Upgrading Credit Pricing and Risk Assessment through Embeddings*

Xavier Gabaix† Ralph S.J. Koijen‡ Robert J. Richmond§ Motohiro Yogo¶

February 8, 2025

Abstract

Credit ratings are central in fixed income markets, defining mutual fund benchmarks and risk-based capital regulation of insurance companies. Credit ratings explain a large share of the variation in corporate credit spreads, but a surprising amount of variation remains across firms. Asset pricing theory implies that equilibrium bond prices depend not only on credit ratings but all information that investors use to form their portfolios. We extract a high dimensional representation of this information, called firm embeddings, from US corporate bond holdings of mutual funds and insurance companies. Within broad credit rating categories, firm embeddings explain credit spreads and the volatility of credit spreads better than credit ratings and the distance to default. Therefore, firm embeddings can augment (and eventually replace) credit ratings to provide more timely and accurate information for fixed income markets. We illustrate the potential impact of an improved rating system on the risk-based capital regulation of insurance companies.

JEL: G12, G22, G23

Keywords: Artificial intelligence, Asset pricing, Credit rating, Credit spread, Machine learning, Risk-based capital regulation

*For helpful comments and suggestions, we thank Mike Rashes. For financial support, Gabaix thanks the Ferrante Fund, and Koijen thanks the Center for Research in Security Prices at the University of Chicago and the Fama Research Fund at the University of Chicago Booth School of Business.

†Harvard University, NBER, and CEPR (xgabaix@fas.harvard.edu)

‡University of Chicago Booth School of Business, NBER, and CEPR (ralph.koijen@chicagobooth.edu)

§New York University Stern School of Business and NBER (rrichmon@stern.nyu.edu)

¶Princeton University and NBER (myogo@princeton.edu)

1. Introduction

Credit ratings are the most important measure of risk in fixed income markets. They define mutual fund benchmarks such as the Bloomberg US Corporate Investment Grade Index. They also determine risk weights on fixed income securities as part of risk-based capital regulation of insurance companies. Credit rating agencies assign credit ratings primarily based on financial ratios that characterize a firm's business profile, profitability, leverage, and financial policy (Moody's Investor Service, 2021; S&P Global, 2024b). Credit ratings may not be timely or accurate for various reasons including the historical nature of financial statements, the difficulty of forecasting firm performance, and the laborious nature of the rating process. For this reason, credit rating agencies offer market signals such as Moody's KMV, which is a commercial implementation of the distance to default that uses equity prices, to augment credit ratings. Thus, credit ratings and the distance to default form a powerful combination of signals for investors to price fixed income securities (Liu et al., 2020).

Unsurprisingly, credit ratings and the distance to default explain a large share of the variation in US corporate credit spreads (i.e., yield spreads between corporate bonds and duration-matched Treasury bonds). However, a surprising amount of variation in credit spreads remains across firms, conditional on credit ratings. For example, BBB firms have an average credit spread of 1.94% over the sample period from September 2002 to December 2022. The cross-sectional standard deviation of average credit spreads for these firms is 0.74%. The large standard deviation suggests that credit ratings are not sufficient for credit spreads and that investors are pricing other aspects of these firms. We document this motivating fact through a regression approach at the bond level. A regression of credit spreads on credit ratings, the distance to default, their interactions with a market factor, and a full set of bond characteristics attains an $R^2$of 68%. The standard deviation of residuals from this regression is 0.86% for BBB bonds. Adding firm fixed effects to this regression increases the R^2 to 76%, suggesting the presence of information about firms that is difficult to capture with observed characteristics.

We use an asset pricing model to guide our empirical exercise of extracting information about firms from bond holdings data. In every asset pricing model, investors choose optimal portfolios, and market clearing implies equilibrium asset prices. Thus, bond holdings data contain all relevant information for bond prices. This information not only includes credit ratings and the distance to default but also characteristics unobserved by the econometrician. By substituting out equilibrium bond prices, we can write reduced-form demand as the dot product of investor loadings (i.e., principal component loadings) and firm embeddings (i.e.,

principal component scores). Thus, we can estimate firm embeddings through principal component analysis of bond holdings data. Identification requires sufficient cross-sectional variation in the investor loadings, which arises from heterogeneity in risk aversion, beliefs, risk constraints, and investment mandates in our model. For example, insurance companies face a value-at-risk constraint that does not apply to mutual funds. Estimation of reduced-form demand does not require instruments or comprehensive holdings data that satisfy market clearing in contrast to structural demand estimation (Koijen and Yogo, 2019; Bretscher et al., 2023; Chaudhary et al., 2023). Thus, Gabaix et al. (2024) and this paper develop a recipe for prediction exercises using an asset demand system, which are possible under weaker identifying assumptions than those required for counterfactual exercises.

We use US corporate bond holdings of mutual funds and insurance companies over a quarterly sample period from September 2002 to December 2022. In each cross section, we extract firm embeddings of 64 elements. We then estimate a ridge regression of credit spreads on the firm embeddings, where we estimate the regularization parameter by cross validation. We refer to the linear combination of firm embeddings that best explains credit spreads as a trained embedding. Across all credit rating categories, the trained embeddings have higher explanatory power than the combination of credit ratings and the distance to default.[1] In the largest BBB rating category, the trained embeddings explain credit spreads with an R^2 of 67%, beating 60% for the combination of credit ratings and the distance to default.

We use the same trained embeddings to explain quarterly changes in credit spreads and the volatility of credit spreads as two additional benchmarks. We model the change in credit spreads as a factor model, where the factor loading depends on the credit rating, the distance to default, and the trained embedding. Thus, this benchmark asks which risk measures explain comovement of changes in credit spreads. Across all credit rating categories, the trained embeddings have higher explanatory power for changes in credit spreads than credit ratings but lower explanatory power than the combination of credit ratings and the distance to default. Similarly, we model the volatility of credits spreads as a factor model to ask which risk measures explain comovement. With a minor exception of the highest credit rating category, the trained embeddings have higher explanatory power for the volatility of credit spreads than the combination of credit ratings and the distance to default.

The three benchmarks show that the trained embedding is a risk measure that explains comovement in credit spreads, changes in credit spreads, and the volatility of credit spreads. The results suggest that investors have information about differences in risk across firms that credit ratings and the distance to default do not fully capture. We find low-frequency evidence that further supports this interpretation. Among firms that are currently rated

[1]The six credit rating categories are AAA and AA, A, BBB, BB, B, and CCC and below.

investment grade, the trained embeddings predict a downgrade to speculative grade over the subsequent year and default over the subsequent five years. This prediction is robust to controlling for credit ratings, rating watch indicators, the distance to default, and credit spreads.

These results show that the firm embeddings can augment credit ratings to provide more timely and accurate information for fixed income markets. As a proof of concept, we use the firm embeddings to improve credit ratings. For each date, we rank the bonds by factor loadings for credit spreads, which depend on credit ratings and the trained embeddings. We then assign counterfactual ratings that have the same distribution as the actual ratings. We illustrate the potential impact of an improved rating system through risk-based capital regulation of insurance companies. For each insurance company, we compute required capital on the corporate bond portfolio under the counterfactual ratings.[2] Under the improved rating system, the average insurance company would have to increase or decrease its equity by 16 percentage points to maintain the same risk-based capital ratio. Reassuringly, the improved rating system is just as stable as the actual rating system. The distribution of annual changes in required capital are nearly identical under the two rating systems.

This paper is closest to a literature that investigates the determinants of corporate credit spreads in levels (Campbell and Taksler, 2003) and changes (Collin-Dufresne et al., 2001), focusing on observed characteristics that are motivated by the Merton (1974) model. Both the level of and changes in credit spreads have a factor structure, where the factor loadings depend on risk measures such as credit ratings and the distance to default. Our contribution is to use big data on institutional bond holdings and machine learning to improve the estimation of factor loadings and thereby increase explanatory power for credit spreads. We do not attempt to resolve the credit spread puzzle of whether credit spreads are consistent with structural models of credit risk, given the relatively short sample period (Huang and Huang, 2012; Feldhütter and Schaefer, 2018).

Gabaix et al. (2024) developed the methodology for estimating embeddings and testing them on benchmarks, using US stock market data. In this paper, we find that the same general methodology works well for the US corporate bond market. Thus, we have continuity in the research methodology, even though the research questions are entirely different. This continuity validates the asset pricing theory that guides our empirical approach. It also suggests an exciting possibility that our methodology is more widely applicable across different countries, asset classes, and benchmarks. This possibility remains to be proven in

[2]Required capital is the sum of each bond holding times the risk weight corresponding to its credit rating. The risk weights are 0.39% for A and above, 1.26% for BBB, 4.46% for BB, 9.70% for B, and 22.31% for CCC and below (Levy et al., 2021, p. 7).

future research.

The remainder of the paper is organized as follows. In Section 2, we describe the data construction and present summary statistics on the corporate bond market and credit spreads. In Section 3, we present an asset pricing model that guides our empirical approach. In Section 4, we describe the estimation methodology and present summary statistics for the firm embeddings. In Section 5, we train the firm embeddings to explain credit spreads and test the trained embeddings on three benchmarks. In Section 6, we provide additional evidence to help us understand the trained embeddings, by connecting them to low-frequency risk and the investor loadings. In Section 7, we illustrate the potential impact of an improved rating system on the risk-based capital regulation of insurance companies. Section 8 concludes.

2. Data Construction and Summary Statistics

We describe the construction of US corporate bond market data. We then present summary statistics on the corporate bond market and credit spreads that motivate our study.

2.1. Corporate Bond Market Data

We construct US corporate bond market data at the quarterly frequency from September 2002 to December 2022. Our data sources are Mergent (2024) for the bond characteristics, FINRA (2024) for the bond prices, and Thomson Reuters (2024) for the bond holdings of institutional investors. We clean these datasets as described in Appendix B and merge them by CUSIP.

The firm characteristics are from the Global Stock Returns and Characteristics Data (Jensen et al., 2023), which clean and merge the Center for Research in Securities Prices (CRSP) US Stock Database and the Compustat Database. As we describe in Appendix B, we construct a link file between a bond's issuer (6-digit) CUSIP and gvkey (Compustat Global Company Key) to merge with the CRSP-Compustat Database. Throughout the paper, our definition of a "firm" is an issuer CUSIP, which is the primary unit of analysis for credit ratings and default. The link between issuer CUSIP and gvkey can change over time due to corporate events such as mergers and acquisitions.

2.1.1. Bond-Level Data

Our sample consists of corporate bonds that appear in all four datasets. We further filter the sample to bonds that have at least $1 million outstanding, have a maturity between 1 and 30 years, have a credit rating by one of the three major rating agencies (i.e., Moody's,

S&P, or Fitch), are not in default, and have a traded price in the quarter of observation. These criteria are intended to eliminate small or illiquid bonds.

Throughout the paper, our definition of an "investor" is a subaccount ID in Thomson Reuters (2024). This definition generally corresponds to a mutual fund (e.g., Vanguard Wellington Fund) or an insurance subsidiary (e.g., MetLife Investors USA Insurance Company). We keep mutual funds and insurance companies, eliminating pension funds that have low and inconsistent coverage in Thomson Reuters (2024). We further filter the sample to investors that own at least 20 firms (i.e., issuer CUSIP) and bonds that are owned by at least 20 investors. These criteria eliminate investors with highly concentrated portfolios or bonds with highly concentrated ownership, for which estimation of firm embeddings is challenging.

Since we restrict our sample to corporate bonds held by mutual funds and insurance companies, we have better coverage of investment-grade bonds (rated BBB− and above) than speculative-grade bonds (rated BB+ and below). We present the results for speculative-grade bonds for completeness, but we focus our discussion on the results for investment-grade bonds throughout the paper. By using nonpublic or foreign data, future research may be able to improve the sample coverage to include pension funds, banks, and other types of investors.

We construct the credit spread as the semiannually compounded yield minus the duration-matched zero-coupon Treasury yield (Gürkaynak et al., 2007). We construct the change in credit spreads between the next quarter and this quarter. In this construction, we use the credit spread in the next quarter only if the trade occurs within the last 10 trading days of the quarter to ensure that the change in credit spreads is not based on stale prices. We construct the volatility of credit spreads as the annualized standard deviation of daily credit spreads within a quarter.

We use credit ratings and rating watch indicators, prioritized in the order of Moody's, S&P, and Fitch. Credit rating agencies issue a positive or a negative rating watch to indicate that a credit rating is under review for a possible upgrade or downgrade. According to the Merton (1974) model, a firm's default probability is decreasing in its distance to default, which is the asset value minus the face value of debt, divided by the standard deviation of asset value. Moody's KMV, which is a commercial implementation of the distance to default, is used to price corporate bonds (Liu et al., 2020). Because historical data from Moody's KMV are unavailable for academic research, we follow the procedure in Campbell et al. (2008, pp. 2936–2937) to estimate the distance to default. The inputs in this procedure are the ratio of market equity to the book value of debt, the 252-day equity volatility, and the 1-year Treasury yield.

We use bond characteristics as controls in some of our empirical specifications. They

are the Macaulay duration, log market value outstanding, indicator variables for embedded options (i.e., callable, putable, and credit enhancement), and indicator variables for covenants (i.e., bondholder protective, issuer restrictive, and subsidiary restrictive).

2.1.2. Firm-Level Data

By aggregating the bond-level data, we construct firm-level data with unique observations by date, firm (i.e., issuer CUSIP), and investor (i.e., subaccount ID). We construct the firm-level credit spread as an average credit spread over the firm's bonds, weighted by market value outstanding. We do the same for the change in and the volatility of credit spreads. We aggregate credit ratings and rating watch indicators by using the most recent rating update by firm.

2.2. Summary of the Corporate Bond Market

Figure 1 summarizes the corporate bond market over the sample period from September 2002 to December 2022. In Panel A, the aggregate market value of corporate bonds grows from \$1.281 trillion in September 2002 to \$4.059 trillion in December 2022. In Panel B, the share of the aggregate market value held by mutual funds grows from 4% in September 2002 to 17% in December 2022. The share of the aggregate market value held by insurance companies is more constant, declining slightly from 30% in September 2002 to 24% in December 2022. Thus, the bond holdings of mutual funds and insurance companies in Thomson Reuters (2024) cover 41% of the corporate bond market in December 2022.

2.3. Summary of Credit Spreads

Panel A of Figure 2 shows the time series of the average credit spread by credit rating category. In the cross section, the average credit spread is decreasing in the credit rating. In the time series, the average credit spread has a factor structure, where the factor loading is decreasing in the credit rating.

Panel B of Figure 2 shows the time series of the cross-sectional mean of the annualized volatility of credit spreads by credit rating category. In the cross section, the volatility of credit spreads is decreasing in the credit rating. In the time series, the volatility of credits spreads has a factor structure, where the factor loading is decreasing in the credit rating.

Table 1 summarizes firm-level credit spreads by credit rating. For BBB firms, credit spreads have a mean of 1.94% and a standard deviation of 1.27%. Some of this variation comes from the factor structure in the time series. To isolate the cross-sectional variation, we compute the standard deviation of credit spreads by date and credit rating, which we then

average over the sample period. The cross-sectional standard deviation of credit spreads is 0.74% for BBB firms. The surprising fact is that there is large variation in credit spreads across firms that credit ratings do not fully capture.

We document the key insights from Figure 2 and Table 1 more formally in a regression framework. We estimate a regression model for the credit spread on bond n issued by firm f at date t:

$$Y_{n,f,t} = \left(\beta_Y + \boldsymbol{\beta}'_{C,Y}\boldsymbol{C}_{n,f,t}\right) Y_t + \boldsymbol{\beta}'_C\boldsymbol{C}_{n,f,t} + \boldsymbol{\beta}'_D\boldsymbol{D}_{n,f,t} + \epsilon_{n,f,t}. \tag{1}$$

The vector $\boldsymbol{C}_{n,f,t}$ contains indicator variables for credit rating categories and the distance to default. The variable Y_t is the market factor at date t, which is the value-weighted average of credit spreads. Thus, we model both the level and the factor loading to depend on the credit rating and the distance to default. The vector $\boldsymbol{D}_{n,f,t}$ contains bond characteristics that include duration and its interaction with credit rating categories, log market value outstanding, and indicator variables for embedded options and covenants. $\epsilon_{n,f,t}$ is an error term.

The first column of Table 2 reports regression (1) with only credit ratings and their interaction with the market factor. The omitted group of AAA and AA bonds have a factor loading of 0.54. The factor loading increases by 0.17 for A bonds, 0.50 for BBB bonds, 1.27 for BB bonds, 1.65 for B bonds, and 2.65 for CCC and below bonds. The second column adds the distance to default and its interaction with the market factor. A standard deviation decrease in the distance to default increases the factor loading by 0.37.

The third column of Table 2 adds the bond characteristics, which increase the R^2 slightly from 65% to 68%. These bond characteristics have small economic effects. For example, a credit enhancement increases the credit spread by 9 basis points, and a bondholder protective covenant increases the credit spread by 7 basis points. This paper focuses on the variation in credit spreads across firms, which is the primary unit of analysis for credit ratings and default. We refer to Mota and Siani (2024) for a complementary and detailed analysis of variation in credit spreads within firms across bond characteristics.

We take the residuals from the regression in the third column of Table 2 and compute their standard deviation by credit rating category. The standard deviation is 0.56% for AAA and AA bonds, 0.66% for A bonds, 0.86% for BBB bonds, 1.56% for BB bonds, 2.50% for B bonds, and 5.58% for CCC and below bonds. Thus, the large standard deviation in credit spreads is robust to controlling for credit ratings, the distance to default, and a full set of bond characteristics.

The fourth column of Table 2 adds firm fixed effects, which increase the R^2 from 68% to 76%. Both Tables 1 and 2 show firm-level variation in credit spreads that credit ratings and

the distance to default do not fully capture. This finding suggests the presence of information about firms that is difficult to capture with observed characteristics. In Section 3, we develop an asset pricing model to guide our empirical exercise of extracting this information from bond holdings data.

3. Asset Pricing Model

We present an asset pricing model that guides our empirical approach. We derive reduced-form demand as the dot product of investor loadings (i.e., principal component loadings) and firm embeddings (i.e., principal component scores). Identification of firm embeddings requires sufficient cross-sectional variation in the investor loadings, which arises from heterogeneity in risk aversion, beliefs, risk constraints, and investment mandates in our model. Firm embeddings contain information that is difficult to capture with observed characteristic and help explain bond prices and risk measures.

3.1. Bond Market

There are two periods, indexed as t and $t+1$. There are N bonds, indexed by $n = 1, \ldots, N$. These bonds are issued by different firms, indexed by $f = 1, \ldots, F$. We denote the set of bonds issued by firm f as $\mathcal{F}$. We normalize the face value of each bond to \$1. Let $P_{n,f,t}$ be the price of bond n issued by firm f at time t. Let $B_{n,f,t}$ be the face value outstanding of bond n issued by firm f at time t. We use a bold letter to denote a vector that stacks the corresponding variable. For example, $\boldsymbol{P}_t$ is an N-dimensional vector of bond prices at time t.

Bond n issued by firm f pays off $D_{n,f,t+1}$ at time $t+1$, which could be less than \$1 because of default risk. The payoffs have a factor structure:

$$D_{n,f,t+1} = \phi_{n,f,t} + \psi_{n,f,t} M_{t+1} + \omega_{n,f,t+1}, \tag{2}$$

where $\phi_{n,f,t}$ is the expected payoff and $\psi_{n,f,t}$ is the factor loading. The market factor M_{t+1} has zero mean and unit variance. The idiosyncratic shock $\omega_{n,f,t+1}$ is uncorrelated with the market factor, is uncorrelated across bonds, and has zero mean and variance of σ^2. The assumption of constant idiosyncratic variance across bonds is inessential but simplifies the presentation.

Let $\boldsymbol{C}_{n,f,t}$ be a vector of observed bond and firm characteristics, including the credit rating and the distance to default. Let $\boldsymbol{Z}_{f,t}$ be a vector of firm characteristics that are observed by investors but unobserved by the econometrician. Thus, investors have information relevant

to default risk that is not fully reflected in credit ratings. The expected payoff of bond n issued by firm f at time t is

$$\phi_{n,f,t} = \boldsymbol{\Phi}_C' \boldsymbol{C}_{n,f,t} + \boldsymbol{\Phi}_Z' \boldsymbol{Z}_{f,t}. \tag{3}$$

For example, the expected payoff is increasing in the credit rating and decreasing in the distance to default. The factor loading of bond n issued by firm f at time t is

$$\psi_{n,f,t} = \boldsymbol{\Psi}_C' \boldsymbol{C}_{n,f,t} + \boldsymbol{\Psi}_Z' \boldsymbol{Z}_{f,t}. \tag{4}$$

For example, the factor loading is decreasing in the credit rating and increasing in the distance to default. Let $\boldsymbol{\psi}_t$ be an N-dimensional vector of factor loadings at time t.

In addition to the risky bonds, there is a riskless asset in perfectly elastic supply with a constant interest rate normalized to zero.

3.2. Investors

There are I investors, indexed by $i = 1, \ldots, I$. The investors choose an optimal portfolio of bonds at time t. Let $Q_{i,n,f,t}$ be the face value of bond n issued by firm f that investor i holds at time t. Let $O_{i,t}$ be investor i's dollar holding of the riskless asset at time t. The investor's wealth at time t is

$$A_{i,t} = \boldsymbol{P}_t' \boldsymbol{Q}_{i,t} + O_{i,t}. \tag{5}$$

The investor's wealth at time $t+1$ is

$$A_{i,t+1} = A_{i,t} + (\boldsymbol{D}_{t+1} - \boldsymbol{P}_t)' \boldsymbol{Q}_{i,t}. \tag{6}$$

We model heterogeneity in risk aversion, beliefs, risk constraints, and investment mandates to capture the full range of investors within and across sectors (i.e., mutual funds and insurance companies). Investors have mean-variance expected utility with heterogeneous risk aversion, where $\gamma_i > 0$ is investor i's coefficient of absolute risk aversion. Investors also have heterogeneous beliefs and agree to disagree. Investor i believes that the expected payoff of bond n issued by firm f at time t is

$$\phi_{i,n,f,t} = \boldsymbol{\Phi}_{i,C}' \boldsymbol{C}_{n,f,t} + \boldsymbol{\Phi}_{i,Z}' \boldsymbol{Z}_{f,t}. \tag{7}$$

For simplicity, we assume that the investors agree about the factor loadings and idiosyncratic

risk.[3] Let $\boldsymbol{\phi}_{i,t}$ be an N-dimensional vector of investor i's perceived factor loadings at time t.

Each investor solves a portfolio choice problem:

$$\max_{\boldsymbol{Q}_{i,t}} \ \mathbb{E}_{i,t}\left[A_{i,t+1}\right] - \frac{\gamma_i}{2}\mathrm{Var}_{i,t}\left(A_{i,t+1}\right) - \theta_i \boldsymbol{\psi}_t' \boldsymbol{Q}_{i,t}, \tag{8}$$

subject to the intertemporal budget constraint (6). We subscript the expectation and the variance by i to denote heterogeneous beliefs. The last term in the objective function (8) represents a risk constraint or an investment mandate. The parameter $\theta_i \geq 0$ captures the strength of the risk constraint or the investment mandate for investor i. For example, insurance companies face a value-at-risk constraint that penalizes portfolios with higher systematic risk, which can be equivalent to a lower weighted credit rating through equation (4).[4] Mutual funds face heterogeneous investment mandates, where investment-grade funds have higher θ_i than speculative-grade funds.

3.3. Equilibrium

We solve the portfolio choice problem (8) in Appendix A. Investor i's demand for bond n issued by firm f at time t is

$$Q_{i,n,f,t} = -\pi_{i,t} P_{n,f,t} + \boldsymbol{\kappa}_{i,t}' \boldsymbol{C}_{n,f,t} + \boldsymbol{\zeta}_{i,t}' \boldsymbol{Z}_{f,t}, \tag{9}$$

where

$$\pi_{i,t} = \frac{1}{\gamma_i \sigma^2}, \tag{10}$$

$$\boldsymbol{\kappa}_{i,t} = \frac{1}{\gamma_i \sigma^2}\left(\boldsymbol{\Phi}_{i,C} - \Theta_{i,t}\boldsymbol{\Psi}_C\right), \tag{11}$$

$$\boldsymbol{\zeta}_{i,t} = \frac{1}{\gamma_i \sigma^2}\left(\boldsymbol{\Phi}_{i,Z} - \Theta_{i,t}\boldsymbol{\Psi}_Z\right), \tag{12}$$

$$\Theta_{i,t} = \frac{\boldsymbol{\psi}_t'\left(\boldsymbol{\phi}_{i,t} - \boldsymbol{P}_t\right) + \theta_i \sigma^2}{\boldsymbol{\psi}_t'\boldsymbol{\psi}_t + \sigma^2}. \tag{13}$$

According to equation (9), asset demand decreases in the bond price $P_{n,f,t}$, increases in observed characteristics $\boldsymbol{C}_{n,f,t}$ that relate to a higher expected payoff or lower systematic

[3] We could generalize the model to allow for disagreement about the factor loadings and idiosyncratic risk (Koijen and Yogo, 2019). The factor loadings that enter the risk constraint (8) could still be homogeneous across investors under the assumption that rating agencies or regulators have different beliefs than the investors (Chaudhary, 2024).

[4] We refer to Koijen and Yogo (2022) for a microfoundation of heterogeneity in θ_i across insurance companies, based on their liability structure.

risk, and increases in unobserved firm characteristics $\boldsymbol{Z}_{f,t}$ that relate to a higher expected payoff or lower systematic risk. If beliefs were homogeneous and θ_i were constant across investors, the two-fund separation theorem applies, and investors hold identical portfolios of risky bonds. Therefore, heterogeneity in either beliefs or θ_i is necessary for nontrivial portfolio heterogeneity across investors. Investors with higher θ_i tilt their portfolio toward bonds with observed or unobserved characteristics that relate to lower systematic risk (e.g., higher credit ratings).

Market clearing for bond n issued by firm f at time t is

$$B_{n,f,t} = \sum_{i=1}^{I} Q_{i,n,f,t}.$$

Substituting asset demand (9), we solve for the equilibrium price as

$$P_{n,f,t} = \frac{1}{\overline{\pi}_t}\left(-B_{n,f,t} + \overline{\boldsymbol{\kappa}}_t' \boldsymbol{C}_{n,f,t} + \overline{\boldsymbol{\zeta}}_t' \boldsymbol{Z}_{f,t}\right), \tag{14}$$

where $\overline{\pi}_t = \sum_{i=1}^{I} \pi_{i,t}$, $\overline{\boldsymbol{\kappa}}_t = \sum_{i=1}^{I} \boldsymbol{\kappa}_{i,t}$, and $\overline{\boldsymbol{\zeta}}_t = \sum_{i=1}^{I} \boldsymbol{\zeta}_{i,t}$. The equilibrium price decreases in supply $B_{n,f,t}$, increases in observed characteristics $\boldsymbol{C}_{n,f,t}$ that relate to a higher expected payoff or lower systematic risk, and increases in unobserved firm characteristics $\boldsymbol{Z}_{f,t}$ that relate to a higher expected payoff or lower systematic risk.

3.4. Firm Embeddings

Substituting the equilibrium price (14) in asset demand (9), we have

$$Q_{i,n,f,t} = \frac{\pi_{i,t}}{\overline{\pi}_t} B_{n,f,t} + \left(\boldsymbol{\kappa}_{i,t} - \frac{\pi_{i,t}}{\overline{\pi}_t}\overline{\boldsymbol{\kappa}}_t\right)' \boldsymbol{C}_{n,f,t} + \left(\boldsymbol{\zeta}_{i,t} - \frac{\pi_{i,t}}{\overline{\pi}_t}\overline{\boldsymbol{\zeta}}_t\right)' \boldsymbol{Z}_{f,t}. \tag{15}$$

Aggregating across all bonds issued by firm f, we have

$$Q_{i,f,t} = \sum_{n \in \mathcal{F}} Q_{i,n,f,t} = \boldsymbol{\lambda}_{i,t}' \boldsymbol{x}_{f,t}, \tag{16}$$

where

$$\boldsymbol{\lambda}_{i,t} = \left[\tfrac{\pi_{i,t}}{\overline{\pi}_t}, \quad \left(\boldsymbol{\kappa}_{i,t} - \tfrac{\pi_{i,t}}{\overline{\pi}_t}\overline{\boldsymbol{\kappa}}_t\right)', \quad \left(\boldsymbol{\zeta}_{i,t} - \tfrac{\pi_{i,t}}{\overline{\pi}_t}\overline{\boldsymbol{\zeta}}_t\right)'\right]', \tag{17}$$

$$\boldsymbol{x}_{f,t} = \left[\textstyle\sum_{n \in \mathcal{F}} B_{n,f,t}, \quad \sum_{n \in \mathcal{F}} \boldsymbol{C}_{n,f,t}', \quad \boldsymbol{Z}_{f,t}'\right]'. \tag{18}$$

Equation (16) shows that a reduced-form representation of investor i's demand for bonds

issued by firm f is the dot product of an investor loading $\boldsymbol{\lambda}_{i,t}$ and a firm embedding $\boldsymbol{x}_{f,t}$. The firm embedding is a vector that contains all relevant information for the pricing of bonds issued by firm f, including observed characteristics (e.g., credit ratings and the distance to default) and unobserved firm characteristics $\boldsymbol{Z}_{f,t}$. The terminology embedding follows the artificial intelligence literature that represents words, images, and now firms with high dimensional vectors (Gabaix et al., 2024).

By principal component analysis, we can estimate (16) as the dot product of principal component loadings $\boldsymbol{\lambda}_{i,t}$ and principal component scores $\boldsymbol{x}_{f,t}$. If firm embeddings contained only observed characteristics, this estimation exercise would be an inefficient way to recover observed characteristics like credit ratings and the distance to default. Thus, the primary purpose of estimating firm embeddings is to recover information that is difficult to capture with observed characteristics.

Reduced-form demand (16) substitutes out the endogenous bond price. Therefore, estimation of reduced-form demand does not require instruments or comprehensive holdings data that satisfy market clearing in contrast to structural estimation of asset demand (9) in demand system asset pricing (Koijen and Yogo, 2019; Bretscher et al., 2023; Chaudhary et al., 2023). However, identification requires variation in the investor loadings $\boldsymbol{\lambda}_{i,t}$ through heterogeneity in risk aversion, beliefs, risk constraints, or investment mandates. For example, variation in bond holdings between insurance companies that are more regulated (i.e., higher θ_i) and mutual funds that are less regulated (i.e., lower θ_i) may be useful for identification. Compared with mutual funds, insurance companies may tilt their portfolios toward bonds with observed or unobserved characteristics that relate to lower systematic risk.

After estimating firm embeddings, we can estimate a cross-sectional regression model for bond prices:

$$P_{n,f,t} = \beta_B B_{n,f,t} + \boldsymbol{\beta}_C' \boldsymbol{C}_{n,f,t} + \boldsymbol{\beta}_x' \boldsymbol{x}_{f,t} + \epsilon_{n,f,t}. \tag{19}$$

According to equation (14), firm embeddings should explain bond prices better than observed characteristic alone. According to equation (4), firm embeddings should also explain risk measures better than observed characteristics alone. Equation (18) shows that the sum of bond characteristics for a given firm is a subvector of firm embeddings. Thus, firm embeddings are also useful controls for heterogeneity in bond characteristics (e.g., embedded options and covenants) that observed characteristics do not fully capture.

4. Estimating Firm Embeddings

We describe the estimation methodology and present summary statistics for the firm embeddings.

4.1. Estimation Methodology

We derived equation (16) in a linear asset pricing model to illustrate our empirical approach in a simple setting. In empirical implementation, we specify a loglinear demand function since the cross section of bond holdings are closer to a lognormal distribution. If the true model of portfolio choice has nonlinearities that arise for various reasons (e.g., portfolio constraints), the loglinear demand function is misspecified. However, the spirit of machine learning is that a sufficiently high dimension of firm embeddings could extract a lot of useful information for the sole purpose of prediction. Alternatively, we could move away from principal component analysis and consider nonlinear models (Gabaix et al., 2024), but we keep the methodology relatively simple in this paper.

We model investor i's log aggregate dollar holding of bonds issued by firm f at date t as

$$h_{i,f,t} = \boldsymbol{\lambda}_{i,t}' \boldsymbol{x}_{f,t} + \delta_{i,t} + \delta_{f,t} + \nu_{i,f,t}. \tag{20}$$

The first term on the right side is a dot product of the investor loadings $\boldsymbol{\lambda}_{i,t}$ and the firm embeddings $\boldsymbol{x}_{f,t}$. The next two terms are investor fixed effects $\delta_{i,t}$ and firm fixed effects $\delta_{f,t}$. The error term $\nu_{i,f,t}$ is uncorrelated with the investor loadings and the firm embeddings.

For each date, we prepare the bond holdings data as follows.

1. We winsorize both tails of log bond holdings at the 2.5 percentile.

2. We estimate the cross-sectional mean by firm and remove the firm fixed effects.

3. We estimate the cross-sectional mean by investor and remove the investor fixed effects.

4. We construct a balanced panel across investors and firms by filling zero holdings with the minimum observed value by investor (after taking out the fixed effects in steps 2 and 3). The working assumption is that a firm that an investor does not hold must not be more desirable than the smallest position that the investor does hold.

We estimate the investor loadings and the firm embeddings by principal component analysis. We set the dimension of firm embeddings to 64 elements.

Gabaix et al. (2024) find that the performance in various stock market benchmarks is robust to alternative procedures for preparing the stock holdings data. They consider filling zero holdings with the mean observed value (after taking out the investor fixed effects). They also consider ranked holdings instead of log holdings, filling zero holdings with a rank of zero.

4.2. Summary of Firm Embeddings

Figure 3 shows the cumulative share of the cross-sectional variance of bond holdings that the principal components explain. The figure shows the median and the interquartile range across all dates in the sample period. The first principal component alone explains 46% of the variation in bond holdings for the median date. Adding the second principal component increases the cumulative share of the cross-sectional variance to 54%. The fact that bond holdings have a strong factor structure is perhaps unsurprising, given the importance of credit ratings. Additional principal components steadily increase the cumulative share of the cross-sectional variance to 73% for all 64 elements.

According to the asset pricing model in Section 5, the firm embeddings contain all relevant information for bond prices. The firm embeddings serve a dual purpose of selecting the most relevant characteristics and extracting information that may be difficult to capture with observed characteristics. We choose the dimension 64 to be sufficiently high to capture most of the relevant information in bond holdings for our benchmarks in Section 5. We train the firm embeddings to explain credit spreads by ridge regression with cross validation. Gabaix et al. (2024) find good performance with this approach in various stock market benchmarks. A potential alternative is to estimate firm embeddings for each benchmark by supervised principal component analysis. We prefer our approach for its computational simplicity, provided that we find satisfactory performance.

5. Fixed Income Benchmarks

An important application of credit ratings is pricing and risk assessment of fixed income securities. We test whether the firm embeddings explain credit spreads, changes in credit spreads, and the volatility of credit spreads as three benchmarks around this application. In each benchmark, we compare the explanatory power of firm embeddings to that of credit ratings and the distance to default.

5.1. Explaining Credit Spreads

We train the firm embeddings to explain credit spreads. We model the weighted average credit spread for firm f at date t as

$$Y_{f,t} = \boldsymbol{\gamma}_t' \boldsymbol{x}_{f,t} Y_t + \eta_{f,t}, \tag{21}$$

where Y_t is the market factor (i.e., the value-weighted average of credit spreads) and $\eta_{f,t}$ is an error term. For each date t, we estimate the coefficients $\boldsymbol{\gamma}_t$ by ridge regression with cross

validation to prevent overfitting.[5] We estimate the regularization parameter by ten-fold cross validation. We use nine folds for training and the remaining fold for validation and choose the model that minimizes the mean squared error across the ten validation samples. We use the estimated coefficients to construct the trained embedding as

$$x_{f,t} = \boldsymbol{\gamma}_t' \boldsymbol{x}_{f,t}, \tag{22}$$

which is a linear combination of firm embeddings that best explains credit spreads.

In equation (21), the multiplication by the market factor is a normalization that serves two purposes. First, this normalization stabilizes the estimated coefficients across dates because credit spreads have a factor structure. Therefore, we can define a tighter range for the regularization parameter that works across dates, improving computational efficiency. Second, the trained embedding $x_{f,t}$ has an economic interpretation as a factor loading on the market factor. However, we cannot rule out the presence of multiple factors because the factors and the factor loadings are not separately identified.

We compare the explanatory power of the trained embeddings with that of credit ratings and the distance to default. We model the credit spread for bond n issued by firm f at date t as

$$Y_{n,f,t} = \left(\beta_Y + \boldsymbol{\beta}_{C,Y}' \boldsymbol{C}_{n,f,t} + \beta_{x,Y} x_{f,t}\right) Y_t + \boldsymbol{\beta}_C' \boldsymbol{C}_{n,f,t} + \boldsymbol{\beta}_D' \boldsymbol{D}_{n,f,t} + \epsilon_{n,f,t}. \tag{23}$$

The vector $\boldsymbol{C}_{n,f,t}$ contains indicator variables for credit ratings and the distance to default. The vector of bond characteristics $\boldsymbol{D}_{n,f,t}$ contains duration, log market value outstanding, and indicator variables for embedded options and covenants. $\epsilon_{n,f,t}$ is an error term. Regression (23) captures the factor structure in credit spreads, shown in Panel A of Figure 2.

The first column of Table 3 reports regression (23) with only the market factor and the bond characteristics. The second column adds credit ratings and their interaction with the market factor. We estimate the regression separately by credit rating category. We focus our discussion on the BBB rating category, which is the largest credit rating category by the number of observations, because the results are similar for the other credit rating categories.

[5] A potential concern with cross validation of cross-sectional models is cross-sectional correlation of the error term through a factor structure. This concern does not apply to our application. Under the null that the model is correctly specified, the firm embeddings remove the factor structure so that the residuals are cross-sectionally uncorrelated. Under the alternative that the model is misspecified (e.g., the firm embeddings are not sufficiently high dimensional), the use of cross validation is still valid for a prediction exercise. By construction, the residuals are uncorrelated with the firm embeddings, even if they retain some factor structure. We are ultimately interested in the explanatory power of this potentially misspecified model, even if we cannot interpret $\boldsymbol{\gamma}_t$ as a structural parameter (i.e., the causal impact of particular elements of the firm embedding on credit spreads).

In the BBB rating category, the omitted group of BBB+ bonds have a factor loading of 0.95. The factor loading increases by 0.06 for BBB bonds and by 0.30 for BBB− bonds. Credit ratings and their interaction with the market factor explain a large share of the variation in credit spreads with an R^2 of 51%. The third column adds the distance to default and its interaction with the market factor. The factor loading increases by 0.43 for a standard deviation decrease in the distance to default. The distance to default increases the R^2 from 51% to 60%.

The fourth column of Table 3 reports regression (23) with only the market factor, its interaction with the trained embedding, and the bond characteristics. The trained embeddings explain credit spreads with an R^2 of 67%, beating 60% for combination of credit ratings and the distance to default. The fifth column adds credit ratings to the fourth column to verify that the trained embeddings retain explanatory power. The trained embeddings remain statistically significant, and the R^2 increases from 67% to 68%. The sixth column adds credit ratings and the distance to default to the fourth column to verify that the trained embeddings retain explanatory power. The trained embeddings remain statistically significant, and the R^2 increases from 67% to 71%. These results confirm that the firm embeddings contain information about credit spreads that credit ratings and the distance to default do not.

Figure 4 summarizes the R^2 from Table 3. For each credit rating category, the first four bars report the R^2 corresponding to the first four columns of Table 3. For example, the R^2 is 46%, 51%, 60%, and 67% for the BBB rating category. Figure 4 also shows the R^2 for the speculative-grade rating categories. For completeness, we report the corresponding regressions for the speculative-grade rating categories in Appendix C. Across all credit rating categories, the trained embeddings have higher explanatory power than combination of credit ratings and the distance to default.

5.2. Placebo Embeddings

We trained the firm embeddings to explain credit spreads by ridge regression with cross validation to prevent overfitting. We design a placebo experiment to verify that the high explanatory power of the firm embeddings is genuine. We mimic the firm embeddings by generating a normally distributed random vector of dimension 64, which we call placebo embeddings. We train the placebo embeddings to explain credit spreads by ridge regression with cross validation. We then estimate regression (23) with only the market factor, its interaction with the trained placebo embedding, and the bond characteristics. That is, we repeat the same regression as the fourth column of (23), replacing the firm embeddings with the placebo embeddings.

We add the R^2 from the placebo experiment as a fifth bar in each panel of Figure 4. The

placebo experiment attains the same R^2 as the baseline specification with only the market factor and the bond characteristics, represented by the first bar. Thus, the trained placebo embedding does not add any explanatory power. By comparing the fourth bar with the fifth bar, we verify that the firm embeddings have genuine explanatory power that is not an artifact of its high dimension and overfitting.

5.3. Explaining Changes in Credit Spreads

As a second benchmark, we use the trained embeddings (22) to explain changes in credit spreads. We model the quarterly change in credit spreads for bond n issued by firm f from date t to $t+1$ as

$$\Delta Y_{n,f,t+1} = \left(\beta_Y + \boldsymbol{\beta}'_{C,Y}\boldsymbol{C}_{n,f,t} + \beta_{x,Y}x_{f,t}\right)\Delta Y_{t+1} + \epsilon_{n,f,t+1}, \tag{24}$$

where $\epsilon_{n,f,t+1}$ is an error term. This regression is equation (23) in first differences, assuming that the credit ratings, the distance to default, the trained embeddings, and the bond characteristics remain constant over time. Changes in credit spreads have a factor structure (Collin-Dufresne et al., 2001), and measures of dealer inventory and intermediary distress explain a large share of the factor (Friewald and Nagler, 2019; He et al., 2022). Regression (24) takes the factor structure as given and models the factor loadings as a function of the credit ratings, the distance to default, and the trained embeddings. This benchmark tests whether the trained embedding is a risk measure that explains comovement of changes in credit spreads.

The first column of Table 4 reports regression (24) with only the market factor and the bond characteristics. The second column adds credit ratings and their interaction with the market factor. We focus our discussion on the BBB rating category because the results are similar for the other credit rating categories. In the BBB rating category, the omitted group of BBB+ bonds have a factor loading of 0.82. The factor loading increases by 0.16 for BBB bonds and by 0.70 for BBB− bonds. Credit ratings and their interaction with the market factor explain a large share of changes in credit spreads with an R^2 of 44%. The third column adds the distance to default and its interaction with the market factor. The factor loading increases by 0.35 for a standard deviation decrease in the distance to default. The distance to default increases the R^2 from 44% to 47%.

The fourth column of Table 4 reports regression (24) with only the market factor, its interaction with the trained embedding, and the bond characteristics. The trained embeddings explain changes in credit spreads with an R^2 of 45%. The fifth column adds credit ratings to the fourth column to verify that the trained embeddings retain explanatory power.

The trained embeddings remain statistically significant, and the R^2 increases from 45% to 46%. The sixth column adds credit ratings and the distance to default to the fourth column to verify that the trained embeddings retain explanatory power. The trained embeddings remain statistically significant, and the R^2 increases from 45% to 48%. These results confirm that the firm embeddings contain information about changes in credit spreads that credit ratings and the distance to default do not.

Figure 5 summarizes the R^2 from Table 4. For each credit rating category, the first four bars report the R^2 corresponding to the first four columns of Table 4, and the fifth bar reports the placebo experiment. Figure 4 also shows the R^2 for the speculative-grade rating categories. For completeness, we report the corresponding regressions for the speculative-grade rating categories in Appendix C. Across all credit rating categories, the trained embeddings have higher explanatory power than credit ratings but lower explanatory power than the combination of credit ratings and the distance to default.

5.4. Explaining the Volatility of Credit Spreads

As a third benchmark, we use the trained embeddings (22) to explain the volatility of credit spreads. We model log volatility of credit spreads for bond n issued by firm f over a quarterly period between dates t and $t+1$ as

$$v_{n,f,t+1} = \rho v_{n,f,t} + \left(\beta_Y + \boldsymbol{\beta}'_{C,Y}\boldsymbol{C}_{n,f,t} + \beta_{x,Y}x_{f,t}\right) Y_t + \boldsymbol{\beta}'_C\boldsymbol{C}_{n,f,t} + \boldsymbol{\beta}'_D\boldsymbol{D}_{n,f,t} + \epsilon_{n,f,t+1}, \quad (25)$$

where $\epsilon_{n,f,t+1}$ is an error term. This regression replaces the credit spread in equation (23) with the lead value of log volatility and adds current log volatility as a regressor to model its persistence. Regression (25) captures the factor structure in the volatility of credit spreads, shown in Panel B of Figure 2. This benchmark tests whether the trained embedding is a risk measure that explains comovement of the volatility of credit spreads.

The first column of Table 5 reports regression (25) with only the market factor and the bond characteristics. The second column adds credit ratings and their interaction with the market factor. We focus our discussion on the BBB rating category because the results are similar for the other credit rating categories. The volatility of credit spreads is persistent with an autoregressive coefficient of 0.39. Credit ratings and their interaction with the market factor explain the volatility of credit spreads with an R^2 of 42%. The third column adds the distance to default and its interaction with the market factor. The distance to default increases the R^2 from 42% to 43%.

The fourth column of Table 5 reports regression (25) with only the market factor, its interaction with the trained embeddings, and the bond characteristics. The trained embed-

dings explain the volatility of credit spreads with an R^2 of 44%. The fifth column adds credit ratings to the fourth column to verify that the trained embeddings retain explanatory power. The trained embeddings remain statistically significant, and the R^2 remains 44%. The sixth column adds credit ratings and the distance to default to the fourth column to verify that the trained embeddings retain explanatory power. The trained embeddings remain statistically significant, and the R^2 remains 44%. These results confirm that the firm embeddings contain information about the volatility of in credit spreads that credit ratings and the distance to default do not.

Figure 6 summarizes the R^2 from Table 5. For each credit rating category, the first four bars report the R^2 corresponding to the first four columns of Table 5, and the fifth bar reports the placebo experiment. Figure 6 also shows the R^2 for the speculative-grade rating categories. For completeness, we report the regressions for the speculative-grade rating categories in Appendix C. With a minor exception of the highest credit rating category, the trained embeddings have higher explanatory power than combination of credit ratings and the distance to default.

6. Understanding Firm Embeddings

We present additional evidence to help us understand the information content of the firm embeddings. First, we find that the firm embeddings capture low-frequency risk, predicting rating downgrades and default. Second, we find systematic differences in the investor loadings between mutual funds and insurance companies, which help identify the firm embeddings.

6.1. Predicting Low-Frequency Risk

The trained embedding (21) is a linear combination of firm embeddings that explains credit spreads better than the combination of credit ratings and the distance to default. Moreover, the three benchmarks show that the trained embedding is a risk measure that explains comovement in credit spreads, changes in credit spreads, and the volatility of credit spreads. The results suggest that investors have information about differences in risk across firms that credit ratings and the distance to default do not fully capture. We look for low-frequency evidence that further supports this interpretation, by testing whether the trained embeddings predict rating downgrades and default.

We first examine rating downgrades. Among firms that are currently rated investment grade (i.e., BBB− and above), we construct a downgrade indicator as one if a firm is downgraded to speculative grade (i.e., BB+ and below) over the subsequent year. In Table 6, we estimate a logit model to predict the downgrade indicator. In the first column, we start with

indicator variables for credit rating categories and rating watch indicators, which attain a pseudo R^2 of 16%. In the second column, we add the distance to default, which increases the pseudo R^2 from 16% to 24%. In the third column, we add the firm-level weighted average credit spread, which increases the pseudo R^2 from 24% to 27%. In the fourth column, we instead add the trained embedding, which increases the pseudo R^2 from 24% to 26%. In the fifth column, we include both the credit spread and the trained embedding, and they are both statistically significant predictors of rating downgrades.

We next examine default. Among firms that are currently rated investment grade (i.e., BBB− and above), we construct a default indicator if a firm defaults on any of its bonds over the subsequent five years. In Table 7, we estimate a logit model to predict the default indicator. In the first column, we start with indicator variables for credit rating categories and rating watch indicators, which attain a pseudo R^2 of 3%. In the second column, we add the distance to default, which increases the pseudo R^2 from 3% to 13%. In the third column, we add the firm-level weighted average credit spread, which is a statistically insignificant predictor of default. In the fourth column, we instead add the trained embedding, which is a statistically significant predictor of default. In the fifth column, we include both the credit spread and the trained embedding, neither of which are statistically significant predictors of default.[6]

This evidence suggests that the trained embedding is a risk measure that relates to expected rating downgrades and default. However, we view the evidence in this section as only suggestive, given the relatively short sample period. A much longer sample period is necessary to accurately estimate default rates (Feldhütter and Schaefer, 2018).

6.2. Investor Loadings

In the asset pricing model in Section 3, cross-sectional variation in the investor loadings arises from heterogeneity in risk aversion, beliefs, risk constraints, and investment mandates. Moreover, this variation in the investor loadings is important for the identification of the firm embeddings. We introduce simple statistics that summarize the variation in the investor loadings.

Let $H_{i,f,t}$ be investor i's dollar holding of bonds issued by firm f at date t. Let $M_{f,t}$ be the total market value of bonds issued by firm f at date t. For each investor i at date t, we compute the holdings-weighted average of the trained embeddings (21) minus the market-

[6]According to the asset pricing model in Section 3, equilibrium bond prices (14) depend on the firm embeddings. Therefore, it is not surprising that credit spreads and the trained embeddings do not have independent explanatory power.

weighted average as

$$\overline{x}_{i,t} = \frac{\sum_{f=1}^{F} H_{i,f,t} x_{f,t}}{\sum_{f=1}^{F} H_{i,f,t}} - \frac{\sum_{f=1}^{F} M_{f,t} x_{f,t}}{\sum_{f=1}^{F} M_{f,t}}. \tag{26}$$

This statistic tells us whether investor i tilts its portfolio toward firms that have embeddings associated with a higher credit spread.

Panel A of Figure 7 shows the median (solid) and the interquartile range (dots) of the portfolio-weighted trained embeddings (26) by sector. The unit of the vertical axis is a standard deviation of credit spreads by date and credit rating category. The median mutual fund tilts its portfolio toward firms that have embeddings associated with a higher credit spread, compared with the median insurance company. This finding is consistent with insurance companies choosing more conservative portfolios due to risk-based capital regulation. The wide interquartile range implies significant variation in how much investors tilt toward bonds with higher credit spreads.

7. An Improved Rating System

The results in Section 5 show that firm embeddings can augment credit ratings to provide more timely and accurate information for fixed income markets. Credit ratings that are inaccurate can create perverse incentives and weaken market discipline. Insurance companies may hold onto bonds with inaccurately high ratings or avoid purchasing bonds with inaccurately low ratings (Becker and Ivashina, 2015; Ellul et al., 2015; Becker et al., 2022). Since credit ratings determine risk-based capital through the risk weights, inaccurate ratings can misguide investors and regulators about the financial health of insurance companies.

As a proof of concept, we use the firm embeddings to improve credit ratings. Based on regression (23), we have estimated factor loadings $\beta_Y + \boldsymbol{\beta}'_{C,Y}\boldsymbol{C}_{n,f,t} + \beta_{x,Y}x_{f,t}$ for each bond at each date. For each date, we rank the bonds by these factor loadings and assign counterfactual ratings that have the same distribution as the actual ratings. Table 8 reports the joint distribution of the actual ratings and the counterfactual ratings, pooled over the sample period. Under the actual rating system, 42% of bonds are rated BBB. Under the improved rating system, 29% would remain BBB, 1% would be upgraded to AAA or AA, 9% would be upgraded to A, and 3% would be downgraded to BB.

For each insurance company, we compute the required capital as the sum of each bond holding times the risk weight corresponding to its credit rating.[7] We also compute the average risk weight as the required capital divided by the total bond holding. Panel A of

[7]The risk weights are 0.39% for A and above, 1.26% for BBB, 4.46% for BB, 9.70% for B, and 22.31% for CCC and below (Levy et al., 2021, p. 7).

Figure 8 is a scatter plot of log required capital under the actual versus the counterfactual ratings. Above the 45 degree line are insurance companies whose required capital increases under the improved rating system. The mean absolute change in log required capital is 0.16. Under the improved rating system, the average insurance company would have to increase or decrease its equity by 16 percentage points to maintain the same risk-based capital ratio. Panel B is a scatter plot of the average risk weight under the actual versus the counterfactual ratings. The mean absolute change in the average risk weight is 0.20%.

Insurance companies and regulators prefer capital requirements that are insensitive to short-term fluctuations in asset prices, given the long-term nature of insurance liabilities. From this perspective, credit ratings that are slow to adjust may be desirable to prevent excessive changes in risk-based capital. We show that the improved rating system retains this feature of the actual rating system. Figure 9 shows the distribution of annual changes in the average risk weight under the actual versus the counterfactual ratings. The distributions are nearly identical, which implies that the annual changes in required capital are similar under the improved rating system.

Our counterfactual exercise corresponds to a relatively modest proposal to improve the rating system, holding the distribution of credit ratings constant. A more radical proposal is to create a new rating system by artificial intelligence, based on firm characteristics and firm embeddings. A new rating system need not be ordinal but could be cardinal, tied to risk measures such as factor loadings and volatility. Thus, credit ratings would be a more direct input into economically relevant risk measures for insurance companies such as value at risk. Given the current dominance of credit ratings in fixed income markets, the transition to a new rating system should be incremental with continuous input from regulators and market participants.

8. Conclusion

We use the corporate bond holdings of mutual funds and insurance companies to estimate firm embeddings, which are a high dimensional representation of information that investors use to form their portfolios. We then train the firm embeddings to explain credit spreads by ridge regression with cross validation. Across all credit rating categories, the trained embeddings explain credit spreads better than the combination of credit ratings and the distance to default. We use the same trained embeddings to explain changes in credit spreads and the volatility of credit spreads as two additional benchmarks. The trained embeddings explain comovement in these two outcome variables that credit ratings and the distance to default do not explain. Therefore, firm embeddings can augment credit ratings

to provide more timely and accurate information for pricing and risk assessment of fixed income securities.

Artificial intelligence is transforming the financial industry by automating labor intensive tasks. Credit analysis is a prime example of such tasks. The critical input for artificial intelligence is big data, and this paper shows that bond holdings data could play that role in credit analysis. Asset pricing theory implies that bond holdings data contain all relevant information for pricing and risk assessment. Thus, an interpretation of firm embeddings is that it is like crowd sourcing credit analysis to a large group of institutional investors. Imagine a fixed income market in which index providers define benchmarks and regulators define risk weights, based on risk measures such as factor loadings and volatility. By estimating these risk measures as a function of firm and bond characteristics and firm embeddings, index providers and regulators could cease dependence on credit ratings.

Firm embeddings could have other applications in fixed income markets. Bond underwriters could use firm embeddings to price bonds and to find potential buyers, based on a comparison group of bonds issued by similar firms. Broker-dealers could use firm embeddings for matrix pricing in the secondary market, based on a comparison group of traded bonds. Research on the predictability of corporate bond returns could find higher performance by using firm embeddings in addition to bond and firm characteristics (Bali et al., 2022; Bell et al., 2024). Firm embeddings, which relate to the systematic risk in credit spreads, could be useful for predicting firm-level investment and macroeconomic activity (Gilchrist and Zakrajšek, 2012). We leave these and other potential applications for future research.

References

Bali, Turan G., Amit Goyal, Dashan Huang, Fuwei Jiang, and Quan Wen, "Predicting Corporate Bond Returns: Merton Meets Machine Learning," 2022. Working paper.

Becker, Bo and Victoria Ivashina, "Reaching for Yield in the Bond Market," *Journal of Finance*, 2015, *70* (5), 1863–1902.

_ **, Marcus M. Opp, and Farzad Saidi**, "Regulatory Forbearance in the U.S. Insurance Industry: The Effects of Removing Capital Requirements for an Asset Class," *Review of Financial Studies*, 2022, *35* (12), forthcoming.

Bell, Sebastian, Ali Kakhbod, Martin Lettau, and Abdolreza Nazemi, "Glass Box Machine Learning and Corporate Bond Returns," 2024. NBER Working Paper 33320.

Bretscher, Lorenzo, Lukas Schmid, Ishita Sen, and Varun Sharma, "Institutional Corporate Bond Pricing," 2023. Working paper.

Campbell, John Y. and Glen B. Taksler, "Equity Volatility and Corporate Bond Yields," *Journal of Finance*, 2003, *58* (6), 2321–2350.

_ **, Jens Hilscher, and Jan Szilagyi**, "In Search of Distress Risk," *Journal of Finance*, 2008, *63* (6), 2899–2939.

Center for Research in Security Prices, *CRSP-Compustat Linking Table* 2024.

_ , *Stock Events - Names History* 2024.

Chaudhary, Manav, "Regulator Beliefs," 2024. Working paper.

_ **, Zhiyu Fu, and Jian Li**, "Corporate Bond Multipliers: Substitutes Matter," 2023. Working paper.

Collin-Dufresne, Pierre, Robert S. Goldstein, and J. Spencer Martin, "The Determinants of Credit Spread Changes," *Journal of Finance*, 2001, *56* (6), 2177–2207.

Dick-Nielsen, Jens, "How to Clean Enhanced TRACE Data," 2014. Working paper.

Ellul, Andrew, Chotibhak Jotikasthira, Christian T. Lundblad, and Yihui Wang, "Is Historical Cost Accounting a Panacea? Market Stress, Incentive Distortions, and Gains Trading," *Journal of Finance*, 2015, *70* (6), 2489–2538.

Feldhütter, Peter and Stephen M. Schaefer, "The Myth of the Credit Spread Puzzle," *Review of Financial Studies*, 2018, *31* (8), 2897–2942.

FINRA, *TRACE Enhanced Historical Data* 2024.

Friewald, Nils and Florian Nagler, "Over-the-Counter Market Frictions and Yield Spread Changes," *Journal of Finance*, 2019, *74* (6), 3217–3257.

Gabaix, Xavier, Ralph S. J. Koijen, Robert J. Richmond, and Motohiro Yogo, "Asset Embeddings," 2024. Working paper.

Gilchrist, Simon and Egon Zakrajšek, "Credit Spreads and Business Cycle Fluctuations," *American Economic Review*, 2012, *102* (4), 1692–1720.

Gürkaynak, Refet S., Brian Sack, and Jonathan H. Wright, "The U.S. Treasury Yield Curve: 1961 to the Present," *Journal of Monetary Economics*, 2007, *54* (8), 2291–2304.

He, Zhiguo, Paymon Khorrami, and Zhaogang Song, "Commonality in Credit Spread Changes: Dealer Inventory and Intermediary Distress," *Review of Financial Studies*, 2022, *35* (10), 4630–4673.

Huang, Jing-Zhi and Ming Huang, "How Much of the Corporate-Treasury Yield Spread Is Due to Credit Risk?," 2012, *2* (2), 153–202.

Jensen, Theis Ingerslev, Bryan T. Kelly, and Lasse Heje Pedersen, "Is There a Replication Crisis in Finance?," *Journal of Finance*, 2023, *78* (5), 2465–2518.

Koijen, Ralph S. J. and Motohiro Yogo, "A Demand System Approach to Asset Pricing," *Journal of Political Economy*, 2019, *127* (4), 1475–1515.

_ and _, "The Fragility of Market Risk Insurance," *Journal of Finance*, 2022, *77* (2), 815–862.

Levy, Amnon, Pierre Xu, Andy Zhang, Akshay Gupta, Libor Pospisil, Mark Li, and Kamal Kumar, "Revisions to the RBC C1 Bond Factors Prepared for the and the NAIC and the ACLI," 2021. Moody's Analytics.

Liu, Peter, Zhong Zhuang, Douglas Dwyer, Yukyung Choi, and Samuel Malone, "Moody's Analytics EDF-Based Bond Valuation Model Version 2.0," 2020. Moody's Analytics Modeling Methodology.

Mergent, *Fixed Income Securities Database* 2024.

Merton, Robert C., "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates," *Journal of Finance*, 1974, *29* (2), 449.

Moody's Investor Service, "Rating Methodology: Manufacturing," 2021. Moody's Investors Service Rating Methodology.

Mota, Lira and Kerry Y. Siani, "Financially Sophisticated Firms," 2024. Working paper.

S&P Global, *Capital IQ Linking Tables* 2024.

_ , "S&P Global Ratings Corporate and Infrastructure Finance Criteria," 2024. Corporate Methodology.

Thomson Reuters, *eMAXX Bond Holdings Data* 2024.

Wharton Research Data Services, "WRDS Corporate Bond Database: Data Overview and Construction Manual," 2017. Working paper.

Table 1. Credit Spreads by Credit Rating

		Standard deviation		
Rating	Mean	Pooled	Cross-sectional	Observations
AAA	0.68	0.60	0.40	745
AA+	0.94	0.70	0.33	344
AA	1.05	0.84	0.53	1,440
AA-	1.10	0.81	0.54	2,813
A+	1.16	0.88	0.50	5,497
A	1.30	1.04	0.65	9,001
A-	1.51	1.21	0.73	9,754
BBB+	1.74	1.18	0.71	12,595
BBB	1.94	1.27	0.74	15,145
BBB-	2.34	1.52	0.90	11,460
BB+	3.26	2.04	1.20	4,239
BB	3.82	2.08	1.23	3,828
BB-	4.11	2.20	1.33	4,426
B+	4.80	2.67	1.61	4,339
B	5.29	2.89	1.80	3,971
B-	6.16	3.66	2.34	4,276
CCC+ and below	9.67	6.20	4.66	3,086

This table reports the pooled mean and standard deviation of firm-level credit spreads by credit rating. The firm-level credit spread is an average credit spread over the firm's bonds, weighted by market value outstanding. The cross-sectional standard deviation by date and credit rating is averaged over the sample period. The quarterly sample period covers September 2002 to December 2022.

Table 2. Credit Spreads and Observed Characteristics

	(1)	(2)	(3)	(4)
Market	0.54	0.60	0.66	0.67
	(0.06)	(0.04)	(0.04)	(0.04)
Rating × Market:				
A	0.17	0.24	0.20	0.17
	(0.04)	(0.03)	(0.03)	(0.03)
BBB	0.50	0.51	0.46	0.44
	(0.10)	(0.08)	(0.08)	(0.08)
BB	1.27	1.03	0.97	0.97
	(0.20)	(0.19)	(0.19)	(0.18)
B	1.65	1.32	1.23	1.28
	(0.23)	(0.22)	(0.22)	(0.22)
CCC and below	2.65	2.17	1.99	2.38
	(0.33)	(0.31)	(0.28)	(0.27)
Distance to default × Market		-0.37	-0.37	-0.37
		(0.03)	(0.03)	(0.03)
Rating:				
A	-0.05	-0.23	-0.23	-0.15
	(0.07)	(0.05)	(0.05)	(0.06)
BBB	-0.04	-0.21	-0.30	-0.45
	(0.18)	(0.14)	(0.15)	(0.15)
BB	0.17	0.22	0.27	-0.26
	(0.34)	(0.33)	(0.36)	(0.35)
B	1.00	1.11	2.13	1.16
	(0.41)	(0.40)	(0.47)	(0.48)
CCC and below	3.30	3.56	7.06	4.43
	(0.66)	(0.63)	(0.67)	(0.69)
Distance to default		0.37	0.31	0.26
		(0.05)	(0.05)	(0.05)
Duration			0.06	0.06
			(0.00)	(0.00)
Rating × Duration:				
A			0.00	0.00
			(0.00)	(0.00)
BBB			0.01	0.02
			(0.01)	(0.00)
BB			-0.00	0.01
			(0.01)	(0.01)
B			-0.19	-0.15
			(0.03)	(0.03)
CCC and below			-0.74	-0.58
			(0.06)	(0.06)
Amount outstanding			-0.19	-0.24
			(0.02)	(0.03)
Callable			-0.01	-0.04
			(0.03)	(0.02)
Putable			-0.03	0.02
			(0.04)	(0.03)
Credit enhancement			0.09	0.01
			(0.02)	(0.02)
Covenants:				
Bondholder protective			0.07	-0.02
			(0.01)	(0.01)
Issuer restrictive			0.02	0.02
			(0.01)	(0.01)
Subsidiary restrictive			-0.05	-0.08
			(0.01)	(0.01)
Issuer FE				Yes
R^2	0.62	0.65	0.68	0.76
Observations	360,487	360,487	360,487	360,487

The market factor is the value-weighted average of credit spreads. The omitted credit rating category is AAA and AA. The distance to default is standardized in the cross section of firms. In columns (3) and (4), the bond characteristics are duration and its interaction with credit rating categories, log market value outstanding, indicator variables for embedded options (i.e., callable, putable, and credit enhancement), and indicator variables for covenants (i.e., bondholder protective, issuer restrictive, and subsidiary restrictive). The coefficients for the bond characteristics are reported in percentage points. The standard errors in parentheses are robust to clustering by date. The quarterly sample period covers September 2002 to December 2022.

Table 3. Credit Spreads on Investment-Grade Bonds

Panel A. Rated AAA and AA						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	0.57	0.39	0.46	0.55	0.48	0.52
	(0.05)	(0.08)	(0.09)	(0.05)	(0.07)	(0.08)
Rating × Market:						
AA+		0.12	0.07		0.04	0.02
		(0.07)	(0.09)		(0.05)	(0.06)
AA		0.22	0.20		0.06	0.08
		(0.06)	(0.09)		(0.04)	(0.05)
AA−		0.23	0.21		0.10	0.11
		(0.04)	(0.06)		(0.03)	(0.04)
Distance to default × Market			-0.17			-0.12
			(0.03)			(0.02)
Embedding × Market				1.04	1.00	0.78
				(0.08)	(0.09)	(0.06)
Rating:						
AA+		-0.03	0.00		-0.00	0.04
		(0.11)	(0.15)		(0.08)	(0.11)
AA		-0.14	-0.15		-0.05	-0.06
		(0.11)	(0.16)		(0.06)	(0.09)
AA−		-0.16	-0.19		-0.10	-0.10
		(0.07)	(0.10)		(0.04)	(0.07)
Distance to default			0.20			0.16
			(0.05)			(0.04)
R^2	0.54	0.56	0.66	0.69	0.69	0.72
Observations	29,601	29,601	29,601	29,601	29,601	29,601

Panel B. Rated A						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	0.72	0.69	0.84	0.80	0.75	0.81
	(0.03)	(0.05)	(0.05)	(0.03)	(0.03)	(0.03)
Rating × Market:						
A		-0.03	-0.05		0.03	0.01
		(0.04)	(0.03)		(0.02)	(0.02)
A−		0.14	0.09		0.10	0.09
		(0.06)	(0.05)		(0.04)	(0.03)
Distance to default × Market			-0.31			-0.15
			(0.05)			(0.02)
Embedding × Market				0.96	0.93	0.78
				(0.06)	(0.06)	(0.04)
Rating:						
A		0.10	0.14		0.01	0.04
		(0.07)	(0.05)		(0.03)	(0.03)
A−		0.03	0.08		-0.03	-0.00
		(0.09)	(0.08)		(0.06)	(0.05)
Distance to default			0.41			0.18
			(0.08)			(0.04)
R^2	0.46	0.48	0.58	0.65	0.66	0.67
Observations	116,955	116,955	116,955	116,955	116,955	116,955

Panel C. Rated BBB						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	1.05	0.95	1.09	1.11	1.04	1.11
	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)	(0.05)
Rating × Market:						
BBB		0.06	0.00		0.05	0.01
		(0.07)	(0.04)		(0.04)	(0.03)
BBB−		0.30	0.16		0.18	0.11
		(0.08)	(0.06)		(0.05)	(0.05)
Distance to default × Market			-0.43			-0.26
			(0.04)			(0.02)
Embedding × Market				0.99	0.92	0.79
				(0.03)	(0.03)	(0.02)
Rating:						
BBB		0.12	0.15		0.03	0.07
		(0.12)	(0.07)		(0.07)	(0.05)
BBB−		0.15	0.23		0.05	0.13
		(0.13)	(0.11)		(0.08)	(0.08)
Distance to default			0.46			0.30
			(0.07)			(0.04)
R^2	0.46	0.51	0.60	0.67	0.68	0.71
Observations	152,327	152,327	152,327	152,327	152,327	152,327

The market factor is the value-weighted average of credit spreads. The omitted credit rating is AAA in Panel A, A+ in Panel B, and BBB+ in Panel C. The distance to default is standardized in the cross section of firms. The trained embedding is a linear combination of firm embeddings that best explains credit spreads. All specifications include duration, log market value outstanding, indicator variables for embedded options (i.e., callable, putable, and credit enhancement), and indicator variables for covenants (i.e., bondholder protective, issuer restrictive, and subsidiary restrictive). The coefficients for these bond characteristics are not reported for brevity. The standard errors in parentheses are robust to clustering by date. The quarterly sample period covers September 2002 to December 2022.

Table 4. Changes in Credit Spreads on Investment-Grade Bonds

Panel A. Rated AAA and AA						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	0.59	0.37	0.55	0.61	0.53	0.58
	(0.12)	(0.10)	(0.19)	(0.13)	(0.20)	(0.21)
Rating × Market:						
AA+		0.05	-0.04		-0.00	-0.05
		(0.09)	(0.14)		(0.12)	(0.14)
AA		0.14	0.10		-0.02	0.04
		(0.09)	(0.12)		(0.18)	(0.14)
AA−		0.37	0.29		0.19	0.23
		(0.11)	(0.08)		(0.07)	(0.08)
Distance to default × Market			-0.19			-0.16
			(0.10)			(0.08)
Embedding × Market				1.32	1.18	0.46
				(0.81)	(0.84)	(0.48)
R^2	0.29	0.31	0.35	0.32	0.33	0.36
Observations	27,732	27,732	27,732	27,732	27,732	27,732

Panel B. Rated A						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	0.68	0.66	0.79	0.72	0.70	0.80
	(0.03)	(0.10)	(0.11)	(0.05)	(0.12)	(0.12)
Rating × Market:						
A		-0.02	-0.04		-0.00	-0.02
		(0.11)	(0.09)		(0.09)	(0.08)
A−		0.10	0.06		0.06	0.04
		(0.13)	(0.11)		(0.14)	(0.13)
Distance to default × Market			-0.24			-0.20
			(0.07)			(0.04)
Embedding × Market				0.78	0.76	0.47
				(0.40)	(0.43)	(0.35)
R^2	0.34	0.34	0.38	0.36	0.36	0.39
Observations	105,529	105,529	105,529	105,529	105,529	105,529

Panel C. Rated BBB						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	1.06	0.82	0.97	1.12	0.93	1.01
	(0.08)	(0.04)	(0.04)	(0.08)	(0.05)	(0.05)
Rating × Market:						
BBB		0.16	0.07		0.10	0.05
		(0.04)	(0.03)		(0.03)	(0.03)
BBB−		0.70	0.51		0.53	0.43
		(0.18)	(0.17)		(0.20)	(0.17)
Distance to default × Market			-0.35			-0.27
			(0.02)			(0.02)
Embedding × Market				0.97	0.81	0.56
				(0.11)	(0.16)	(0.14)
R^2	0.41	0.44	0.47	0.45	0.46	0.48
Observations	137,579	137,579	137,579	137,579	137,579	137,579

The market factor is the value-weighted average quarterly change in credit spreads. The omitted credit rating is AAA in Panel A, A+ in Panel B, and BBB+ in Panel C. The distance to default is standardized in the cross section of firms. The trained embedding is a linear combination of firm embeddings that best explains credit spreads. The standard errors in parentheses are robust to clustering by date. The quarterly sample period covers September 2002 to December 2022.

Table 5. Volatility of Credit Spreads on Investment-Grade Bonds

Panel A. Rated AAA and AA						
	(1)	(2)	(3)	(4)	(5)	(6)
Volatility	0.41	0.41	0.39	0.38	0.38	0.37
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Market	20.87	15.93	19.79	21.76	19.47	21.40
	(4.81)	(5.30)	(5.04)	(4.97)	(5.56)	(5.10)
Rating × Market:						
AA+		3.64	1.77		1.94	0.98
		(3.48)	(4.39)		(3.44)	(4.16)
AA		4.13	3.60		0.46	1.39
		(2.33)	(3.21)		(2.40)	(2.70)
AA−		7.93	7.22		4.85	5.37
		(3.18)	(3.17)		(2.83)	(2.86)
Distance to default × Market			-6.85			-5.77
			(2.87)			(2.94)
Embedding × Market				24.63	25.86	15.36
				(5.77)	(6.37)	(4.98)
Rating:						
AA+		-7.10	-5.22		-6.68	-4.69
		(7.70)	(9.26)		(7.52)	(8.94)
AA		-8.69	-8.64		-6.86	-7.06
		(4.68)	(6.50)		(4.17)	(5.62)
AA−		-13.75	-14.15		-12.55	-12.67
		(5.70)	(6.10)		(5.09)	(5.58)
Distance to default			9.31			8.51
			(6.30)			(6.47)
R^2	0.45	0.45	0.47	0.46	0.46	0.47
Observations	28,785	28,785	28,785	28,785	28,785	28,785

Panel B. Rated A						
	(1)	(2)	(3)	(4)	(5)	(6)
Volatility	0.40	0.39	0.38	0.36	0.36	0.35
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Market	21.31	21.00	24.05	24.98	24.12	24.64
	(5.25)	(5.75)	(5.41)	(5.18)	(5.58)	(5.26)
Rating × Market:						
A		0.98	0.64		2.16	1.96
		(1.66)	(1.54)		(1.59)	(1.54)
A−		0.44	-0.14		-0.14	-0.14
		(1.88)	(1.77)		(1.80)	(1.79)
Distance to default × Market			-4.89			-1.36
			(1.03)			(1.14)
Embedding × Market				20.31	19.90	17.42
				(2.77)	(2.76)	(2.71)
Rating:						
A		-0.82	-0.56		-2.51	-2.56
		(2.94)	(2.76)		(2.86)	(2.77)
A−		6.14	6.25		5.13	4.54
		(3.28)	(3.24)		(3.08)	(3.21)
Distance to default			4.01			-1.16
			(2.15)			(2.43)
R^2	0.43	0.43	0.44	0.44	0.45	0.45
Observations	112,022	112,022	112,022	112,022	112,022	112,022

Panel C. Rated BBB						
	(1)	(2)	(3)	(4)	(5)	(6)
Volatility	0.40	0.39	0.36	0.35	0.34	0.33
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Market	23.27	23.79	26.38	27.51	28.34	28.76
	(6.64)	(6.48)	(6.47)	(6.47)	(6.36)	(6.34)
Rating × Market:						
BBB		0.46	0.07		0.04	0.17
		(0.98)	(0.85)		(0.77)	(0.78)
BBB−		0.40	-0.65		-2.34	-1.96
		(1.62)	(1.53)		(1.43)	(1.47)
Distance to default × Market			-4.23			-0.69
			(0.81)			(0.72)
Embedding × Market				19.90	18.81	17.00
				(2.21)	(2.05)	(1.97)
Rating:						
BBB		3.72	2.57		2.54	1.24
		(2.13)	(1.85)		(1.84)	(1.76)
BBB−		14.68	12.61		14.01	11.19
		(3.49)	(3.23)		(3.09)	(3.04)
Distance to default			-0.28			-3.92
			(1.75)			(1.76)
R^2	0.41	0.42	0.43	0.44	0.44	0.44
Observations	145,842	145,842	145,842	145,842	145,842	145,842

The market factor is the value-weighted average of credit spreads. The omitted credit rating is AAA in Panel A, A+ in Panel B, and BBB+ in Panel C. The distance to default is standardized in the cross section of firms. The trained embedding is a linear combination of firm embeddings that best explains credit spreads. All specifications include duration, log market value outstanding, indicator variables for embedded options (i.e., callable, putable, and credit enhancement), and indicator variables for covenants (i.e., bondholder protective, issuer restrictive, and subsidiary restrictive). The coefficients for these bond characteristics are not reported for brevity. The standard errors in parentheses are robust to clustering by date. The quarterly sample period covers September 2002 to December 2022.

Table 6. Downgrade Risk for Investment-Grade Bonds

	(1)	(2)	(3)	(4)	(5)
Rating:					
A	1.77	1.91	1.57	1.75	1.53
	(0.71)	(0.71)	(0.72)	(0.72)	(0.73)
BBB	4.25	4.22	3.90	4.05	3.86
	(0.71)	(0.70)	(0.70)	(0.71)	(0.71)
Rating watch:					
Positive	-2.15	-2.00	-2.22	-1.98	-2.19
	(0.62)	(0.62)	(0.65)	(0.64)	(0.69)
Negative	1.89	1.74	1.58	1.71	1.59
	(0.10)	(0.10)	(0.11)	(0.10)	(0.11)
Distance to default		-1.29	-1.00	-1.01	-0.89
		(0.10)	(0.09)	(0.09)	(0.09)
Credit spread			27.96		22.38
			(2.05)		(2.31)
Embedding				2.23	1.40
				(0.23)	(0.25)
Pseudo R^2	0.16	0.24	0.27	0.26	0.28
Observations	64,575	64,575	64,575	64,575	64,575

This table reports the coefficients from a logit model for a downgrade to speculative grade over the subsequent year, among firms that are currently rated investment grade. The omitted credit rating category is AAA and AA. The distance to default is standardized in the cross section of firms. The trained embedding is a linear combination of firm embeddings that best explains credit spreads. The standard errors in parentheses are robust to clustering by firm. The pseudo R^2 is one minus the log likelihood of a given model divided by the log likelihood of a null model with only an intercept. The quarterly sample period covers September 2002 to December 2021.

Table 7. Default Risk for Investment-Grade Bonds

	(1)	(2)	(3)	(4)	(5)
Rating:					
A	1.41	1.52	1.45	1.46	1.45
	(1.02)	(1.03)	(1.04)	(1.03)	(1.04)
BBB	0.81	0.54	0.44	0.45	0.42
	(1.01)	(1.01)	(1.02)	(1.01)	(1.02)
Rating watch:					
Positive	-1.10	-0.93	-0.96	-0.93	-0.94
	(0.73)	(0.73)	(0.73)	(0.73)	(0.73)
Negative	0.81	0.37	0.31	0.29	0.28
	(0.30)	(0.32)	(0.33)	(0.33)	(0.34)
Distance to default		-1.47	-1.36	-1.32	-1.30
		(0.30)	(0.32)	(0.32)	(0.33)
Credit spread			7.89		2.89
			(5.69)		(6.59)
Embedding				1.14	1.02
				(0.58)	(0.63)
Pseudo R^2	0.03	0.13	0.14	0.14	0.14
Observations	43,817	43,817	43,817	43,817	43,817

This table reports the coefficients from a logit model for default on any of its bonds over the subsequent five years, among firms that are currently rated investment grade. The omitted credit rating category is AAA and AA. The distance to default is standardized in the cross section of firms. The trained embedding is a linear combination of firm embeddings that best explains credit spreads. The standard errors in parentheses are robust to clustering by firm. The pseudo R^2 is one minus the log likelihood of a given model divided by the log likelihood of a null model with only an intercept. The quarterly sample period covers September 2002 to December 2017.

Table 8. An Improved Rating System

	Counterfactual rating						
Actual rating	AAA and AA	A	BBB	BB	B	CCC and below	Total
AAA and AA	4	4	0	0	0	0	8
A	3	19	11	0	0	0	32
BBB	1	9	29	3	0	0	42
BB	0	0	2	5	2	0	9
B	0	0	1	1	3	1	6
CCC and below	0	0	0	0	0	1	1
Total	8	32	42	9	6	1	100

This table reports the joint distribution of the actual ratings and the counterfactual ratings in percent. The counterfactual ratings rank bonds based on the estimated factor loadings for credit spreads, which depend on credit ratings and the trained embeddings. The quarterly sample period covers September 2002 to December 2022.

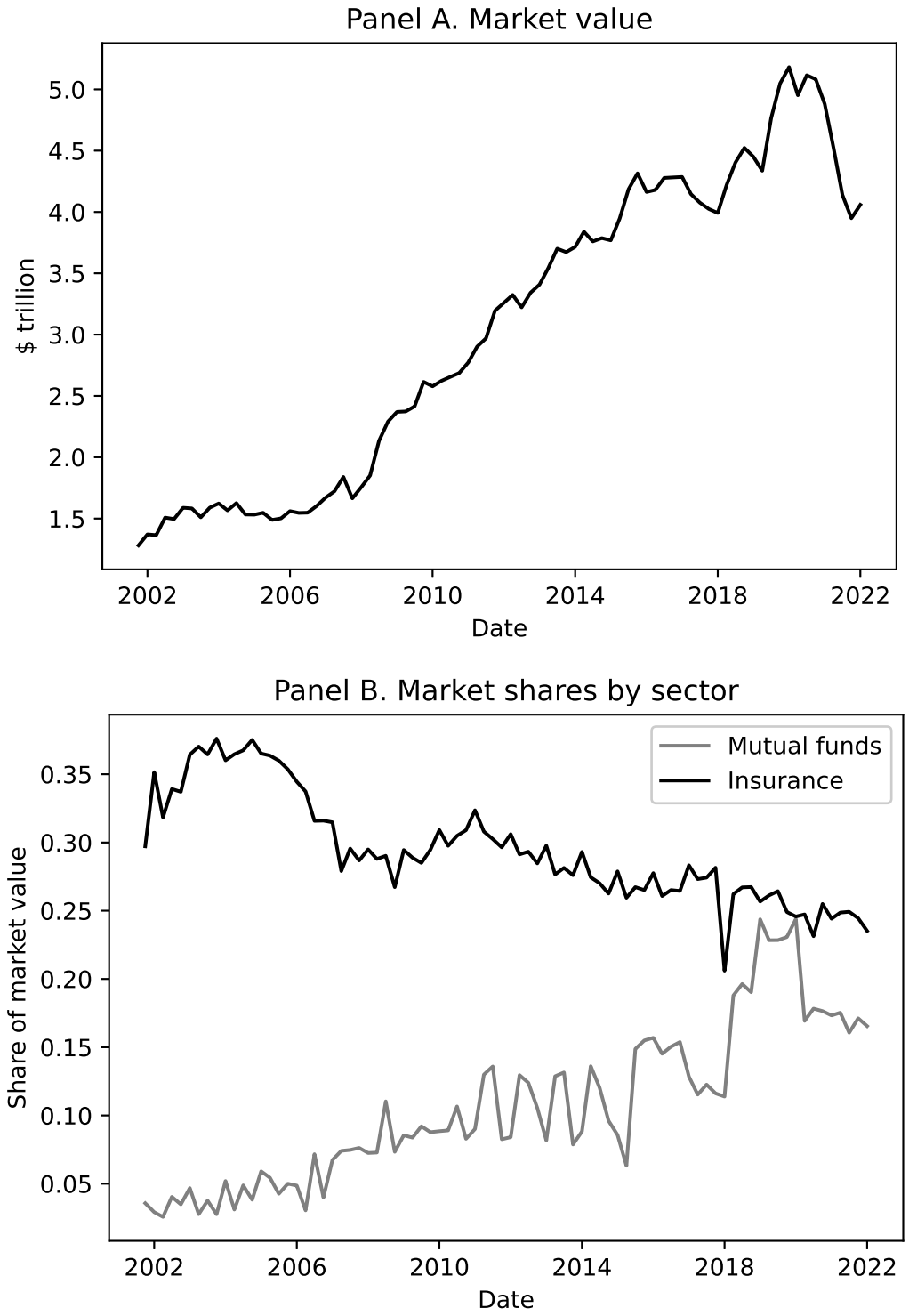

Figure 1. Corporate Bond Market. Panel A shows the aggregate market value of US corporate bonds that satisfy our sample criteria. Panel B shows the shares of the aggregate market value held by mutual funds and insurance companies. The quarterly sample period covers September 2002 to December 2022.

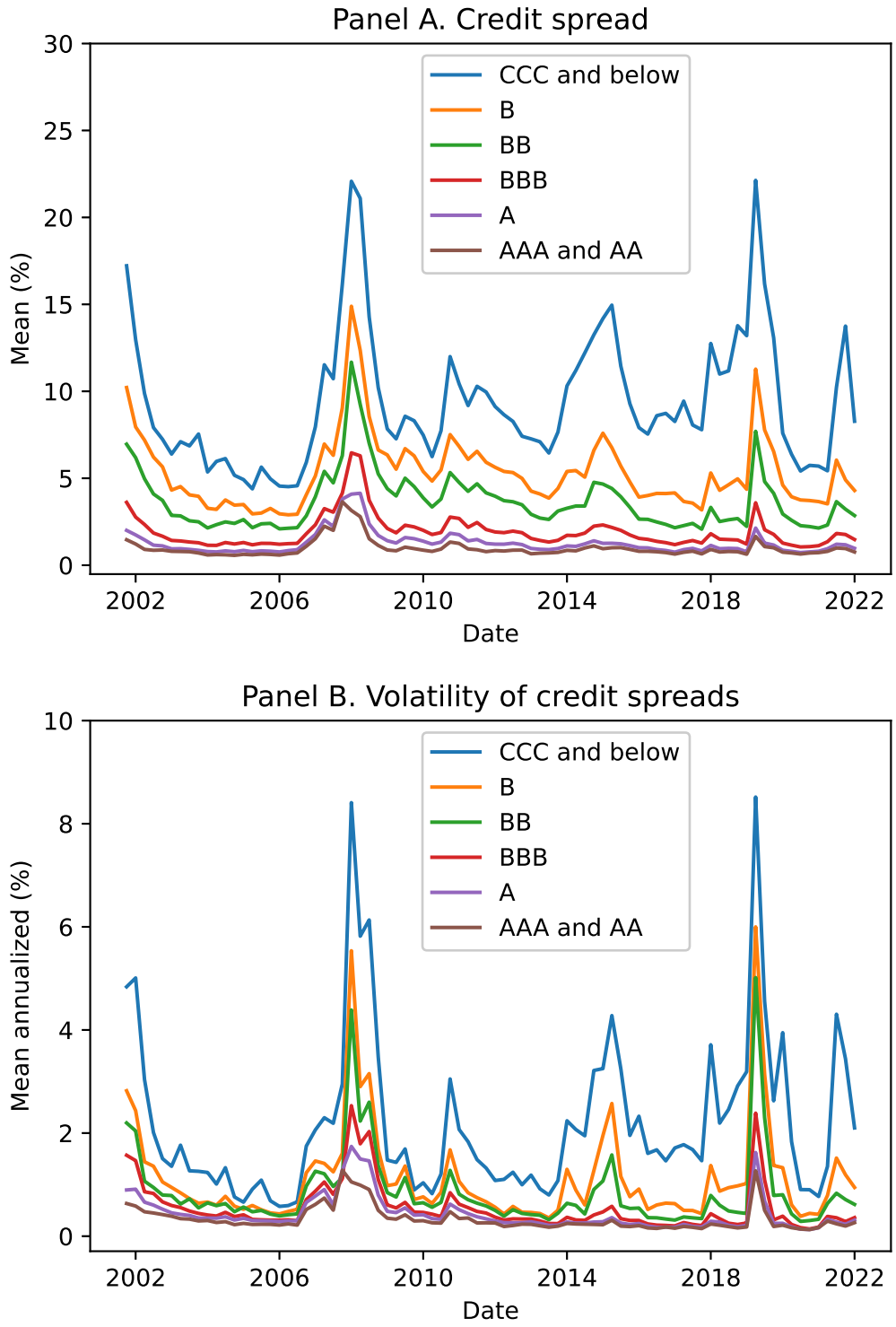

Figure 2. Factor Structure in Credit Spreads. Panel A shows the average credit spread by credit rating category. Panel B shows the cross-sectional mean of the annualized volatility of credit spreads by credit rating category. The quarterly sample period covers September 2002 to December 2022.

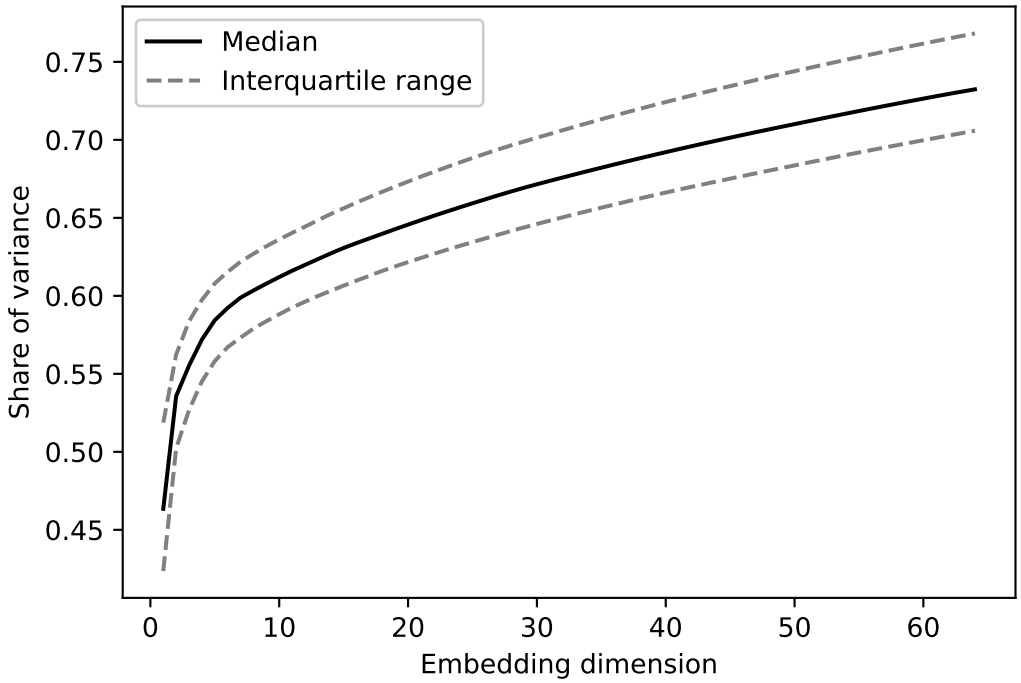

Figure 3. Explained Variation in Bond Holdings. This figure shows the cumulative share of the cross-sectional variance of bond holdings that the principal components explain. It shows the median and the interquartile range across all dates in the quarterly sample period from September 2002 to December 2022.

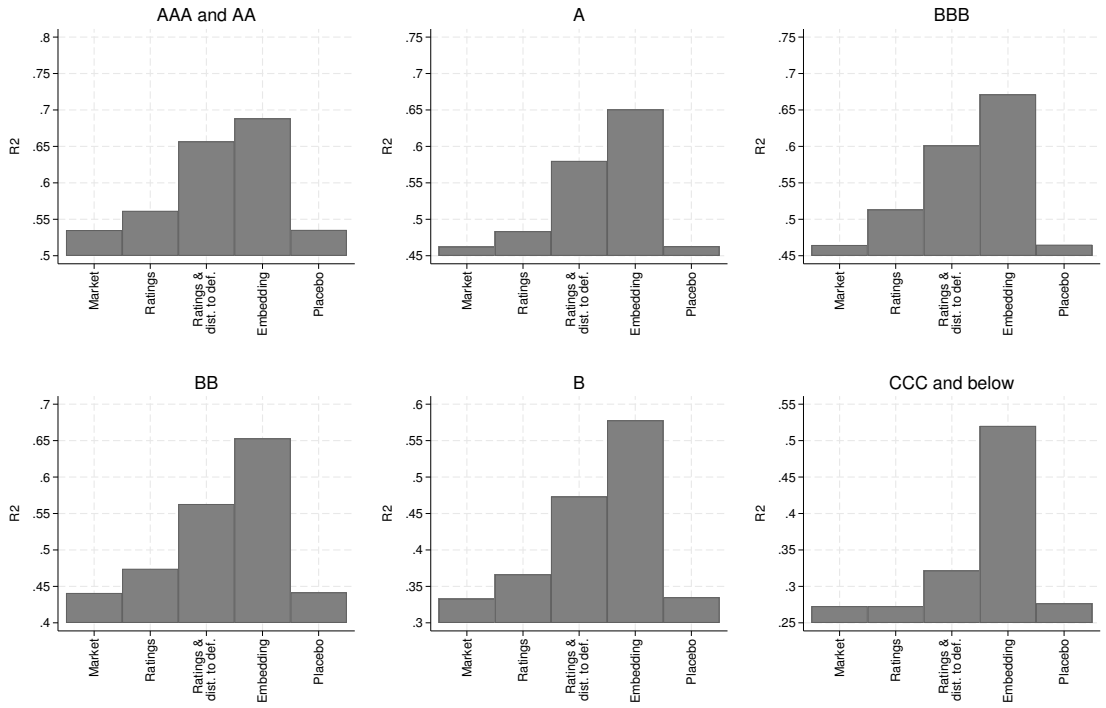

Figure 4. Model Fit for Credit Spreads. The first four bars show the R^2 corresponding to the first four columns of Table 3 and Table C1 in Appendix C. The fifth bar shows the R^2 from a regression of credit spreads on the market factor, its interaction with the trained placebo embedding, and the bond characteristics. The market factor is the value-weighted average of credit spreads. The bond characteristics are duration, log market value outstanding, indicator variables for embedded options (i.e., callable, putable, and credit enhancement), and indicator variables for covenants (i.e., bondholder protective, issuer restrictive, and subsidiary restrictive).

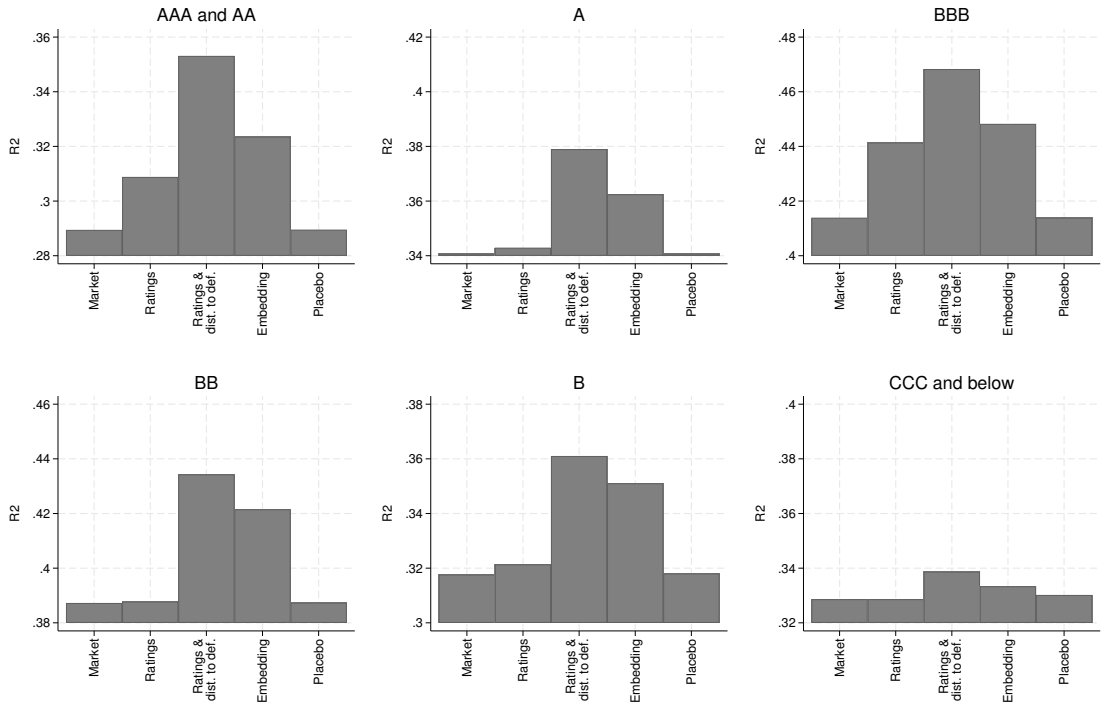

Figure 5. Model Fit for Changes in Credit Spreads. The first four bars show the R^2 corresponding to the first four columns of Table 4 and Table C2 in Appendix C. The fifth bar shows the R^2 from a regression of changes in credit spreads on the market factor and its interaction with the trained placebo embedding. The market factor is the value-weighted average quarterly change in credit spreads.

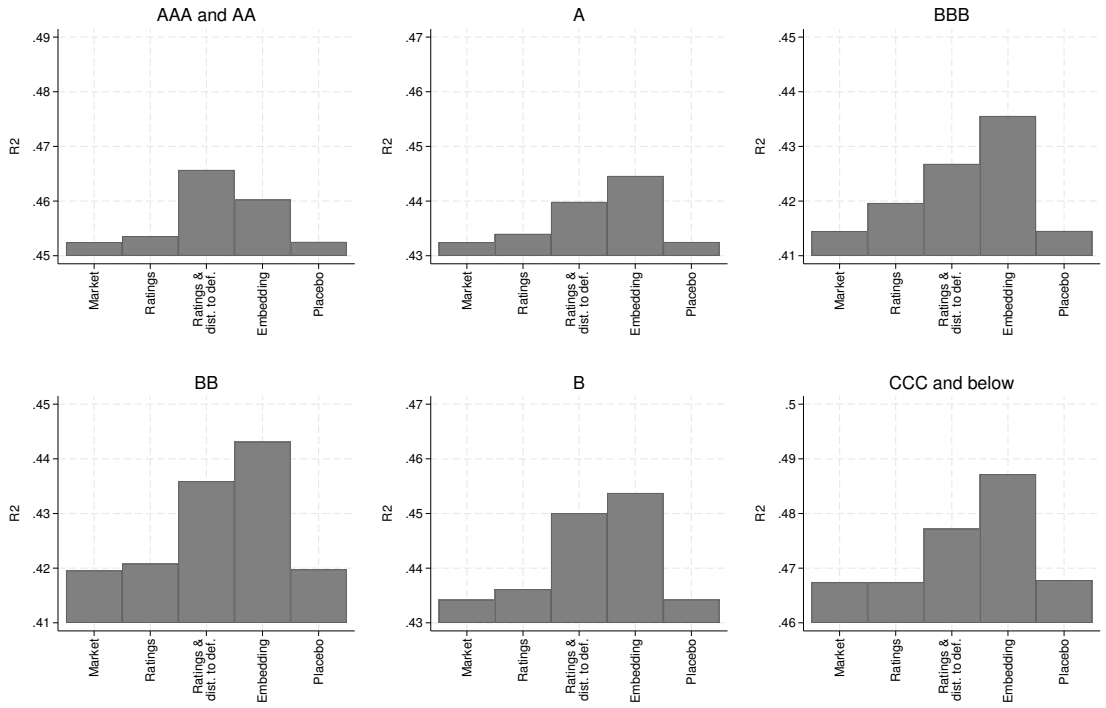

Figure 6. Model Fit for the Volatility of Credit Spreads. The first four bars show the R^2 corresponding to the first four columns of Table 5 and Table C3 in Appendix C. The fifth bar shows the R^2 from a regression of the volatility of credit spreads on the market factor, its interaction with the trained placebo embedding, and the bond characteristics. The market factor is the value-weighted average of credit spreads. The bond characteristics are duration, log market value outstanding, indicator variables for embedded options (i.e., callable, putable, and credit enhancement), and indicator variables for covenants (i.e., bondholder protective, issuer restrictive, and subsidiary restrictive).

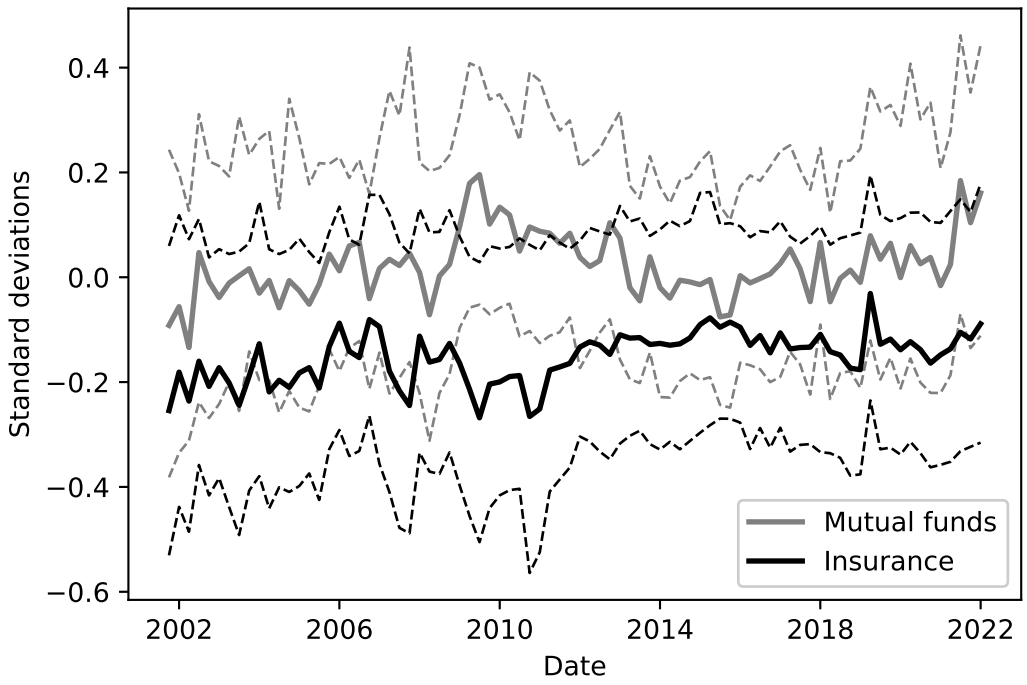

Figure 7. Portfolio-Weighted Trained Embeddings. This figure shows the median (solid) and the interquartile range (dots) of the portfolio-weighted trained embeddings by sector. The trained embedding is a linear combination of firm embeddings that best explains credit spreads. The unit for the vertical axis is a standard deviation of credit spreads by date and credit rating category.

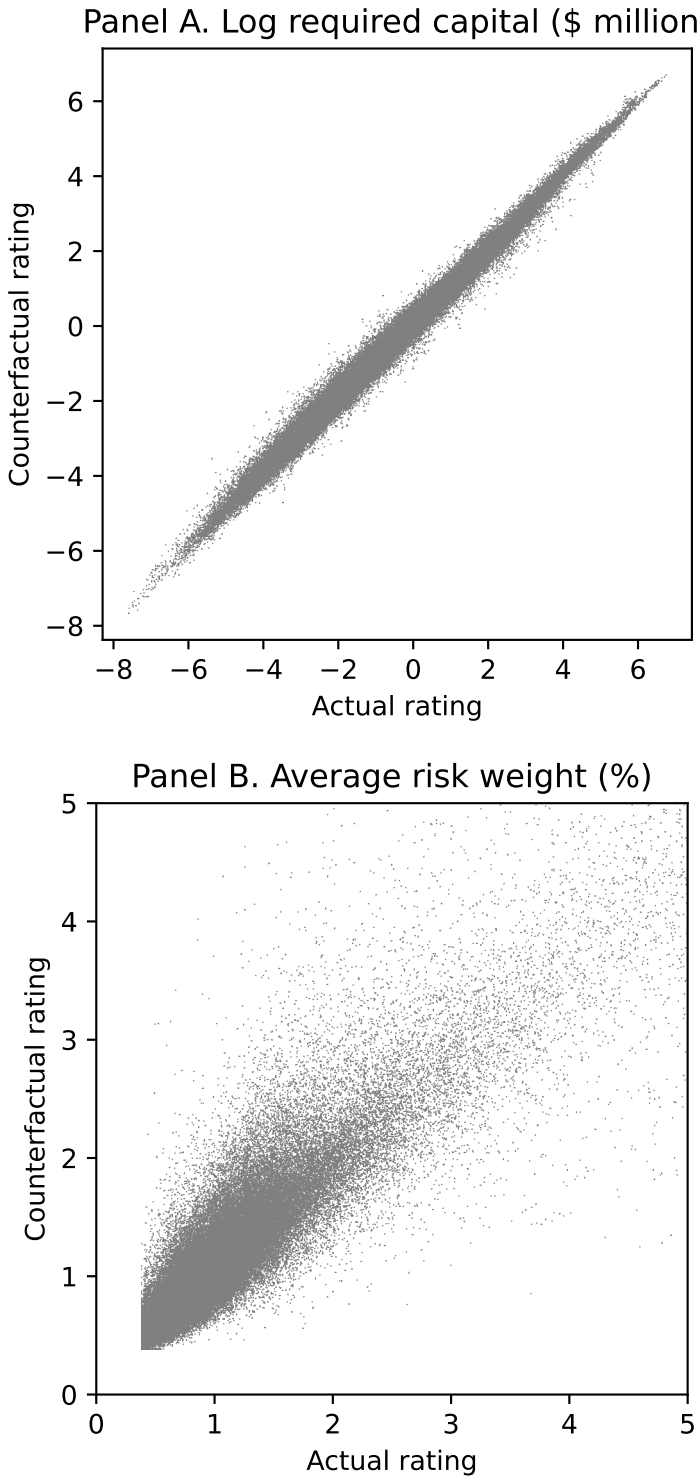

Figure 8. Required Capital under an Improved Rating System. Panel A is a scatter plot of log required capital for insurance companies under the actual versus the counterfactual ratings. Panel B is a scatter plot of the average risk weight for insurance companies under the actual versus the counterfactual ratings. The counterfactual ratings rank bonds based on the estimated factor loadings for credit spreads, which depend on credit ratings and the trained embeddings. The quarterly sample period covers September 2002 to December 2022.

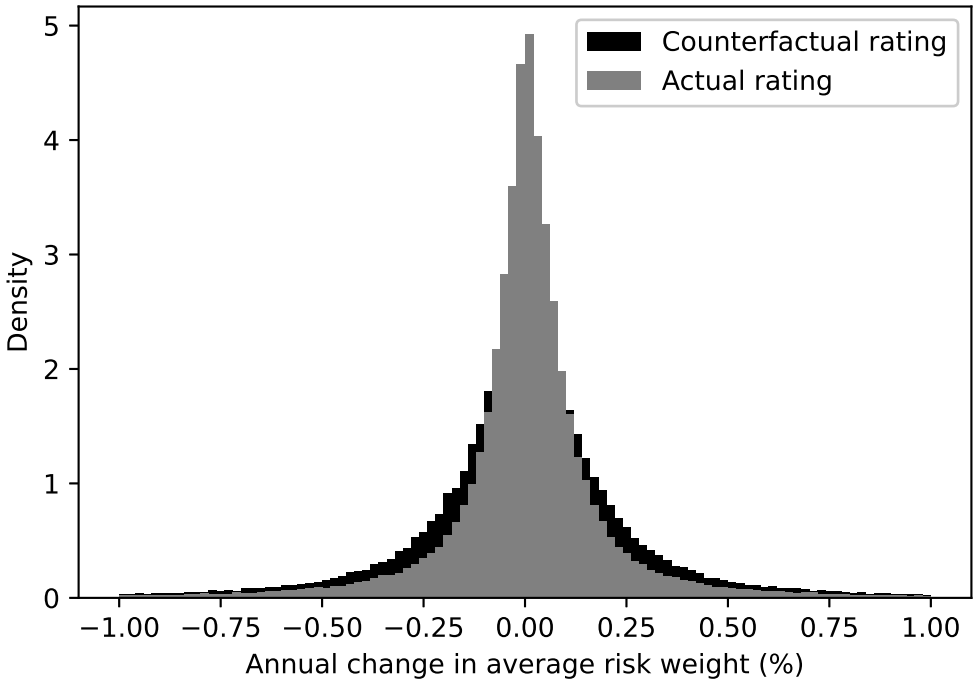

Figure 9. Changes in Required Capital under an Improved Rating System. This figure shows the distribution of annual changes in the average risk weight for insurance companies under the actual versus the counterfactual ratings. The counterfactual ratings rank bonds based on the estimated factor loadings for credit spreads, which depend on credit ratings and the trained embeddings. The quarterly sample period covers September 2002 to December 2022.

A. Solution of the Portfolio Choice Problem

We solve the portfolio choice problem 8. The first-order condition for portfolio choice is

$$\boldsymbol{\phi}_{i,t} - \mathbf{P}_t - \gamma_i \left(\boldsymbol{\psi}_t \boldsymbol{\psi}_t' + \sigma^2 \mathbf{I}\right) \boldsymbol{Q}_{i,t} - \theta_i \boldsymbol{\psi}_t = \mathbf{0}. \tag{A1}$$

We solve for the optimal portfolio as

$$\begin{aligned} \boldsymbol{Q}_{i,t} &= \frac{1}{\gamma_i} \left(\boldsymbol{\psi}_t \boldsymbol{\psi}_t' + \sigma^2 \mathbf{I}\right)^{-1} \left(\boldsymbol{\phi}_{i,t} - \mathbf{P}_t - \theta_i \boldsymbol{\psi}_t\right) \\ &= \frac{1}{\gamma_i \sigma^2} \left(\mathbf{I} - \frac{\boldsymbol{\psi}_t \boldsymbol{\psi}_t'}{\boldsymbol{\psi}_t' \boldsymbol{\psi}_t + \sigma^2}\right) \left(\boldsymbol{\phi}_{i,t} - \mathbf{P}_t - \theta_i \boldsymbol{\psi}_t\right) \\ &= \frac{1}{\gamma_i \sigma^2} \left(\boldsymbol{\phi}_{i,t} - \mathbf{P}_t - \Theta_{i,t} \boldsymbol{\psi}_t\right), \end{aligned} \tag{A2}$$

where the second line follows by the Woodbury matrix identity. Substituting equations (7) and (4), we have

$$Q_{i,n,f,t} = \frac{1}{\gamma_i \sigma^2} \left(-P_{n,f,t} + \left(\boldsymbol{\Phi}_{i,C} - \Theta_{i,t} \boldsymbol{\Psi}_C\right)' \boldsymbol{C}_{n,f,t} + \left(\boldsymbol{\Phi}_{i,Z} - \Theta_{i,t} \boldsymbol{\Psi}_Z\right)' \boldsymbol{Z}_{f,t}\right), \tag{A3}$$

which implies equation (9).

B. Data Construction

We describe the steps for cleaning and merging the Mergent Fixed Income Securities Database (FISD), the TRACE Enhanced Historical Data, and the eMAXX Bond Holdings Data. We also describe the construction of a link file between issuer CUSIP and gvkey.

B.1. Preparing the Mergent FISD

We clean and merge the necessary files from Mergent (2024). We start with the bond issue file and keep corporate bonds, identified by a bond type of CDEB, CMTN, CMTZ, or CZ. We exclude bonds that have variable coupons, are convertible, are in foreign currency, are Rule 144A bonds, or are asset-backed securities. We merge the data files for coupon information, bond redemption, and covenants (i.e., bondholder protective, issuer restrictive, and subsidiary restrictive). We keep bonds with a valid CUSIP, offering date, maturity date, interest frequency, and coupon rate. We keep bonds with a face value of \$10, \$50, \$100, \$1,000, \$2,000, \$5,000, \$10,000, \$100,000, or \$200,000.

We combine the data files on the offering amount and the historical amount outstanding to construct the amount outstanding at the monthly frequency. We combine the data files on the current rating and the historical rating to construct the credit rating and the rating watch indicator (positive or negative) for Moody's, S&P, and Fitch at the monthly frequency. We use the issue default file to construct a default indicator at the monthly frequency. We define a bond in default if the default type is B (bankruptcy), I (interest default), or P (principal default) or one of the credit ratings indicate default (C for Moody's, D for S&P, and D for Fitch). We recode the credit rating if necessary to ensure mutual consistency with the default indicator.

B.2. Preparing the TRACE Enhanced Historical Data

We clean FINRA (2024), separately for the pre-data (i.e., transactions before February 6, 2012) and the post-data (i.e., transactions on or after February 6, 2012) due to changes in the reporting requirements. The cleaning procedure for the post-data follows Dick-Nielsen (2014). The cleaning procedure for the pre-data follows Dick-Nielsen (2014), except for the cleaning procedure for chained corrections in Wharton Research Data Services (2017).

For the pre-data, we start with the original trade reports (trade status = T) and the corrected trade reports (trade status = W), excluding reversals (as of indicator = R). To handle chained corrections, we keep the most recent report for each original message sequence number. We remove cancellations by matching on CUSIP, quantity, price, execution date and time, buy/sell indicator, contra party indicator, and message sequence number. We remove reversals by matching on CUSIP, quantity, price, execution date and time, buy/sell indicator, contra party indicator, and message sequence number. We remove the buy side of interdealer trades (buy/sell indicator = B and contra party indicator = D) to avoid double counting. We keep trade reports with positive reported prices.

For the post-data, we start with the original trade reports (trade status = T) and the corrected trade reports (trade status = R). We remove cancellations (trade status = X or C) and reversals (trade status = Y) by matching on CUSIP, quantity, price, execution date and time, buy/sell indicator, contra party indicator, and message sequence number. We remove the buy side of interdealer trades (buy/sell indicator = B and contra party indicator = D) to avoid double counting. We keep trade reports with positive reported prices.

We merge the cleaned trade reports with the bond characteristics from Mergent (2024). We then use QuantLib to compute the semiannually compounded yield for each reported trade, based on the clean price, trading date, offering date, maturity date, first interest date, interest frequency, coupon rate, day count basis, and face value. We remove reported trades with extreme yields that are likely to be erroneous in two steps. We first remove negative

yields and yields greater than 30%. We then remove yields greater than the 99.9 percentile by credit rating (prioritized in the order of Moody's, S&P, and Fitch). Based on the remaining sample, we compute the volume-weighted average price and the corresponding yield at the daily frequency.

B.3. Preparing the eMAXX Bond Holdings Data

We use the North American corporate bond holdings (covering North American investors) and the European bond holdings (covering European investors) in Thomson Reuters (2024). We take three steps to avoid double counting. First, we remove aggregate reports for co-managed funds. Second, we keep the first record by report date, subaccount ID, and managing firm ID. Third, we keep the last report by year-quarter, subaccount ID, and managing firm ID. We keep reports that are within five days of the quarter-end date, removing observations with unusual report dates. We aggregate bond holdings by date, CUSIP, and subaccount ID.

We merge the bond holdings with Mergent (2024) by date and CUSIP to filter corporate bonds that are still outstanding. We remove observations where the par amount held exceeds the par amount outstanding. We merge the bond holdings with FINRA (2024) by date and CUSIP to construct the market value of the bond holding as the par amount held times the most recent volume-weighted average price within the quarter.

B.4. Merging the Datasets

We merge the three datasets by CUSIP. Panel A of Figure B1 starts with the total number of bonds in the Mergent FISD that satisfy our sample criteria. The total number of bonds decreases after merging with the TRACE Enhanced Historical Data because some bonds do not trade. The total number of bonds further decreases after merging with the eMAXX Bond Holdings Data because mutual funds and insurance companies do no hold some bonds. Panel B shows that the total face value of bonds decreases slightly after merging with the TRACE Enhanced Historical Data and the eMAXX Bond Holdings Data. Thus, the bonds that drop out of our sample are small in terms of face value outstanding.

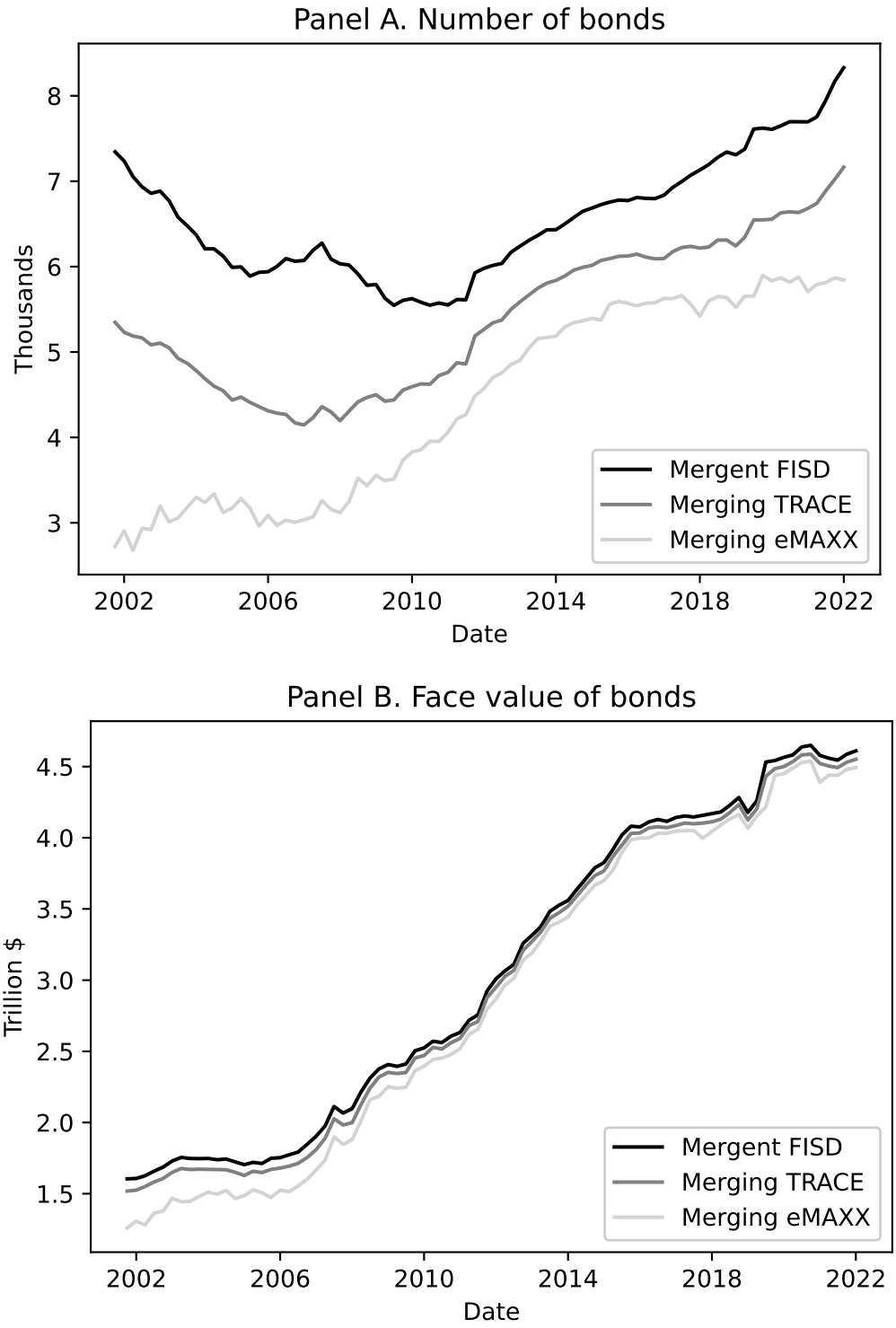

Figure B1. Sample of Corporate Bonds. Panel A shows the total number of bonds in the Mergent FISD that satisfy our sample criteria and the total number of bonds after merging with the TRACE Enhanced Historical Data and the eMAXX Bond Holdings Data. Panel B shows the same for the total face value of bonds. The quarterly sample period covers September 2002 to December 2022.

B.5. Link File between Issuer CUSIP and gvkey

A challenge in linking bonds to the CRSP-Compustat Database is that a firm can issue bonds under multiple CUSIP. Consider an example of General Electric in December 2022. General Electric has a permno of 12060, a CUSIP of 36960430 for its equity, and debt outstanding under the following issuer CUSIP: 369604, 36959C, 36962G, 057224, 05723K, 05724B, 35180P, 36164Q, 81413P, and 957674. Similarly to Mota and Siani (2024), we construct valid links from multiple sources to improve upon the link file in Wharton Research Data Services (2017).

We start with the set of gvkey (Compustat Global Company Key) and permno (CRSP permanent identifier) linked to primary securities (linkprim = P or C) in Center for Research in Security Prices (2024a). We construct a link file between a bond's issuer CUSIP and gvkey in the following order of priority.

1. We start with all active links between issuer CUSIP (i.e., the first six digits of CUSIP) and gvkey in Center for Research in Security Prices (2024b). When there are multiple valid links on a given date, we prioritize in the order of a longer duration, an earlier link start date, or an earlier link end date.

2. We start with all active links between issuer CUSIP (i.e., the first six digits of CUSIP) and gvkey in S&P Global (2024a). We find (and have verified with S&P Global) that these linking tables are incomplete before February 2008. When there are multiple valid links on a given date, we prioritize in the order of a longer duration, an earlier link start date, or an earlier link end date.

3. We start with all active links between issuer CUSIP (i.e., the first six digits of CUSIP) and ticker symbol, based on the trade reports in FINRA (2024). We merge these data with Center for Research in Security Prices (2024b) by ticker symbol and Center for Research in Security Prices (2024a) by permno. When there are multiple valid links on a given date, we prioritize in the order of a longer duration, an earlier link start date, or an earlier link end date.

C. Supplemental Results

Tables C1–C3 report regressions for the speculative-grade rating categories. Table C1 reports regressions of credit spreads on the market factor and its interactions with credit ratings, the distance to default, and the trained embeddings. Table C2 reports regressions of changes in credit spreads on the market factor and its interactions with credit ratings, the distance to

default, and the trained embeddings. Table C3 reports regressions of the volatility of credit spreads on the market factor and its interactions with credit ratings, the distance to default, and the trained embeddings.

Table C1. Credit Spreads on Speculative-Grade Bonds

Panel A. Rated BB						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	1.79	1.77	1.50	1.77	1.73	1.57
	(0.16)	(0.15)	(0.12)	(0.16)	(0.14)	(0.13)
Rating × Market:						
BB		0.07	0.00		0.04	0.01
		(0.11)	(0.12)		(0.09)	(0.10)
BB−		-0.03	-0.14		0.06	-0.01
		(0.12)	(0.11)		(0.07)	(0.07)
Distance to default × Market			-0.74			-0.43
			(0.10)			(0.04)
Embedding × Market				1.10	1.05	0.91
				(0.04)	(0.05)	(0.04)
Rating:						
BB		0.51	0.52		0.18	0.24
		(0.17)	(0.19)		(0.15)	(0.17)
BB−		1.02	1.04		0.32	0.43
		(0.20)	(0.19)		(0.13)	(0.13)
Distance to default			0.51			0.32
			(0.15)			(0.07)
R^2	0.44	0.47	0.56	0.65	0.66	0.69
Observations	33,501	33,501	33,501	33,501	33,501	33,501
Panel B. Rated B						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	2.10	1.90	1.23	2.18	2.00	1.56
	(0.19)	(0.16)	(0.14)	(0.23)	(0.23)	(0.23)
Rating × Market:						
B		0.15	-0.12		0.22	0.04
		(0.12)	(0.14)		(0.07)	(0.06)
B−		0.60	0.04		0.45	0.12
		(0.13)	(0.17)		(0.07)	(0.06)
Distance to default × Market			-1.31			-0.82
			(0.18)			(0.11)
Embedding × Market				1.20	1.16	1.00
				(0.03)	(0.03)	(0.03)
Rating:						
B		0.27	0.54		-0.11	0.13
		(0.24)	(0.26)		(0.13)	(0.11)
B−		0.12	0.71		-0.11	0.31
		(0.27)	(0.31)		(0.15)	(0.13)
Distance to default			0.89			0.68
			(0.30)			(0.20)
R^2	0.33	0.37	0.47	0.58	0.59	0.62
Observations	22,831	22,831	22,831	22,831	22,831	22,831
Panel C. Rated CCC and below						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	2.96	2.96	2.36	2.92	2.92	2.12
	(0.27)	(0.27)	(0.61)	(0.29)	(0.29)	(0.50)
Distance to default × Market			-0.72			-0.80
			(0.57)			(0.46)
Embedding × Market				1.46	1.46	1.39
				(0.05)	(0.05)	(0.06)
Distance to default			-2.00			-0.48
			(1.02)			(0.86)
R^2	0.27	0.27	0.32	0.52	0.52	0.54
Observations	5,272	5,272	5,272	5,272	5,272	5,272

The market factor is the value-weighted average of credit spreads. The omitted credit rating is BB+ in Panel A and B+ in Panel B. The distance to default is standardized in the cross section of firms. The trained embedding is a linear combination of firm embeddings that best explains credit spreads. All specifications include duration, log market value outstanding, indicator variables for embedded options (i.e., callable, putable, and credit enhancement), and indicator variables for covenants (i.e., bondholder protective, issuer restrictive, and subsidiary restrictive). The coefficients for these bond characteristics are not reported for brevity. The standard errors in parentheses are robust to clustering by date. The quarterly sample period covers September 2002 to December 2022.

Table C2. Changes in Credit Spreads on Speculative-Grade Bonds

Panel A. Rated BB						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	1.73	1.79	1.39	1.76	1.93	1.57
	(0.22)	(0.23)	(0.16)	(0.22)	(0.25)	(0.18)
Rating × Market:						
BB		-0.04	-0.17		-0.26	-0.30
		(0.17)	(0.18)		(0.17)	(0.18)
BB−		-0.15	-0.27		-0.30	-0.35
		(0.14)	(0.15)		(0.15)	(0.16)
Distance to default × Market			-0.95			-0.76
			(0.14)			(0.16)
Embedding × Market				1.12	1.17	0.82
				(0.44)	(0.41)	(0.44)
R^2	0.39	0.39	0.43	0.42	0.42	0.45
Observations	31,541	31,541	31,541	31,541	31,541	31,541

Panel B. Rated B						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	2.17	1.99	1.09	2.22	2.10	1.35
	(0.27)	(0.27)	(0.22)	(0.30)	(0.29)	(0.20)
Rating × Market:						
B		0.08	-0.29		0.08	-0.21
		(0.14)	(0.18)		(0.11)	(0.14)
B−		0.54	-0.07		0.34	-0.08
		(0.21)	(0.23)		(0.17)	(0.20)
Distance to default × Market			-1.45			-1.15
			(0.19)			(0.18)
Embedding × Market				1.18	1.15	0.81
				(0.21)	(0.20)	(0.18)
R^2	0.32	0.32	0.36	0.35	0.35	0.37
Observations	21,182	21,182	21,182	21,182	21,182	21,182

Panel C. Rated CCC and below						
	(1)	(2)	(3)	(4)	(5)	(6)
Market	3.87	3.87	1.69	3.95	3.95	1.79
	(0.43)	(0.43)	(0.73)	(0.48)	(0.48)	(0.73)
Distance to default × Market			-1.79			-1.77
			(0.59)			(0.62)
Embedding × Market				0.42	0.42	0.41
				(0.35)	(0.35)	(0.32)
R^2	0.33	0.33	0.34	0.33	0.33	0.34
Observations	4,636	4,636	4,636	4,636	4,636	4,636

The market factor is the value-weighted average quarterly change in credit spreads. The omitted credit rating is BB+ in Panel A and B+ in Panel B. The distance to default is standardized in the cross section of firms. The trained embedding is a linear combination of firm embeddings that best explains credit spreads. The standard errors in parentheses are robust to clustering by date. The quarterly sample period covers September 2002 to December 2022.

Table C3. Volatility of Credit Spreads on Speculative-Grade Bonds

Panel A. Rated BB						
	(1)	(2)	(3)	(4)	(5)	(6)
Volatility	0.48	0.47	0.42	0.41	0.41	0.38
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Market	15.95	15.50	17.32	19.11	18.47	20.41
	(7.28)	(7.67)	(7.64)	(6.98)	(7.43)	(7.35)
Rating × Market:						
BB		1.91	1.79		1.15	1.53
		(3.47)	(3.49)		(3.17)	(3.25)
BB−		0.14	-0.83		0.77	0.37
		(2.91)	(3.13)		(2.88)	(3.12)
Distance to default × Market			-4.26			-0.75
			(1.67)			(1.70)
Embedding × Market				13.36	13.30	11.49
				(1.30)	(1.35)	(1.25)
Rating:						
BB		-0.22	-1.75		-2.74	-4.42
		(7.75)	(7.94)		(7.64)	(7.88)
BB−		6.19	5.19		-0.53	-1.33
		(6.76)	(7.20)		(6.94)	(7.46)
Distance to default			-8.20			-11.25
			(3.81)			(3.73)
R^2	0.42	0.42	0.44	0.44	0.44	0.45
Observations	32,750	32,750	32,750	32,750	32,750	32,750

Panel B. Rated B						
	(1)	(2)	(3)	(4)	(5)	(6)
Volatility	0.56	0.55	0.50	0.49	0.49	0.46
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Market	13.30	13.61	17.37	17.15	17.36	21.76
	(7.82)	(7.61)	(8.18)	(7.31)	(6.80)	(7.44)
Rating × Market:						
B		-0.44	-0.78		-0.01	0.29
		(1.52)	(1.43)		(1.77)	(1.75)
B−		0.94	0.11		0.01	0.65
		(2.16)	(1.79)		(2.52)	(2.29)
Distance to default × Market			-0.97			2.32
			(2.94)			(3.13)
Embedding × Market				9.47	9.27	8.09
				(1.58)	(1.57)	(1.52)
Rating:						
B		6.64	4.49		4.33	1.63
		(4.17)	(3.80)		(4.27)	(4.13)
B−		6.61	3.47		5.38	0.73
		(4.48)	(3.85)		(4.96)	(4.71)
Distance to default			-18.41			-20.76
			(6.55)			(6.56)
R^2	0.43	0.44	0.45	0.45	0.45	0.46
Observations	22,373	22,373	22,373	22,373	22,373	22,373

Panel C. Rated CCC and below						
	(1)	(2)	(3)	(4)	(5)	(6)
Volatility	0.61	0.61	0.57	0.55	0.55	0.52
	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
Market	12.76	12.76	26.66	15.04	15.04	28.55
	(7.52)	(7.52)	(10.43)	(7.39)	(7.39)	(10.18)
Distance to default × Market			8.69			8.81
			(7.01)			(6.79)
Embedding × Market				6.80	6.80	6.48
				(1.24)	(1.24)	(1.24)
Distance to default			-41.33			-38.86
			(15.58)			(15.54)
R^2	0.47	0.47	0.48	0.49	0.49	0.50
Observations	5,001	5,001	5,001	5,001	5,001	5,001

The market factor is the value-weighted average of credit spreads. The omitted credit rating is BB+ in Panel A and B+ in Panel B. The distance to default is standardized in the cross section of firms. The trained embedding is a linear combination of firm embeddings that best explains credit spreads. All specifications include duration, log market value outstanding, indicator variables for embedded options (i.e., callable, putable, and credit enhancement), and indicator variables for covenants (i.e., bondholder protective, issuer restrictive, and subsidiary restrictive). The coefficients for these bond characteristics are not reported for brevity. The standard errors in parentheses are robust to clustering by date. The quarterly sample period covers September 2002 to December 2022.