

---

# GEOFAC: Graph-Geometric Factor Models for Portfolio Construction

---

## Abstract

We propose GEOFAC (Graph-Geometric Factor Models for Portfolio Construction), an unsupervised graph–geometric framework for discovering latent, orthogonal sources of actionable information and constructing portfolios using such information in prediction markets. Starting with raw trades scraped from Polymarket, we construct co-trading graphs informed by trades that occur in similar time frames. We augment our graph with geometric features which encode critical information. Using this graph, we construct SAGEConv models that assign weights to wallets learned by trading returns. This construction of GEOFAC mimics the traditional factor portfolio framework which seeks to maximize a portfolio’s excess return over its variance – the Sharpe ratio ([Fernando \[2025\]](#)) – by gaining varying levels of exposure to different sources of risk in a market. GEOFAC supplants the traditionally supervised construction of risk factors found in the factor modeling framework using geometric methods to define and trade on factors in a self-supervised manner, allowing the model to learn what common sources of information people use for their trades and when a trader might have insider information. We observed that such natively geometric approaches to feature construction and training yielded the greatest recognition of informed trade activity, yielding improvements in resulting in Sharpe of up to 100% over baseline performance ([Fernando \[2025\]](#)).

## Introduction

In the past several years, Polymarket (2020) and Kalshi (2022) have built functionality for individuals across the world to trade on events, from elections and geopolitics to sports and pop culture. The companies’ value is the enablement of near instant, market-driven information propagation of provably accurate information. For example, essentially all major news outlets from the New York Times to Politico have referenced Kalshi metrics to benchmark political polls and world events after the platforms predicted electoral outcomes better than many polls. For instance, [Morrow \[2024\]](#) credits Polymarket with predicting the re-election of President Donald Trump in 2024 in the face of most expert opinions.

As participation in such prediction markets has grown – especially in the United States – many hedge funds, high frequency trading shops (HFTs), and individuals have developed largely successful (e.g. [Polymarket \[2025a\]](#), [Polymarket \[2025b\]](#), [Polymarket \[2025c\]](#)) trading strategies for Polymarket markets using historical and real time markets information. At the same time, many such instances of gross outperformance have brought forth speculation of trading on non-public information. [Osipovich \[2025\]](#) highlights an example of this, where hours prior to Venezuelan opposition leader María Corina Machado being announced as the recipient the 2025 Nobel Peace Prize, a newly activated Polymarket account loaded up on low-probability bets. Within the day, the position was up over 50,000 dollars, prompting accusations of insider trading. With our project, we seek to better understand and parameterize the information that markets trade on – from the most public and high frequency of sources such as crypto price volatility to the most privileged and unfair information such as who will win the Nobel Peace Prize – and develop methods to predict the associated likelihood of trading on insider information.

In our work, we draw inspiration from the traditional Factor Portfolio framework which seeks to explain portfolio returns through the lens of exposure to risk factors ([BlackRock \[2025\]](#)). However,

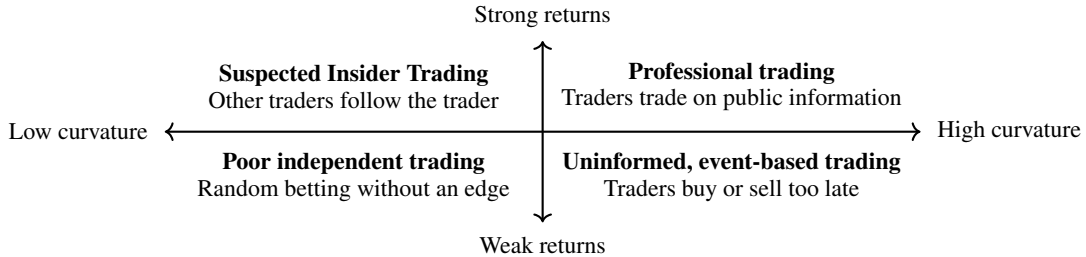
while in the traditional factor portfolio approach the practitioner has to define risk factors a-priori, our project uses the inherent cluster-like geometry of the graphs that we construct to help the SAGEConv learn latent information source representations that can be combined to maximize Sharpe. Under such construction, tightly-connected, positive curvature components of the graph correspond to wallets looking to similar information sources while other component types reflect idiosyncratic or weakly shared behavior.

## Background and Related Work

Practitioners have long believed that one of the most accurate sources of information to exist is prediction markets. [Wolfers and Zitzewitz \[2004\]](#) provides one of the strongest cases explaining prediction markets’ out-performance of traditional forecasting strategies. They find that the financial incentives, liquidity, and continuous trading ensure that the probabilities generated by prediction markets provide estimates of probabilities that events happen that are very close to reality. [Ng et al. \[2025\]](#) extends such analysis, finding that the deeper liquidity pools of Polymarket during the 2024 election yielded faster adjustments of probabilities and more accurate forecasting as well. However, due to the anonymity of specifically Polymarket’s platform, [Wood \[2025\]](#) has argued that insider trading is likely very pervasive on the platform.

With respect to viewing trades within a market as a network, [Betancourt et al. \[2017\]](#) finds that an effective strategy for understanding and modeling future returns over trading regimes is using network modeling. [Seabrook \[2023\]](#) extends this by formalizing the idea of an influencer in trading networks while [Oechssler et al. \[2018\]](#) analyzes copy-trading to find that people under-appreciate the risk associated with following successful traders. At the same time, [Sia et al. \[2019\]](#) proposes novel architectures for finding curvature and communities in complex network structures.

Given preexisting work into community detection in financial trading networks, we seek to replicate the explanatory power of trading communities in predicting returns. Our hypothesis was predicated on the idea that traders can be grouped by the sources that they get their information from – whether that is news, their own statistical models, alternative media, in-person experiences, privileged and non-public sources, etc. Given that this is the case, traders are likely to trade on different markets using that information. For instance, if a news alert comes that says the president is going to implement new tariffs on countries, then one trader might go to trade on a market explicitly about tariffs while another might use that information to bet on the number of travelers that enter the country next year. Traders that repeatedly trade on the same market at the same times are likely to be using the same information from similar meta-sources to make their trades, and thus if we create a co-trading graph these trades will show up with positive curvature. On the other hand, if there is a specific trader that other traders copytrade (buying similar portfolios to the original trader), then this is likely to show up with negative curvature because the following wallets are unlikely to be trading on similar markets except for when they copy trade.



We validate these hypothesis in Appendix C with some examples of wallets of interest that ended up fitting into our trading portfolio.

## Methods and Models

Our goal is to construct a portfolio of traders to copy from, which involves computing and aggregating what we call *information-based* beta factors, which ideally represent exposure to different assets and

strategies across the market <sup>1</sup> (Quantpedia [2025]). Taking  $\chi$  to be weights assigned to beta factors, we can naively express our resulting portfolio strategy as  $\pi = \chi\beta$ .

To do this, we construct a directed co-trading graph  $G = (V, E)$  where nodes  $V$  represent wallets and edges  $E$  encode trading relationships. For each wallet-condition pair, we identify continuum windows of temporal sequences of trades where consecutive trades are grouped into windows such that each trade in the window is separated by at most  $\Delta t$  seconds, and no other trade from that wallet occurs within  $\Delta t$  seconds of any trade in that window. For each continuum window  $w$  with start time  $t_s$  and end time  $t_e$ , we then define a co-trading window as  $[t_s, t_e + 40]$  seconds. Two wallets  $u$  and  $v$  are connected by a directed edge ( $u \rightarrow v$ ) if wallet  $v$  trades in wallet  $u$ 's co-trading windows at least  $\mu$  percent of the time for the same market. Edge weights are set to the fraction of overlapping windows, creating a weighted adjacency matrix. Each wallet node is represented by a feature vector  $x_v \in \mathbb{R}^d$ , with features including curvature of the local region within the graph, connectivity, trade count, historical mean return and variance, among other geometric components. We also include a one-hot encoding for the top-20 most traded series categories.

### Curvature-Informed GNN for Factor Portfolio Construction Model

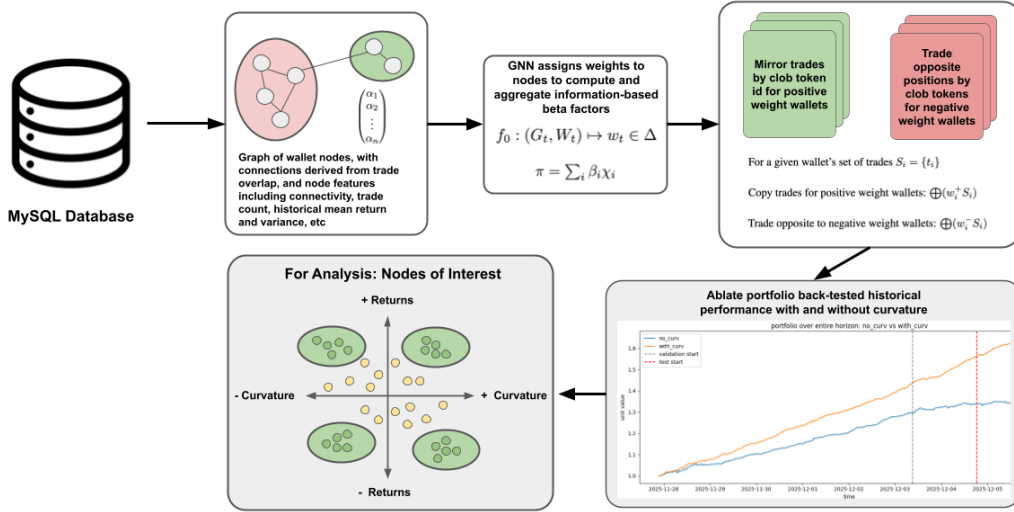


Figure 1: Schema Diagram

For a given time-frame, we then construct a SAGEConv-based GNN, consisting of 8 SAGEConv layers with hidden dimension 16, followed by a 2-layer MLP head ( $16 \rightarrow 16 \rightarrow 1$ ) that outputs raw scores  $s_v$  for each wallet. This model assigns node weights across our graph, yielding the map  $f_0 : (G_t, W_t) \mapsto w_t \in \Delta$ . For a given wallet's set of trades  $S_i = \{t_i\}$ , we then copy trades for positive weight wallets  $\oplus(w_i^+ S_i)$ , and trade opposite to negative weight wallets:  $\oplus(w_i^- S_i)$ , proportional to weight magnitudes. The idea is for the model to learn to combine disparate information sources in the latent space from the multitude of potential sources. Further, it should understand that it is non-optimal to highly weight wallets that are closely connected in the graph we construct because their covariance will be high without providing meaningful incremental information in their combination. <sup>2</sup>.

We then use a step-forward, rolling-window approach to train and evaluate our GNN portfolio model. We divide the time series into sequential chunks and construct targets by concatenating these blocks together.<sup>3</sup> For each chunk  $k$ , we compute returns over that block. We optimize the model to maximize the portfolio Sharpe returns, with our loss function is defined as  $\mathcal{L} = -\frac{\mu_p}{\sigma_p + \epsilon}$

<sup>1</sup>We can think of beta factors like principal components, both because they constitute directions of common movement that summarize how individual assets co-move with the broader system, and because they are usually difficult to map to exactly a single asset class or trading strategy.

<sup>2</sup>We also include several portfolio normalization strategies that apply CAPM-informed constraints. We've included details on this in the Appendix F

<sup>3</sup>For example, our training segment runs from chunks 2 through NUM\_CHUNKS-2, leaving the remaining data for validation and testing.

where  $\mu_p = \frac{1}{T} \sum_{t=1}^T r_{p,t}$  and  $\sigma_p = \sqrt{\frac{1}{T} \sum_{t=1}^T (r_{p,t} - \mu_p)^2}$  are the mean and standard deviation of portfolio returns  $r_{p,t} = \sum_v w_v r_{v,t}$  over  $T$  time steps. We train using Adam optimizer with learning rate  $10^{-3}$  and weight decay  $10^{-5}$  for up to 2000 epochs, with early stopping based on validation loss, and a patience of 300 epochs.

## Empirical and Theoretical Results

We observed a statistically significant difference in model performance when we included curvature as an input feature versus when we didn't. Our base setup obtained a Sharpe ratio of 0.302 without curvature, and 0.547 with curvature.

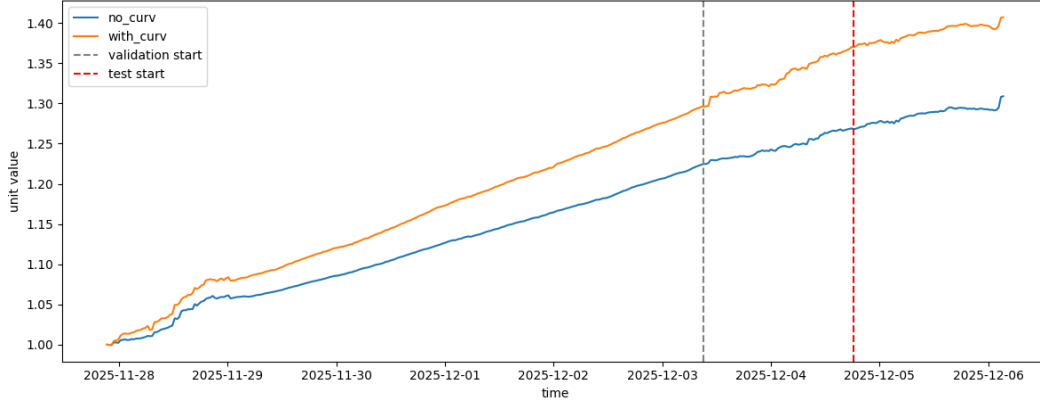


Figure 2: Returns of Factor Portfolio Using Curvature vs. Without Curvature

We then ran ablations over two variables: MIN\_RELATIONSHIP\_WEIGHT – the minimum threshold for considering a relationship (likely between wallets or trades) as significant, and MAX\_GAP\_SECONDS, the maximum allowed time gap (in seconds) between consecutive trades to group them as part of the same continuum window:

MAX_GAP_SECONDS	MIN_RELATIONSHIP_WEIGHT		
	0.10	0.30	0.65
Short (15)	0.1779 (0.3034)	0.2660 (0.4826)	0.3020 ( <b>0.5470</b> )
Medium (120)	-0.1495 (0.0924)	0.2654 (0.2005)	<b>0.3741</b> (0.4830)
Long (600)	0.0924 (-0.0538)	0.2123 (0.2364)	0.2738 (0.4507)

Table 1: Validation Sharpe by MIN\_RELATIONSHIP\_WEIGHT and MAX\_GAP\_SECONDS no curvature (with curvature in parentheses). Each test was run 3 times.

Our results indicate strongest performance when utilizing large thresholds for constructing graph edges, as well as short intervals between trades, with curvature-informed results leading to superior performance a majority of the time. We also find that incorporating curvature as a feature yields the greatest performance gains at higher relationship weights.

Further, we found that the model is able to effectively discriminate whether it should be long or short certain wallets very effectively, with the most influential wallets (both on the long and the short side) being found in Appendix E. It also constructed a portfolio with strong diversification across asset classes – a hallmark of a portfolio with a strong Sharpe as seen in Appendix D.

Finally, we suspected that disparate markets would have varying geometric structure, as the composition and activity patterns of market participants would itself differ. Our intuition was confirmed. Running a local curvature measure, we computed curvature for the top 50 markets by trade volume, and found that the markets with the lowest curvature happened to be high-frequency crypto markets. We suspect this is because these markets are often dominated by algorithmic traders, who can systematically follow the wallets which perform well. Another way to express this intuition is through a measure we made called leader ratio: mathematically defined as  $\frac{\text{out degree}}{\text{in degree} + 1}$ , it roughly measures the

relative frequency of leaders and followers per market. The correlation of leader ratio with curvature is  $-0.897$  which indicates that in markets driven primarily by public, highly-available information it is hard to find a subset of the wallets that are first to act or are leaders. On the other hand, in negative curvature markets we find that there is frequently a subset of the traders that are leaders or act first – likely on privileged information. We show more information on this in Appendices H and I.

## Conclusion and Future Work

Our results indicate a meaningful contribution to returns and Sharpe ratio when implementing curvature information into our weight-assigning GNN, indicating that our graph constructions carry underlying geometric data that aids portfolio construction. One structural limitation of our work is the inability to map disparate wallets to common individuals or groups, since new accounts can be opened up anonymously, and money can be transferred freely between wallets. Kalshi, Polymarket’s aforementioned competitor, alleviates this issue by requiring SSN-based identity verification when creating a new account, and while names are still not publicly available, the additional layer of accountability likely influences the frequency of informed trading.

Another limitation of our works stems from arbitrage trading between different prediction market platforms, where traders exploit price discrepancies between platforms in a manner that guarantees positive returns once the markets resolve ([Wang \[2025\]](#)). This can result in artificially high or low returning wallets which aren’t representative of the quality of information that wallets use. We were also bottlenecked by compute as the dataset we work with is massive.

A geometric limitation to our work is that our computed node curvature is only well-defined up to how we construct our trader graphs, and the hyperparameters we choose, rather than being some intrinsic, constant property of the nodes themselves. This is a natural limitation of constructing our own graphs rather than working with graphical data or data that has some canonical way of being composed into a graph.

Looking to other approaches that could have been taken for this problem, we see that while geometric methods are more difficult than other options that could be pursued, the inherent geometric nature of information flow can be best modeled in the graph structures that we propose.

One alternative approach that we could have taken is a simple traditional factor-based portfolio based on heuristics of markets or wallets. This would have entailed looking at the trends of the markets, order flows, and order books to understand where the price is likely to go in the future. This strategy is advantageous on account of more interpretable factors; we can better understand what we have exposure to and why returns fall the way they do. This could allow us to have a more robust strategy for hedging against tail risk to ensure that our portfolio continues to perform well.

One of the key disadvantages of such an approach is that we may trade performance for inheritability. If one believes the key claim of prediction markets that they are closer to ground truth because of the financial incentive involved, then it stands that it is very hard to beat the general market. As more people become interested in these markets, that will only continue to become harder as markets become more efficient. Our strategy of parameterizing and discovering information sources in the latent space using a geometric approach yields an edge over the traditional linear approaches that traditional factor portfolios offer. The traditional approach simply has less expressive power than our graph-based non-linear approach.

Another competing approach that could be explored would be to construct category-based models that are able to perform well within specific domains of markets instead of trading across the board. This supervised approach would likely yield stronger returns in terms of the Sharpe ratio, but would be exposed to less markets and thus be able to deploy less capital overall. Further, such a model would require serious upfront effort in terms of gathering data and creating pipelines. While our method is unable to capture this level of information at the speed and efficiency by which bespoke data pipelines might be able to, when wallets start trading on this information it should be possible to follow along as we could detect the presence of new information. Thus, because of the breadth of our strategy and the return on effort, the geometric-based trade monitoring seems like the optimal approach in this domain.

The code we used for model training and evaluation is available at [Brigham et al. \[2025\]](#) while the Youtube video for the presentation is at [Brigham \[2025\]](#).

## References

- B. Betancourt, A. Rodríguez, and N. Boyd. Modelling and prediction of financial trading networks: An application to the NYMEX natural gas futures market. 2017. URL <https://arxiv.org/abs/1710.01415>.
- BlackRock. What is factor investing?, 2025. URL <https://www.blackrock.com/us/individual/investment-ideas/what-is-factor-investing>.
- T. Brigham. Cpsc 6440 final. <https://youtu.be/FbfGFX0wgbA>, 2025. YouTube video.
- T. Brigham, T. Dimov, and O. Laskov. CPSC6440\_Final: Geometric gnn portfolio construction on polymarket, Dec. 2025. URL [https://github.com/TristanB22/CPSC6440\\_Final](https://github.com/TristanB22/CPSC6440_Final).
- C. Bürgi, W. Deng, and K. Whelan. Makers and takers: The economics of the kalshi prediction market, 2025. URL <https://www.karlwhelan.com/Papers/Kalshi.pdf>. University College Dublin.
- F. Feng, X. He, X. Li, M. Zhang, and T.-S. Chua. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems*, 2019.
- J. Fernando. Sharpe ratio: Definition, formula, and examples, 9 2025. URL <https://www.investopedia.com/terms/s/sharperatio.asp>.
- W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 2017. URL <https://arxiv.org/abs/1706.02216>.
- M. Haugh. Mean-variance optimization and the CAPM. Lecture notes, IEOR E4706: Foundations of Financial Engineering, 2016. URL <https://www.columbia.edu/~mh2078/FoundationsFE/MeanVariance-CAPM.pdf>.
- A. Morrow. How prediction markets saw something the polls and pundits didn’t. *CNN Business*, Nov. 2024. Nightcap newsletter.
- H. Ng, L. Peng, Y. Tao, and D. Zhou. Price discovery and trading in prediction markets. *SSRN Electronic Journal*, Nov. 2025. doi: 10.2139/ssrn.5331995. URL <https://ssrn.com/abstract=5331995>.
- J. Oechssler, S. Weidenholzer, and J. Apesteguia. Financial social trading networks: The case of copy trading platforms, Sept. 2018. URL <https://cepr.org/voxeu/columns/financial-social-trading-networks-case-copy-trading-platforms>. VoxEU column.
- A. Osipovich. A mystery trader scored with prescient bets on the nobel peace prize. *The Wall Street Journal*, Oct. 2025. Markets & Finance.
- Plotly. Plotly open source graphing library for python, 2025. URL <https://plotly.com/python/>. Version 6.5.0 documentation.
- Polymarket. User profile: Firstorder, 2025a. URL <https://polymarket.com/@FirstOrder>.
- Polymarket. User profile: minglee2, 2025b. URL <https://polymarket.com/@minglee2>.
- Polymarket. User profile: piastri, 2025c. URL <https://polymarket.com/@piastri>.
- Quantpedia. Beta factor, 2025. URL <https://quantpedia.com/strategy-tags/beta-factor/>.
- O. Saguillo, V. Ghafouri, L. Kiffer, and G. Suarez-Tangil. Unravelling the probabilistic forest: Arbitrage in prediction markets. *arXiv preprint arXiv:2508.03474*, 2025. URL <https://arxiv.org/abs/2508.03474>.
- I. Seabrook. *Influencers in Dynamic Financial Networks*. Doctoral thesis (ph.d.), University College London, London, United Kingdom, 2023. URL <https://discovery.ucl.ac.uk/id/eprint/10167705>. Open access thesis available via UCL Discovery.

- J. Sia, E. Jonckheere, and B. Bogdanovic. Ollivier-ricci curvature-based method to community detection in complex networks. *Scientific Reports*, 2019. doi: 10.1038/s41598-019-46030-3.
- G. Wang. How to programmatically identify arbitrage opportunities on polymarket (and why i built a portfolio betting agent instead), Aug. 2025. Accessed: 2025-12-12.
- J. Wolfers and E. Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004. doi: 10.1257/0895330041371321.
- A. Wood. How prediction markets raise insider trading and credit risks, Dec. 2025. URL <https://cointelegraph.com/news/prediction-markets-insider-trading-credit-risks>. Analysis article.

## A Dataset Description

We use a proprietary dataset of 108 million Polymarket trades scraped via the platform’s public API. The data is stored in a MySQL environment on a local storage system and structured as follows:

- **activity\_trades:** Complete record of all recorded transactions with details including size, price, position, market, and time stamp. Totals 108,233,011 entries.
- **events:** Names and descriptions of real world occurrences which define markets that users can trade on. Totals 60,614 entries.
- **event\_series:** Mapping table relating events to series.
- **markets:** Information about markets that have existed on the Polymarket platform, including descriptions, duration, volume, and outcome. Totals 623,206 entries and 83 variables.
- **market\_events:** Key that maps markets to larger events. Totals 623,206 entries and 5 variables.
- **market\_outcomes:** Outcome prices for markets, including resolved and ongoing markets. 1,246,882 entries across 7 variables.
- **market\_sourced\_trades:** For each market, a record of the last 1,500 trades that occurred. Totals 17,863,970 entries across 15 variables.
- **market\_summary:** Titles and descriptions of markets with corresponding unique token ids for YES and NO outcomes.
- **outcomes:** A list of 25 possible market outcomes, almost exclusively YES or NO in recent markets.
- **series:** Descriptions of series (repeating markets) such as Elon Tweets or NFL, including frequency and volume. 240 entries across 31 variables.
- **trade\_types:** 8 trade types from CONVERSION, MARKET, MERGE, REDEEM, REWARD, SPLIT, TRADE, YIELD.
- **wallets:** Wallet addresses and associated pseudonyms. 797,241 entries across 14 variables.

This dataset is sufficiently sized and representative of Polymarket activity over several years. Completeness is ensured by the decentralized, auditable nature of the exchange and the comprehensive wallet-graph-based scraping approach.

## B SAGEConv

We use SAGEConv, an extension of GraphSAGE first proposed by [Hamilton et al. \[2017\]](#), as an inductive graph neural network to learn wallet-level representations from the co-trading graph. Let  $G = (V, E)$  denote the wallet graph with node features  $x_v$ . GraphSAGE computes node embeddings by iteratively aggregating information from local neighborhoods rather than learning node-specific parameters. Initial representations are set as  $h_v^{(0)} = x_v$ , and for each layer  $k = 1, \dots, K$ , neighborhood information is aggregated and combined with the node’s previous representation:

$$h_v^{(k)} = \sigma \left( W^{(k)} \cdot \text{CONCAT} \left( h_v^{(k-1)}, \text{AGGREGATE}^{(k)} \left( \{h_u^{(k-1)} : u \in \mathcal{N}(v)\} \right) \right) \right),$$

where  $\text{AGGREGATE}^{(k)}$  is a permutation-invariant operator such as mean or pooling, and  $\sigma$  is a nonlinearity. Our final node embedding is given by  $z_v = h_v^{(K)}$ .

Model parameters are then trained using an unsupervised, contrastive objective that encourages nearby nodes in the graph to have similar embeddings while pushing apart unrelated nodes. For a node  $u$ , GraphSAGE uses the following loss criterion:

$$\mathcal{L}_u = -\log \sigma(z_u^\top z_v) - Q \mathbb{E}_{v_n \sim P_n} [\log \sigma(-z_u^\top z_{v_n})],$$

where  $v$  is sampled from short random walks starting at  $u$ ,  $v_n$  are negative samples, and  $Q$  controls the number of negatives. In GEOFAC, we use resulting embeddings to assign wallet weights, allowing the model to incorporate local graph structure and geometry when constructing copy-trading portfolios.

## C Trader Portfolio Examples

Below are some examples of wallets, their associated returns, and our models assigned weighting. First, we see an example of a high curvature, positive returning wallet, which our model correctly assigned positive weight to when composing the portfolio (wallet 0xa20b482f97063f4f88ef621c9203e60814399940 has an assigned weight of 0.025, and a curvature of 0.99).

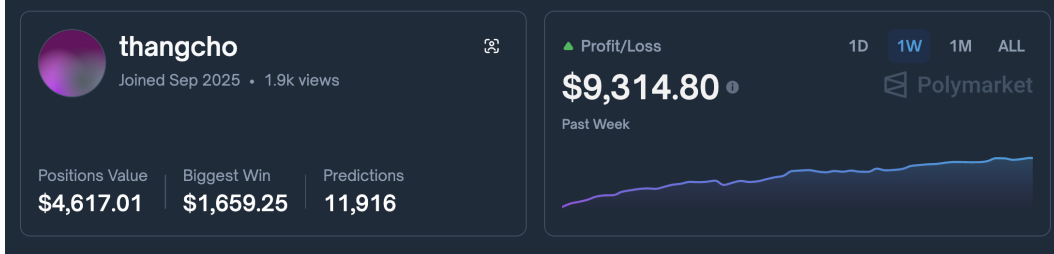


Figure 3: Positive Returns, Positive Weight Example

Additionally, we see examples of high curvature, negative returning wallets, which our model correctly assigned negative weights to (wallet 0xe3726a1b9c6ba2f06585d1c9e01d00afaedaeb38 has an assigned weight of -0.029, and a curvature of 1.0).

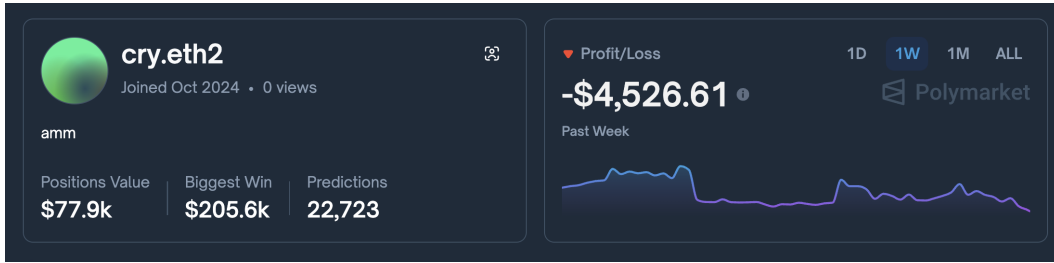


Figure 4: Negative Returns, Negative Weight Example

We also see that within these high curvature clusters, our model only selected several wallets rather than spreading weight equally across all nodes, which is exactly the result we intended for.

0	0x00090e8b4fa8f88dc9c1740e460dd0f670021d43	+1.000000	+0.000000	+0.000000
1	0xa9736e281441255fc815a65ae4f9ba4c0b77f0f6	+1.000000	+0.000000	+0.000000
2	0xa9335a2b5e88645ee5bb552d8675626ed0d94e84	+1.000000	+0.000000	+0.000000
3	0xa952437457ad260b8f47305ac35e4bde4d0f5ca1	+1.000000	+0.000000	+0.000000
4	0xa9587187a9194682bf115ff4d64fa7a86cc8fecf	+1.000000	+0.000000	+0.000000
5	0xa95c277f15519e5b4c6833bd9144f3a0f395517a	+1.000000	+0.000000	+0.000000
6	0xa95f45fe6ee731ae5c631402fb86852700e52bdb	+1.000000	+0.000000	+0.000000
7	0xa9650fe4301f45e7f090ada7252f9c1268183565	+1.000000	+0.000000	+0.000000
8	0xa976d4a7248126439a189bcd5538c4e501735c43	+1.000000	-0.002372	-0.000671
9	0xa9b44dca52ed35e59ac2a6f49d1203b8155464ed	+1.000000	+0.010569	+0.003543
10	0xa97cb2eb997156154e3506ee0c069d49e5aa55b2	+1.000000	+0.000000	+0.000000

Figure 5: Many positive curvature wallets received weights of zero

## D Disparate Categories Drive Returns

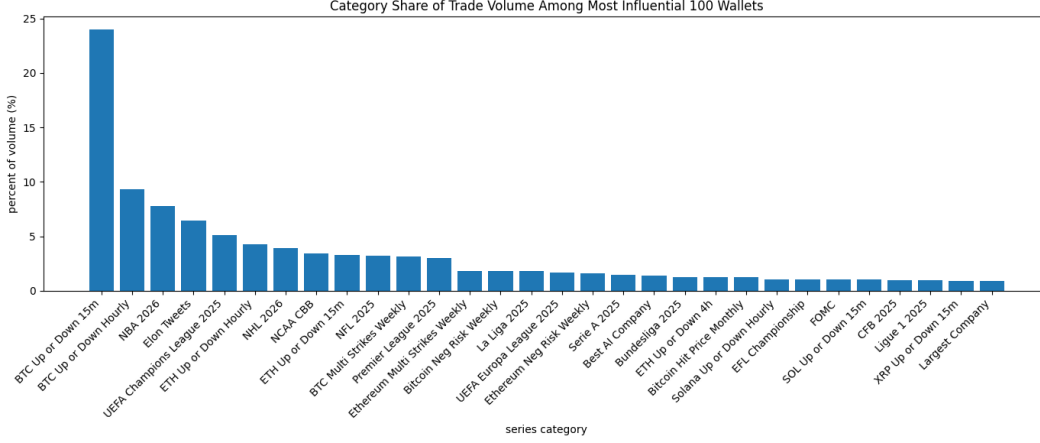


Figure 6: The Model Learned to Integrate Several Sources of Information

## E Weightings of Portfolio Allocations

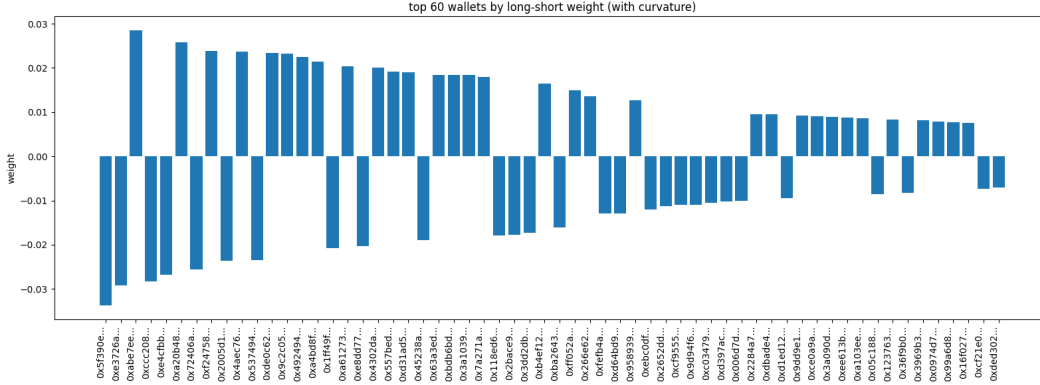


Figure 7: Varying Long/Short Exposures for Wallets

## F CAPM-based Portfolio Normalizations

In order to construct portfolios, we first ensure that we have market-neutral portfolios constructed. We do so by de-meaning the raw scores that we end up predicting in the model for each of the wallets. That is, for the predicted weight  $z_i$  for wallet  $i$ , we end up rescaling the weight such that we have:

$$\tilde{z}_i = \frac{1}{|V|} \sum z_i$$

We then control for the leverage that the model is allowed to take on to ensure that the model does not artificially boost results by taking on more leverage than is allowed which gives us our final weight for each portfolio and ensures that we adhere to a unit portfolio constraint:

$$w_i = \frac{\tilde{z}_i}{\sum_{v \in V} |\tilde{z}_i|}$$

Besides this set of constraints, we also tried other methods of normalizing the portfolio to get a unit portfolio that could effectively show us how we performed. For instance, we attempted to construct a

long-only portfolio (that is, shorting is not allowed) which yielded worse returns than other methods that we tried.

We also tried taking the softmax of the weights using the softmax function given by the following, but found that that ended up over-weighting some portfolios compared to others which decreased Sharpe:

$$\mathcal{S}(w_i) = \frac{e^{w_i}}{\sum_{j=1}^n e^{w_j}}$$

## G Graph Visualizations

To evidence the fact that our constructed graphs are encoding meaningful information (and are neither too sparse or fully connected), we create Plotly visualizations with samples of 300 random data wallets in our dataset ([Plotly \[2025\]](#)). Notably, we also observe different clustering structures with different samples of data, indicating diverse wallet activity.

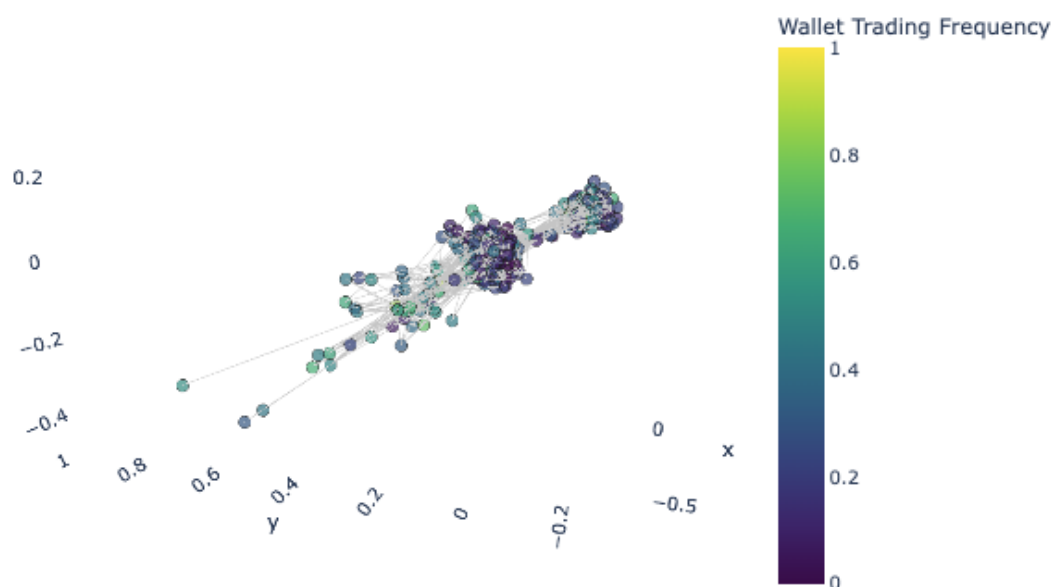


Figure 8: Sample 1 Visualization

## H Curvature by Market

Curvature is heterogeneous across markets. Markets which are disproportionately smart, where a lot of following is taking place, are low curvature. This is confirmed in the data, as the lowest 10 curvature markets are all high frequency crypto. We suspect this due to the presence of algorithmic traders acting as a source of privileged information due to the inaccessibility of creating high-performing crypto models due to data gaps.

Top Markets by Leader Ratio			
	Market	Leader Ratio	Curvature
0	BTC Up or Down Hourly	3.7346	0.5243
1	BTC Up or Down 15m	3.2832	0.5251
2	ETH Up or Down Hourly	1.8685	0.5421
3	ETH Up or Down 15m	1.8421	0.6520
4	Solana Up or Down Hourly	1.1456	0.7475
5	ETH Up or Down 4h	1.0347	0.8456
6	XRP Up or Down 15m	0.9629	0.7927
7	SOL Up or Down 15m	0.9543	0.7756
8	BTC Up or Down Dally	0.9063	0.7328
9	BTC Up or Down 4h	0.7454	0.9533

Figure 9: Market Curvatures

## I Correlation of Curvature and Leader Ratio

Leader ratio is measured as  $\frac{\text{out degree}}{\text{in degree}+1}$ . It broadly captures how many leaders to followers are in a given market with respect to the trade ordering. The correlation of this measure to curvature across all markets is  $-0.897$ . Qualitatively, we see is that the number of 'leaders' is the highest in low curvature high frequency crypto markets. This would imply that algorithmic traders in these markets often copy well-performing wallets.

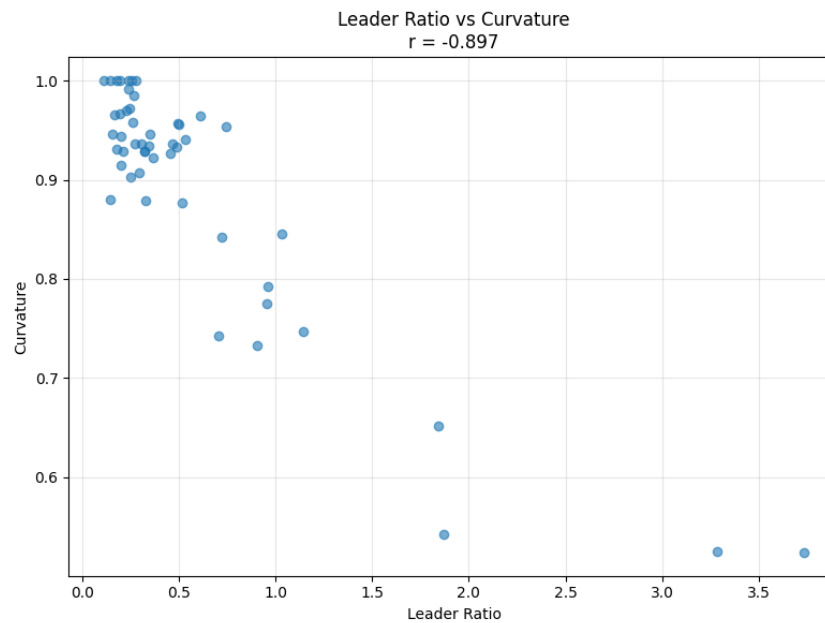


Figure 10: Leader Ratio Against Curvature