The Optimal Use of AI in Financial Regulation

Christopher Clayton* Antonio Coppola†

September 2026

Abstract

We study whether AI methods applied to large-scale portfolio holdings data can improve financial stability monitoring and inform macroprudential regulation. We build a deep learning model tailored to security-level data on the holdings of financial intermediaries, which treats these data as a network linking investors to the assets they hold. The model builds in economic structure and estimates latent representations of both assets and investors. Applied to the universe of non-bank financial intermediaries, covering nearly $40 trillion in wealth, the model substantially outperforms a broad set of existing approaches in out-of-sample forecasts of intermediary trading behavior, including in crisis episodes. The model's incremental explanatory power for the cross-section of asset returns during stress events is more than ten times that of traditional holdings-based approaches, and it outperforms existing systemic risk metrics in identifying risky institutions. This approach shows that the holdings network encodes rich, economically interpretable information relevant for forecasting the dynamics of financial stress episodes. We embed our empirical approach into a macroprudential optimal policy framework to formalize why these objects matter for policy and welfare. We characterize when the predictive information from the model improves welfare by sharpening the cross-sectional targeting of policy interventions, and we demonstrate a complementarity between prediction and structural knowledge.

Keywords: Financial Stability, Holdings Data, Fire Sales, Macroprudential Policy, Asset Pricing, Deep Learning.

JEL Codes: C4, G1, G2.

*Yale School of Management and NBER; christopher.clayton@yale.edu.

†University of Chicago Booth School of Business, NBER, and CEPR; antonio.coppola@chicagobooth.edu.

We thank Saleem Bahaj, Hui Chen, Maryam Farboodi, Paul Fontanier, Xavier Gabaix, Stefano Giglio, Paul Huebner, Guido Imbens, Hanno Lustig, Matteo Maggiori, Stefan Nagel, Pablo Ottonello, Markus Pelger, Rob Richmond, Jann Spiess, and Ansgar Walther for helpful comments. We also thank Oliver Wang and Zhongyu Yin for outstanding research assistance. This paper was previously circulated under the title *"Financial Regulation and AI: A Faustian Bargain?"*. Financial support from the Stanford GSB Business, Government, and Society (BGS) Initiative is gratefully acknowledged.

1 Introduction

A central task of macroprudential financial regulation is to identify where shocks will be amplified through the financial system and to intervene before the resulting price dislocations become socially costly. Central banks and other financial supervisors today face an evolving information environment characterized by both greater observability of financial positions throughout the financial system and by the availability of better information processing technologies. On the data front, supervisors now have access to granular data on portfolio positions, showing who owns what across a large share of the financial system. On the technological front, high-dimensional predictive methods have become much more capable. This paper brings together these new data and methods to show that deep learning models, when properly designed for the rich cross-section of portfolio holdings data that is now available, can substantially improve financial stability monitoring across multiple dimensions: forecasting how financial intermediaries will trade, particularly during market stress periods; explaining which markets and assets suffer price dislocations during crises; and identifying risky financial institutions. We show that the structure of overlapping portfolio positions across financial intermediaries contains more information about fire-sale vulnerabilities than is captured by the low-dimensional summaries used by existing approaches, and that this information can be recovered in an economically interpretable way. We complement our empirical analysis with a macroprudential optimal policy framework that provides one disciplined interpretation of when and why such predictive information has policy value and can improve the targeting of regulatory interventions.

Learning effectively from data on institutional portfolios requires building models whose architecture appropriately reflects the structure of the problem. With unlimited data, any sufficiently flexible model could in principle learn equivalently well from portfolio positions; with the finite data available in practice, how the model is built makes a large difference. This motivates our focus on model design: that is, we ask how to construct economics-informed deep learning approaches, which use the economics of the problem to inform the structure of the model itself, so that it learns efficiently from limited data. Fire-sale dynamics, which are a core channel of financial crises, are inherently a networked phenomenon: they propagate through the network of overlapping portfolios that connects investors and assets. For example, a fund that is forced to liquidate corporate bonds depresses prices for all other financial institutions holding those same bonds, which in turn might trigger further rounds of forced selling by other investors. The dynamics of who is likely to sell what in a crisis therefore depend on the full high-dimensional structure of portfolio overlaps across the financial system, beyond just the characteristics of individual investors or individual assets in isolation. Much of the modern deep learning toolkit, including sequence-based architectures that have been recently applied in asset pricing contexts, is designed for sequential or grid-structured inputs such as text and images. In domains where the data has an important network structure, however, architectures tailored to learn from it have led to major advances in performance (Jumper et al. 2021). We therefore design our deep learning approach to explicitly capture and make use of

the full network structure of portfolios.

Specifically, we build and train a graph transformer, a model within the class of graph neural networks, which are deep learning architectures designed for network-structured data. The graph transformer incorporates an attention mechanism, a parametric component which lets the model learn how to best combine information across the portfolio holdings network. By doing this, the model learns underlying latent representations of both investors and assets known as embeddings: these are vectors of latent characteristics which summarize everything the model has learned about each investor and asset. We build economic priors into this deep learning model, for instance by using portfolio exposures as biases in its attention mechanism, so that positions of greater economic importance receive correspondingly greater weight in the information aggregation. Attention is the same building block that underlies modern large language models, but our model is built for data on portfolios rather than text.

We train the deep learning model on a quarterly panel of portfolio holdings data covering the positions of roughly 37,000 non-bank financial intermediaries in 85,000 assets, totaling nearly $40 trillion in wealth by the end of our sample. Our empirical application focuses on non-bank financial intermediaries since these are an increasingly important source of financial fragility, partly as a result of regulatory reform following the Great Financial Crisis which prompted a migration of risk intermediation from the banking sector to the non-bank sector. The model is trained on the foundational task of predicting the portfolio adjustments of the intermediaries in the data. It takes as input the current cross-section of portfolio holdings and a realized or scenario value of aggregate financial stress, and it then produces a forecast for how each investor will trade in each asset. Hence our estimation target is the cross-sectional pattern of trading conditional on the aggregate state. The model does not produce a forecast of the aggregate state itself: the primary value of portfolio holdings data in our approach is in predicting which investors will adjust which positions given the prevailing stress level, not in predicting whether a crisis will occur.

We demonstrate the performance of the model in three complementary dimensions that are relevant for financial stability monitoring and macroprudential regulation. First, we show that the model is able to successfully predict how investors will trade. Under a walk-forward out-of-sample evaluation, in which the model is retrained each year on an expanding window and then tested strictly on future data, the graph transformer substantially outperforms conventional balance-sheet measures, standard econometric methods, and alternative deep learning benchmarks in predicting financial intermediaries' trading behavior, including during crisis episodes. The model outperforms benchmarks on both the extensive margin (predicting which investor-asset positions are traded) and the intensive margin (how large trades are, conditional on trading). We show that the model generalizes well out of sample, including when making forecasts for new assets and new investors that were not in its training data, suggesting that the model learns underlying patterns of trading behavior that generalize beyond the specific sample seen during training.

Second, and of particular interest from an asset pricing perspective, we show that the model's

learned representations (embeddings) of assets and investors explain the cross-section of asset price dislocations during stress episodes, despite the model never having been trained on price or returns data. This asset pricing exercise provides a test of whether the model has learned the dimensions of trading that matter for fire sales. Traditional approaches to explaining crisis-period asset returns rely on collapsing the high-dimensional holdings data into a few low-dimensional summary statistics motivated by economic theory. We show that the graph transformer's embeddings deliver more than ten times the incremental explanatory power of these traditional summaries for the cross-section of bond returns in stress periods: 20 percentage points of incremental adjusted R-squared beyond standard asset characteristics, compared to roughly 1 percentage point from representative existing low-dimensional approaches. Intuitively, unlike traditional approaches, the model can draw on the full high-dimensional structure of the holdings data that is relevant for trading behavior and fire sale dynamics: for example, which positions are most crowded, which specific institutions are most exposed to funding instability, or where forced selling is likely to concentrate.

We show empirically how to open the black box of the model's predictions to shed light on the underlying economic mechanisms driving the predictions. For example, we document that an important dimension of the learned embeddings is economically interpretable: it captures the interaction of position crowdedness (concentration among flow-sensitive holders) and asset illiquidity. This single dimension accounts for nearly one fifth of the embeddings' explanatory power for crisis-period returns. Hence the model is recovering, from the data's relational structure, one of the economic forces that theory identifies as governing fire-sale amplification.

Third, we show how to use the model's predictions to identify risky financial institutions before a stress event. As a systemic risk metric, the model's predicted selling intensity outperforms existing institution-level measures—including ΔCoVaR, fund outflows, trailing portfolio returns, and portfolio concentration—by a wide margin in the non-bank setting we study. We also show how to quantify the uncertainty in the model's point predictions: we apply nonparametric conformal methods to construct prediction intervals around the model's forecasts. We show that the resulting prediction intervals obtained are well-calibrated across both calm and stress market regimes. Together, these results demonstrate that the network of financial positions contains a substantial amount of predictive information about fire-sale dynamics that conventional low-dimensional summaries miss. Our optimal policy framework shows how this recovered heterogeneity can be used to target macroprudential interventions more precisely.

Two additional properties of our deep learning architecture make it particularly well-suited to our setting. First, the model is permutation-invariant: predictions do not depend on the arbitrary ordering of investors or assets. This is a key distinction from sequence-based architectures such as text transformer architectures that underpin modern large language models, in which the order of tokens is an important part of the information used by the model. In our setting, permutation invariance respects a fundamental symmetry of the problem. Second, the graph transformer model is inductive: as all model parameters are shared across nodes, the model is able to generalize to

previously unseen investors or assets without retraining. We verify the inductive generalization capability of the model empirically. The ability of a model to generalize inductively is important in panel settings such as the one we consider, where new entities continuously enter and exit the data. Moreover, parameter sharing also enforces a form of regularization on the model, preventing the model from overfitting to the training data and leading to strong generalization out of sample. Together, these properties lead to more sample-efficient learning, which distinguishes graph-based architectures from alternatives that treat portfolio data as an unstructured grid or sequence.

While the empirical objects that we study are of direct, independent interest for financial stability monitoring and macroprudential surveillance, we also develop a complementary theoretical framework that offers one disciplined interpretation of the policy and welfare value of these empirical objects. In particular, the theory lets us analyze the extent to which the predictive models we introduce here can be useful for informing policy in an equilibrium environment subject to behavioral adaptation of market participants. This is the familiar Lucas (1976) critique, which is a natural concern with deploying any reduced-form predictive model for policy: i.e., the model might forecast distress accurately, but the empirical relationships it learns need not be invariant to the policy itself. We characterize the regulator's optimal intervention and show that it decomposes into two inputs. The first is a set of policy transmission elasticities, which characterize how the regulator's policy instruments impacts liquidation prices and quantities in equilibrium. The second is a set of cross-sectional targeting statistics: the forecasted pattern of forced selling across investors and assets that would prevail absent intervention (matching the empirical focus on predicting trading). The key implication is that a purely predictive model—i.e., one that is uninformative about policy counterfactuals—can improve welfare, provided it sharpens the regulator's forecast of these targeting statistics and the regulator has prior knowledge of the relevant policy transmission elasticities. For example, a regulator who knew before 2008 that stress would concentrate in mortgage-backed securities, and in the institutions most exposed to them, could in principle have used that information to target its regulation specifically at those markets and institutions.

We consider a standard fire-sale environment in which intermediaries, faced with binding collateral constraints, are forced to sell assets prior to maturity to second-best users. Intermediaries can choose which assets to sell, subject to their constraints, and assets may face different degrees of fire sales. In this environment, we consider intervention by a regulator that can use a set of regulatory wedges in order to manage which assets are sold by intermediaries, for example, pushing intermediaries to sell assets with smaller fire sale impacts. The key information friction is that the regulator faces uncertainty about the underlying parameters that govern what assets intermediaries sell, and what the fire sale spillovers are. The regulator can choose to adopt a *model*, which is a process for drawing a signal of the true parameters, before intervening. The regulator's core question is where to intervene and by how much. The importance of fire-sale externalities varies across markets. For example, a dollar of forced selling in an illiquid corporate bond market is likely to depress prices far more than the same quantity of selling pressure in a more liquid market, such as the market

for government bonds. So the regulator needs to know both which markets are under the most selling pressure and how its instruments translate into price changes in those markets. These are two different kinds of information.

We formalize this distinction by showing that under certain conditions on the information structure in the model, the regulator's optimal intervention decomposes into two components,[1]

$$\underbrace{\tau^*}_{\text{Optimal Intervention}} = \underbrace{\Omega}_{\text{Policy Elasticities}} \quad \underbrace{\hat{L}}_{\text{Predicted Selling}} .$$

The first term Ω corresponds to the structural elasticities capturing the regulator's policy transmission: how the regulator's instruments change equilibrium selling, and how those changes move market-clearing prices. These depend on many features of the economic environment, for example the depth of secondary markets, the binding constraints facing intermediaries, as well as the particular structure of demand. The second term $\hat{L}$ consists of the regulator's forecast of the pattern of forced selling across investors and assets that would prevail absent policy intervention, capturing the predictive channel. The prediction of liquidations is important to the regulator because the social benefit of intervening on any margin is proportional to the amount of forced selling on that margin, which determines the amount of assets being revalued in the fire sale. A model that accurately forecasts where selling will concentrate therefore allows the regulator to target its intervention more precisely. This is true even if the model says nothing about how policy transmission, i.e. about how intervention changes outcomes. This point also clarifies why the empirical task we focus on is to predict the cross-sectional pattern of investor-level trading conditional on the aggregate state.

We show that under the information conditions in our framework, the structural loadings that translate portfolios and policy wedges into equilibrium outcomes are invariant to the regulator's model choice, even though intermediaries' ex ante portfolio decisions respond endogenously to the anticipated intervention, making the use of a predictive model robust to the Lucas critique in this environment. Intuitively, the robustness comes from the fact that in this framework, predictive models are not used for estimating policy counterfactuals, but simply to improve the cross-sectional targeting of the policy. The theory yields a complementarity result: the accuracy of a predictive model is most valuable precisely on the margins where the regulator knows that policy transmission is strongest in a causal sense. We also show that a weighted predictive R-squared, where the weights reflect the causal leverage of the policy, is a sufficient statistic for the welfare gains from a given predictive model, which provides a bridge between out-of-sample forecast accuracy and the value of a predictive model for regulation. We further study the feedback of model choice on ex ante portfolio decisions, characterize optimal ex ante macroprudential regulation, and extend the analysis to subsidy-based (bailout) interventions.

The broader contribution of this paper is to clarify the role of prediction in macro-policy prob-

[1]The structure of this optimal wedge formula is familiar from the macroprudential policy literature (e.g., Farhi and Werning 2016), with our main insight being the implication for the use of predictive models in optimal policy.

lems with rich cross-sectional heterogeneity. We show that in these environments, reduced-form prediction can be valuable even though it is not a substitute for structural modeling. Its role is to estimate the sufficient statistics that policy rules can use for more accurate targeting. Structural approaches remain necessary for performing disciplined counterfactuals and welfare analysis, while predictive models are valuable because they are able to recover heterogeneity that is difficult to capture with structural models alone. Our analysis provides a template for how to combine these two different types of modeling approaches in a policy problem. In this sense, the paper speaks both to macroprudential regulation and to a broader class of economic policy problems in which governments observe increasingly rich data but still need structurally disciplined rules for intervention.

Related Literature. Our analysis relates to several areas of the literature. First, we connect to the literature on applications of machine learning and deep learning to finance. Gu, Kelly and Xiu (2020, 2021), Kozak, Nagel and Santosh (2020), Nagel (2021), and Bryzgalova, DeMiguel, Li and Pelger (2023) apply machine learning techniques to the canonical empirical asset pricing problem of valuation in the cross-section of assets. Similarly, Chen, Pelger and Zhu (2024) introduce a deep learning architecture for modeling stock returns. Giglio, Kelly and Xiu (2022) review the intersection of empirical asset pricing and machine learning. Gabaix, Koijen, Richmond and Yogo (2024) use holdings data together with a variety of deep learning sequence models (e.g., Word2Vec and BERT) and recommender systems to estimate asset embeddings, and Gabaix, Koijen, Richmond and Yogo (2025) apply these to explain variation and volatility in credit spreads. Chen, Cheng, Liu and Tang (2023) study transfer learning problems to bridge structural and machine learning models in asset pricing. Dolphin, Smyth and Dong (2022) and Sarkar (2025) also construct asset embeddings, respectively from return data and textual sources. Bybee (2025) generates expectations from large language models and studies bubble dynamics. Cohen, Lu and Nguyen (2026) apply a recurrent network to predict the directions of mutual fund trades. Methodologically, our paper employs graph neural network architectures (Scarselli, Gori, Tsoi, Hagenbuchner and Monfardini 2008; Hamilton, Ying and Leskovec 2017; Xu, Hu, Leskovec and Jegelka 2018; Wu, Pan, Chen, Long, Zhang and Yu 2020) to learn representations from the relational structure of asset holdings data, and we show that explicitly modeling the graph structure of overlapping portfolios yields large predictive gains.[2]

Our applications of machine learning models for financial regulation relate to a longstanding debate, from Koopmans (1947) to Friedman (1953) and Lucas (1976), about whether predictive relationships that are not themselves structural can inform policy. We show that sufficiently capable predictive models can have a role in the macro-regulatory toolbox, not as substitutes for structural models, but as complements that recover high-dimensional targeting statistics while structure disciplines counterfactuals.

Second, we connect to the literature on fire sales, and macroprudential regulation and ex-post

[2]See also Elliott, Golub and Jackson (2014) and Acemoglu, Ozdaglar and Tahbaz-Salehi (2015) for theory highlighting the importance of the network structure of positions for financial stability.

interventions. Theoretical contributions include Bernanke and Gertler (1986), Williamson (1988), Kiyotaki and Moore (1997), Caballero and Krishnamurthy (2001), Lorenzoni (2008), Uhlig (2010), Bianchi (2011, 2016), Stein (2012), Farhi and Werning (2016), Chari and Kehoe (2016), Bianchi and Mendoza (2018), Dávila and Korinek (2017), and Clayton and Schaab (2022, 2025). On the empirical side, prior work has focused on particular variables to explain and predict the occurrence and magnitude of fire sales, such as leverage and maturity mismatches (Fisher 1933; Kindleberger and Aliber 1978; Brunnermeier and Oehmke 2013; Schularick and Taylor 2012; Adrian and Shin 2014; Adrian and Brunnermeier 2016; Acharya, Pedersen, Philippon and Richardson 2017; Krishnamurthy and Muir 2025) as well as investor composition (Brainard and Tobin 1968; Coval and Stafford 2007; Haddad, Moreira and Muir 2021; Coppola 2025; Fang, Hardy and Lewis 2025): the empirical contribution of this paper asks whether there is additional value from a more agnostic approach that does not impose an ex-ante focus on particular dimensions of the data.

Third, we relate to literature studying the deployment of machine learning and large-scale data for economic analysis and policy problems more broadly, including Einav and Levin (2014b,a), Kleinberg, Ludwig, Mullainathan and Obermeyer (2015), Athey (2017, 2018), Mullainathan and Spiess (2017), Kleinberg, Lakkaraju, Leskovec, Ludwig and Mullainathan (2018), Athey and Imbens (2019), Gillis and Spiess (2019), Farboodi and Veldkamp (2020), Athey and Wager (2021), and Fuster, Goldsmith-Pinkham, Ramadorai and Walther (2022). Relatedly, a recent literature in macroeconomics has highlighted how empirically identified causal effect estimates can be used to inform policy (Nakamura and Steinsson 2018, McKay and Wolf 2023) and has developed sufficient-statistics approaches to macro policy problems (Barnichon and Mesters 2023). The question of how non-policy invariant relationships should be exploited by regulators also has earlier conceptual parallels, for instance by Barro and Gordon (1983) in the context of the Phillips curve. Lastly, the model and information design aspect of our theoretical approach connects to work on stress testing (Shapiro and Skeie 2015; Faria-e Castro, Martinez and Philippon 2017; Goldstein and Leitner 2018; Leitner and Williams 2023; Orlov, Zryumov and Skrzypacz 2023; Parlatore and Philippon 2025).

2 Economics-Informed Deep Learning for Holdings Data

We begin our analysis by showing how to design deep learning models that are effective in learning from portfolio holdings data, informed by the nature of the economic environment. The defining feature of holdings data is rich relational structure, meaning that the relationships between entities in the data (i.e., investors and assets) are a core part of the information contained in it. Formally, we represent portfolio holdings data as a bipartite graph connecting investors and assets, with positions as edges. We build a graph transformer, an attention-augmented graph neural network that learns by iteratively propagating information along this holdings network (Scarselli et al. 2008; Hamilton

et al. 2017; Wu et al. 2020).[3]

We emphasize three properties that distinguish this architecture from alternative econometric approaches. First, the model learns directly from the network structure of overlapping portfolios: through an iterative process known as *message passing*, each investor's latent representation (its embedding) updates based on the embeddings of the assets it holds, and vice versa. This process gradually propagates information through the network of portfolios as the model is trained and evaluated. As we show in our empirical analysis, the ability of the model to learn from the relational structure of the data is the single largest source of its predictive advantage. Second, the model is permutation-invariant, meaning that its predictions do not depend on the arbitrary ordering convention applied to investors or assets to represent portfolios, unlike sequence-based architectures in which token order is essential. Third, because all of the model's parameters are shared across nodes in the graph, this architecture is inductive: it can produce informative embeddings and forecasts for new, unseen investors or assets without retraining, as we verify empirically.

Estimation Objective. The model's core task is to learn how each investor in the financial system trades each asset in response to the prevailing state of the world. This forecasting task is of direct empirical interest for financial stability monitoring, stress testing, and systemic risk measurement. As we discuss in Section 4, it also has a natural interpretation in terms of the targeting statistics that are relevant for informing optimal policy.

Formally, we let $\mathcal{G}_t$ denote the observed holdings data at time t: this is the full cross-section of who holds what, and how much, at the end of period t. We model $\mathcal{G}_t$ as a graph connecting investors to assets via their positions. We also let ζ_{t+1} be a scalar summary of the aggregate state at period $t+1$: this is a single number that summarizes the extent to which the period represents a stress regime or a calm regime. In our implementation, we use the Financial Stress Index (FSI) published by the St. Louis Fed as our aggregate state summary ζ_{t+1}, which is the first principal component of 18 time series indicators of financial market stress including the VIX and yield spreads.

We distinguish between market-value positions $w_{ia,t}$ and notional-quantity positions $q_{ia,t}$. The position $w_{ia,t}$ is the market value of investor i's holdings in asset a at time t, while $q_{ia,t}$ denotes the notional quantity (shares for equities, par value for bonds). For each investor-asset pair (i, a), we define the trade outcomes

$$y_{ia,t+1} = \log\left(\frac{q_{ia,t+1}}{q_{ia,t}}\right), \tag{1}$$

which capture the pure quantity change in investor i's position in asset a between the end of period t and the end of period $t+1$.[4] The central estimation target is the conditional expectation of trading

[3]Deep learning models on graphs have achieved strong results in domains where relational structure is paramount, including protein structure prediction (Jumper et al. 2021), traffic forecasting (Derrow-Pinion et al. 2021), and drug discovery (Jiang et al. 2021).

[4]We adopt the log transformation to facilitate the numerical training of the model, as it places buys and sells on a symmetric scale: i.e., doubling a position and halving it yield targets of equal magnitude with opposite sign. In our implementation, we clip the values of $y_{ia,t+1}$ to the range $[-5, 5]$ to limit the influence of

behavior given this state information:

$$\hat{y}_{ia,t+1} = \hat{\mathbb{E}}\left[\underbrace{y_{ia,t+1}}_{\text{Trade from } t \text{ to } t+1} \;\middle|\; \underbrace{\mathcal{G}_t}_{\text{Positions at } t}, \underbrace{\zeta_{t+1}}_{\text{Aggregate state}} \right] \quad \text{for all } (i,a). \tag{2}$$

That is, for each investor-asset pair (i, a), the model takes as inputs the current cross-section of portfolio positions $\mathcal{G}_t$ and a realized (or scenario) value of aggregate financial stress ζ_{t+1}, and produces a (conditional) forecast of the quantity change in each position over the subsequent period. Hence, given the configuration of the financial system at time t, these conditional expectations describe how each investor adjusts each position for a given level of aggregate financial stress.

Several features of the estimation target defined in equation (2) are worth emphasizing. First, the model does not forecast the aggregate state ζ_{t+1} itself. Rather, it takes ζ_{t+1} as given, either as a real-time observable that the regulator can measure when designing the intervention, or as a hypothetical stress scenario that the regulator can vary to produce scenario-contingent forecasts. Given this aggregate state ζ_{t+1}, the model then predicts the cross-sectional pattern of trading responses conditional on that state. Second, the information set $\mathcal{G}_t$ consists entirely of time-t observables (current portfolio positions as well as the corresponding investor and asset characteristics) while the prediction target $y_{ia,t+1}$ is a forward-looking quantity that will be realized in the future. To make this concrete, suppose we stand at the end of 2019. The model sees every position held at that date, and it is told the level of financial stress in the coming quarter, 2020Q1, when the Covid-19 shock hit. It then forecasts how each investor will trade each of its positions during that quarter.

We train the model on the full cross-section of trading behavior $\{y_{ia,t+1}\}$ rather than crisis-period forced liquidations directly, for two reasons. First, forced liquidations are not directly observable in the data: realized trades reflect the superposition of many motives, and isolating forced selling necessarily requires imposing additional assumptions upon the raw data. Second, aggregate fire sale episodes (and financial crises more generally) are rare events, so a model targeting crisis liquidations directly would face a severe small-sample problem. Training instead on the broader task of predicting all trading behavior exploits far more data and learns general patterns of portfolio adjustment that are informative about the specific adjustments constituting fire sales.[5] Our approach learns general-purpose representations of investors and assets from the full distribution of trading behavior, and we then verify empirically that these representations are informative about fire sale dynamics by showing that the learned embeddings explain the cross-section of asset price dislocations during

extreme outliers from very small initial positions. The target $y_{ia,t+1}$ is only defined for continuing positions with $q_{ia,t} > 0$: new entries ($q_{ia,t} = 0$) are excluded from model training and evaluation. Thus, the baseline forecasting task focuses on adjustments of existing positions, while entry into entirely new positions is left outside the main target.

[5] A helpful intuition is that while there are only very few aggregate crises in the data, there is much more information available on how investors tend to react to shocks more generally, including idiosyncratic shocks: by observing the full cross-section of investors and assets over time, we can make full use of all this information to learn behavior.

market stress episodes (as shown in Section 3). When we consider empirical proxies for forced sales in Section 3.5, the model predicts these at least as well as trades overall.

Core Architecture Specification. We now formally specify the graph transformer architecture that we use to learn the conditional expectation functions in equation (2). The holdings graph is $\mathcal{G}_t = (\mathcal{I}_t, \mathcal{A}_t, \mathcal{E}_t)$, where $\mathcal{I}_t = \{1, \ldots, I_t\}$ is the set of investors, $\mathcal{A}_t = \{1, \ldots, N_t\}$ is the set of assets, and $\mathcal{E}_t$ is the set of edges (i.e., positions). Whenever investor i holds asset a at time t, then $(i, a, w_{ia,t}, q_{ia,t}) \in \mathcal{E}_t$.[6] We write $\mathcal{V}_t = \mathcal{I}_t \cup \mathcal{A}_t$ as the set of all nodes, both assets and investors, and let $\mathcal{N}_t(v)$ be the neighborhood set for a given node v, which is the set of all nodes connected to v at time t. For assets, $\mathcal{N}_t(v)$ corresponds to the investors holding the asset, while for investors this corresponds to the set of assets in the investor's portfolio. Each node $v \in \mathcal{V}_t$ also has an associated set of characteristics $x_{v,t}$. The characteristics can be either numerical (e.g., the total amount outstanding of a given security) or categorical (e.g., the type of institutional investor).[7]

The model learns latent vector representations of each investor and asset called embeddings, following the language typically used in the deep learning literature. These latent embeddings are built in a series of steps which aggregate from progressively higher-order connections over the network. An initial set of embeddings $h_{v,t}^{(0)}$ maps each node's observable characteristics $x_{v,t}$ into a d_h-dimensional vector through a learned function ϕ:

$$h_{v,t}^{(0)} \;=\; \phi(x_{v,t})\,. \tag{3}$$

These initial embeddings are then enriched through L successive layers of message passing, the core of the model, in which each node's representation is updated using the representations of its neighbors in the holdings network.

At each message-passing layer ℓ, every node first constructs a *message* by passing its prior-layer embedding through a learned function $M^{(\ell)} : \mathbb{R}^{d_h} \to \mathbb{R}^{d_h}$:

$$M_{v,t}^{(\ell)} = M^{(\ell)}(h_{v,t}^{(\ell-1)}). \tag{4}$$

Each node then aggregates the messages of its neighbors into a weighted average,

$$m_{v,t}^{(\ell)} = \sum_{u \in \mathcal{N}_t(v)} \alpha_{vu,t}^{(\ell)}\, M_{u,t}^{(\ell)}, \tag{5}$$

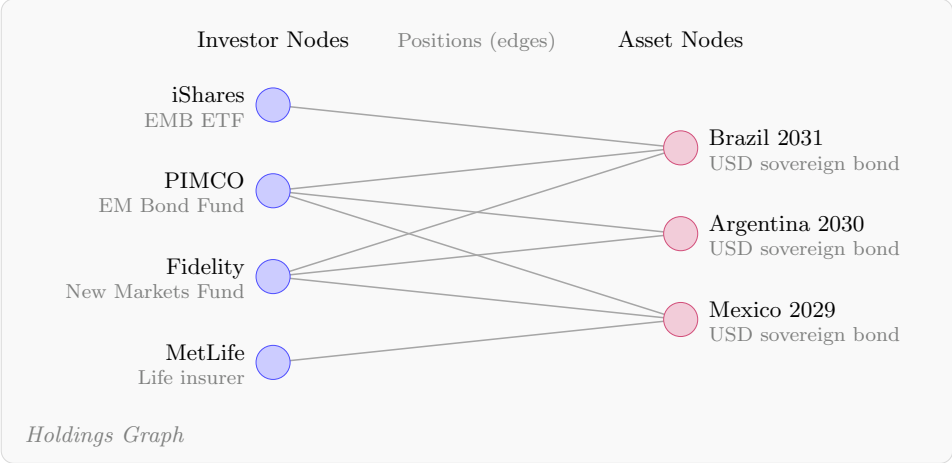
where the weights $\alpha_{vu,t}^{(\ell)}$ sum to one over $\mathcal{N}_t(v)$ and come from a multi-headed attention mechanism

[6] The positions can alternatively be written as $w_{ai,t} = w_{ia,t}$ and $q_{ai,t} = q_{ia,t}$. We allow for both notations so as to keep the rest of the formal architecture specification symmetric.

[7] Categorical features are embedded into a vector space of dimension d_c using functions that are learned jointly with the rest of the model's parameters, and then they are concatenated with numerical features to form a single characteristics vector $x_{v,t}$.

that learns which positions should receive greater or lesser weight (Vaswani et al., 2017).[8]

Figure 1: **Holdings data as a bipartite graph: illustrative example**



Notes: Stylized example of the holdings graph. Blue nodes are investors, pink nodes are assets, and each line is a position. This is an illustrative example constructed without using actual positions data.

The embeddings are then updated by combining each node's current embedding with its aggregated message via a learned function $U^{(\ell)} : \mathbb{R}^{2d_h} \to \mathbb{R}^{d_h}$:

$$h_{v,t}^{(\ell)} = U^{(\ell)}(h_{v,t}^{(\ell-1)}, m_{v,t}^{(\ell)}). \tag{7}$$

After L layers, a readout layer $\rho : \mathbb{R}^{d_h} \to \mathbb{R}^{d_e}$ produces the final asset and investor embeddings, which are the ones we refer to in subsequent sections unless otherwise specified:

$$e_{v,t} = \rho(h_{v,t}^{(L)}). \tag{8}$$

To build intuition, Figure 1 illustrates the learning over the portfolio graph structure with a small hypothetical example. Investors and assets form the two sets of nodes, and each position links an investor to an asset it holds. Consider the dollar-denominated Argentine sovereign bond maturing in 2030. Its immediate neighbors in the graph are the investors that hold it: in the

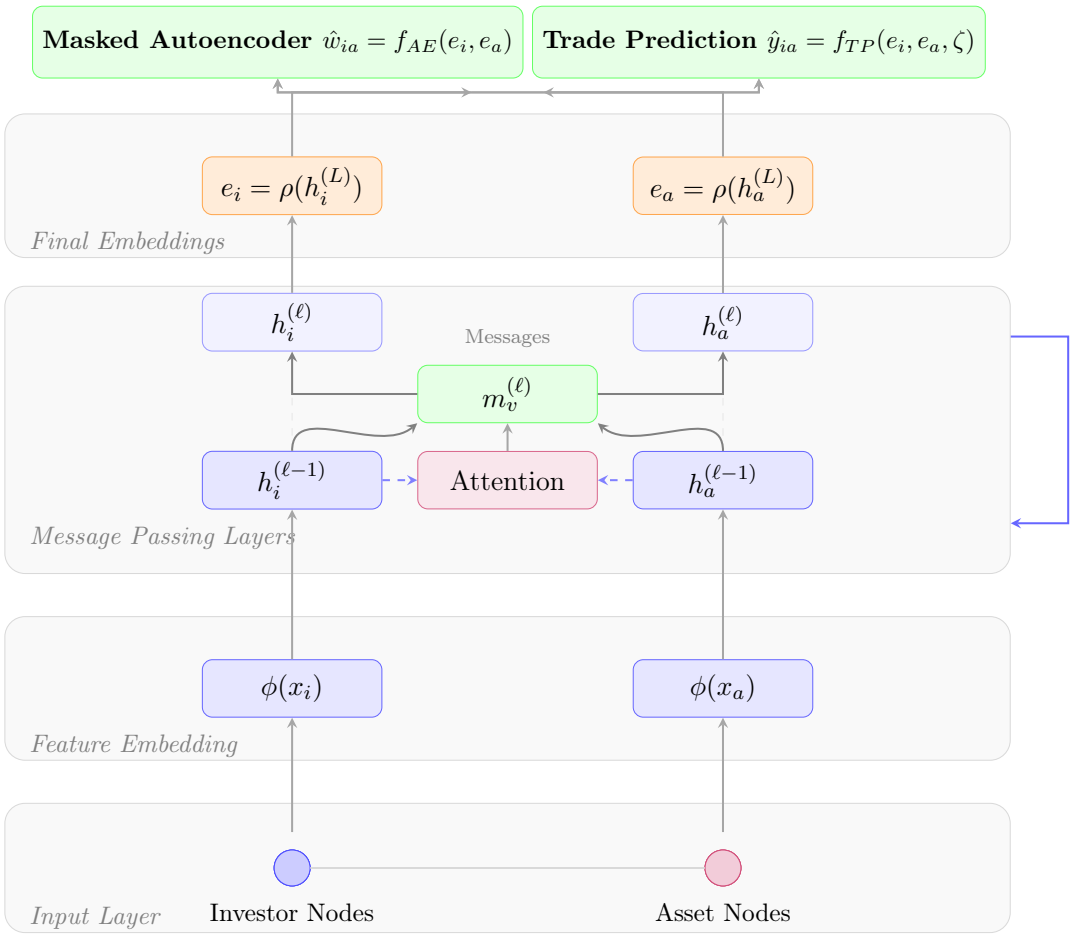
[8]We use S_A attention heads with scaled dot-product queries and keys in each message passing layer. Each head has its own query and key projections of the node embeddings and its own value projection of the neighbors' messages. We incorporate the economic importance of each position (an economic prior) through the normalized edge exposure

$$\tilde{w}_{vu,t} = \frac{w_{vu,t}}{\sum_{k \in \mathcal{N}_t(v)} w_{vk,t}} \qquad \text{for } u \in \mathcal{N}_t(v), \tag{6}$$

which enters the attention weights as an additive log-bias, so that a node's attention is concentrated on its economically important positions. Equation (5) shows a single attention head for simplicity; Appendix B gives the full multi-head computation, including the query, key, and value projections.

figure, these are illustrative emerging-market bond funds managed by PIMCO and Fidelity. One step further out are the other assets in those investors' same portfolios, such as specific Brazilian and Mexican sovereign bonds. One more step out are the other investors that hold those assets, e.g. an emerging-market bond ETF and a life insurer. Each round of message passing extends the Argentine bond's representation by one step: after the first layer, it reflects the bond's holders. After the second, it also reflects information about the other assets those holders own; after the third, the other investors in those assets. The final embedding $e_{v,t}$ of the Argentine bond therefore draws on all of the connections shown in the figure, not just on the bond's own characteristics.

Figure 2: **The graph transformer architecture**



Notes: We visualize the graph transformer architecture. The observable characteristics associated with each investor and asset (i.e, each node in the graph) are mapped into initial embeddings via a learned function ϕ. These embeddings are then refined through L rounds of message passing in which nodes exchange information with their neighbors in the holdings network. The multi-headed attention mechanism determines the weight each position receives. A layer ρ produces final embeddings e_i and e_a, which feed into two heads: a masked autoencoder for pretraining on co-holding patterns, and a trade prediction head for the supervised forecasting task. All learned functions (ϕ, $M^{(\ell)}$, $U^{(\ell)}$, ρ, and the prediction heads) are parameterized as feed-forward neural networks.

Training Strategy. We train the model in two sequential phases (Devlin et al. 2019; He et al. 2022). In the first phase (self-supervised pretraining), we train the graph transformer backbone using a masked autoencoder (MAE) objective: the model randomly masks a subset of edges in the graph and learns to reconstruct the masked position sizes. This forces the model to learn which positions tend to co-occur—discovering latent similarities among investors and assets from the pattern of who holds what—before it sees any information about trades. In the second phase (supervised training), we freeze the autoencoder head and fine-tune the graph transformer backbone using the supervised trade prediction head, which provides our estimates of the conditional response functions defined in equation (2).

Prediction Heads. The supervised prediction head is a hurdle model, motivated by the fact that approximately 40% of existing positions see no change in quantity in any given quarter. Without separating the two margins, the model would effectively average over the zero and nonzero trades, performing poorly at both predicting which positions are traded and predicting the size of trades that do occur. The hurdle model decomposes the prediction into two components: (i) an extensive margin prediction of whether a trade occurs at all, and (ii) an intensive margin prediction of the trade's magnitude, conditional on it occurring. The trade prediction head uses the concatenated investor and asset embeddings $(e_{i,t}, e_{a,t})$ together with the aggregate state ζ_{t+1} as inputs, branching into classification logits $\hat{c}_{ia,t+1}$ and a regression prediction $\hat{m}_{ia,t+1}$. The combined prediction is:

$$\hat{y}_{ia,t+1} = \sigma(\hat{c}_{ia,t+1}) \cdot \hat{m}_{ia,t+1}, \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function. The supervised objective combines binary cross-entropy for the extensive margin and mean squared error for the intensive margin. The MAE and supervised losses, along with the full specification of the training procedure, are detailed in Appendix B. We optimize all model parameters using the AdamW optimizer (Kingma 2014; Loshchilov and Hutter 2018).

The model architecture is summarized in Figure 2. We validate these architectural choices empirically in Section 3.2, where ablation experiments confirm that the graph structure is the primary source of predictive information.

3 Using the Model for Stability Monitoring Tasks

We implement and evaluate the graph transformer model on portfolio holdings microdata covering nearly $40 trillion in non-bank intermediary wealth by the end of the sample. We assess the model along three dimensions relevant for financial stability monitoring: forecasting institutional trading behavior (particularly during crisis events), explaining the cross-section of asset price dislocations during stress episodes, and constructing effective systemic risk indicators at the institution level. We also evaluate the model's ability to generalize outside of its immediate training data by providing forecasts for unseen investors and assets, and we show how to quantify uncertainty in the model's

Table 1: **Summary statistics**

	Mean	Std	P10	Median	P90	*N*
Panel A: Panel structure (per quarter)						
Investors	17,751	6,290	9,121	17,408	25,389	72
Assets	33,302	11,623	16,060	35,178	46,609	72
Positions	2,310,818	1,172,503	900,260	2,043,472	4,091,039	72
Positions per investor	122.8	22.6	98.7	115.8	160.5	72
Holders per asset	65.6	13.2	55.7	58.5	88.2	72
Panel B: Holdings characteristics (pooled)						
Log holding value	12.60	3.58	9.44	13.11	16.01	166,378,929
Holding value ($M)	5.8	29.9	0.0	0.5	9.0	166,378,929
Trade target ($y_{ia,t}$)	0.005	0.899	-0.277	0.074	0.348	149,755,712
Panel C: Investor characteristics (pooled)						
Total AUM ($M)	790.5	4929.1	13.4	120.3	1259.0	1,270,517
Number of positions	130.26	323.75	5.00	49.00	282.00	1,278,103
Equity share	0.49	0.47	0.00	0.56	1.00	1,278,103
Bond share	0.33	0.45	0.00	0.00	1.00	1,278,103
Panel D: Asset characteristics (pooled)						
Total held value ($M)	418.9	2149.7	7.7	78.2	644.7	2,397,769
Number of holders	69.40	145.68	2.00	23.00	165.00	2,397,769
Coupon (%)	4.60	7.81	1.25	4.75	7.31	1,335,690
Years to maturity	9.10	9.93	1.37	5.83	24.30	778,007

Notes: Panel A reports summary statistics of the holdings data at the quarterly level, across 72 quarters from 2004Q1 to 2021Q4. Panels B–D report summary statistics pooled across all investor-asset-quarter observations. Holdings data are from Factset and Morningstar and cover non-bank financial intermediaries including mutual funds, ETFs, separate accounts, pensions, and insurance companies. Holding value is the dollar value of each position. "Total AUM" is total assets under management; "Total held value" is the aggregate dollar amount held by all investors in a given security.

forecasts by constructing prediction intervals around its point predictions.

3.1 Setting and Data: Non-Bank Financial Intermediary Portfolios

Since the Great Financial Crisis, tighter bank regulation (e.g., leverage and proprietary-trading constraints) has reduced dealer balance-sheet intermediation and pushed a larger share of risk-taking and liquidity transformation into non-bank financial intermediaries (NBFIs), such as money market funds, funds holding bonds and collateralized loan obligations, ETFs, insurers, and pensions. These institutions now intermediate a substantial fraction of tradable credit and duration risk. While non-bank intermediaries typically tend to be less levered than banks, they are subject to different fire-sale amplification mechanisms such as runnable or flow-sensitive funding and mark-to-market constraints, which have been the focus of a rapidly growing literature (Chen et al. 2010; Goldstein et al. 2017; Aramonte et al. 2023; Coppola 2025). Therefore NBFIs have become central to episodes where widespread price dislocations occur: for instance, acute liquidity needs by bond mutual funds played a key role in debt market disruptions in the first quarter of 2020 (Haddad, Moreira and Muir 2021; Ma, Xiao and Zeng 2022).[9]

For these reasons, we implement our empirical approach in the setting of non-bank financial intermediation. Specifically, we train the graph transformer model on quarterly security-level institutional holdings data from Factset and Morningstar, starting in the first quarter of 2004, which cover a broad range of non-bank intermediaries including mutual funds, ETFs, separate accounts, pensions, and insurance companies. The holdings data gives us observations of the positions graph $\mathcal{G}_t$ for each quarter's cross-section, and we also use it to construct the trade outcomes $y_{ia,t+1}$. Table 1 reports summary statistics for the holdings data. Across all quarters, the data covers the positions of roughly 37,000 unique non-bank financial intermediaries in 85,000 unique assets.

Model Training and Evaluation. We evaluate the model using two complementary out-of-sample strategies, and we report results under both. The first is a static split strategy, which randomly partitions quarters into training and validation sets. This approach maximizes the available training data and provides a pooled measure of predictive accuracy. The second strategy is a walk-forward evaluation, in which the model is retrained from scratch each year on an expanding window of past data, and then it is tested strictly on the subsequent year's quarters, consistent with how a model would be trained and evaluated in practice by a regulator. For example, the walk-forward forecasts for 2016 come from a model trained only on data through 2015. The walk-forward evaluation spans six annual windows from 2015 to 2020, with the final window providing a particularly salient test during a crisis event: the model is trained only on pre-crisis data and then evaluated on the Covid-19 stress episode (occurring in the first quarter of 2020).

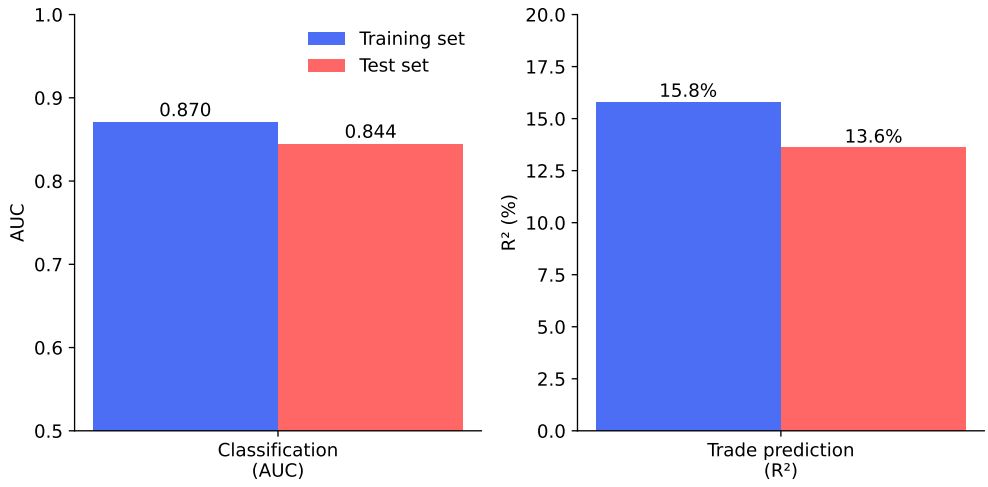
[9] Another notable recent stress episode that originated in the NBFI sector is the 2022 gilt market crisis in the United Kingdom, whose dynamics centered around Liability-Driven Investment (LDI) vehicles of UK pension funds.

Table 2: **Model hyperparameters**

A. Architecture	
Hidden dimension (d_h)	512
Embedding dimension (d_e)	128
Message-passing layers (L)	3
Attention heads (S_A)	8
Categorical embedding dimension (d_c)	16
Ensemble members (P)	4
Dropout rate	0.1
B. Training	
MAE pretrain epochs	50
Supervised fine-tune epochs	500 (early stopping)
Learning rate	10^{-3}
Weight decay	10^{-2}
Edge masking rate (MAE)	0.35

Notes: We list the hyperparameters used for our graph transformer architecture and in the two-phase training procedure.

Figure 3: **Performance metrics: static OOS split**



Notes: We plot the out-of-sample performance of the trade prediction hurdle model's two components under the static split: the classification AUC for predicting whether a trade occurs (left panel) and the regression R^2 on nonzero $y_{ia,t+1}$ targets (right panel). The blue bars show performance on the training set, while the red bars show out-of-sample performance on the validation set.

We construct the node feature vectors x_v using reference information from Factset, Morningstar, and GCAP. For assets, the feature vectors x_v include asset class, currency of denomination, amount outstanding, number of holders, average position size, standard deviation of position size, as well as bond sub-class and coupon for debt securities. For investors, they include institution type (such as open-end mutual funds, ETFs, separate accounts, etc.), manager style (including flags for active vs. passive portfolio management and strategy types), total AUM, number of positions, average position size, and standard deviation of position size.[10]

Hyperparameters are chosen as in Table 2. We set the hidden dimension to $d_h = 512$ and the final embeddings dimensionality to $d_e = 128$. We use $L = 3$ layers of message-passing with $S_A = 8$ attention heads per layer. We do not increase L beyond this number to prevent the over-smoothing problem, whereby the node embeddings would converge to similar values.[11] The final model ensembles over $P = 4$ independent learners.[12]

3.2 Predicting the Trading Behavior of Financial Intermediaries

We evaluate how well the graph transformer model forecasts intermediary trading behavior under both out-of-sample strategies. We evaluate performance on the extensive margin of trading by using the Area Under the Curve (AUC) of the model's predictions, which measures how well the model distinguishes positions that will be traded from those that will not, with an AUC of 1 indicating perfect separation and 0.5 indicating no discriminatory power. On the intensive margin, we measure the regression R^2 of the predictions. The model performs strongly on both the extensive and the intensive margins of trading. Under the static split evaluation strategy, the graph transformer achieves an out-of-sample classification AUC of 0.84 and an intensive margin R^2 of 13.6% on nonzero trades, with minimal degradation from training to validation sets (Figure 3). Under the walk-forward evaluation, the model achieves a pooled R^2 of 8.8% and AUC of 0.814 across the six annual windows (Figure 4, Table 3), with stable performance throughout the evaluation period. These values should be interpreted in the context of the nature of the task, which involves forecasting millions of individual trades each quarter, and trading at this granular level is very noisy. As we show below, they represent a five- to six-fold improvement in out-of-sample performance relative to simple econometric approaches that use just asset and investor characteristics.

[10] In principle, our architecture also allows for the use of global features that vary over time but not across nodes during message-passing: these can be introduced as vectors which enter message-passing in the same way for all nodes in a given time period.

[11] Intuitively, over-smoothing occurs for higher numbers of message-passing iterations since each consecutive iteration increases the receptive field of each node, i.e. the set of nodes that the final-layer embeddings attend to, and for high L values the receptive fields of all nodes converge to the largest possible field, which is the set of all nodes in the graph.

[12] We introduce dropout during training for additional regularization, with a dropout rate of 0.1. The AdamW optimizer uses a starting learning rate of 10^{-3} with weight decay of 10^{-2} and learning rate reduction on plateau. The self-supervised MAE pretraining phase runs for 50 epochs with a 35% edge masking rate; the supervised fine-tuning phase runs for up to 500 epochs with early stopping (patience of 5 epochs) based on the validation regression correlation. We train the model on four NVIDIA H100 GPUs.

The walk-forward R^2 is naturally lower than the static-split R^2 due to two factors. First, the walk-forward windows train on fewer quarters, reducing the model's effective training sample. Second, the random static split creates potential information leakage through the persistence of the holdings graph. The walk-forward design eliminates this channel entirely and therefore provides a more conservative measure of the model's predictive accuracy in a real-time regulatory deployment. In the 2020 window, the model achieves a walk-forward R^2 of 8.9%, and the per-quarter R^2 for 2020Q1, the peak of the Covid-19 market stress episode, is similarly high at 9.8%. The classification AUC is similarly high across both evaluation settings. The fully inductive design, sharing parameters across nodes, helps the model perform consistently on unseen investors and assets and limits overfitting to the training data.

Table 3: **Walk-forward temporal out-of-sample test, detailed statistics**

Test Year	Training Qs	AUC	Reg. Corr.	R^2 (%)	Comb. Corr.	Observations
2015	43	0.808	0.299	8.85	0.261	9,202,340
2016	47	0.815	0.290	8.30	0.230	10,311,507
2017	51	0.814	0.303	9.07	0.247	10,689,133
2018	55	0.809	0.274	7.46	0.225	11,627,911
2019	59	0.810	0.283	7.98	0.244	13,034,368
2020	63	0.825	0.304	8.92	0.262	13,358,675
Pooled	—	**0.814**	**0.297**	**8.80**	**0.244**	68,223,934

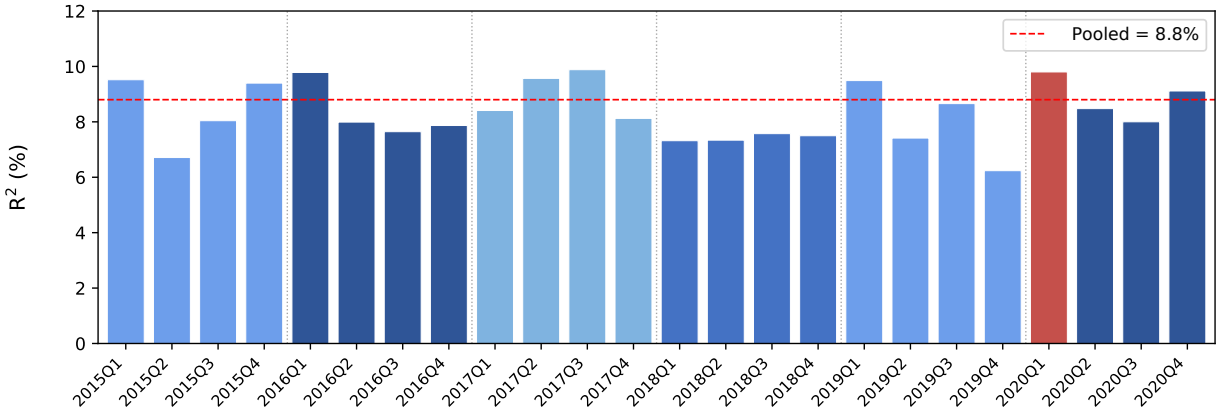
Notes: We show detailed results from the walk-forward out-of-sample test for the trade prediction task. The model is trained from scratch in each window. For each window, we show the number of training quarters. AUC measures the extensive margin classification of trade vs. no-trade. Regression correlation (Reg. Corr.) and R^2 are computed on non-zero target edges. Combined correlation (Comb. Corr.) uses combined predictions on the intensive and extensive margins from the full hurdle model.

We assess how well the model's trade predictions are calibrated by considering different quantiles of the outcomes. Figure 5a shows a binned scatter plot of predicted vs. realized trades $y_{ia,t+1}$: the OLS slope through the bin means is 0.85, confirming that the model's predictions are informative for all levels of trading outcomes (both buying behavior and selling behavior) and that they are approximately well-calibrated. Figure 5b provides a portfolio sort of realized $y_{ia,t+1}$ by decile of the model's predicted $\hat{y}_{ia,t+1}$: the monotonic increase from the bottom decile to the top decile further confirms that the predictions are informative across the full range of trade outcomes.

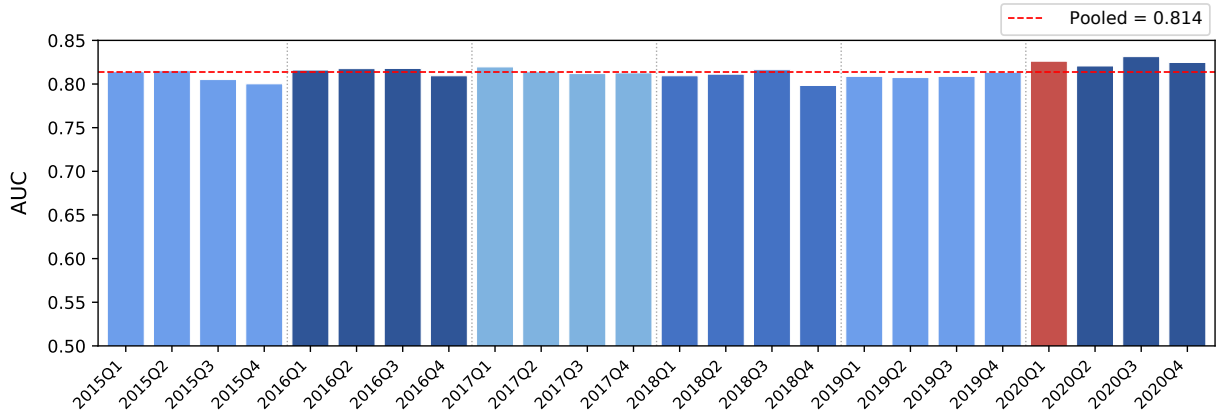
To place these performance metrics into context, Table 4 compares the graph transformer against a suite of benchmark models, all trained on identical trade prediction targets and evaluated out-of-sample on the same held-out quarters using the walk-forward evaluation strategy. We organize these comparisons to gradually isolate the sources of the model's predictive power. We start with simple forecasting rules and linear methods which use only observable investor and asset characteristics. We then proceed to more flexible nonlinear methods, and finally to deep learning architectures that

Figure 4: **Walk-forward out-of-sample trade forecasting performance, by quarter**

(a) Intensive margin predictions (R^2)



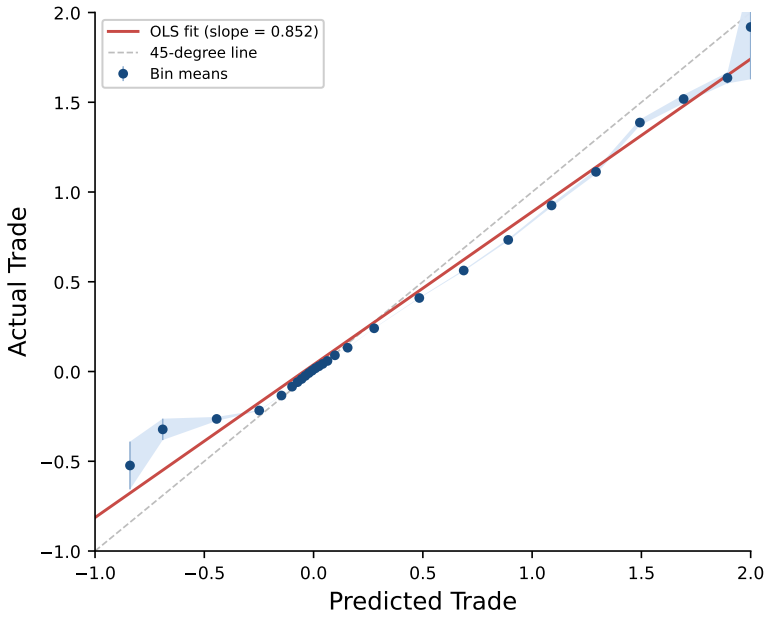
(b) Extensive margin predictions (AUC)



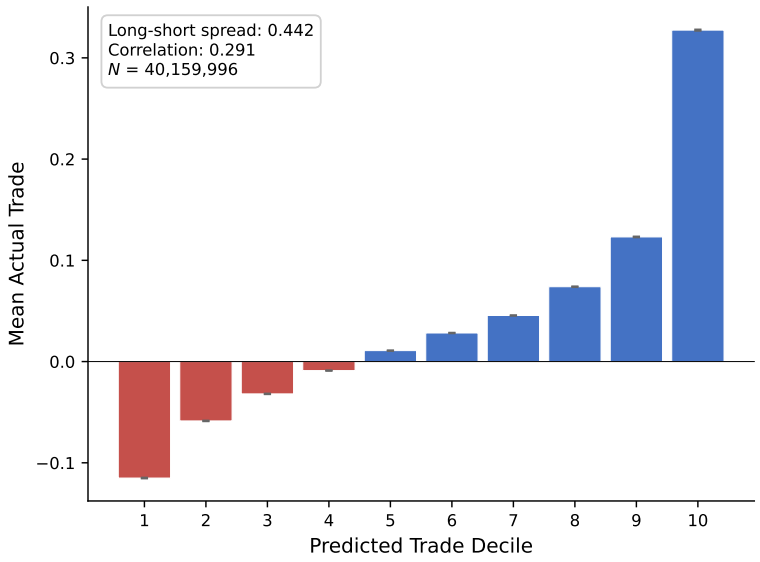
Notes: Per-quarter R^2 (top panel) and classification AUC (bottom panel) from the walk-forward trade prediction evaluation. Each bar represents one quarter that is strictly out of sample: the model used to generate each prediction was trained only on preceding quarters. Different shades of blue indicate different walk-forward windows. The red bar highlights 2020Q1, the onset of the Covid-19 crisis. Red dashed lines show pooled metrics.

Figure 5: **Assessing the calibration of the model's predictions**

(a) Actual vs. predicted trades $y_{ia,t+1}$



(b) Portfolio sorts



Notes: Panel A plots a binned scatter plot of out-of-sample predicted trades $\hat{y}_{ia,t+1}$ (horizontal axis) against realized trades $y_{ia,t+1}$ (vertical axis). Each dot is the mean of one bin; shaded band shows 95% confidence intervals. The red line is the OLS fit through the bin means. Panel B sorts observations into deciles based on the model's predicted trades $\hat{y}_{ia,t+1}$, and it plots the mean realized trade $y_{ia,t+1}$ within each decile. Error bars show 95% confidence intervals for the decile means. Realized outcomes are winsorized at the 1st and 99th percentiles to limit the influence of extreme trades from small initial positions. The long-short spread is the difference in mean realized $y_{ia,t+1}$ between the top and bottom deciles.

progressively exploit more of the relational structure of the holdings data.

We first consider a range of traditional econometric and statistical models, which reveal the difficulty of forecasting trades from characteristics alone. A persistence forecast—which predicts that each investor will repeat last quarter's trade in each asset—produces a negative out-of-sample R^2: investors who bought last quarter are, if anything, less likely to buy again. Ridge regression on investor and asset characteristics achieves an R^2 of just 1.7%, and gradient-boosted trees, which can capture nonlinear interactions among the same features, reach only 1.4%. The lesson from these baselines is that observable characteristics of investors and assets, even combined with flexible nonlinear methods, explain little of the high-dimensional cross-sectional variation in trading behavior. The predictive signal lies predominantly in the structure of who holds what, and how those positions relate to one another across the network.

Next, we consider embeddings approaches which learn latent representations of assets and investors. The simplest such approach is a recommender system based on factorization of the holdings matrix, which achieves a walk-forward R^2 of 1.6%. We then introduce deep learning models of our own that progressively exploit the holdings network's structure. Before going to the complete graph transformer, we also build a two-tower recommender system, a standard architecture in large-scale recommender systems, which we adapt to holdings data. The two-tower recommender learns separate latent embeddings for investors and assets and then scores their interactions via an inner product, in the spirit of collaborative filtering models. It achieves an R^2 of 4.2%, roughly a threefold improvement over the simpler baselines. The two-tower model is inductive and generalizes to unseen entities, and it implicitly captures some relational structure through the shared embedding space: in this sense, it implements some but not all of the features of the full graph transformer approach. The graph transformer captures the full network structure by explicitly propagating information across the holdings network. This yields an R^2 of 8.8%, a substantial improvement over all benchmarks.

For comparison, we also examine how an off-the-shelf pretrained LLM would do on this same task. Specifically, we structure a few-shot prediction task using Anthropic's Claude Opus 4.8 model. The model is prompted with a series of examples of historical portfolio data and trading outcomes, then presented with the same information as the other models and asked to generate forecasts of trading behavior. The off-the-shelf LLM achieves a negative R^2 in predicting trading. This is not unexpected, as this task concerns specialized data on portfolios and trading behavior that are unlikely to be in the training corpus of a generic LLM, but it highlights the benefits of developing domain-specific deep learning models for financial stability monitoring.

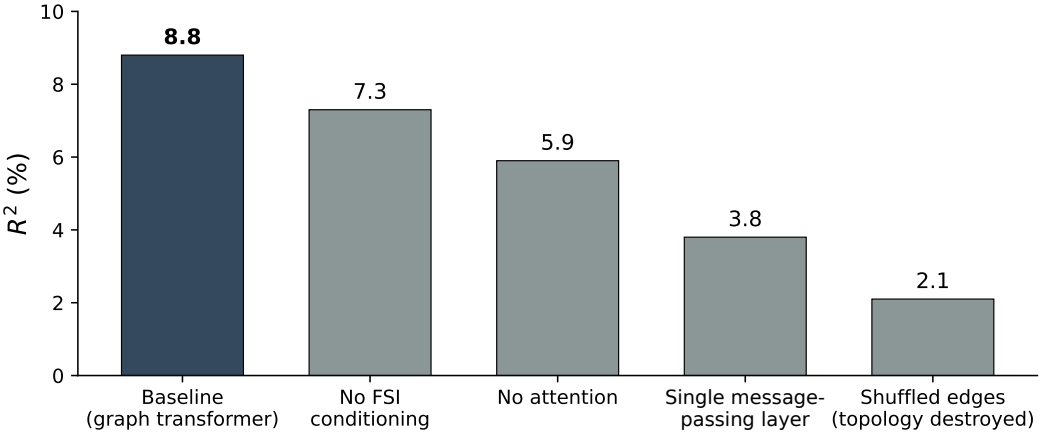
Ablation Analysis. To investigate and isolate the sources of the empirical performance of the graph transformer model, we perform a series of ablation experiments, where we remove features of the model one at a time. Figure 6 reports walk-forward R^2 from these ablation experiments. The ablations show that the network information (the topology of the network) is a central source of

Table 4: **Trade prediction: comparison with alternative benchmarks**

Model	Ours	Graph	Inductive	AUC	WF R^2 (%)
A. Traditional models					
Persistence (lagged $y_{ia,t+1}$)				0.70	<0
Ridge regression				0.59	1.7
Gradient boosting				0.72	1.4
B. Latent-representation models					
Matrix-factorization recommender system				0.72	1.6
Two-tower recommender system	✓		✓	0.80	4.2
Pre-trained LLM (Claude Opus 4.8, few-shot)			✓	0.68	<0
Graph transformer	✓	✓	✓	**0.81**	**8.8**

Notes: All models predict trades $y_{ia,t+1}$ on a held-out test set, using a walk-forward out-of-sample evaluation. "AUC" is the area under the ROC curve for predicting whether a trade occurs (nonzero target). "WF R^2" is the pooled R^2 on intensive margin trading predictions. "Ours" indicates models that we develop or adapt for holdings data as novel contributions of this paper. "Graph" indicates models that use the portfolio holdings network structure via message passing. "Inductive" indicates models that can generalize to unseen investors or assets without retraining.

Figure 6: **Architecture ablation study**



Notes: Walk-forward R^2 (%) for the baseline graph transformer and four ablation variants, each retrained from scratch in every expanding window. "No attention" replaces attention-based aggregation with symmetric mean pooling. "Single message-passing layer" restricts message passing to one-step connections over the network. "No FSI conditioning" removes the aggregate stress index ζ_{t+1} from the prediction head. "Shuffled edges" randomly permutes edge indices to destroy network topology while preserving node features and degree distributions.

the model's predictive performance: randomly shuffling edges, while preserving node features and degree distributions, reduces walk-forward R^2 from 8.8% to 2.1%, eliminating more than three quarters of the model's explanatory power and bringing the performance in line with models based only on characteristics. This confirms that the holdings network itself, not merely the flexible function approximation capacity of a neural network, is the primary driver of performance. Higher-order network structure is very important as well: restricting the model to a single layer of message passing, so that each node can only observe its immediate neighbors rather than aggregating information across higher degrees of network connections, reduces R^2 to 3.8%. The attention mechanism, which allows the model to weight neighbors by economic importance rather than aggregating them symmetrically, contributes meaningfully, with its removal lowering R^2 to 5.9%. Finally, removing aggregate stress conditioning (ζ_{t+1}), which forces the model to infer the macroeconomic state implicitly from the high-dimensional portfolio data alone, reduces R^2 to 7.3%. That the model retains substantial predictive power even without this input suggests that the holdings network itself contains information about the economic state that is relevant for forecasting trading behavior, though providing the state variable explicitly remains beneficial.

Interpreting the Out-of-Sample Results. We emphasize that our approach, both in the theory and in the empirical implementation, focuses on the particular amplification mechanism of fire sales. While this channel has played an important role in many episodes of financial stress, our analysis is not intended as a general model of other channels of financial crises. More generally, as with any predictive approach, the usefulness of our method requires some persistence in the underlying relationship between the observed state (in our case, portfolio positions) and subsequent outcomes (here, trading behavior). The graph transformer is designed to flexibly learn these relationships and it can be retrained as new data arrives, but it cannot guarantee unchanged predictive performance under an arbitrary shift in the underlying economic environment. Our out-of-sample exercises provide an empirical test showing that the predictive structure learned from holdings generalizes across the changes observed in our data, but this should not be interpreted as evidence of complete invariance of the model's predictive performance across all regime changes.

3.3 Explaining the Cross-Section of Asset Returns in Stress Episodes

We show next that the graph transformer's embeddings, though never trained on price or returns data, are informative about the cross-sectional variation in asset returns during stress episodes. This exercise addresses a natural concern with the results so far. Much of the trading in the data is not necessarily relevant from a regulator's perspective, as the focus of the theory is ultimately on fire sales during stress states and the pecuniary externalities generated by their price impact. A model could forecast trades well in an unconditional sense without having learned the dimensions of trading that drive fire sales. We shift focus from quantities to prices to test for this explicitly: if the model has picked up the forces behind fire sales, its internal representations should explain

which assets suffer the largest price dislocations in stress states. Fire-sale events are rare, so we do not fit the model to them during training. Instead, we ask whether the structure learned from the broader task of predicting trading behavior generalizes to explaining crisis-period price dislocations out of sample.

The question of whether the structure of positions explains stress-period bond returns has precedents in the literature. Traditional approaches to this question rely on collapsing the high-dimensional holdings data into a single low-dimensional summary statistic motivated by economic theory, and testing whether that statistic has predictive power for the cross-section of returns. For example, Coppola (2025) shows that the insurer share of a bond's investor base—i.e., the share of the bond held by insurance companies, a scalar summary of who holds the asset—predicts crisis-period returns because insurance companies are relatively stable holders that dampen fire-sale dynamics.[13] Such a summary is transparent, but by construction it discards most of the information in the holdings data: an asset typically has many holders, each with its own trading patterns and constraints. This paper takes a different methodological approach: we run the same kind of cross-sectional regression as this literature, but replace the hand-constructed summary with the full embedding vector that the model has learned for each bond. We compare the two approaches and highlight the advantages of each. We show that there is substantial additional information in the full high-dimensional holdings network, beyond what any single theoretically motivated summary captures.

Specifically, we focus on the cross-section of bond returns in the first quarter of 2020, during the Covid crisis—an episode featuring rapid and widespread disruptions in debt markets, partly driven by acute liquidity needs of investment funds and other non-bank intermediaries (Haddad et al. 2021; Ma et al. 2022). We focus on the Covid episode here since the out-of-sample evaluation is particularly clean: as it is at the end of the sample chronologically, we can train the model on an expanding window strictly on past data and then estimate our pricing regressions. However, our conclusions are not limited to the Covid quarter and they extend to other stress episodes.[14]

For each bond a, we let $r_{a,t}$ be its return during the quarter, and we let $e_{a,t-1}$ denote its prior-quarter asset embedding, constructed by applying the trained graph transformer to the preceding quarter's holdings and characteristics.[15] We estimate and compare cross-sectional specifications of

[13] We focus on this approach as a representative example of a broader class of analyses which extract information from the holdings data using low-dimensional indicators constructed using explicit economic hypotheses.

[14] In particular, we also train a model on post-GFC periods and then apply the learned embeddings in a cross-sectional regression of bond returns during the GFC using the specification in equation (10). In this additional test, the embeddings are computed from 2007Q4 holdings using a model trained only on post-2009 data. We find that the embeddings deliver 27 percentage points in adjusted R^2 gains incrementally to asset characteristics for the GFC cross-section of returns, broadly consistent with the results shown in Table 5.

[15] We use the sample of bonds that are reportable to TRACE, the transactions database maintained by FINRA. We apply to TRACE the data cleaning steps in Dick-Nielsen (2014), which are common in the literature. TRACE lets us observe the prices at which bonds transact. Quarterly bond returns account for price changes, coupons, and accrued interest, and they are constructed as in Coppola (2025).

the following form:

$$r_{a,t} = f(e_{a,t-1},\, X_{a,t-1}) + \varepsilon_{a,t} \tag{10}$$

$$r_{a,t} = f(\phi_{a,t-1},\, X_{a,t-1}) + \varepsilon_{a,t}, \tag{11}$$

where $X_{a,t-1}$ is a bond characteristics vector including bond rating and residual maturity and $\phi_{a,t-1}$ is the bond's insurer share. The first specification uses standard bond characteristics plus the graph transformer's learned embeddings. The second replaces the embeddings with the insurer share in the bond's investor base ($\phi_{a,t-1}$), a representative example of the traditional approach of using a single theory-motivated summary of the holdings data. This way, we can compare the relative explanatory power of these two approaches, which both use portfolio positioning to explain the cross-section of asset fire sales but in different (and, as we argue, complementary) methodological ways.

Table 5: **Explaining bond price dislocations**

	Embeddings Only	Embeddings and Characteristics	Traditional Approach
Cross-validated adjusted R^2			
Total	.39	.42	.22
Incremental to characteristics	—	.20	.01
Bond characteristics ($X_{a,t-1}$)		✓	✓
Model embeddings ($e_{a,t-1}$)	✓	✓	
Insurer share ($\phi_{a,t-1}$)			✓
Observations (N)	6,249	6,249	6,249

Notes: Out-of-fold adjusted R-squared from gradient boosting models for the cross-section of bond returns in 2020Q1. "Embeddings" uses the graph transformer's prior-quarter asset embeddings $e_{a,t-1}$. "Traditional" uses the insurer share of each bond's investor base as a single theory-motivated summary of the holdings data, following Coppola (2025). Bond characteristics include rating and residual maturity. The incremental R-squared is the difference in adjusted R-squared between each specification and the characteristics-only baseline.

Given the high-dimensional nature of the embeddings $e_{a,t-1}$, we estimate a non-linear, nonparametric model: specifically, we estimate these specifications using gradient boosting with five-fold cross-validation. Table 5 reports the results. We report the out-of-sample adjusted R-squared both from a model which includes only the embeddings (column 1), a model which includes both the embeddings and bond characteristics (column 2), and a model which includes the insurer share and bond characteristics. For the latter two, we also report the incremental adjusted R-squared, which is the difference in the adjusted R-squared between the full model (including both $X_{a,t-1}$ and either $e_{a,t-1}$ or $\phi_{a,t-1}$) and a model including only characteristics. The incremental adjusted R-squared coming from the embeddings in the full model is 20 percentage points. The traditional approach—adding insurer share to the same characteristics—yields an incremental adjusted R-squared of just 1 percentage point from the insurer share relative to characteristics.

The graph transformer's embeddings thus deliver more than ten times the incremental explanatory power compared to a simple low-dimensional summary. This gap reflects the differences between the two approaches, which we view as two complementary methodological and economic perspectives. The insurer share collapses the entire holdings network into a single scalar per asset, motivated by a particular theoretical hypothesis. This approach has many advantages that the deep learning approach lacks, including the fact that it is highly transparent about the economic mechanism, and it offers relative ease in applying reduced-form identification techniques to probe for causality. The deep learning approach, on the other hand, can exploit a much richer set of information present in the high-dimensional holdings data in a more mechanism-agnostic way: the graph transformer's embeddings encode the high-dimensional structure of who holds each bond, how those holders are connected through overlapping portfolios, and what the resulting patterns imply for the concentration of selling pressure during stress. In Section 3.6, where we perform interpretability analyses on the trained model, we investigate the economic channels underlying the embeddings' explanatory power for the cross-section of crisis-time asset returns. We show, for example, that part of their explanatory power reflects a specific, economically interpretable dimension capturing the presence of correlated ("crowded") positions by flighty investors in illiquid assets.

3.4 Identifying Risky Financial Institutions

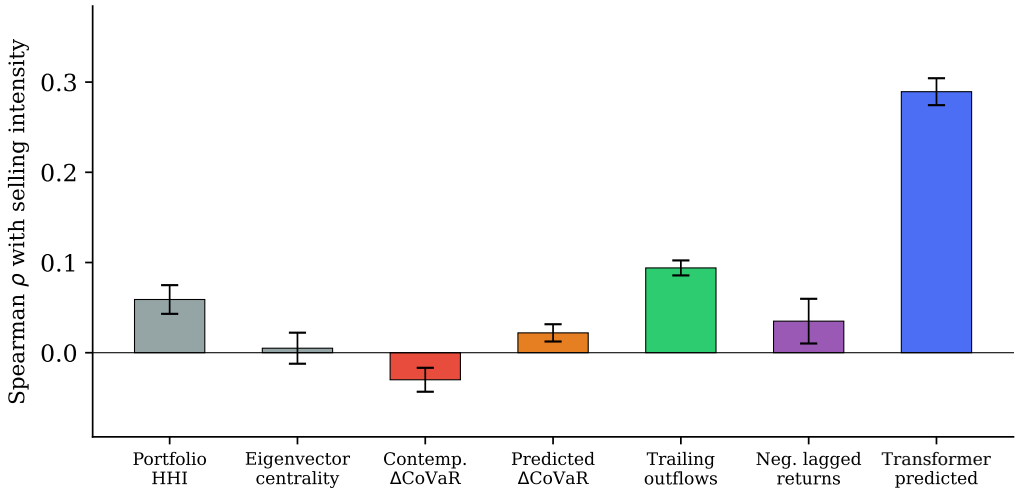
Next, we consider the perspective of a regulator who is interested in monitoring systemic risk using institution-level metrics. We ask how our deep learning based approach compares to selected existing measures of systemic risk and other indicators in identifying institutions that contribute most to selling pressure during stress periods in the non-bank setting that we consider. Our comparison therefore focuses on a specific policy-relevant outcome, the cross-sectional ranking of institutions according to their subsequent contribution to aggregate selling pressure. For this comparison, for each investor i and quarter t, we define the realized investor selling intensity as

$$S_{i,t} \;=\; \frac{1}{\mathrm{AUM}_{i,t-1}} \sum_{a} w_{ia,t-1} \, \max\!\left(-\frac{q_{ia,t} - q_{ia,t-1}}{q_{ia,t-1}},\, 0 \right), \tag{12}$$

which measures the fraction of assets under management that an investor sells in a given quarter. Scaling by $1/\mathrm{AUM}_{i,t-1}$ removes the mechanical size channel (larger investors sell more in absolute terms) and focuses on variation in selling behavior relative to size. For each out-of-sample quarter, we compute the Spearman rank correlation between each indicator and the realized selling intensity across investors.

We compare seven indicators in our non-bank setting. The first is an adaptation of the ΔCoVaR measure introduced by Adrian and Brunnermeier (2016), which quantifies the increase in the financial system's tail risk conditional on a given institution being in distress. We estimate ΔCoVaR using monthly fund NAV returns, with a median of 119 months of data per fund over the 2003–2018 period, defining the system return as the AUM-weighted average monthly return across all funds,

Figure 7: **Predicting institution-level selling pressure**



Notes: We plot the average Spearman rank correlation between each indicator and realized AUM-normalized selling intensity $S_{i,t}$ across investors, computed out of sample. Contemporaneous ΔCoVaR adapts the Adrian and Brunnermeier (2016) methodology to monthly fund NAV returns. Forward ΔCoVaR is a linear prediction from lagged observable fund characteristics. Trailing outflows uses the negative of the percentage fund flow lagged one year. Negative lagged returns uses the negative of the previous quarter's compound gross return. Graph transformer predicted selling uses the model's out-of-sample trade predictions.

analogous to the value-weighted financial sector index used by Adrian and Brunnermeier (2016). For each fund, we estimate the 5th-percentile quantile regression of the system return on the individual fund return.[16] The second metric is a forward measure of ΔCoVaR, obtained by regressing the contemporaneous ΔCoVaR_i on lagged fund characteristics using OLS estimated on training data only, and applying the fitted model out of sample.

The third measure is portfolio concentration, as measured by the HHI of holdings. The fourth is eigenvector centrality of the portfolio overlap network, a simple theoretically motivated summary statistic for network structure. The fifth is trailing fund outflows, motivated by the flow-driven fire sale channel of Coval and Stafford (2007): funds experiencing outflows must liquidate holdings, so we use the lagged percentage fund outflow as a predictor of selling intensity. The sixth is negative lagged fund returns, motivated by the flow-performance relationship that is well-documented in the NBFI sector (Chevalier and Ellison 1997): poor past returns trigger outflows, which in turn force selling. The seventh is the predicted selling intensity from our graph transformer model, $\hat{S}_i$, where $\hat{S}_{i,t} = \text{AUM}_{i,t-1}^{-1} \sum_a w_{ia,t-1} \max(-\hat{y}_{ia,t}, 0)$ uses the model's out-of-sample trade predictions.

[16] Let $R_{i,t}$ denote fund i's monthly gross return and $R_{\text{sys},t} = \sum_j w_j R_{j,t} / \sum_j w_j$ the AUM-weighted system return, where w_j is fund j's average assets under management. We estimate $Q_{0.05}(R_{\text{sys},t} \mid R_{i,t}) = \alpha_i + \beta_i R_{i,t}$ and compute $\Delta\text{CoVaR}_i = \hat{\beta}_i(\text{VaR}_{0.05}(R_i) - \text{median}(R_i))$, where $\text{VaR}_{0.05}(R_i)$ is the 5th percentile of fund i's monthly return distribution. Funds with fewer than 36 monthly observations are excluded. The resulting ΔCoVaR measures capture fund-level tail risk contributions estimated using approximately 20,800 funds with sufficient data.

Table 6: **Predicting institution-level selling pressure**

	Spearman ρ with selling intensity		
Method	All quarters	Stress	Calm
Portfolio HHI	0.059	0.045	0.060
	(0.007)	(0.007)	(0.007)
Eigenvector centrality	0.005	0.003	0.005
	(0.004)	(0.003)	(0.005)
Contemporaneous ΔCoVaR	-0.030	-0.018	-0.031
	(0.005)	(0.010)	(0.005)
Forward ΔCoVaR	0.022	0.038	0.021
	(0.006)	(0.001)	(0.006)
Trailing fund outflows	0.094	0.083	0.096
	(0.005)	(0.011)	(0.005)
Neg. lagged fund returns	0.035	0.006	0.038
	(0.009)	(0.000)	(0.010)
Transformer predicted selling intensity	**0.289**	**0.202**	**0.297**
	(0.015)	(0.013)	(0.015)

Notes: Each cell reports the average Spearman rank correlation between the given measure and realized selling intensity $S_{i,t}$, with standard errors across quarters in parentheses. Graph transformer predictions are out of sample. Stress quarters are defined by the St. Louis Fed Financial Stress Index (FSI) being above 1.5.

Figure 7 reports the results for the Spearman rank correlation (across investors) between each indicator and realized selling intensity. Table 6 also shows the same metrics separately for stress quarters (defined as quarters in which the financial stress index exceeds a threshold, FSI > 1.5) and non-stress quarters. The graph transformer achieves the highest rank correlation with selling intensity ($\rho = 0.29$). The next strongest measure is trailing fund outflows ($\rho = 0.09$), consistent with the flow-driven fire sale channel. All other measures achieve smaller correlations: negative lagged returns ($\rho = 0.04$), forward ΔCoVaR from fund characteristics ($\rho = 0.02$), portfolio HHI ($\rho = 0.06$), eigenvector centrality ($\rho = 0.01$), and contemporaneous ΔCoVaR ($\rho = -0.03$). The results are stable across stress and calm quarters.

These results highlight the value of holdings-based approaches for non-bank financial institutions. For non-bank intermediaries such as mutual funds, the cross-sectional variation in selling pressure contribution can depend importantly on the structure of who holds what: which funds hold concentrated positions in illiquid assets, which investor bases are dominated by flow-sensitive holders, where portfolio overlaps create channels for contagion, and so on. This is the type of information that the graph transformer extracts from the holdings network. Market-based systemic risk measures such as ΔCoVaR (Adrian and Brunnermeier 2016), MES (Acharya et al. 2017), and SRISK (Brownlees and Engle 2017) capture institutions' tail comovement or conditional capital shortfalls. These measures are informative about financial fragility among both banks and non-bank financial

institutions (Benoit et al. 2017, Acharya et al. 2025). Our comparison evaluates a different outcome, selling intensity, and it demonstrates the usefulness of holdings-based forecasts for this particular monitoring task. We view these forecasts as complementary to market-based monitoring.

Table 7: **Trade prediction: heterogeneity by investor type and asset class**

	AUC	WF R^2 (%)
A. Investor type		
Open-end mutual funds	0.811	8.08
ETFs	0.827	9.87
Non-fund institutions	0.785	5.72
Other funds	0.842	7.16
Active investors only	0.809	7.77
B. Asset class		
Common equity	0.753	8.47
Corporate bonds	0.778	7.91
Sovereign bonds	0.767	9.03
Fund shares	0.768	9.23
Other assets	0.866	7.22
All	**0.814**	**8.80**

Notes: Out-of-sample trade prediction performance decomposed by investor type (panel A) and asset class (panel B) using the graph transformer model, with a walk-forward out-of-sample evaluation. AUC measures classification of trading vs non-trading edges. Regression R^2 is computed on the nonzero-target subset.

3.5 Heterogeneity and Inductive Generalization

Next, we investigate heterogeneity in the predictive performance of the graph transformer, and we also test how well our approach generalizes inductively out of sample by conducting "cold start" tests where we form predictions for unseen assets and unseen investors. A natural question is whether the predictive ability of the model comes primarily from relatively more mechanical aspects of the data, such as by correctly assessing the trades of passive index-tracking investors. Table 7 addresses this directly: the "active funds only" row evaluates the model only on trades by funds that are classified as active (i.e., that do not passively track an index), and it shows an AUC of 0.809 and regression R^2 of 7.8% in the walk-forward evaluation. This is quantitatively very similar to the full-sample results. The model's predictive ability is thus not driven by mechanical passive rebalancing. Table 7 also further decomposes the model's out-of-sample trade prediction performance by investor type and asset class. Panel A shows results separately for different types of institutions. Panel B shows the results by asset class. The model performs well across investor types and asset classes, including those (like open-end funds and corporate bonds) that have been at the center of fire sale dynamics

in recent episodes of market stress.[17]

We conduct additional tests to validate the model's inductive generalization, i.e. its ability to make predictions for entities never seen during training. For each experiment, we remove a category of investors or assets from the training data and we retrain the graph transformer model from scratch, then evaluate the trained model on the complete (unmodified) holdings graphs. Table 8 reports the results on both the extensive and intensive margins of trade prediction. When removing random sets of investors or assets, the model's accuracy on seen edges is not strongly affected by the holdout: for these random holdout tests, the gap between held-out and seen walk-forward R^2 is small (e.g., 8.3% vs. 7.7% for the cold start test where we randomly exclude 20% of assets during training, and 8.0% vs. 6.5% for the test which excludes a random 20% of investors), indicating that the model generalizes to held-out investors and assets with limited degradation. We also consider a more demanding cold-start test where we drop an entire asset class (corporate bonds). We show that in this case, the performance drops meaningfully as expected, but the graph transformer continues to make meaningful predictions on the held-out group of assets: in this test, despite not having been trained on any corporate bond observations, the model still generalizes to the unseen asset class and achieves an R^2 of 3.4%. We interpret this as evidence of partial transfer of the learning to a genuinely unseen asset class: the model can form informative predictions for new investors and assets using patterns learned in related markets (for instance, through commonalities between how investors tend to trade in sovereign bond markets and corporate bond markets), but the predictive performance still meaningfully deteriorates when the new entities come from a materially different economic environment.

3.6 Economic Content of the Learned Embeddings

In this section we focus on "unpacking the black box" of the trained deep learning model: that is, can we understand the underlying economic channels that the model has learned in a way that is more transparent and interpretable? We show that we can at least partially shed light on the economic structure encoded by the trained model, which reflects several interpretable mechanisms.[18]

[17] The theory of fire-sale externalities that motivates our framework implies that the welfare-relevant trades are forced liquidations: i.e., sales driven by funding constraints rather than discretionary portfolio rebalancing. A natural question is whether the model's predictive ability extends specifically to these economically important trades. Following Coval and Stafford (2007), we define a trade as "plausibly forced" if it occurs at an open-end mutual fund experiencing concurrent negative fund flows: such funds must sell assets to meet redemptions, regardless of their assessment of asset fundamentals. We use Morningstar quarterly fund flow data to identify OEFs with negative percentage flows in crisis quarters. The graph transformer achieves an R^2 of 10.1% on forced trades, higher than the baseline performance for voluntary trades.

[18] In developing the economic hypotheses examined below, we also used the trained model itself as an exploratory tool, building upon the approach of Ludwig and Mullainathan (2024). Specifically, we examined contrasting pairs of investor and asset observations with similar characteristics but different model-predicted trading behavior, to help formulate candidate economic explanations. In ongoing work, we are formalizing and expanding on this approach, which uses the predictive model as a tool for exploring and understanding the underlying economic mechanisms systematically.

Table 8: **Cold-start inductive generalization test**

	% Edges	WF R^2 (%)		AUC	
Held-out group	held out	Held-Out	Seen	Held-Out	Seen
Corporate bonds	16.2	3.40	7.91	0.711	0.778
Random 20% investors	19.3	6.47	8.01	0.802	0.815
Random 20% assets	19.2	7.68	8.26	0.807	0.812

Notes: We show inductive generalization ("cold start") test using the graph transformer model. For each row, the model is trained from scratch with the indicated entity group entirely removed from all training graphs, then evaluated on the complete (unmodified) graphs. "Held-out" edges are those touching at least one held-out entity; "Seen" edges connect only entities present during training. We show extensive margin AUC and intensive margin R^2 from the walk-forward out-of-sample evaluation. The random investor test holds out 20% of individual investors (consistent across quarters); the random asset test holds out 20% of individual securities. For the corporate bonds row, "Seen" refers to the baseline model evaluated on all corporate bond observations.

To start, we show that an economically interpretable dimension of the embeddings learned by the model is their encoding of the interaction of position crowdedness and asset illiquidity. We highlight this dimension both because it provides an explicit example of an interpretable channel captured by the model, and because it is quantitatively salient. The fire-sale externality in Section 4 is proportional to the product of selling pressure (which is larger when positions are concentrated among flow-sensitive holders) and price impact (which is larger for illiquid assets). This mechanism has indeed been recognized and been a focus in the literature, both theoretically and empirically (Greenwood and Thesmar 2011, Anton and Polk 2014, Brown, Howard and Lundblad 2022).

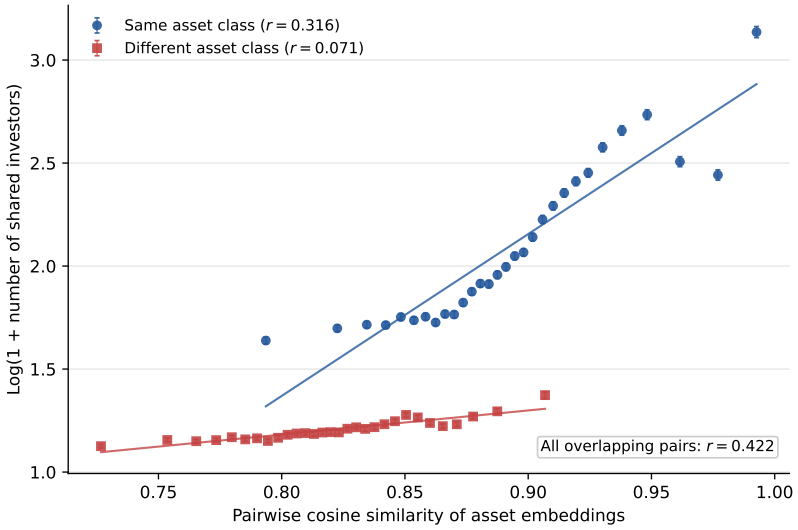
To test whether the embeddings pick up this mechanism on their own, we construct what we term a "fire-sale vulnerability" ($\text{FSV}_{a,t}$) index for each asset, a simple statistic measuring the product of the fraction of holdings attributable to flow-sensitive investors (open-end mutual funds and ETFs) and a measure of asset illiquidity (inverse average daily volume). Importantly, we never use this measure as part of the training of the deep learning model; instead we only use it ex-post as a means to interpret the economic content of the embeddings learned by the model. We then project the asset embeddings onto the direction that best predicts the $\text{FSV}_{a,t}$ score using ridge regression.

We find that the embeddings predict the index $\text{FSV}_{a,t}$ with a cross-validated R^2 of 11.5%, showing that the learned representations encode this information. Moreover, this direction accounts for a meaningful share of the embeddings' explanatory power for crisis-period asset returns. In the cross-sectional pricing specification of Section 3.3, removing the FSV direction from the embedding vector (i.e., orthogonalizing the asset embeddings to it) reduces the incremental R^2 from the embeddings by roughly 3 percentage points.[19] This shows that the model recovers a familiar and economically

[19] A simple placebo test using random embedding directions yields near-zero R^2. Figure A.II (in the Appendix) visualizes the asset embedding space, using a two-dimensional t-SNE projection of the embedding vectors, colored by quintiles of the $\text{FSV}_{a,t}$ score. High-vulnerability assets (Q5) cluster in distinct regions, consistent with the graph transformer learning how to discriminate assets according to this information.

important mechanism of fire-sale risk identified in the existing literature in an interpretable manner, even though the model is not trained to reproduce it.

Figure 8: **Asset embedding similarity predicts co-holding**



Notes: Binned scatter plot (30 equal-frequency bins) of pairwise co-holding (log number of shared investors) against pairwise asset embedding cosine similarity, for a sample of 3,000 assets restricted to pairs with at least one shared holder. Blue circles show same-class asset pairs; red squares show different-class pairs.

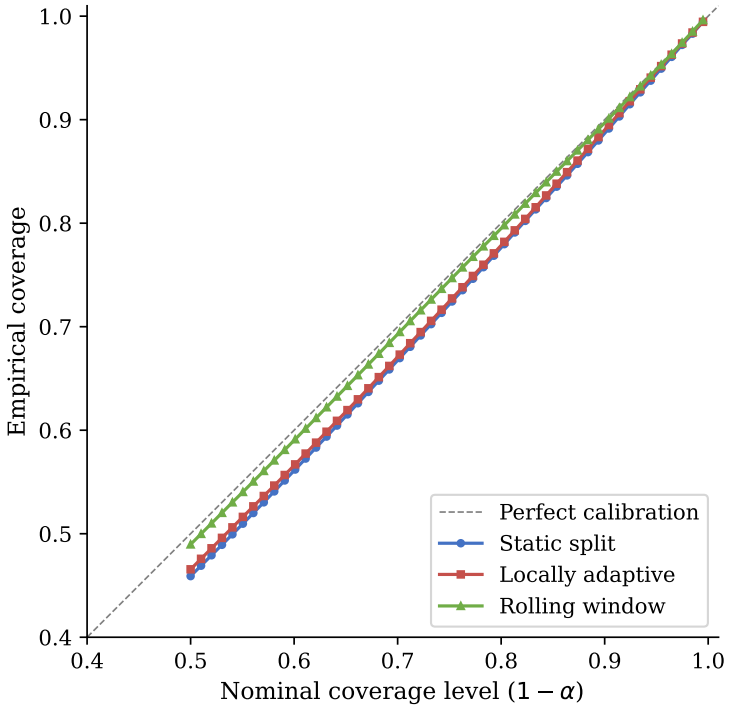
Next, we implement two additional exercises to corroborate that the embeddings capture economically meaningful structure beyond the fire-sale vulnerability dimension introduced above. First, we show that roughly half of the embedding variation reflects novel structure not captured by observable characteristics: when we jointly regress each embedding dimension on all observable characteristics, the variance-weighted R^2 is 0.54 for investor embeddings and 0.50 for asset embeddings. The remaining variation is learned from the holdings network itself.

Second, we show that the embeddings capture co-holding structure. Figure 8 plots the relationship between the pairwise cosine similarity of asset embeddings and a measure of pairwise co-holding for a random sample of assets. These two measures are positively correlated and this positive correlation is especially strong within same-class asset pairs ($r = 0.32$, $p < 0.001$), confirming that embedding similarity captures patterns of common ownership both within and across asset classes. Assets held by the same investors are likely to be traded together during portfolio rebalancing, and the graph transformer's learned embeddings internalize this common-ownership signal.

3.7 Quantifying Uncertainty in the Model's Predictions

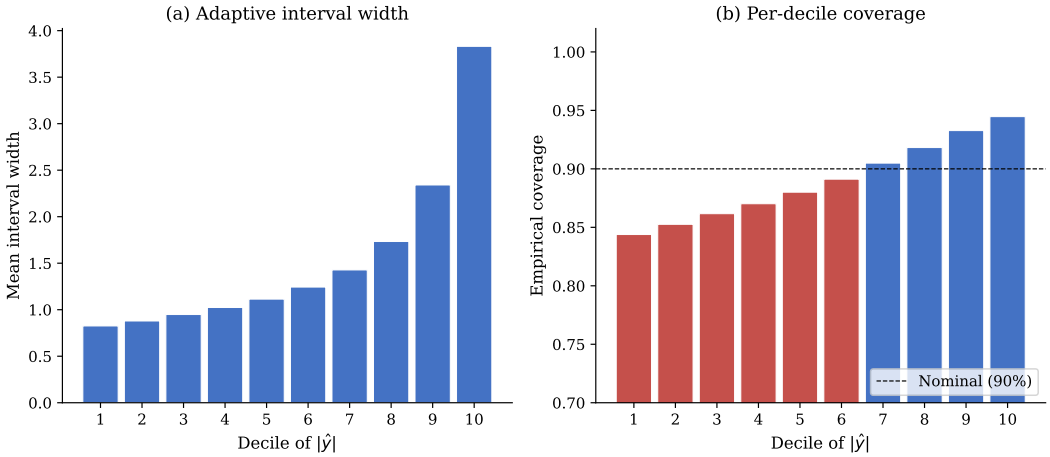
Point predictions are especially useful for regulatory decision-making if they are supplemented with measures of uncertainty. To provide useful ranges, we accompany each point forecast with a prediction interval using split conformal prediction methods (Vovk et al., 2005; Lei et al., 2018). Given

Figure 9: **Conformal prediction: coverage calibration**



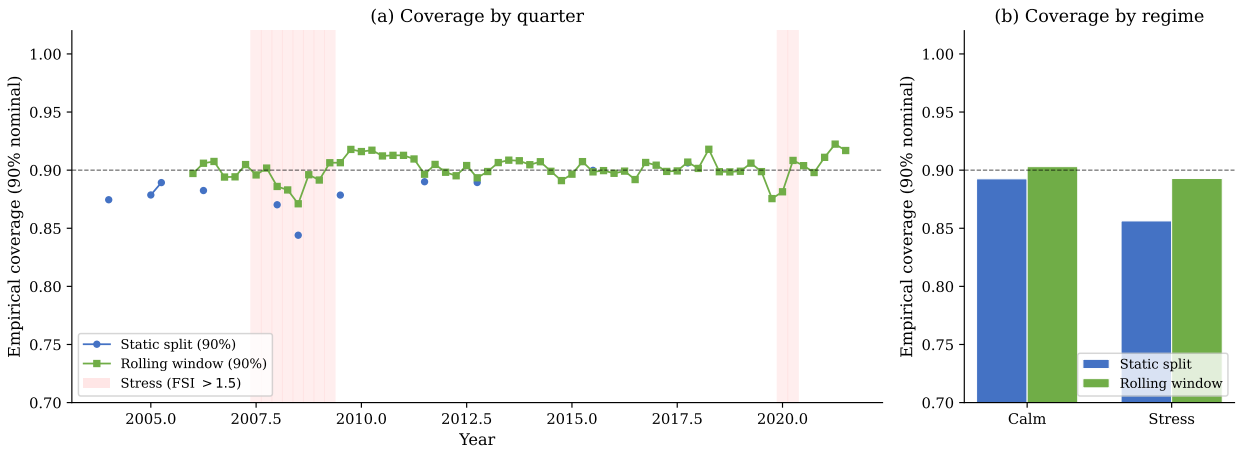
Notes: We show empirical coverage against nominal coverage for the prediction intervals obtained using three conformal methods: static split (calibrating the intervals once on 11 fixed validation quarters), locally adaptive (heteroskedastic intervals via residual normalization), and rolling window (recalibrating on the preceding 8 quarters before each evaluation quarter). The dashed line indicates perfect calibration.

Figure 10: **Adaptive interval width and per-decile coverage**



Notes: We consider the locally-adaptive conformal intervals at 90% nominal coverage. Panel A shows the mean interval width by decile of $|\widehat{y}|$. Panel B shows per-decile empirical coverage; bars below the dashed 90% line are highlighted.

Figure 11: **Coverage robustness: time series and stress regimes**



Notes: Panel A shows the empirical coverage for each quarter obtained using the static split and rolling-window conformal prediction intervals at 90% nominal coverage. The shaded regions indicate stress quarters (defined as FSI > 1.5). Panel B shows the empirical coverage by market regime (calm vs. stress) for both methods. The rolling window intervals achieve near-nominal coverage in both regimes.

a desired nominal coverage level $1 - \alpha$, a prediction interval is an interval such that the outcome falls within the interval with probability $1 - \alpha$: for example, outcomes should fall within a 90% prediction interval 90% of the time. Standard split conformal prediction guarantees finite-sample marginal coverage under the assumption that the calibration and test observations are exchangeable (Angelopoulos and Bates, 2023). Since our panel data are subject to both cross-sectional dependencies and distributional shifts over time, the formal exchangeability assumption may not hold exactly. Because of this, we treat this conformal prediction exercise simply as a distribution-free calibration device, and we evaluate coverage validity empirically.[20]

We partition the data into training, calibration, and held-out test sets. For each calibration pair $(i, a) \in \mathcal{C}_{\text{cal}}$, we compute the absolute residual $s_{ia} = |y_{ia,t+1} - \widehat{y}_{ia,t+1}|$ and construct the split conformal interval $C_{1-\alpha} = [\widehat{y}_{ia,t+1} \pm \widehat{q}_{1-\alpha}]$, where $\widehat{q}_{1-\alpha}$ is the appropriate quantile of the calibration scores.[21] We report three variants: a static split calibrating once on a fixed set of 11 validation quarters; a locally adaptive variant that normalizes residuals by the local median absolute deviation within deciles of $|\widehat{y}_{ia,t+1}|$ to produce heteroskedastic intervals (Lei et al., 2018); and a rolling window variant that recalibrates on the preceding 8 quarters before each evaluation quarter.

Figure 9 reports empirical coverage for the estimated prediction intervals against nominal coverage. At the 95% level, all three methods track the nominal level closely. Some differences emerge at 90%: the static split exhibits mild under-coverage, consistent with distributional shift, while rolling-window recalibration tracks the nominal level more accurately. Figure 11 examines coverage

[20] Extensions to covariate-shift conformal methods (Tibshirani et al., 2019) and conformal procedures for dependent data (Barber et al., 2023; Chernozhukov et al., 2018) are a natural direction for future work.

[21] Specifically, $\widehat{q}_{1-\alpha}$ is the $\lceil (n+1)(1-\alpha) \rceil / n$ quantile of $\{s_{ia}\}$ and $n = |\mathcal{C}_{\text{cal}}|$.

stability across time and regimes at 90% nominal. The rolling window achieves coverage ranging from 88.6% to 91.6% across all evaluated quarters. The static split shows a calm-stress gap (88.9% vs. 86.9%), which the rolling window largely closes (90.3% vs. 89.9%), showing near-uniformity across regimes.[22]

4 A Theory of Optimal Policy with Predictive Models

The empirical results in the preceding sections demonstrate that suitably designed deep learning models can accurately forecast institutional trading and recover economically meaningful measures of systemic risk. These objects are of independent interest for financial stability monitoring. We now develop a tractable equilibrium model of fire sales and optimal macroprudential intervention that provides a framework for interpreting the policy and welfare value of these empirical forecasts. We characterize conditions such that even in an environment subject to the usual Lucas critique, high-dimensional predictive models have value to the regulator by allowing for better cross-sectional policy targeting.

The key insight is that reduced-form predictive models can be useful for informing policy even though they do not by themselves identify policy counterfactuals. Instead, the theory clarifies a division of labor between reduced-form prediction and structural models, and it provides a template for optimally combining these two approaches in the same optimal policy rule. We show that under certain conditions, the optimal policy rule takes a sufficient-statistics form that combines a matrix of policy transmission elasticities (which we take as given, and could for example be obtained by calibrating a micro-founded structural model) with forecasts of where fire sales will occur in a crisis under the baseline of no policy intervention. The latter objects tell the regulator how to target the policy in the cross-section.

4.1 Setup

Time is discrete, $t = 1, 2$. There are N assets. There are I intermediary types, each of equal measure, with a representative intermediary of each type i. There is also a representative arbitrageur. For simplicity in the main text, intermediaries start date $t = 1$ with an inherited portfolio of assets. In Appendix C, we introduce a date 0 in which intermediaries optimally choose asset allocations, and study implications for optimal regulation.

At date 1, intermediaries face binding constraints that force them to sell assets prior to maturity to arbitrageurs, who are second best users. At date 2, payoffs are distributed and consumption

[22]Importantly, we note that these are marginal coverage rates pooled across all investor-asset pairs (i, a) within each quarter. Conditional coverage for specific investor types, asset classes, or individual entities may of course differ. Figure 10 provides partial evidence: locally-adaptive intervals at 90% nominal show coverage by decile of $|\widehat{y}_{ia,t+1}|$ that is approximately uniform, though the lowest deciles fall to roughly 83%, indicating that positions with the smallest predicted trades have somewhat less reliable intervals.

occurs. We assume that at date 1, intermediaries and arbitrageurs know the true values of all underlying parameters. The regulator's information structure will be introduced below.

Intermediaries. Intermediaries are risk neutral. Intermediaries enter date 1 with a vector $q_i = (q_{i1}, \ldots, q_{iN})^T$ of asset positions. If $q_{in} < 0$, then intermediary i has undertaken a negative investment (shorting) asset n. If held to maturity, intermediary i's holdings of asset n will produce a (stochastic) per-unit return R_{in} at date 2. We denote $R_i = (R_{i1}, \ldots, R_{iN})^T$ to be the vector of asset returns.

At date 1, intermediary i can sell assets prior to maturity. We denote $\ell_{in} \geq 0$ to be sales by intermediary i at endogenous price γ_n, with $\ell_i = (\ell_{i1}, \ldots, \ell_{iN})^T$ and $\gamma = (\gamma_1, \ldots, \gamma_N)^T$. Intermediary i faces an adjustment cost $\frac{1}{2}\ell_i^T H_i^\ell \ell_i$ on asset sales, where H_i^ℓ is $N \times N$. Assets will be sold at a discount on their fundamental value ($\gamma \leq R_i$, see below), and intermediary i faces a set of "rollover constraints" that force asset sales. This set of M constraints is given by

$$A_i^q q_i + \rho_i \leq A_i^\ell \ell_i, \tag{13}$$

where A_i^q, A_i^ℓ are $M \times N$ and ρ_i is $M \times 1$. For example, equation 13 can capture constraints on asset positions (e.g., a limit on debt) or a requirement to raise funds based on asset holdings (e.g., a cost of maintaining the project).[23]

At date 2, intermediary i realizes payoff on assets held to maturity and consumes. Intermediary i's total payoff (consumption) at date 2, inclusive of adjustment costs, is

$$U_i = q_i^T R_i - \ell_i^T (R_i - \gamma) - \frac{1}{2}\ell_i^T H_i^\ell \ell_i. \tag{14}$$

Arbitrageurs. A representative arbitrageur is a second-best user of intermediary assets. If the arbitrageur purchases a vector $L = (L_1, \ldots, L_N)^T$ of intermediary assets at date 1, the arbitrageur can use them in a production technology to produce $\mathcal{F}(L) = L^T\overline{\gamma} - \frac{1}{2}L^T\Gamma L$ units of the consumption good at date 2, where $\overline{\gamma}$ is $N \times 1$ and Γ is $N \times N$. The representative arbitrageur's payoff is

$$U_A = L^T(\overline{\gamma} - \gamma) - \frac{1}{2}L^T\Gamma L. \tag{15}$$

Market Clearing. The market for liquidations of each asset must clear at date 1, that is,

$$L = \sum_i \ell_i, \tag{16}$$

which specifies N market clearing conditions (one for each asset n).

[23] We simplify analysis by not having equation 13 depend on the liquidation price γ, as for example in price-dependent collateral constraints. This enables us to maintain a linear-quadratic structure throughout the paper. It is straightforward to extend analysis to include prices in constraints, but the characterization of the ex-ante optimal portfolio would no longer admit a closed-form solution.

Competitive Equilibrium. We define a competitive equilibrium as follows.

Definition 1. *A competitive equilibrium of the model is a vector of prices γ and a set of allocations $\{\ell_i, L\}$ such that:*

1. *Intermediary i chooses ℓ_i to maximize utility (equation 14) subject to the rollover constraint (13), taking prices γ as given.*
2. *Arbitrageurs choose L to maximize utility (equation 15), taking prices γ as given.*
3. *The liquidation markets clear (equation 16).*

4.2 Regulator's Model Design and Intervention

At date 1, the regulator can intervene in order to try to manage the fire sale price impact of liquidations. Formally, we assume that the regulator can impose a vector $\tau_i = (\tau_{i1}, \ldots, \tau_{iN})^T$ of revenue-neutral liquidation wedges on each intermediary i, which alter the intermediary's perceived price for selling the asset.[24] The wedge τ_{in} is formally a revenue-neutral tax on intermediary i from selling asset n, with $\tau_{in} < 0$ being a subsidy for sale. Due to its flexibility, the wedge approach is commonly employed in the macroprudential policy literature (Farhi and Werning 2016). Wedges can be interpreted as taxes and also as quantity restrictions (e.g., Clayton and Schaab 2022). As a result, intermediary i's payoff at date 2 is modified to be

$$U_i - (\ell_i^T - \ell_i^{*T})\tau_i \tag{17}$$

where $\ell_i^{*T}\tau_i$ is equilibrium revenue remissions to intermediary i based on the equilibrium asset liquidations ℓ_i^* of intermediaries of type i. Intermediaries take revenue remissions as given. We allow for the possibility that the regulator has incomplete instruments, represented by a regulatory cost $\delta(\tau) = \frac{1}{2}\tau^T\Delta\tau$, where Δ is $NI \times NI$.[25] To simplify the exposition, we introduce an assumption of matrix symmetry that we maintain throughout the paper.

Assumption 1. *The matrices H_i^ℓ, Γ, Δ are symmetric.*

Information Structure and Timing. In contrast to private agents, the regulator does not know the true values of the parameters $\Phi = \{A_i^q, A_i^\ell, H_i^\ell, \rho_i, R_i, \overline{\gamma}\}$. That is, the regulator is uncertain as to the true parameters underlying the rollover constraint (equation 13), the asset return R_i, the

[24] In Appendix C.4, we instead assume the regulator must use asset holding subsidies rather than liquidation taxes. The results at date 1 are identical, but the asset holding subsidies can give rise to additional moral hazard at date 0 when asset allocations q are chosen.

[25] One could instead model incompleteness as a restriction $\Delta(\tau) \leq 0$, which will imply a shadow cost on the regulatory incompleteness constraint measured by the Lagrange multiplier λ. This shadow cost is akin to the reduced-form cost function $\delta(\tau) = \lambda\Delta(\tau)$ except that λ is endogenous.

arbitrageurs' baseline productivity $\overline{\gamma}$, and the liquidation adjustment cost H_i^ℓ. We denote the regulator's prior to be $\Phi \sim \mu_0$.[26]

Before setting wedges, the regulator can use a *model* to make an inference about the true parameters. Formally, a model $M \in \mathcal{M}$ is a process of drawing a signal s of the parameters Φ. When the regulator draws a signal s from running model M, the regulator's posterior distribution is given by $\Phi \sim \mu_p|s, M$. After running the model and drawing the signal, the regulator chooses its wedges τ. After the regulator chooses its wedges, private agents move and the equilibrium is determined. Because the regulator sets wedges before the equilibrium is determined, the regulator cannot obtain any new information from observing the market before setting regulation.

4.2.1 Equilibrium Outcome for Given Wedges

We solve the regulator's optimum by first characterizing the equilibrium outcome, given the regulator's wedges $\tau = (\tau_1^T, \ldots, \tau_I^T)^T$ on the equilibrium at date 1. Our model is substantially simplified by the information structure: because private agents directly observe Φ, private agents do not learn from the regulator's choice of model or the signal drawn. As a result, the realized equilibrium outcome depends on the wedges τ and the true model parameters Φ, but not (for a given vector of asset allocations $q = (q_1^T, \ldots, q_I^T)^T$) on the regulator's choice of model M or the realized signal s.

The linear-quadratic structure of preferences and the rollover constraint leads to a characterization of the equilibrium as a simple linear system of equations. The following Lemma characterizes this equilibrium.

Lemma 1. *Given regulatory wedges τ, the equilibrium at date 1 is given by*

$$\ell_i = \overline{\ell}_i + \Lambda_i^q q_i - \Lambda_i^\tau \tau_i - \sum_{j=1}^{N} \left[\Lambda_j^{q,e} q_j^* - \Lambda_j^{\tau,e} \tau_j^* \right] \tag{18}$$

$$\gamma = \overline{\gamma} - \Gamma \sum_i \ell_i \tag{19}$$

where $\overline{\ell}_i$ $(N \times 1)$ and the loadings $\Lambda_i^q, \Lambda_i^\tau, \Lambda_i^{q,e}, \Lambda_i^{\tau,e}$ $(N \times N)$ are defined in the proof.

The characterization of the equilibrium in Lemma 1 is intuitive. Liquidations start from a baseline level $\overline{\ell}_i$ and increase (for positive loadings Λ_i^q) in asset holdings while decreasing (for positive loadings Λ_i^τ) in the liquidation wedge. Higher liquidations result in a lower liquidation price (for positive Γ) due to more assets having to be absorbed by arbitrageurs (equation 19). This means

[26] Assuming that private agents are perfectly informed is a simplifying assumption. Assuming that Γ is known to the regulator simplifies the analysis by eliminating the ability of a regulator to learn about the price impact of liquidations from observing a signal of γ or ℓ. This will allow us to fully separate predictive and causal channels in our framework. Absent this assumption, predictive models that inform a regulator about γ and ℓ might have even more value because they would also allow the regulator to learn about these structural parameters that determine the causal impact of a policy intervention (even if very little information is learned).

that liquidations by intermediary i decrease in asset holdings and increase in the liquidation wedges applied to other intermediaries within the same sector and across different sectors (for positive matrix loadings), a result of a substitution effect resulting from the increase in equilibrium price. In a model without fire sales, the loadings satisfy $\Lambda_j^{q,e} = \Lambda_j^{\tau,e} = 0$. We distinguish in equation 18 between the asset choices of intermediary i and the wedges applied to intermediary i's liquidations (denoted with no $*$) as opposed to equilibrium objects (denoted with $*$) that enter because they determine the equilibrium liquidation price.

Crucially, given the information structure that private agents have a finer information partition than the regulator, the regulator's model choice can alter the endogenous portfolios q_i that feed into the determination of the equilibrium liquidations and liquidation prices through choice of the regulatory wedges (see Appendix C), but not by changing the structural loadings (or agents' beliefs about them) $\Lambda_i = \{\Lambda_i^q, \Lambda_i^\tau, \Lambda_i^{q,e}, \Lambda_i^{\tau,e}\}$ that translate portfolio quantities and wedges into outcomes. In this sense, the equilibrium determination and the regulator's forecasting problem satisfy a key form of policy invariance.

The assumption that private agents observe the full state at date 1 can be weakened substantially. To obtain our policy invariance result, it is sufficient that any regulator-generated public information be informative only about the predictive component of the state (i.e., the baseline pattern of liquidation pressure) and not about the transmission component which determines how portfolios and policy wedges map into equilibrium liquidations and prices. Under Bayesian updating, private agents may learn from the regulator about the former component of the state and in that case, the policy invariance result continues to hold. Appendix C.5 formalizes this extension.

4.3 Regulator's Optimal Wedges

We solve the regulator's problem by backward induction: first, we characterize the regulator's optimal wedges τ, given a choice of model M and a realized signal s. We then characterize the regulator's welfare given model choice M in terms of sufficient statistics.[27] To simplify analysis, we assume that the regulator places equal weight on all intermediaries but places a welfare weight of zero on arbitrageurs.[28]

Given that liquidation wedges are revenue-neutral, the regulator's optimal choice of τ solves

$$\max_{\tau} \mathbb{E}\left[\sum_i U_i \,\middle|\, s, M\right] - \delta(\tau),$$

subject to equilibrium determination (Lemma 1).

As preliminaries to the proposition below, we define $\overline{\ell}_i(q) = \overline{\ell}_i + \Lambda_i^q q_i - \sum_i \Lambda_i^{q,e} q_i$ to be the $N \times 1$ vector of liquidations by intermediary i of each asset if there is no regulatory intervention ($\tau = 0$).

[27] In Appendix C, we study the regulator's optimal choice of model.

[28] This focuses attention on a distributive externality from shifting wealth between arbitrageurs and intermediaries, who effectively have different marginal values of wealth at date 1 (Dávila and Korinek 2017).

We define $L(q) = \sum_i \bar{\ell}_i(q)$ to be the $N \times 1$ vector capturing total liquidations of each asset (across all intermediaries) if there is no intervention. The following proposition characterizes the regulator's optimal liquidation wedges in terms of these objects and model parameters.

Proposition 1. *Given a model M and signal s, the regulator's optimal wedges are*

$$\tau^* = \mathbb{E}\left[\Xi \middle| s, M\right]^{-1} \mathbb{E}\left[(\sum_i \overline{\Lambda}_i^\tau)^T \Gamma L(q) \middle| s, M\right] \tag{20}$$

where $\Xi = \overline{\Lambda}^{\tau T} + (\sum_i \overline{\Lambda}_i^\tau)^T \Gamma (\sum_i \overline{\Lambda}_i^\tau) + \Delta$, where $\overline{\Lambda}_i^\tau = (e_i \otimes I_N)^T \Lambda_i^\tau - (\Lambda_1^{\tau,e}, \ldots, \Lambda_I^{\tau,e})$ is $N \times NI$ (where e_i is the standard basis vector whose i^{th} element is 1 and $\otimes$ is the Kronecker product), and where $\overline{\Lambda}^\tau = (\overline{\Lambda}_1^{\tau T}, \ldots, \overline{\Lambda}_I^{\tau T})^T$ is $NI \times NI$.

The optimal wedges of Proposition 1 are familiar from the macroprudential policy literature on fire sales, and intuitively encode an expected marginal cost-marginal benefit trade-off. The marginal cost of the policy is captured in the (conditional expectation of the) inverse matrix Ξ, while the marginal benefit is captured in the expectation.

There are three components to the private and regulatory marginal cost of liquidations, reflecting the distortion of intermediaries' activities away from their private optimum. First, increasing liquidation wedges directly distorts the intermediaries' activities (the first term of Ξ). Second, by using wedges to alter equilibrium prices, the regulator changes the incentives for intermediaries to sell different assets (the second term). Finally, there is the regulatory cost Δ.

The social marginal benefit of regulation arises from the mitigation of the fire sale. This has two components that are central to our analysis. First is the causal effect of the policy intervention on equilibrium liquidation prices, captured by the term $(\sum_i \overline{\Lambda}_i^\tau)^T \Gamma$. The utility consequence of these price changes is proportional to the total value of assets being sold by intermediaries, $L(q)$.

Importantly, this means the social benefit of regulation derives from a product of the causal impact of the policy intervention, $(\sum_i \overline{\Lambda}_i^\tau)^T \Gamma^T$, and the social value of changing the price, here given by $L(q)$. This means that in designing regulation, there is value not only to identifying causal policy impacts, but also to predicting the magnitude of liquidations $L(q)$. To fully disentangle these two mechanisms, we can assume independence of the predictive and causal components of the system.[29]

Definition 2 (Predictive-Causal Independence)**.** *We have predictive-causal independence if $\{\overline{\Lambda}_i^\tau\}$ and $\{\bar{\ell}_i, \Lambda_i^q, \Lambda_i^{q,e}\}$ are independent of one another.*

Under predictive-causal independence, we obtain the following corollary to Proposition 1.

[29] Naturally absent this assumption, we could perform an analogous decomposition to Corollary 1, but also include the covariance between the two terms.

Corollary 1. *Under predictive-causal independence, the regulator's optimal wedges are*

$$\tau^* = \mathbb{E}\left[\Xi \middle| s, M\right]^{-1} \underbrace{\mathbb{E}\left[(\sum_i \overline{\Lambda}_i^\tau)^T \Gamma \middle| s, M\right]}_{\text{Causal: Policy Impact on Liquidation Price}} \overbrace{\mathbb{E}\left[L(q) \middle| s, M\right]}^{\text{Predictive: Total Liquidations}} \tag{21}$$

An important implication of Proposition 1 and Corollary 1 is that a model that is purely predictive—that is, a model which can predict liquidations $L(q)$ but is uninformative about the loadings $\{\overline{\Lambda}_i^\tau\}$ that govern the policy counterfactual—can be valuable to a regulator, provided that the regulator has some prior or posterior knowledge over the causal impact of the policy. This causal component of the optimal policy rule could be obtained through a fully-specified equilibrium structural model (e.g., as reviewed in Christiano et al. 2005) or, under certain conditions, from causal effects that are estimated empirically in reduced form (Guren et al. 2021; McKay and Wolf 2023; Caravello et al. 2025). The predictive information informs the regulator as to the magnitude and margins for intervention by informing the regulator about the social benefit of the intervention. Corollary 1 implies that, holding fixed the causal impact of the policy, the regulator would design larger-magnitude interventions when the regulator's model predicted that total liquidations would be larger.

Formally, we will define a predictive model to be a model that delivers information only on the predictive component of Corollary 1.

Definition 3. *We say a model M is* predictive *if, under predictive-causal independence, the posterior joint distribution of $\{\overline{\Lambda}_i^\tau\}$ is the same as the prior for all signals s.*

4.4 Predictive Accuracy and Welfare

We next ask what the welfare benefit to a regulator at date 1 is based on the model M the regulator runs.[30] Intuitively, for a predictive model, Proposition 1 suggests that the welfare gains obtained from using the model should be related to the predictive accuracy of that model. In this section, we formalize this intuition, showing that the predictive correlation of a predictive model's forecasts—weighted appropriately according to the causal leverage of the policy—is a sufficient statistic for the welfare gains from the model.

As a preliminary to the proposition below, we define a few objects. Given a square matrix W, we define variance and covariance operators that use W as a weighting matrix:

$$\text{Cov}_W(X, Y) = \mathbb{E}\left[(X - E[X])^T W (Y - E[Y])\right], \quad \text{Var}_W(X) = \text{Cov}_W(X, X).$$

The maximum ex-ante expected welfare gain (relative to relying on the prior) within the class of

[30] In Appendix C, we consider how the regulator's model choice affects intermediaries' optimal portfolio allocations when they are originally chosen.

predictive models, $\Delta V_{\max}$, is attained by a perfect predictive signal—that is, a predictive model M whose forecasts $\widehat{L}(q) = \mathbb{E}\left[L(q) \mid s, M\right]$ satisfy $\widehat{L}(q) = L(q)$. The following proposition characterizes welfare losses relative to the perfect forecast benchmark $\Delta V_{\max}$ in terms of the model's predictive accuracy.

Proposition 2. *Under predictive-causal independence, the regulator's expected welfare gain from employing a predictive model M, relative to relying on the prior, is*

$$\Delta V(M) = \Delta V_{\max} \cdot \rho^2_{\Theta_0}(M), \tag{22}$$

where

$$\rho_{\Theta_0}(M) \equiv \frac{\mathrm{Cov}_{\Theta_0}(L, \widehat{L})}{\sqrt{\mathrm{Var}_{\Theta_0}(L)\,\mathrm{Var}_{\Theta_0}(\widehat{L})}},$$

where the weighting matrix Θ_0 is given $\Theta_0 = \mathbb{E}_0[\Upsilon^T \Psi \Upsilon | M]$, where $\Psi = \mathbb{E}_0[2(\sum_i \overline{\Lambda}_i^\tau)^T \Gamma (\sum_i \overline{\Lambda}_i^\tau) + \Delta + \overline{\Lambda}^{\tau T} H^\ell \overline{\Lambda}^\tau]$, and where $\Upsilon = (\mathbb{E}_0[\Xi])^{-1} \mathbb{E}_0[(\sum_i \overline{\Lambda}_i^\tau)^T \Gamma]$.

Proposition 2 shows that the welfare gains from a predictive model scale linearly with the Θ_0-weighted R-squared achieved by the model in predicting liquidations. The weighting matrix Θ_0 has a natural interpretation as the metric that converts variation in predicted liquidations $\hat{L}(q)$ into variation in welfare after the regulator applies the optimal wedges τ^*: therefore, assets that generate larger externalities or for which the intervention has a higher causal leverage on outcomes receive a larger weight in the statistic $\rho_{\Theta_0}(M)$. The predictive R-squared $\rho^2_{\Theta_0}(M)$ has in fact an exact interpretation as the fraction of the maximum achievable welfare gain within the class of predictive models that is actually attained. The preceding empirical sections showed that high-dimensional predictive models can deliver accurate cross-sectional forecasts of trading behavior, and this result provides a way to bridge the empirical performance of welfare.

Proposition 2 completes the characterization of the regulator's ex-post problem. In Appendix C.2, we also study the equilibrium feedback of model choice and ex-post intervention on intermediaries' ex-ante portfolio decisions.

5 Conclusion

This paper asks whether the combination of granular data on portfolio holdings and modern deep learning methods can improve financial stability monitoring. We show that the network structure of portfolio holdings contains more information relevant for forecasting the dynamics of stress episodes than conventional low-dimensional summaries capture. We introduce a deep learning architecture that we show recovers this information in an economically interpretable way. The model learns latent representations of investors and assets that deliver strong out-of-sample forecasts of trading behavior including during crisis episodes, that explain the cross-section of asset price dislocations

during stress events with more than ten times the incremental explanatory power beyond standard asset characteristics of existing approaches, and that outperform conventional systemic risk metrics in identifying risky institutions. Our complementary theoretical framework clarifies when and how these predictive objects have policy and welfare value: we show that in certain models of macroprudential intervention, they serve as sufficient statistics that govern the cross-sectional targeting of policy interventions. More broadly, the paper suggests an approach for combining prediction and structure in policy problems with rich cross-sectional heterogeneity. Predictive models are not substitutes for structural analysis or causal inference, but they can be powerful complements, helping to capture heterogeneity in the data better and to enable more precise policy targeting. As AI methods continue to advance and regulatory data grows richer, this complementarity is likely to become increasingly important for policy design.

References

Acemoglu, Daron, Asuman Ozdaglar, and Alireza Tahbaz-Salehi, "Systemic risk and stability in financial networks," *American Economic Review*, 2015, *105* (2), 564–608.

Acharya, Viral V, Lasse H Pedersen, Thomas Philippon, and Matthew Richardson, "Measuring systemic risk," *The review of financial studies*, 2017, *30* (1), 2–47.

Acharya, Viral V, Markus K Brunnermeier, and Diane Pierret, "Systemic risk measures: From the panic of 1907 to the banking stress of 2023," *Annual Review of Financial Economics*, 2025.

Adrian, Tobias and Hyun Song Shin, "Procyclical leverage and value-at-risk," *The Review of Financial Studies*, 2014, *27* (2), 373–403.

Adrian, Tobias and Markus K Brunnermeier, "CoVaR," *The American Economic Review*, 2016, *106* (7), 1705.

Angelopoulos, Anastasios N and Stephen Bates, "Conformal Prediction: A Gentle Introduction," *Foundations and Trends in Machine Learning*, 2023, *16* (4), 494–591.

Anton, Miguel and Christopher Polk, "Connected stocks," *The Journal of Finance*, 2014, *69* (3), 1099–1127.

Aramonte, Sirio, Andreas Schrimpf, and Hyun Song Shin, "Non-bank financial intermediaries and financial stability," *BIS Working Paper*, 2023.

Athey, Susan, "Beyond prediction: Using big data for policy problems," *Science*, 2017, *355* (6324), 483–485.

Athey, Susan, "The impact of machine learning on economics," in "The economics of artificial intelligence: An agenda," University of Chicago Press, 2018, pp. 507–547.

Athey, Susan and Guido W Imbens, "Machine learning methods that economists should know about," *Annual Review of Economics*, 2019, *11* (1), 685–725.

Athey, Susan and Stefan Wager, "Policy learning with observational data," *Econometrica*, 2021, *89* (1), 133–161.

Barber, Rina Foygel, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani, "Conformal Prediction Beyond Exchangeability," *The Annals of Statistics*, 2023, *51* (2), 816–845.

Barnichon, Régis and Geert Mesters, "A sufficient statistics approach for macro policy," *American Economic Review*, 2023, *113* (11), 2809–2845.

Barro, Robert J and David B Gordon, "A positive theory of monetary policy in a natural rate model," *Journal of political economy*, 1983, *91* (4), 589–610.

Benoit, Sylvain, Jean-Edouard Colliard, Christophe Hurlin, and Christophe Pérignon, "Where the Risks Lie: A Survey on Systemic Risk," *Review of Finance*, 2017, *21* (1), 109–152.

Bernanke, Ben S and Mark Gertler, "Agency costs, collateral, and business fluctuations," 1986.

Bianchi, Javier, "Overborrowing and Systemic Externalities in the Business Cycle," *American Economic Review*, December 2011, *101* (7), 3400–3426.

Bianchi, Javier, "Efficient Bailouts?," *American Economic Review*, December 2016, *106* (12), 3607–59.

Bianchi, Javier and Enrique G. Mendoza, "Optimal Time-Consistent Macroprudential Policy," *Journal of Political Economy*, 2018, *126* (2), 588–634.

Brainard, William C and James Tobin, "Pitfalls in financial model building," *The American economic review*, 1968, *58* (2), 99–122.

Brown, Gregory W, Philip Howard, and Christian T Lundblad, "Crowded trades and tail risk," *The Review of Financial Studies*, 2022, *35* (7), 3231–3271.

Brownlees, Christian and Robert F Engle, "SRISK: A Conditional Capital Shortfall Measure of Systemic Risk," *The Review of Financial Studies*, 2017, *30* (1), 48–79.

Brunnermeier, Markus K and Martin Oehmke, "Bubbles, financial crises, and systemic risk," *Handbook of the Economics of Finance*, 2013, *2*, 1221–1288.

Bryzgalova, Svetlana, Victor DeMiguel, Sicong Li, and Markus Pelger, "Asset-pricing factors with economic targets," *Available at SSRN*, 2023, *4344837*.

Bybee, J Leland, "The ghost in the machine: Generating beliefs with large language models," *Working Paper*, 2025.

Caballero, Ricardo J. and Arvind Krishnamurthy, "International and domestic collateral constraints in a model of emerging market crises," *Journal of Monetary Economics*, December 2001, *48* (3), 513–548.

Caravello, Tomás, Alisdair McKay, and Christian Wolf, "Evaluating Monetary Policy Counterfactuals:(When) Do We Need Structural Models?," *Working Paper*, 2025.

Chari, V. V. and Patrick J. Kehoe, "Bailouts, Time Inconsistency, and Optimal Regulation: A Macroeconomic View," *American Economic Review*, September 2016, *106* (9), 2458–93.

Chen, Hui, Yuhan Cheng, Yanchu Liu, and Ke Tang, "Teaching economics to the machines," *Available at SSRN 4642167*, 2023.

Chen, Luyang, Markus Pelger, and Jason Zhu, "Deep learning in asset pricing," *Management Science*, 2024, *70* (2), 714–750.

Chen, Qi, Itay Goldstein, and Wei Jiang, "Payoff complementarities and financial fragility: Evidence from mutual fund outflows," *Journal of financial economics*, 2010, *97* (2), 239–262.

Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu, "Exact and Robust Conformal Inference Methods for Predictive Machine Learning with Dependent Data," *Proceedings of the 31st Conference on Learning Theory*, 2018, *75*, 732–749.

Chevalier, Judith and Glenn Ellison, "Risk taking by mutual funds as a response to incentives," *Journal of political economy*, 1997, *105* (6), 1167–1200.

Christiano, Lawrence J, Martin Eichenbaum, and Charles L Evans, "Nominal rigidities and the dynamic effects of a shock to monetary policy," *Journal of political Economy*, 2005, *113* (1), 1–45.

Clayton, Christopher and Andreas Schaab, "Multinational Banks and Financial Stability," *The Quarterly Journal of Economics*, 01 2022, *137* (3), 1681–1736.

Clayton, Christopher and Andreas Schaab, "Bail-Ins, Optimal Regulation, and Crisis Resolution," *The Review of Financial Studies*, 03 2025, p. hhaf002.

Cohen, Lauren, Yiwen Lu, and Quoc H Nguyen, "Mimicking Finance," *Working Paper*, 2026.

Coppola, Antonio, "In safe hands: The financial and real impact of investor composition over the credit cycle," *The Review of Financial Studies*, 2025, p. hhaf017.

Coval, Joshua and Erik Stafford, "Asset fire sales (and purchases) in equity markets," *Journal of Financial Economics*, 2007, *86* (2), 479–512.

Derrow-Pinion, Austin, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire et al., "Eta prediction with graph neural networks in google maps," in "Proceedings of the 30th ACM international conference on information & knowledge management" 2021, pp. 3767–3776.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in "Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)" 2019, pp. 4171–4186.

Dick-Nielsen, Jens, "How to clean enhanced TRACE data," *Available at SSRN 2337908*, 2014.

Dolphin, Rian, Barry Smyth, and Ruihai Dong, "Stock embeddings: Learning distributed representations for financial assets," *arXiv preprint arXiv:2202.08968*, 2022.

Dávila, Eduardo and Anton Korinek, "Pecuniary Externalities in Economies with Financial Frictions," *The Review of Economic Studies*, 02 2017, *85* (1), 352–395.

Einav, Liran and Jonathan Levin, "The data revolution and economic analysis," *Innovation Policy and the Economy*, 2014, *14* (1), 1–24.

Einav, Liran and Jonathan Levin, "Economics in the age of big data," *Science*, 2014, *346* (6210), 1243089.

Elliott, Matthew, Benjamin Golub, and Matthew O Jackson, "Financial networks and contagion," *American Economic Review*, 2014, *104* (10), 3115–3153.

Fang, Xiang, Bryan Hardy, and Karen K Lewis, "Who holds sovereign debt and why it matters," *The Review of Financial Studies*, 2025, p. hhaf031.

Farboodi, Maryam and Laura Veldkamp, "Long-run growth of financial data technology," *American Economic Review*, 2020, *110* (8), 2485–2523.

Farhi, Emmanuel and Iván Werning, "A theory of macroprudential policies in the presence of nominal rigidities," *Econometrica*, 2016, *84* (5), 1645–1704.

Faria-e Castro, Miguel, Joseba Martinez, and Thomas Philippon, "Runs versus lemons: Information disclosure and fiscal capacity," *The Review of Economic Studies*, 2017, *84* (4), 1683–1707.

Fisher, Irving, "The debt-deflation theory of great depressions," *Econometrica: Journal of the Econometric Society*, 1933, pp. 337–357.

Fontanier, Paul, "Optimal policy for behavioral financial crises," *Journal of Financial Economics*, 2025, *166*, 104005.

Friedman, Milton, "The methodology of positive economics," *Essays In Positive Economics*, 1953.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably unequal? The effects of machine learning on credit markets," *The Journal of Finance*, 2022, *77* (1), 5–47.

Gabaix, Xavier, Ralph SJ Koijen, Robert Richmond, and Motohiro Yogo, "Asset embeddings," *Available at SSRN 4507511*, 2024.

Gabaix, Xavier, Ralph SJ Koijen, Robert Richmond, and Motohiro Yogo, "Upgrading Credit Pricing and Risk Assessment through Embeddings," *Available at SSRN*, 2025.

Giglio, Stefano, Bryan Kelly, and Dacheng Xiu, "Factor models, machine learning, and asset pricing," *Annual Review of Financial Economics*, 2022, *14* (1), 337–368.

Gillis, Talia B and Jann L Spiess, "Big data and discrimination," *The University of Chicago Law Review*, 2019, *86* (2), 459–488.

Goldstein, Itay and Yaron Leitner, "Stress tests and information disclosure," *Journal of Economic Theory*, 2018, *177*, 34–69.

Goldstein, Itay, Hao Jiang, and David T Ng, "Investor flows and fragility in corporate bond funds," *Journal of Financial Economics*, 2017, *126* (3), 592–613.

Greenwood, Robin and David Thesmar, "Stock price fragility," *Journal of Financial Economics*, 2011, *102* (3), 471–490.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, 2020, *33* (5), 2223–2273.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, "Autoencoder asset pricing models," *Journal of Econometrics*, 2021, *222* (1), 429–450.

Guren, Adam, Alisdair McKay, Emi Nakamura, and Jón Steinsson, "What do we learn from cross-regional empirical estimates in macroeconomics?," *NBER Macroeconomics Annual*, 2021, *35* (1), 175–223.

Haddad, Valentin, Alan Moreira, and Tyler Muir, "When selling becomes viral: Disruptions in debt markets in the COVID-19 crisis and the Fed's response," *The Review of Financial Studies*, 2021, *34* (11), 5309–5351.

Hamilton, Will, Zhitao Ying, and Jure Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, 2017, *30*.

He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, "Masked Autoencoders Are Scalable Vision Learners," in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition" 2022, pp. 16000–16009.

Hendrycks, Dan and Kevin Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

Jiang, Dejun, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou, "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models," *Journal of cheminformatics*, 2021, *13*, 1–23.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko et al., "Highly accurate protein structure prediction with AlphaFold," *nature*, 2021, *596* (7873), 583–589.

Kindleberger, Charles and Robert Aliber, *Manias, panics and crashes: a history of financial crises*, Springer, 1978.

Kingma, Diederik P, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Kiyotaki, Nobuhiro and John Moore, "Credit cycles," *Journal of political economy*, 1997, *105* (2), 211–248.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The quarterly journal of economics*, 2018, *133* (1), 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer, "Prediction policy problems," *American Economic Review*, 2015, *105* (5), 491–495.

Koopmans, Tjalling C, "Measurement without theory," *The Review of Economics and Statistics*, 1947, *29* (3), 161–172.

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

Krishnamurthy, Arvind and Tyler Muir, "How credit cycles across a financial crisis," *The Journal of Finance*, 2025, *80* (3), 1339–1378.

Laffont, Jean-Jacques and Jean Tirole, "Using Cost Observation to Regulate Firms," *Journal of Political Economy*, 1986, *94* (3), 614–641.

Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman, "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 2018, *113* (523), 1094–1111.

Leitner, Yaron and Basil Williams, "Model secrecy and stress tests," *The Journal of Finance*, 2023, *78* (2), 1055–1095.

Lorenzoni, Guido, "Inefficient Credit Booms," *The Review of Economic Studies*, 07 2008, *75* (3), 809–833.

Loshchilov, Ilya and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2018.

Lucas, Robert E, "Econometric policy evaluation: A critique," in "Carnegie-Rochester conference series on public policy," Vol. 1 North-Holland 1976, pp. 19–46.

Ludwig, Jens and Sendhil Mullainathan, "Machine Learning as a Tool for Hypothesis Generation," *The Quarterly Journal of Economics*, 2024, *139* (2), 751–827.

Ma, Yiming, Kairong Xiao, and Yao Zeng, "Mutual fund liquidity transformation and reverse flight to liquidity," *The Review of Financial Studies*, 2022, *35* (10), 4674–4711.

McKay, Alisdair and Christian K Wolf, "What can time-series regressions tell us about policy

counterfactuals?," *Econometrica*, 2023, *91* (5), 1695–1725.

Mullainathan, Sendhil and Jann Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

Nagel, Stefan, *Machine learning in asset pricing*, Princeton University Press, 2021.

Nakamura, Emi and Jón Steinsson, "Identification in macroeconomics," *Journal of Economic Perspectives*, 2018, *32* (3), 59–86.

Orlov, Dmitry, Pavel Zryumov, and Andrzej Skrzypacz, "The design of macroprudential stress tests," *The Review of Financial Studies*, 2023, *36* (11), 4460–4501.

Parlatore, Cecilia and Thomas Philippon, "Designing stress scenarios," *The Journal of Finance*, 2025, *80* (2), 833–873.

Sarkar, Suproteem K, "Economic representations," 2025.

Scarselli, Franco, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, 2008, *20* (1), 61–80.

Schularick, Moritz and Alan M Taylor, "Credit booms gone bust: monetary policy, leverage cycles, and financial crises, 1870–2008," *American Economic Review*, 2012, *102* (2), 1029–1061.

Shapiro, Joel and David Skeie, "Information management in banking crises," *The Review of Financial Studies*, 2015, *28* (8), 2322–2363.

Stein, Jeremy C., "Monetary Policy as Financial Stability Regulation*," *The Quarterly Journal of Economics*, 01 2012, *127* (1), 57–95.

Tibshirani, Ryan J., Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas, "Conformal Prediction Under Covariate Shift," *Advances in Neural Information Processing Systems*, 2019, *32.*

Uhlig, Harald, "A model of a systemic bank run," *Journal of Monetary Economics*, 2010, *57* (1), 78–96.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017, *30.*

Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer, *Algorithmic Learning in a Random World*, Springer, 2005.

Williamson, Oliver E, "Corporate finance and corporate governance," *The journal of finance*, 1988, *43* (3), 567–591.

Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, 2020, *32* (1), 4–24.

Xu, Keyulu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka, "How powerful are graph neural networks?," *arXiv preprint arXiv:1810.00826*, 2018.

Online Appendix for
"The Optimal Use of AI in Financial Regulation"

Christopher Clayton Antonio Coppola

September 2026

A Proofs of Propositions in Main Text

A.1 Proof of Lemma 1

Intermediary i's Lagrangian over ℓ_i is

$$\mathcal{L}_i = q_i^T(R_i - p_i) - \ell_i^T(R_i + \tau_i - \gamma) - \frac{1}{2}q_i^T H_i^q q_i - \frac{1}{2}\ell_i^T H_i^\ell \ell_i - \lambda_i^T(A_i^\ell \ell_i - A_i^q q_i - \rho_i),$$

yielding FOCs

$$0 = -(R_i + \tau_i - \gamma) - H_i^\ell \ell_i - A_i^{\ell T}\lambda_i,$$

$$\ell_i = -H_i^{\ell-1}(R_i + \tau_i - \gamma) - H_i^{\ell-1}A_i^{\ell T}\lambda_i.$$

Assuming the rollover constraint binds,

$$A_i^\ell\Big[- H_i^{\ell-1}(R_i + \tau_i - \gamma) - H_i^{\ell-1}A_i^{\ell T}\lambda_i\Big] = A_i^q q_i + \rho_i,$$

$$-\lambda_i = (A_i^\ell H_i^{\ell-1}A_i^{\ell T})^{-1}A_i^\ell H_i^{\ell-1}(R_i + \tau_i - \gamma) + (A_i^\ell H_i^{\ell-1}A_i^{\ell T})^{-1}A_i^q q_i + (A_i^\ell H_i^{\ell-1}A_i^{\ell T})^{-1}\rho_i,$$

and so we have

$$\begin{aligned}\ell_i = {} & H_i^{\ell-1}A_i^{\ell T}(A_i^\ell H_i^{\ell-1}A_i^{\ell T})^{-1}\rho_i + H_i^{\ell-1}A_i^{\ell T}(A_i^\ell H_i^{\ell-1}A_i^{\ell T})^{-1}A_i^q q_i \\ & + \Big[H_i^{\ell-1}A_i^{\ell T}(A_i^\ell H_i^{\ell-1}A_i^{\ell T})^{-1}A_i^\ell H_i^{\ell-1} - H_i^{\ell-1}\Big](R_i + \tau_i - \gamma).\end{aligned}$$

We therefore write

$$\ell_i = \overline{\ell}_i^0 + \Lambda_i^q q_i - \Lambda_i^\tau(\tau_i - \gamma),$$

where

$$\overline{\ell}_i^0 = H_i^{\ell-1}A_i^{\ell T}(A_i^\ell H_i^{\ell-1}A_i^{\ell T})^{-1}\rho_i - \Lambda_i^\tau R_i,$$

$$\Lambda_i^q = H_i^{\ell-1}A_i^{\ell T}(A_i^\ell H_i^{\ell-1}A_i^{\ell T})^{-1}A_i^q,$$

$$\Lambda_i^\tau = H_i^{\ell-1} - H_i^{\ell-1}A_i^{\ell T}(A_i^\ell H_i^{\ell-1}A_i^{\ell T})^{-1}A_i^\ell H_i^{\ell-1}.$$

Arbitrageur maximization yields

$$\gamma = \overline{\gamma} - \Gamma L,$$

$$\gamma = \overline{\gamma} - \Gamma \sum_i \overline{\ell}_i^0 - \Gamma \sum_i \Lambda_i^q q_i + \Gamma \sum_i \Lambda_i^\tau (\tau_i - \gamma),$$

$$\gamma = \left(\mathbb{I} + \Gamma \sum_i \Lambda_i^\tau\right)^{-1} \left(\overline{\gamma} - \Gamma \sum_i \overline{\ell}_i^0\right) - \sum_i \left[\left(\mathbb{I} + \Gamma \sum_i \Lambda_i^\tau\right)^{-1} \Gamma \Lambda_i^q q_i - \left(\mathbb{I} + \Gamma \sum_i \Lambda_i^\tau\right)^{-1} \Gamma \Lambda_i^\tau \tau_i\right],$$

$$\gamma = \overline{\gamma}^e - \Gamma^e \sum_i \left[\Lambda_i^q q_i^* - \Lambda_i^\tau \tau_i^*\right],$$

where we have defined $\overline{\gamma}^e = \left(\mathbb{I} + \Gamma \sum_i \Lambda_i^\tau\right)^{-1} \left(\overline{\gamma} - \Gamma \sum_i \overline{\ell}_i^0\right)$ and $\Gamma^e = \left(\mathbb{I} + \Gamma \sum_i \Lambda_i^\tau\right)^{-1} \Gamma$. Finally substituting back in,

$$\ell_i = \overline{\ell}_i + \Lambda_i^q q_i - \Lambda_i^\tau \tau_i - \sum_i \left[\Lambda_i^{q,e} q_i^* - \Lambda_i^{\tau,e} \tau_i^*\right],$$

where $\overline{\ell}_i = \overline{\ell}_i^0 + \Lambda_i^\tau \overline{\gamma}^e$, $\Lambda_i^{q,e} = \Lambda_i^\tau \Gamma^e \Lambda_i^q$, and $\Lambda_i^{\tau,e} = \Lambda_i^\tau \Gamma^e \Lambda_i^\tau$.

A.2 Proof of Proposition 1

From Lemma 1, equilibrium liquidations given (q, τ) are

$$\ell_i = \overline{\ell}_i + \Lambda_i^q q_i - \Lambda_i^\tau \tau_i - \sum_i \left[\Lambda_i^{q,e} q_i - \Lambda_i^{\tau,e} \tau_i\right].$$

We define $\overline{\Lambda}_i^\tau = (e_i \otimes I_N)^T \Lambda_i^\tau - (\Lambda_1^{\tau,e}, \ldots, \Lambda_I^{\tau,e})$ (where e_i is the standard basis vector whose i^{th} element is 1 and $\otimes$ is the Kronecker product) and similarly define $\overline{\Lambda}_i^q = (e_i \otimes I_N)^T \Lambda_i^q q_i - (\Lambda_1^{q,e}, \ldots, \Lambda_I^{q,e})$. We then write

$$\ell_i = \overline{\ell}_i + \overline{\Lambda}_i^q q - \overline{\Lambda}_i^\tau \tau.$$

Stacking over i, we have

$$\ell = \overline{\ell} + \overline{\Lambda}^q q - \overline{\Lambda}^\tau \tau.$$

Next, we can write the regulator's objective as

$$\mathbb{E}\left[\sum U_i - \delta(\tau) \middle| s, M\right] = \mathbb{E}\left[\sum_i \mathcal{L}_i \middle| s, M\right] + \mathbb{E}\left[\tau^T \ell \middle| s, M\right] - \frac{1}{2} \tau^T \Delta \tau,$$

where as in the proof of Lemma 1 we have defined the intermediary i Lagrangian by

$$\mathcal{L}_i = q_i^T (R_i - p_i) - \ell_i^T (R_i + \tau_i - \gamma) - \frac{1}{2} q_i^T H_i^q q_i - \frac{1}{2} \ell_i^T H_i^\ell \ell_i - \lambda_i^T (A_i^\ell \ell_i - A_i^q q_i - \rho_i).$$

The regulator thus solves

$$\max_\tau \mathbb{E}\left[\sum_i \mathcal{L}_i \middle| s, M\right] + \mathbb{E}\left[\tau^T \ell \middle| s, M\right] - \frac{1}{2} \tau^T \Delta \tau$$

subject to

$$\ell = \overline{\ell} + \overline{\Lambda}^q q - \overline{\Lambda}^\tau \tau,$$

$$\gamma = \overline{\gamma} - \Gamma L.$$

By the Envelope Theorem, we have the regulator's FOC given by

$$0 = \mathbb{E}\Big[-\ell + \sum_i \ell_i \frac{\partial \gamma}{\partial \tau}\Big| s, M\Big] + \mathbb{E}\Big[\frac{\partial \tau^T \ell}{\partial \tau}\Big| s, M\Big] - \Delta\tau,$$

$$0 = \mathbb{E}\Big[\sum_i (\overline{\Lambda}_i^\tau)^T \Gamma^T L \Big| s, M\Big] + \mathbb{E}\Big[-\overline{\Lambda}^{\tau T}\tau \Big| s, M\Big] - \Delta\tau.$$

Writing $L = L(q) - \sum_i (\overline{\Lambda}_i^\tau)\tau$ and using the assumption that Γ is symmetric,

$$\mathbb{E}[\overline{\Lambda}^{\tau T}|s, M] + \Delta\tau = \mathbb{E}\Big[\sum_i (\overline{\Lambda}_i^\tau)^T \Gamma \Big(L(q) - \sum_i (\overline{\Lambda}_i^\tau)\tau \Big) \Big| s, M\Big],$$

$$\tau = \mathbb{E}\Big[\Xi \Big| s, M\Big]^{-1} \mathbb{E}\Big[(\sum_i \overline{\Lambda}_i^\tau)^T \Gamma L(q) \Big| s, M\Big],$$

where $\Xi = \overline{\Lambda}^{\tau T} + (\sum_i \overline{\Lambda}_i^\tau)^T \Gamma^T (\sum_i \overline{\Lambda}_i^\tau) + \Delta$.

A.3 Proof of Proposition 2

We begin with an intermediate Lemma characterizing the regulator's welfare function. As a preliminary to the Lemma, we define $\theta_i(q) = R_i - (\overline{\gamma} - \Gamma L(q))$ to be the fire sale losses to intermediary i from selling assets prior to maturity in the event of no regulatory intervention. We define $\hat{L}(q) = \mathbb{E}[L(q)|s, M]$ to be predicted total liquidations, given a model M and signal s.

Lemma 2. *Under predictive-causal independence, the regulator's ex-ante expected welfare given positions q and a model M is*

$$\begin{aligned} V(q, M) =& \mathbb{E}_0\Big[q^T(R - p) - \ell(q)^T \theta(q) - \frac{1}{2} q^T H^q q - \frac{1}{2}\ell(q)^T H^\ell \ell(q)\Big] \\ &+ \frac{1}{2} L_0(q)^T \Theta_0 L_0(q) + \frac{1}{2}\mathrm{tr}\Big(\Theta_0\, cov_0(\hat{L}(q))\Big), \end{aligned} \tag{A.1}$$

where $\Theta_0 = \mathbb{E}_0[\Upsilon^T \Psi \Upsilon | M]$, *where* $\Psi = \mathbb{E}\Big[2(\sum_i \overline{\Lambda}_i^\tau)^T \Gamma (\sum_i \overline{\Lambda}_i^\tau) + \Delta + \overline{\Lambda}^{\tau T} H^\ell \overline{\Lambda}^\tau \Big| s, M\Big]$, *and where* $\Upsilon = \Big(\mathbb{E}\Big[\Xi \Big| s, M\Big]\Big)^{-1} \mathbb{E}\Big[(\sum_i \overline{\Lambda}_i^\tau)^T \Gamma \Big| s, M\Big]$.

Using Lemma 2, the welfare gain of a predictive model relative to using just the prior is[1]

$$\Delta V = \frac{1}{2}\operatorname{tr}(\Theta_0 \operatorname{cov}(\widehat{L})).$$

The maximum welfare gain within the class of predictive models is attained by a perfect predictive signal ($\widehat{L} = L$), so:

$$\Delta V_{\max} = \frac{1}{2}\operatorname{tr}(\Theta_0 \operatorname{cov}(L)).$$

Given a weighting matrix W, the W-weighted correlation between the predictions $\widehat{L}$ and the true liquidations L is:

$$\rho_W \equiv \frac{\operatorname{Cov}_W(L, \widehat{L})}{\sqrt{\operatorname{Var}_S(L)\operatorname{Var}_W(\widehat{L})}}.$$

For W deterministic and since $\mathbb{E}L = \mathbb{E}\widehat{L}$ by the law of iterated expectations, we have that:

$$\operatorname{Cov}_W(L, \widehat{L}) := \mathbb{E}\left[(L - \mathbb{E}L)^T W(\widehat{L} - \mathbb{E}L)\right] = \mathbb{E}\left[(\mathbb{E}[L \mid s] - \mathbb{E}L)^T W(\mathbb{E}[L \mid s] - \mathbb{E}L)\right] = \operatorname{Var}_W(\widehat{L}).$$

Therefore:

$$\rho_W^2 = \frac{\operatorname{Var}_W(\widehat{L})^2}{\operatorname{Var}_W(\widehat{L})\operatorname{Var}_W(L)} = \frac{\operatorname{Var}_W(\widehat{L})}{\operatorname{Var}_W(L)}.$$

Using the fact that $\operatorname{Var}_W(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^T W(X - \mathbb{E}[X])\right] = \operatorname{tr}(W \operatorname{cov}(X))$, we then have that:

$$\rho_W^2 = \frac{\operatorname{tr}(W \operatorname{cov}(\widehat{L}))}{\operatorname{tr}(W \operatorname{cov}(L))}.$$

Therefore, $\Delta V = \Delta V_{\max} \cdot \rho^2_{\Theta_0}$.

A.3.1 Proof of Lemma 2

As in the proof of Proposition 1, the regulator's ex-ante expected welfare given positions q, a model M, and a signal s is

$$V(q, M, s) = \mathbb{E}\left[q^T(R - p) - \frac{1}{2}q^T H^q q \middle| s, M\right] + \mathbb{E}\left[\sum_i \left[-\ell_i^T(R_i - \gamma) - \frac{1}{2}\ell_i^T H_i^\ell \ell_i\right] - \frac{1}{2}\tau^T \Delta\tau \middle| s, M\right],$$

where we define H^q to be the block diagonal matrix whose diagonal elements are H_i^q. We define $\theta_i(q) = R_i - \overline{\gamma} + \Gamma L(q)$ to obtain:

$$V(q, M, s) = \mathbb{E}\left[q^T(R - p) - \frac{1}{2}q^T H^q q \middle| s, M\right] + \mathbb{E}\left[-\ell^T\theta(q) + L^T\Gamma\sum_i(\overline{\Lambda}_i^\tau)\tau - \frac{1}{2}\ell_i^T H_i^\ell \ell_i - \frac{1}{2}\tau^T\Delta\tau \middle| s, M\right],$$

[1] For a predictive model, Θ_0 is invariant to the model M. If the regulator only uses the prior, then $cov_0(\hat{L}(q)) = 0$ since $\hat{L}(q) = L_0(q)$.

$$V(q,M,s) = \mathbb{E}\Bigg[q^T(R-p) - \frac{1}{2}q^T H^q q - \ell(q)^T\theta(q)\Bigg|s,M\Bigg] + \mathbb{E}\Bigg[\tau^T\overline{\Lambda}^{\tau T}\theta(q) + \tau^T(\sum_i \overline{\Lambda}_i^\tau)^T\Gamma\Bigg(L(q) - (\sum_i \overline{\Lambda}_i^\tau)\tau\Bigg) - \frac{1}{2}\ell_i^T H_i^\ell \ell_i - \frac{1}{2}\tau^T\Delta\tau\Bigg|s,M\Bigg],$$

$$V(q,M,s) = \mathbb{E}\Bigg[q^T(R-p) - \frac{1}{2}q^T H^q q - \ell(q)^T\theta(q)\Bigg|s,M\Bigg] + \mathbb{E}\Bigg[\tau^T\Bigg(\overline{\Lambda}^{\tau T}\theta(q) + (\sum_i \overline{\Lambda}_i^\tau)^T\Gamma L(q)\Bigg) - \tau^T\Bigg((\sum_i \overline{\Lambda}_i^\tau)^T\Gamma(\sum_i \overline{\Lambda}_i^\tau) + \frac{1}{2}\Delta\Bigg)\tau - \frac{1}{2}\ell^T H^\ell \ell\Bigg|s,M\Bigg].$$

From here, we can expand out

$$\ell^T H^\ell \ell = \Bigg(\ell(q)^T - \tau^T\overline{\Lambda}^{\tau T}\Bigg)H^\ell\Bigg(\ell(q) - \overline{\Lambda}^\tau\tau\Bigg) = \ell(q)^T H^\ell \ell(q) - 2\tau^T\overline{\Lambda}^{\tau T}H^\ell\ell(q) + \tau^T\overline{\Lambda}^{\tau T}H^\ell\overline{\Lambda}^\tau\tau,$$

and substitute back in to obtain

$$\begin{aligned}V(q,M,s) = \mathbb{E}\Bigg[q^T(R-p) - \frac{1}{2}q^T H^q q - \ell(q)^T\theta(q) - \frac{1}{2}\ell(q)^T H^\ell\ell(q)\Bigg|s,M\Bigg] \\ + \mathbb{E}\Bigg[\tau^T\Bigg(\overline{\Lambda}^{\tau T}\theta(q) + (\sum_i \overline{\Lambda}_i^\tau)^T\Gamma L(q) + \overline{\Lambda}^{\tau T}H^\ell\ell(q)\Bigg) \\ - \frac{1}{2}\tau^T\Bigg(2(\sum_i \overline{\Lambda}_i^\tau)^T\Gamma(\sum_i \overline{\Lambda}_i^\tau) + \Delta + \overline{\Lambda}^{\tau T}H^\ell\overline{\Lambda}^\tau\Bigg)\tau\Bigg|s,M\Bigg].\end{aligned}$$

Note that this welfare function is necessarily equivalent to that obtained in Proposition 1. Therefore, an optimal tax formula that is equivalent to that in Proposition 1 yields:

$$\tau = \mathbb{E}\Bigg[2(\sum_i \overline{\Lambda}_i^\tau)^T\Gamma(\sum_i \overline{\Lambda}_i^\tau) + \Delta + \overline{\Lambda}^{\tau T}H^\ell\overline{\Lambda}^\tau\Bigg|s,M\Bigg]^{-1}\mathbb{E}\Bigg[\overline{\Lambda}^{\tau T}\theta(q) + (\sum_i \overline{\Lambda}_i^\tau)^T\Gamma L(q) + \overline{\Lambda}^{\tau T}H^\ell\ell(q)\Bigg|s,M\Bigg].$$

Accordingly substituting back into welfare, we have

$$\begin{aligned}V(q,M,s) =& \mathbb{E}\Bigg[q^T(R-p) - \frac{1}{2}q^T H^q q - \ell(q)^T\theta(q) - \frac{1}{2}\ell(q)^T H^\ell\ell(q)\Bigg|s,M\Bigg] \\ &+ \frac{1}{2}\tau^T\mathbb{E}\Bigg[2(\sum_i \overline{\Lambda}_i^\tau)^T\Gamma(\sum_i \overline{\Lambda}_i^\tau) + \Delta + \overline{\Lambda}^{\tau T}H^\ell\overline{\Lambda}^\tau\Bigg|s,M\Bigg]\tau.\end{aligned}$$

We can then characterize the expected welfare gains by

$$\begin{aligned}V(q,M) =& \mathbb{E}_0\Bigg[q^T(R-p) - \frac{1}{2}q^T H^q q - \ell(q)^T\theta(q) - \frac{1}{2}\ell(q)^T H^\ell\ell(q)\Bigg|M\Bigg] \\ &+ \frac{1}{2}\mathbb{E}_0\Bigg[\tau^T\mathbb{E}\Bigg[2(\sum_i \overline{\Lambda}_i^\tau)^T\Gamma(\sum_i \overline{\Lambda}_i^\tau) + \Delta + \overline{\Lambda}^{\tau T}H^\ell\overline{\Lambda}^\tau\Bigg|s,M\Bigg]\tau\Bigg|M\Bigg],\end{aligned}$$

where τ is given as in Proposition 1.

From here, we define $\Psi = \mathbb{E}\left[2(\sum_i \overline{\Lambda}_i^\tau)^T\Gamma(\sum_i \overline{\Lambda}_i^\tau)+\Delta+\overline{\Lambda}^{\tau T}H^\ell\overline{\Lambda}^\tau\middle| s, M\right]$, and note that $\Psi^T = \Psi$. We can then write more compactly,

$$\begin{aligned}V(q, M) =&\mathbb{E}_0\left[q^T(R-p) - \frac{1}{2}q^T H^q q - \ell(q)^T\theta(q) - \frac{1}{2}\ell(q)^T H^\ell \ell(q)\middle| M\right]\\ &+ \frac{1}{2}\mathbb{E}_0\left[\tau^T\Psi\tau\middle| M\right].\end{aligned}$$

Predictive-Causal Independence. Under predictive-causal independence, we have

$$\tau^* = \mathbb{E}\left[\Xi\middle| s, M\right]^{-1}\mathbb{E}\left[(\sum_i \overline{\Lambda}_i^\tau)^T\Gamma\middle| s, M\right]\mathbb{E}\left[L(q)\middle| s, M\right].$$

We therefore define $\Upsilon = \mathbb{E}\left[\Xi\middle| s, M\right]^{-1}\mathbb{E}\left[(\sum_i \overline{\Lambda}_i^\tau)^T\Gamma\middle| s, M\right]$ and define $\Theta = \Upsilon^T\Psi\Upsilon$. Then, defining $\hat{L}(q) = \mathbb{E}[L(q)|s, M]$ to be predicted liquidations, we have:

$$\begin{aligned}V(q, M) =&\mathbb{E}_0\left[q^T(R-p) - \frac{1}{2}q^T H^q q - \ell(q)^T\theta(q) - \frac{1}{2}\ell(q)^T H^\ell \ell(q)\middle| M\right]\\ &+ \frac{1}{2}\mathbb{E}_0\left[\hat{L}(q)^T\Theta\hat{L}(q)\middle| M\right].\end{aligned}$$

Using predictive-causal independence, we have:

$$\begin{aligned}V(q, M) =& \mathbb{E}_0\left[q^T(R-p) - \frac{1}{2}q^T H^q q - \ell(q)^T\theta(q) - \frac{1}{2}\ell(q)^T H^\ell \ell(q)\middle| M\right]\\ &+ \frac{1}{2}\mathbb{E}_0\left[\hat{L}(q)^T\Theta_0\hat{L}(q)\middle| M\right].\end{aligned}$$

Finally, expanding the second line, we obtain

$$\begin{aligned}V(q, M) =&\mathbb{E}_0\left[q^T(R-p) - \frac{1}{2}q^T H^q q - \ell(q)^T\theta(q) - \frac{1}{2}\ell(q)^T H^\ell \ell(q)\middle| M\right]\\ &+ \frac{1}{2}L_0(q)^T\Theta_0 L(q) + \frac{1}{2}\mathrm{tr}(\Theta_0 cov_0(\hat{L}(q))).\end{aligned}$$

B Architecture and Training Details

This appendix provides the detailed specifications deferred from Section 2.

Multi-Head Attention Computation. We allow for S_A distinct and independent attention heads. For each head $s \in \{1, \ldots, S_A\}$, we compute query, key, and value projections using learnable projection matrices $W_q^{(\ell,s)}, W_k^{(\ell,s)} \in \mathbb{R}^{d_k \times d_h}$ and $W_v^{(\ell,s)} \in \mathbb{R}^{d_v \times d_h}$, where $d_k = d_h/S_A$ and $d_v = d_h$.

Define

$$Q_{v,t}^{(\ell,s)} = W_q^{(\ell,s)} h_{v,t}^{(\ell-1)}, \qquad K_{u,t}^{(\ell,s)} = W_k^{(\ell,s)} h_{u,t}^{(\ell-1)}, \qquad V_{u,t}^{(\ell,s)} = W_v^{(\ell,s)} M_{u,t}^{(\ell)}. \tag{A.2}$$

For each edge (v, u), attention weights are formed by a neighborhood softmax, incorporating the normalized edge exposure $\tilde{w}_{vu,t}$ as an additive log-bias:

$$\alpha_{vu,t}^{(\ell,s)} = \frac{\exp\left(\frac{\left(Q_{v,t}^{(\ell,s)}\right)^T K_{u,t}^{(\ell,s)}}{\sqrt{d_k}} + \log \tilde{w}_{vu,t} \right)}{\sum_{u' \in \mathcal{N}_t(v)} \exp\left(\frac{\left(Q_{v,t}^{(\ell,s)}\right)^T K_{u',t}^{(\ell,s)}}{\sqrt{d_k}} + \log \tilde{w}_{vu',t} \right)}. \tag{A.3}$$

This ensures $\sum_{u \in \mathcal{N}_t(v)} \alpha_{vu,t}^{(\ell,s)} = 1$ for each head and node. The aggregated message for each node is

$$m_{v,t}^{(\ell,s)} = \sum_{u \in \mathcal{N}_t(v)} \alpha_{vu,t}^{(\ell,s)} V_{u,t}^{(\ell,s)}, \qquad m_{v,t}^{(\ell)} = \frac{1}{S_A} \sum_{s=1}^{S_A} m_{v,t}^{(\ell,s)}. \tag{A.4}$$

Loss Functions. For the masked autoencoder (MAE) objective, masked position sizes are predicted as $\hat{w}_{ia,t} = f_{\text{AE}}(e_{i,t}, e_{a,t})$. The autoencoder loss is $\mathcal{L}_{\text{AE}} = \sum_{(i,a) \in \mathcal{E}_{\text{masked},t}} (g(w_{ia,t}) - g(\hat{w}_{ia,t}))^2$, where $\mathcal{E}_{\text{masked},t} \subset \mathcal{E}_t$ is the set of masked edges and $g(\cdot) = \text{sgn}(\cdot) \log(|\cdot|)$. The supervised trade prediction loss combines binary cross-entropy for the extensive margin and mean squared error for the intensive margin:

$$\mathcal{L}_{\text{TP}} = \sum_{(i,a) \in \mathcal{E}_t} \text{BCE}(\hat{c}_{ia,t+1}, \mathbf{1}[y_{ia,t+1} \neq 0]) + \sum_{\substack{(i,a) \in \mathcal{E}_t \\ y_{ia,t+1} \neq 0}} (y_{ia,t+1} - \hat{m}_{ia,t+1})^2, \tag{A.5}$$

where $\text{BCE}(\hat{c}, y) = -[y \log \sigma(\hat{c}) + (1 - y) \log(1 - \sigma(\hat{c}))]$. The overall training procedure minimizes these losses sequentially:

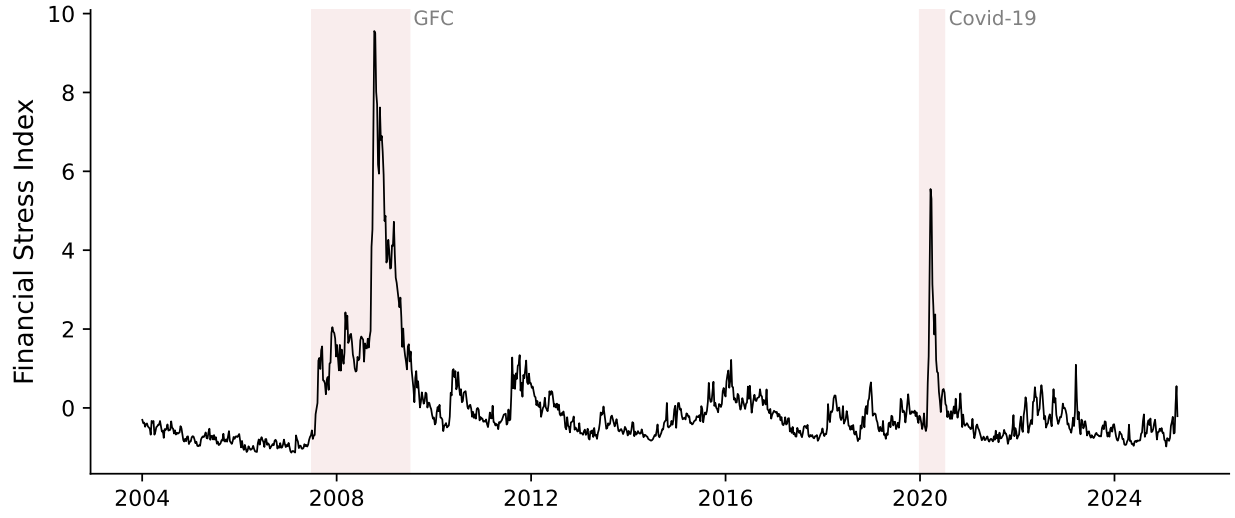
$$\text{Phase 1: } \min_{\Theta_{\text{Backbone}}, \Theta_{\text{AE}}} \mathcal{L}_{\text{AE}}, \qquad \text{Phase 2: } \min_{\Theta_{\text{Backbone}}, \Theta_{\text{TP}}} \mathcal{L}_{\text{TP}}, \tag{A.6}$$

where Θ_{Backbone}, Θ_{AE}, and Θ_{TP} denote the learnable parameter vectors for the graph transformer backbone, masked autoencoder head, and trade prediction head, respectively.

Model Parameters. The trainable components include the feature map ϕ, the pre-embedding functions for categorical characteristics, the message feed-forward layers $M^{(\ell)}$, the attention mechanism projection matrices $W_q^{(\ell,s)}$, $W_k^{(\ell,s)}$, and $W_v^{(\ell,s)}$, the node update function $U^{(\ell)}$, the embeddings projection layer ρ, and the task-specific prediction heads f_{AE} and f_{TP}. We collect the set of parameters in all these learnable components in the vector Θ.[2]

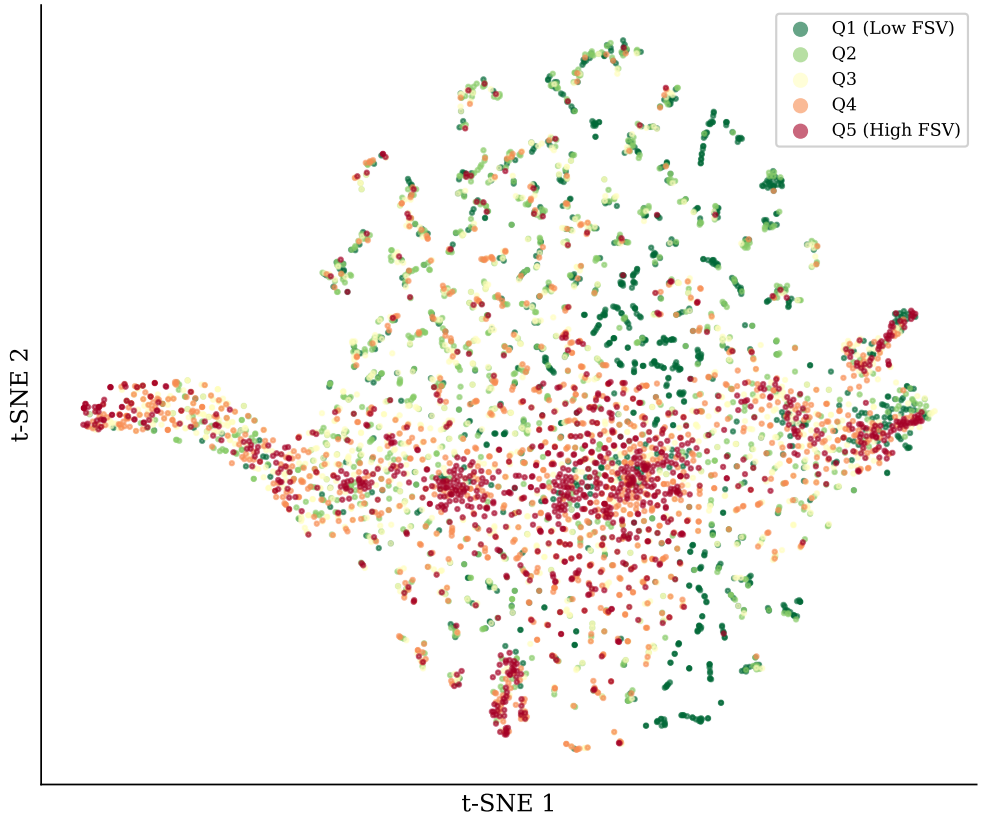
[2] All feed-forward layers in our architecture use GELU non-linearities (Hendrycks and Gimpel 2016). These help avoid dead gradients during training.

Figure A.I: **St. Louis Fed Financial Stress Index**



Notes: Weekly observations; shaded regions indicate the Great Financial Crisis (2007–2009) and the Covid-19 crisis (2020). The index is centered at zero.

Figure A.II: **Asset embeddings encode fire-sale vulnerability**



Notes: t-SNE projection of asset embeddings from the 2019Q4 cross-section, colored by quintile of the fire-sale vulnerability (FSV) score. FSV is defined as the product of flow-sensitive holder share (i.e., the fraction of the security that is held by open-end mutual funds or ETFs) and asset illiquidity (measured by 1/ADV). Each point represents an asset's embedding, projected down to two dimensions via t-SNE (perplexity = 50, 2,000 iterations).

C Theory Appendix

In this appendix, we extend the setup to incorporate a date 0 in which asset positions are chosen. The structure of dates 1 and 2 are the same as before.

Intermediaries. At date 0, intermediary i invests in a vector $q_i = (q_{i1}, \ldots, q_{iN})^T$ of assets. If $q_{in} < 0$, then intermediary i has undertaken a negative investment (shorting) asset n.[3] The cost to intermediary i of producing the asset vector q_i is $C_i(q_i) = p_i^T q_i - \frac{1}{2} q_i^T H_i^q q_i$, where p_{in} is the per-unit cost to i of producing asset n and where $p_i = (p_{i1}, \ldots, p_{iN})^T$. The cost component $q_i^T H_i^q q_i$ is a quadratic adjustment/holding cost, where H_i^q is an $N \times N$ matrix. Intermediary i's total payoff (consumption) at date 2, including costs incurred at date 0, is

$$U_i = q_i^T (R_i - p_i) - \ell_i^T (R_i - \gamma) - \frac{1}{2} q_i^T H_i^q q_i - \frac{1}{2} \ell_i^T H_i^\ell \ell_i. \tag{A.7}$$

Information Structure at Date 0. We assume that although nature draws the true parameters at date 0, all agents at date 0 are uncertain about the true values of the parameters $\Phi = \{A_i^q, A_i^\ell, H_i^\ell, \rho_i, R_i, \overline{\gamma}\}$. All agents, including the regulator, have a common prior $\Phi \sim \mu_0$ at date 0.[4] At date 1, as in the baseline setup, the true model parameters become common knowledge of private agents, and the problem unfolds as in the main text.[5]

Competitive Equilibrium. We define a competitive equilibrium as follows.

Definition 4. *A competitive equilibrium of the model, given true model parameters Φ, is a vector of prices γ and a set of allocations $\{q_i, \ell_i, L\}$ such that:*

1. *At date 1, taking as given asset allocations $\{q_i\}$ and model parameters Φ:*
   (a) *Intermediary i chooses ℓ_i to maximize utility (equation A.7) subject to the rollover constraint (13), taking prices γ as given.*
   (b) *Arbitrageurs choose L maximize utility (equation 15), taking prices γ as given.*
   (c) *The liquidation markets clear (equation 16).*
2. *At date 0:*
   (a) *Intermediary i chooses q_i to maximize expected utility $\mathbb{E}_0[U_i]$, where $\mathbb{E}_0$ denotes the expectation given the prior μ_0 over model parameters Φ.*

[3] We could endow intermediaries with a stock of assets, but for expositional convenience we set that stock to 0.

[4] In practice, intermediaries may hold more precise priors over model parameters than the regulator, in which case the regulator could also infer structural information from observed portfolio choices. One advantage of the ex-post intervention in such an extension would be its ability to react to information revealed by intermediaries' choices, potentially reducing risks of moral hazard or imperfectly calibrated regulation ex ante.

[5] Assuming that private agents and the regulator have a common prior at date 0 simplifies analysis because it prevents the regulator from learning information about these parameters from inference about private sector beliefs based on the asset allocation. It also eliminates a regulatory incentive based on different beliefs the regulator and agents (e.g., Fontanier 2025.)

C.1 Expected Welfare from a Model and Optimal Model Choice

We next ask how a regulator at date 1 would choose a model $M \in \mathcal{M}$ under discretion, taking as given the asset holdings q of intermediaries. We maintain simplicity by assuming predictive-causal independence for the main text (Definition 2) and by focusing on predictive models (Definition 3).

We assume that the regulator faces a separable utility cost $\mathcal{C}(M, q)$ from adopting (predictive) model $M \in \mathcal{M}$. In Lemma 2, for a predictive model (Definition 3) the key sufficient statistic of the model from a welfare perspective is the prior covariance matrix over the ex-post prediction of liquidations, $\Sigma_0^{\hat{L}} = cov_0(\hat{L}(q))$. The matrix Θ_0 and the averaged expected liquidations $L_0(q)$ are invariant to model choice. For symmetric matrices $\Sigma_0^{\hat{L}}$ that are attainable from some model M, we can therefore re-represent costs as $C(\Sigma_0^{\hat{L}}, q) = \inf_{M \in \mathcal{M} | cov_0(\hat{L}(q)) = \Sigma_0^{\hat{L}}} \mathcal{C}(M, q)$. We assume that C is differentiable in $(\Sigma_0^{\hat{L}}, q)$ over an open ball that contains its optimal value.

As a result, for a predictive model we can write the regulator's optimization problem as

$$\max_{\Sigma_0^{\hat{L}}} \frac{1}{2} \text{tr}\left(\Theta_0 \Sigma_0^{\hat{L}}\right) - C(\Sigma_0^{\hat{L}}, q)$$

subject to the constraint that $\Sigma_0^{\hat{L}}$ is a symmetric matrix. We obtain the following result on the optimal predictive model.

Proposition 3. *The regulator's optimal predictive model satisfies*

$$\frac{\partial C(\Sigma_0^{\hat{L}}, q)}{\partial \Sigma_0^{\hat{L}}} + \left(\frac{\partial C(\Sigma_0^{\hat{L}}, q)}{\partial \Sigma_0^{\hat{L}}}\right)^T = \Theta_0 \tag{A.8}$$

where $\frac{\partial C(\Sigma_0^{\hat{L}}, q)}{\partial \Sigma_0^{\hat{L}}}$ *is a square matrix whose ij-th element is* $\frac{\partial C(\Sigma_0^{\hat{L}}, q)}{\partial (\Sigma_0^{\hat{L}})_{ij}}$.

Proposition 3 yields an intuitive trade-off on optimal predictive model choice in terms of choice of covariance matrix of the policy intervention. The regulator is willing to pay a higher marginal cost to increase precision on dimensions where Θ_0 is larger, that is when the policy has greater impact on aggregate liquidations and prices, when the regulatory cost is higher, or when the impact through holding costs is higher. In this sense, there is a complementarity between predictive power and knowledge of the causal policy impact: the regulator is willing to pay larger costs to acquire precise predictions of aggregate liquidations precisely on dimensions on which the policy impact on liquidations and liquidation prices is anticipated to be largest.[6]

[6] In a dynamic extension where the one-shot model is embedded as a stage game played repeatedly, the regulator would face a trade-off between acquiring predictive information useful for the current crisis and uncovering structural parameters valuable for future interventions. We conjecture that predictive models would be relatively myopically useful, providing substantial information for current intervention but limited updating about the true distribution of underlying parameters. This could face the regulator with an interesting dynamic trade-off between myopically acquiring more predictive information, and trying to uncover the true structural parameters that would also be useful for designing regulation and interventions during the next crisis.

C.2 Privately Optimal Asset Allocations

To examine how model choice and ex-post intervention shape the ex-ante asset allocations of intermediaries, we begin by studying the private optimum of individual intermediaries, and then study socially optimal interventions. Intermediary i at date 0 takes as given the model choice M^* of the regulator and the resulting possible equilibria,[7] and solves

$$\max_{q_i} \mathbb{E}_0\left[q_i^T(R_i - p_i) - \ell_i^T(R_i + \tau_i^* - \gamma) - \frac{1}{2}q_i^T H_i^q q_i - \frac{1}{2}\ell_i^T H_i^\ell \ell_i\right].$$

Intermediary i's asset allocation is therefore affected both through the specific intervention τ_i^* anticipated, and also through the liquidation price (which in turn also affects equilibrium liquidations). Intermediary i knows that equilibrium liquidations are determined as in Lemma 1, but only internalizes the effect of its own q_i on its own liquidations (and not on equilibrium liquidation price or equilibrium liquidations, and not on the optimal model choice or intervention).

The following proposition characterizes intermediaries' privately optimal asset allocations.[8]

Proposition 4. *The privately optimal asset allocation satisfies*

$$\mathbb{E}_0\left[H_i^q q_i^* + \Lambda_i^{qT} H_i^{\ell} \overline{\ell}_i(q^*)\right] = \mathbb{E}_0\left[R_i - p_i - \Lambda_i^{qT}\theta_i(q^*)\right] - \mathbb{E}_0\left[\Lambda_i^{qT}\left(\tau_i^* - \left(H_i^{\ell}\overline{\Lambda_i^{\tau}} + \Gamma(\sum_i \overline{\Lambda_i^{\tau}})\right)\tau^*\right]\right. \tag{A.9}$$

Proposition 4 expresses the optimal portfolio choice in the form of a marginal cost-marginal benefit trade-off. The left hand side captures the marginal cost of increasing holdings of an asset, which includes both the ex-ante and ex-post adjustment costs of holding more and liquidating more of an asset. This term is evaluated at the no-intervention benchmark, that is as-if we had $\tau^* = 0$.

The right hand side captures the marginal benefit, which is decomposed into a marginal benefit absent the ex-post regulatory intervention (the first term) plus the marginal benefit arising from the impact of intervention (the second term). The first term on the RHS is the baseline expected asset return, $R_i - p_i$, net of costs of liquidations. The liquidation costs are the amount liquidations are changed by increasing holdings, Λ_i^{qT}, times the fire sale loss in liquidation, $\theta_i(q^*)$.

The final term on the RHS captures the impacts of model choice and the ex-post intervention. This is the only term in equation A.9 that depends on model choice and intervention. It captures the cost of holding an asset induced through regulation. The expectation of ex-post policy intervention has two impacts. First, a higher wedge τ_{in} on intermediary i increases the cost of liquidating asset n, discouraging the intermediary from holding portfolios that result in it liquidating that asset ex post. There are also equilibrium costs of the vector of wedges τ through the equilibrium price. In contrast, here a higher wedge τ_{in} encourages the intermediary i to hold more of an asset when the asset price rises as a result, which partially offsets the benefit from raising the price. This is a standard

[7] Recall that at date 0, the intermediary does not know the true parameters.

[8] Note that Proposition 4 does not rely on predictive-causal independence or a predictive model.

channel of moral hazard. This moral hazard channel is amplified when the regulator's instruments do not cover all holders of a given asset, as unregulated intermediaries capture the benefit of higher liquidation prices without bearing any cost, so their ex-ante incentive to accumulate the asset is undeterred by the prospect of intervention.

It is clear that the ex-post intervention affects the optimal asset allocation: a higher expected tax on asset n ex post directly discourages its purchase ex ante, but a higher liquidation price ex post encourages its purchase ex ante. As a result, even a purely predictive model can impact the asset allocation ex ante. In particular, even though under a purely predictive model (Definition 3) the expectation of τ^* is the same as if the regulator ran no model, there is a covariance induced between the tax itself, τ^*, and the impact on liquidations, Λ_i^q, as long as the predictive model is loading at least some on Bayesian inference on the impact of asset holdings on liquidations Λ_i^q. That is to say, focusing on the direct term $\mathbb{E}_0[\Lambda_i^{qT}\tau_i^*]$, for a purely predictive model we can write

$$\mathbb{E}_0\left[\Lambda_i^{qT}\tau_i^*\right] = \mathbb{E}_0\left[\Lambda_i^{qT}\right]\mathbb{E}_0\left[\tau_i\right] + \mathrm{cov}_0\left(\Lambda_i^{qT}, \mathbb{E}_0[\Xi]^{-1}\mathbb{E}_0\left[(\sum_i \overline{\Lambda}_i^{\tau})^T\Gamma\right]\mathbb{E}[L(q)|s,M]\right),$$

where the second term reflects how the true value of Λ_i^q varies with the prediction of liquidations. In a similar fashion, moral hazard can also be exacerbated if the impact of asset allocations on liquidations Λ_i^q is a key driver of the predictability of ex post liquidations.

More broadly, both the virtuous deterrence and the moral hazard channels are bounded by the precision of the regulator's model: a less granular model generates smaller covariance between the intervention and the structural loadings, and hence attenuates both the beneficial and the adverse feedback of the ex-post intervention on ex-ante portfolio choice.[9]

C.3 Socially Optimal Asset Allocation

Finally, we study the constrained efficient asset allocation q of the regulator. Formally, we assume that the regulator can impose wedges on asset holdings at date 0, that is a revenue-neutral tax $t_i = (t_{i1}, \ldots, t_{iN})^T$. Because the regulator has a complete set of wedges ex ante, this is equivalent to the regulator directly picking portfolio allocations ex ante.[10] The following proposition characterizes the optimal wedges $t = (t_1^T, \ldots, t_I^T)^T$ when the regulator chooses an optimal predictive model ex post.

[9] This gives rise to a potential time-consistency problem in model choice, since the model choice affects ex-ante allocations. The tension is likely more pronounced when the regulator cannot regulate all intermediaries or all assets ex ante. We leave the implications of commitment versus discretion for optimal model design to future work.

[10] It is straightforward to extend results to incomplete ex-ante instruments, but for brevity we focus on complete instruments.

Proposition 5. *For a predictive model, the social planner's optimal date* 0 *portfolio wedges* t *are*

$$\begin{aligned} t =& \mathbb{E}_0\left[(\sum_j \overline{\Lambda}_j^q)^T \Gamma L(q) - \Lambda^{q,eT} \sum_i \left(\theta_i(q) + H_i^\ell \ell_i(q)\right)\right] \\ &+ \frac{dC(\Sigma_0^{\hat{L}}, q)}{dq} - \mathbb{E}_0\left[(\sum_i \overline{\Lambda}_i^q)^T\right]\Theta_0 \mathbb{E}_0[L(q)] - \mathbb{E}_0\left[\Psi\tau\right] \end{aligned} \tag{A.10}$$

where $\mathbb{E}_0[\Psi\tau]$ *is a stacked vector whose* i^{th} *block is* $\mathbb{E}_0\left[\Lambda_i^{qT}\left(\tau_i - \left(\Gamma(\sum_i \overline{\Lambda}_i^\tau) + H_i^\ell \overline{\Lambda}_i^\tau\right)\tau\right)\right]$.

The optimal ex-ante asset tax formula is intuitive and is decomposed into two lines. The first line captures the uninternalized effects of intermediaries increasing holdings of an asset in absence of regulatory intervention. First, as q changes, total liquidations change across intermediaries, resulting in changes in the liquidation price and hence changes in revenue proceeds from total liquidations $L(q)$. Second, changes in forced liquidations through changes in the equilibrium price result in losses due to the size of the discount $\theta_i(q)$ and due to changes in marginal holding costs $H_i^\ell \ell_i(q)$. In absence of an ex-post intervention, this first line would capture the standard macroprudential regulation of asset positions.[11]

The second line captures the interaction of the ex-ante asset regulation with the ex-post model design and intervention. The first term, dC/dq, reflects how the changing assets held by intermediaries affects the cost of acquiring information through the model ex post. The regulator imposes larger holding taxes on asset positions that make running the model more costly ex post. For example, the regulator might discourage holdings of non-transparent or hard-to-assess assets.

The second term on the second line captures the effect on the expected intervention size (the term $L_0(q)^T \Theta_0 L_0(q)$ in equation A.1). Intuitively, concentrations into a position q increase total liquidations on the margin by $\sum_i \overline{\Lambda}_i^q$, which prompts a larger ex post intervention by the regulator. This increase in ex-post intervention helps to manage the fire sale and so mitigates the impact of the ex-ante increase in asset allocation. This gives a first dimension whereby the ex-post intervention provides a partial substitute for the ex-ante asset tax, and leads the regulator to require potentially smaller interventions ex ante, knowing that fire sales will be managed ex post.

Finally, there are the direct and moral hazard effects on intermediaries, $\mathbb{E}_0[\Psi\tau^*]$, are those in the final term in equation A.9 of Proposition 4. First, the prospect of the ex-post tax on liquidations directly discourages holdings of the asset ex ante when holding more promotes liquidations ex post. Second, there is the moral hazard effect: as the liquidation price rises, intermediaries' perceived cost of liquidations $\theta_i(q)$ falls, and so they become more willing to hold more of an asset even if that forces them to liquidate. It is interesting again to observe that for assets for which the direct effect dominates the indirect price effect, the ex-post tax actually serves as a substitute for ex-ante regulation, leading to a smaller intervention t ex ante.

[11] Since our framework allows negative positions $q_{in} < 0$, we can think of these negative positions as liabilities. As such, it also captures regulation of such liabilities.

C.4 Ex-Post Subsidy-Based Interventions

Our baseline analysis assumes that the regulator directly manages liquidations through a wedge/tax on liquidations. In practice, ex-post interventions can also involve "bailout" interventions that subsidize retaining assets rather than taxing selling them (e.g. lender of last resort, debt guarantees). Suppose that rather than taxing liquidations ℓ_i, the regulator instead applies a subsidy on the amount $q_i - \ell_i$ of the asset that is held to maturity. Formally, the payoff to intermediary i at date 2 is now

$$U_i = U_i + (q_i^T - q_i^{*T})\tau_i - (\ell_i^T - \ell_i^{*T})\tau_i,$$

so that the regulator continues to apply revenue-neutral interventions.[12]

Because the regulator at date 1 takes the asset allocations as given, the regulator's optimization problem over both model choice and the ex-post intervention are formally the same as before. Thus, Lemma 1 and Propositions 1-3 continue to apply. However, the privately optimal asset allocation is affected by the subsidy on asset retention. In particular, in Proposition 4, the expected return on holding asset q_i is raised from R_i to $R_i + \tau_i^*$. Intuitively an ex-post subsidy promotes overinvestment in that asset. This gives rise to a familiar channel of moral hazard.

Relative to the case of a liquidation tax, it is worthwhile to note that the only new term for private intermediaries in their optimization problem is the revenue benefits $q_i^T \mathbb{E}_0[\tau_i^*]$ of their asset position. Therefore, relative to the liquidation tax, the regulator's model choice only affects intermediaries' asset allocation ex ante to the extent it changes the expected size of the intervention, $\mathbb{E}_0[\tau_i^*]$. In particular for a predictive model under predictive-causal independence (Definitions 2 and 3), we know that $\mathbb{E}_0[\tau_i^*]$ does not depend on the model used; instead, it is only the covariance matrix of the intervention (its precision) that depends on the model. It is thus interesting and surprising that purely predictive models that help with prediction but not with causal inference do not exacerbate moral hazard relative to the case of the liquidation tax. In contrast, models that are informative about the causal impact of the policy intervention change expectations of the policy intervention, and are potentially associated with moral hazard. The logic is reminiscent of that of Laffont and Tirole (1986) in the context of regulation of a firm with unobservable effort and uncertain costs.

C.5 Relaxed Information Structure for Policy Invariance

The baseline theory assumes that private agents observe the full state Φ in the Middle. This section shows how the policy invariance result can be relaxed under a substantially weaker condition that regulator-generated public information is informative about the predictive component of the state and not about the transmission component.

We illustrate the relaxation as follows. Suppose that private agents are uncertain over the return

[12] Since individual intermediaries take revenue remissions as given, there is still moral hazard from the subsidy.

R_i, sharing a common prior with the regulator, and that the regulator makes public its model M and signal s before private agents choose liquidations; we denote this regulator-supplied information $d = (M, s)$. Let $B \equiv \sum_i \Lambda_i^\tau$, so that $\Xi \equiv \Lambda_\tau^\top + B^\top \Gamma B + \Delta$, where the matrices Λ_i^τ and Ξ are those defined in Proposition 1. Suppose that:

(i) the transmission objects $\Lambda \equiv \{\Lambda_i^q, \Lambda_i^\tau, \Lambda_i^{q,e}, \Lambda_i^{\tau,e}\}_{i=1}^I$ are observed by private agents in the Middle;

(ii) after observing d, private agents update Bayesianly from the common prior.

Define the posterior-predictive no-intervention liquidation vector

$$\widehat{L}_d(q) \equiv \mathbb{E}[L(q) \mid \Lambda, \Xi, d].$$

Then, the proof of Lemma 1 follows as before, with $\mathbb{E}[R_i|d]$ replacing R_i. As a result, we can write

$$L(q, \tau; d) = \widehat{L}_d(q) - B\tau,$$

where $L(q)$ is liquidations absent intervention. All dependence of equilibrium outcomes on the regulator-generated public signal d enters through the posterior-predictive baseline liquidation vector $\widehat{L}_d(q)$; the slope matrices mapping portfolios and wedges into outcomes are unchanged. Then under predictive-causal independence, the regulator's optimal wedges satisfy

$$\tau^*(d) = \Xi^{-1} B^\top \Gamma \widehat{L}_d(q).$$

C.6 Proofs

C.6.1 Proof of Proposition 3

The optimal model choice solves

$$\max_{\Sigma_0^{\hat{L}}} \frac{1}{2} \text{tr}\left(\Theta_0 \Sigma_0^{\hat{L}}\right) - C(\Sigma_0^{\hat{L}}, q)$$

subject to the restriction that $\Sigma_0^{\hat{L}}$ is symmetric. Taking the first order condition, we have

$$\frac{\partial C(\Sigma_0^{\hat{L}}, q)}{\partial \Sigma_0^{\hat{L}}} + \left(\frac{\partial C(\Sigma_0^{\hat{L}}, q)}{\partial \Sigma_0^{\hat{L}}}\right)^T = \Theta_0$$

since Θ_0 is symmetric.

C.6.2 Proof of Proposition 4

The intermediary i ex ante optimization problem is

$$\max_{q_i} \mathbb{E}_0\Big[\sum_i \Big[q_i^T(R_i - p_i) - \ell_i^T(R_i + \tau_i^* - \gamma) - \frac{1}{2}q_i^T H_i^q q_i - \frac{1}{2}\ell_i^T H_i^\ell \ell_i\Big],$$

where we have

$$\ell_i = \overline{\ell}_i + \Lambda_i^q q_i - \sum_j \Lambda_j^{q,e} q_j^* - \overline{\Lambda}_i^\tau \tau^*.$$

So, we have the FOC

$$0 = \mathbb{E}_0\Big[R_i - p_i - \Lambda_i^{qT}(R_i + \tau_i^* - \gamma) - H_i^q q_i - \Lambda_i^{qT} H_i^\ell \ell_i\Big].$$

Substituting in for ℓ_i, we have

$$\mathbb{E}_0\Big[H_i^q\Big]q_i = \mathbb{E}_0\Big[R_i - p_i - \Lambda_i^{qT}(R_i + \tau_i^* - \gamma) - \Lambda_i^{qT} H_i^\ell\Big(\overline{\ell}_i + \Lambda_i^q q_i - \sum_j \Lambda_j^{q,e} q_j^* - \overline{\Lambda}_i^\tau \tau^*\Big)\Big],$$

$$\mathbb{E}_0\Big[H_i^q + \Lambda_i^{qT} H_i^\ell \Lambda_i^q\Big]q_i = \mathbb{E}_0\Big[R_i - p_i - \Lambda_i^{qT}(R_i - \gamma) - \Lambda_i^{qT} H_i^\ell\Big(\overline{\ell}_i - \sum_j \Lambda_j^{q,e} q_j^*\Big)\Big] + \mathbb{E}_0\Big[-\Lambda_i^{qT}\tau_i^* + \Lambda_i^{qT} H_i^\ell \overline{\Lambda}_i^\tau \tau^*\Big].$$

We use that $R_i - \gamma = \theta_i(q^*) - \Gamma(\sum_i \overline{\Lambda}_i^\tau)\tau^*$ to obtain

$$\mathbb{E}_0\Big[H_i^q q_i^* + \Lambda_i^{qT} H_i^\ell \overline{\ell}_i(q^*)\Big] = \mathbb{E}_0\Big[R_i - p_i - \Lambda_i^{qT}\theta_i(q^*)\Big] - \mathbb{E}_0\Big[\Lambda_i^{qT}\Big(\tau_i^* - \Big(H_i^\ell \overline{\Lambda}_i^\tau + \Lambda_i^{qT}\Gamma(\sum_i \overline{\Lambda}_i^\tau)\Big)\tau^*\Big].$$

C.6.3 Proof of Proposition 5

We have that:

$$\begin{aligned} V(q, M) = \mathbb{E}\Big[q^T(R - p) - \frac{1}{2}q^T H^q q - \ell(q)^T\theta(q) - \frac{1}{2}\ell(q)^T H^\ell \ell(q)\Big] \\ + \frac{1}{2}L_0(q)\Theta_0 L_0(q) + \frac{1}{2}\mathrm{tr}(\Theta_0 cov_0(\hat{L}(q)). \end{aligned}$$

Recalling that $\ell_i(q) = \overline{\ell}_i + \overline{\Lambda}_i^q q$, $\theta_i(q) = R_i - \overline{\gamma} + \Gamma L(q)$, and $L(q) = \sum_i \overline{\ell}_i + \sum_i \overline{\Lambda}_i^q q$, then we can expand:

$$\sum_i \ell_i(q)^T\theta_i(q) = \sum_i \overline{\ell}_i^T\theta_i(0) + q^T\sum_i \overline{\Lambda}_i^{qT}\theta_i(0) + q^T(\sum_j \overline{\Lambda}_j^q)^T\Gamma L(0) + q^T\sum_i\Big(\overline{\Lambda}_i^{qT}\Big)\Big(\Gamma\sum_j \overline{\Lambda}_j^q\Big)q,$$

$$\ell(q)H^\ell\ell(q) = \sum_i \overline{\ell}_i^T H_i^\ell \overline{\ell}_i + 2q^T\Big(\sum_i \overline{\Lambda}_i^{qT} H_i^\ell \overline{\ell}_i\Big) + q^T\Big(\sum_i \overline{\Lambda}_i^{qT} H_i^\ell \overline{\Lambda}_i^q\Big)q.$$

For a predictive model, Θ_0 is invariant to the model choice M and to q. Therefore by the

Envelope Theorem,

$$0 = \mathbb{E}_0\Bigg[R - p - H^q q - \Bigg[\sum_i \overline{\Lambda}_i^{qT}\theta_i(0) + (\sum_j \overline{\Lambda}_j^q)^T \Gamma L(0) + 2\sum_i \Big(\overline{\Lambda}_i^{qT}\Big)\Big(\Gamma \sum_j \overline{\Lambda}_j^q\Big)q\Bigg] - \Bigg[\Big(\sum_i \overline{\Lambda}_i^{qT} H_i^\ell \overline{\ell}_i\Big) + \Big(\sum_i \overline{\Lambda}_i^{qT} H_i^\ell \overline{\Lambda}_i^q\Big)q\Bigg]\Bigg] - \frac{dC(\Sigma_0^{\hat{L}}, q)}{dq} + \mathbb{E}_0\Bigg[(\sum_i \overline{\Lambda}_i^q)^T\Bigg]\Theta_0 \mathbb{E}_0[L(q)],$$

$$0 = \mathbb{E}_0\Bigg[R - p - H^q q - \Bigg[\sum_i \overline{\Lambda}_i^{qT}\Big(\theta_i(0) + \Big(\Gamma \sum_j \overline{\Lambda}_j^q\Big)q\Big) + (\sum_j \overline{\Lambda}_j^q)^T \Gamma\Big(L(0) + \sum_j \overline{\Lambda}_j^q q\Big)\Bigg] - \Bigg[\Big(\sum_i \overline{\Lambda}_i^{qT} H_i^\ell \overline{\ell}_i\Big) + \Big(\sum_i \overline{\Lambda}_i^{qT} H_i^\ell \overline{\Lambda}_i^q\Big)q\Bigg]\Bigg] - \frac{dC(\Sigma_0^{\hat{L}}, q)}{dq} + \mathbb{E}_0\Bigg[(\sum_i \overline{\Lambda}_i^q)^T\Bigg]\Theta_0 \mathbb{E}_0[L(q)]\,.$$

We have

$$\theta_i(0) + \Gamma\Big(\sum_j \overline{\Lambda}_j^q\Big)q = \theta_i(q),$$

$$L(0) + \Big(\sum_j \overline{\Lambda}_j^q\Big)q = L(q),$$

so we get:

$$\mathbb{E}_0\Big[H^q q\Big] = \mathbb{E}_0\Big[R - p - \sum_i \overline{\Lambda}_i^{qT}\Big(\theta_i(q) + H_i^\ell \ell_i(q)\Big)\Big] - \mathbb{E}_0\Big[(\sum_j \overline{\Lambda}_j^q)^T \Gamma L(q)\Big] - \frac{dC(\Sigma_0^{\hat{L}}, q)}{dq} + \mathbb{E}_0\Bigg[(\sum_i \overline{\Lambda}_i^q)^T\Bigg]\Theta_0 \mathbb{E}_0[L(q)].$$

Recall that we have

$$\ell_i(q) = \overline{\ell}_i + \Lambda_i^q q_i - \sum_j \Lambda_j^{q,e} q_{ij},$$

then

$$\sum_i \overline{\Lambda}_i^{qT}\Big(\theta_i(q) + H_i^\ell \ell_i(q)\Big) = \Lambda^{qT}\Big(\theta(q) + H^\ell \ell(q)\Big) - \sum_i (\Lambda_1^{q,e}, \ldots, \Lambda_I^{q,e})^T\Big(\theta_i(q) + H_i^\ell \ell_i(q)\Big),$$

so we can then write

$$\mathbb{E}_0\Big[H^q q\Big] = \mathbb{E}_0\Big[R - p - \Lambda^{qT}\Big(\theta(q) + H^\ell \ell(q)\Big)\Big] + \mathbb{E}_0\Big[\sum_i (\Lambda_1^{q,e}, \ldots, \Lambda_I^{q,e})^T\Big(\theta_i(q) + H_i^\ell \ell_i(q)\Big) - (\sum_j \overline{\Lambda}_j^q)^T \Gamma L(q)\Big] - \frac{dC(\Sigma_0^{\hat{L}}, q)}{dq} + \mathbb{E}_0\Big[(\sum_i \overline{\Lambda}_i^q)^T\Big]\Theta_0 \mathbb{E}_0[L(q)].$$

Now, recall that we can write the private FOC, subject to an ex-ante asset tax t_i, as

$$\mathbb{E}_0\Big[H_i^q\Big]q_i = \mathbb{E}_0\Big[R_i - (p_i + t_i) - \Lambda_i^{qT}\Big(\theta_i(q^*) + H_i^\ell \ell_i(q)\Big)\Big] - \mathbb{E}_0\Big[\Lambda_i^{qT}\Big(\tau_i - \Big(\Gamma(\sum_i \overline{\Lambda}_i^\tau) + H_i^\ell \overline{\Lambda}_i^\tau\Big)\tau\Big)\Big].$$

Therefore, stacking and combining the two equations we obtain:

$$\begin{aligned} -t - \mathbb{E}_0\Big[\Psi\tau\Big] = \mathbb{E}_0\Big[\sum_i (\Lambda_1^{q,e}, \ldots, \Lambda_I^{q,e})^T\Big(\theta_i(q) + H_i^\ell \ell_i(q)\Big) \\ - (\sum_j \overline{\Lambda}_j^q)^T \Gamma L(q)\Big] - \frac{dC(\Sigma_0^{\hat{L}}, q)}{dq} + \mathbb{E}_0\Big[(\sum_i \overline{\Lambda}_i^q)^T\Big]\Theta_0 \mathbb{E}_0[L(q)], \end{aligned}$$

where $\mathbb{E}_0\Big[\Psi\tau\Big]$ is the stacking of $\mathbb{E}_0\Big[\Lambda_i^{qT}\Big(\tau_i - \Big(\Gamma(\sum_i \overline{\Lambda}_i^\tau) + H_i^\ell \overline{\Lambda}_i^\tau\Big)\tau\Big)\Big]$. Thus defining $\Lambda^{q,e} = (\Lambda_1^{q,e}, \ldots, \Lambda_I^{q,e})$, we have

$$t = \mathbb{E}_0\Big[(\sum_j \overline{\Lambda}_j^q)^T \Gamma L(q) - \Lambda^{q,eT} \sum_i \Big(\theta_i(q) + H_i^\ell \ell_i(q)\Big)\Big].$$