Alpha Signal Discovery via Numerical Representation Learning

Deep Mathematical Analysis of Time-Series and Graph Structures for Robust Alpha Generation

By Ace Prasad
Regnum: 2306629

- Table of Contents:

- Key Terms:

- Representation Learning: ML process that allows algorithms to automatically discover representations (patterns from raw data)

- Alpha Signals: financial metric representing the excess return of an investment relative to a benchmark index (a.k.a active returns)

- Self-Supervised Learning (SSL) Disentanglement: learning distinct representations from data without labelled examples (separating various market dynamics, enhancing interpretability / stability of generated alpha signals from time-series / graph-structured data.

- Advanced econometric models: (ARIMA, GARCH, VAR, FF3F) statistical techniques to analyse financial time seriescontributing towards alpha generation

- factor models: statistical frameworks used to explain asset returns via underlying market factors / charachteristics

- stochastic processes: models for system evolution over time time in probabilistic manner (Random Walks – Basic Model, Geometric Brownian Motion – Asset Price Modelling, Mean Reverting Processes – Ornstein-Uhlenbeck Processes)

- Spectral analysis: used to analyse frequency components of time-series data, helping to comprehend underlying patterns, trends and cycles in asset prices.

- wavelet transforms: analysing time series data by decomposing signals into their constituent components at different scales and resolutions, useful for capturing non-stationary behaviours

- fractional differencing: technique used to transform non-stationary time-series data into stationary data while preserving long memory properties.

- Cointegration: statistical concept used to analyze the long-term relationships between non-stationary time-series variables.

- Numerical Optimization: Techniques for signal construction and portfolio allocation.

- Proposed Research Problem Statements:

Primary Question:

- Can advanced numerical representation learning methods, applied to financial time series and graph-structured data, extract stable and interpretable alpha signals that significantly outperform traditional quantitative benchmarks?

Supporting Questions:

- How do disentangled numerical representations, derived from time-series and graph data, capture distinct market dynamics (e.g., momentum, value, carry, liquidity) and contribute to alpha generation?

- Can self-supervised learning objectives, designed for numerical data, effectively learn robust features that generalize across diverse market conditions and asset classes?

- What is the impact of incorporating complex mathematical relationships (e.g., non-linear dependencies, higher-order moments, network effects) from numerical data on the stability and profitability of the generated alpha signals?

- How can advanced numerical optimization techniques be integrated with learned representations to construct portfolios with superior risk-adjusted returns?

- Tech Stack:

Component	Tools / Libraries
Base Language	Python
ML Frameworks	PyTorch, TensorFlow, PyTorch Geometric
Time Series / Econometrics	Statsmodels, Arch, Tsfresh, Pywt
Numerical Optimization	SciPy, Cvxpy
Graph Analysis	networkx, igraph
Data Source / API's	Yahoo Finance, Alpha Vantage, Quandl, WRDS (Wharton Research Data Services), FRED (Federal Reserve Economic Data), EDGAR API.
Backtesting & Eval.	Bt, Backtrader, QuantConnect
Visualization	Matplotlib, Seaborn, Plotly

- Prior Knowledge Required:

Quantitative Finance:

Asset pricing models (Capital Asset Pricing Model (CAPM), Arbitrage Pricing Theory (APT), and multi-factor models), factor investing, Information Coefficient (IC) /Sharpe & portfolio construction).

Machine Learning:

Deep learning architectures (feedforward neural networks, convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers), model training dynamics, evaluation metrics (accuracy, precision, recall, F1 score, and area under the ROC curve).

Representation Learning:

Embedding techniques (PCA (Principal Component Analysis), t-SNE (t-distributed Stochastic Neighbor Embedding), and autoencoder usage, disentanglement, inductive bias.

Self-Supervised Learning:

Pretext tasks, contrastive learning (positive and negative samples, e.g SimCLR and MoCo), masked modeling.

Graph Learning:

(Node2Vec, DeepWalk, GCNs, GATs, financial graphs).

Advanced Econometrics:

Time-series econometrics (ARIMA, GARCH, VAR models), panel data analysis, cointegration, causality testing (Granger causality).

Stochastic Calculus:

(e.g Geometric Brownian motion, martingales, and Ito's lemma) to facilitate understanding and modeling complex financial processes, mean-reverting processes (e.g., Ornstein-Uhlenbeck processes).

Numerical Linear Algebra:

For efficient computation with large numerical datasets (multiplication, inversion, and eigenvalue decomposition) w/t optomization techniques (gradient descent)

* Optimization Theory - For portfolio construction and signal weighting (Optional)

- Data & API’s:

Data Type	Sources
Price Data	Yahoo Finance, Quandl (Nasdaq Data Link - free datasets)
Fundamental Metrics	P/E, EV/EBITDA, ROE (from Yahoo Finance)
Volatility & Risk Metrics	CBOE (VIX, Skew Index - limited data available)
High-Frequency / Intra-Day Data	Tiingo (limited free tier), Polygon.io (limited free tier)
Sentiment & Textual Signals	Twitter API (free tier), CBOE Sentiment Index, News API (e.g., GNews, NewsAPI - free tier)
Macro Indicators	FRED, OECD (some datasets available for free), World Bank API (some datasets available for free)
Graph Data	GICS classification, correlation matrices, ownership networks (constructed using free libraries)
Alternative Data	Web scraping for e-commerce data, social media analytics (using free tools)
Economic Indicators	IMF Data (some datasets available for free), Conference Board Leading Economic Index (LEI - free)
Technical Indicators	Custom calculations (momentum, moving averages, Bollinger Bands) using free libraries like TA-Lib
Market Depth Data	Level 2 data from exchanges (limited free access), IBKR API (limited free tier)
Consumer Behavior Data	Web traffic analytics (e.g., SimilarWeb - limited free tier)

- **Feature Set:**

Fundamental Metrics

- P/E Ratio (Price-to-Earnings)
- EV/EBITDA
- ROE (Return on Equity)
- Debt/Equity Ratio:
- Cash Flow Ratios

Technical Indicators

- Momentum Indicators
- Volatility Indicators
- Volume-Based Indicators (OBV, Accumulation/Distribution Line)
- Mean Reversion Signals

Graph-Based Features

- Node Embeddings (from GNNs)
- Centrality Measures (Degree, Betweenness & Eigenvector Centrality
- Community Membership
- Graph Density
- Clustering Coefficients

Time-Series Specific Features

- Partial/Auto correlation
- Hurst Exponent
- Wavelet Coefficients
- Statistical Moments of Returns (Skewness, Kurtosis)

Regime Indicators

- VIX (Volatility Index)
- Drawdown Markers
- Yield Curve Slope
- Credit Spreads
- Macroeconomic Cycles

- Data & API Constraints:

Data Source/API	Constraint
Alpha Vantage	500 calls/day, 5 calls/minute limit on the free tier.
WRDS (CRSP/ Compustat)	Requires institutional access (through a university subscription).
Quandl (Free Tiers)	Limits on accessible datasets, daily API calls, data history depth.
Tiingo (Free Tier)	50 requests/hour and 500 symbols/month.
Polygon.io (Free Tier)	5 requests/minute on the free tier; minute-level data via paywall.
FRED API	~ 120 calls/minute; subject to undocumented throttling.
Yahoo Finance	Undocumented rate limits unstable for automated scraping. Not official API.

- Potential Issues:

Problem Category	Potential Problem
Granularity & Consistency	Financial data in various frequencies requiring careful alignment
Validity & Quality	Lookahead bias -> inflated backtesting results
Noise & Non-Stationarity	Historical financial data is noisy and volatile, necessitating robust preprocessing techniques.
Corporate Actions & Adjustments	Raw price data may not account for stock splits or dividends,
Cost & Licensing	Premium data sources = costs, limits access to complete datasets.
High-Dimensionality & Compute	Financial graphs with many nodes increase model complexity and require significant computational resources.

- ## Development Timeline:

Phase	Month	Week	Activities	Deliverables
Phase 1: Foundation & Data Engineering	Month 1-2		**Goal:** - Establish robust data pipelines. - preprocess core numerical datasets. - set up the foundational ML environment.	
	Month 1	Week 1-2	- **Define Asset Universe:** Select a manageable asset universe (e.g., S&P 500 / specific sector). - **Core Data Acquisition:** automate scripts to download daily/weekly OHLCV data, fundamental metrics, key macro indicators from chosen sources. - **Initial Data Cleaning:** Handle missing values, basic outlier detection, and corporate action adjustments.	

		Week 3-4	- **Temporal Alignment:** Implement robust methods to align data across different frequencies. - **Feature Engineering (Basic Numeric):** Compute standard technical indicators and fundamental ratios. - **Environment Setup:** Finalize Python environment and install necessary libraries.	- Cleaned and aligned core numerical dataset. - Functional data ingestion pipeline. - Basic feature set. - Configured ML environment.
Phase 2: Graph Construction & Baseline Modeling	Months 3-4		**Goal:** - Build initial financial graphs. - Establish baseline alpha signal models for comparison.	

	Month 3	Week 1-4	- **Graph Definition:** Decide on initial graph types. - **Graph Construction:** Implement scripts to build dynamic graphs. - **Graph-based Feature Engineering:** Extract basic graph features for each time step.	- Dynamic financial graphs. - Initial graph-based features.
	Month 4	Week 1-4	- **Baseline Alpha Models:** Implement simple traditional alpha models and basic ML model using only numerical time-series features. - **Initial Backtesting Framework:** Set up a basic backtesting framework to evaluate baseline models.	- Implemented baseline alpha models. - Functional backtesting framework.
Phase 3: Deep Numerical Representa-tion Learning	Months 5-7		**<u>Goal:</u>** - Develop and train core DL models (GNNs) for numerical representation learning.	

	Month 5	Week 1-2	- **GNN Architecture Selection:** Choose a suitable GNN architecture. - **GNN Integration:** Integrate GNNs into the alpha signal generation pipeline.	
		Week 3-4	- **Initial GNN Training:** Train the GNN model on historical data. - **Signal Generation:** Generate alpha signals from GNN embeddings.	
	Month 6	Week 1-2	- **Advanced Time-Series Model Fusion:** Implement more sophisticated time-series model. - **Disentangle-ment Strategy:** Begin exploring methods for disentangling numerical factors.	

		Week 3-4	- **Numeric Fusion:** Develop a strategy for fusing representations from GNNs and advanced time-series models. - **Combined Signal Generation:** Generate alpha signals from the fused numerical representations.	
	Month 7	Week 1-2	- **Rolling-Window Validation:** Implement robust rolling-window backtesting for the DL models. - **Performance Metrics:** Evaluate signals using key metrics.	
		Week 3-4	- **Hyperparameter Tuning (Initial):** Conduct initial hyperparameter optimization. - **Model Iteration:** Identify AOI based on initial results.	- Trained GNN + advanced time-series models. - Initial alpha signals from DL models.

Phase 4: Robustness, Interpreta-bility & Manuscript Drafting	Months 8-10		**<u>Goal:</u>** - Rigorous model test for robustness. - Enhance interpretability. - Begin drafting the conference paper.	
	Month 8	Week 1-2	- **Stress Testing:** Evaluate signal performance under different market regimes. - **Sensitivity Analysis:** Test sensitivity to key parameters or data inputs.	
		Week 3-4	- **Out-of-Sample Generalization:** Ensure signals generalize well to unseen periods. - **Ablation Studies:** Understand the contribution of different numerical feature sets or model components.	

	Month 9	Week 1-2	- **Disentangle-ment Analysis:** Quantify disentanglement of learned numerical factors. - **Signal Interpretation:** Analyze economic intuition behind the generated alpha signals.	
		Week 3-4	- **Model Refinement:** Incorporate insights from robustness and interpretability analysis to further refine models. - **Final Performance Evaluation:** Comprehensive backtest and final performance stat generation.	
	Month 10	Week 1-4	- **Outline Paper:** Create a detailed outline for the conference paper. - **Draft Sections:** Begin drafting the methodology, data, and results sections.	- Extensive robustness analysis. - Interpretability insights. - Refined DL models. - First draft of conference paper.

Phase 5: Finalization & Submission	Months 11-12		**Goal:** - Polish paper. - Prepare presentation materials. - Submit to target conferences.	
	Month 11	Week 1-2	- **Review & Edit:** Thoroughly review and edit the paper for academic rigor. - **Figures & Tables:** Create high-quality figures and tables to illustrate key findings.	
		Week 3-4	- **Presentation Draft:** Prepare a draft of the conference presentation. - **Rehearsal:** Practice presenting the work.	
	Month 12	Week 1-2	- **Final Checks:** Perform final checks on data, code, and paper. - **Submission:** Submit the paper to the target conferences.	- Finalized conference paper. - Presentation slides. - Successful submission to target conf.
		Week 3-4	- **Contingency/ Next Steps:** Potential revisions for journal submission if conf. paper is accepted.	

- Lean Cost Breakdown:

Category	Resource	Cost	Notes / Substitutes
Core Financial Data	Yahoo Finance, Alpha Vantage (free tier), Quandl (free tiers), IEX Cloud (free tier).	$0	Provides daily OHLCV data, market cap, sector information, and historical returns for a wide range of equities. Sufficient for prototyping and backtesting.
Fundamental Financial Data	Yahoo Finance, Alpha Vantage (free tier), public company filings (EDGAR API - free).	$0	Key fundamental metrics are often available via free APIs or can be extracted from public filings.
Macroeconomic Indicators	FRED API (Federal Reserve Economic Data), OECD Public API, Quandl (free tiers).	$0	Comprehensive and reliable macroeconomic data is freely available from official sources.
Graph-Related Data	Calculated from price data (correlation matrices), GICS classifications, publicly available industry reports/studies.	$0	No need for expensive proprietary graph databases. Graphs can be constructed from readily available numerical data.
Options Data	Yahoo Finance (limited options chain data), EODHistoricalData (free tier), Polygon.io (limited free).	$50	Public options chain data for liquid underlyings includes strikes, maturities, and implied volatilities. Tick-level data is expensive, but EOD data is often sufficient for features.
Modeling & Computation		$0 - $100	

Category	Resource	Cost	Notes / Substitutes
Deep Learning Frameworks	PyTorch, TensorFlow, scikit-learn, HuggingFace, PyTorch Geometric, DGL.	$0	All major deep learning and graph neural network frameworks are open-source and free to use.
Time-Series & Econometrics Libraries	pandas, numpy, statsmodels, arch, tsfresh, pywt, linearmodels.	$0	Comprehensive libraries for numerical data manipulation, statistical analysis, and econometric modeling are open-source.
Numerical Optimization Libraries	SciPy, CVXPY (for convex optimization).	$0	Standard Python libraries for numerical methods and optimization.
Compute Infrastructure	Kaggle Kernels (free GPU access), Google Colab Pro (optional), University GPU Clusters / HPC.	$100	Deep learning models benefit significantly from GPUs. Free tiers are available, but Colab Pro ($10/month) offers more robust access. University resources are ideal if available. Apply for academic cloud credits if needed.
Backtesting & Evaluation		Estimated Cost: $0	
Backtesting Frameworks	bt (Python library), Backtrader, QuantConnect (free tier), custom Python scripts.	$0	Open-source libraries provide robust tools for simulating trading strategies and evaluating performance metrics.
Visualization	Matplotlib, Seaborn, Plotly (for interactive plots).	$0	Standard Python visualization libraries are free and powerful.
Literature Access	arXiv, SSRN, Semantic Scholar, Google Scholar.	$0	Vast majority of academic papers are accessible for free.

- Potential Co-Authors:

Lan Zhang – University of Illinois Chicago
Specialization: High-frequency data, realized volatility.
Potential Contribution: Guides the use of high-frequency numerical data for robust volatility estimation and feature engineering, crucial for time-series analysis.

Yacine Aït-Sahalia – Princeton University
Specialization: Stochastic volatility, diffusion models.
Potential Contribution: Deep expertise in the mathematical modeling of stochastic processes

Damiano Brigo – Imperial College London
Specialization: Stochastic modeling & financial theory.
Potential Contribution: Expertise in rigorous mathematical modeling and ensuring financial soundness (e.g., no-arbitrage principles) in DL frameworks for numeric data.

Damir Filipović – EPFL / Swiss Quant Institute
Specialization: Stochastic models and ML in finance.
Potential Contribution: Bridges the gap between advanced stochastic modeling and machine learning applications, directly relevant to your project's core.

John C. Hull – University of Toronto
Specialization: Derivatives & ML in quantitative finance.
Potential Contribution: Provides foundational knowledge in quantitative finance and practical applications of ML, ensuring the project's relevance to market practice.

Blanka Horvath – Oxford Man Institute
Specialization: Rough volatility, deep hedging.
Potential Contribution: Guides the modeling of volatility dynamics and their impact on alpha signals, especially in complex numerical environments.

Stefan Zohren – Oxford Man Institute
Specialization: Deep learning, RL, NLP trading signals.
Potential Contribution: Expertise in deep learning architectures (including time-series Transformers), reinforcement learning for signal generation and portfolio allocation.

Qinkai Chen – École Polytechnique
Specialization: GNNs and realized multivariate volatility models.
Potential Contribution: Directly relevant for integrating Graph Neural Networks to capture inter-asset numerical relationships and their impact on volatility and returns.

Julien Guyon – École des Ponts ParisTech
Specialization: Embedding models & path-dependent volatility.
Potential Contribution: Expertise in creating effective embeddings from numerical data and understanding complex volatility structures, key for representation learning.

Zeda Xu – Carnegie Mellon University
Specialization: GARCH-informed neural networks (GINN).
Potential Contribution: Brings experience in blending statistical time-series models (e.g. GARCH) with neural networks, enhancing mathematical depth of numerical analysis

Achintya Gopal – Columbia / Bloomberg
Specialization: Generative models, latent variable modeling in quant finance.
Potential Contribution: Ideal for guiding the development of disentangled numerical representations and generative models for financial time series.

Magnus Wiese – University of Kaiserslautern
Specialization: Quantitative GANs & return modeling.
Potential Contribution: Expertise in generative models for numerical data, which can be used for robust signal generation or data augmentation.

Marcos López de Prado – Cornell / ADIA
Specialization: Systematic strategies & ML in quant finance.
Potential Contribution: Ensure the alpha signals are robust, reliable, and avoid common pitfalls like overfitting, crucial for practical systematic strategies.

Igor Halperin – NYU Tandon / Fidelity
Specialization: ML-based pricing & risk modeling.
Potential Contribution: Bridges deep learning with practical financial applications, helping to ensure the signals are relevant for pricing and risk management.

Andrew Ang – BlackRock & Columbia
Specialization: Factor investing, macro systematic strategies.
Potential Contribution: Provides strong understanding of traditional factor models and macro influences, essential for benchmarking/contextualizing numerical alpha signals.

Gordon Ritter – Columbia MFE
Specialization: Statistical ML & portfolio optimization.
Potential Contribution: Offers expertise in applying statistical machine learning to financial data and optimizing portfolios based on generated signals.

Rama Cont – University of Oxford
Specialization: Generative modeling and stochastic finance.
Potential Contribution: Deep expertise in stochastic processes and generative models, highly relevant for modeling complex numerical financial data.

Charles-Albert Lehalle – Imperial College London
Specialization: ML bias, stochastic control, order flow.
Potential Contribution: Can advise on mitigating biases in ML models applied to numerical data and understanding the impact of market microstructure.

Yada Zhu – MIT-IBM Watson AI Lab
Specialization: Graph models, time-series ML, privacy-preserving finance.
Potential Contribution: Direct expertise in graph models and time-series ML, which are central to your numerical project.

David Byrd – Bowdoin College
Specialization: Market-making ML and spoofing detection.
Potential Contribution: Offers insights into market microstructure and latent regime inference from numerical data, which can inform robust signal generation.

- Potential Publication & Conference Venues:

Category	Journal/ Conference	Focus
Top-Tier Academic Finance Journals	Journal of Financial Economics (JFE)	Top-tier finance journal with emphasis on theoretical and empirical asset pricing. Highly competitive.
	Journal of Finance (JF)	Flagship journal of the American Finance Association, covering all areas of financial economics. Extremely competitive.
	Journal of Financial and Quantitative Analysis (JFQA)	Leading academic finance journal for empirical and theoretical research in financial economics. Strong fit for quantitative methods.
	Review of Financial Studies (RFS)	Top-tier journal covering a broad range of financial topics, including asset pricing and quantitative methods.

Category	Journal/ Conference	Focus
Leading Applied Finance & Quantitative Finance Journals	Journal of Financial Data Science (JFDS)	Perfect fit for applied AI/ML methods in quantitative finance, especially those dealing with novel data sources and computational techniques.
	International Review of Financial Analysis (IRFA)	Strong applied finance focus, often publishing empirical studies and quantitative analyses. (Impact Factor ≈ 9.8)
	Journal of Banking & Finance (JBF)	Broad finance scope, ideal for empirical studies and market microstructure analysis. Good for projects with practical implications.
	Quantitative Finance	Dedicated to mathematical and computational methods in finance. Strong fit for the "deep mathematical analysis" aspect.
	Journal of Investment Management (JIM)	Bridges academic research and practical investment management. Good for projects with clear implications for alpha generation.
Machine Learning & AI Conferences/ Workshops	NeurIPS / ICML / ICLR Workshops (e.g., ML4Finance, SSL, Causal ML)	Cutting-edge AI-finance intersection. Workshops are excellent for presenting early-stage or focused research, getting feedback, and networking.
	AAAI / IJCAI Workshops (e.g., AI in Finance)	Similar to NeurIPS/ICML workshops, but under the umbrella of broader AI conferences.
	KDD Workshops (e.g., FinKDD)	Data mining and knowledge discovery, with specific workshops for finance applications. Good for projects with novel data processing or feature engineering.

Category	Journal/ Conference	Focus
Specialized Quantitative Finance Conferences	Global Derivatives / QuantMinds / RiskMinds	Practitioner-oriented conferences covering quantitative finance, risk management, and derivatives. Excellent for presenting practical applications and getting industry feedback.
	IAQF (International Association for Quantitative Finance) Annual Conference	Brings together academics and practitioners in quantitative finance.
	CQF Institute Research Seminars / Webinars	While not a traditional publication, presenting here can gain visibility within the quant community.
Pre-print Servers	arXiv (q-fin.CP / cs.LG)	Pre-print server for quantitative finance (Computational Finance) and Machine Learning (Learning). Widely read by academics and practitioners.
	SSRN (Social Science Research Network)	Pre-print archive for finance/economics research. Also widely read.
Academic Conferences	American Finance Association (AFA) Annual Meeting	The largest and most prestigious finance conference. Highly competitive for full paper presentations, but poster sessions or related society meetings might be an option.
	Financial Management Association (FMA) Annual Meeting	Another large finance conference with various tracks, including quantitative finance.
	SIAM Journal on Financial Mathematics (SIAM JFM)	If your "deep mathematical analysis" leads to significant theoretical contributions or novel numerical methods, this journal could be a target.

- ## Notes for Publication:

1. **Focus on Core Quantitative Finance**: Many foundational and cutting-edge quantitative finance research areas are deeply rooted in time-series analysis, econometric modeling, and the mathematical properties of financial data. By emphasizing these aspects, the research aligns with a long-standing tradition of rigorous academic inquiry in finance, enhancing its credibility and relevance in the field.

2. **Novelty in Numerical Methods**: While textual analysis in finance is a growing field, there remains immense scope for novelty in applying advanced mathematical and machine learning techniques, such as sophisticated Graph Neural Networks (GNNs), advanced time-series models, and causal inference on numerical data. The "Very High" innovation/novelty score for "Alpha Signal Discovery" and "Cross-Sectional Return Prediction," which are heavily numerical and graph-based, indicates a strong potential for impactful contributions to the literature.

3. **Addressing Fundamental Challenges**: The research tackles fundamental challenges such as robust signal extraction from noisy time series, understanding complex inter-asset relationships through graphs, and ensuring model generalization across varying market regimes. A deep mathematical approach to these problems is inherently valuable and contributes to the advancement of quantitative finance.

4. **Interpretability and Explainability**: A focus on disentangled numerical representations can lead to more interpretable alpha signals, which is a significant area of research in both academia and industry. Understanding why a signal works is often as important as its effectiveness, making this aspect crucial for both practical applications and academic scrutiny.

5. **Avoids Data Licensing Hurdles**: High-quality textual data can be expensive and difficult to obtain. By concentrating on numerical data, much of which is publicly available or accessible via academic licenses, the research becomes more feasible and reproducible. This accessibility is a significant advantage for academic soundness and enhances the potential for broader dissemination of findings.

6. **Stronger Link to Econometrics**: A deeper dive into numerical analysis naturally incorporates more econometric principles, which are highly valued in academic finance for their statistical rigor and theoretical grounding. This connection not only strengthens the research's academic foundation but also enhances its appeal to a wider audience within the finance community.