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

Heiner Beckmeyer[†] Timo Wiedemann[‡]

This version: April 24, 2022

Abstract

Firm characteristics are often missing. We set up an attention-based machine learning model borrowing ideas from state-of-the-art research in natural language processing to understand how characteristics relate to the cross-section of other – observed – firm characteristics and their historical evolution. Our model reconstructs firm characteristics with high accuracy and comfortably outperforms competing approaches. Revisiting the vast literature on risk factors in financial research reveals that disregarding the influence of missing observations likely overestimates the magnitude of factor premia. We also provide the recovered values for missing entries of firm characteristics for all characteristics for future research.

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

JEL classification: G10, G12, G14, C1, C55

[*] We want to thank Turan G. Bali, Andrea Barbon, Dmitriy Muravyev, Ilias Filippou, Yuri Tserlukevich, Florian Weigert, Julian Thimme, Christian Schlag, Simon Rottke, Johannes Kriebel, Jörn Debener and Nicole Branger for valuable feedback and discussions.

[†] 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

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. Figure 1 shows the evolution of missing values for a large set of 151 firm-level characteristics from the dataset provided by Jensen, Kelly, and Pedersen (2021). At the start of our sample in the 60s, an average of 58 characteristics – more than 38% – are missing per firm×month observation. While this number has declined considerably in the following decades, the average firm still misses 17 characteristics in the most recent decade, with many firms providing less information.

The study at hand is devoted to recovering missing firm characteristics, drawing on the informational content of the cross-section of other – observed – characteristics, as well as how a given firm's characteristics have evolved through time. We apply state-of-the-art advances from the field of natural language processing 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.

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 insight to the case of missing firm characteristics. 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 and their historical evolution, 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 set of 151 characteristics provided by Jensen et al.

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

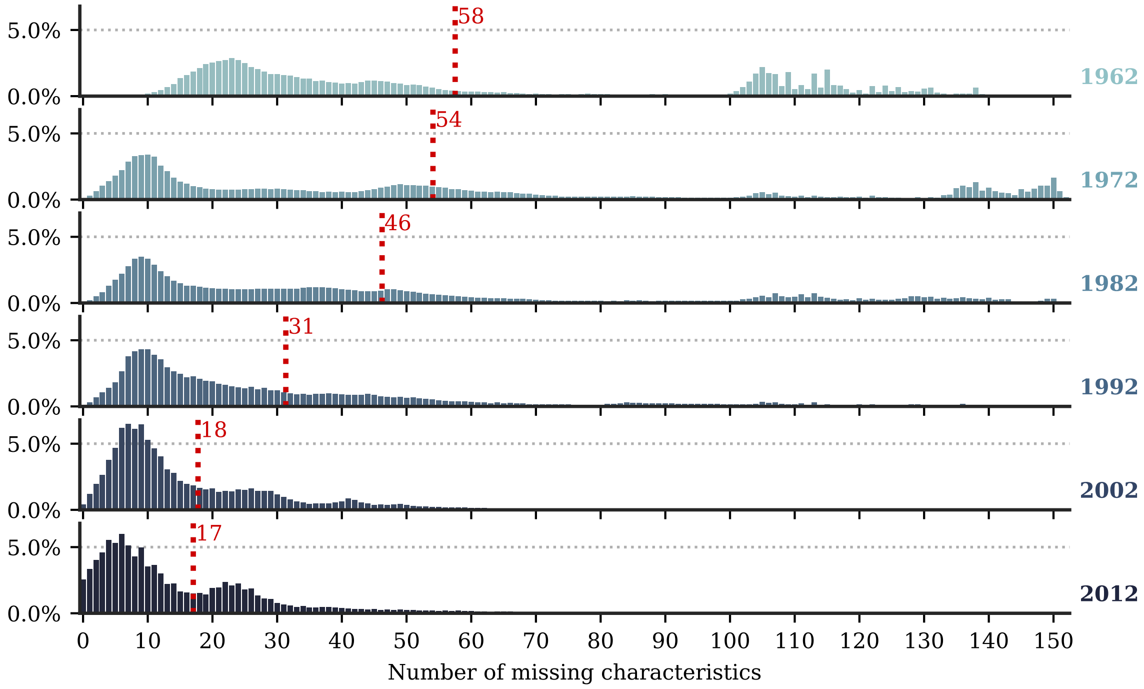

Fig. 1. Distribution of Missing Firm Characteristics over Time.

The figure shows the distribution of the number of missing firm characteristics per observation for each decade in our dataset. We use 151 firm level characteristics from the dataset provided by Jensen et al. (2021) with common filters applied (cf. Section 3.1). The dashed red lines indicate the mean number of missing characteristics per firm-month observation.

(2021) for the years of 1962 through 2020. To assure that our model learns the latent structure governing firm characteristics in an optimal environment, we train it using the most recent and most complete 15 years. Machine learning algorithms require a careful consideration of techniques that assure that the estimated model fit of the training sample carries over to unseen data. We employ a battery of regularization mechanisms and find little degradation of the model's accuracy in out-of-sample tests.

For a brief demonstration of our model's success in reconstructing firm characteristics, Figure 2 shows Apple's actual and the model's prediction of Apple's market capitalization over time. The reconstruction is highly accurate at all times and follows the actual distribution of the characteristic well.

Our main metric to assess the model's accuracy for a broader set of characteristics and firms is the *expected percentile deviation (EPD)*, which measures how many percentiles our prediction is off on average. We obtain perfect accuracy for about 32% and more

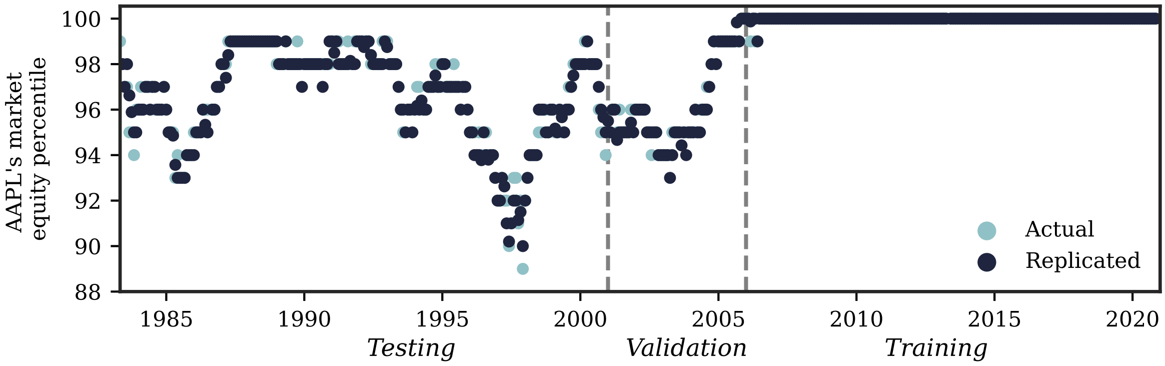

Fig. 2. Replication of Apple's (AAPL) Market Capitalization Percentile.

The figure shows the evolution of Apple's (ticker AAPL) actual market capitalization percentile over time, as well as our model's reconstruction of it. We find that our model manages to replicate the values to a high degree.

than 25% of cases in the training and testing samples, respectively. The expected percentile deviation amounts to 3.77 during the training and 4.51 during the testing sample. Separately considering accounting-, and market-based, as well as hybrid characteristics, we find that reconstructing the latter is easiest. The model has the hardest time reconstructing market-based characteristics. The EPD for these variables, however, is still fairly low, at or below five percentiles. We generally find that the model 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 `qmj`, among others. The characteristics that the model most struggles with are particularly those using daily information and seasonal returns. The model operates at the monthly frequency and is never fed intra-month information, which explains the former. The latter are missing for the majority of observations. Uncovering a suitable structure for seasonal returns is consequently hampered.

A consideration of the model's absolute accuracy is one thing. Investigating in how far it outperforms competing approaches potentially more fruitful. We compare our model's accuracy with a special case using only cross-sectional information. In contemporaneous research, Bryzgalova, Lettau, Lerner, and Pelger (2021) and Freyberger, Höppner, Neuhierl, and Weber (2021) propose using linear methods to impute missing values, which leverage information about other characteristics, but disregard their historical evolution.

We compare our model's accuracy with a nested case, which uses only this cross-sectional information. Given that we allow for nonlinearities, and most importantly interaction effects between input characteristics, this approach subsumes the method proposed in Bryzgalova et al. (2021). We also consider two simple approaches: the mean imputation advocated by Green, Hand, and Zhang (2017), as well as an estimator using the last available historic value for the target characteristics.

The results highlight the importance of incorporating historical information: not only is our model by far the most accurate, with a 2.3× decrease in the expected percentile deviation compared to the model using only cross-sectional information (4.25 vs. 9.86). We also find that simply imputing the last available value for a target characteristic performs comparable to the cross-sectional model, and even manages to outperform it for accounting characteristics. Accounting information fluctuates little over time. Using historical information thus leads to fairly accurate predictions. For hybrid characteristics, which incorporate information from both market and accounting sources, we again find that imputing the last value outperforms the cross-sectional model. This outperformance is especially severe in the out-of-sample testing data. While our model manages to produce highly accurate reconstructions with an EPD of just 2.19, the cross-sectional model's EPD stands at 8.87 – twice as high as the EPD of using the last value (4.42). Not considering temporal information degrades how well the model's ability to reconstruct firm characteristics carries over to unseen data.

We want to stress that our model is not a black box, but produces interpretable outputs.[2] The rigorous use of the attention mechanism allows us to understand the internals of the model to a high degree. Consistent with the idea that characteristics of a certain group should be impacted by other characteristics of said group, we find a high intra-group attention weight. At the same time, we report on the benefits of including firm characteristics of various kinds when estimating the latent structure that governs them. Reconstructing characteristics of each group also requires information from all other groups, with market-based characteristics, which rapidly adjust to changes in the informational landscape, being the most important. The reliance on information about the characteristic itself, in contrast, is fairly limited.

The model comparison has highlighted the importance of incorporating past informa-

[2] 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).

tion. Using temporal attention weights, we can show that it is mostly information about firm characteristics *within the last year* that is used in the model's predictions. This is in line with efficient and timely-updated information through financial markets and adequate accounting and reporting standards.

In an extensive simulation study, we highlight that our model setup manages to *simultaneously* uncover multiple processes governing the evolution of a firm's characteristics. In one unified framework, it accurately predicts masked entries for auto-regressive and cross-sectional characteristics, as well as characteristics which are a mixture of both. Furthermore, we can show that it accurately recovers the temporal information patterns of more complex autoregressive processes.

We find sensible distributions of recovered entries of firm characteristics that were previously missing. For example, the distribution of firms with missing book-to-market (`be_me`) is relatively balanced with slightly more firms being identified as "value" firms. Considering that `be_me` is often missing due to a negative book-value, this is inline with Brown, Lajbcygier, and Li (2008) who finds that stocks with a negative book-value are more similar to value than they are to growth stocks. Furthermore, the in-fill distribution for standard momentum (Carhart, 1997) and its first part (Novy-Marx, 2012) are comparable, showing adequate *internal consistency* of the model's recoveries. Other tests of internal consistency agree with this assessment. We also propose a new way to gauge the confidence with which our model recovers certain characteristics. A fully uninformed prediction will converge to an equal probability assigned to each percentile of the target characteristic. We use this insight and compute a similarity of the model's predictions to a uniform distribution using the *Jensen-Shannon Divergence*. In most cases our model is highly confident when recovering missing characteristics.

After establishing that a) our model is highly accurate in reconstructing masked characteristics, and b) confidently produces sensible distributions of missing firm characteristics, we investigate the impact of using the now-completed dataset for the application of high-minus-low factor returns. We see a clear trend that incorporating this additional information pushes existing estimates of factor premia towards zero. 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. We also provide the recovered percentiles and estimates for the raw firm characteristics using various interpolation methods for future research.[3]

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 simply dropped missing observations. This ad-hoc approach not only vastly reduces the sample size, but also results in biased inference if data points are not missing at random. It is straightforward to see that 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 prominent examples using this approach. Bali, Beckmeyer, Moerke, and Weigert (2021b) and Bali, Goyal, Huang, Jiang, and Wen (2021a) also use this approach when using a joint stock-option or stock-bond dataset, respectively. Afifi and Elashoff (1966) argues 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 to dealing with the issue of missing firm characteristics.

The paper most closely related to ours is the work by Bryzgalova et al. (2021). Using the model setup of Xiong and Pelger (2019), the authors propose the use of principal component analysis to estimate the latent factor structure in the characteristics space and impute missing values using the common components.[4] Hence, they leverage the information content of other observable characteristics to impute those that are missing. We deviate in many important aspects: first, we explicitly incorporate temporal information about the evolution of firm characteristics. In an extensive model comparison, we find this to be imperative for the model's success. Second, the model of Bryzgalova et al. (2021) only includes first-order effects, and thus disregards interaction effects between characteristics. Current asset pricing research stresses the importance of interactions between

[3] The data is accessible at `https://sites.google.com/view/beckmeyer/data-code`.

[4] At the time of writing this paper, their paper is not publicly available. We thus base this part on Appendix A.2 of Kaniel, Lin, Pelger, and Van Nieuwerburgh (2021), which outlines the procedure.

characteristics to explain stock returns (Gu et al., 2020; Chen and Zimmermann, 2020; Kozak, Nagel, and Santosh, 2020b). We too find that interactions between characteristics are important to recover missing values.

Another contemporaneous attempt at leveraging the informational content of missing firm characteristics is provided by Freyberger et al. (2021). The authors use moment conditions in a generalized method of moments framework to estimate missing characteristics based on a pre-selected set of 18 characteristics, which they require to be observable. They fill missing observations in a joint setup to explain stock returns. Implicitly, the approach thus requires that characteristics are relevant return predictors. Put differently, the recovered missing entries are not the true value of the characteristic, but rather the value that best helps explain the stock's return. Different from ours, their method disregards temporal information and remains linear. Furthermore, by keeping the required data filters to a minimum, we are able to work on a much larger set of firm characteristics, without strict assumptions about which characteristics ultimately drive the evolution of others.

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 famous natural language processing models and is now an integral part of Google's research 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. In an analogous fashion, we apply this idea to the task of predicting missing firm characteristics by leveraging the information content of other – observed – characteristics.

An Illustrative Example Consider a simple example to understand how we leverage information from observed firm characteristics to recover those that are missing. Figure 3 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.

Before we discuss the model architecture in detail, we introduce the two central building blocks used in our model: attention and gated skip connections.

Attention To recover missing firm characteristics through the characteristics that are available to us in an interpretable fashion, we rely on *attention* – a machine learning technique that allows the model to focus on the most important parts of the input data, while fading out the rest. The rigorous use of attention in machine learning as a standalone technique was proposed by Vaswani et al. (2017) and gave rise to "Transformers", which are by now the backbone of most state-of-the-art models in natural language and sequence processing. Attention computes how similar a tensor of search queries $\mathbf{Q}$ is to an internally-updated tensor of keys $\mathbf{K}$. Both $\mathbf{Q}$ and $\mathbf{K}$ are learned linear transformations of the same input x. This gives rise to the name "self-attention". Using the resulting attention 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 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, given a certain query. In economic terms, how important is Apple's market capitalization to recover Apple's book-to-market ratio? We can express attention

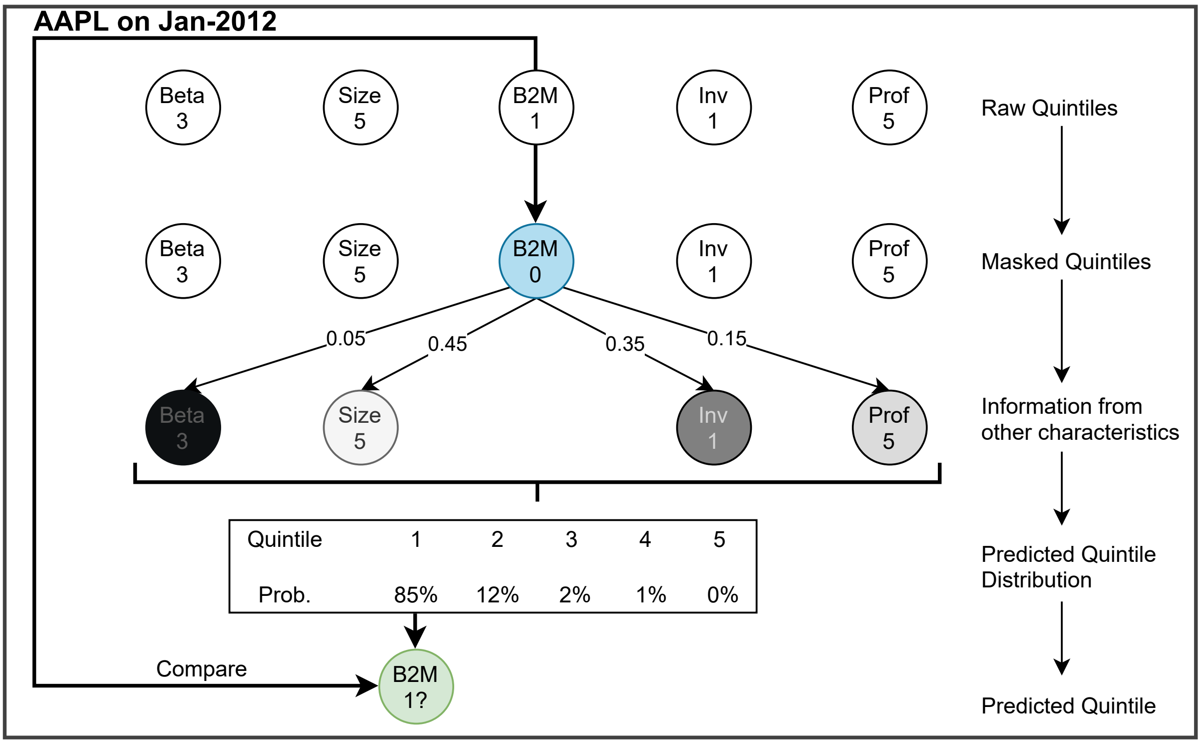

Fig. 3. Exemplary Workflow to Recover Firm Characteristics

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.

as,

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

where N^{A} denotes the number of units to attend to. In the temporal case, this is the number of lookback months, which we set to $T = 60$, covering the historical evolution of firm characteristics over the last five years. In the case of feature attention, N^{A} equals the number of features $F = 151$. 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 up to 1, with values between 0 and 1, thereby mapping from $\mathbb{R}^d$ to probability space

Δ^d.[5] We consider normalization 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. $p_j \geq 0 \ \forall \ j$). Martins and Astudillo (2016) introduce Sparsemax ($\alpha = 2$), which aggressively pushes small weights towards 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.[6]

To increase the learning capacity, multiple attention heads – each with its own attention matrix – are commonly employed. We opt for a total of $N^{\text{heads}} = 8$ temporal and feature attention heads per processing unit. We follow Lim et al. (2021) and use *interpretable multi-head attention (IMHA)* throughout this paper. It averages the attention matrices of each attention head before multiplying it with a single learned value matrix $\mathbf{V}$:

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

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

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. This 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 a target characteristic.

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

[6] Results are shown in Table 7.

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

$$\boldsymbol{\omega}(\mathbf{x}) = \frac{1}{1 + e^{-\text{Linear}(\mathbf{x})}}, \tag{6}$$

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 input $\mathbf{x}$ and processed input $f(\mathbf{x})$:

$$\mathbf{y} = \boldsymbol{\omega}(\mathbf{x}) \cdot f(\mathbf{x}) + [1 - \boldsymbol{\omega}(\mathbf{x})] \cdot x \tag{7}$$

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).

2.1. Model Architecture

To simplify the task of reconstructing missing firm characteristics, while at the same time retaining a high-degree of expressivity 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 discretize each characteristic into percentiles, yielding a total of 100 classes per characteristic.[7] This approach has the added benefit that it provides a natural way for dealing with missing data: we simply add an additional class to each characteristic, which captures the informativeness of a missing input. Features masked for reconstructed are instead denoted by class "0" in Figure 4, for a total of 102 classes. A common approach in the literature is to

[7] This approach is commonly employed in gradient-boosted trees, such as Microsoft's LightGBM (Ke, Meng, Finley, Wang, Chen, Ma, Ye, and Liu, 2017).

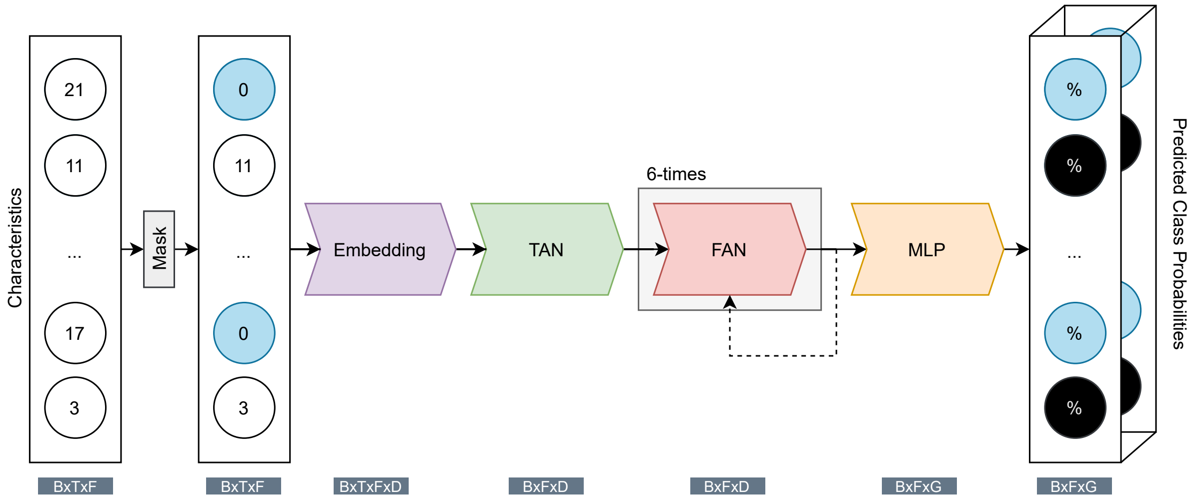

Fig. 4. Model Setup

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.

impute the cross-sectional mean of the characteristic for missing values. We instead treat missing values independently, allowing a missing entry to carry information about *why* it is missing. When training the model, we ask it to reconstruct a random subset of 20% of the available characteristics per firm, denoted by the blue circles. We then feed this masked input tensor through our model, which generates a probability distribution across the percentiles of each masked characteristic. Once we have properly trained the model in a controlled environment, we use it to recover missing firm characteristics for the whole sample.

Our model architecture consists of four main processing units shown in Figure 4: *feature embeddings* create a high-dimensional representation of each input characteristic and push dissimilar characteristic percentiles 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. For a detailed description of the model setup, we refer to Appendix A.

3. Estimation & Data

3.1. Dataset of Firm Characteristics

We analyze the dataset studied in Jensen et al. (2021), which contains monthly firm characteristics computed from prominent outlets, such as CRSP and Compustat, for all stocks trading on the NYSE, NASDAQ, and AMEX exchanges. The dataset contains 406 characteristics in total and can be downloaded at `https://github.com/bkelly-lab/ReplicationCrisis`. For our main analyses, we focus on the 153 characteristics identified by Jensen et al. (2021) and further exclude characteristics `seas_11_15na` and `seas_16_20na`, which are missing for more than 90% of observations in the sample. 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. We furthermore require that the return information has been obtained from CRSP.[8] Our model extracts information about the likely value of a missing characteristic from other, 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. Our data covers July-1962 through December-2020, for a total of 57 years, providing information about 151 characteristics on 25,118 unique firms, for a total of more than 3.2 million firm-month observations. We furthermore split the characteristics into three groups, conditional on their information source. We consider a group of market-based

[8] We have noticed that Compustat returns are highly volatile and that they fall vastly out of line when compared to the much larger CRSP-sample.

[9] We have also estimated a model without this filtering steps 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.

(57 characteristics), a group of accounting-based (75), and a hybrid group (21).

Observations in the data set are frequently missing. Our objective is to fill these gaps and thus provide a full picture of firm characteristics. Table 4 in the Appendix gives a full overview of how often each characteristic in the dataset is missing.

3.2. Training Routine

We set up a controlled environment to prime the model on recovering missing characteristics. During the training phase, we randomly mask 20% of the *available* input characteristics and ask the model to recover their percentiles by leveraging information about other firm characteristics. The model is flexible enough to understand the release cycle of accounting variables. We therefore mask not only the month-t value of these variables, but also the two preceding months ($t-1$ and $t-2$). The general approach is known as *masked language modeling* and follows Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, and Stoyanov (2019). We are the first to apply it to the case of recovering *missing* variables. At this self-supervised stage, we have full control knowing 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.

By 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 of Lin, Goyal, Girshick, He, and Dollár (2017):

$$\mathrm{FL}(\mathbf{p}) = \frac{1}{|\mathbf{p}|} \sum_{c} -(1-p_c)^{\gamma} \cdot \log(p_c), \tag{8}$$

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 set $\gamma = 2$ (with $\gamma = 0$ we obtain the standard cross-entropy loss) and optimize over the mean loss for all masked (and thus reconstructed) characteristics.

Figure 1 illustrates that the sample has grown considerably more complete in recent years, with an average of 17/151 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 will be able to extract. We thus flip the common train/validate/testing split and train the model using the most recent 15 years of data (2006-2020) and validate the resulting fit on a five-year validation sample (2001-2005).

Optimization and Regularization Neural networks are typically trained using stochastic gradient descent, which uses a subset of the entire dataset in each iteration to evaluate the gradient and consequently 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). This learning rate is then increased for the first 30% of training epochs, up to a high number (lr = 0.005). This ramp-up helps Adam find good estimates of the moments of the gradient, which aids the algorithm in making informed decisions for the epochs with the highest learning rates. Afterwards, we gradually decrease the learning rate once more up to the total number of training epochs 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 two days on four Nvidia Tesla A100 40GB GPUs. A list of the hyperparameters, their search ranges and optimal values is given in Table 7 in the appendix.

To assure that the latent structure found by the model carries over to unseen data, we employ a number of regularization techniques. **Explicit regularization** techniques include proper weight decay for Adam (Loshchilov and Hutter, 2017). Weight decay adds a certain fraction the L2 norm of the model parameters to the loss function, which forces the model to choose small and conservative parameters. To that, we add amsgrad (Tran et al., 2019), which 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

[10] Setting the maximum number of epochs to 750 or 1000 yield similar results.

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.

Some regularization is baked into the model **implicitly**: by randomly masking 20% of the *available* characteristics per firm-month observation, we reconstruct characteristics with a higher availability more often. At the same time, this overweights the number of reconstructed characteristics from firms with a higher overall availability of characteristics. By treating missing firm characteristics as an individual class, we force the model to weigh a characteristic's informativeness against its availability when using it as an input for reconstruction. Furthermore, in the extreme case in which all but one characteristic are missing for a firm, the predicted probability distribution across percentiles converges to a uniform distribution. We later use this insight to directly gauge how "confident" the model is when imputing missing firm characteristics.

4. Reconstructing Firm Characteristics

When fitting the model, we randomly mask a fixed percentage of non-missing characteristics, which we try to reconstruct with the remaining characteristics, and their historical evolution. In this controlled environment, we can directly infer the model's *internal validity*, by considering how well it reconstructs observed percentiles of masked characteristics. Gauging the model's *external* validity, in contrast, is much more involved. To do so, we assess its performance in reconstructing masked features in a holdout testing sample, which covers more than 30 years. These approaches tell us how well the model works in an absolute sense: how far off are our predictions of characteristics percentiles on average? We also assess in how far modelling both cross-sectional *and* temporal dependencies between characteristics helps in the reconstruction, by comparing our model with competing methods. We compare a simple time-imputation, the common mean- or median-imputation, as well as a special case of our model, which considers only the (nonlinear) information embedded in the cross-section of other firm characteristics, but disregards their temporal evolution.

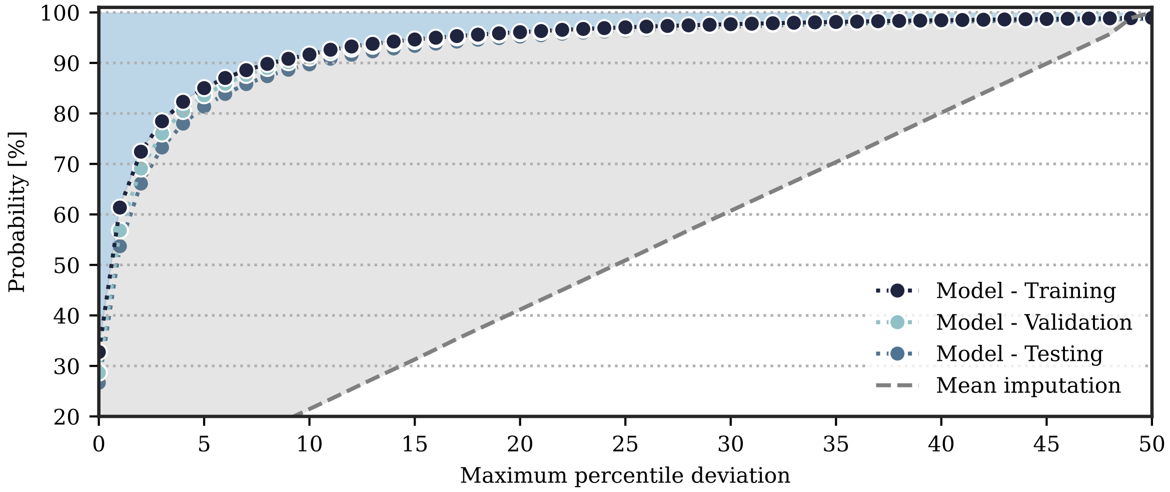

Fig. 5. Model Accuracy Curve.

The figure shows the cumulative distribution function of the model error $|\Delta|$ defined in Eq. (9) for the training, validation and testing sample. We also show the result for imputing the median for comparison. Section 4.2 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. (10).

4.1. Model Performance

The primary metric used 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. In a first step, 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,

$$p(|\Delta| = k) = p(|y - \hat{y}| = k) \qquad \text{where} \qquad |\Delta| \; \in \; [0, 1, \ldots, 99]. \tag{9}$$

Figure 5 shows the resulting cumulative distribution function $p(|\Delta| \leq k)$, for $k \in [0..50]$, for the training, validation and testing samples. For more than 32% 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 5 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 less than 9% of cases. We also find that our model's performance is highly consistent and carries over well to the validation and the true ouf-of-sample testing data.

The blue area above the lines denotes the cumulative classification error, or the *expected percentile deviation (EPD)*,[11]

$$\mathrm{E}\left[|\Delta|\right] = \sum_{k=0}^{99} p(|\Delta| = k) \cdot k. \tag{10}$$

A perfect model produces EPD = 0. Fully random predictions yield an EPD of $33.\bar{3}$. EPD is a neat way to summarize the information provided by the ROC curve in a single number. In the out-of-sample testing data, we achieve an EPD of 4.51. In other words, our model predictions are on average off by less than five percentiles. This compares well with the EPD of the mean imputation of roughly 25. For the validation (training) sample, the EPD amounts to 4.04 (3.77). [12] These numbers are of course averages across a wide range of 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. At the same time, however, Figure 1 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.

While we do find evidence of better performance in recent years in Figure 6, the EPD is fairly stable over time and still really low in the testing sample starting in 1973. For better interpretability, we have split the EPD numbers for market, accounting and

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

[12] We may also express the performance in a reconstruction R^2, for example advocated by Bryzgalova et al. (2021), which amounts to 88.9 % (88.3 %, 86.1 %) for the training (validation, testing) sample. A full comparison using this metric for the extended set of 151 characteristics (Bryzgalova et al. (2021) use 46) is provided in Table 6 in the appendix. The authors achieve a full in-sample R^2 of almost 75% on the subset of 46 characteristics considered.

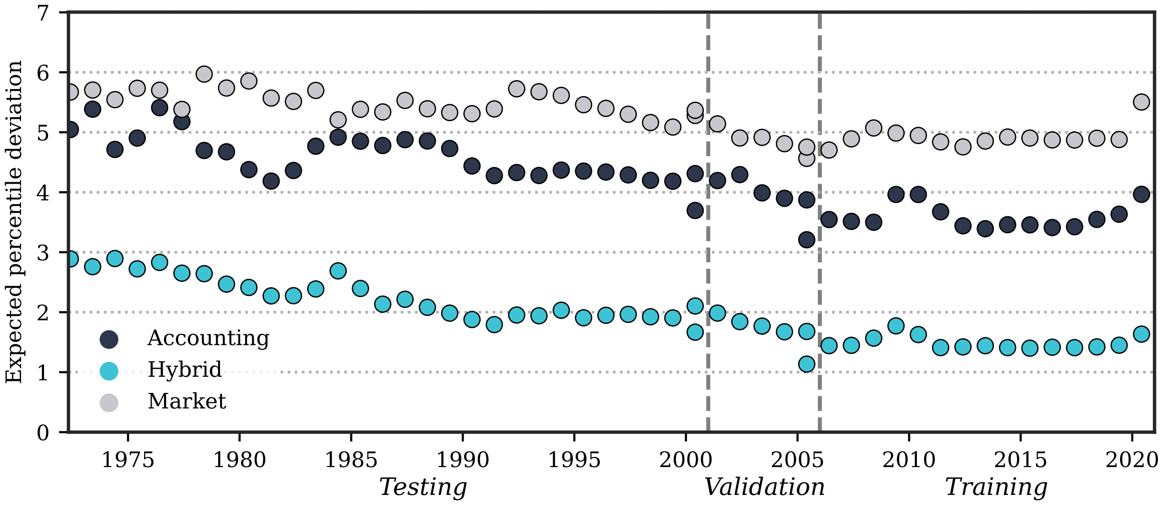

Fig. 6. Model Accuracy over Time.

The figure shows the model's accuracy as measured by the expected percentile deviation defined in Eq. (10) over time for accounting- and market-based, as well as hybrid characteristics.

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 3 in the early years of our sample, it trends downward to around 2 by the start of the validation sample and now stands at around 1.5 percentiles. The trends for the other groups are comparable. We generally find the best performance for hybrid characteristics, which have a comparatively high availability, and the worst for market characteristics, which generally vary the most. Interestingly, 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.

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 so, we sort each firm×month observation by the number of available characteristics and compare the reconstruction performance across different *missingness* buckets. Figure 7 shows the results.

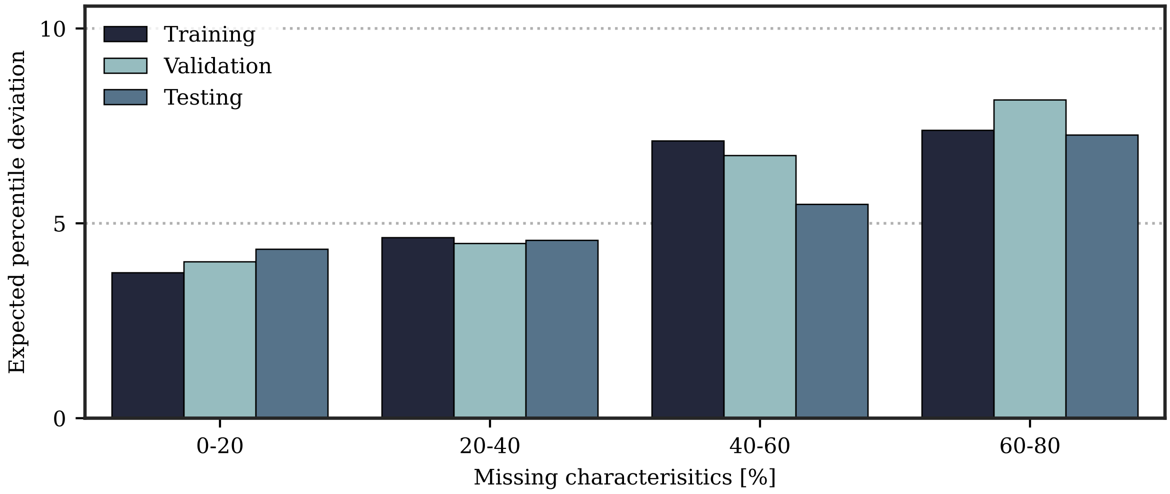

Fig. 7. Model Accuracy as a Function of Available Information.

The figure shows the model's accuracy measured by the expected percentile deviation defined in Eq. (10) 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.

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 8 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 4. 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 exploit. 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`

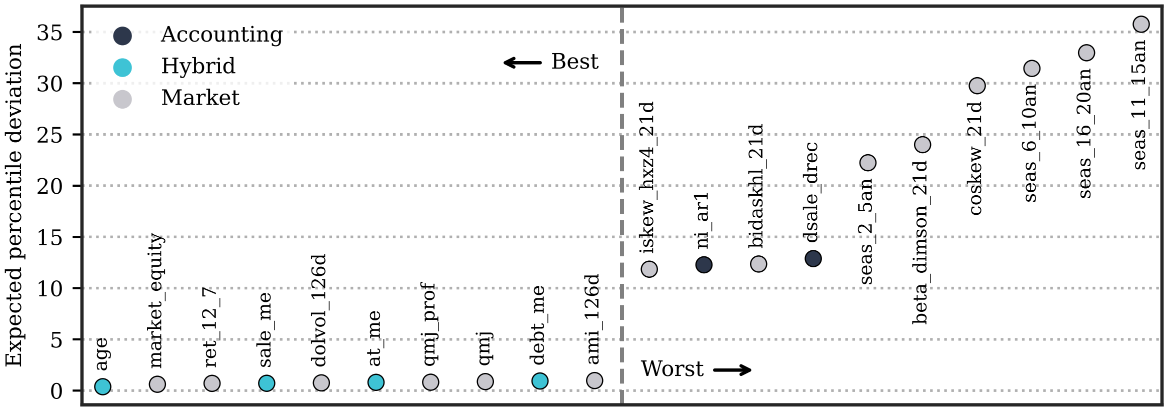

Fig. 8. Model Accuracy by Characteristic.

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. (10). A complete overview can be found in Table 4 in the Appendix. We further categorize characteristics into three groups, accounting-, hybrid- and market-based.

(Novy-Marx, 2012), and quality-minus-junk `qmj` (Asness, Frazzini, and Pedersen, 2019). We find two distinct clusters of characteristics among those *worst* reconstructed. Four 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, simply because we restrict it to consider only monthly information itself. The second cluster is that of seasonal returns (Heston and Sadka, 2010), again comprising four out of the ten characteristics. Seasonal returns are missing quite often and consequently hard to reconstruct. Internally, the model learns to focus on reconstructing the other characteristics and instead uses seasonal returns as "internal placeholders" for interim computations. Overall, the EPD for only six out of 151 characteristics is above 20 percentiles, or in other words, less than accurate to the quintile.

4.2. 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, Pruitt, and Su, 2019). This will not only bias the results if the missing-at-random assumption is violated but also becomes infeasible if the number of characteristics considered is large. In light of the multiple testing problem, this is however exactly what is needed, i.e. testing preferably

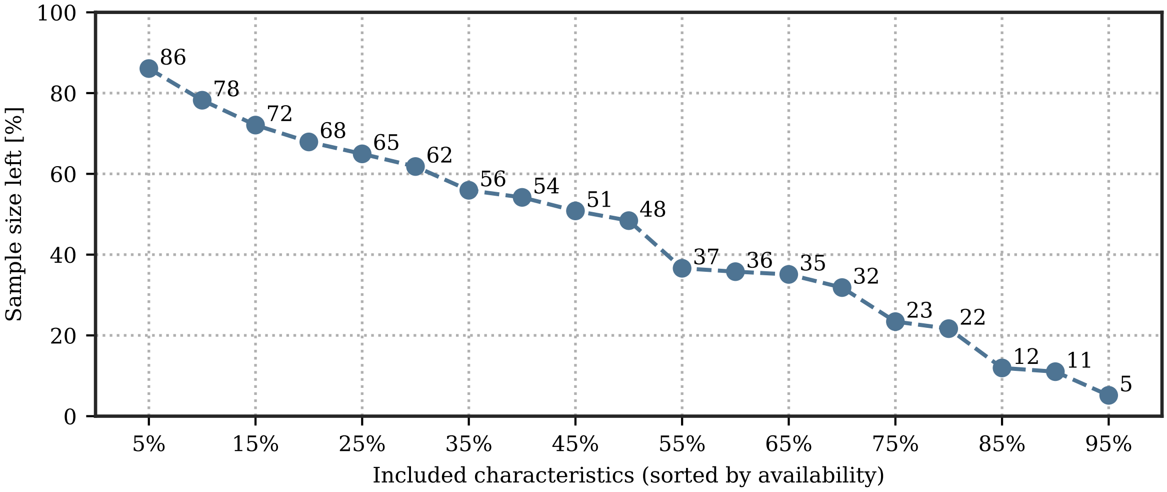

Fig. 9. Sample Size when Excluding Observations with Missing Characteristics.

The figure shows the remaining sample size when firm×month observations are dropped if *any* characteristic is missing, as a function of the number of characteristics included. We sort characteristics by their overall availability and calculate the theoretical sample size for several subsets of these 153 characteristics when all observations are dropped if any characteristic is missing. The full data set, obtained from Jensen et al. (2021), comprises 3,390,340 observations with 153 characteristics.

all possible characteristics simultaneously. To elaborate on the fact that this methods is infeasible when the number of characteristics is large, Figure 9 shows the remaining sample size for characteristic sub-sets, when all observations are excluded for which any characteristic is missing.

For example, including only the 10 % (16) characteristics that are most often available already decreases the sample size by 22 %. More importantly, if one were to use only those observations for which 95% of the 153 characteristics are available, only five percent of the sample would be left. This circumstance invariably leads to false inference, considering that the amount of information available depends on a firm's characteristic. Larger firms typically have stricter reporting requirements, requiring them to reveal more information and information of higher quality.

More apt methods are thus required. As competing benchmark models we thus consider three different approaches of varying complexity. The least complex model is the mean-imputation, which imputes the cross-sectional mean for missing characteristics, fully ignoring both temporal dependencies of the missing characteristic and the latent structure across different characteristics. The second approach exploits available time-series data by imputing missing characteristics by the last value if available and the cross-sectional median otherwise. The idea behind this approach is to harness the auto-correlation in

Table 1: Model Comparison – Accuracy by Imputation Method.

The table shows the imputation accuracy measured by the expected percentile deviation defined in Eq. (10). We differentiate our model's accuracy from that of a cross-sectional model, which disregards temporal information. We further consider imputing masked features with the last time-series observation or with the cross-sectional median as competing approaches. If the last value is not available, we set the value to the cross-sectional median. 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	**4.25**	**3.77**	**4.04**	**4.51**
X-Sectional model	9.86	8.14	8.92	10.83
Last	10.09	10.16	10.07	10.06
Mean imputation	25.08	25.08	25.13	25.07
		Accounting		
Full model	**4.22**	**3.59**	**4.05**	**4.53**
X-Sectional model	10.20	8.45	9.45	11.12
Last	9.52	9.45	9.46	9.56
Mean imputation	24.75	24.74	24.68	24.77
		Market		
Full model	**5.27**	**4.92**	**4.92**	**5.50**
X-Sectional model	10.46	9.13	9.49	11.28
Last	13.32	13.44	13.13	13.30
Mean imputation	25.01	25.03	25.05	24.99
		Hybrid		
Full model	**1.96**	**1.48**	**1.80**	**2.19**
X-Sectional model	7.38	4.58	5.70	8.87
Last	4.39	4.30	4.43	4.42
Mean imputation	26.32	26.36	26.87	26.20

many – especially accounting – characteristics.[13] The last competing approach is a nested version of our full model which is able to infer information from the cross-section of firm characteristics but disregards historical information altogether. This *X-sectional model* can be considered a non-linear extension of the model proposed by Bryzgalova et al. (2021). For this, we simply discard the temporal attention (TAN) block as shown in the model setup in Figure 4. Results for the model comparison are shown in Table 1.

By far the worst performing model is the mean imputation method which has an

[13] We have also considered using the historical 12-month average, but find that using the last available value leads to better results.

expected percentile deviation of about 25.[14] That is, the expected error induced by this method when characteristics are missing at random are 25 percentiles. Much better performance is achieved both by the cross-sectional model and the last imputation method. Interestingly, we find that the cross-sectional model only manages to beat the simple "Last" approach for the subset of market-based characteristics. There are strong temporal patterns in most characteristics. This finding is not new to the literature (Keloharju, Linnainmaa, and Nyberg, 2021) but it is worth highlighting that any model trying to impute missing values should consider the time-series dynamics of the characteristic itself, not just information about other characteristics. The model we propose incorporates information from both sources. This yields by far the best performance overall and across all characteristic groups. The expected percentile deviation for the entire sample amounts to 4.25 – a 2.3-fold improvement over the cross-sectional (9.86) and last-imputation (10.09) methods. The expected deviation amounts to roughly four percentiles for the average characteristic. Many common applications using firm characteristics instead rely on decile or quintile information – a level of accuracy we can accommodate readily.

4.3. Interpretability

We have so far highlighted that our model performs well in reconstructing percentiles of firm characteristics, and outperforms competing methods. We now wish to understand the model's internals: how does it come up with its predictions, which information about other characteristics is important at which stage, and how important is modeling the historical evolution of firm characteristics to the model's success?

Feature Importance In a first step, we consider how the reconstruction of a target characteristic is influenced by information about itself, information from other characteristics of the same group, and information from characteristics of the two other groups. To do so, we express the resulting *feature attention matrices* as directed weights, where each row indicates how much target characteristic c is influenced by all the others. Since the row-wise sum of the attention matrix is always 1, we can simply add up the values for

[14] The maximum absolute deviation possible amounts to $|\Delta| = 50$ in this case. With characteristics being uniformly distributed between $[1, 100]$ the mean will theoretically amount to $50/2 = 25$. As we randomly mask 20% of characteristics for reconstruction, we observe a slight empirical deviation to this theoretical EPD.

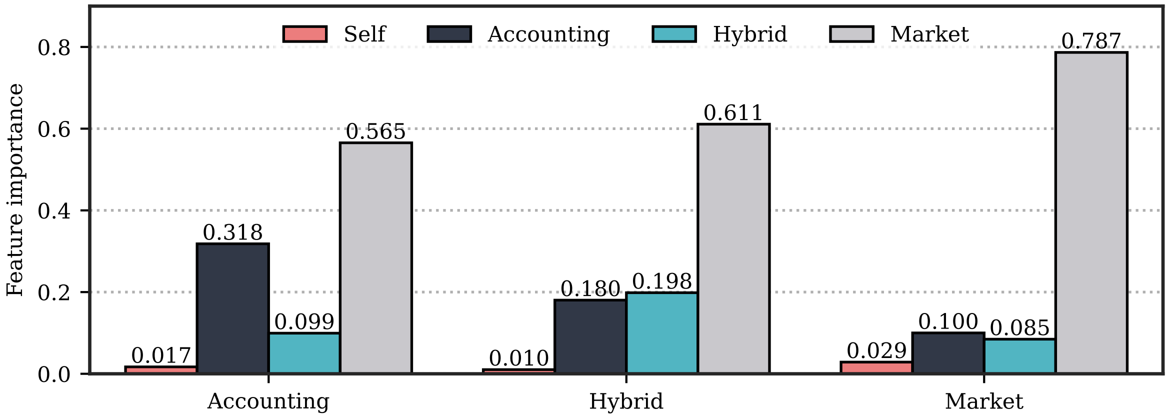

Fig. 10. Feature Importance by Characteristic Type.

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.

characteristics belonging to each of the three groups and thus obtain a breakdown of the group's importance.[15] We also provide the estimate for the "self-importance", indicating how important a characteristic's own evolution through time is for its reconstruction. If each characteristic was equally informative about all others, this self-importance would amount to $1/N = 1/151 \approx 0.006$.

Figure 10 provides this breakdown: starting with market variables, we find other market characteristics most important with an average weight of 78.7%, which is significantly larger than the weight for accounting or hybrid characteristics at 10.0% and 8.5%, respectively. The self-importance for market characteristics is highest at 2.9%, which is about 4.8-times as large as it would be in the case of equal information. The results are in line with our prior: market variables change most through time, such that contemporaneous information from other market variables provide valuable information. For hybrid characteristics, we find an outsized impact of other characteristics of the same group and market-based information. Note that hybrid is the smallest group, comprising only 21 characteristics. The average hybrid characteristic thus contributes 0.94% to the model estimation, the average market-based characteristic 1.07%. The most important

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

Table 2: 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	0.978	0.818	0.917	0.973	0.991	0.999	1.000	1.000
Year-2	0.013	0.000	0.000	0.000	0.005	0.017	0.049	0.101
Year-3	0.005	0.000	0.000	0.000	0.000	0.005	0.024	0.053
Year-4	0.002	0.000	0.000	0.000	0.000	0.001	0.014	0.034
Year-5	0.001	0.000	0.000	0.000	0.000	0.000	0.006	0.027

block is that of market-based characteristics, which captures the fast-changing informational landscape for hybrid characteristics. For accounting-based characteristics, we still find market-based characteristics to be most important, with an average total weight of 56.5%. The next-biggest block is that of accounting variables, at 31.8%, which is more than three times higher when compared to market characteristics. The characteristic itself has an influence of 1.7%. The results highlight that all three groups of characteristics rely on all other groups to make informed reconstructions, but that the fast-changing nature of market-based characteristics is most informative for the reconstruction of *all* characteristics.

Time Importance We explicitly account for the historical evolution of input characteristics in a flexible fashion, such that is may incorporate varying levels of temporal information for per target characteristic. Especially accounting-based characteristics may benefit from this inclusion, given that they seldom fluctuate heavily from quarter-to-quarter. As an 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 1 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 2 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 97.8%, with comparatively little variation over time. The first percentile still amounts to 81.8%. In contrast, the second year receives on average less than 2% of the total attention, with occasional spikes above 4.9% in just five percent of cases. 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 all necessary information, highlighting the efficiency of modern financial markets and financial reporting.

The model focuses primarily on most recent year, but the inclusion of this temporal information is vital for its performance. Section 4.2 has highlighted that disregarding how characteristics evolve through time severely impedes the model's performance.

5. Model Validation: A Simulation Study

We have shown so far that our model performs well in reconstructing masked firm characteristics and that it comfortably outperforms competing approaches. We have further shown that our model's performance is driven by an explicit inclusion of temporal *and* cross-characteristic information, while being fully agnostic about the underlying structure. We deliberately rely on the capability of our model to extract this structural information on its own. To showcase how well the proposed method learns about different types of (mixed) processes governing the evolution of firm characteristics, we set up a simulation study.

Mixture Processes The benefit of our model setup is that it allows to simultaneously model characteristics with different underlying processes. Some may rely more on temporal, others more on cross-characteristic information. At the same time, the inclusion of this large number of characteristics facilitates cross-learning effects, wherein one type of characteristic is also used in the reconstruction of characteristics of another type, see Figure 10. To see how our model manages to *simultaneously* deal with characteristics driven by multiple types of processes, we simulate three types of characteristics with different properties. The first set of characteristics c follows an AR(1) process:

$$c_{i,t}^{\mathrm{AR}} = \gamma_i \cdot c_{i,t-1}^{\mathrm{AR}} + \varepsilon_{i,t}, \tag{11}$$

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

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

$$c_{i,t}^{\mathrm{XS}} \sim \mathcal{N}(\mathbf{0}, \mathbf{cov}), \tag{12}$$

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:

$$\mathrm{ar}_{i,t} = \gamma_i \cdot \mathrm{ar}_{i,t-1} + \varepsilon'_{i,t} \tag{13}$$

$$c_{i,t}^{\mathrm{AR+XS}} = \omega^{\mathrm{AR}} \cdot \mathrm{ar}_{i,t} + \left(1 - \omega^{\mathrm{AR}}\right) \cdot \mathrm{xs}_{i,t}, \tag{14}$$

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, each. We again compare how well our model competes against the three competitors, imputing the mean, using the last value, or only considering cross-sectional information. We optimize the model over 100 epochs.

Table 3 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 by far produces the lowest EPD. Considering the subcategories of characteristics, we find that the performance of imputing the last value and of our model is virtually the same for characteristics following an AR(1) process. As we chose a high auto-correlation within the model setup, we explicitly aimed at giving the *last* imputation method a fair chance to beat our full model. Finding that our model has virtually the same performance shows that the time-attention mechanism is capable to extract the important time-series information from the data. In contrast, the X-Sectional model fails to uncover any meaningful structure within these variables, and

Table 3: Simulation: Model Comparison – Accuracy by Imputation Method.

The table shows the imputation accuracy measured by the expected percentile deviation defined in Eq. (10) for the simulation study outlined in Section 5. We differentiate our model's accuracy from that of a cross-sectional model, which disregards temporal information. We further consider imputing masked features with the last time-series observation or with the cross-sectional median as competing approaches. Results are shown for characteristics following in AR(1) process, a one-factor cross-sectional model, or a combination of the two. The best performing model is highlighted in bold for each case.

	Expected percentile deviation			
	Full	Training	Validation	Testing
		All		
Full model	**5.55**	**5.54**	**5.55**	**5.56**
X-Sectional model	23.25	23.25	23.25	23.26
Last	20.38	20.37	20.38	20.38
Mean imputation	24.99	24.99	24.99	24.99
		AR(1)		
Full model	6.70	6.69	6.70	6.71
X-Sectional model	40.71	40.71	40.71	40.71
Last	**6.55**	**6.55**	**6.55**	**6.55**
Mean imputation	25.00	25.00	25.00	25.00
		XS		
Full model	**3.03**	**3.02**	**3.03**	**3.03**
X-Sectional model	3.54	3.53	3.54	3.54
Last	33.47	33.46	33.48	33.47
Mean imputation	25.00	25.01	25.00	25.00
		AR(1) + XS		
Full model	**6.93**	**6.92**	**6.93**	**6.94**
X-Sectional model	25.51	25.50	25.51	25.52
Last	21.10	21.10	21.10	21.09
Mean imputation	24.97	24.97	24.97	24.97

even underperforms simply imputing the mean. This is to be expected, as the model has no information about a characteristic's evolution through time. Also considering a fourth set of characteristics governed by an AR(12) process leads to a similar conclusion, with our model beating all competitors by a large margin.

As anticipated, we find the exact opposite for cross-sectional characteristics. Our model performs best, with no meaningful difference to the X-sectional model. Last and mean imputation fail to uncover the structure in these variables. For the process combining both an autoregressive and a cross-sectional component, we again find that our model performs best by a large margin. Interestingly, both the Last imputation and the X-sectional model do not produce sensible estimates here performing similarly to the mean imputation method. Finding that all methods - except our model - exhibit a low performance even for a seemingly simple underlying process highlights the importance of bringing both temporal and cross-characteristic information into the fold. In summary, the simulation study stresses the flexibility of our model setup in uncovering the latent structure governing observable firm characteristics for a variety of underlying processes.

Temporal Patterns Generally - and most likely - the temporal evolution of firm characteristics is more intricate than being possibly modeled by an AR(1) process. The temporal attention mechanism enables our model to draw information from lagged values within the specified look-back window with no prior restrictions on where to draw information from. This makes the model flexible, in a way that it can theoretically accommodate all possible dependencies within the temporal evolution of the characteristics space. To test how well our model can fit the temporal evolution of characteristics we again return to a controlled simulation environment. We model a AR(12) process with exponentially decaying weights, i.e. the evolution of a target characteristic $c_{i,t}^{\mathrm{AR}(12)}$ follows

$$c_{i,t}^{\mathrm{AR}(12)} = \gamma_i \cdot \sum_{k=1}^{12} w_k \cdot c_{i,t-k}^{\mathrm{AR}(12)} + \varepsilon_{i,t} \tag{15}$$

with ϵ and γ as before. We choose exponentially-decaying weights w_k,

$$w_k = C \cdot e^{-0.25 \cdot k} \quad k \in [1, 12], \qquad \text{with } C \text{ s.t.} \qquad \sum_{k=1}^{12} w_k = 1, \tag{16}$$

and fit our model to the simulated data.

Figure 11 shows the learned temporal attention weights. 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 the 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 focus solely on the temporal evolution.

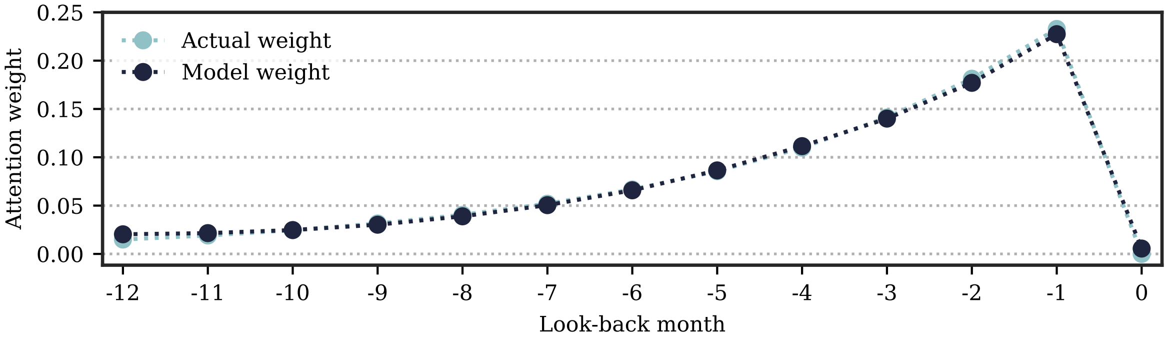

Fig. 11. Temporal Attention Weights – Simulated AR(12) Process

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. (16) 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.

6. Recovering Missing Firm Characteristics

The ultimate objective of our study is to provide a way to recover the distribution of missing firm characteristics. While a detailed discussion of the resulting recovered distribution for all 151 firm characteristics is beyond the scope of this paper, we now highlight the implications for a selected set. A discussion of the model's internal validity using an extended set of characteristics is provided in Appendix C.

Figure 12 shows the recovered distribution of *previously missing* entries for the book-to-market ratio (Fama and French, 1993, `be_me`), the Piotroski (2000) F-score (`f_score`), as well as momentum (Carhart, 1997, `ret_12_1`) and its constituent `ret_12_7` (Novy-Marx, 2012). The distribution of `be_me` is relatively balanced with slightly more firms with missing values being considered "value" firms. Our model reveals that the remaining firm characteristics, as well as their historical evolution lead to predictions of both low and high percentiles of `be_me`. Considering the evidence that `be_me` is often "missing" if the company's book value is negative, this finding makes sense as there is no prior on whether to put stocks with a negative book value in a particular `be_me` class. Our results are however in line with prior literature. For example, Brown et al. (2008) shows that firms with missing `be_me` are more likely value than growth firms. In addition, one may argue that a negative book-value, and consequently a missing entry for `be_me`, is associated with lower financial health. This interpretation is however not confirmed by the distribution of the `f_score`. The score by Piotroski (2000) is designed to grasp the

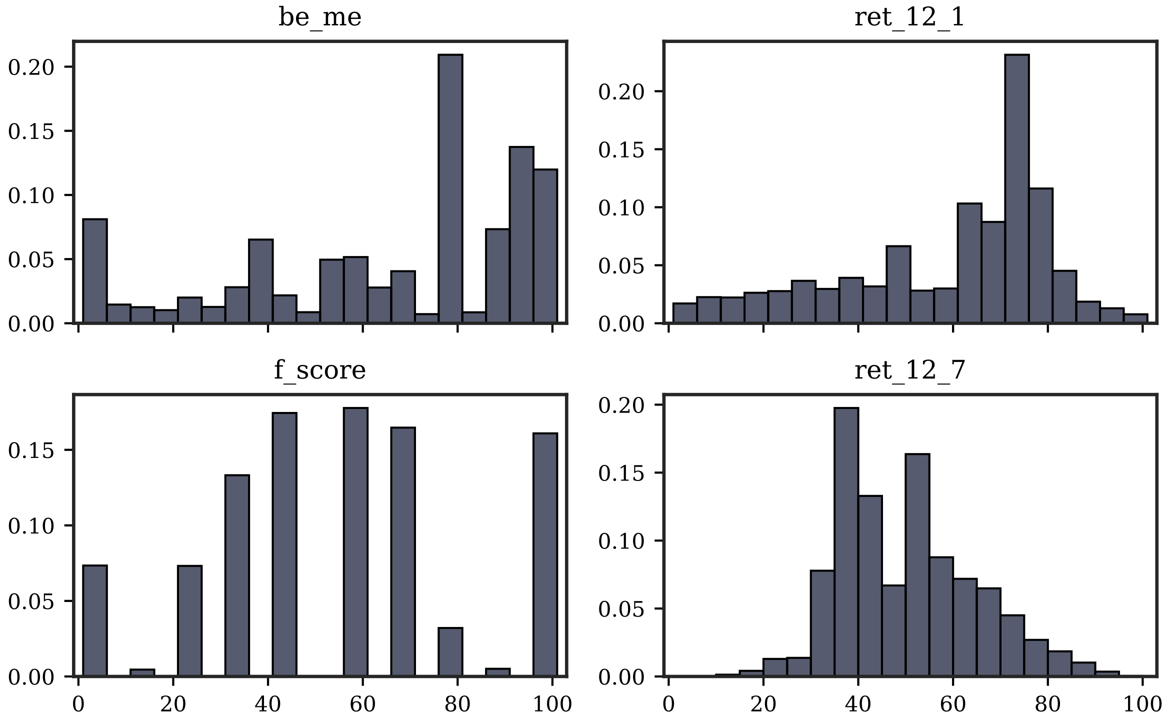

Fig. 12. Recovered Distribution of Missing Firm Characteristics.

The figure shows the distribution of recovered entries of previously missing inputs for the characteristics `be_me`, `f_score`, `ret_12_1`, and `ret_12_7`. The distribution is given in percentiles, from 1 to 100.

financial health status of a firm with lower values pointing towards financial distress. Our model predicts that firms with missing values are likely financially healthy. The same applies to observations with missing `be_me` which is in line with Luo, Liu, and Tripathy (2021), who argue that negative book equity firms are indeed financially healthy but increase their debt to meet current investment demand. Beside this interpretation, the `f_score` also provides a first internal sanity check of our model. Note the almost discrete distribution of the `f_score`, showing 10 peaks. This follows the score's definition as the sum of nine binary signals, consequently ranking from 0 to 9. Our model is able to infer even discrete distributions reliably. We further provide a secondary internal sanity check through momentum `ret_12_1`. The model primarily places firms with a missing entry for momentum in lower percentiles, suggesting that a significant portion of "Loser" stocks may be missing from the sample. This is confirmed by `ret_12_7`, which Novy-Marx (2012) shows is responsible for momentum profits. Together, these results show that the latent structure uncovered by the model in a self-supervised fashion produces sensible distributions of recovered firm characteristics. It further points out that the omission of missing characteristics is likely to skew the true underlying distribution.

Since we lack the controlled environment we had when reconstructing masked characteristic, we cannot directly assess the quality of the recovered characteristics. As a replacement, we propose a novel way to gauge the model's confidence about the recovery of a certain missing characteristic. In the extreme case of a total lack of both historical and cross-sectional information, the model's prediction will be random over the characteristic's percentile. This results in a uniform distribution, predicting each possible percentile with a probability of $1/100 = 1\%$. We can compute the similarity of this uniform distribution P, arising from uninformed guesses by the model, with the actual distribution Q across percentiles. Let both distributions be defined on the same probability space $\mathcal{X}$. We then calculate the *Jensen-Shannon divergence*,

$$\mathrm{JSD}(P \,||\, Q) = \frac{1}{2}D_{\mathrm{KL}}(P \,||\, M) + \frac{1}{2}D_{\mathrm{KL}}(Q \,||\, M), \qquad \text{where} \quad M = \frac{1}{2}(P+Q), \tag{17}$$

which is a symmetric version of the *Kullback-Leibler Divergence* D_{KL},

$$D_{\mathrm{KL}} = \sum_{x \in \mathcal{X}} \log\left(\frac{P(x)}{Q(x)}\right), \tag{18}$$

and bounded by 0 and log(2). A value of 0 indicates that $P = Q$, a value of log(2) that they are maximally different, i.e. all probability mass lies on one percentile. Adapted to our problem, the higher the value, the more confident the model in the prediction. Table 4 provides the average JSD for each characteristic.

Figure 13 shows the distribution of the resulting JSD confidence levels. For most recoveries, we find fairly high JSD values. The model is confident in its predictions. To put the values into perspective, consider the case in which the model assigns a probability of 10% to the true class, and $\frac{100-10}{N-1} = 0.6\%$ to the remaining classes. The JSD value in this case equals 0.037, which corresponds to the *lower edge* of our model's confidence. Higher JSD values indicate that our model produces distributions with a high probability mass at or around a single predicted percentile. We would generally not expect confidence levels at the upper limit of log(2), given that firm characteristics are likely measured with noise. The model therefore *should* place probability mass not on a single percentile, but in the vicinity of the most likely percentile. It is also more confident in its prediction than the benchmark case we have outlined, suggesting that the model places its bets with conviction.

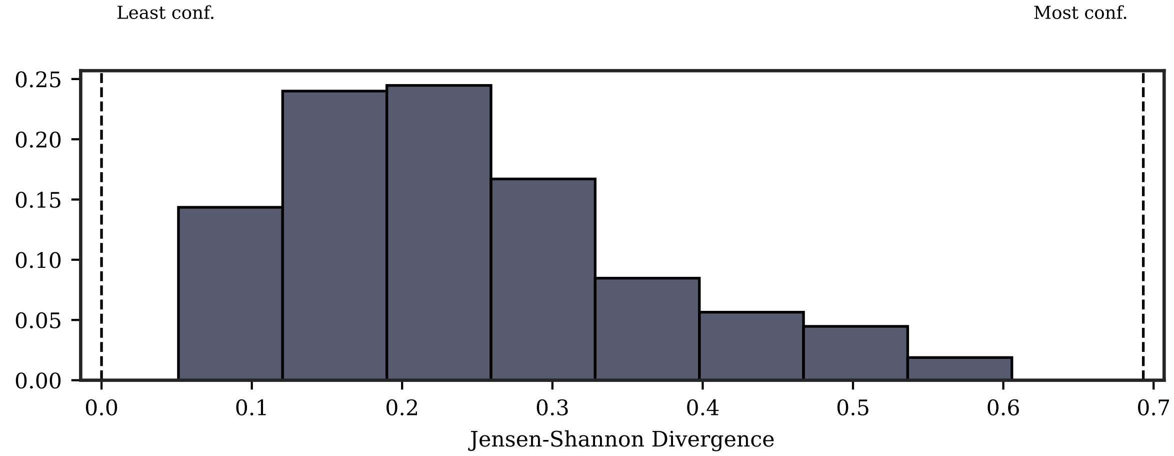

Fig. 13. Model Confidence.

The figure shows a histogram of the recovery confidence of our model, measured by the Jensen-Shannon divergence, defined in Eq. (17). The divergence is bounded between 0 and log(2), with higher values indicating a higher "conviction" in the model's attempt to recover missing 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. Fortunately, we can back out reasonable estimates for the *raw characteristics*, under the assumption that at least some information about the target characteristic is provided through 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 established results in economic research using this completed dataset is beyond the scope of this paper. We provide, however, the data for future research.[16]

[16] The data can be downloaded at `https://sites.google.com/view/beckmeyer/data-code`.

7. Application: Factor Portfolios in Finance

What is the impact of changes in the distribution of firm characteristics after filling in missing values? We study high-minus-low factor portfolios, which are a common application of this data and have been the cornerstone in financial research (Fama and French, 1993). The completed dataset can of course be used in many other applications, extending to research in Accounting, Market, Management, and Economics.

Figure 14 shows the changes in factor premia pre- and post-recovery of missing firm characteristics. For this exercise, we first sort stocks into deciles for a given characteristic. We then calculate the returns for each decile portfolio, first discarding missing values ("Pre"), and then using the imputed values using our best-fit model ("Post"). Changes in portfolio returns arise if the firms with recovered characteristics have different return patterns than the average return in the portfolio. To focus on the outright changes arising due to a change in the portfolio decomposition after imputing missing values, we consider equally-weighted returns. Using value-weights invariably masks part of the impact of missing values, as information about large stocks tends to be more complete and overall of better quality. We then form the zero-cost factor portfolio as the difference between the highest and lowest decile portfolio.

Figure 14 shows the resulting changes in the factor premia for the 20 characteristics with the lowest and the 20 characteristics with the highest "Pre"-premium in black. The red circles denote the factor premium *after* considering the impact of imputed missing values. An obvious trend emerges: using the completed set of firm characteristics almost uniformly pushes factor premia towards zero. The most severe examples for this are the premia for `f_score` (pre: 24%, post: 6%), residual momentum `resff3_12_1` (pre: 16%, post: 10%), and the change in net-operating assets `noa_gr1a` (pre: −18%, post: −13%). An interesting counterpoint to this is the change in the factor premium for `rd_me`, a measure of research and development expenditures. Here we find that using the completed dataset pushes the premium slightly upward, increasing it to 25%. Across all 151 characteristics, we find that the absolute change in the post- to pre-premium amounts to 2.35% per year.

Other characteristics are less affected. For example, the factor returns for `market_equity`, or intermediate momentum `ret_12_7` have barely changed. In conclusion, we can say that accounting for missing characteristics potentially leads to severe changes in the magnitude of factor premia. A careful consideration of this impact is warranted, and provides an

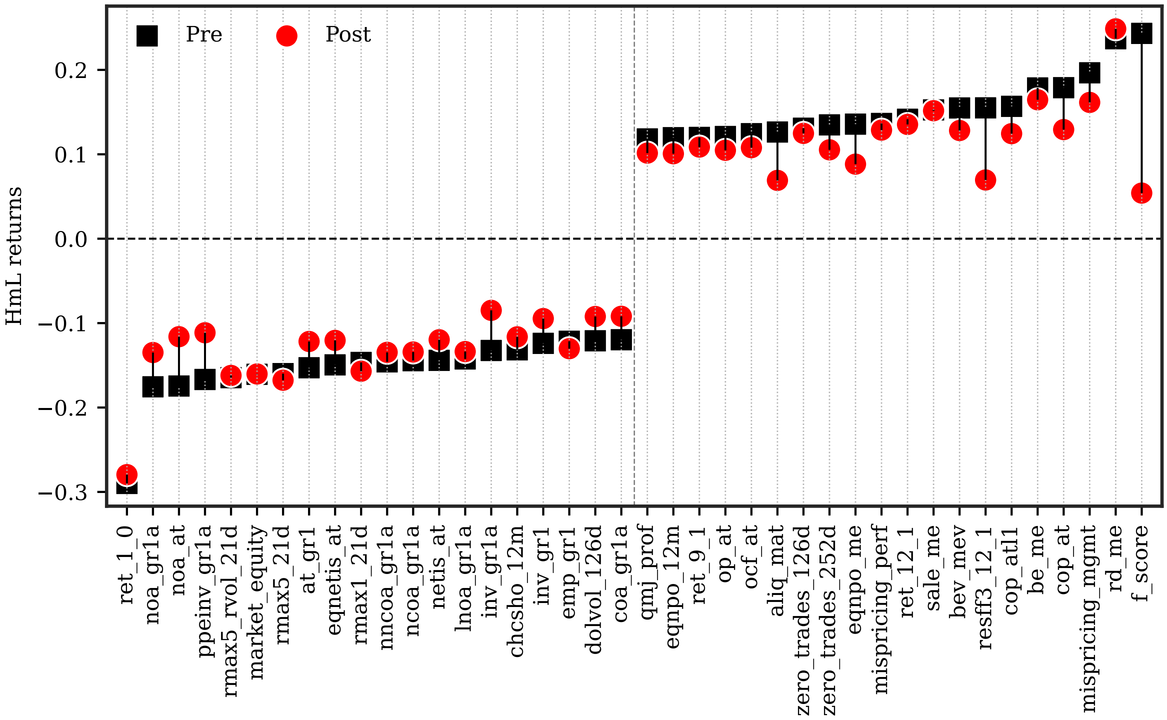

Fig. 14. Impact of Missing Observations on Factor Portfolio Returns.

The figure shows the change in high-minus-low factor portfolio returns for the 20 characteristics with the ex-ante ("Pre") smallest and largest premia. The premia without incorporating the information of imputed missing characteristics is given in black ("Pre"),. the premia after the inclusion of this information in red ("Post"). A complete list of these changes is provided in Table 5.

additional hurdle for newly proposed 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 13 of the 151 factor premia lose their significance. Still, most factor premia that are significant before the imputation remain significant thereafter, and others gain significance (see Table 5). In contrast to Freyberger et al. (2021), our method does not consider information from stock returns. It is instead agnostic about how the recovered firm characteristics change factor premia. The analysis provided here thus 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 (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 would argue in favor of reproducibility in Finance.

8. Conclusion

Missing firm characteristics are a pervasive issue in economic research. We propose a novel machine learning method to suitably impute these missing values, using both information about the historical evolution, as well as the cross-section of other observed characteristics. Our model produces predictions that are stable over time and significantly outperforms ad-hoc approaches, as well as recent advances in the financial literature. We highlight the importance of both sources of information – from observable characteristics, and over time. We highlight the efficacy of our modeling approach in a simulation study. The flexible setup is able to identify multiple processes underlying a vast array of input characteristic in a unified framework.

Not accounting for missing characteristics leads to flawed inference. We show that risk premia of many prominent stock market factors are likely much smaller than previously thought, after the inclusion of suitably filled missing entries of characteristics. We provide the imputed values for missing firm characteristics, as well as our fitted model for future research.

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.

Altman, E. I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance 23, 589–609.

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

Asness, C. S., Frazzini, A., Pedersen, L. H., 2019. Quality minus junk. Review of Accounting Studies 24, 34–112.

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

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.

Brown, S. J., Lajbcygier, P., Li, B., 2008. Going negative: What to do with negative book equity stocks. The journal of Portfolio Management 35, 95–102.

Bryzgalova, S., Lettau, M., Lerner, S., Pelger, M., 2021. Asset pricing with missing data. Working Paper .

Carhart, M. M., 1997. On persistence in mutual fund performance. The Journal of finance 52, 57–82.

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 .

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 .

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

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.

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 .

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.

Kaniel, R., Lin, Z., Pelger, M., Van Nieuwerburgh, S., 2021. Machine-learning the skill of mutual fund managers. Available at SSRN 3977883 .

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30, 3146–3154.

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 .

Liu, Y., Zhu, H., 2016. Campbell r. harvey. The Review of Financial Studies 29.

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.

Luo, H., Liu, I., Tripathy, N., 2021. A Study on Firms with Negative Book Value of Equity. International Review of Finance 21, 145–182.

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.

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

Ohlson, J. A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of accounting research pp. 109–131.

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.

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

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.

Xiong, R., Pelger, M., 2019. Large dimensional latent factor modeling with missing observations and applications to causal inference. arXiv preprint arXiv:1910.08273 .

Appendix

A. Model Setup in Detail

Feature Embeddings The financial literature points towards 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, Khetan, Cvitkovic, and Karnin, 2020; Somepalli, Goldblum, Schwarzschild, Bruss, and Goldstein, 2021; Lim et al., 2021; Gorishniy, Rubachev, Khrulkov, and Babenko, 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 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 15. 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.

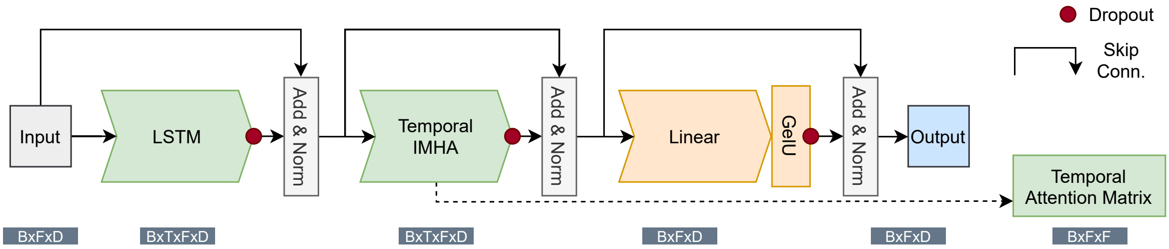

Fig. 15. Temporal Attention Network

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.

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

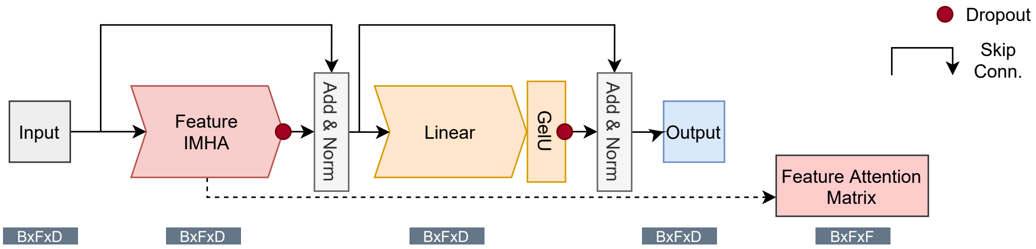

Fig. 16. Feature Attention Network

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.

characteristic's percentiles for all firm-month observations in the batch of size B. We then regard the most probable percentile as the predicted class and compare it with the true (unmasked) percentile.

B. *Accuracy, Missingness and Model Confidence*

The following Table 4 provides summary information about the model accuracy as measured by the expected percentile deviation defined in Eq. (10) 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. Conf. stands for model confidence which provides information about the value of the Jensen-Shannon divergence defined in Eq. (17). We further classified characteristics in accounting (A), hybrid (H) and market (M) variables.

Table 4: Missingness, accuracy and model confidence per characteristic.

	Expected percentile deviation						
	Full	Training	Validation	Testing	Miss. [%]	Conf.	Class
age	0.42	0.26	0.39	0.49	0.00	-	H
market_equity	0.67	0.49	0.61	0.75	0.48	-	M
sale_me	0.77	0.60	0.69	0.85	11.83	0.322	H
ret_12_7	0.79	0.53	0.65	0.93	17.03	0.402	M
dolvol_126d	0.86	0.55	0.69	1.03	10.69	0.149	M
at_me	0.87	0.67	0.78	0.96	11.44	0.305	H
qmj_prof	0.90	0.64	0.78	1.02	12.13	0.343	M
qmj	0.90	0.82	0.85	0.96	35.60	0.475	M
debt_me	1.01	0.83	0.87	1.10	11.70	0.285	H
corr_1260d	1.07	0.87	0.93	1.22	34.74	0.199	M
be_me	1.07	0.90	1.11	1.13	14.21	0.271	H
ami_126d	1.12	0.68	0.87	1.39	16.35	0.253	M
prc	1.12	0.74	1.24	1.25	0.48	-	M
div12m_me	1.12	0.91	0.77	1.27	7.69	0.486	H
netdebt_me	1.13	0.84	0.85	1.29	11.70	0.252	H
ivol_capm_252d	1.18	1.12	1.10	1.22	16.21	0.177	M
betabab_1260d	1.24	1.18	1.01	1.34	35.24	0.190	M
qmj_safety	1.25	1.02	1.07	1.38	8.56	0.266	M
rd_me	1.30	0.99	1.20	1.46	61.26	0.214	H
zero_trades_252d	1.35	0.70	0.91	1.72	12.68	0.085	M
op_atl1	1.36	0.90	1.15	1.60	14.75	0.329	A
bev_mev	1.37	1.21	1.51	1.41	16.07	0.323	H
gp_atl1	1.38	0.87	1.18	1.63	14.80	0.350	A
ni_me	1.40	0.92	1.03	1.66	11.60	0.513	H
ebitda_mev	1.49	1.12	1.37	1.67	13.56	0.341	H
eqnpo_me	1.51	1.14	1.82	1.63	28.12	0.389	H
eqnpo_12m	1.53	1.16	1.23	1.74	9.51	0.148	H
gp_at	1.54	0.90	1.18	1.87	11.81	0.243	A
rd5_at	1.55	1.14	1.52	1.83	73.83	0.273	A
qmj_growth	1.56	1.38	1.54	1.65	35.60	0.305	M
turnover_126d	1.56	0.56	0.99	2.11	10.69	0.192	M
op_at	1.56	0.99	1.24	1.86	11.74	0.172	A

Continued on next page.

Table 4: Missingness, accuracy and model confidence per characteristic.

eqpo_me	1.57	1.41	1.97	1.59	31.77	0.454	H
cop_at	1.59	0.98	1.61	1.87	19.91	0.337	A
ocf_me	1.60	0.87	1.05	2.02	13.49	0.318	H
at_turnover	1.62	0.89	1.12	2.02	12.74	0.289	A
rd_sale	1.64	1.00	1.26	2.01	62.01	0.329	A
opex_at	1.66	1.07	1.24	1.97	11.81	0.188	A
mispricing_mgmt	1.66	1.56	1.62	1.71	15.12	0.301	M
chcsho_12m	1.69	1.28	1.30	1.93	8.57	0.170	H
zero_trades_126d	1.71	0.62	1.05	2.32	10.69	0.113	M
ebit_sale	1.72	1.22	1.37	1.99	13.11	0.587	A
cop_atl1	1.75	0.99	1.65	2.13	20.47	0.354	A
mispricing_perf	1.77	1.15	1.30	2.09	5.20	0.208	M
ebit_bev	1.85	1.40	1.70	2.05	15.49	0.359	A
ret_12_1	1.90	1.16	1.62	2.29	17.10	0.268	M
ope_bel1	1.90	1.46	1.80	2.09	28.49	0.330	A
ope_be	1.91	1.33	1.71	2.16	25.25	0.259	A
eq_dur	1.98	1.59	1.80	2.18	24.23	0.207	A
sale_bev	2.00	1.42	1.80	2.26	15.43	0.262	A
ni_be	2.01	1.24	1.43	2.42	14.29	0.422	M
ivol_capm_21d	2.10	2.47	1.87	1.96	15.31	0.110	M
rvol_21d	2.27	2.20	2.03	2.35	15.31	0.156	M
nncoa_gr1a	2.28	1.84	2.25	2.48	23.42	0.152	A
ivol_ff3_21d	2.29	2.45	2.17	2.23	15.31	0.103	M
fcf_me	2.29	1.22	1.62	2.89	18.90	0.326	H
ret_9_1	2.32	1.67	2.13	2.64	15.17	0.265	M
seas_1_1na	2.32	1.94	2.47	2.53	35.88	0.306	M
rmax5_21d	2.49	2.15	2.17	2.73	15.32	0.155	M
at_be	2.55	1.68	1.87	3.03	14.00	0.433	A
beta_60m	2.56	2.52	1.65	2.76	26.26	0.053	M
aliq_mat	2.60	1.84	2.65	2.90	28.06	0.443	M
turnover_var_126d	2.61	2.29	2.40	2.79	10.69	0.193	M
noa_gr1a	2.62	2.11	2.39	2.88	24.38	0.089	A
dolvol_var_126d	2.66	2.26	2.51	2.86	10.69	0.138	M
ivol_hxz4_21d	2.66	2.83	2.60	2.59	24.50	0.132	M
ocfq_saleq_std	2.69	2.11	2.42	3.18	47.18	0.150	A
o_score	2.73	2.40	2.83	2.85	24.04	0.232	A
ocf_at	2.78	1.38	1.77	3.57	13.31	0.227	A
capx_gr3	2.79	2.45	2.88	2.95	35.12	0.227	A
ncoa_gr1a	2.79	2.74	3.00	2.77	21.98	0.148	A
ret_6_1	2.83	2.08	2.66	3.20	13.18	0.335	M
oaccruals_at	2.94	1.82	2.36	3.58	19.83	0.134	A
niq_at	3.00	2.40	2.68	3.35	29.13	0.202	A
z_score	3.08	2.42	3.17	3.33	25.61	0.200	A
capx_gr2	3.08	2.71	3.22	3.23	29.70	0.149	A
capex_abn	3.11	2.57	3.36	3.34	36.68	0.522	A
intrinsic_value	3.24	2.89	3.52	3.33	35.16	0.347	H
at_gr1	3.29	2.62	3.09	3.61	14.42	0.203	A
sale_gr3	3.30	2.82	3.25	3.53	26.84	0.143	A
ret_60_12	3.32	2.54	3.42	3.72	41.62	0.132	M

Continued on next page.

Table 4: Missingness, accuracy and model confidence per characteristic.

emp_gr1	3.33	3.14	3.51	3.38	29.45	0.150	A
niq_be	3.39	2.48	2.94	3.92	31.38	0.344	A
sale_emp_gr1	3.49	2.91	3.57	3.75	25.91	0.121	A
ret_3_1	3.52	2.32	2.94	4.16	11.13	0.350	M
tangibility	3.66	2.49	3.14	4.20	21.64	0.452	A
noa_at	3.69	2.51	3.08	4.29	23.92	0.252	A
kz_index	3.74	3.41	3.88	3.85	24.04	0.217	A
inv_gr1a	3.77	3.24	4.27	3.90	17.09	0.104	A
taccruals_at	3.80	2.17	2.77	4.76	20.36	0.233	A
lti_gr1a	3.83	3.11	3.78	4.16	21.55	0.255	A
eqnetis_at	3.88	2.80	3.85	4.39	27.74	0.468	H
fnl_gr1a	3.88	3.26	3.80	4.16	14.69	0.242	A
aliq_at	3.95	2.29	3.93	4.65	22.73	0.210	A
sale_gr1	4.06	3.07	3.47	4.59	16.66	0.213	A
cowc_gr1a	4.06	3.54	3.92	4.31	23.51	0.144	A
coa_gr1a	4.25	4.35	4.44	4.17	22.95	0.131	A
rmax5_rvol_21d	4.31	3.86	3.67	4.68	19.88	0.202	M
seas_2_5na	4.32	3.85	5.25	4.72	72.56	0.237	M
oaccruals_ni	4.32	2.67	3.49	5.25	19.87	0.231	A
nfna_gr1a	4.33	2.57	4.42	5.09	14.69	0.316	A
ni_ivol	4.41	4.07	4.01	4.68	36.38	0.222	A
inv_gr1	4.56	4.43	4.80	4.56	32.18	0.156	A
be_gr1a	4.61	4.19	4.30	4.84	18.81	0.203	A
taccruals_ni	4.66	3.09	3.48	5.63	20.42	0.255	A
niq_be_chg1	4.78	4.10	4.34	5.23	37.99	0.269	A
sti_gr1a	5.06	5.13	4.57	5.14	31.68	0.198	A
ret_1_0	5.12	4.14	4.47	5.67	9.72	0.287	M
netis_at	5.15	3.83	4.62	5.88	27.75	0.366	H
niq_at_chg1	5.27	4.83	4.86	5.60	35.07	0.216	A
zero_trades_21d	5.27	3.22	3.78	6.46	9.22	0.314	M
cash_at	5.30	3.36	4.14	6.31	12.44	0.203	A
capx_gr1	5.33	4.58	5.25	5.70	24.37	0.086	A
ni_inc8q	5.50	5.41	5.22	5.62	39.71	0.583	A
ppeinv_gr1a	5.66	5.79	6.04	5.54	24.09	0.105	A
prc_highprc_252d	5.66	4.65	5.50	6.18	16.23	0.208	M
dsale_dinv	5.75	5.24	5.80	5.94	35.99	0.198	A
lnoa_gr1a	5.77	4.69	7.02	6.08	25.93	0.146	A
resff3_12_1	5.86	5.55	5.30	6.10	19.23	0.123	M
betadown_252d	5.96	5.32	5.59	6.36	17.45	0.104	M
rmax1_21d	5.98	4.96	4.80	6.73	15.32	0.088	M
seas_6_10na	6.10	6.03	5.89	6.32	84.04	0.304	M
col_gr1a	6.12	5.17	5.91	6.55	21.63	0.061	A
seas_1_1an	6.59	4.57	5.56	7.65	13.08	0.214	M
debt_gr3	6.64	5.98	6.61	6.94	33.92	0.478	A
resff3_6_1	7.02	6.52	6.56	7.32	19.32	0.111	M
ncol_gr1a	7.16	6.73	7.53	7.27	22.50	0.109	A
dbnetis_at	7.23	5.42	6.86	8.05	12.42	0.326	H
f_score	7.40	6.66	6.85	7.82	29.08	0.526	A
dsale_dsga	7.45	6.41	6.74	8.07	34.26	0.156	A

Continued on next page.

Table 4: Missingness, accuracy and model confidence per characteristic.

saleq_su	7.46	7.13	7.73	7.58	35.95	0.177	A
dgp_dsale	7.65	6.19	7.33	8.36	25.08	0.181	A
saleq_gr1	7.66	6.00	6.63	8.63	24.49	0.261	A
tax_gr1a	7.76	7.84	7.40	7.79	15.30	0.105	A
pi_nix	7.87	7.33	7.50	8.12	33.81	0.111	A
niq_su	8.12	7.34	8.02	8.57	34.84	0.163	A
iskew_capm_21d	8.31	8.80	7.88	8.17	15.32	0.168	M
ocf_at_chg1	8.55	6.62	7.22	9.68	17.12	0.288	A
earnings_variability	8.62	8.96	9.59	8.21	37.51	0.116	A
rskew_21d	8.88	8.19	8.06	9.38	15.32	0.293	M
iskew_ff3_21d	10.09	9.75	10.08	10.26	15.31	0.108	M
ni_ar1	11.89	12.68	12.88	11.27	36.38	0.196	A
iskew_hxz4_21d	11.91	11.41	11.99	12.14	24.50	0.106	M
bidaskhl_21d	12.30	12.65	12.27	12.15	13.86	0.188	M
dsale_drec	12.68	13.18	13.11	12.36	24.50	0.124	A
seas_2_5an	23.48	16.35	22.50	27.87	44.89	0.105	M
beta_dimson_21d	24.57	23.56	23.06	25.39	15.31	0.124	M
coskew_21d	29.98	28.93	29.81	30.53	15.31	0.078	M
seas_6_10an	31.49	28.02	32.53	33.79	63.74	0.133	M
seas_16_20an	33.18	32.56	32.61	33.75	81.45	0.257	M
seas_11_15an	36.06	35.14	35.22	36.91	74.22	0.140	M
Average	4.75	4.11	4.52	5.09	23.58	0.239	

C. Additional Discussion of the Model's In-fill Distribution

In this part of the paper's Appendix we want to discuss the plausibility of the model's predictions for missing firm characteristics in more detail. To do so, we exploit the rich structure of the 151 firm characteristics in our dataset and identify two sets of characteristics that allow us to perform a visual sanity check of our results. The first group comprises *quality-minus-junk* Asness et al. (2019) and its *growth*, *safety*, and *profitability* components. Since `qmj` is simply the combination of the other three, the resulting in-fill distributions should be internally consistent. The second group comprises composite characteristics, that are obtained by combining the information from multiple "base" characteristics.

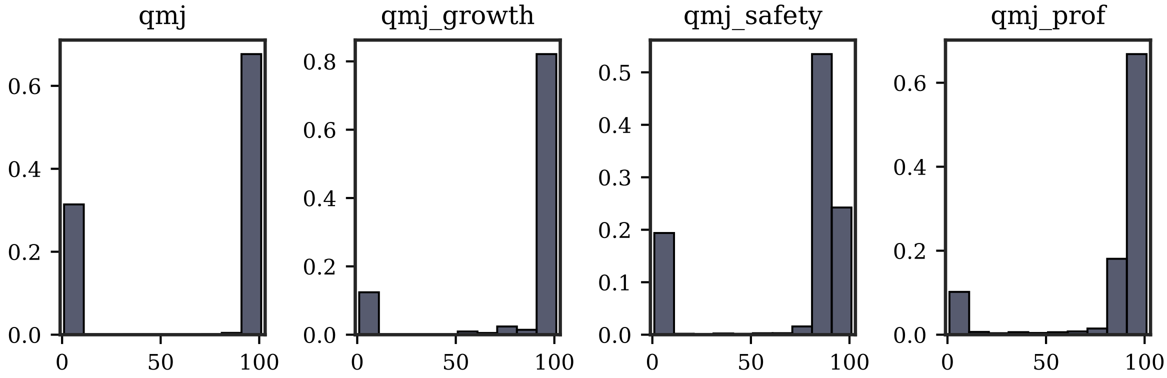

Fig. 17. Predicted Distribution of Missing Firm Characteristics – Quality-Minus-Junk.

The figure shows the distribution of recovered entries of previously missing inputs for the characteristics `qmj`, `qmj_growth`, `qmj_safety`, and `qmj_prof`. The distribution is given in percentiles, from 1 to 100.

Quality-Minus-Junk and its Parts The model generally estimates a bi-modal distribution at the two extremes – missing entries of `qmj` are predominantly placed in the lowest or highest percentiles. The distribution of the other characteristics, as well as their historical evolution disproportionately leads the model to only recover small and large percentiles for `qmj`. Looking at its constituents, we find a similar picture for all its components. All distributions, i.e. for `qmj_growth`, `qmj_safety` and `qmj_prof`, are bi-modal with more weight being put on the highest percentiles. The results clearly show, that the model's predictions of the recovered distribution for `qmj` and its constituents is internally consistent.

Composite Scores We next consider three composite scores found in the data set. The `f_score` by Piotroski (2000) was already discussed in Section 6. The `o_score` by Ohlson (1980) and the `z_score` by Altman (1968) both measure a firm's proximity to default. Higher values correlate to a higher chance of near-term default. For both variables, we find that the model predicts a diverse set of values, with a high mass around the median, but also considerable mass at both tails. In fact, missing entries for `z_score` are more often placed in high percentiles – a high chance of default – than they are placed in low percentiles. For `o_score` the probability of allocation to the highest and lowest percentile is comparable. The estimated distributions of the composite scores are internally consistent. They furthermore highlight the need for modern imputation methods as the one put forth in this paper, as the missing-at-random assumption is clearly violated.

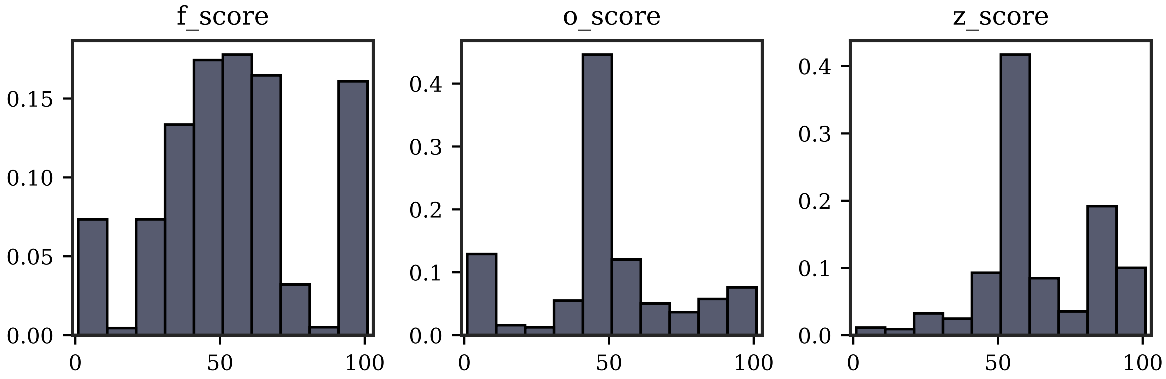

Fig. 18. Predicted Distribution of Missing Firm Characteristics – Composite Scores

The figure shows the distribution of recovered entries of previously missing inputs for the characteristics `f_score`, `o_score`, and `z_score`. The distribution is given in percentiles, from 1 to 100.

D. *Change in Factor Premia*

The following Table 5 reports changes in factor portfolio returns after the inclusion of firms with previously missing values. 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 11 occasions. We note, however, that the *total number* of significant factors is fairly constant at 120 before and 117 after imputation.

Table 5: Change in Factor Premia.

	HmL^{Pre}		HmL^{Post}		ΔHmL		Not sig.
f_score	0.24	***	0.05	**	−0.19	***	
seas_16_20an	0.08	***	−0.04	-	−0.12	***	Y
seas_1_1na	0.03	-	−0.08	**	−0.11	***	
seas_2_5an	0.10	***	−0.00	-	−0.10	***	Y
resff3_12_1	0.16	***	0.07	***	−0.09	***	
seas_11_15an	0.07	***	0.00	-	−0.07	***	Y
qmj	0.10	***	0.03	**	−0.06	***	
prc_highprc_252d	0.04	-	−0.02	-	−0.06	***	
aliq_mat	0.13	***	0.07	***	−0.06	***	
saleq_su	0.07	***	0.02	**	−0.06	***	
niq_be_chg1	0.10	***	0.05	***	−0.05	***	
ni_ivol	0.00	-	−0.05	-	−0.05	***	
cop_at	0.18	***	0.13	***	−0.05	***	
eqnpo_me	0.14	***	0.09	***	−0.05	***	
dsale_dinv	0.05	***	0.01	-	−0.05	***	Y
qmj_growth	0.06	***	0.01	-	−0.05	***	Y
nfna_gr1a	0.10	***	0.06	***	−0.04	***	
ebitda_mev	0.10	***	0.07	**	−0.04	***	
betabab_1260d	−0.05	-	−0.09	***	−0.04	***	
dolvol_var_126d	0.03	-	−0.00	-	−0.04	***	
mispricing_mgmt	0.20	***	0.16	***	−0.04	***	
fcf_me	0.11	***	0.07	***	−0.03	***	
niq_at_chg1	0.08	***	0.05	***	−0.03	***	
cop_atl1	0.16	***	0.12	***	−0.03	***	
seas_1_1an	0.09	***	0.06	***	−0.03	***	
zero_trades_252d	0.13	***	0.11	***	−0.03	***	
resff3_6_1	0.09	***	0.06	***	−0.03	***	
sale_bev	0.10	***	0.07	***	−0.03	***	
ocf_me	0.10	***	0.08	***	−0.03	***	
niq_su	0.10	***	0.07	***	−0.03	***	
ocf_at_chg1	0.05	***	0.02	**	−0.03	***	
bev_mev	0.15	***	0.13	***	−0.03	***	
iskew_ff3_21d	−0.02	**	−0.05	***	−0.03	***	
bidaskhl_21d	0.02	-	−0.00	-	−0.03	***	

Continued on next page.

Table 5: Change in Factor Premia.

turnover_var_126d	0.05	**	0.02	-	−0.02	***	Y
ni_me	0.05	-	0.02	-	−0.02	***	
tangibility	0.09	***	0.06	***	−0.02	**	
beta_60m	−0.01	-	−0.03	-	−0.02	-	
ivol_capm_252d	−0.06	-	−0.08	**	−0.02	**	
eq_dur	−0.11	***	−0.13	***	−0.02	**	
seas_6_10an	0.09	***	0.06	***	−0.02	*	
eqnpo_12m	0.12	***	0.10	***	−0.02	***	
iskew_hxz4_21d	−0.03	***	−0.05	***	−0.02	***	
ope_bel1	0.06	**	0.04	*	−0.02	*	
sale_emp_gr1	−0.02	***	−0.04	***	−0.02	***	
qmj_prof	0.12	***	0.10	***	−0.02	***	
ocf_at	0.12	***	0.11	***	−0.02	***	
op_at	0.12	***	0.10	***	−0.02	***	
dgp_dsale	0.06	***	0.05	***	−0.02	**	
rd_sale	0.01	-	−0.00	-	−0.01	-	
ope_be	0.07	**	0.06	**	−0.01	**	
pi_nix	0.01	-	−0.00	-	−0.01	*	
ebit_bev	0.06	*	0.04	-	−0.01	**	Y
be_me	0.18	***	0.16	***	−0.01	***	
iskew_capm_21d	−0.05	***	−0.06	***	−0.01	**	
ebit_sale	0.05	-	0.04	-	−0.01	**	
saleq_gr1	−0.01	-	−0.03	**	−0.01	-	
div12m_me	0.02	-	0.01	-	−0.01	-	
coskew_21d	−0.00	-	−0.01	-	−0.01	*	
ret_9_1	0.12	***	0.11	***	−0.01	***	
opex_at	0.03	-	0.01	-	−0.01	***	
rmax1_21d	−0.15	***	−0.16	***	−0.01	-	
sti_gr1a	−0.03	**	−0.04	***	−0.01	-	
o_score	−0.02	-	−0.03	-	−0.01	**	
dsale_dsga	0.01	-	−0.00	-	−0.01	-	
ivol_hxz4_21d	−0.11	***	−0.12	***	−0.01	-	
at_turnover	0.04	***	0.03	**	−0.01	**	
emp_gr1	−0.12	***	−0.13	***	−0.01	***	
mispricing_perf	0.14	***	0.13	***	−0.01	***	
rmax5_21d	−0.16	***	−0.17	***	−0.01	-	
qmj_safety	0.05	*	0.05	-	−0.01	***	Y
at_be	−0.01	-	−0.02	-	−0.01	-	
niq_be	0.12	***	0.11	***	−0.01	-	
gp_atl1	0.04	*	0.03	*	−0.01	-	
op_atl1	0.09	***	0.08	***	−0.01	-	
at_me	0.11	***	0.10	***	−0.01	***	
ret_12_1	0.14	***	0.14	***	−0.01	***	
zero_trades_126d	0.13	***	0.12	***	−0.01	-	
gp_at	0.09	***	0.09	***	−0.01	**	
ret_6_1	0.10	***	0.09	**	−0.01	***	
beta_dimson_21d	−0.03	-	−0.04	*	−0.00	-	
ret_3_1	0.06	**	0.06	*	−0.00	-	
earnings_variability	0.01	-	0.01	-	−0.00	-	

Continued on next page.

Table 5: Change in Factor Premia.

cash_at	0.04	*	0.04	*	−0.00	-	
eqpo_me	0.05	**	0.04	***	−0.00	-	
ivol_ff3_21d	−0.11	**	−0.11	***	−0.00	-	
debt_me	0.01	-	0.00	-	−0.00	-	
niq_at	0.09	***	0.09	***	−0.00	-	
sale_me	0.15	***	0.15	***	−0.00	-	
rd5_at	0.06	*	0.06	**	−0.00	-	
ret_12_7	0.11	***	0.11	***	−0.00	-	
age	0.01	-	0.01	-	0.00	-	
market_equity	−0.16	***	−0.16	***	0.00	-	
prc	−0.08	*	−0.07	*	0.00	-	
lti_gr1a	−0.04	***	−0.04	***	0.00	-	
ni_be	0.06	*	0.06	**	0.00	-	
rmax5_rvol_21d	−0.16	***	−0.16	***	0.00	-	
kz_index	−0.04	**	−0.04	*	0.00	-	
seas_2_5na	−0.08	***	−0.07	***	0.00	-	
zero_trades_21d	0.06	*	0.06	**	0.00	-	
turnover_126d	−0.10	***	−0.10	***	0.00	-	
tax_gr1a	0.02	-	0.02	**	0.01	*	
taccruals_ni	−0.04	***	−0.03	***	0.01	-	
dsale_drec	0.01	-	0.02	**	0.01	-	
ami_126d	0.08	**	0.08	***	0.01	-	
lnoa_gr1a	−0.14	***	−0.13	***	0.01	*	
ncol_gr1a	−0.01	-	−0.01	-	0.01	-	
seas_6_10na	−0.09	***	−0.08	***	0.01	-	
ncoa_gr1a	−0.14	***	−0.13	***	0.01	**	
ret_1_0	−0.29	***	−0.28	***	0.01	***	
ni_inc8q	0.01	-	0.02	*	0.01	**	
nncoa_gr1a	−0.15	***	−0.13	***	0.01	***	
dbnetis_at	−0.10	***	−0.08	***	0.01	***	
ivol_capm_21d	−0.11	**	−0.10	***	0.01	-	
debt_gr3	−0.10	***	−0.09	***	0.01	-	
rd_me	0.24	***	0.25	***	0.01	**	
z_score	−0.05	*	−0.03	-	0.01	***	Y
ocfq_saleq_std	−0.04	-	−0.02	-	0.01	-	
sale_gr3	−0.08	***	−0.07	***	0.02	-	
chcsho_12m	−0.13	***	−0.12	***	0.02	***	
netdebt_me	−0.08	***	−0.07	***	0.02	***	
oaccruals_at	−0.11	***	−0.09	***	0.02	***	
ni_ar1	−0.01	-	0.01	-	0.02	***	
taccruals_at	−0.04	**	−0.02	-	0.02	***	Y
capx_gr1	−0.10	***	−0.08	***	0.02	***	
betadown_252d	−0.05	-	−0.03	-	0.02	**	
netis_at	−0.14	***	−0.12	***	0.02	***	
fnl_gr1a	−0.12	***	−0.09	***	0.03	***	
sale_gr1	−0.10	***	−0.07	***	0.03	***	
corr_1260d	−0.04	-	−0.01	-	0.03	**	
coa_gr1a	−0.12	***	−0.09	***	0.03	***	
col_gr1a	−0.05	***	−0.02	**	0.03	***	

Continued on next page.

Table 5: Change in Factor Premia.

dolvol_126d	−0.12	***	−0.09	***	0.03	***	
eqnetis_at	−0.15	***	−0.12	***	0.03	***	
inv_gr1	−0.12	***	−0.09	***	0.03	***	
at_gr1	−0.15	***	−0.12	***	0.03	***	
be_gr1a	−0.12	***	−0.08	***	0.03	***	
intrinsic_value	−0.04	-	−0.01	-	0.03	***	
rskew_21d	−0.06	***	−0.03	**	0.03	***	
aliq_at	−0.11	***	−0.08	***	0.03	**	
cowc_gr1a	−0.10	***	−0.06	***	0.04	***	
capex_abn	−0.06	***	−0.03	***	0.04	***	
capx_gr2	−0.11	***	−0.07	***	0.04	***	
capx_gr3	−0.11	***	−0.07	***	0.04	***	
noa_gr1a	−0.18	***	−0.13	***	0.04	***	
ret_60_12	−0.09	***	−0.04	***	0.04	**	
oaccruals_ni	−0.12	***	−0.07	***	0.05	***	
inv_gr1a	−0.13	***	−0.08	***	0.05	***	
ppeinv_gr1a	−0.17	***	−0.11	***	0.06	***	
noa_at	−0.17	***	−0.12	***	0.06	***	
rvol_21d	−0.11	***	−0.05	-	0.07	***	Y
Σ		120		117		108	11

E. Model Comparison - R^2 by Imputation Method

The following Table 6 shows the imputation accuracy by R^2, which measures how much variation in observable characteristics a method can explain. As we discretize each characteristic into percentiles we calculate the R^2 in the following fashion:

$$R^2_{\mathrm{x}} = 1 - \frac{\sum_{k=0}^{99} \; p_{\mathrm{x}}(|\Delta| = k) \cdot (k/100)^2}{\sum_{k=0}^{99} \; p_{\mathrm{MI}}(|\Delta| = k) \cdot (k/100)^2} \tag{19}$$

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 a cross-sectional model, which disregards temporal information. We further consider imputing masked features with the last time-series observation or with the cross-sectional median as competing approaches. If the last value is not available, we set the value to the cross-sectional median. Results are shown for market- and accounting-based, as well as hybrid characteristics. The best performing model is highlighted in bold for each case.

Table 6: Model Comparison – R^2 by Imputation Method.

	R^2			
	Full	Training	Validation	Testing
		All		
Full model	**0.863**	**0.878**	**0.874**	**0.854**
X-Sectional model	0.599	0.674	0.652	0.554
Last	0.574	0.567	0.581	0.576
Mean imputation	0.000	0.000	0.000	0.000
		Accounting		
Full model	**0.868**	**0.890**	**0.877**	**0.857**
X-Sectional model	0.559	0.640	0.599	0.514
Last	0.605	0.609	0.615	0.601
Mean imputation	0.000	0.000	0.000	0.000
		Market		
Full model	**0.811**	**0.822**	**0.828**	**0.803**
X-Sectional model	0.577	0.634	0.636	0.537
Last	0.393	0.375	0.401	0.400
Mean imputation	0.000	0.000	0.000	0.000
		Hybrid		
Full model	**0.956**	**0.969**	**0.964**	**0.949**
X-Sectional model	0.764	0.867	0.846	0.705
Last	0.869	0.874	0.876	0.865
Mean imputation	0.000	0.000	0.000	0.000

F. Hyperparameters

The following Table 7 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, Lyu, Tutunov, Wang, Grosnit, Griffiths, Maraval, Jianye, Wang, Peters, et al. (2020).

Table 7: 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	$\in [\textbf{Soft}, \text{Ent}, \text{Sparse}]\text{Max}$	
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	$\in [\text{Soft}, \textbf{Ent}, \text{Sparse}]\text{Max}$	
TAN Dropout	$\in [\mathbf{0.0}, 0.1, 0.3]$	
TAN Linear Dropout	$\in [0.0, 0.1, \mathbf{0.3}]$	
MLP dropout	$\in [\mathbf{0.0}, 0.1, 0.3]$	