From Prediction to Causal Interpretation: A DML Case Study in Financial Economics

Peilin Rao
Department of Economics
University of California, Los Angeles
Los Angeles, CA 90024
jackrao@g.ucla.edu

Randall R. Rojas
Department of Economics
University of California, Los Angeles
Los Angeles, CA 90024
rrojas@econ.ucla.edu

Abstract

This paper presents a case study on bridging the translational gap between advanced causal machine learning and scientific practice in financial economics, directly addressing the core questions of the CauScien workshop. We tackle a fundamental scientific question: what are the causal drivers of stock market troughs? Moving beyond the "black box" prediction paradigm, we implement a novel, two-stage comparative causal analysis designed for a complex, real-world setting. We first establish a baseline using Double/Debiased Machine Learning (DML) for a standard partially linear model. Recognizing the limitations of this assumption in a non-linear domain, we then employ a more flexible DML specification to estimate the Average Partial Effect (APE), which is better suited to our binary, interactive setting. The comparison reveals that conclusions about economic causality are critically sensitive to model specification. The more flexible APE model corrects the economic interpretation of key indicators and uncovers robust causal roles for the volatility in options-implied risk appetite and market liquidity—relationships obscured or misrepresented by the simpler linear model. By integrating these findings with intermediary asset pricing theories, we demonstrate how translating modern causal inference methods to a complex social science domain can yield new scientific insights.

1 Introduction

Understanding the forces that trigger stock market troughs is a problem of immense economic importance where causality inherently manifests. However, moving from pure prediction to credible causal inference in this domain presents a formidable challenge, exemplifying the translational gap between causal learning theory and applied science. The complex, non-linear, and high-dimensional nature of financial markets can easily lead to spurious conclusions if overly simplistic models are used. This paper directly confronts this challenge, asking: how can we best integrate causality with domain expertise and real-world scientific data to accelerate discovery in complex, high-stakes domains like finance?

The "credibility revolution" in econometrics, powered by tools like Double/Debiased Machine Learning (DML) from Chernozhukov et al. [2018], offers a path forward. DML provides a framework for obtaining statistically valid causal estimates even in the presence of high-dimensional confounding. While these methods are gaining traction in finance [Feng et al., 2020], their application to macro-finance questions remains nascent. Crucially, as recent research underscores, causal conclusions drawn from these sophisticated methods can themselves be highly sensitive to specification choices and the potential for unobserved confounding [Chernozhukov et al., 2022].

Accepted to the NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science.

This paper tackles this challenge directly through a novel, comparative causal framework, presenting a case study on achieving robust causal inference in financial economics. We first establish a baseline using DML for the canonical partially linear model (DML-PLR). Recognizing its limitations, we then implement an advanced DML framework to estimate the Average Partial Effect (APE). This comparison reveals that the more flexible APE model is essential for credible inference, correcting misinterpretations from the linear model and uncovering new causal pathways. By interpreting these robust findings through the lens of intermediary asset pricing theories [He and Krishnamurthy, 2013], we demonstrate a successful translation of causal ML from a black-box tool to an instrument for generating new scientific insights within our domain, aligning with the workshop's goal of fostering research that starts with concrete, real-world problems.

2 From Prediction to a Causal Target Variable

A credible causal analysis first requires a well-defined target variable and a robust predictive signal. We identify significant market troughs in the S&P 500 from 2013-2025 using a modified Bry and Boschan [1971] algorithm. A key methodological challenge is that identifying a trough is inherently retrospective, creating a data-leakage paradox. We resolve this by framing our objective as a nowcast: estimating in real-time the probability that the current period will eventually be identified as a trough, using only features available up to that day. This provides a timely signal ($\mathbf{y}_t = 1$ if day t is in a trough period, 0 otherwise) suitable for causal analysis.

To capture the state of the market, we engineer a high-dimensional feature set (>200 indicators) based on established factors from the financial economics literature, categorized into structural and sentiment indicators. Full details on the construction and rationale for these features are provided in Appendix A.4. A Support Vector Machine (SVM) classifier, with hyperparameters tuned via a strict forward-chaining time-series cross-validation, trained on these features demonstrates strong out-of-sample predictive power (ROC AUC of 0.89). The success of this predictive model is not the end goal, but a necessary prerequisite to ensure our subsequent causal analysis is based on a meaningful and well-specified relationship. As detailed in Appendix A.2, we conducted extensive testing to confirm this signal is stable and robust against common failure modes like covariate shift and concept drift.

3 A Comparative Causal Framework

While the SVM model is predictive, it does not establish causality. To do so, we employ a DML framework, which we implement in two stages to test the sensitivity of our conclusions to model assumptions. Our causal identification for both models relies on the unconfoundedness assumption, detailed further in Appendix A.1.

3.1 Baseline Model: DML for a Partially Linear Model (DML-PLR)

Our baseline causal model is the Partially Linear Regression (PLR) [Chernozhukov et al., 2018], a common benchmark:

$$\mathbf{Y} = \theta\mathbf{D} + g(\mathbf{X}) + \epsilon$$

Here, $\mathbf{Y}$ is the binary trough outcome, $\mathbf{D}$ is the treatment variable (a specific indicator), and $\mathbf{X}$ is a high-dimensional vector of all other features serving as confounders. DML provides a $\sqrt{N}$-consistent estimate for the constant treatment effect θ by using machine learning to flexibly model the nuisance functions $g(\mathbf{X})$ and $\mathbb{E}[\mathbf{D}|\mathbf{X}]$. In simple terms, this model assumes that a one-unit change in a treatment has the exact same impact on trough probability, regardless of whether the market is calm or in a full-blown panic. However, this model's core assumption—that the causal effect θ is constant and additively separable—is highly restrictive in financial markets where interactions are paramount.

3.2 Primary Model: DML for the Average Partial Effect (DML-APE)

To address the PLR's limitations, our primary analysis uses a more flexible DML estimator for an interactive model, which is better suited to a binary outcome:

$$P(\mathbf{Y} = 1|\mathbf{D} = d, \mathbf{X} = x) = l(d, x)$$

The causal parameter of interest is the Average Partial Effect (APE), θ_0, representing the average change in the trough probability for a one-unit increase in the treatment, averaged across the entire data distribution:

$$\theta_0 = \mathbb{E}_{\mathbf{D},\mathbf{X}} \left[\frac{\partial l(\mathbf{D}, \mathbf{X})}{\partial \mathbf{D}} \right]$$

Estimating the APE requires a Neyman-orthogonal score function, the derivation of which is detailed in Appendix A.6. The estimation process requires learning three nuisance functions (the outcome, treatment mean, and treatment variance models) using machine learning. A key practical challenge is that noise in these nuisance models can create extreme outliers in the score distribution; we address this by using the sample median as our point estimator—a robust approach justified in Appendix A.7—to prevent outliers from biasing our results. The resulting estimate for θ_0 is robust to first-order estimation errors and allows the treatment effect to vary non-linearly with $\mathbf{X}$. This θ_0 thus measures the average effect across all possible market conditions, allowing for the possibility that a change in a treatment has a huge effect during a panic but a negligible effect during a calm period.

3.3 Robustness and Identification

For both models, we implement a strict protocol to avoid multicollinearity and "bad controls" by excluding features that are mechanistic components of the treatment. Crucially, all statistically significant causal claims are subjected to a formal sensitivity analysis using the method of Cinelli and Hazlett [2020]. This quantifies how strong an unobserved confounder would need to be to invalidate our results, ensuring our findings are robust.

4 Empirical Results: The Importance of Model Specification

Our comparative causal analysis reveals that the choice of model specification is critical for drawing credible economic conclusions. The DML-APE model's ability to capture non-linear interactions uncovers a richer, more intuitive set of causal drivers than the restrictive DML-PLR baseline. Table 1 highlights four key insights from this comparison, while the complete sets of robust causal estimates for both models are provided in Appendix A.3. The full, unabridged results for all 200+ features are available in the online appendix, detailed in Appendix A.8.

First, a small set of core drivers are robust to either specification. For example, both models find that an upward trend in the Fed Funds futures slope (ffr_slope_scaled_trend), which signals market expectations of future monetary easing, has a statistically significant, negative causal impact on trough probability. This agreement suggests a powerful, unambiguous stabilizing force.

Second, the move to a more flexible specification helps discard potentially spurious findings. The DML-PLR model finds that higher volatility in credit spreads (credit_spread_scaled_std) is stabilizing. This counter-intuitive effect vanishes in the DML-APE model, suggesting the linear model's finding was an artifact of its failure to account for interactions with broader market volatility.

Third, and most critically, the DML-APE model identifies new causal pathways entirely missed by the linear model. It finds that the *volatility* of options-based risk appetite measures (e.g., Gamma Exposure (GEX), Volatility Risk Premium (VRP)) are robust causal drivers. This points to a more sophisticated mechanism where it is not just the level of fear, but its rate of change and persistence, that causally contributes to market capitulation.

Finally, the DML-APE model reverses the sign of several key estimates, resolving counter-intuitive results from the linear model. For instance, the PLR model finds that volatility in the trend of market illiquidity (amihud_illiquidity_trend_z_scaled_std) is stabilizing. The APE model reverses this, finding a robust positive effect. This correction aligns with economic theory: rising instability in market liquidity is a causal precursor to a trough.

5 Interpretation and Scientific Contribution

Our work demonstrates how causal ML, when integrated with domain expertise, can produce specific scientific discoveries in financial economics. The robust findings from our DML-APE model are not just a list of important features; they provide high-frequency empirical validation for modern

Table 1: Comparative DML Causal Estimates: PLR vs. APE Models

Theme	Treatment Variable (D)	Model	Coeff. ($\hat{\theta}$)	p-value	Robust?
Finding 1: Consistent Negative Effect of Easing Expectations					
Monetary Policy	ffr_slope_scaled_trend	PLR	-0.1436	0.0010	Yes
		APE	-0.0073	<0.0001	Yes
Finding 2: Effect of Credit Spread Volatility Lost Robustness					
Credit Conditions	credit_spread_scaled_std	PLR	-0.0524	<0.0001	Yes
		APE	-	-	No
Finding 3: New Volatility-Based Drivers Gained Robustness					
Options Risk Appetite	gex_oi_trend_z_scaled_std	PLR	-	-	No
		APE	0.0773	<0.0001	Yes
Volatility Risk Premium	vrp_roc63_scaled_std	PLR	-	-	No
		APE	-0.0021	0.0099	Yes
Finding 4: Causal Sign Reversal for Liquidity and Sentiment					
Market Liquidity	amihud_illiquidity_trend_z_scaled_std	PLR	-0.0608	0.0001	Yes
		APE	0.0160	<0.0001	Yes
Market Sentiment	pcr_oi_roc63_scaled_std	PLR	-0.0549	0.0057	Yes
		APE	0.0241	<0.0001	Yes

Notes: The comparison shows how moving to the flexible APE model is essential for credible inference, revealing findings that are consistent, lose robustness, gain robustness, or reverse in sign. The coefficient magnitudes between PLR and APE models are not directly comparable due to different parameter interpretations; the analysis focuses on sign and statistical significance

intermediary asset pricing theories [He and Krishnamurthy, 2013]. This framework posits that market stability is determined by the risk-bearing capacity of a specialized financial intermediary sector. A market trough represents a phase transition into a constrained, non-linear regime where this capacity is exhausted.

Our causal findings paint a clear, empirical picture of this theoretical state. The APE model's discovery that the *volatility* of options-implied risk measures causally drives troughs points to an erratic market price of risk, precisely as predicted when constrained intermediaries cannot smoothly absorb shocks. Similarly, the corrected, positive causal effect for illiquidity volatility is the empirical signature of intermediaries withdrawing from the market, triggering fire-sale dynamics. The stabilizing effect of monetary easing expectations fits perfectly, as the prospect of easier future funding conditions causally boosts intermediaries' risk-bearing capacity today.

By moving beyond linear assumptions, our causal analysis provides a more theoretically coherent account of how latent risks described by structural economic models manifest as observable market events. This confirms the value of translating flexible causal methods to social science, where they can correct misinterpretations from simpler models and provide a richer empirical validation of scientific theory. This integration of methods and domain knowledge directly embodies the bottom-up research paradigm advocated by the workshop.

6 Conclusion

This paper presents a successful translation of advanced causal machine learning to financial economics, addressing the workshop's central theme of bridging theory and scientific practice. Our primary contribution is a comparative causal analysis demonstrating that credible scientific conclusions about the drivers of market troughs are critically dependent on using flexible models that can capture non-linear interactions. The DML-APE model proved superior, correcting interpretations from a simpler linear model and identifying the causal role of volatility in risk appetite and liquidity. These findings provide novel, high-frequency empirical support for intermediary asset pricing theories. This work serves as a practical case study illustrating how the dual tools of prediction and causal inference can be combined to move from black-box models to robust scientific discovery in a complex, real-world domain.

Broader Impacts Statement

This research has the potential for positive societal impacts by contributing to a better understanding of financial stability, which could aid regulators and policymakers in developing more effective tools to mitigate market crises. However, potential negative impacts must also be considered. The causal factors identified could, in principle, be exploited to build predatory trading algorithms designed to profit from or even exacerbate market instability. We believe this risk is partially mitigated by the "nowcasting" nature of our target variable, which serves more as a real-time risk indicator than a simple, forward-looking predictive signal for automated trading.

References

Yakov Amihud. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56, jan 2002. doi: 10.1016/s1386-4181(01)00024-6.

Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Paul Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625, mar 2003. doi: 10.1111/1468-0262.00418.

Gurdip Bakshi, Nikunj Kapadia, and Dilip Madan. Stock return characteristics, skew laws, and the differential pricing of individual option contracts. *The Review of Financial Studies*, 16(1):101–143, mar 2003. doi: 10.1093/rfs/16.1.0101.

Randall S. Billingsley and Don M. Chance. Put–call ratios and market timing effectiveness. *The Journal of Portfolio Management*, 15(1):25–28, fall 1988. doi: 10.3905/jpm.1988.409184.

Tim Bollerslev, George Tauchen, and Hao Zhou. Expected stock returns and variance risk premia. *The Review of Financial Studies*, 22(11):4463–4492, nov 2009. doi: 10.1093/rfs/hhp008.

Gerhard Bry and Charlotte Boschan. *Cyclical Analysis of Time Series: Selected Procedures and Computer Programs*. National Bureau of Economic Research, New York, 1971. ISBN 0870142232.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, feb 2018. doi: 10.1111/ectj.12097.

Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Working Paper 30058, National Bureau of Economic Research, may 2022.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, feb 2020. doi: 10.1111/rssb.12348.

Eugene F. Fama and Kenneth R. French. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, 25(1):23–49, nov 1989. doi: 10.1016/0304-405x(89)90095-0.

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370, jun 2020. doi: 10.1111/jofi.12883.

Zhiguo He and Arvind Krishnamurthy. Intermediary asset pricing. *The American Economic Review*, 103(2):732–770, apr 2013. doi: 10.1257/aer.103.2.732.

Jun Pan and Allen M. Poteshman. The information in option volume for future stock prices. *The Review of Financial Studies*, 19(3):871–908, fall 2006. doi: 10.1093/rfs/hhj024.

Robert E. Whaley. The investor fear gauge. *The Journal of Portfolio Management*, 26(3):12–17, spring 2000. doi: 10.3905/jpm.2000.319728.

A Appendix: Supplementary Materials

A.1 Causal Identification and Estimation Details

Our causal analysis rests on a standard identification strategy and a robust, data-driven estimation procedure for the nuisance functions required by the Double/Debiased Machine Learning (DML) framework.

- **Identification Assumption**: Our causal identification relies on the assumption of unconfoundedness (also known as conditional ignorability). This assumption posits that, conditional on the high-dimensional set of covariates $\mathbf{X}$ (which includes over 200 engineered features capturing market structure, dealer positioning, sentiment, and macroeconomic conditions), the assignment of the treatment variable $\mathbf{D}$ is independent of the potential outcomes for a market trough. Formally, $\mathbf{Y}(d) \perp \mathbf{D}|\mathbf{X}$ for all values d in the support of $\mathbf{D}$. While this assumption is untestable, its plausibility is strengthened by the comprehensive nature of our feature set, which is designed to control for a wide range of potential confounding factors.
- **Time-Series Cross-Fitting and Nuisance Learners:** A critical concern with time-series data is preventing data leakage. Our DML procedure explicitly avoids this by using a forward-chaining (or 'rolling-origin') cross-fitting scheme (implemented via 'sklearn.model_selection.TimeSeriesSplit'). For K folds, each fold (k) trains on all data from time 1 to T_k and generates out-of-sample (OOS) predictions for the subsequent block of data from $T_k + 1$ to T_{k+1}. This strictly preserves the temporal order and ensures that nuisance models are only ever fit on past data to generate predictions for the "future" validation fold.
- **Within-Fold Horse Race:** The "horse-race" mentioned in the main text is conducted on the OOS predictions generated by this time-aware procedure. For each nuisance function (e.g., $\mathbb{E}[\mathbf{D}|\mathbf{X}]$), we trained multiple learners ('GradientBoostingRegressor' and 'LassoCV') on the training portion of each fold. We then computed the out-of-sample R^2 for each learner on the validation portion. The learner with the superior OOS R^2 across all folds was dynamically selected, and its OOS predictions were used to construct the final Neyman-orthogonal score. This ensures our nuisance model selection is both data-driven and robust against look-ahead bias.
- **Inference for Correlated Data:** We acknowledge that while our use of a median estimator (Appendix A.7) and a standard non-parametric bootstrap provides robustness to outliers in the score distribution, it does not explicitly account for potential autocorrelation in the scores themselves. The suggestion to verify inference coverage using time-series-specific methods, such as a block-bootstrap or simulations with autocorrelated pseudo-outcomes, is a valuable direction for future research to further strengthen the statistical validity of the confidence intervals.

A.2 Predictive Model Robustness and Stability Analysis

A critical challenge for any predictive model in finance is structural breaks. To validate the resilience of our primary SVM prediction model, we conducted a series of diagnostic tests on the hold-out sample (July 2023 - June 2025) to detect common failure modes, namely performance degradation, covariate shift, and concept drift.

- Model Performance Stability: We calculated the model's Brier score over a 63-day rolling window to track its accuracy and calibration over time. As shown in Figure 1, the model is highly stable. The rolling Brier score remains exceptionally low for the vast majority of the test period. The score exhibits brief, sharp spikes that correctly coincide with the actual trough events, and crucially, it quickly reverts to its low baseline afterward. This demonstrates that the model's performance does not persistently degrade after a crisis event.
- Covariate Shift Analysis: We tested for covariate shift by comparing the distributions of our most important input features between the training and testing periods. Figure 2 shows that the distributions exhibit a high degree of overlap. This absence of significant covariate shift provides strong evidence that the statistical properties of the key predictors did not fundamentally change in the hold-out period, enhancing the credibility of the model's test set performance.

- Concept Drift Analysis: We tested for concept drift, where the relationship between features and the outcome changes, by analyzing the stability of SHAP feature importance over time. We split the hold-out test set chronologically and generated SHAP plots independently for each half. As shown in Figure 3, the feature importance rankings are highly consistent across both periods. This stability is strong evidence that the underlying economic relationships the model learned remained valid throughout the test period, confirming its robustness against concept drift.

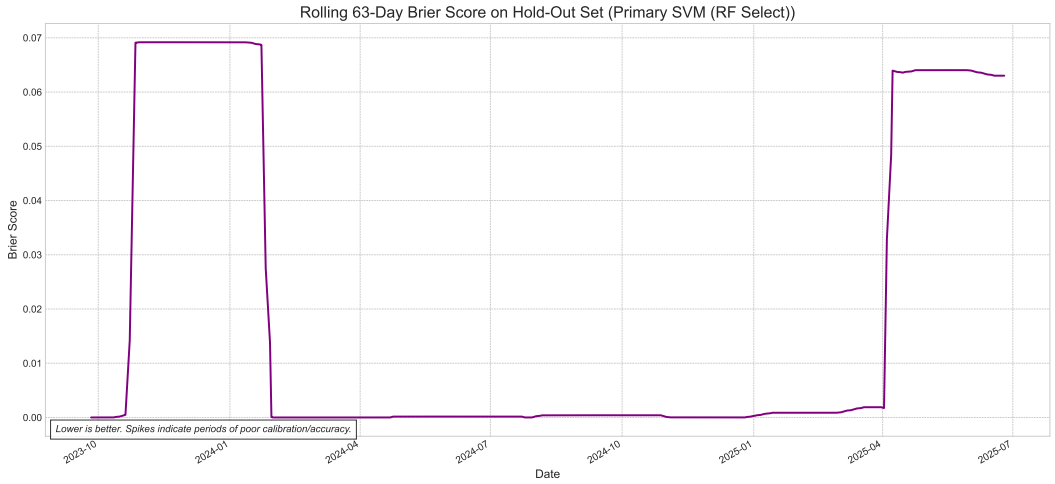


Figure 1: Model Performance Stability on the Hold-Out Test Set. The figure plots the Brier score of the primary SVM model's calibrated probability forecasts, calculated over a 63-day rolling window. The rapid return to a near-zero baseline following trough events (spikes) demonstrates performance stability.

A.3 Full DML Estimation and Sensitivity Analysis Results

This section contains the complete set of treatment variables for which the DML analysis yielded a statistically significant causal estimate ($p < 0.05$) that was also robust to the formal sensitivity analysis of Cinelli and Hazlett [2020]. Table 2 lists the robust findings from the baseline DML-PLR model. Table 3 lists the robust findings from our primary DML-APE model.

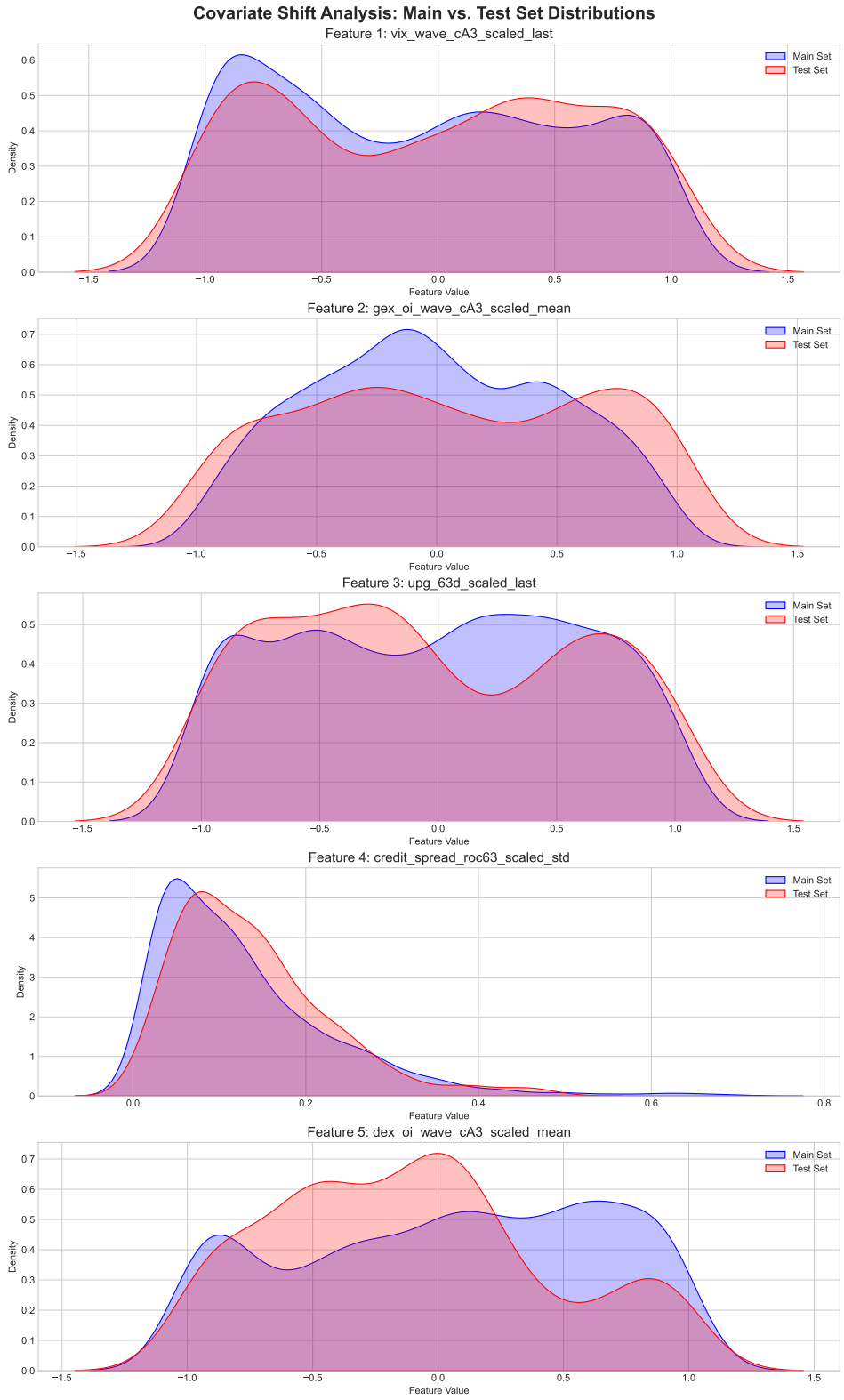
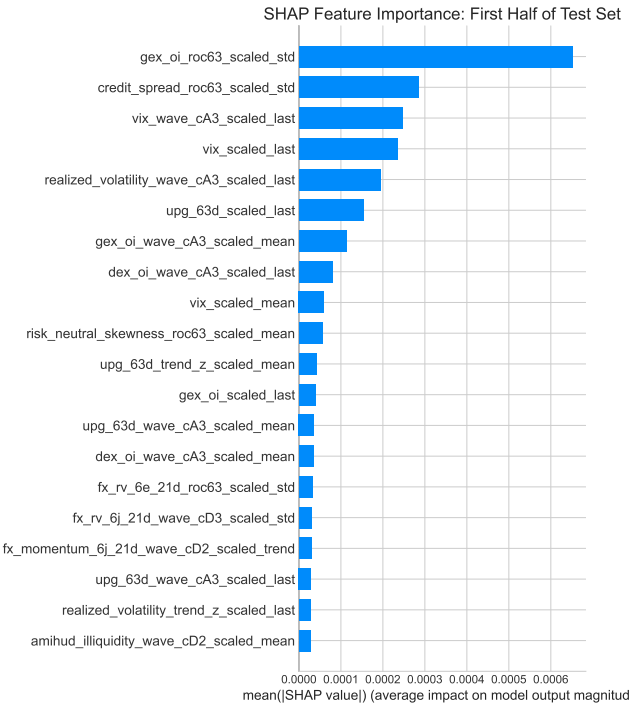
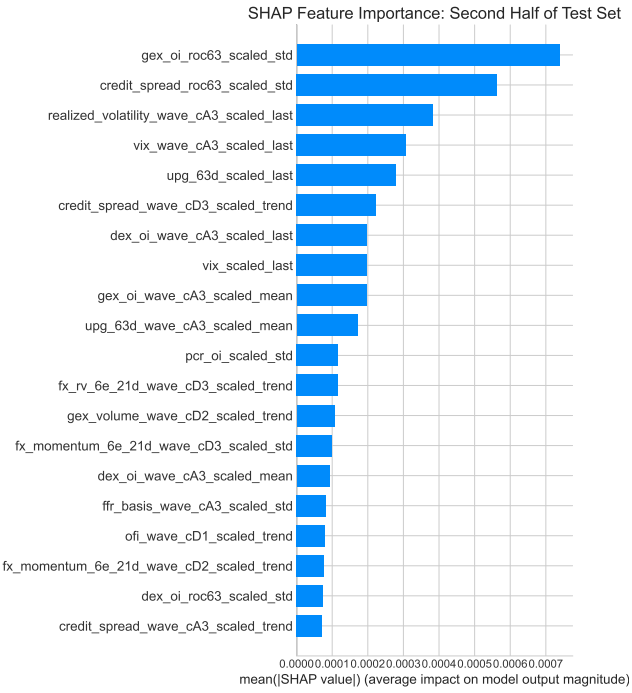


Figure 2: Covariate Shift Analysis for Top Predictive Features. The figure compares kernel density estimates (KDEs) for key features between the Main Set (training/validation, blue) and the Test Set (hold-out, orange). The high degree of overlap suggests the absence of significant covariate shift.



(a) First half of test set.



(b) Second half of test set.

Figure 3: Stability of SHAP Feature Importance on the Hold-Out Test Set. The high degree of consistency in feature rankings and their relative magnitudes between the two periods indicates that the model's learned relationships are stable and robust against concept drift.

Table 2: Complete Robust Causal Estimates from the DML-PLR Model

Treatment Variable	**Coeff.** ($\hat{\theta}$)	p**-value**	**bias_phi**	**Adj. 95% CI Lower**	**Adj. 95% CI Upper**	**Benchmark** R_Y^2	**Benchmark** R_D^2
ffr_slope_trend_z_scaled_last	-0.0157	0.0000	0.0000	-0.0216	-0.0098	0.0520	0.0000
ffr_slope_roc63_scaled_last	-0.0228	0.0000	0.0048	-0.0368	-0.0087	0.0518	0.0159
credit_spread_scaled_std	-0.0524	0.0000	0.0000	-0.0743	-0.0306	0.0433	0.0000
credit_spread_trend_z_scaled_std	-0.0522	0.0000	0.0000	-0.0747	-0.0297	0.0424	0.0000
fx_rv_6j_21d_trend_z_scaled_mean	-0.0143	0.0000	0.0000	-0.0207	-0.0078	0.0664	0.0000
ffr_slope_roc63_scaled_mean	-0.0122	0.0001	0.0000	-0.0181	-0.0062	0.0520	0.0000
ffr_slope_scaled_last	-0.0154	0.0001	0.0000	-0.0230	-0.0079	0.0518	0.0000
amihud_illiquidity_trend_z_scaled_std	-0.0608	0.0001	0.0000	-0.0917	-0.0299	0.0569	0.0000
ffr_slope_scaled_mean	-0.0145	0.0003	0.0000	-0.0224	-0.0066	0.0520	0.0000
fx_rv_6j_21d_trend_z_scaled_last	-0.0153	0.0005	0.0000	-0.0238	-0.0067	0.0656	0.0000
ffr_slope_scaled_trend	-0.1436	0.0010	0.0000	-0.2293	-0.0579	0.0515	0.0000
fx_rv_6j_21d_scaled_mean	-0.0128	0.0025	0.0038	-0.0248	-0.0007	0.0656	0.0069
pcr_oi_roc63_scaled_std	-0.0549	0.0057	0.0000	-0.0938	-0.0160	0.0932	0.0000
risk_neutral_skewness_scaled_trend	-0.1346	0.0069	0.0000	-0.2321	-0.0370	0.0448	0.0000
risk_neutral_skewness_trend_z_scaled_trend	-0.1533	0.0082	0.0000	-0.2669	-0.0397	0.0446	0.0000
flow_concentration_10d_scaled_std	0.0636	0.0124	0.0000	0.0138	0.1134	0.0533	0.0000
fx_rv_6j_21d_scaled_last	-0.0096	0.0135	0.0000	-0.0172	-0.0020	0.0682	0.0000
ffr_basis_roc63_scaled_mean	0.0111	0.0177	0.0000	0.0019	0.0202	0.0520	0.0000
risk_neutral_kurtosis_trend_z_scaled_mean	0.0186	0.0196	0.0000	0.0030	0.0342	0.0448	0.0000
flow_concentration_10d_trend_z_scaled_mean	-0.0093	0.0202	0.0000	-0.0172	-0.0015	0.0523	0.0000
flow_concentration_10d_trend_z_scaled_std	0.0404	0.0219	0.0000	0.0059	0.0750	0.0546	0.0000
flow_concentration_10d_roc63_scaled_std	0.0515	0.0228	0.0000	0.0072	0.0958	0.0594	0.0000
ffr_basis_roc63_scaled_last	0.0081	0.0260	0.0000	0.0010	0.0152	0.0526	0.0000
ffr_slope_trend_z_scaled_trend	-0.0834	0.0306	0.0000	-0.1590	-0.0078	0.0520	0.0000
risk_neutral_kurtosis_scaled_mean	0.0173	0.0306	0.0000	0.0016	0.0329	0.0448	0.0000
risk_neutral_skewness_roc63_scaled_trend	-0.0946	0.0359	0.0000	-0.1829	-0.0062	0.0465	0.0000
pcr_oi_trend_z_scaled_std	-0.0337	0.0420	0.0000	-0.0662	-0.0012	0.1126	0.0000

Notes: This table reports the set of statistically significant ($p < 0.05$) causal estimates from the DML-PLR model that are robust to unobserved confounding. Robustness is assessed using the formal sensitivity analysis of Cinelli and Hazlett [2020]. Benchmark R_Y^2 and Benchmark R_D^2 report the out-of-sample partial R^2 of the outcome and the treatment explained by the observed confounders, respectively. These values serve as a benchmark for the plausible strength of an unobserved confounder. The results are deemed robust if the adjusted 95% confidence interval, which accounts for potential bias from a hypothetical confounder as strong as the observed ones, still excludes zero. These results are a subset of the full analysis; the complete results for all features can be found in the online appendix referenced in Appendix A.8.

Table 3: Complete Robust Causal Estimates from the DML-APE Model

Treatment Variable	**Coeff.** ($\hat{\theta}$)	p**-value**	**bias_phi**	**Adj. 95% CI Lower**	**Adj. 95% CI Upper**	**Benchmark** R_Y^2	**Benchmark** R_D^2
fx_rv_6j_21d_roc63_scaled_std	0.0057	0.0000	0.0000	0.0046	0.0069	0.1085	0.0000
risk_neutral_kurtosis_trend_z_scaled_std	0.0485	0.0000	0.0000	0.0413	0.0557	0.0366	0.0000
fx_rv_6j_21d_trend_z_scaled_std	0.0072	0.0000	0.0000	0.0061	0.0083	0.1085	0.0000
risk_neutral_kurtosis_roc63_scaled_std	0.0382	0.0000	0.0000	0.0339	0.0425	0.0366	0.0000
ffr_slope_trend_z_scaled_std	0.0092	0.0000	0.0000	0.0083	0.0100	0.0427	0.0000
fx_rv_6e_21d_scaled_std	0.0100	0.0000	0.0000	0.0089	0.0110	0.0913	0.0000
amihud_illiquidity_trend_z_scaled_std	0.0160	0.0000	0.0000	0.0129	0.0191	0.0375	0.0000
risk_neutral_skewness_trend_z_scaled_std	0.0343	0.0000	0.0000	0.0301	0.0384	0.0366	0.0000
risk_neutral_skewness_scaled_std	0.0359	0.0000	0.0000	0.0320	0.0397	0.0366	0.0000
risk_neutral_skewness_roc63_scaled_std	0.0399	0.0000	0.0000	0.0320	0.0479	0.0366	0.0000
pcr_oi_trend_z_scaled_std	0.0213	0.0000	0.0000	0.0196	0.0231	0.1064	0.0000
fx_rv_6e_21d_roc63_scaled_mean	-0.0016	0.0000	0.0000	-0.0020	-0.0012	0.0913	0.0000
ffr_slope_roc63_scaled_std	0.0063	0.0000	0.0000	0.0057	0.0070	0.0427	0.0000
pcr_oi_scaled_std	0.0191	0.0000	0.0000	0.0178	0.0203	0.1064	0.0000
fx_rv_6e_21d_roc63_scaled_std	0.0061	0.0000	0.0000	0.0048	0.0074	0.0913	0.0000
pcr_oi_roc63_scaled_std	0.0241	0.0000	0.0000	0.0214	0.0268	0.1064	0.0000
ffr_slope_scaled_std	0.0103	0.0000	0.0000	0.0099	0.0107	0.0427	0.0000
ffr_slope_scaled_last	0.0177	0.0000	0.0000	0.0146	0.0209	0.0427	0.0000
fx_rv_6j_21d_scaled_std	0.0051	0.0000	0.0000	0.0039	0.0063	0.1085	0.0000
risk_neutral_kurtosis_scaled_std	0.0957	0.0000	0.0000	0.0726	0.1189	0.0366	0.0000
fx_momentum_6e_21d_trend_z_scaled_std	0.0047	0.0000	0.0000	0.0035	0.0058	0.0913	0.0000
ffr_slope_scaled_trend	-0.0073	0.0000	0.0000	-0.0092	-0.0054	0.0427	0.0000
gex_oi_trend_z_scaled_std	0.0773	0.0000	0.0186	0.0381	0.1165	0.0331	0.0391
fx_momentum_6e_21d_scaled_std	0.0038	0.0000	0.0000	0.0028	0.0049	0.0913	0.0000
fx_rv_6j_21d_trend_z_scaled_last	-0.0018	0.0000	0.0000	-0.0023	-0.0012	0.1085	0.0000
fx_momentum_6j_21d_trend_z_scaled_std	0.0026	0.0000	0.0000	0.0018	0.0035	0.1085	0.0000
fx_rv_6e_21d_roc63_scaled_last	-0.0012	0.0000	0.0000	-0.0017	-0.0008	0.0913	0.0000
ffr_slope_trend_z_scaled_trend	-0.0066	0.0000	0.0000	-0.0090	-0.0042	0.0427	0.0000
fx_rv_6e_21d_trend_z_scaled_mean	-0.0012	0.0000	0.0000	-0.0017	-0.0008	0.0913	0.0000
fx_rv_6e_21d_trend_z_scaled_last	-0.0009	0.0000	0.0000	-0.0013	-0.0006	0.0913	0.0000
fx_rv_6j_21d_trend_z_scaled_trend	0.0037	0.0000	0.0000	0.0020	0.0054	0.1085	0.0000
fx_rv_6j_21d_trend_z_scaled_mean	-0.0017	0.0001	0.0000	-0.0025	-0.0009	0.1085	0.0000
ffr_slope_roc63_scaled_last	-0.0006	0.0001	0.0000	-0.0009	-0.0003	0.0427	0.0000
fx_rv_6e_21d_scaled_mean	-0.0009	0.0003	0.0000	-0.0014	-0.0004	0.0913	0.0000
flow_concentration_10d_trend_z_scaled_mean	-0.0013	0.0003	0.0000	-0.0021	-0.0006	0.0424	0.0000
risk_neutral_kurtosis_scaled_mean	0.0051	0.0008	0.0000	0.0021	0.0081	0.0366	0.0000
fx_rv_6e_21d_trend_z_scaled_trend	0.0024	0.0009	0.0000	0.0010	0.0038	0.0913	0.0000
flow_concentration_10d_scaled_std	0.0021	0.0013	0.0000	0.0008	0.0034	0.0424	0.0000
ffr_basis_roc63_scaled_trend	-0.0020	0.0034	0.0000	-0.0033	-0.0006	0.0427	0.0000
ffr_slope_scaled_mean	0.0011	0.0085	0.0000	0.0003	0.0019	0.0427	0.0000
ffr_basis_scaled_last	0.0007	0.0096	0.0000	0.0002	0.0012	0.0427	0.0000
vrp_roc63_scaled_std	-0.0021	0.0099	0.0000	-0.0036	-0.0005	0.0367	0.0000
flow_concentration_10d_roc63_scaled_std	0.0018	0.0114	0.0000	0.0004	0.0031	0.0424	0.0000
risk_neutral_kurtosis_trend_z_scaled_trend	-0.0060	0.0233	0.0000	-0.0111	-0.0008	0.0366	0.0000
ffr_basis_roc63_scaled_last	-0.0003	0.0299	0.0000	-0.0006	0.0000	0.0427	0.0000
risk_neutral_kurtosis_scaled_trend	-0.0043	0.0313	0.0000	-0.0082	-0.0004	0.0366	0.0000
ffr_basis_scaled_mean	0.0006	0.0345	0.0000	0.0000	0.0011	0.0427	0.0000
flow_concentration_10d_trend_z_scaled_std	0.0011	0.0360	0.0000	0.0001	0.0021	0.0424	0.0000

Notes: This table reports the set of statistically significant ($p < 0.05$) causal estimates from the DML-APE model that are robust to unobserved confounding. The coefficient ($\hat{\theta}$) is the Average Partial Effect (APE). The point estimate is the median of the Neyman-orthogonal scores, and the 95% confidence intervals and p-values are derived from a non-parametric bootstrap of these scores. Benchmark R_Y^2 and Benchmark R_D^2 report the out-of-sample partial R^2 of the outcome and the treatment explained by the observed confounders, respectively. These values serve as a benchmark for the plausible strength of an unobserved confounder. The results are deemed robust if the adjusted 95% confidence interval, which accounts for potential bias from a hypothetical confounder as strong as the observed ones, still excludes zero. The complete results for all features can be found in the online appendix referenced in Appendix A.8.

A.4 Feature Engineering and Descriptive Statistics

This section provides details on the key indicators engineered for the predictive and causal models, along with their descriptive statistics. We categorize indicators into physical/structural and psychological/sentiment groups. Tables 4 and 5 detail the construction and economic rationale for key parent indicators. Table 6 provides summary statistics for these parent indicators, revealing the high persistence and non-normality that motivate the use of non-parametric scaling and non-linear models.

Table 4: Physical/Structural Indicators ($\mathbf{z}_t$) and Economic Rationale

Name	Mathematical Definition	Economic Intuition	Reference(s)
GEX (OI)	$\sum_i(\Gamma_{C,i} \cdot \mathbf{OI}_{C,i} - \Gamma_{P,i} \cdot \mathbf{OI}_{P,i}) \times 100$	Measures dealer gamma exposure from open positions. High positive GEX may suppress volatility, while low or negative GEX can amplify it. Capitulation troughs often occur in negative gamma regimes.	SqueezeMetrics
GEX (Volume)	$\sum_i(\Gamma_{C,i} \cdot \mathbf{V}_{C,i} - \Gamma_{P,i} \cdot \mathbf{V}_{P,i}) \times 100$	Measures dealer gamma exposure from the day's trading volume, capturing intraday hedging pressures.	
Delta Exposure	$\sum_i(\Delta_{C,i} \cdot \mathbf{OI}_{C,i} + \Delta_{P,i} \cdot \mathbf{OI}_{P,i}) \times 100$	Measures net market delta positioning. Extremely low or negative values indicate bearish positioning and potential for short covering, often seen near troughs.	SqueezeMetrics
Credit Spread	$\mathbf{Yld}_{HY} - \mathbf{Yld}_{RF}$	The premium for bearing credit risk. A widening spread signals deteriorating economic conditions and heightened risk aversion, which peaks near market troughs.	Fama and French [1989]
Amihud Illiquidity	$\frac{\|\mathbf{R}_{daily}\|}{\mathbf{V}_{\$,\text{daily}}}$	Measures price impact. High values indicate illiquidity, as small volumes cause large price changes. Liquidity often vanishes near troughs.	Amihud [2002]
FFR Slope	$\mathbf{P}_{C1} - \mathbf{P}_{C3}$	Spread between 1st and 3rd Fed Funds futures. A steepening (more positive slope) can signal expectations of easier future policy, often a response to market stress.	

Table 5: Psychological/Sentiment Indicators ($\mathbf{u}_t$) and Economic Rationale

Name	Mathematical Definition	Economic Intuition	Reference(s)
Realized Volatility	$\sqrt{252 \cdot (\sum_{i=1}^{M-1} r_{i,intra}^2 + r_{overnight}^2)}$	Historical volatility from high-frequency data. Spikes in RV indicate panic and forced liquidation, which characterize market bottoms.	Andersen et al. [2003]
VIX	CBOE VIX Index methodology	Market's expectation of 30-day implied volatility. High VIX signals fear and demand for portfolio insurance, peaking at market troughs.	Whaley [2000]
Volatility Risk Premium	$\mathbf{VIX}_t - \mathbf{RV}_t$	The premium investors pay for protection against volatility. A negative VRP (realized > implied) often signals panic and deleveraging, a common feature of troughs.	Bollerslev et al. [2009]
PCR (OI)	$\frac{\sum \text{Put } \mathbf{OI}}{\sum \text{Call } \mathbf{OI}}$	Ratio of open put to call contracts. High values indicate extreme bearish sentiment and hedging, which often precedes a market reversal.	Billingsley and Chance [1988]
PCR (Volume)	$\frac{\sum \text{Put } \mathbf{Volume}}{\sum \text{Call } \mathbf{Volume}}$	Ratio of traded put to call volume. Spikes indicate intense intraday fear and panic buying of puts, characteristic of capitulation lows.	Pan and Poteshman [2006]
RN Skewness	$\mathbb{E}_Q[(\frac{K-\mu_K}{\sigma_K})^3]$	Third moment of the risk-neutral distribution. Highly negative skew indicates high demand for OTM puts (crash protection), which is most pronounced at bottoms.	Bakshi et al. [2003]
RN Kurtosis	$\mathbb{E}_Q[(\frac{K-\mu_K}{\sigma_K})^4]$	Fourth moment of the risk-neutral distribution. High kurtosis ("fat tails") indicates the market is pricing in a high probability of extreme moves.	Bakshi et al. [2003]

Table 6: Descriptive Statistics for Parent Indicators (2013-2025)

Indicator	Mean	Std. Dev.	Skewness	Kurtosis	Min	Max	$\rho(1)$
Panel A: Physical/Structural							
gex_oi	6.65e+04	2.21e+06	41.417	1790.665	-8.14e+06	1.03e+08	0.682
credit_spread	0.049	0.016	0.258	0.024	0.014	0.114	0.998
amihud_illiquidity	9.94e-12	3.47e-11	0.000	0.000	0.000	1.17e-09	-0.049
ffr_slope	0.066	0.237	1.807	6.425	-0.615	1.203	0.995
Panel B: Psychological/Sentiment							
RV	12.700	9.840	4.105	32.356	0.758	133.842	0.669
VIX	17.812	6.942	2.730	13.973	9.140	82.690	0.970
VRP	3.336	5.042	-3.574	30.486	-58.725	16.729	0.662
PCR_OI	1.819	0.185	0.178	-0.554	1.389	2.489	0.987

Notes: This table reports summary statistics for the untransformed "parent" indicators. The final column, $\rho(1)$, reports the first-order autocorrelation coefficient.

A.5 Time-Series Properties and Stationarity

A valid concern for financial data is the high persistence in many indicators, as evidenced by the high $\rho(1)$ values in Table 6. The DML framework's i.i.d. assumption is a potential concern for such time-series.

Our feature engineering methodology—which aggregates features over a rolling lookback window (e.g., calculating _std, _trend)—is explicitly designed to mitigate this by transforming non-stationary parent series into stationary features. To validate this, we performed Augmented Dickey-Fuller (ADF) tests on both the raw parent indicators and the final set of aggregated features used as inputs for our models. The results, summarized in Table 7, confirm the effectiveness of this approach. While a portion of the parent indicators are non-stationary, 100% of the aggregated features used in our analysis are stationary ($p < 0.05$), making them far more appropriate inputs for the DML estimation and better satisfying its underlying statistical assumptions.

Table 7: Augmented Dickey-Fuller (ADF) Stationarity Test Results

Feature Set	**Features Tested**	**% Stationary ($p < 0.05$)**
Parent Indicators	147	90.62%
Aggregated Features (Model Inputs)	588	100.00%

Notes: The table shows the percentage of features in each set that are stationary according to the ADF test. The aggregation process (calculating mean, std, trend, last) successfully transforms the persistent parent indicators into a 100% stationary feature set for the model.

A.6 Derivation of the Neyman-Orthogonal Score for the APE

This appendix outlines the Neyman-orthogonal score function used for the estimation of the Average Partial Effect (APE), following the double/debiased machine learning framework of Chernozhukov et al. [2018].

A.6.1 Model Setup and Parameter of Interest

We consider a structural model where the outcome $\mathbf{Y}$ is determined by a continuous treatment $\mathbf{D}$ and confounders $\mathbf{X}$ through a general function $l_0(\mathbf{D}, \mathbf{X}) = \mathbb{E}[\mathbf{Y}|\mathbf{D}, \mathbf{X}]$. The analysis rests on the standard unconfoundedness assumption.

The causal parameter of interest is the Average Partial Effect (APE), θ_0:

$$\theta_0 = \mathbb{E}\left[\frac{\partial l_0(\mathbf{D}, \mathbf{X})}{\partial \mathbf{D}}\right]$$

The expectation is taken over the joint distribution of $(\mathbf{D}, \mathbf{X})$.

A.6.2 The Neyman-Orthogonal Score

To obtain a robust, $\sqrt{n}$-consistent estimate of θ_0, we rely on a Neyman-orthogonal score function. This property ensures that first-order estimation errors in the nuisance functions do not bias the final estimate of θ_0. The general score function for the APE is:

$$\psi(\mathbf{W}; \theta, \eta) = \frac{\partial l(\mathbf{D}, \mathbf{X})}{\partial \mathbf{D}} - \theta - \frac{1}{p(\mathbf{D}|\mathbf{X})}\frac{\partial p(\mathbf{D}|\mathbf{X})}{\partial \mathbf{D}}\left(\mathbf{Y} - l(\mathbf{D}, \mathbf{X})\right)$$

where $\mathbf{W} = (\mathbf{Y}, \mathbf{D}, \mathbf{X})$ and $\eta = (l, p)$ is the set of nuisance functions, including the conditional density of the treatment $p(\mathbf{D}|\mathbf{X})$.

A.6.3 A Practical Score via a Semi-Parametric Assumption

Directly estimating the conditional density and its derivative is challenging. Following standard practice, we adopt a flexible semi-parametric assumption that the treatment is conditionally Gaussian:

$$\mathbf{D}|\mathbf{X} \sim \mathcal{N}(m_0(\mathbf{X}), v_0(\mathbf{X}))$$

Under this assumption, the general score simplifies to the practical form we implement, which only requires estimating the conditional mean $m(\mathbf{X})$ and variance $v(\mathbf{X})$ as nuisance functions:

$$\psi(\mathbf{W};\theta,\eta) = \frac{\partial l(\mathbf{D},\mathbf{X})}{\partial \mathbf{D}} - \theta + \frac{\mathbf{D} - m(\mathbf{X})}{v(\mathbf{X})}\left(\mathbf{Y} - l(\mathbf{D},\mathbf{X})\right)$$

where the nuisance functions are now $\eta = (l, m, v)$. A formal proof of Neyman-orthogonality involves showing that the Gateaux derivative of the expected score with respect to the nuisance paths is zero at the true values, a standard result in this literature.

A.7 Robustness of the Median Estimator for the APE

A key practical challenge in implementing the APE score function is its sensitivity to estimation noise in the nuisance models. The bias correction term is inversely proportional to the estimated conditional variance of the treatment, $\hat{v}(\mathbf{X})$. When the ML model predicts a value of $\hat{v}(\mathbf{X})$ that is close to zero for some observations, this term can generate extreme outliers in the distribution of the estimated scores, $\hat{\psi}_i$.

In such cases, the sample mean becomes an unreliable estimator of the distribution's central tendency. Figure 4 illustrates this problem for one of our treatment variables. The distribution is clearly heavy-tailed and asymmetric. A naive interpretation of the sample mean would be misleading.

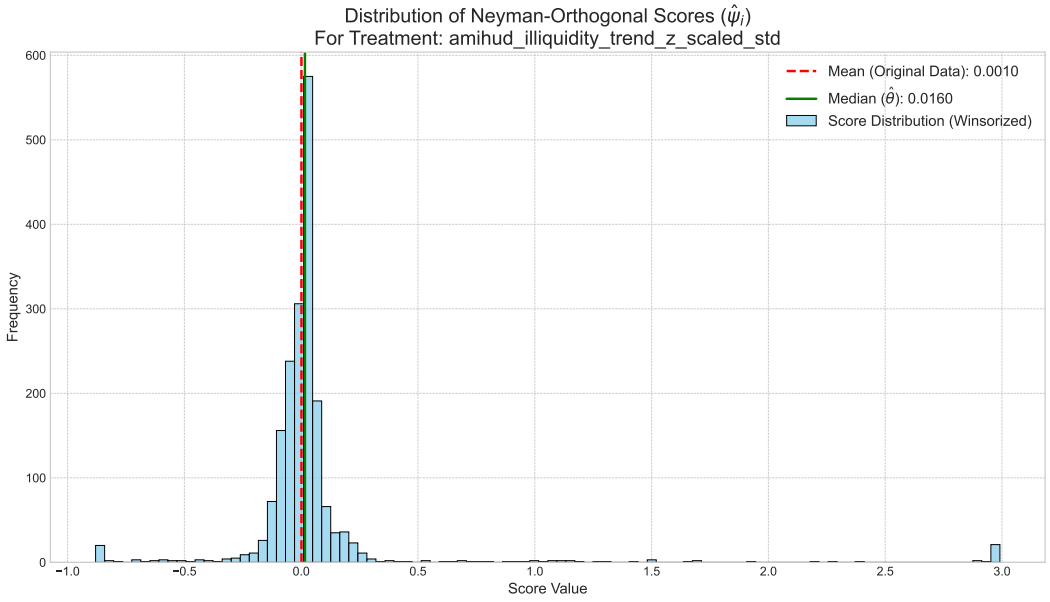


Figure 4: Distribution of Neyman-Orthogonal Scores for the APE. The histogram shows the calculated scores ($\hat{\psi}_i$) for the treatment amihud_illiquidity_trend_z_scaled_std. The red dashed line indicates the sample mean (0.0010), while the green solid line marks the sample median (0.0160). The mean is pulled toward zero by outliers, whereas the median robustly captures the positive central tendency.

To overcome this, we employ the sample median of the scores as our robust point estimator for θ_0. The median is insensitive to the magnitude of extreme outliers in the tails. Furthermore, since analytical formulas for the standard error are also unreliable for heavy-tailed distributions, we use a non-parametric bootstrap on the calculated scores to construct robust confidence intervals and p-values. This median-of-scores and bootstrap inference framework ensures our causal estimates are resilient to the noisy outputs of the nuisance models, reducing the risk of reporting spurious null findings.

A.8 Online Appendix with Complete Results

The complete, unabridged results for the DML-PLR and DML-APE estimation and sensitivity analyses, covering all 200+ features, are provided in a supplementary online appendix. github.com/jackraorpl/market-trough-prediction-appendix.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim that model specification is critical for causal inference in this domain and that a flexible APE model corrects and refines the conclusions from a simpler linear model. Sections 4 and 5 directly support this claim through a comparative analysis and by linking the APE model's findings to established economic theory.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly states in Appendix A.1 that the core identification relies on the untestable unconfoundedness assumption. We acknowledge that while our high-dimensional feature set makes this assumption more plausible, the potential for unobserved confounders can never be fully eliminated.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Appendix A.1 details the model setup and the core unconfoundedness assumption. Appendix A.5 presents the derivation of the practical score function, explicitly stating the semi-parametric assumption of a Gaussian conditional distribution for the treatment variable, which is required for its implementation.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides a comprehensive description of the methodology. Appendix A.1 details the causal identification and nuisance learner selection, Appendix A.4 provides definitions for all parent indicators, and Section 3 describes the DML estimation procedure, including the use of cross-fitting and the specific causal models being compared.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The financial data used in this study is proprietary and subject to licensing restrictions, preventing its public release. The code will not be released at this time to maintain alignment with the blind review process.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix A.1 details our data-driven approach for selecting nuisance learners ('GradientBoostingRegressor' vs 'LassoCV') within each cross-fitting fold. Appendix A.2 describes the hold-out test set (July 2023 - June 2025) used for validating the predictive model's stability.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: All causal estimates reported in Table 1, as well as the full results in Tables 2 and 3, are presented with p-values and 95% confidence intervals. The caption for Table 3 clarifies that the inference for the APE model is based on a non-parametric bootstrap of the Neyman-orthogonal scores.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The computational resources required for this analysis are not substantial. The entire DML pipeline for all treatment variables can be executed in under an hour on a standard multi-core CPU workstation, and no specialized hardware like GPUs was required.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses anonymized market data and does not involve human subjects. It adheres to all aspects of the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper has a "Broader Impacts Statement" section after the conclusion. For positive impacts, this research could lead to a better understanding of financial stability, potentially aiding regulators and policymakers in creating more effective tools to mitigate market crises. For negative impacts, the causal factors identified could potentially be exploited to build predatory trading algorithms that seek to profit from or even exacerbate market instability. However, the "nowcasting" nature of the target makes this more of a real-time risk indicator than a simple predictive signal for trading.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not release a high-risk model or dataset (e.g., a large language model or scraped personal data

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All software packages used (e.g., scikit-learn, statsmodels, NumPy) are standard open-source libraries, and their licenses (e.g., BSD, MIT) are respected. We credit the authors of the key methodological papers, such as Chernozhukov et al. [2018]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce or release any new datasets, models, or software

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This research does not involve crowdsourcing or human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This research does not involve human subjects and therefore does not require IRB approval

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Large language models were not used as part of the core research methodology or data generation process in this paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.