf{y} = \mathbf{W}^O \mathbf{x}, \qquad \text{with} \quad \mathbf{W}^O \in \mathbb{R}^{(100 \times D)} \tag{8}$$

We then optionally apply additional nonlinear processing along that dimension, wherein the optionality is governed by skip-gates described in Appendix IA1.1. Finally, we apply a Softmax normalization (Eq. (2)), which transforms the model's output to a probability distribution over the percentiles.

3. Data, Training & Performance Metrics

Firm Characteristics We analyze the dataset studied in Jensen et al. (2021), which contains monthly firm characteristics computed from CRSP and Compustat, for all stocks trading on NYSE, NASDAQ, and AMEX. We focus on the 153 characteristics identified by Jensen et al. (2021), from which we drop seasonal returns (Heston and Sadka, 2010), which are more often missing than they are available. Similar to Gu et al. (2021), we require only a minimum set of filters in order to work on the largest possible dataset. For a firm-month observation to be included, we require that it refers to common equity and the firm's primary security.

Our model extracts information about the likely value of a missing characteristic from observed characteristics and their evolution through time. We therefore require that each firm×month observation has valid information about at least 20% of the input characteristics. We specifically do not dictate which characteristics have to be available, or which are informative about missing entries of other characteristics, but rather let the data speak for itself. This filtering step discards 0.2% of observations in the joint training and validation sample, and 7.2% in the testing sample.[9] We follow the standard procedure in the literature and lag quarterly accounting data by three months and annual accounting data by half a year. In total, our data covers the period from July-1962 through December-2020, for a total of 57 years, providing information about 143 characteristics

[9] We have also estimated a model without this filter in a previous version of the paper. All results shown in this version carry over to the unfiltered sample. However, requiring a minimum amount of information for a given firm seems plausible in our opinion, if we want to leverage the information that is available from these characteristics to recover entries that are not.

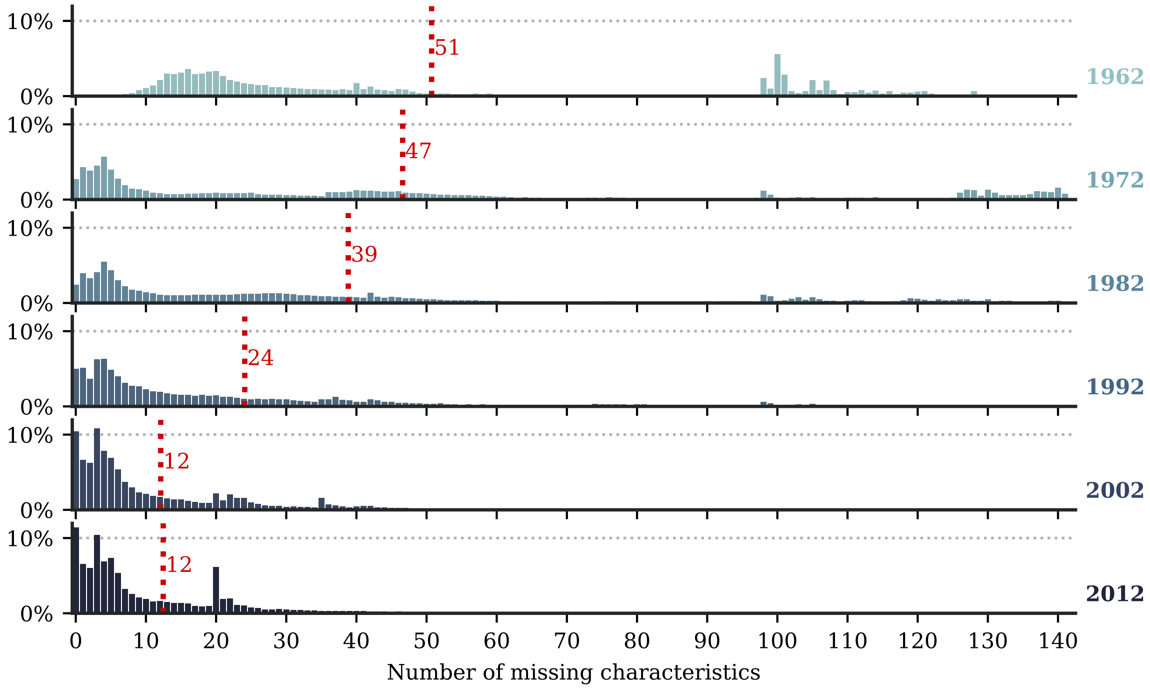

Fig. 3. Distribution of Missing Firm Characteristics over Time.

The figure shows the distribution of the number of missing firm characteristics per observation for each decade in our dataset. We use 143 firm level characteristics from the dataset provided by Jensen et al. (2021) with common filters applied, see Section 3. The dashed red lines indicate the mean number of missing characteristics per firm-month observation.

on 25,118 unique firms, for a total of more than 3.2 million firm-month observations. Of the 143 characteristics, 47 are based on market information alone, 75 on accounting information, and 21 are a hybrid that use information from both sources.

Figure 3 and Figure 4 discuss two important dimensions of missingness patterns in our dataset: First, Figure 3 shows the evolution of missing values for the 143 firm-level characteristics over time. During the 60s, an average of 51 characteristics – more than 35% – are missing for a firm×month observation. Despite this number declining considerably in the following decades, the average firm still misses 12 characteristics today. Dropping missing entries of firm characteristics potentially limits the historic sample that a researcher can use. Most tests in empirical asset pricing, however, crucially depend on

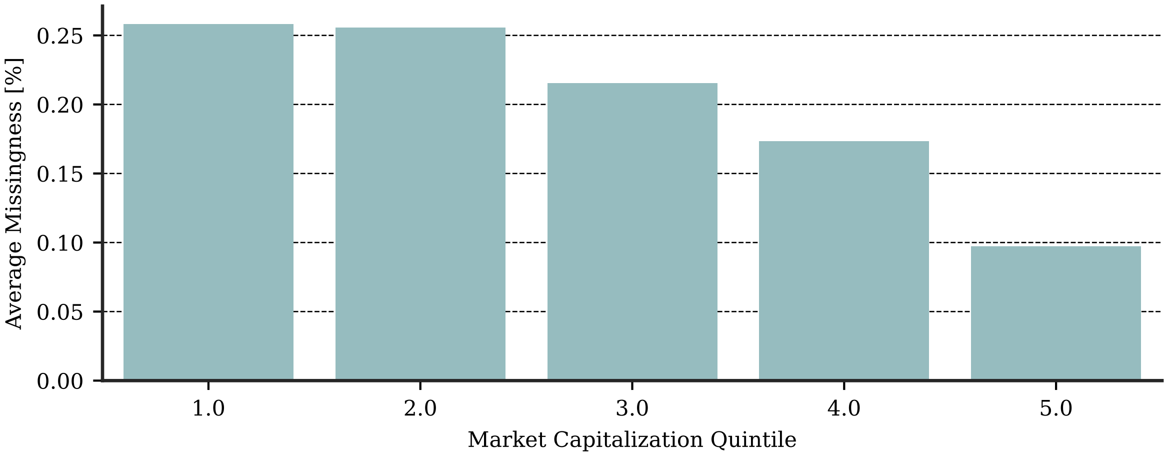

Fig. 4. Percentage of Missing Characteristics by Market Capitalization Quintile.

The figure shows the average proportion of missing entries that belong to firm×month observations within a certain market capitalization quintile. For this, we sort stocks by their market capitalization each month and report the proportion of missing firm characteristics within each quintile. The numbers are scaled to sum to 1.

long time-series. As a second dimension of missingness, Figure 4 shows the average cross-sectional proportion of missing entries across the 143 firm characteristics considered, shown separately for market capitalization quintiles formed in each month t. The 40% smallest firms on average make up more than half of all missing characteristics. This number steadily drops to below 10% for the 20% largest firms. Dropping firm×month observations with missing characteristics therefore systematically excludes information about smaller firms. At the same time, however, missing information is an issue that is not exclusive to these smaller issues but rather pervasive across the entire CRSP universe. Table B2 provides a list of the characteristics used, how often each is missing, and how well our model is able to reconstruct each characteristic.

Training Routine As we are discretizing the input characteristics into percentiles, we can formulate the problem of recovering firm characteristics as a multi-class classification. The standard approach to solving these is by minimizing the *cross-entropy loss*, which is the negative log-probability of the target class. To force the model to also get the predictions right for characteristics that are harder to recover, we use the focal loss proposed by Lin, Goyal, Girshick, He, and Dollár (2017):

$$\mathrm{FL}(\mathbf{p}) = \frac{1}{|\mathbf{p}|} \sum_{c} -(1 - p_c)^{\gamma} \cdot \log(p_c), \tag{9}$$

which reduces the influence of examples that the model classifies well already. Here, p_c denotes the predicted probability of the target percentile for masked characteristic c. We optimize over the mean loss for all masked (and thus reconstructed) characteristics. For this, we set $\gamma = 25$, which saturates estimated probabilities p_c at around 20%. This does two things: first, it forces the model to come up with good predictions for all characteristics, instead of focusing on a few that are easiest to reconstruct. And second, $\gamma = 25$ is lenient enough to accommodate different scales at which the input characteristics operate. While most are represented as real numbers, which we then discretize to percentiles, some characteristics, notably the `f_score` by Piotroski (2000), are discrete to begin with, ranging between 0 and 9. If γ is set to aggressively, the model is forced to produce seemingly "random" predictions for `f_score`. Instead, our choice of $\gamma = 25$ strikes a balance between the two.

Figure 3 illustrates that the sample has grown considerably more complete in recent years, with an average of 12/143 characteristics missing in the last decade, compared to more than a third in the 60s. The more characteristics that are available to us, the more information we are able to extract. As a consequence, we flip the common train/validate/testing split and train the model using the most recent 15 years of data (2006-2020) and validate the choice of hyperparameters in a five-year validation sample (2001-2005). The remaining 37 years (1962-2000) are used for out-of-sample tests regarding the robustness of the model fit in never-before-seen data.

Optimization Neural networks are typically trained using stochastic gradient descent, which uses a subset of the data in each iteration to evaluate the gradient and update the model weights. The key parameter governing the success of this training procedure is the learning rate, which controls the size of each step taken in the opposite direction of the gradient. We use the *adaptive moment estimation algorithm (Adam)*, introduced by Kingma and Ba (2014), that individually changes the learning rate for each model parameter by estimating the first two moments of the gradient. To help Adam converge to good solutions, we furthermore adopt the "OneCycle" learning rule by Smith and Topin (2017), which starts with a very low learning rate (lr = 0.00001), which is then increased for the first 30% of training epochs, up to a high number (lr = 0.005). This ramp-up phase with a low learning rate helps Adam find good estimates of the moments of the gradient, which aids the algorithm in making informed decisions for the epochs with the higher learning rates. After the increase of the learning rate, we gradually decrease it until the total number of training epochs is reached to refine the fit. We set the maximum number of training epochs to 400.[10] With a batch size $B = 2400$ and a total of approximately 780,000 observations in the 15 years-long training sample, we update the model parameters with stochastic gradient descent more than 130,000 times. Training each hyperparameter-combination takes about a day on eight Nvidia Tesla A100 80GB GPUs. A list of the hyperparameters, their search ranges and optimal values is given in Table IA2.1 in the appendix.

Regularization To assure that the latent structure found by the model carries over to unseen data, we employ a number of regularization techniques: we include proper weight decay for Adam (Loshchilov and Hutter, 2017), which adds a fraction of the L2 norm of the model parameters to the loss function, forcing the model to choose small and conser-

[10] Increasing the number of maximum epochs does not change the results.

vative parameters. amsgrad (Tran et al., 2019) adds theoretical convergence guarantees to the ad-hoc effectiveness of Adam. During training, we furthermore randomly drop the activation of connections in the model. This *dropout* helps the model find general solutions (Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov, 2014). Lastly, we use layer normalization after each skip-connection. This assures that each processing unit operates on roughly the same data range (Ba, Kiros, and Hinton, 2016). Layer normalization tends to work better than batch normalization for sequence and time-aware modeling tasks.

Performance Measures The way the model is fitted – that is by randomly masking a fixed percentage of non-missing characteristics which we try to reconstruct – provides us with a controlled environment to assess the model's performance by quantifying its accuracy in reconstructing observed percentiles.

The primary metric we propose to evaluate the model's ability to reconstruct firm characteristics follows the ROC curve (*reveiver operating characteristic*) – a staple in machine learning research for evaluating classification problems. For that, we obtain the sampling frequencies p of the model error $|\Delta|$ as the absolute difference between the observed class y and the model predicted class $\hat{y}$ for the set of masked characteristics,

$$p(|\Delta| = k) = p(|y - \hat{y}| = k) \qquad \text{where} \qquad |\Delta| \;\in\; [0, 1, \ldots, 99]. \tag{10}$$

The cumulative distribution function $p(|\Delta| \leq k)$, for $k \in [0..50]$, or *ROC curve*, then tells us about the model's true-positive rate for a given threshold of the allowed model error.

We also consider the average number of percentiles we deviate from the truth, which we call *expected percentile deviation (EPD)*:[11]

$$\mathrm{E}\left[|\Delta|\right] = \sum_{k=0}^{99} p(|\Delta| = k) \cdot k. \tag{11}$$

A perfect model produces EPD = 0. Fully random predictions yield an EPD of $33.\bar{3}$. EPD is a convenient way to summarize the information provided by the ROC curve in a single number.

4. Model Validation: A Simulation Study

As outlined in the previous sections, we have set up our model with explicit jobs assigned to each building block. It is designed to include both temporal *and* cross-characteristic information, while being fully agnostic about the underlying processes governing the evolution of firm characteristics. We deliberately rely on the capability of our model to extract this structural information on its own. To showcase how well the proposed model learns about different types of processes, we set up a controlled environment in a simulation study.

Considered Processes The benefit of our model setup is that we can simultaneously model characteristics with different underlying processes. In the simulation, we consider characteristics that may be driven by an autoregressive process, rely heavily on cross-characteristic information, or a blend of the two. The inclusion of a large number of characteristics, $F = 143$ in our case, facilitates cross-learning effects, wherein one type of characteristic can be used in the reconstruction of characteristics of another type.[12] To see

[11] EPD is directly linked to the area under the ROC curve, commonly used in machine learning (AUROC), where EPD = 1−AUROC.

[12] Empirical evidence of this is provided in Figure 14 and Figure 15.

how our model manages to *simultaneously* deal with characteristics driven by different types of processes, we simulate three types of characteristics with different properties. The first set of characteristics c follows an AR(12) process:

$$c_{i,t}^{\mathrm{AR}(12)} = \gamma_i \cdot \sum_{k=1}^{12} w_k \cdot c_{i,t-k}^{\mathrm{AR}(12)} + \varepsilon_{i,t} \tag{12}$$

where $\varepsilon \sim \mathcal{N}(0,1)$ and $\gamma \sim \mathcal{U}(0.9,1)$ denotes a high level of auto-correlation. We choose exponentially-decaying weights w_k,

$$w_k = C \cdot e^{-0.25 \cdot k} \quad k \in [1,12], \qquad \text{with } C \text{ s.t.} \qquad \sum_{k=1}^{12} w_k = 1, \tag{13}$$

The second set of characteristics is cross-sectionally dependent, following a multivariate normal distribution (Freyberger et al., 2021):

$$c_{i,t}^{\mathrm{XS}} \sim \mathcal{N}(\mathbf{0}, \mathbf{cov}), \tag{14}$$

where $\mathrm{cov}_{i,j} = 0.99^{|i-j|}$, for characteristics i and j.

Finally, the third set of characteristics combines the two cases from above:

$$\mathrm{ar}_{i,t} = \gamma_i \cdot \mathrm{ar}_{i,t-1} + \varepsilon'_{i,t} \tag{15}$$

$$c_{i,t}^{\mathrm{AR+XS}} = \omega^{\mathrm{AR}} \cdot \mathrm{ar}_{i,t} + \left(1 - \omega^{\mathrm{AR}}\right) \cdot \mathrm{xs}_{i,t}, \tag{16}$$

where $\mathrm{ar}_{i,t}$ governs the autoregressive component, $\varepsilon' \sim \mathcal{N}(0,1)$, and $\mathrm{xs} \sim \mathcal{N}(\mathbf{0}, \mathbf{cov})$, with the same covariance matrix as above. ω^{AR} denotes the relative weight of the AR-component, which we set to 0.25.

We simulate a sample of 100 firms with 50 characteristics of each category for 25 years of monthly data, of which we use 15 for training, and 5 for validation and testing,

Table 2: Simulation: Model Accuracy

The table shows the imputation accuracy measured by the expected percentile deviation defined in Eq. (11) for the simulation study outlined above. We differentiate our model's accuracy from that of a Last imputation method which is a good competitor for autoregressive characteristics. We also include the Mean imputation method which is frequently used in the finance literature (e.g. Green et al., 2017; Gu et al., 2020, 2021).

	Expected percentile deviation			
	All	AR(12)	XS	AR(12) + XS
Full model	5.58	6.71	3.08	6.94
Last	20.38	6.55	33.48	21.12
Mean imputation	25.01	25.01	25.00	25.01

each. We assess how well our model is able to reconstruct firm characteristics using two simple benchmarks for comparison. First, we consider imputing the cross-sectional mean, regardless of the underlying process. As a second benchmark, we impute the last available value for each characteristics, which will naturally work well for autoregressive characteristics.

Table 2 shows the results. Overall, we find that our model manages to uncover the latent structure governing *all* types of characteristics and that within a single model. Pooled across all characteristics, we find that it produces by far the lowest EPD, showcasing its flexibility. For autoregressive characteristics *AR(12)*, we find that it performs on par with imputing the last entry, with EPDs of 6.71 vs. 6.55, respectively. For fully cross-sectional characteristics *XS*, imputing the characteristic's previous entry leads to random guesses and an EPD of $\approx$ 33. Our model, however, manages to predict the true percentiles with high accuracy. The same applies to characteristics that are governed by a joint autoregressive and cross-sectional process *AR(12) + XS*. The inclusion of cross-sectional dependencies between characteristics renders the imputation of the last available value a poor choice, being off by more than 20 percentiles on average. Imputing the cross-sectional mean naturally fails to uncover the intricacies of the underlying processes and unconditionally produces an EPD of 25. In sum, our model is highly flexible in accommodating multiple processes governing firm characteristics *at the same time.* It frees researchers and investors alike from taking a stand ex-ante about which process governs a target characteristics and instead learns to approximate the process directly from the data.

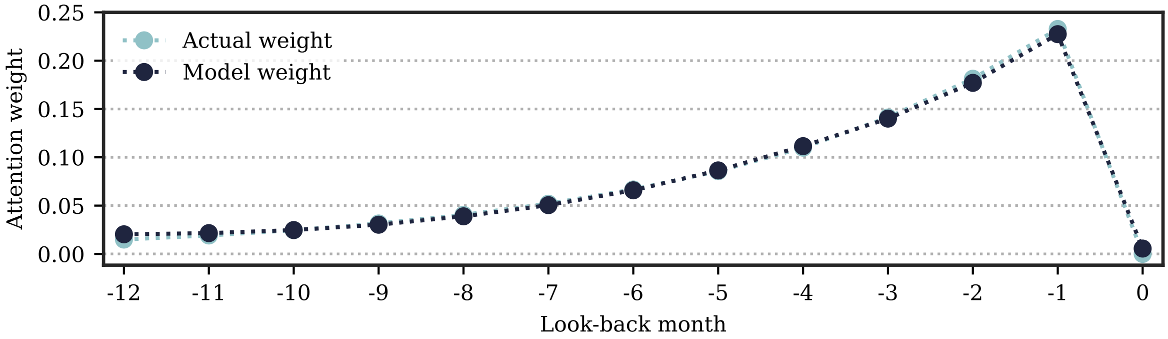

Fig. 5. Temporal Attention Weights – Simulated AR(12) Process

The figure shows temporal attention weights for a simulated AR(12) process in the specified look-back window of 12 month. Temporal attention weights measure how much information from each historical time-step is incorporated in the final prediction of the model. In that sense, it is directly comparable to the weights w_k specified in Eq. (13) for the simulated AR(12) process. We added both the actual (i.e. pre-specified) and the model predicted weight to the graph for comparison.

Temporal Patterns The temporal attention mechanism enables our model to use information from lagged values within the specified look-back window with no prior restrictions on where to draw information from. Table 2 has already shown that this enables the model to accurately reconstruct autoregressive firm characteristics. Figure 5 shows the learned temporal attention weights showcasing from which time lag the model uses information. The weights of the AR(12) process and the extracted temporal attention weights of the model perfectly line up. The model is capable of exactly identifying the temporal dependencies governing the evolution of these characteristics. Importantly, we find a weight of ≈ 0 placed on information from time $t = 0$. Despite being presented with contemporaneous information about other characteristics, our model has learned to disregard it and instead focuses solely on the temporal evolution.

5. Reconstructing Firm Characteristics

In this section, we apply our model to real-world data and ask it to reconstruct a set of randomly masked firm characteristics using the information embedded in other characteristics and their historical evolution. We first assess how well the model performs in an absolute sense: how far off are the predictions of the true percentile, across time, for different characteristics, and for different levels of available information for a firm? We then compare how well nested model cases, described in Section 3, fare against the full model. This investigation allows us to highlight the importance of incorporating a vast array of information from multiple dimensions when filling missing entries. Lastly, we zoom into the model and understand how it comes up with its predictions by exploiting the proposed architecture, which is fairly interpretable through the heavy use of the attention mechanism.

5.1. Model Performance

Figure 6 shows the resulting cumulative distribution function $p(|\Delta| \leq k)$, for $k \in [0..50]$, for the training, validation and testing samples (Eq. (10)). For a quarter of the cases in the training sample, our model manages to recover the masked characteristic's percentile exactly. For comparison, we have also included the performance of using the common mean- or median-imputation. Numerous studies use this approach to deal with missing firm characteristics (Green et al., 2017; Gu et al., 2020, 2021). The gray area in Figure 6 directly denotes our model's outperformance over this ad-hoc approach. Simply inferring the characteristic's mean is insufficient and disregards important variation in firm characteristics. In fact, for about 77% of cases, the mean imputation produces a deviation of more than a decile. Our model instead deviates by that much in fewer than one in ten cases. We also find that our model's performance is highly consistent and carries over

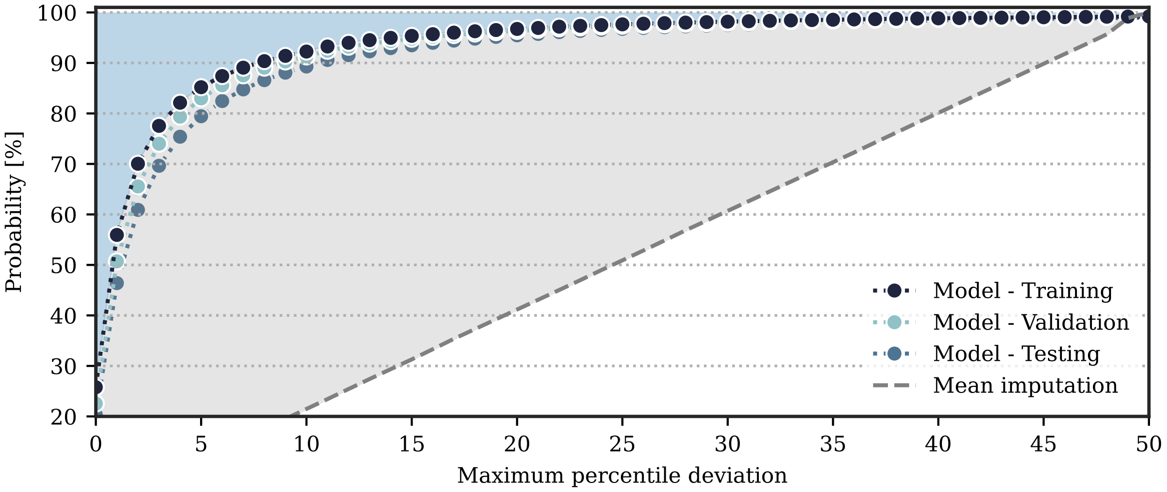

Fig. 6. Model Accuracy Curve.

The figure shows the cumulative distribution function of the model error $|\Delta|$ defined in Eq. (10) for the training, validation and testing sample. We also show the result for imputing the median for comparison. Section 5.2 provides a detailed model comparison. The gray-shaded area denotes the outperformance of our model compared to this ad-hoc method. The blue shaded area denotes the expected portfolio deviation (EPD) defined in Eq. (11).

well to the validation and the ouf-of-sample testing data.

The blue area above the curve is the EPD defined in Eq. (11). In the out-of-sample testing data, we achieve an EPD of 4.31. In other words, our model predictions are on average off by less than five percentiles, which significantly outperforms simply imputing the mean, with an EPD of 25. For the validation (training) sample, the EPD drops slightly, to 4.03 (3.63). [13] These numbers of course differ vastly across characteristics, each with differences in how hard they are to reconstruct, how often they are missing, and when they are missing. We now investigate the reconstruction performance across these dimensions.

Accuracy over Time. We first consider how well the model predictions stack up over time. Preferably, the prediction quality should be unaffected by temporal progression.

[13] We may also express the performance in a reconstruction R^2 which amounts 87% across the joint training, validation and testing sample, see Table A1 in the Appendix.

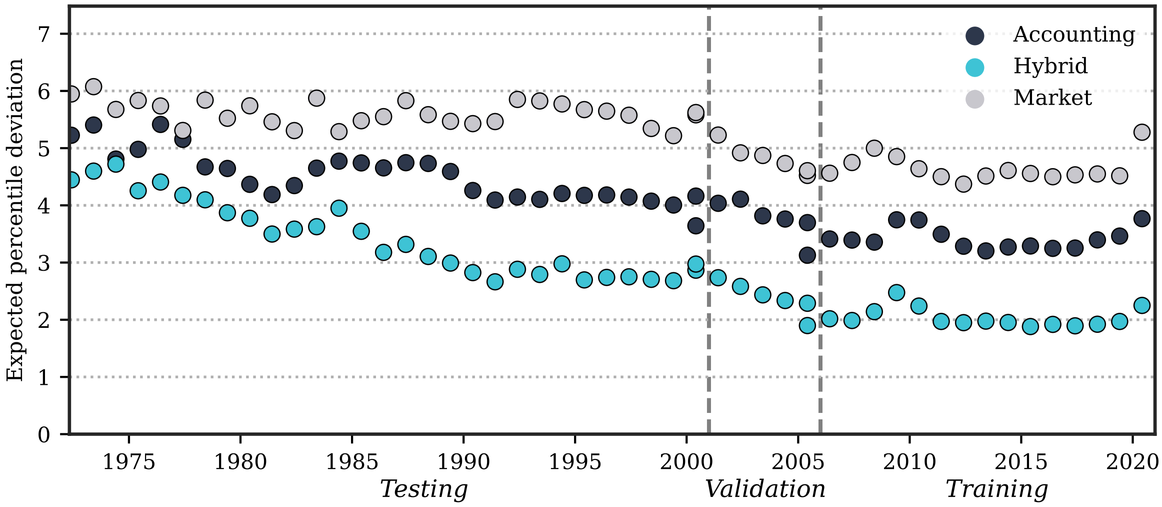

Fig. 7. Model Accuracy over Time.

The figure shows the model's accuracy as measured by the expected percentile deviation defined in Eq. (11) over time for accounting- and market-based, as well as hybrid characteristics.

At the same time, however, Figure 3 shows that the degree of missingness has decreased considerably over time. Likewise, we use the most recent 15 (+5) years to train (+validate) the model and its parameters. It is natural to assume some form of generalization gap to unseen testing data, which in our case is also the data with the highest degree of overall missingness.

While we do find evidence of better performance in recent years in Figure 7, the EPD is fairly stable over time and still low throughout the testing sample starting in 1973. For better interpretability, we have split the EPD numbers for market, accounting and hybrid characteristics, wherein hybrid characteristics use information from both sources (an example is the book-to-market ratio). We find that the average performance for all three groups of characteristics has improved slightly over time. For example, while the EPD for hybrid characteristics is around 4 in the early years of our sample, it trends downward to around 3 by the start of the validation sample and now stands at around 2 percentiles. The trends for the other groups are comparable. We find the best performance for hybrid characteristics, which have a comparatively high availability, and the

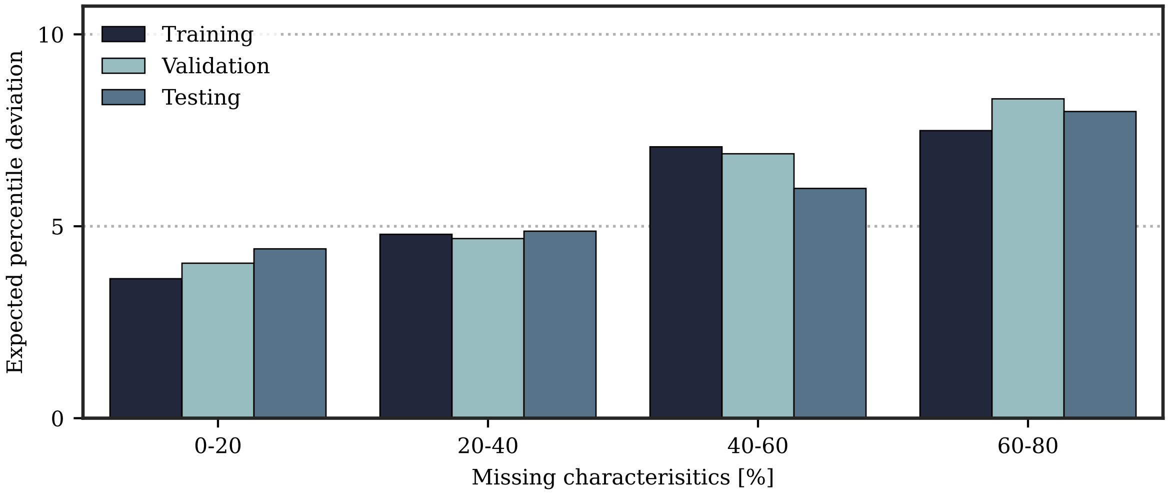

Fig. 8. Model Accuracy as a Function of Available Information.

The figure shows the model's accuracy measured by the expected percentile deviation defined in Eq. (11) as a function of the number of missing characteristics per firm×month observation. We group observations into quintiles depending on how many characteristics are missing and show results separately for the training, validation and testing sample.

worst for market characteristics, which generally vary the most. In fact, Figure IA5.1 in the Internet Appendix highlights the percentile migration for the three types of characteristics. Market-based characteristics fluctuate most from quarter to quarter. Still, the EPD is at or below 5 for all groups, suggesting that our predictions are on average off by less than five percentiles, even in the out-of-sample tests.[14]

This temporal stability shows that our model is able to pick up on, and ultimately exploit, a strong latent structure governing the evolution of firm characteristics. The slight increase in the EPD in the testing sample is likely not driven by shifts in this structure, but rather by the higher degree of missing information for that period.

Accuracy by Available Information. We therefore investigate how well the model is able to reconstruct characteristics when the degree of available information varies. To do

[14] Note that we employ a full out-of-sample test to assess how well our model's predictions hold up in data never used in its estimation. The sample accuracy would improve further had we decided to train the model on the full dataset.

so, we sort each firm×month observation by the number of available characteristics and compare the reconstruction performance across different *missingness* buckets. Figure 8 shows the results.

We find that the reconstruction error is increasing in the respective missingness buckets. More cross-characteristic information allows the model to better pick up on interactions with other firm characteristics and consequently achieve better predictions for the target characteristic. This effect, however, is fairly modest throughout. In fact, even for the firm×month observations with 60-80% missing characteristics, we find an EPD of 6-7 percentiles, still achieving better-than-decile accuracy.

Accuracy by Characteristics. We have seen that the average prediction of characteristics for firms with only few other characteristics is still precise. The lower the missingness, however, the better the predictions tend to be. To follow up on this, we now investigate the characteristics that we predict the best, and those that are hardest to predict. Figure 9 provides a breakdown of the ten characteristics with the lowest EPD and the ten characteristics with the highest EPD. A complete list is provided in Table B2. We furthermore indicate the group that each characteristic belongs to.

Among the characteristics *best* reconstructed is `age`, which deterministically increases by 1 each quarter, a behavior our flexible model architecture is able to identify. We can also reconstruct certain market-based characteristics very well, with a EPD of near zero. Notable examples are `market_equity` (Banz, 1981), momentum in the form of `ret_12_7` (Novy-Marx, 2012), and the dollar volume over the last half year `dolvol_126d`. We find a distinct cluster of characteristics among those *worst* reconstructed. Six out of the ten characteristics use daily information in their construction. They rapidly change from month to month – an evolution the model is not able to pick up on, because we never feed it information at a higher frequency than monthly. Other characteristics that the

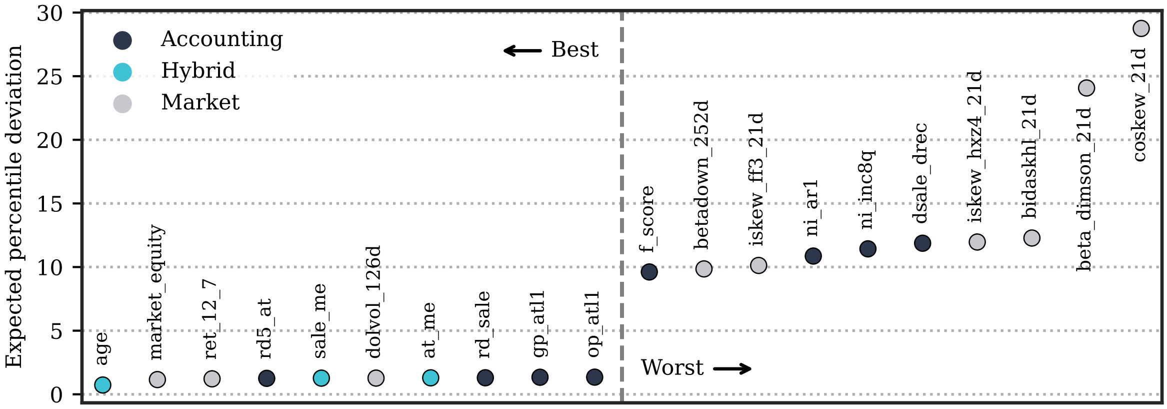

Fig. 9. Model Accuracy by Characteristic.

The figure shows the model's accuracy for the ten characteristics that the model reconstructs the best and the worst, measured by the expected portfolio deviation defined in Eq. (11). A complete overview can be found in Table B2 in the Appendix. We further categorize characteristics into three groups, accounting-, hybrid- and market-based.

model has a hard time reconstructing are measures of earnings persistence `ni_ar1` and the number of consecutive quarters with earnings increases `ni_inc8q`. Overall, the EPD for only two out of the 143 characteristics is above 20 percentiles, or in other words, less than accurate to the quintile.

Unbiased Predictions. We have so far focused on absolute deviations from the true percentile. While the predictions of our model are quite accurate, we want to understand if any systematic biases exist in the reconstruction process. For this, Figure 10 shows the median and 10 and 90% quantiles for the signed estimation error Δ across different percentiles for the true class. The figure provides three interesting insights into the prediction process. First, the median signed estimation error across all percentiles of the target characteristic is indistinguishable from zero. In other words, the model's predictions are unbiased. Second, the 10th and 90th percentiles are fairly symmetric around the median. Interestingly, the range in which the estimation errors fluctuate is smaller for percentiles in the tail of the distribution. Of course, for a true class of "0", the model can only deviate by predicting classes too high. However, this smaller fluctuation

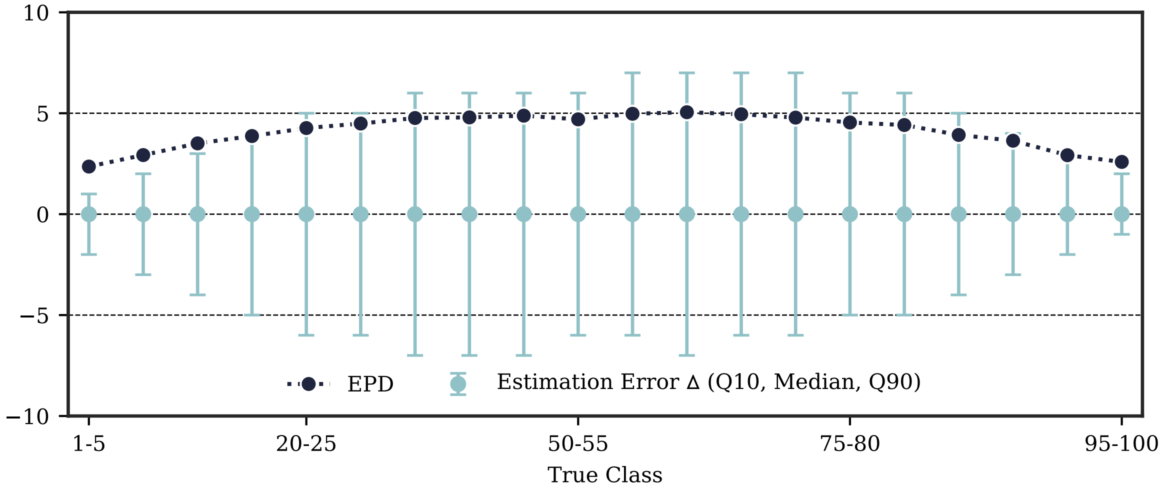

Fig. 10. Signed Estimation Error for Different Target Percentiles.

The figure shows both the distribution of the signed estimation error with error bars indicating the 10 % and 90 % quantile and the EPD as defined in Eq. (11) for different target percentiles.

for the tail percentiles is not only driven by the lower bound (positive deviations for classes 1–5 in Figure 10), but also evident in smaller deviations in the other direction (negative deviations in Figure 10). Third, this pattern is also found for EPD, which is generally smallest for tail percentiles.

5.2. Competing Approaches

The economic literature has come up with different ad-hoc methods for dealing with missing firm characteristics. The simplest method restricts the analysis to a sub-sample for which all information is available (Lewellen, 2015; Kelly et al., 2019). This will not only bias the results if the missing-at-random assumption is violated (Afifi and Elashoff, 1966; Freyberger et al., 2021) but also becomes infeasible if the number of characteristics considered is large and the subsample with all characteristics available shrinks. For example, retaining only the subsample of firms for which all 143 characteristics in our sample are available would reduce the sample size from roughly 3.1 million observations to only 186,158 – a reduction of more than 95%. In light of the multiple testing prob-

Table 3: Competing Approaches – Description

	Short description
Last	Last value imputation. For variables with quarterly updates we use the value of the previous quarter. If the last value is not available we impute the cross-sectional mean.
Hist. mean	Historical mean imputation which imputes the mean value of the last 12 months.
Mean by size	For each point in time, stocks are grouped into deciles by their market equity. We then impute the mean value of the corresponding size group for a specific characteristic.
Mean by industry	We use Kenneth French's industry classification to form 12 distinct industry groups at each point in time. We then impute the mean value of the corresponding group for a specific characteristic.
Mean imputation	Cross-sectional mean imputation method.
Mean by type	Characteristics are grouped by their type (Accounting, Market or Hybrid). We then impute the mean value of the corresponding group of characteristics for the target characteristic.
Mean by theme	Characteristics are grouped by theme (cf. Jensen et al., 2021) to capture similar aspects of a company. We then impute the mean value of the corresponding group of characteristics for the target characteristic.

lem, however, we would optimally test *all* possible characteristics and their combinations simultaneously.

To circumvent the issue of a decreasing sample size, the finance literature has come up with a simple ad-hoc method, the mean imputation method (e.g. Green et al., 2017; Gu et al., 2020, 2021), which simply imputes the cross-sectional mean and hence discards time-series variation if characteristics are cross-sectionally standardized. We describe this and alternative methods for imputing missing characteristics, which are tailored to capture a researcher's prior knowledge about how different characteristics may evolve in Table 3. In contrast, our machine learning method is agnostic about the underlying processes, as we have emphasized in Section 4, and thus does not require selecting a bespoke imputation approach for each characteristic.

We present methods that harness time-series information of a specific characteristic, either imputing the last available value of a target characteristic or a historical mean (Last, Hist. mean). Many characteristics are correlated over time such that past values

Table 4: Competing Approaches – EPD

The table shows the imputation accuracy as measured by the expected percentile deviation defined in Eq. (11) for different approaches to compare them with our model. The approaches are shortly described in Table 3. Characteristics are grouped by their theme as defined by Jensen et al. (2021) and results show the average EPD for all characteristics attributed to the same theme. The best method for each theme is highlighted in bold.

	Expected percentile deviation							
	Full model	Last	Hist. mean	Mean by size	Mean by industry	Mean imputation	Mean by type	Mean by theme
Accruals	**3.99**	12.95	19.82	24.74	24.10	24.99	27.89	41.41
Debt issuance	**4.55**	10.22	17.60	24.14	23.92	24.60	26.68	31.74
Investment	**3.98**	9.93	17.70	24.20	24.02	24.90	27.82	35.68
Leverage	**3.59**	4.81	8.56	23.27	21.36	26.08	26.29	25.61
Low risk	**4.99**	9.20	13.00	21.56	22.88	24.88	28.40	31.90
Momentum	**5.05**	14.39	23.36	23.98	24.01	25.27	22.68	13.95
Profit growth	**5.37**	14.39	21.77	23.88	23.59	24.26	23.49	21.09
Profitability	**3.32**	6.26	12.25	19.77	22.91	24.29	23.18	20.42
Quality	**3.15**	6.41	12.57	24.40	22.63	25.76	23.47	17.72
Seasonality	**9.06**	9.84	15.19	21.65	22.24	23.39	25.04	25.51
Size	**1.80**	1.99	7.23	11.62	22.85	25.61	26.53	32.93
Skewness	**8.27**	31.98	26.52	24.69	24.68	25.00	28.73	41.25
Value	**2.53**	3.65	11.11	24.12	22.84	26.84	21.57	20.75
Mean	**4.59**	10.46	15.90	22.46	23.23	25.07	25.52	27.69

provide information for future realizations. Of course this requires time-series information to be available. Alternative methods leverage information of clusters of stocks, i.e., stocks clustered by market capitalization or industry (Mean by size, Mean by industry). Characteristics of firms of similar size or within the same industry potentially exhibit similar dynamics, such that averages of these groups may serve as a good proxy to capture these dynamics. Lastly we consider information provided from other characteristics (Mean by type, Mean by theme). Grouping characteristics in terms of type or theme aims at identifying characteristics with similar content to come up with predictions for missing characteristics.[15] These approaches will work well if characteristics with similar

[15] To assure that characteristics of one type or within a theme are comparable, we sort them such that higher values indicate higher returns unconditionally, as measured by the characteristic-sorted high-minus-low return spread.

informational content have a sufficiently high correlation such that they are placed in similar cross-sectional percentiles. To test these ad-hoc approaches for imputing missing values we compare their EPD to that of our model for different characteristic themes. Results are shown in Table 4.

Our model outperforms the other approaches for all characteristic themes and for most themes it does so by a large margin. The best competitor is the Last imputation method which on average still produces an EPD twice that of our model. While imputing the last value comes close to our model's accuracy for "Size" and "Value" characteristics it fails to produce meaningful imputations for characteristics that proxy for "Profit growth", "Momentum", and most notably "Skewness" (EPD of 31.98 vs. our model's 8.27). The third best model again leverages temporal information and imputes a historical mean. This produces an EPD averaged across the twelve characteristic themes of 15.90, more than tripling our model's 4.59. The remaining methods all perform poorly, with EPDs around or even above 25. An EPD of 25 is obtained when imputing the cross-sectional mean.[16] It seems as though clustering information by different portions of the cross-section of stocks or the information content of characteristics is insufficient in finding suitable imputations for missing firm characteristics. Instead, the success of our proposed model architecture stresses that time-series and cross-sectional information should be combined and that the informational content of observable firm characteristics is more complex than accounted for by the aggregations considered here.

6. Recovering Missing Firm Characteristics

The objective of our study is to present a method to recover missing firm characteristics and provide a completed data for future research. Naturally, these predictions are as-

[16] Considering the amount of variation explained (R^2) shown in Table A1 confirms these findings.

sociated with uncertainty. The choice of predicting cross-sectional percentiles produces a probability distribution across the percentiles of a target characteristic. We use this probability distribution to quantify the uncertainty of each prediction – a feature unavailable to other imputation methods, such as the common mean-imputation or alternative approaches proposed by contemporaneous studies.

Model Uncertainty. When reconstructing masked characteristics in the previous section, we have full knowledge about the desired outcome and a direct way to assess the quality of the reconstruction. Instead, when applying the model to missing entries we do not. However, we can quantify the uncertainty associated with each prediction. "Good" predictions have two properties: 1) reconstructing the characteristic generally works well within the fully controlled environment of Section 5, i.e. it produces a low EPD, and 2) the estimated probability distribution across percentiles of a missing firm characteristic is significantly different from an uninformed guess.

To address 2), we proceed in two ways: for an uninformed guess the model is unable to produce a meaningful distribution across percentiles. In these cases, it approximately defaults to a uniform distribution, predicting each percentile with equal probability. However, even if the model's estimated probability distribution is statistically indistinguishable from a uniform distribution, the prediction may still correctly identify the true percentile. We argue that the model's predictions are internally consistent, whenever the percentiles which are assigned the highest probabilities are in close proximity.

We apply the Kolmogorov-Smirnov test to formally assess if the estimated distribution is significantly different from a uniform distribution (Massey Jr., 1951).[17] Let $F_U(x)$ denote the cumulative distribution function of a uniformly distributed random variable,

[17] Note that the Kolmogorov-Smirnov test likely underrejects the null hypothesis in our setting, given that we discretize the input space to 100 percentiles. The test-statistic scales with $\sqrt{1/N}$, where $N = 100$ in our setting.

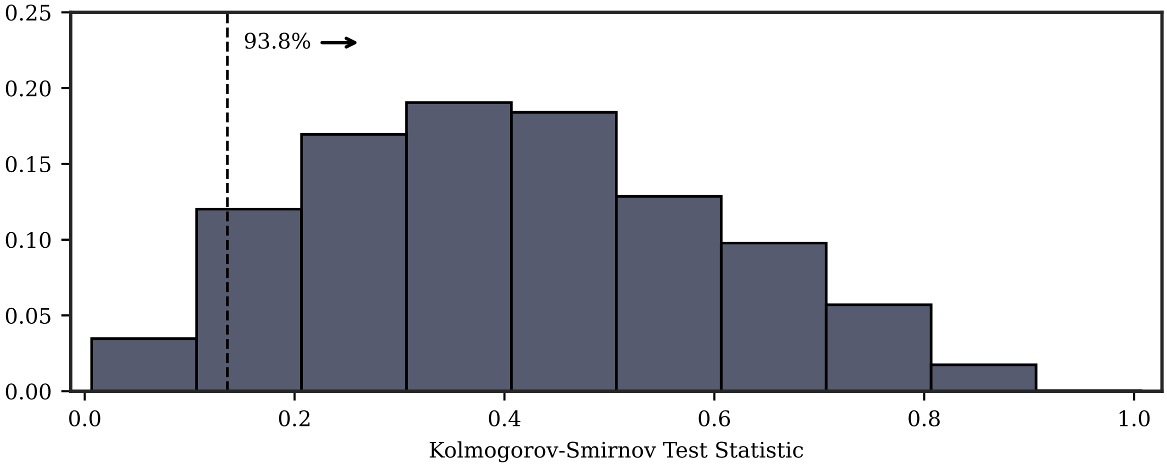

Fig. 11. Model Confidence.

The figure shows a histogram of the recovery confidence of our model, measured by the Kolmogorov-Smirnov test statistic. The test measures whether the probability distribution across all classes for each recovered characteristic as predicted by our model is significantly different from a uniform distribution. The dashed vertical line indicates the 5 % confidence level. Based on this measure we find 93.8 % of the reconstructed characteristics to be predicted with confidence.

on the interval $x \in [1, 100]$. Likewise, let $F^c_{\mathrm{M}}(x)$ denote the cumulative step-function estimated by our model when generating the probability distribution across percentiles for target characteristic c. Then, the Kolmogorov-Smirnov test statistic d^c for the model prediction is given by

$$d^c = \max_x |F_U(x) - F^c_{\mathrm{M}}(x)| \tag{17}$$

In our case, $F^c_M(x)$ significantly differs form a uniform distribution at the 5 % significance-level if d^c exceeds $d_{\max} = 0.136$ (Massey Jr., 1951).

Figure 11 shows the histogram of the Kolmogorov-Smirnov statistic for the pooled set of filled entries of firm characteristics. For 93.8% of the recovered entries of firm characteristics the model produces a probability distribution across the characteristic's percentiles that is significantly different from a uniform distribution.

For the remaining 6.2% of missing entries, we now analyze how far apart the percentiles with the highest and next-highest assigned probability lie. In 58.6% of cases in which we fail to reject the null of a uniform assignment across percentiles, the percentile with the highest and the percentile with the next-highest probability are just one percentile apart. In 69.1% of cases they differ by no more than five percentiles. In total, for only 1.9% of the recovered missing entries the predicted probability distribution is indistinguishable from a uniform distribution *and* the two classes with the highest assigned probability disagree with regards to the general location of the true percentile. Of course other researchers may want want to apply different benchmarks to identify recovered firm characteristics with low model uncertainty. Hence, we provide the output probability distribution of our model alongside each recovered characteristics to encourage future research on this topic.

Discussion One may argue that the model should produce a probability distribution with a single peak – a predicted percentile with 100% probability associated. However, this result, while theoretically appealing, is highly unrealistic empirically and in no way expedient. For one, predictions depend on the observable characteristics of a firm. Small changes in single input characteristics or in combinations of characteristics may lead to vastly different predictions of missing characteristics. We require the model to take these nuances into account and balance their informational content. Inevitably, this leads to a dispersion in the probability distribution. At the same time, the input characteristics, even if observable, are measured with noise. This noise directly translates to uncertainty in the model's predictions. Lastly, we use the so-called focal loss, described in Section 3, to force the model to also learn about reconstructing characteristics for which a reconstruction is harder. Above a certain probability threshold, assigning a larger probability to the true percentile does not improve the estimation loss. In our case, the threshold amounts to roughly 20%. We are in the unique position to quantify the model's uncertainty and believe it to be an important aspect of imputing missing firm characteristics.

Recovering Raw Firm Characteristics To recover missing firm characteristics, we have considered the characteristic's distribution, discretized to a fine-grid of percentiles. Instead, many applications in Accounting, Management, and Marketing research require the actual values, not just their cross-sectional distribution. We can back out reasonable estimates for the *raw characteristics*, by interpolating between values of the characteristic in the predicted percentiles observed for other firms.

We consider three methods to come up with estimates of raw firm characteristics. For a given recovered percentile for the target characteristic, we first identify all firms that fall within said percentile. Within this set, we then identify the firms which have the lowest and highest value of the characteristic. The first method simply linearly interpolates between these two edge points, and reports the "mid" value therein. The second and third methods give the "mean" and "median" of all observed values within the respective percentile instead. Revisiting the multitude of established results in economic research using this completed dataset is beyond the scope of this paper. However, as an application we test how the recovered firm characteristics influence factor premia in characteristic-based asset pricing.

7. Application: Factor Portfolios in Finance

What is the impact of changes in the distribution of firm characteristics after filling in missing values? The answer of course depends on the research question and can be addressed in many ways. For one attempt at providing an answer, we choose to study the impact on high-minus-low factor portfolios, which are a common application of this data in characteristic-based asset pricing and have been the cornerstone of financial research since at the very latest Fama and French (1993).

We first sort stocks into deciles for a given characteristic. We then calculate the returns for each decile portfolio weighted by each included firm's market capitalization in the past month, and first discard missing values ("Pre"), and then use the completed data set with imputed values ("Post"). Consequently, we form the zero-cost factor portfolio as the difference between the highest and lowest decile portfolio.[18] Changes in portfolio returns arise if the firms with recovered characteristics have different return patterns than the average firm in the portfolio before recovering missing values. Note that we purposefully separate the imputation of missing firm characteristics and the calculation of factor returns. This is different from Freyberger et al. (2021), who impute firm characteristics conditional on explaining stock returns. Instead, our approach is agnostic about whether missing entries of firm characteristics should relate to return patterns and only investigates this after the imputation.

Figure 12 shows the change in high-minus-low factor portfolio returns for the 30 characteristics with the ex-post largest change in the factor premium due to the inclusion of recovered firm-characteristics. The high-minus-low returns discarding stock observations with a missing entry for the sorting characteristic are given in black ("Pre"), whereas the red circles denote the returns *after* considering the impact of imputed missing values ("Post"). A clear trend emerges: using the completed set of firm characteristics pushes most high-minus-low spreads towards zero. Examples with large changes in the average high-minus-low spreads are `fcf_me` (pre: 13%, post: 6%), the change in net-operating assets `noa_gr1a` (pre: 13%, post: 7%), and momentum `ret_12_1` (pre: 21%, post: 16%). A complete list can be found in the Internet Appendix Table IA3.1. Across all 143 characteristics, we find that the absolute return spread decreases by an average of -1.26 % with a Newey and West (1987) t-value of -6.67.

[18] We follow Jensen et al. (2021) and cap the value-weights by the 80th percentile of the market capitalization of all firms in any given months to limit the influence of large outliers. To focus on the outright changes arising due to a change in the portfolio decomposition after imputing missing values, we also consider equally-weighted returns in Appendix IA4. Using value-weights invariably masks part of the impact of missing values, as information for large stocks is more complete and overall of better quality.

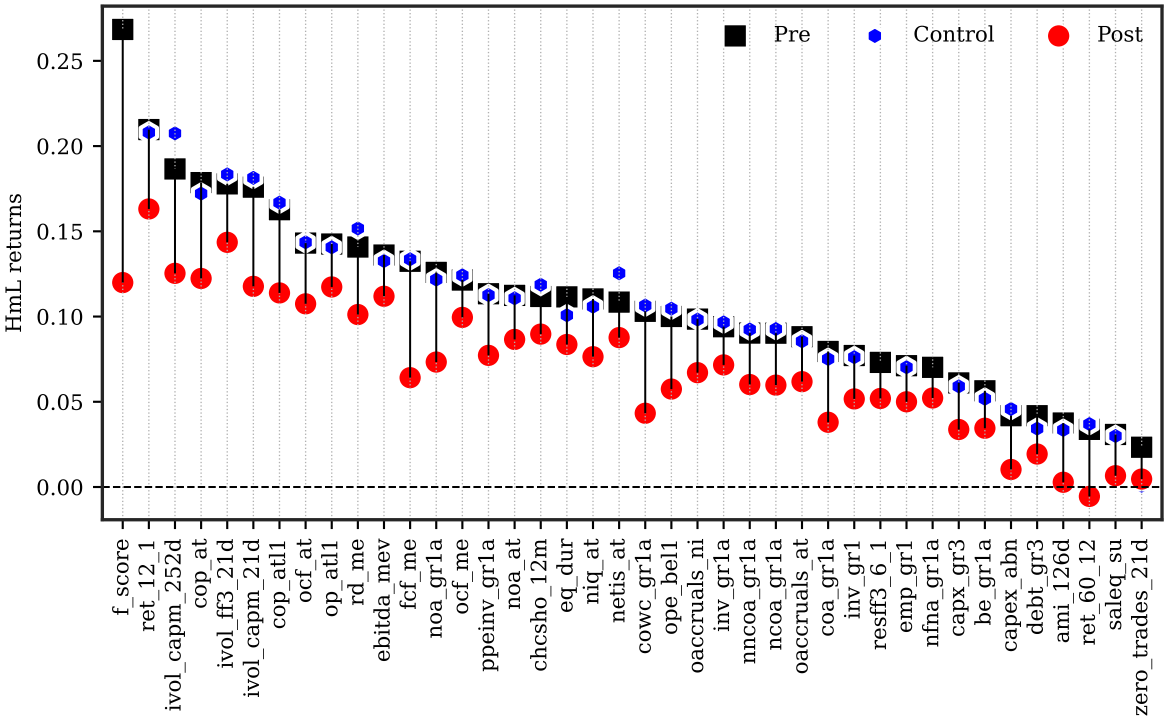

Fig. 12. Impact of Missing Observations on Factor Portfolio Returns.

The figure shows the change in high-minus-low factor portfolio returns for the 30 characteristics with the ex-post largest change in the factor premium due to the inclusion of recovered firm-characteristics. The premium without incorporating the information of imputed missing characteristics is given in black ("Pre"), the premium after the inclusion of this information in red ("Post"). Blue data points display the control which consists of all observations of the "Pre"-group but using characteristic values *reconstructed* by our model. This is to show that our model does not mechanically drive down the high-minus-low returns. A complete list of these changes is provided in Table IA3.1.

We investigate two potential issues, which may give rise to a mechanical reduction in absolute return spreads: first, one may argue that even if high-minus-low return spreads are correctly identified using the uncompleted dataset, through the imputation approach some stocks will be wrongfully allocated to the high and low portfolio. Assuming that returns are a monotonic function of the true portfolio, this will push down on the high-minus-low return spreads after imputation. A first indication that this is less of a concern using our model is given in Figure 10: the reconstruction accuracy is generally highest in the high and low portfolios. Another potential concern relates to the use of past information about characteristics to impute today's missing values. If the imputation relies heavily on this past information, imputed values are potentially less informative about future returns.[19] Baba Yara et al. (2020) investigate the predictive power of “old” information and generally finds that past characteristics still generate large return spreads.

To address these points, we proceed as follows: we create a *reconstructed* data set which reconstructs the observations of the uncompleted dataset, using our model, in the same fashion as in Section 5: we randomly mask and reconstruct 20 % of the available input characteristics. We repeat this step 25 times to ensure that $\geq 99\%$ of characteristics have been reconstructed. The resulting high-minus-low return spreads for this reconstructed data set are also shown in Figure 12 as the blue circles (“Control”). The results clearly show that there is no mechanical downward-bias induced by our model. Regression results further support this evidence: high-minus-low return spreads using the reconstructed and uncompleted dataset are statistically indistinguishable across the 143 characteristics.

The average impact on high-minus-low return spreads is large and quite heterogeneous. Some return spreads are barely impacted, others by a lot. How much of this impact is driven by changes in the average returns of the long vs. the short component of the characteristic-sorted factor portfolios? The post-completion change in the average high-minus-low return is given by

$$\Delta\text{HmL} = \text{HmL}^{\text{Post}} - \text{HmL}^{\text{Pre}} = \Delta\text{Hi} - \Delta\text{Lo} \tag{18}$$

The red line in Figure 13 provides the distribution of ΔHmL across the 143 characteristics considered and visually confirms the evidence that average long-short portfolio returns tend to decrease using the completed dataset. For ease of exposition, we show the Pre-to-Post change in the negative returns of “Lo”, which makes it comparable to how changes to “Hi” are presented and conforms with the profits a long-short investor would make

[19] We thank Andrea Barbon for pointing out this potential issue to us.

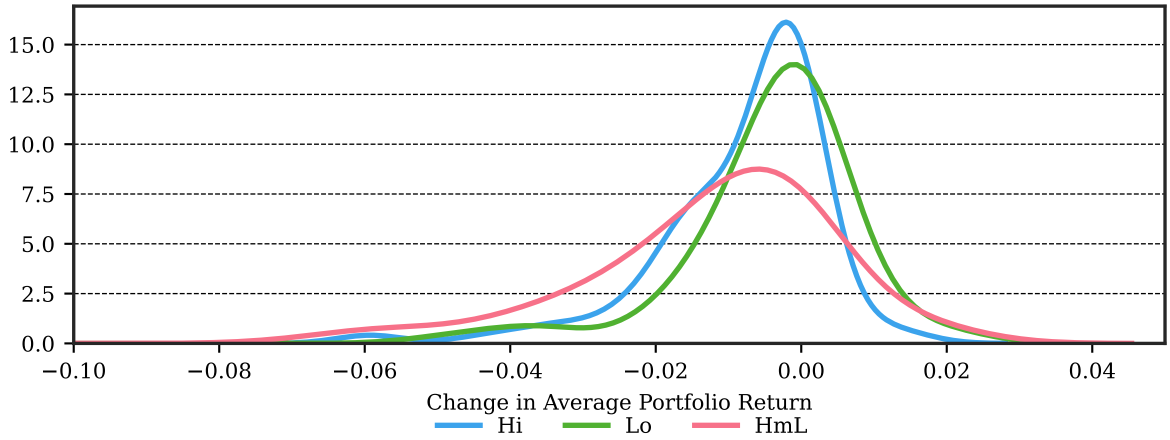

Fig. 13. Distribution of Changes in Factor Returns

The figure shows a kernel density estimate of the change in portfolio returns post-completion, as defined in Eq. (18). The red line shows the distribution of ΔHmL across the 143 firm characteristics considered, while the blue (green) line show the results for the high (low) portfolio, respectively.

when shorting portfolio "Lo". We find that the average profits of both the high and low portfolio decrease for the average characteristic but that the return change for the high portfolio is on average more negative. Average returns to the low (high) portfolio on average decrease by −0.52% (−0.80%) per year. Interestingly, the changes in average returns of the low and high portfolio are uncorrelated, with a correlation coefficient of just 0.028 across characteristics. This shows that our completion procedure does not merely lead to a uniform reduction in the returns of the long and short leg of the factor portfolios.

Our results stress the importance of carefully considering the impact of missing firm characteristics and provide an additional hurdle for newly proposed risk factors to pass: the factors should survive not only in the sample in which they are available outright, but also using the extended sample including firms with missing observations. In total we find that 6 of the 143 factor premia lose their significance. Still, most factor premia that are significant before the imputation remain significant thereafter, and others even gain significance (see Table IA3.1). We have already highlighted that our approach is agnostic about whether firm characteristics are informative about stock returns, in contrast to the study by Freyberger et al. (2021). The described analysis therefore adds a new out-of-sample test for assessing the validity of newly-found and existing risk factors. With this, we complement the recent debate on whether financial research experiences a replication crisis (Harvey, Liu, and Zhu, 2016; Jensen et al., 2021). Judging by the relative stability of most factor premia with respect to the impact of missing values (in terms of their significance, not necessarily their magnitude), we argue in favor of replicability in Finance.

8. Which Information Set Matters?

We have so far shown that our model performs well in reconstructing percentiles of firm characteristics and that completing the panel of 143 firm characteristics has profound implications for risk factors in asset pricing. We now investigate the importance of incorporating nonlinear dependencies between characteristics, between assets, and over time in more detail, by looking at the model's internal structure to assess how it comes up with its predictions, as well as by considering simpler – nested – models that include only a subset of the information available to the full model specification.

Feature Importance To assess which characteristics the model uses to reconstruct missing entries of target characteristic c, we express the *feature attention matrices* described in Section 2 as directed weights, wherein each row indicates how much information about each characteristic is necessary to reconstruct a masked entry of c. First, we assess for each characteristic type (accounting, hybrid, market) the importance of information about characteristics of the same type, as well as the other two. Second, we once again follow Jensen et al. (2021) and cluster the 143 characteristics into twelve themes, such as "Accruals" or "Skewness". We then investigate for each characteristic theme, how impor-

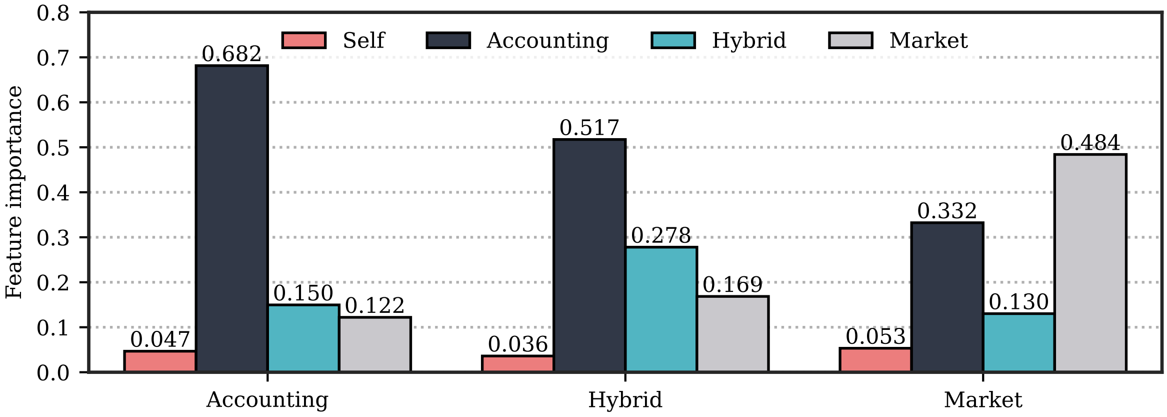

Fig. 14. Feature Importance by Characteristic Type.

The figure shows the average feature importance weights for information drawn from the target characteristic itself ("Self"), as well as the joint importance of characteristics of each group, including accounting-based, market-based, as well as hybrid characteristics. We also split the target characteristic by these groups to show how the information flow changes. Note that the model has no information about these groupings, they arise organically from the data. The feature attention per characteristic group naturally sums up to one.

tant information about characteristics of the same theme is in comparison to information from the other eleven themes. For example, we may be interested in the importance of "Size" characteristics, when reconstructing a characteristic from the "Value" group. This analysis also provides an intuitive justification for including the 143 firm characteristics jointly, should information from the other themes be important in the model's reconstruction of firm characteristics.

Figure 14 provides this breakdown at the level of characteristic types. Since the row-wise sum of the attention matrix is always 1, we can simply add up the values for characteristics belonging to each of the three types.[20] We also separately highlight the importance of historical information about the characteristic itself. If each characteristic was equally informative about all others, this self-importance would amount to $1/N = 1/143 \approx 0.007$.

[20] Diebold and Yılmaz (2014) estimate directed networks between firms using observable returns in a VAR-framework. As an analogy, the attention would be the connectedness matrix in their framework. We are interested in the "From"-connectedness, i.e. how much the internal representation of a target characteristic is influenced by the others.

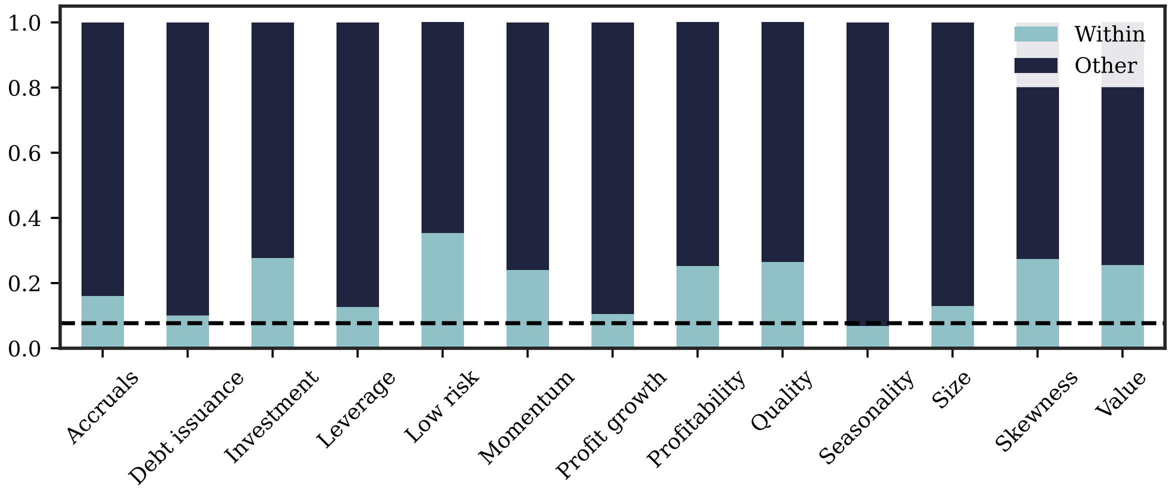

Fig. 15. Feature Importance by Theme of Characteristic.

The figure shows the average feature importance weights for the twelve themes of characteristics identified by Jensen et al. (2021), split by information drawn from characteristics of the same theme ("Within") versus other themes ("Other"). The dotted line denotes the theoretical feature importance if each theme was equally informative about all other themes ($1/N^{\text{themes}} = 1/12$).

We find a high importance of same-type information: to reconstruct accounting-based characteristics, information about other accounting-based characteristics is most important. The same applies to market-based characteristics. As expected, reconstructing hybrid characteristics requires a mix of all other characteristic types. Importantly, the model places a high weight on the historical evolution of target characteristic c, from 3.6% for hybrid variables up to 5.3% for market-based characteristics. While we find a high importance of same-group information, the weight placed on this information is far below 100% for all characteristic types. This highlights both the benefits of including as many characteristics as possible, as well as the tremendous flexibility of our modeling approach in assessing and using this vast amount of information.

In Figure 15 we show how characteristics of different themes relate to one another. Consistent with the previous evidence from types of characteristics, the model consistently leverages information about other themes when reconstructing a characteristic of a target theme. Still, the feature importance of information from the same theme ("Within") is

Table 5: Temporal Attention Weights.

The table shows average temporal attention weights for each year in the specified look-back window of 5 years. Similar to feature importance weights, temporal attention weights measure how much information from each historical time-steps is incorporated in the final prediction of the model. Quantiles are calculated from the cross-section of firms for each month and consequently averaged across time. The mean of temporal attention naturally sums up to one.

		Quantiles						
	Mean	1	5	25	50	75	95	99
				Full				
Year-1	0.941	0.828	0.883	0.921	0.942	0.965	0.999	1.000
Year-2	0.037	0.000	0.000	0.021	0.036	0.049	0.076	0.114
Year-3	0.015	0.000	0.000	0.003	0.013	0.023	0.039	0.059
Year-4	0.006	0.000	0.000	0.000	0.003	0.010	0.021	0.033
Year-5	0.001	0.000	0.000	0.000	0.000	0.001	0.005	0.013

fairly high, between 8% for “Seasonality” and 35% for “Low risk”. These results highlight the benefits of including characteristics that capture different facets of a firm’s finances, measured both from a market- and accounting-based point of view.

Time Importance We explicitly account for the historical evolution of input characteristics in a flexible fashion, such that the model may incorporate varying levels of temporal information per target characteristic. Especially accounting-based characteristics may benefit from this inclusion, given that they typically fluctuate little from quarter to quarter. For example, Gonçalves (2021) models the evolution of a firm’s equity duration using a vector autoregressive process with lag 1. But this inclusion may also provide fruitful information for market-based characteristics. Keloharju et al. (2021) have recently shown that it is not today’s value for firm characteristics that has explanatory power over returns, but rather a characteristic’s deviation from its long-run mean. Table 6 shows how identifying both this long-run mean, as well as how a characteristic fluctuates around this mean is beneficial when recovering missing firm characteristics.

Table 5 provides the results. While we allow the model to incorporate information from up to five historical years, we find overwhelming evidence that most information is drawn from the past year. The mean attention put on this year amounts to 94.1%, with comparatively little variation over time. The first percentile of how much weight is placed on the most recent year still amounts to 82.8%. In contrast, the second year receives on average about 3.7% of the total attention, with occasional spikes above 11.4%. In line with this, tuning the hyperparameters for the model reveals a preference for sparse temporal attention weights, using the EntMax normalization function outlined in Eq. (2). Last year's information is imperative when making informed predictions of missing characteristics. Current and near-term values of firm characteristics already incorporate most information necessary, highlighting the efficiency of modern financial markets and financial reporting.

Restricting the Information Set Besides our full model architecture outlined in Section 2, we may also consider a number of nested models, which restrict the reconstruction of firm characteristics to using different dimensions of the input data. We have noted that firm characteristics correlate over time, across assets, and between different characteristics of the same firm. For example, we have just highlighted in the previous section that information from the most recent year is most important. But how important is the inclusion of this temporal information and how well would a model work that only operates on the cross-section of observable firm characteristics in month t?

To assess the importance of incorporating the temporal evolution of firm characteristics, we consider two nested models: the first is restricted to information about the characteristics masked for reconstruction, blending out all other information. At the same time, we disallow the model to interact input characteristics, but add the embedding step. The model may therefore dissect the cross-section of stocks per target characteristic and apply different levels of processing for stocks in different percentiles of the target characteristic.[21] The model is thus forced to reconstruct characteristics using only their historical information, capturing their autoregressive component. We call it the "Temporal model". The second model is given information about all $F = 143$ characteristics, but measured only at time t, fully disregarding their temporal evolution. We denote it as the "X-sectional model".

In the last nested case ("no-self model"), we assess how important historic information about the target (masked) characteristics are. We allow the model to incorporate information about non-masked characteristics, as well as their historical evolution, but provide no historic information about the masked characteristics themselves. Note that this is a highly restrictive model setup, which assumes that we lack any (historic) information about 20% of the input characteristics (those we have randomly masked). Once again, we include the mean imputation approach for comparative purposes, and separately show the accuracy for the three types of characteristics considered (accounting, market, hybrid).

The full model performs best with an EPD of 4.31 for the entire sample – a 2-fold improvement over each of the nested cases. In fact, we find little variation of the full-sample scores for any of the nested models. However, when considering characteristic types individually, we do find slight differences between the models. For example, the temporal model slightly outperforms both the cross-sectional and no-self model for accounting and hybrid variables. For market variables instead, the temporal model performs slightly worse with an EPD of 9.86, which is still a 2.5-fold improvement over imputing the mean. Notably, the full model specifications consistently outperforms the nested models for each characteristic type individually and across all characteristics. Keep in mind that we reconstruct masked entries of all 143 characteristics in a joint model and do not

[21] A simple example is to apply different levels of processing to large and small stocks, which differ in many dimensions, not the least of which is the degree of missing information, which is systematically higher for small stocks, see Figure 4.

Table 6: Nested Models – Accuracy by Imputation Method.

The table shows the imputation accuracy measured by the expected percentile deviation defined in Eq. (11). We differentiate our model's accuracy from that of a cross-sectional model, which disregards temporal information and a temporal model, which disregard information from other characteristics. The no-self model disregards all information on the target characteristics but attends to information about other characteristics and their temporal evolution. We further consider imputing masked features with the cross-sectional median as the benchmark. Results are shown for market- and accounting-based, as well as hybrid characteristics. The best performing model is highlighted in bold for each case.

	Expected percentile deviation			
	Full	Training	Validation	Testing
		All		
Full model	**4.31**	**3.63**	**4.03**	**4.67**
X-Sectional model	9.02	7.19	8.14	10.01
Temporal model	8.30	7.40	7.82	8.80
No-self model	8.38	6.66	7.52	9.31
Mean imputation	25.08	25.08	25.13	25.07
		Accounting		
Full model	**4.10**	**3.42**	**3.89**	**4.45**
X-Sectional model	9.56	7.73	8.81	10.52
Temporal model	7.99	7.13	7.63	8.45
No-self model	8.98	7.18	8.24	9.92
Mean imputation	24.75	24.74	24.67	24.77
		Market		
Full model	**5.29**	**4.65**	**4.90**	**5.65**
X-Sectional model	9.16	7.66	8.30	10.03
Temporal model	9.86	8.81	9.02	10.51
No-self model	7.91	6.66	7.16	8.63
Mean imputation	25.01	25.03	25.05	24.99
		Hybrid		
Full model	**2.89**	**2.04**	**2.50**	**3.32**
X-Sectional model	6.96	4.30	5.47	8.35
Temporal model	5.97	5.11	5.72	6.38
No-self model	7.43	4.92	5.92	8.76
Mean imputation	26.32	26.35	26.87	26.21

refit the models for only accounting or market-based characteristics, providing a unified framework for dealing with missing firm information.

9. Conclusion

A vast literature in empirical asset pricing uses observable firm characteristics as proxies for differences in expected stock returns. However, firm characteristics are frequently missing, some more often than others. Basing inference on the set of stocks for which a target characteristic is available potentially biases the results due to variation in the sample considered. For example, Figure 4 shows that small firms are consistently missing more information. Therefore, results obtained from a dataset in which missing information is discarded or incorrectly dealt with may not generalize well to the case of small stocks.

In this paper, we propose a comprehensive machine learning method, which borrows from recent advances in natural language processing and adapts their insights to the case of financial data to fill these missing entries. For this, we use three types of information: about other characteristics of a target firm, how these characteristics evolved over time, and from the cross-section of other firms. We show in a first step that the proposed model vastly outperforms ad-hoc methods, such as imputing the cross-sectional mean, but also more involved methods, tailored to leverage ex-ante information a researcher may have about how firm characteristics may evolve. We can also show that predictions are unbiased and highlight that the inclusion of all three types of information is vital to the model's success. Second, our model setup allows us to explicitly quantify the uncertainty attached to each prediction. Fully uninformed predictions will produce a uniform distribution across a target characteristic's percentiles. We show that in most cases the model's predictions significantly differ from a uniform distribution. Third, we use the completed dataset, i.e., a panel with imputed missing entries, to investigate changes in average returns of common risk factors. For most characteristics, absolute high-minus-low return spreads are lower using the completed dataset. We carefully assure that this is not driven by sorting on "old" information. Still, most return spreads remain significant, adding another piece of evidence in favor of replicability in finance (Jensen et al., 2021). Finally, we investigate the model's internal mechanism to come up with its predictions for missing firm characteristics and show that it leverages information of all 143 characteristics, stressing a) the model's flexibility and b) benefits of including information about many aspects of a firm's finances. We have made the filled entries of firm characteristics, as well as the associated modeling uncertainty available for future research.[22]

[22] The imputed firm characteristics can be downloaded from the first author's website.

References

Abrevaya, J., Donald, S. G., 2017. A gmm approach for dealing with missing data on regressors. Review of Economics and Statistics 99, 657–662.

Afifi, A. A., Elashoff, R. M., 1966. Missing observations in multivariate statistics i. review of the literature. Journal of the American Statistical Association 61, 595–604.

Arik, S. O., Pfister, T., 2019. Tabnet: Attentive interpretable tabular learning (2019). arXiv preprint arXiv:1908.07442 .

Ba, J. L., Kiros, J. R., Hinton, G. E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450 .

Baba Yara, F., Boons, M., Tamoni, A., 2020. New and old sorts: Implications for asset pricing. Martijn and Tamoni, Andrea, New and Old Sorts: Implications for Asset Pricing (January 31, 2020) .

Bali, T., Goyal, A., Huang, D., Jiang, F., Wen, Q., 2021a. Different strokes: Return predictability across stocks and bonds with machine learning and big data. Swiss Finance Institute, Research Paper Series .

Bali, T. G., Beckmeyer, H., Moerke, M., Weigert, F., 2021b. Option return predictability with machine learning and big data. Available at SSRN 3895984 .

Banz, R. W., 1981. The relationship between return and market value of common stocks. Journal of financial economics 9, 3–18.

Bryzgalova, S., Lerner, S., Lettau, M., Pelger, M., 2022. Missing financial data. Available at SSRN 4106794 .

Cahan, E., Bai, J., Ng, S., 2022. Factor-based imputation of missing values and covariances in panel data of large dimensions. Journal of Econometrics .

Chen, A. Y., Zimmermann, T., 2020. Open source cross-sectional asset pricing. Critical Finance Review, Forthcoming .

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Diebold, F. X., Yılmaz, K., 2014. On the network topology of variance decompositions: Measuring the connectedness of financial firms. Journal of econometrics 182, 119–134.

Fama, E., French, K., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics .

Freyberger, J., Höppner, B., Neuhierl, A., Weber, M., 2021. Missing data in asset pricing panels. Available at SSRN .

Gonçalves, A. S., 2021. The short duration premium. Journal of Financial Economics .

Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A., 2021. Revisiting deep learning models for tabular data. arXiv preprint arXiv:2106.11959 .

Green, J., Hand, J. R., Zhang, X. F., 2017. The characteristics that provide independent information about average us monthly stock returns. The Review of Financial Studies 30, 4389–4436.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

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

Harvey, C. R., Liu, Y., Zhu, H., 2016. … and the cross-section of expected returns. The Review of Financial Studies 29, 5–68.

Heston, S. L., Sadka, R., 2010. Seasonality in the cross section of stock returns: the international evidence. Journal of Financial and Quantitative Analysis 45, 1133–1160.

Huang, X., Khetan, A., Cvitkovic, M., Karnin, Z., 2020. Tabtransformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv:2012.06678 .

Jensen, T. I., Kelly, B. T., Pedersen, L. H., 2021. Is there a replication crisis in finance? Tech. rep., National Bureau of Economic Research.

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

Keloharju, M., Linnainmaa, J. T., Nyberg, P., 2021. Long-term discount rates do not vary across firms. Journal of Financial Economics .

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kozak, S., Nagel, S., Santosh, S., 2020a. Shrinking the cross-section. Journal of Financial Economics 135, 271–292.

Kozak, S., Nagel, S., Santosh, S., 2020b. Shrinking the cross-section. Journal of Financial Economics 135, 271–292.

Lewellen, J., 2015. The cross-section of expected stock returns. Critical Finance Review 4, 1–44.

Lim, B., Arık, S. Ö., Loeff, N., Pfister, T., 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting .

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 .

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 .

Lundberg, S. M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777.

Martins, A., Astudillo, R., 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In: *International conference on machine learning*, PMLR, pp. 1614–1623.

Massey Jr., F. J., 1951. The kolmogorov-smirnov test for goodness of fit. Journal of the American statistical Association 46, 68–78.

Newey, W. K., West, K. D., 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica 55, 703–708.

Novy-Marx, R., 2012. Is momentum really momentum? Journal of Financial Economics 103, 429–453.

Peters, B., Niculae, V., Martins, A. F., 2019. Sparse sequence-to-sequence models. arXiv preprint arXiv:1905.05702 .

Piotroski, J. D., 2000. Value investing: The use of historical financial statement information to separate winners from losers. Journal of Accounting Research pp. 1–41.

Smith, L. N., Topin, N., 2017. Super-convergence: Very fast training of neural networks using large learning rates. arxiv e-prints, page. arXiv preprint arXiv:1708.07120 .

Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., Goldstein, T., 2021. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. arXiv preprint arXiv:2106.01342 .

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15, 1929–1958.

Tran, P. T., et al., 2019. On the convergence proof of amsgrad and a new version. IEEE Access 7, 61706–61716.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: *Advances in neural information processing systems*, pp. 5998–6008.

Wilks, S. S., 1932. Moments and distributions of estimates of population parameters from fragmentary samples. The Annals of Mathematical Statistics 3, 163–195.

Appendix A. R^2 by Imputation Method

Table A1: Model Comparison – R^2 by Imputation Method.

Table A1 shows the imputation accuracy by R^2, which measures how much of the variation in reconstructing masked characteristics a method can explain that is not already explained by imputing the cross-sectional mean. As we discretize each characteristic into percentiles we calculate the R^2 in the following fashion:

$$R^2_{\mathrm{x}} = 1 - \frac{\sum_{k=0}^{99} \, p_{\mathrm{x}}(|\Delta| = k) \cdot (k/100)^2}{\sum_{k=0}^{99} \, p_{\mathrm{MI}}(|\Delta| = k) \cdot (k/100)^2} \tag{A1}$$

with subscript x indicating the current method being evaluated and MI standing for the *Mean Imputation* method. We differentiate our model's accuracy from that of the approaches outlined in Table 3 and separately provide the achieved R^2s for the twelve themes of characteristics from Jensen et al. (2021). The best performing model is highlighted in bold for each case.

	R^2							
	Full model	Last	Hist. mean	Mean by size	Mean by industry	Mean imputation	Mean by type	Mean by theme
Accruals	**0.91**	0.43	0.07	0.01	0.04	0.00	−0.28	−1.82
Debt issuance	**0.87**	0.55	0.15	0.02	0.04	0.00	−0.21	−0.84
Investment	**0.90**	0.60	0.20	0.04	0.04	0.00	−0.30	−1.20
Leverage	**0.92**	0.88	0.70	0.15	0.24	0.00	−0.07	−0.13
Low risk	**0.85**	0.66	0.54	0.19	0.12	0.00	−0.36	−0.76
Momentum	**0.91**	0.49	−0.01	0.08	0.07	0.00	0.16	0.61
Profit growth	**0.82**	0.28	−0.13	0.02	0.03	0.00	0.03	0.18
Profitability	**0.92**	0.78	0.48	0.26	0.08	0.00	0.06	0.04
Quality	**0.91**	0.79	0.50	0.07	0.16	0.00	0.15	0.44
Seasonality	**0.55**	0.48	0.23	0.10	0.06	0.00	−0.18	−0.30
Size	**0.98**	0.98	0.77	0.70	0.14	0.00	−0.11	−1.17
Skewness	**0.80**	−0.88	−0.21	0.02	0.02	0.00	−0.36	−1.85
Value	**0.96**	0.92	0.60	0.12	0.19	0.00	0.30	0.26
Mean	**0.87**	0.53	0.30	0.14	0.09	0.00	−0.09	−0.50

Appendix B. Accuracy, Missingness and Model Confidence

Table B2 provides summary information about the model accuracy as measured by the expected percentile deviation defined in Eq. (11) for each characteristic separately. Characteristics are sorted from best to worst model accuracy. We further include the missingness of each characteristic in the data set for all firm×month observations. "sig. KS-Test" provides information on how many of the recovered firm characteristic exhibit a significant - i.e. on the 5 % level - KS-statistic as defined in Section 6, which is a measure of model confidence for the recovered firm characteristics. We further classified characteristics in accounting (A), hybrid (H) and market (M) variables.

Table B2: Missingness, accuracy and model confidence per characteristic.

	Expected percentile deviation						
	Full	Training	Validation	Testing	Miss. [%]	sig. KS-Test [%]	Class
age	0.74	0.64	0.77	0.78	0.00	-	H
market_equity	1.26	0.90	1.15	1.42	0.48	99.86	M
rd5_at	1.28	0.93	1.33	1.50	73.83	98.66	A
ret_12_7	1.36	0.96	1.09	1.60	17.03	99.99	M
sale_me	1.39	1.02	1.21	1.57	11.83	98.06	H
at_me	1.41	1.04	1.24	1.60	11.44	99.92	H
dolvol_126d	1.44	0.93	1.18	1.71	10.69	99.40	M
rd_sale	1.47	0.92	1.21	1.77	62.01	59.47	A
gp_atl1	1.48	1.00	1.29	1.73	14.80	98.37	A
op_atl1	1.50	1.03	1.27	1.74	14.75	56.36	A
ivol_capm_252d	1.52	1.42	1.40	1.58	16.21	93.30	M
gp_at	1.55	0.98	1.24	1.85	11.81	64.53	A
at_turnover	1.62	0.94	1.16	1.99	12.74	99.59	A
op_at	1.65	1.10	1.36	1.92	11.74	90.06	A
cop_at	1.68	1.14	1.74	1.93	19.91	69.19	A
ebit_sale	1.70	1.23	1.37	1.95	13.11	97.46	A
opex_at	1.72	1.12	1.29	2.04	11.81	79.22	A
qmj	1.75	1.58	1.75	1.85	35.60	97.39	M
qmj_prof	1.80	1.33	1.58	2.04	12.13	52.68	M
be_me	1.84	1.44	1.83	2.01	14.21	83.90	H
corr_1260d	1.85	1.47	1.75	2.08	34.74	94.25	M
cop_atl1	1.85	1.13	1.72	2.21	20.47	73.07	A
ebit_bev	1.91	1.45	1.78	2.11	15.49	79.15	A
ope_bel1	1.92	1.47	1.82	2.11	28.49	96.43	A
prc	1.97	1.25	2.28	2.21	0.48	99.19	M
ope_be	1.99	1.37	1.71	2.26	25.25	94.91	A
debt_me	2.00	1.55	1.68	2.23	11.70	98.70	H
ni_be	2.03	1.32	1.52	2.41	14.29	88.88	M
sale_bev	2.03	1.45	1.84	2.30	15.43	98.19	A

Continued on next page.

Table B2: Missingness, accuracy and model confidence per characteristic.

netdebt_me	2.07	1.48	1.58	2.40	11.70	67.70	H
ret_12_1	2.08	1.58	2.07	2.31	17.10	99.96	M
mispricing_perf	2.11	1.57	1.76	2.39	5.20	90.69	M
div12m_me	2.12	1.78	1.47	2.36	7.69	95.73	H
ami_126d	2.13	1.21	1.62	2.69	16.35	93.35	M
ivol_capm_21d	2.17	2.50	1.89	2.07	15.31	93.65	M
rd_me	2.20	1.51	1.98	2.56	61.26	90.13	H
zero_trades_252d	2.21	0.90	1.41	2.95	12.68	77.18	M
betabab_1260d	2.21	1.95	1.94	2.43	35.24	83.44	M
nncoa_gr1a	2.29	1.80	2.25	2.50	23.42	85.25	A
qmj_safety	2.30	1.82	1.97	2.56	8.56	47.44	M
at_be	2.32	1.62	1.82	2.70	14.00	80.50	A
ivol_ff3_21d	2.34	2.47	2.17	2.31	15.31	92.88	M
rvol_21d	2.36	2.24	2.12	2.47	15.31	80.59	M
bev_mev	2.37	1.93	2.41	2.53	16.07	77.38	H
ret_9_1	2.47	1.80	2.34	2.79	15.17	99.83	M
o_score	2.50	2.12	2.45	2.67	24.04	86.26	A
rmax5_21d	2.54	2.20	2.25	2.78	15.32	90.67	M
intrinsic_value	2.60	2.31	2.85	2.67	35.16	91.89	H
niq_at	2.63	2.05	2.32	2.99	29.13	99.70	A
ivol_hxz4_21d	2.65	2.83	2.56	2.57	24.50	90.33	M
noa_gr1a	2.66	2.11	2.43	2.94	24.38	79.12	A
ni_me	2.68	1.66	1.91	3.25	11.60	73.75	H
capx_gr3	2.70	2.38	2.75	2.87	35.12	99.36	A
ebitda_mev	2.71	1.92	2.40	3.10	13.56	68.84	H
ncoa_gr1a	2.77	2.62	2.91	2.80	21.98	77.72	A
ocfq_saleq_std	2.77	2.14	2.46	3.30	47.18	98.94	A
z_score	2.82	2.24	2.92	3.05	25.61	89.57	A
mispricing_mgmt	2.90	2.15	2.45	3.31	15.12	78.44	M
capex_abn	2.93	2.45	3.21	3.12	36.68	97.46	A
eqnpo_me	2.94	2.12	3.36	3.25	28.12	97.81	H
niq_be	2.95	2.09	2.51	3.46	31.38	97.91	A
ret_6_1	2.96	2.18	2.80	3.34	13.18	92.44	M
ocf_at	2.96	1.45	1.83	3.82	13.31	97.59	A
oaccruals_at	3.01	1.86	2.43	3.67	19.83	95.51	A
sale_emp_gr1	3.02	2.46	3.06	3.27	25.91	98.95	A
emp_gr1	3.03	2.79	3.19	3.12	29.45	92.86	A
capx_gr2	3.08	2.76	3.21	3.21	29.70	98.18	A
ocf_me	3.08	1.52	1.92	3.96	13.49	89.27	H
at_gr1	3.11	2.44	2.93	3.43	14.42	84.86	A
sale_gr3	3.17	2.71	3.13	3.39	26.84	94.85	A
eqnpo_12m	3.17	1.95	2.14	3.84	9.51	99.55	H
ret_3_1	3.35	2.54	3.10	3.76	11.13	91.59	M
noa_at	3.36	2.11	2.64	4.01	23.92	80.26	A
eq_dur	3.42	2.50	3.01	3.89	24.23	89.26	A
qmj_growth	3.46	3.12	3.47	3.63	35.60	82.10	M
tangibility	3.48	2.31	2.95	4.03	21.64	77.50	A
lti_gr1a	3.49	2.85	3.21	3.83	21.55	99.79	A
zero_trades_126d	3.53	1.01	1.97	4.94	10.69	99.33	M

Continued on next page.

Table B2: Missingness, accuracy and model confidence per characteristic.

fnl_gr1a	3.54	2.98	3.43	3.81	14.69	98.57	A
eqnetis_at	3.56	2.56	3.50	4.04	27.74	80.60	H
inv_gr1a	3.61	3.19	4.03	3.72	17.09	96.96	A
aliq_at	3.63	2.12	3.64	4.27	22.73	85.54	A
kz_index	3.67	3.25	3.75	3.83	24.04	68.33	A
sale_gr1	3.69	2.92	3.27	4.10	16.66	74.75	A
eqpo_me	3.76	3.08	4.81	3.90	31.77	89.60	H
taccruals_at	3.80	2.16	2.72	4.78	20.36	93.34	A
chcsho_12m	3.87	2.44	2.62	4.66	8.57	100.00	H
cowc_gr1a	3.91	3.35	3.74	4.18	23.51	91.93	A
niq_be_chg1	3.97	3.40	3.66	4.34	37.99	95.08	A
ni_ivol	3.98	3.61	3.56	4.27	36.38	88.86	A
coa_gr1a	4.04	4.04	4.16	4.02	22.95	79.82	A
aliq_mat	4.07	2.76	4.03	4.62	28.06	97.05	M
inv_gr1	4.14	3.89	4.29	4.20	32.18	72.40	A
nfna_gr1a	4.25	2.50	4.10	5.04	14.69	98.47	A
be_gr1a	4.27	3.65	3.83	4.61	18.81	83.94	A
turnover_126d	4.33	0.92	2.21	6.24	10.69	98.68	M
niq_at_chg1	4.35	3.87	4.09	4.66	35.07	89.56	A
rmax5_rvol_21d	4.36	3.83	3.83	4.75	19.88	85.68	M
fcf_me	4.44	2.27	2.99	5.68	18.90	73.35	H
turnover_var_126d	4.51	3.81	4.14	4.89	10.69	86.93	M
oaccruals_ni	4.53	2.61	3.44	5.63	19.87	95.29	A
dolvol_var_126d	4.55	3.70	4.19	5.00	10.69	88.35	M
sti_gr1a	4.63	4.57	4.18	4.76	31.68	99.83	A
taccruals_ni	4.69	3.03	3.35	5.73	20.42	72.45	A
beta_60m	5.01	4.91	3.23	5.40	26.26	99.89	M
netis_at	5.01	3.54	4.51	5.80	27.75	94.65	H
cash_at	5.06	3.11	3.88	6.08	12.44	75.18	A
capx_gr1	5.08	4.43	5.06	5.39	24.37	95.77	A
dsale_dinv	5.13	4.54	5.05	5.39	35.99	93.78	A
ppeinv_gr1a	5.18	4.95	5.30	5.25	24.09	82.07	A
lnoa_gr1a	5.34	4.58	6.52	5.50	25.93	64.71	A
col_gr1a	5.39	4.55	5.19	5.78	21.63	97.16	A
zero_trades_21d	5.56	3.22	3.97	6.89	9.22	58.61	M
rmax1_21d	5.80	4.85	4.67	6.51	15.32	75.93	M
ret_1_0	5.98	4.78	5.26	6.65	9.72	94.15	M
ocf_at_chg1	6.19	4.33	4.99	7.26	17.12	98.77	A
ret_60_12	6.29	5.92	6.65	6.41	41.62	81.40	M
debt_gr3	6.39	5.72	6.42	6.68	33.92	72.94	A
ncol_gr1a	6.83	6.29	7.14	7.00	22.50	97.99	A
dbnetis_at	6.88	4.91	6.00	7.86	12.42	72.85	H
saleq_su	6.90	6.44	7.24	7.07	35.95	96.84	A
resff3_12_1	7.20	6.86	6.65	7.45	19.23	97.74	M
saleq_gr1	7.22	5.58	6.18	8.19	24.49	88.39	A
dsale_dsga	7.33	5.93	6.47	8.15	34.26	71.73	A
dgp_dsale	7.40	5.81	7.06	8.16	25.08	98.42	A
niq_su	7.45	6.34	7.19	8.11	34.84	78.17	A
tax_gr1a	7.63	7.45	7.13	7.79	15.30	54.93	A

Continued on next page.

Table B2: Missingness, accuracy and model confidence per characteristic.

pi_nix	7.70	6.84	7.28	8.08	33.81	81.43	A
earnings_variability	7.87	7.83	8.59	7.74	37.51	94.01	A
prc_highprc_252d	8.00	6.57	8.09	8.67	16.23	88.54	M
iskew_capm_21d	8.31	8.73	7.83	8.20	15.32	53.11	M
rskew_21d	8.72	7.96	7.84	9.28	15.32	81.25	M
resff3_6_1	9.27	8.30	8.59	9.82	19.32	83.13	M
f_score	10.20	8.81	9.02	11.02	29.08	70.79	A
iskew_ff3_21d	10.20	9.82	10.17	10.40	15.31	41.45	M
betadown_252d	10.35	9.10	9.28	11.19	17.45	70.80	M
ni_ar1	10.72	10.89	11.16	10.55	36.38	83.38	A
ni_inc8q	11.63	11.08	11.17	12.05	39.71	99.98	A
dsale_drec	11.77	11.97	12.02	11.63	24.50	91.19	A
iskew_hxz4_21d	12.03	11.51	12.13	12.27	24.50	27.51	M
bidaskhl_21d	12.15	12.62	12.33	11.89	13.86	99.97	M
beta_dimson_21d	24.70	23.30	23.23	25.70	15.31	73.74	M
coskew_21d	29.16	27.65	28.60	30.01	15.31	31.10	M
Average	4.36	3.63	4.03	4.75	21.61	85.89	

Internet Appendix

(not for publication)

Recovering Missing Firm Characteristics with Attention-based Machine Learning

Table of Contents:

Appendix IA1. Model Setup in Detail

An Illustrative Example Consider a simple example to understand how we leverage information from observed firm characteristics to recover those that are missing. Figure IA1.1 shows the actual quintiles for the Fama and French (2015) characteristics for Apple in January of 2012. Assume that we wish to reconstruct Apple's quintile for the book-to-market ratio "B2M". We first mask it by inserting a "0" as a special class capturing characteristics masked for reconstruction. We then run this masked input through the model, which produces a probabilistic mapping between Apple's B2M and the other four characteristics. Assume for this example that knowing about Apple's market capitalization and growth in total assets is most informative about recovering the book-to-market ratio. The model consequently learns to place a higher weight on these characteristics (45% on "Size" and 35% on "Inv"). In contrast, market-based information, such as Apple's beta is less important for this task (weight of 5%). Using this mapping of how informative a certain characteristic is to reconstruct B2M, the model then produces a probability distribution across the five quintiles for B2M. If it places the highest weight on the first quintile (in this example, 85%) we have successfully reconstructed Apple's book-to-market ratio using only information about Apple's other characteristics measured at time t. In the full model, we also incorporate information about how Apple's characteristics have evolved through time.

IA1.1. Building Blocks

Before we discuss the model architecture in detail, we introduce the two central building blocks used in our model: attention and gated skip connections. For a description of the attention mechanism and its merits see Section 2.2 in the main text.

Gated Skip Connections Gated skip connections control the flow of information in our model by dynamically adjusting the impact that each layer of (non)linear processing has. In a standard fully-connected network, each input is fed through each processing layer. There is no way to skip further processing for simpler, while retaining a high level of processing for the most complex inputs. Instead, with skip connections, the model learns the optimal degree of processing per input from the data itself. Specifically, we let

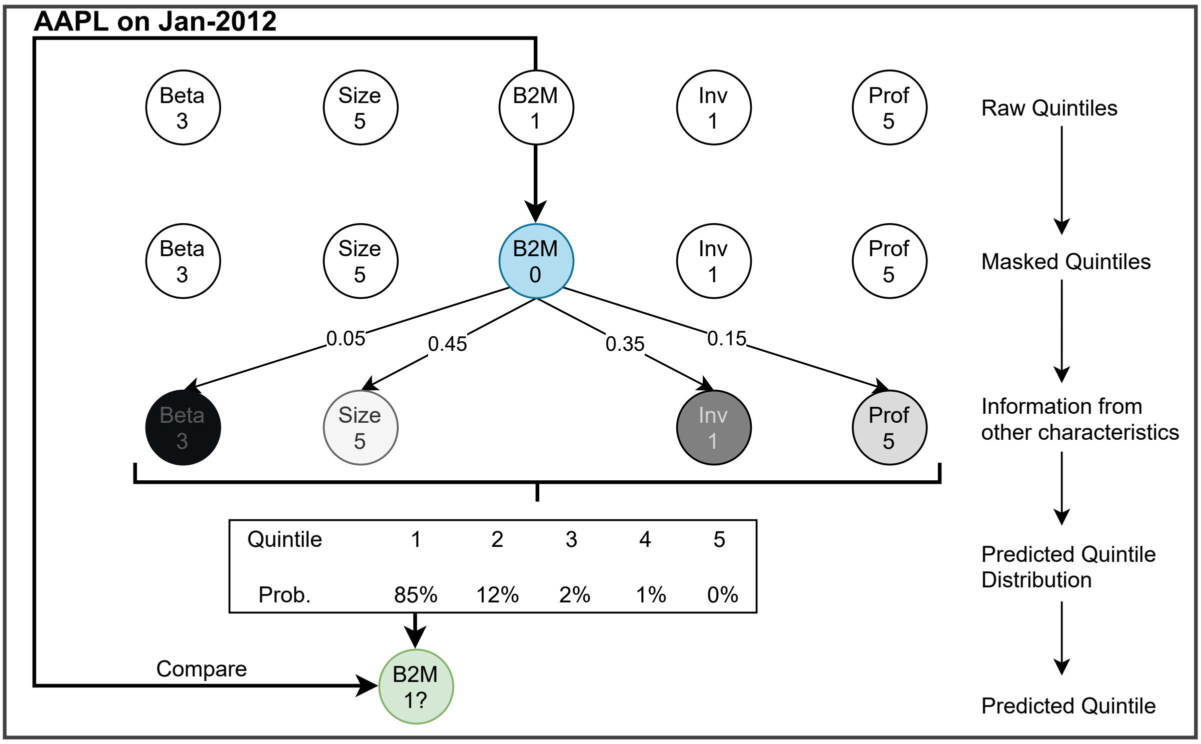

Fig. IA1.1. Exemplary Workflow to Recover Firm Characteristics

The figure shows an example for how our model manages to leverage the information of other firm characteristics to reconstruct a target characteristic, in this case Apple's (ticker AAPL) book-to-market ratio. We first set the characteristic to be reconstructed to a special "masked" token (0), and subsequently ask the model to find an optimally-weighted representation of other firm characteristics to come up with a predicted distribution over possible quintiles for Apple's book-to-market ratio. We then compare the most likely quintile with the actual value, and update the model's parameters through gradient descent, which allows the model to incrementally learn about how to extract information from available characteristics. What is missing from this stylized example is that we also incorporate the historic evolution of firm characteristics in the actual model.

the model decide how much of each additional processing step to skip through weights ω:

$$\boldsymbol{\omega}(\mathbf{x}) = \frac{1}{1 + e^{-\text{Linear}(\mathbf{x})}}, \tag{IA1}$$

where $\text{Linear}(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{b}$ denotes a linear transformation of $\mathbf{x}$. The output $\mathbf{y}$ of a given processing block is then a weighted-average between raw input $\mathbf{x}$ and the processed input $f(\mathbf{x})$:

$$\mathbf{y} = \boldsymbol{\omega}(\mathbf{x}) \cdot f(\mathbf{x}) + [1 - \boldsymbol{\omega}(\mathbf{x})] \cdot x \tag{IA2}$$

Skip connections have been used to improve the performance in many areas, most notably in image processing, spawning the infamous *ResNet* (He, Zhang, Ren, and Sun, 2015). They not only allow for deeper models that generalize well to unseen data but potentially also speed up the estimation. The particular choice of weighted skip connection used for our model follows the "Highway Networks" by Srivastava, Greff, and Schmidhuber (2015).

IA1.2. Model Setup

Figure IA1.2 schematically shows how the model learns to extract information from the cross-section of firms, their characteristics and their historical evolution. The model architecture consists of four main processing units shown in Figure IA1.2, which are further detailed below: *feature embeddings* create a high-dimensional representation of the percentiles of each input characteristic and push dissimilar firms along that characteristic away from each other. The *temporal attention network (TAN)* extracts an optimally-weighted average of the temporal evolution of firm characteristics, and *feature attention networks (FAN)* create a mapping between missing and available characteristics of a given firm. In the last step, we run these extracted connections through a *multi-layer perceptron (MLP)*, which estimates a probability distribution over the percentiles of each characteristic we wish to recover.

Feature Embeddings The financial literature highlights stark differences between stocks with small and large market equity in many aspects (Fama and French, 1993). Recovering Apple's book-to-market ratio using other characteristics may very well lead to a different functional form than recovering Rite Aid Corp.'s book-to-market ratio. To accommodate these differences across the range of a characteristic and to improve the learning capacity of the model, we feed each input characteristic through its own embedding. This is common in machine learning to deal with complex datasets (Huang et al., 2020; Somepalli et al., 2021; Lim et al., 2021; Gorishniy et al., 2021). An embedding is a learned lookup table that represents the percentiles of a target characteristic in a D-dimensional space. Percentiles that are closer in vector space are expected to behave similarly. For example, the model may learn that small stocks should receive different processing from large stocks, by pushing these stocks away form one another in vector space. We choose an internal embedding size of $D = 64$, such that each of the 100 (+1 missing; +1 masked) classes per characteristic is represented by a 64-dimensional vector. Embeddings have the added benefit of increasing the model's internal learning capacity, by adding a fourth "embedding" dimension, leaving the characteristics dimension

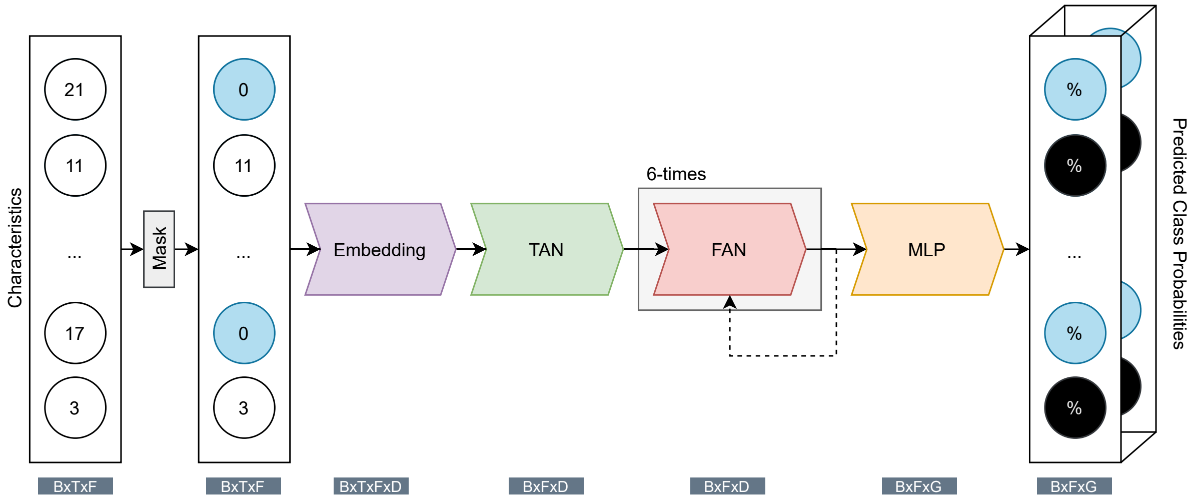

Fig. IA1.2. Model Setup

The figure schematically shows how the model extracts information from the cross-section of firm characteristics, as well as their historical evolution to predict the percentiles of characteristics masked for reconstruction (by the token “0”). We first randomly mask a fixed percentage (20%) of input characteristics for reconstruction, feed the characteristics through embeddings, a temporal attention network (TAN) extracting information about the characteristics’ historical evolution, and multiple feature attention networks (FAN), which extract information from other available characteristics. The last step comprises a multi-layer perceptron (MLP), which generates an informed probability distribution of the true percentile. We then compare how close the model’s predicted percentiles are to the ones actually observed before masking them.

untouched, which increases the model’s interpretability.

Temporal Attention Network To extract temporal patterns across characteristics, we use a simplified version of the temporal attention mechanism put forth by Lim et al. (2021). We feed the input from the embedding layer through an initial *long-short-term memory (LSTM)* network (see Figure IA1.3. The computation of the attention matrix is permutation invariant. It therefore disregards the timing of when information was received. To allow the model to understand that information from four years ago may be less important than the same information obtained last month, we need to add a time-positional encoding to the input. As in Lim et al. (2021), the LSTM serves this purpose. LSTMs have been successfully used by Chen and Zimmermann (2020) to extract macroeconomic states from a large data set of macroeconomic indicators. The effective lookback ability of LSTMs is limited, however, a drawback that temporal attention solves by explicitly attending to past information, without relying on gating mechanisms.

The time-encoded data is fed through the temporal *IMHA* unit with eight attention

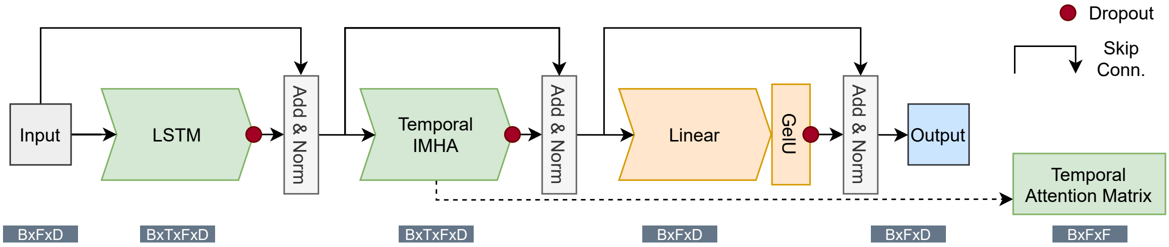

Fig. IA1.3. Temporal Attention Network

The figure shows the setup of the temporal attention network (TAN). We first feed the input through a long-short-term memory network to add positional awareness. We then employ temporal interpretable multi-head attention (IMHA), which extracts a matrix which optimally weights historical time steps. The last step applies a linear fully-connected layer with nonlinear GELU activation function. Each step is skipable in part or in full, through skip connections. We also apply dropout multiple times during training, which increases the stability of the model during inference.

heads, which extracts a weighted importance of past time step in the form of a temporal attention matrix. We follow this up by a simple linear layer with a GELU activation function. GELU has been introduced by Hendrycks and Gimpel (2016) and solves the issue of vanishing gradients occasionally encountered by the standard ReLU.

Feature Attention Network After we have extracted an optimally-weighted temporal representation of the input embeddings, we feed this intermediate data through six FANs. This number follows the original Transformer study by Vaswani et al. (2017). Each FAN creates a feature attention matrix, which tells us which characteristics the model uses to reconstruct a given missing input. The use of multiple consecutive FANs helps the model cover not only simple reconstructions, but also those that require more processing.

We feed the output of the feature *IMHA* with eight attention heads through a linear layer followed by a GelU, and allow for dynamic complexity control through skip-connections.

Multi-Layer Perceptron The last processing unit in our model is a standard MLP. MLPs combine a number of linear layers of varying sizes with activation functions. We use a total of two linear layers, the first of which is followed by a GelU activation function. The last layer takes the internal representation of the input data and creates a $(B \times F \times G)$-dimensional tensor, where G denotes the number of classes. We apply a Softmax function to the last dimension to obtain a probability distribution $\mathbf{p}$ across a characteristic's percentiles for all firm-month observations in the batch of size B. We

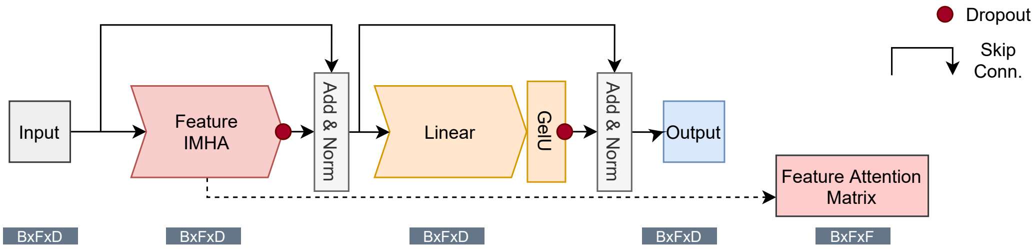

Fig. IA1.4. Feature Attention Network

The figure shows the setup of the feature attention network (FAN). We feed the input through a feature interpretable multi-head attention network (IMHA), followed by a linear fully-connected layer with nonlinear GELU activation function. Each step is skipable in part or in full, through skip connections. We also apply dropout multiple times during training, which increases the stability of the model during inference.

then regard the most probable percentile as the predicted class and compare it with the true (unmasked) percentile.

Appendix IA2. Hyperparameters

The following Table IA2.1 shows the model's hyperparameters, their search ranges and the optimal values using a hyperparameter search with 64 trials and the Bayesian optimization scheme outlined in Cowen-Rivers et al. (2020).

Table IA2.1: Hyperparameters for the Models Considered.

The table shows the hyperparameters and the boundaries from which they are randomly drawn to optimize them for each model considered. Optimal hyperparameter values are shown in **bold**.

	Model to fill missing firm characteristics	
Batch size	2,400	
Training months	180	
Validation months	60	
Testing months	402	
min lr	0.00001	
max lr	0.005	
Weight decay	0.001	
AdamW β_1	0.9	Follows Liu et al. (2019)
AdamW β_2	0.98	Follows Liu et al. (2019)
AdamW eps	$1e{-}6$	Follows Liu et al. (2019)
γ	2	FocalLoss parameter
Mask pct.	20%	Char. to randomly mask for reconstruction
F	151	Number of characteristics
T	60	Number of lookback timesteps
$N^{\text{embedding}}$	64	Internal model size
N^{IMHA}	8 per step	Number of attention heads
FAN steps	6	Number of consecutive FANs
FAN Normalization	$\in [\mathbf{Soft}, \text{Ent}, \text{Sparse}]\text{Max}$	
FAN Dropout	$\in [\mathbf{0.0}, 0.1, 0.3]$	
FAN Linear Dropout	$\in [0.0, \mathbf{0.1}, 0.3]$	
TAN steps	6	Number of consecutive TANs
TAN Normalization	$\in [\text{Soft}, \mathbf{Ent}, \text{Sparse}]\text{Max}$	
TAN Dropout	$\in [\mathbf{0.0}, 0.1, 0.3]$	
TAN Linear Dropout	$\in [0.0, 0.1, \mathbf{0.3}]$	
MLP dropout	$\in [\mathbf{0.0}, 0.1, 0.3]$	

Appendix IA3. Changes in Factor Premia – Value-weighted Returns

The following Table IA3.1 reports changes in factor portfolio returns after the inclusion of firms with previously missing values using market capitalization-weighted returns, following Jensen et al. (2021). Characteristics are sorted by the change in the factor premium ΔHmL. We also provide the premium before (HmL^{Pre}) and after (HmL^{Post}) imputation. Column "Not sig." equals "Y" whenever the factor's premium was significant before inclusion of missing observations, but is not significant thereafter. This happens on 6 occasions. The *total number* of significant factors is fairly constant at 98 before and 95 after imputation.

Table IA3.1: Change in Factor Premia.

	HmL^{Pre}		HmL^{Post}		ΔHmL		Not sig.
f_score	0.27	***	0.12	***	−0.15	***	
fcf_me	0.13	***	0.06	***	−0.07	***	
ivol_capm_252d	0.19	***	0.13	***	−0.06	***	
cowc_gr1a	0.10	***	0.04	***	−0.06	***	
ivol_capm_21d	0.18	***	0.12	***	−0.06	***	
cop_at	0.18	***	0.12	***	−0.06	***	
noa_gr1a	0.13	***	0.07	***	−0.05	***	
cop_atl1	0.16	***	0.11	***	−0.05	***	
ret_12_1	0.21	***	0.16	***	−0.05	***	
ope_bel1	0.10	***	0.06	***	−0.04	***	
coa_gr1a	0.08	***	0.04	***	−0.04	***	
rd_me	0.14	***	0.10	***	−0.04	***	
ret_60_12	0.03	-	−0.01	-	−0.04	***	
ppeinv_gr1a	0.11	***	0.08	***	−0.04	***	
ocf_at	0.14	***	0.11	***	−0.04	***	
ami_126d	0.04	-	0.00	-	−0.03	-	
ivol_ff3_21d	0.18	***	0.14	***	−0.03	***	
niq_at	0.11	***	0.08	***	−0.03	***	
capex_abn	0.04	***	0.01	-	−0.03	***	Y
oaccruals_ni	0.10	***	0.07	***	−0.03	***	
ncoa_gr1a	0.09	***	0.06	***	−0.03	***	
nncoa_gr1a	0.09	***	0.06	***	−0.03	***	
eq_dur	0.11	***	0.08	***	−0.03	***	
capx_gr3	0.06	***	0.03	***	−0.03	**	
oaccruals_at	0.09	***	0.06	***	−0.03	***	
noa_at	0.11	***	0.09	***	−0.03	***	
inv_gr1	0.08	***	0.05	***	−0.03	***	
op_atl1	0.14	***	0.12	***	−0.03	***	
saleq_su	0.03	***	0.01	-	−0.02	***	Y
ebitda_mev	0.14	***	0.11	***	−0.02	***	

Continued on next page.

Table IA3.1: Change in Factor Premia.

debt_gr3	0.04	***	0.02	***	−0.02	***	
inv_gr1a	0.09	***	0.07	***	−0.02	**	
chcsho_12m	0.11	***	0.09	***	−0.02	***	
be_gr1a	0.06	***	0.03	*	−0.02	***	
ocf_me	0.12	***	0.10	***	−0.02	***	
emp_gr1	0.07	***	0.05	***	−0.02	***	
resff3_6_1	0.07	***	0.05	***	−0.02	***	
netis_at	0.11	***	0.09	***	−0.02	***	
zero_trades_21d	0.02	-	0.00	-	−0.02	-	
nfna_gr1a	0.07	***	0.05	***	−0.02	***	
earnings_variability	0.01	-	−0.01	-	−0.02	***	
capx_gr2	0.07	***	0.05	***	−0.02	**	
aliq_mat	0.07	***	0.05	***	−0.02	***	
capx_gr1	0.07	***	0.05	***	−0.02	**	
qmj_growth	0.04	***	0.02	***	−0.02	-	
dolvol_var_126d	0.01	-	−0.01	-	−0.02	**	
prc_highprc_252d	0.10	*	0.08	*	−0.02	***	
mispricing_mgmt	0.14	***	0.12	***	−0.02	***	
niq_be_chg1	0.06	***	0.04	***	−0.02	**	
pi_nix	0.02	*	0.00	-	−0.02	*	Y
eqnpo_12m	0.10	***	0.08	***	−0.02	-	
lnoa_gr1a	0.09	***	0.07	***	−0.01	**	
ope_be	0.11	***	0.09	***	−0.01	-	
kz_index	0.03	-	0.01	-	−0.01	**	
dsale_drec	0.01	-	−0.00	-	−0.01	**	
mispricing_perf	0.18	***	0.17	***	−0.01	**	
qmj_safety	0.05	**	0.04	*	−0.01	**	
taccruals_at	0.03	**	0.02	-	−0.01	**	Y
ni_me	0.12	***	0.11	***	−0.01	-	
dsale_dinv	0.05	***	0.03	***	−0.01	-	
ni_ar1	0.00	-	−0.01	-	−0.01	**	
sale_gr1	0.04	**	0.03	*	−0.01	*	
at_gr1	0.08	***	0.07	***	−0.01	*	
turnover_var_126d	0.01	-	−0.01	-	−0.01	-	
eqpo_me	0.06	**	0.04	**	−0.01	-	
ret_6_1	0.16	***	0.15	***	−0.01	***	
rmax5_rvol_21d	0.09	***	0.08	***	−0.01	*	
at_be	0.04	-	0.03	-	−0.01	**	
ivol_hxz4_21d	0.18	***	0.17	***	−0.01	**	
iskew_ff3_21d	0.01	-	−0.00	-	−0.01	***	
ret_1_0	0.09	***	0.08	***	−0.01	**	
ocf_at_chg1	0.05	***	0.04	***	−0.01	-	
sale_gr3	0.03	*	0.02	-	−0.01	-	Y
bev_mev	0.04	-	0.03	-	−0.01	-	
sale_bev	0.10	***	0.09	***	−0.01	**	
taccruals_ni	0.04	***	0.03	***	−0.01	-	
fnl_gr1a	0.08	***	0.07	***	−0.01	**	
tax_gr1a	0.01	-	−0.00	-	−0.01	-	
rd_sale	0.03	-	0.03	-	−0.01	-	

Continued on next page.

Table IA3.1: Change in Factor Premia.

netdebt_me	0.03	-	0.02	-	−0.01	**	
zero_trades_252d	0.03	-	0.02	-	−0.01	-	
niq_at_chg1	0.05	***	0.05	***	−0.01	-	
turnover_126d	0.02	-	0.01	-	−0.01	-	
eqnetis_at	0.12	***	0.12	***	−0.01	-	
op_at	0.17	***	0.17	***	−0.01	**	
rmax5_21d	0.19	***	0.18	***	−0.01	-	
sale_emp_gr1	0.00	-	−0.00	-	−0.01	-	
sale_me	0.09	***	0.09	***	−0.00	-	
cash_at	0.01	-	0.01	-	−0.00	*	
gp_atl1	0.05	***	0.05	***	−0.00	-	
be_me	0.04	-	0.04	-	−0.00	-	
ebit_sale	0.10	***	0.10	***	−0.00	-	
dgp_dsale	0.05	***	0.04	***	−0.00	-	
resff3_12_1	0.13	***	0.13	***	−0.00	***	
coskew_21d	0.01	-	0.01	-	−0.00	-	
iskew_hxz4_21d	0.01	-	0.01	-	−0.00	-	
ebit_bev	0.09	***	0.09	***	−0.00	-	
ret_3_1	0.12	***	0.12	***	−0.00	-	
beta_dimson_21d	0.01	-	0.00	-	−0.00	-	
z_score	0.03	-	0.03	-	−0.00	-	
lti_gr1a	0.02	*	0.02	-	−0.00	-	Y
div12m_me	0.03	-	0.02	-	−0.00	-	
ret_9_1	0.18	***	0.18	***	−0.00	-	
gp_at	0.09	***	0.09	***	−0.00	-	
niq_su	0.05	***	0.05	***	−0.00	-	
ni_ivol	0.05	-	0.05	-	−0.00	-	
at_me	0.03	-	0.03	-	−0.00	-	
dbnetis_at	0.06	***	0.06	***	−0.00	-	
col_gr1a	0.00	-	0.00	-	−0.00	-	
tangibility	0.02	-	0.02	-	−0.00	-	
bidaskhl_21d	0.11	**	0.10	**	−0.00	-	
aliq_at	0.05	**	0.05	**	−0.00	-	
ret_12_7	0.16	***	0.16	***	−0.00	-	
beta_60m	0.00	-	0.00	-	−0.00	-	
sti_gr1a	0.00	-	0.00	-	−0.00	-	
age	0.01	-	0.01	-	−0.00	-	
ncol_gr1a	0.01	-	0.01	-	0.00	-	
debt_me	0.01	-	0.01	-	0.00	-	
dsale_dsga	0.00	-	0.00	-	0.00	-	
betadown_252d	0.02	-	0.02	-	0.00	-	
rvol_21d	0.19	***	0.19	***	0.00	-	
saleq_gr1	0.01	-	0.01	-	0.00	-	
eqnpo_me	0.14	***	0.14	***	0.00	-	
opex_at	0.03	*	0.03	**	0.00	-	
at_turnover	0.05	***	0.05	***	0.00	-	
rd5_at	0.03	-	0.03	-	0.00	-	
ni_be	0.08	***	0.08	***	0.00	-	
prc	0.06	-	0.06	-	0.00	-	

Continued on next page.

Table IA3.1: Change in Factor Premia.

rskew_21d	0.02	**	0.02	***	0.00	-	
qmj_prof	0.13	***	0.13	***	0.00	-	
rmax1_21d	0.14	***	0.14	***	0.00	-	
ni_inc8q	0.01	-	0.02	-	0.01	-	
niq_be	0.11	***	0.12	***	0.01	-	
dolvol_126d	0.03	-	0.04	-	0.01	-	
o_score	0.07	**	0.08	***	0.01	-	
iskew_capm_21d	0.01	-	0.02	**	0.01	***	
intrinsic_value	0.01	-	0.02	-	0.01	-	
ocfq_saleq_std	0.05	*	0.06	**	0.01	-	
zero_trades_126d	0.03	-	0.04	*	0.01	-	
betabab_1260d	0.04	-	0.06	**	0.02	-	
market_equity	0.08	***	0.10	**	0.02	-	
qmj	0.10	***	0.12	***	0.02	***	
corr_1260d	0.00	-	0.02	-	0.02	-	
Σ		98		95		71	6

Appendix IA4. Changes in Factor Premia – Equally-weighted Returns

The following Table IA4.1 reports changes in factor portfolio returns after the inclusion of firms with previously missing values using equally-weighted returns. Characteristics are sorted by the change in the factor premium ΔHmL. We also provide the premium before (HmL$^{\text{Pre}}$) and after (HmL$^{\text{Post}}$) imputation. Column “Not sig.” equals “Y” whenever the factor’s premium was significant before inclusion of missing observations, but is not significant thereafter. This happens on 6 occasions. We note, however, that the *total number* of significant factors is fairly constant at 113 before and 114 after imputation.

Table IA4.1: Change in Factor Premia.

	HmL$^{\text{Pre}}$		HmL$^{\text{Post}}$		ΔHmL		Not sig.
f_score	0.24	***	0.16	***	−0.08	**	
prc_highprc_252d	0.04	-	−0.03	-	−0.07	***	
noa_gr1a	0.18	***	0.11	***	−0.06	***	
rmax5_rvol_21d	0.16	***	0.11	***	−0.06	***	
debt_gr3	0.10	***	0.04	***	−0.05	***	
ppeinv_gr1a	0.17	***	0.12	***	−0.05	***	
ivol_capm_21d	0.11	**	0.06	-	−0.05	***	Y
ebitda_mev	0.10	***	0.05	*	−0.05	***	
noa_at	0.17	***	0.13	***	−0.05	***	
cop_at	0.18	***	0.13	***	−0.05	***	
cop_atl1	0.16	***	0.11	***	−0.04	***	
inv_gr1	0.12	***	0.08	***	−0.04	***	
saleq_su	0.07	***	0.03	**	−0.04	***	
emp_gr1	0.12	***	0.08	***	−0.04	***	
cowc_gr1a	0.10	***	0.05	***	−0.04	***	
ncoa_gr1a	0.14	***	0.11	***	−0.04	***	
ivol_ff3_21d	0.11	**	0.07	*	−0.04	***	
oaccruals_at	0.11	***	0.07	***	−0.04	***	
zero_trades_252d	0.14	***	0.10	***	−0.04	***	
mispricing_mgmt	0.20	***	0.16	***	−0.04	***	
coa_gr1a	0.12	***	0.08	***	−0.04	***	
resff3_6_1	0.09	***	0.05	***	−0.04	***	
oaccruals_ni	0.12	***	0.08	***	−0.03	***	
eqnpo_12m	0.12	***	0.09	***	−0.03	***	
nncoa_gr1a	0.15	***	0.11	***	−0.03	***	
at_gr1	0.15	***	0.12	***	−0.03	***	
be_gr1a	0.12	***	0.08	***	−0.03	***	
nfna_gr1a	0.10	***	0.07	***	−0.03	***	
ocf_at_chg1	0.05	***	0.02	*	−0.03	***	
aliq_at	0.11	***	0.08	***	−0.03	***	
pi_nix	0.01	-	−0.02	**	−0.03	***	
ret_60_12	0.09	***	0.06	**	−0.03	**	

Continued on next page.

Table IA4.1: Change in Factor Premia.

aliq_mat	0.13	***	0.10	***	−0.03	***	
fcf_me	0.11	***	0.08	***	−0.03	*	
ret_6_1	0.10	***	0.07	**	−0.02	***	
ret_12_1	0.14	***	0.12	***	−0.02	***	
ivol_capm_252d	0.06	-	0.03	-	−0.02	**	
turnover_126d	0.10	***	0.08	**	−0.02	***	
sale_gr1	0.10	***	0.07	***	−0.02	***	
betadown_252d	0.05	-	0.03	-	−0.02	***	
corr_1260d	0.04	-	0.01	-	−0.02	***	
netis_at	0.14	***	0.12	***	−0.02	***	
qmj_growth	0.06	***	0.04	***	−0.02	*	
lnoa_gr1a	0.14	***	0.12	***	−0.02	***	
ivol_hxz4_21d	0.11	***	0.09	**	−0.02	***	
tax_gr1a	0.02	-	−0.00	-	−0.02	***	
capex_abn	0.06	***	0.05	***	−0.02	**	
chcsho_12m	0.13	***	0.11	***	−0.02	***	
capx_gr3	0.11	***	0.09	***	−0.02	*	
inv_gr1a	0.13	***	0.12	***	−0.02	***	
capx_gr1	0.10	***	0.09	***	−0.02	**	
ret_3_1	0.06	**	0.04	-	−0.02	***	Y
sale_emp_gr1	0.02	***	0.01	-	−0.02	***	Y
fnl_gr1a	0.12	***	0.10	***	−0.02	***	
sale_gr3	0.08	***	0.07	***	−0.02	**	
niq_be_chg1	0.10	***	0.09	***	−0.02	**	
op_at	0.12	***	0.11	***	−0.02	***	
be_me	0.18	***	0.16	***	−0.01	**	
dsale_drec	0.01	-	−0.00	-	−0.01	*	
intrinsic_value	0.04	-	0.03	-	−0.01	-	
taccruals_at	0.04	**	0.03	*	−0.01	***	
ocf_at	0.12	***	0.11	***	−0.01	**	
ret_9_1	0.12	***	0.11	***	−0.01	***	
netdebt_me	0.08	***	0.07	***	−0.01	***	
rd5_at	0.06	*	0.05	-	−0.01	-	Y
z_score	0.05	*	0.03	**	−0.01	-	
ret_1_0	0.29	***	0.28	***	−0.01	***	
capx_gr2	0.11	***	0.10	***	−0.01	-	
zero_trades_126d	0.13	***	0.12	***	−0.01	-	
niq_at_chg1	0.08	***	0.07	***	−0.01	*	
niq_at	0.09	***	0.08	***	−0.01	-	
taccruals_ni	0.04	***	0.03	**	−0.01	-	
ebit_sale	0.05	-	0.04	-	−0.01	*	
dgp_dsale	0.06	***	0.05	***	−0.01	**	
sale_bev	0.10	***	0.09	***	−0.01	*	
niq_su	0.10	***	0.09	***	−0.01	**	
rd_me	0.24	***	0.23	***	−0.01	-	
at_me	0.11	***	0.10	***	−0.01	-	
dsale_dsga	0.01	-	−0.00	-	−0.01	-	
ebit_bev	0.06	*	0.05	-	−0.01	*	Y
ami_126d	0.08	**	0.07	***	−0.01	-	

Continued on next page.

Table IA4.1: Change in Factor Premia.

cash_at	0.04	*	0.04	-	−0.01	**	Y
kz_index	0.04	**	0.03	*	−0.01	-	
bev_mev	0.15	***	0.15	***	−0.01	-	
op_atl1	0.09	***	0.08	***	−0.01	-	
eqnetis_at	0.15	***	0.14	***	−0.01	*	
rskew_21d	0.06	***	0.05	***	−0.01	-	
resff3_12_1	0.15	***	0.15	***	−0.01	***	
bidaskhl_21d	0.02	-	0.02	-	−0.01	-	
sale_me	0.15	***	0.15	***	−0.00	-	
tangibility	0.09	***	0.08	***	−0.00	-	
ni_ivol	0.00	-	−0.00	-	−0.00	-	
ocf_me	0.10	***	0.10	***	−0.00	-	
turnover_var_126d	0.05	**	0.04	**	−0.00	-	
dbnetis_at	0.10	***	0.09	***	−0.00	**	
rvol_21d	0.11	***	0.11	***	−0.00	-	
col_gr1a	0.05	***	0.05	***	−0.00	-	
ret_12_7	0.11	***	0.11	***	−0.00	*	
rd_sale	0.01	-	0.01	-	−0.00	-	
beta_dimson_21d	0.03	-	0.03	-	−0.00	-	
dsale_dinv	0.05	***	0.05	***	−0.00	-	
ope_bel1	0.06	**	0.06	***	−0.00	-	
lti_gr1a	0.04	***	0.04	***	−0.00	-	
qmj_safety	0.05	*	0.05	*	−0.00	-	
mispricing_perf	0.14	***	0.14	***	−0.00	-	
prc	0.07	*	0.07	*	−0.00	-	
age	0.01	-	0.01	-	0.00	-	
iskew_hxz4_21d	0.03	***	0.03	***	0.00	-	
ni_inc8q	0.01	-	0.01	-	0.00	-	
beta_60m	0.01	-	0.01	-	0.00	-	
ni_me	0.05	-	0.05	-	0.00	-	
eqpo_me	0.05	**	0.05	***	0.00	-	
market_equity	0.16	***	0.16	***	0.00	-	
saleq_gr1	0.01	-	0.02	-	0.00	-	
eqnpo_me	0.14	***	0.14	***	0.00	**	
ni_be	0.06	*	0.06	*	0.00	-	
rmax5_21d	0.16	***	0.16	***	0.00	-	
opex_at	0.03	-	0.03	-	0.00	-	
div12m_me	0.02	-	0.02	-	0.00	-	
niq_be	0.12	***	0.12	***	0.00	-	
debt_me	0.01	-	0.01	-	0.00	-	
ope_be	0.07	**	0.08	***	0.00	-	
dolvol_126d	0.12	***	0.13	***	0.00	**	
eq_dur	0.11	***	0.11	***	0.01	-	
gp_at	0.09	***	0.10	***	0.01	*	
gp_atl1	0.04	*	0.04	**	0.01	-	
sti_gr1a	0.03	**	0.04	***	0.01	-	
coskew_21d	0.00	-	0.01	-	0.01	-	
at_turnover	0.04	***	0.05	***	0.01	*	
at_be	0.01	-	0.02	-	0.01	-	

Continued on next page.

Table IA4.1: Change in Factor Premia.

iskew_capm_21d	0.05	***	0.06	***	0.01	-	
qmj_prof	0.12	***	0.13	***	0.01	***	
dolvol_var_126d	0.03	-	0.04	**	0.01	*	
zero_trades_21d	0.06	*	0.07	**	0.01	-	
ncol_gr1a	0.01	-	0.02	*	0.01	-	
betabab_1260d	0.05	-	0.07	**	0.01	-	
o_score	0.02	-	0.04	-	0.02	**	
ni_ar1	0.01	-	0.03	***	0.02	**	
iskew_ff3_21d	0.02	**	0.04	***	0.02	***	
earnings_variability	0.01	-	0.03	**	0.02	***	
rmax1_21d	0.15	***	0.17	***	0.02	**	
ocfq_saleq_std	0.04	-	0.06	**	0.03	**	
qmj	0.10	***	0.13	***	0.03	***	
Σ		113		114		89	6

Appendix IA5. Portfolio Migration

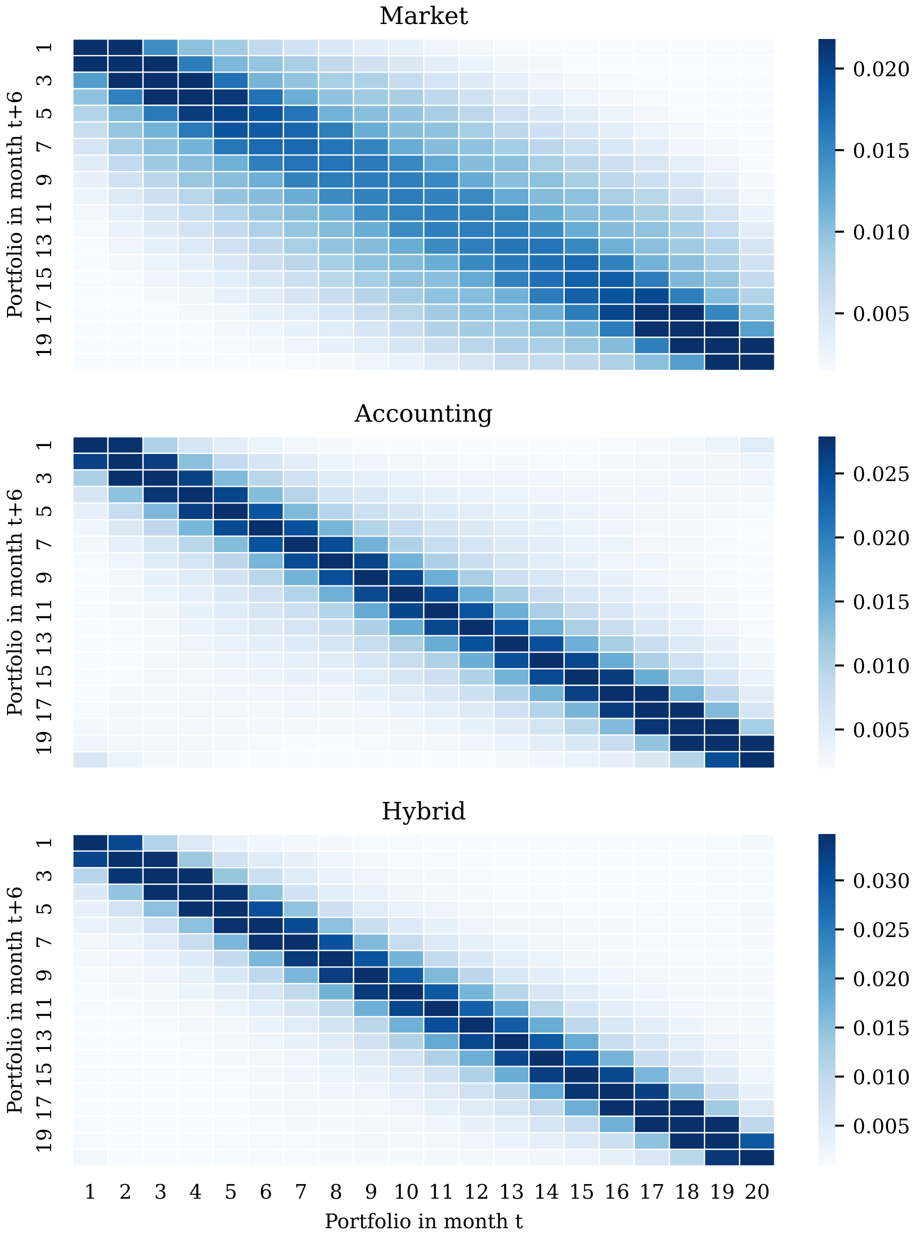

Fig. IA5.1. Portfolio Migration Heatmap

The figure shows how the portfolio allocation of characteristics of different types fluctuates from month t to $t + 6$, i.e., over half a year. Darker shades of blue indicate that a migration from the portfolio on the x-axis to the portfolio on the y-axis is more likely. We separately show the portfolio migration for accounting and market-based, as well as hybrid characteristics. Note that the portfolio migration for market-based characteristics is most dispersed.

References

Chen, A. Y., Zimmermann, T., 2020. Open source cross-sectional asset pricing. Critical Finance Review, Forthcoming .

Cowen-Rivers, A. I., Lyu, W., Tutunov, R., Wang, Z., Grosnit, A., Griffiths, R. R., Maraval, A. M., Jianye, H., Wang, J., Peters, J., et al., 2020. An empirical study of assumptions in bayesian optimisation. arXiv preprint arXiv:2012.03826 .

Fama, E., French, K., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics .

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of financial economics 116, 1–22.

Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A., 2021. Revisiting deep learning models for tabular data. arXiv preprint arXiv:2106.11959 .

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. arxiv 2015. arXiv preprint arXiv:1512.03385 .

Hendrycks, D., Gimpel, K., 2016. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 .

Huang, X., Khetan, A., Cvitkovic, M., Karnin, Z., 2020. Tabtransformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv:2012.06678 .

Lim, B., Arık, S. Ö., Loeff, N., Pfister, T., 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting .

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 .

Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., Goldstein, T., 2021. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. arXiv preprint arXiv:2106.01342 .

Srivastava, R. K., Greff, K., Schmidhuber, J., 2015. Highway networks. arXiv preprint arXiv:1505.00387 .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: *Advances in neural information processing systems*, pp. 5998–6008.