Generative AI and Knowledge-Informed Deep Learning

for Asset Pricing and Investment Management

Lin William Cong[1]

[1] NTU CCDS & NBS, ABFER, CEPR, & NBER

June 15, 2026

Core Themes of Modern AI (Beyond Basic ML and LLMs)

- AI (McCarthy 1955): technologies allowing machines to perform complex tasks that typically require **human intelligence**.
- Human Intelligence (HAI): the ability to learn and perform suitable techniques to solve problems and **achieve goals**, in an uncertain, ever-varying world.
- An economist's perspective of modern AI:
  1. Instruction-driven automation/prediction → end-to-end goal-oriented optimization in enlarged modeling space.
  2. Expertise/theory-driven/reduced-form discriminative modeling → data-driven generative modeling enabling pre-trained foundational models/agents.
- The Tao/Ways of AI:
  1. Larger models, greater computation, general intelligence? Vertical (not-so-) large models for environment and grouped heterogeneity.
  2. Bigger data and associated computation; generating/augmenting data.
  3. General goal-oriented optimization (e.g., RL & Panel Trees) in large spaces.
  4. Data-driven generative counterfactual through large environments and agents. AI agents for experimentation as new/representative subjects/species in data-driven equilibria.

Core Themes of Modern AI (Beyond Basic ML and LLMs)

- AI (McCarthy 1955): technologies allowing machines to perform complex tasks that typically require **human intelligence**.
- Human Intelligence (HAI): the ability to learn and perform suitable techniques to solve problems and **achieve goals**, in an uncertain, ever-varying world.
- An economist's perspective of modern AI:
    1. Instruction-driven automation/prediction → end-to-end goal-oriented optimization in enlarged modeling space.
    2. Expertise/theory-driven/reduced-form discriminative modeling → data-driven generative modeling enabling pre-trained foundational models/agents.
- The Tao/Ways of AI:
    1. Larger models, greater computation, general intelligence? Vertical (not-so-) large models for environment and grouped heterogeneity.
    2. Bigger data and associated computation; generating/augmenting data.
    3. General goal-oriented optimization (e.g., RL & Panel Trees) in large spaces.
    4. Data-driven generative counterfactual through large environments and agents. AI agents for experimentation as new/representative subjects/species in data-driven equilibria.
- AI for Social Sciences (e.g., Finance): (off-the-shelf) ML + text analytics
    - Return prediction, dim. redux, alt. data, etc., trained through examples.
    - Text analytics/LLMs; customized language models (Hoberg & Manela, 2025).
    - Causal ML, Causal Human+Machine Learning (e.g., Cong & Yang, 2025)

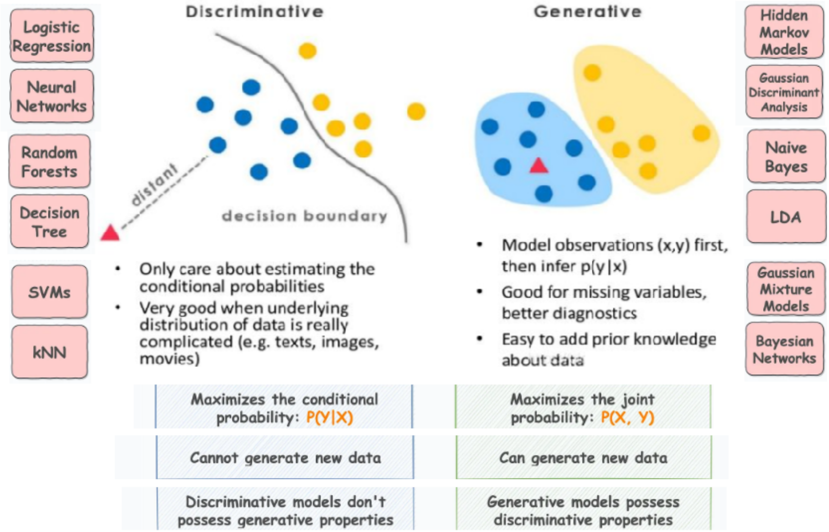
Generative AI for Finance

- What is special about economics/finance research?

Generative AI for Finance

- What is special about economics/finance research?
- I. Complexity, Heterogeneity, and Purpose-Driven
    1. Probabilistic, context-specific (heterogeneous), mergent interactions.
    2. Low signal-to-noise, non-stationary, variable, adaptive, etc.
    3. Value-laden with high ethical and systemic risks.
    4. (Hetergoeneous) goal-oriented; stakeholder interactions and agencies.
- II. Methods and Approach
    1. Multi-modal, qualitative, unstructured, not strictly ordered, non-systematic.
    2. Interpretation and causality.
    3. What-if questions. The future of economics lies not in explaining the past or predicting the world, but in re-imagining and generating the future.
- III. Data limitations:
    1. Scarce (mostly observational), missing, etc.
    2. Privacy-sensitive.
    3. Endogenous and time-varying data generation.
- Limitations of extant generative modeling in economics or social sciences (e.g., DSGE models, Computable General Equilibrium Models, Structural Equation Models, Agent-Based models).

Goal-Oriented Algorithms in Large (Action) Spaces for Optimizations



Discriminative

- Only care about estimating the conditional probabilities
- Very good when underlying distribution of data is really complicated (e.g. texts, images, movies)

Generative

- Model observations (x,y) first, then infer p(y|x)
- Good for missing variables, better diagnostics
- Easy to add prior knowledge about data

Discriminative	Generative
Maximizes the conditional probability: P(Y|X)	Maximizes the joint probability: P(X, Y)
Cannot generate new data	Can generate new data
Discriminative models don't possess generative properties	Generative models possess discriminative properties

Generative Modeling Beyond LLMs

- "Big Four" Families of Generative Models
    1. Variational Autoencoders (VAEs)
    2. Adversarial Networks (GANs)
    3. Autoregressive Models
    4. Diffusion Models / Score-Based Models

Generative Modeling Beyond LLMs

- “Big Four” Families of Generative Models
    1. Variational Autoencoders (VAEs)
    2. Adversarial Networks (GANs)
    3. Autoregressive Models
    4. Diffusion Models / Score-Based Models
    5. Others: Normalizing Flows, Energy-Based Models, etc.

Generative Modeling Beyond LLMs

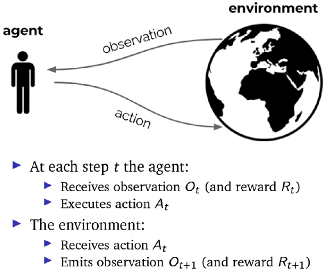
- "Big Four" Families of Generative Models
  1. Variational Autoencoders (VAEs)
  2. Adversarial Networks (GANs)
  3. Autoregressive Models
  4. Diffusion Models / Score-Based Models
  5. Others: Normalizing Flows, Energy-Based Models, etc.
- Types of generative models:
  - **An unconditional generative model** learns the overall data distribution $p(x)$ without relying on any additional input or condition. Its primary objective is to generate samples that resemble the entire data set.
  - **A conditional generative modeling** incorporate extra information or conditions. It learns the conditional distribution $p(x \mid y)$, where y can be any auxiliary information (e.g., class labels, attributes).
- GOALS constitute a new type of generative model beyond LLMs.

Generative Modeling Beyond LLMs

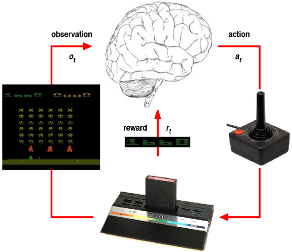
- "Big Four" Families of Generative Models
  1. Variational Autoencoders (VAEs)
  2. Adversarial Networks (GANs)
  3. Autoregressive Models
  4. Diffusion Models / Score-Based Models
  5. Others: Normalizing Flows, Energy-Based Models, etc.
- Types of generative models:
  - **An unconditional generative model** learns the overall data distribution $p(x)$ without relying on any additional input or condition. Its primary objective is to generate samples that resemble the entire data set.
  - **A conditional generative modeling** incorporate extra information or conditions. It learns the conditional distribution $p(x \mid y)$, where y can be any auxiliary information (e.g., class labels, attributes).
- GOALS constitute a new type of generative model beyond LLMs.
- Hallucination or creativity?
  Robust control balances creativity and reliability.

Reinforcement Learning as Efficient Heuristic Search

- Economically guided heuristic search in a large decision/action space.
- People learn by interacting with the environment in an active and sequential way, to optimize some **rewards**.
- Direct construction of portfolios with flexible objectives:
  - “**AlphaPortfolio: Goal-Oriented Investment Management Through Deep Reinforcement Learning**” (Cong, Tang, & Wang, 2026 JFE).
- An alternative framework to studying CF using DL, RL & IRL:
  - “**AlphaManager: A Data-Driven-Robust-Control Approach to Corporate Finance**” (Campello, Cong, & Zhou, 2022).



- At each step t the agent:
  - Receives observation O_t (and reward R_t)
  - Executes action A_t
- The environment:
  - Receives action A_t
  - Emits observation O_{t+1} (and reward R_{t+1})



AlphaPortfolio: Goal-Oriented Investment Management

Through Deep

Reinforcement Learning

Lin William Cong, Nanyang Technological University
Ke Tang, Tsinghua University
Jingyuan Wang, Beihang University

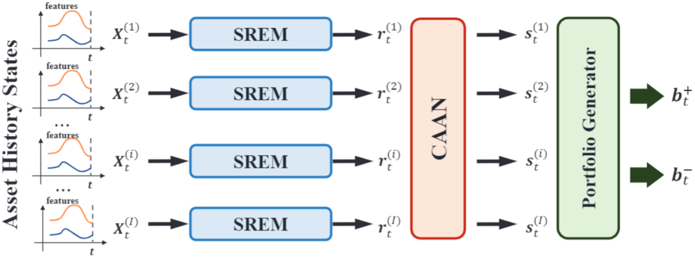
Goal-Oriented Portfolio Management Through Transformer-Based RL

- First “large” model in finance (a few million parameters) with Transformer/offline DRL in 2019.
    - Virtual of complexity/large models (Belkin et al., 2019, Kelly, Malamud, & Zhou, 2022, Fan et al., 2022).
    - Share how humans learn the spirit of Grounded Theory ($> 20,000$ citations).
    - Non-textual, proprietary, but not too large.
- Advantages:
    - Anything objective that can be written as a cumulative reward function; not just utility over terminal wealth.
    - High-dimensional and nonlinear data approximated by deep neural nets instead of linear functions or index functions.
    - Action-state interaction, partial/non-equilibrium, unknown states and payoffs.
    - Cutting-edge sequence modeling (i.e., Transformer) plus cross-sectional information extraction through attention mechanisms.
- Bridge between engineering, AI architecture, and finance.
- RL uniquely fit for portfolio management problems.

Economic Question and AI for Finance

- DRL comes to the rescue:
  - Anything objective that can be written as a cumulative reward function; not just utility over terminal wealth.
  - High-dimensional and nonlinear data approximated by deep neural nets instead of linear functions or index functions.
  - Action-state interaction, partial/non-equilibrium, unknown states and payoffs.
  - Cutting-edge sequence modeling (i.e., Transformer) plus cross-sectional information extraction through attention mechanisms.

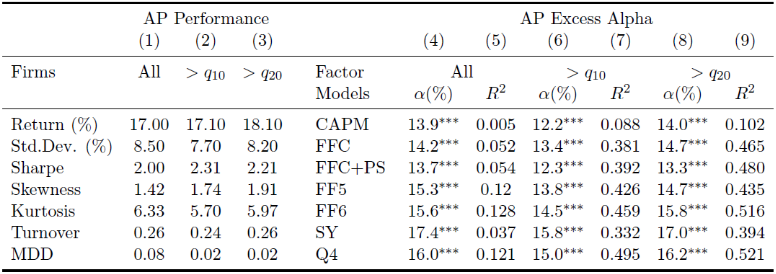
Architecture Innovation: Transformer + Cross-Asset Attention Network + RL



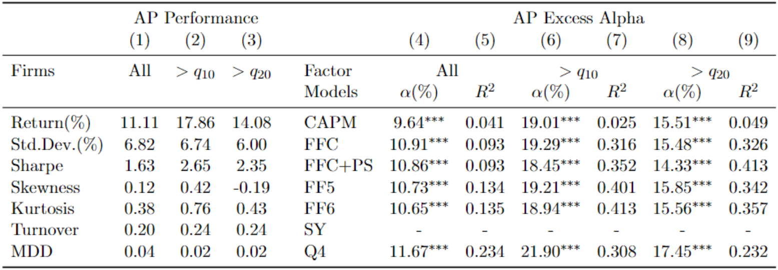
Application to U.S. Equities

- CRSP, Compustat, EDGAR, etc.
- July 1965-June 2016: pre-1990, 1990-2016, post 2001
- 51 characteristics/signals with 1-12 month lags, similar to (Freyberger et.al., 2019).
    - Past-return based.
    - Investment-related characteristics: annual percentage change in total assets (**Investment**),change in inventory over total assets (**IVC**),
    - Profitability-related characteristics such as gross profitability over the book-value of equity (**Prof**) or return on operating assets (**ROA**),
    - Intangibles such as operating accruals (**OA**) and tangibility (**Tan**)
    - Value-related characteristics such as the book-to-market ratio (**BEME**) and earnings-to-price (**E2P**)
    - Trading frictions such as the average daily bid-ask spread (**Spread**) and standard unexplained volume (**SUV**).
- Macro variables and alternative data: MD&A (10-K & 10-Q), Risk Factor Discussions (1A of 10-K), Analyst Reports, etc.

AlphaPortfolio Performance on Test Samples

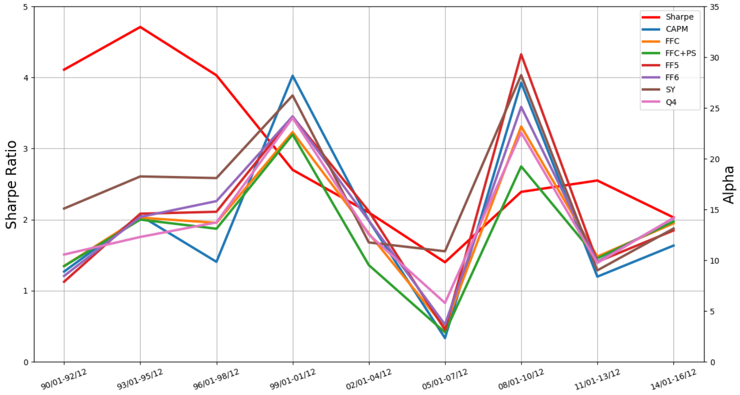


	AP Performance				AP Excess Alpha					
	(1)	(2)	(3)		(4)	(5)	(6)	(7)	(8)	(9)
Firms	All	$> q_{10}$	$> q_{20}$	Factor Models	All $\alpha(\%)$	R^2	$> q_{10}$ $\alpha(\%)$	R^2	$> q_{20}$ $\alpha(\%)$	R^2
Return (%)	17.00	17.10	18.10	CAPM	13.9***	0.005	12.2***	0.088	14.0***	0.102
Std.Dev. (%)	8.50	7.70	8.20	FFC	14.2***	0.052	13.4***	0.381	14.7***	0.465
Sharpe	2.00	2.31	2.21	FFC+PS	13.7***	0.054	12.3***	0.392	13.3***	0.480
Skewness	1.42	1.74	1.91	FF5	15.3***	0.12	13.8***	0.426	14.7***	0.435
Kurtosis	6.33	5.70	5.97	FF6	15.6***	0.128	14.5***	0.459	15.8***	0.516
Turnover	0.26	0.24	0.26	SY	17.4***	0.037	15.8***	0.332	17.0***	0.394
MDD	0.08	0.02	0.02	Q4	16.0***	0.121	15.0***	0.495	16.2***	0.521

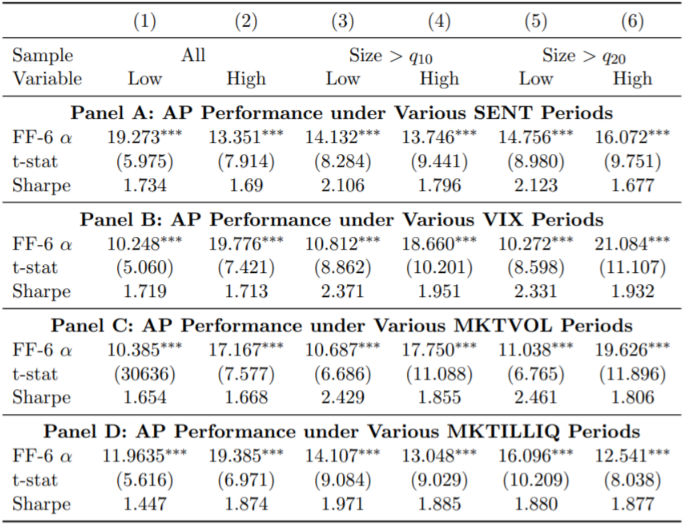


	AP Performance				AP Excess Alpha					
	(1)	(2)	(3)		(4)	(5)	(6)	(7)	(8)	(9)
Firms	All	$> q_{10}$	$> q_{20}$	Factor Models	All $\alpha(\%)$	R^2	$> q_{10}$ $\alpha(\%)$	R^2	$> q_{20}$ $\alpha(\%)$	R^2
Return(%)	11.11	17.86	14.08	CAPM	9.64***	0.041	19.01***	0.025	15.51***	0.049
Std.Dev.(%)	6.82	6.74	6.00	FFC	10.91***	0.093	19.29***	0.316	15.48***	0.326
Sharpe	1.63	2.65	2.35	FFC+PS	10.86***	0.093	18.45***	0.352	14.33***	0.413
Skewness	0.12	0.42	-0.19	FF5	10.73***	0.134	19.21***	0.401	15.85***	0.342
Kurtosis	0.38	0.76	0.43	FF6	10.65***	0.135	18.94***	0.413	15.56***	0.357
Turnover	0.20	0.24	0.24	SY	-	-	-	-	-	-
MDD	0.04	0.02	0.02	Q4	11.67***	0.234	21.90***	0.308	17.45***	0.232

Trends of AlphaPortfolio Performance

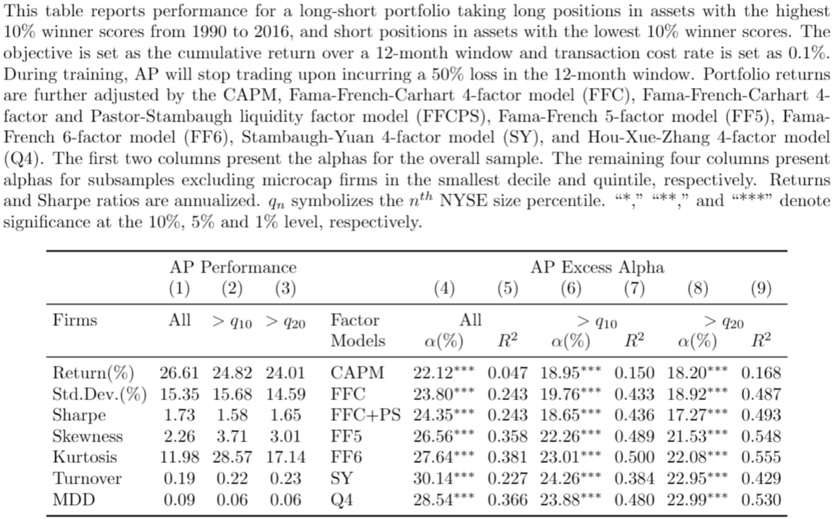


Out-of-Sample Performance under Various Market Conditions



	(1)	(2)	(3)	(4)	(5)	(6)
Sample	All		Size > q_{10}		Size > q_{20}	
Variable	Low	High	Low	High	Low	High
Panel A: AP Performance under Various SENT Periods						
FF-6 α	19.273***	13.351***	14.132***	13.746***	14.756***	16.072***
t-stat	(5.975)	(7.914)	(8.284)	(9.441)	(8.980)	(9.751)
Sharpe	1.734	1.69	2.106	1.796	2.123	1.677
Panel B: AP Performance under Various VIX Periods						
FF-6 α	10.248***	19.776***	10.812***	18.660***	10.272***	21.084***
t-stat	(5.060)	(7.421)	(8.862)	(10.201)	(8.598)	(11.107)
Sharpe	1.719	1.713	2.371	1.951	2.331	1.932
Panel C: AP Performance under Various MKTVOL Periods						
FF-6 α	10.385***	17.167***	10.687***	17.750***	11.038***	19.626***
t-stat	(30636)	(7.577)	(6.686)	(11.088)	(6.765)	(11.896)
Sharpe	1.654	1.668	2.429	1.855	2.461	1.806
Panel D: AP Performance under Various MKTILLIQ Periods						
FF-6 α	11.9635***	19.385***	14.107***	13.048***	16.096***	12.541***
t-stat	(5.616)	(6.971)	(9.084)	(9.029)	(10.209)	(8.038)
Sharpe	1.447	1.874	1.971	1.885	1.880	1.877

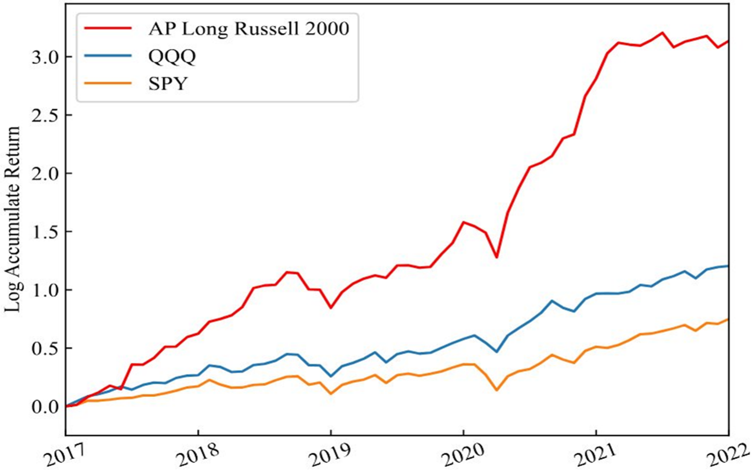
Out-of-Sample Performance of AP Fund



This table reports performance for a long-short portfolio taking long positions in assets with the highest 10% winner scores from 1990 to 2016, and short positions in assets with the lowest 10% winner scores. The objective is set as the cumulative return over a 12-month window and transaction cost rate is set as 0.1%. During training, AP will stop trading upon incurring a 50% loss in the 12-month window. Portfolio returns are further adjusted by the CAPM, Fama-French-Carhart 4-factor model (FFC), Fama-French-Carhart 4-factor and Pastor-Stambaugh liquidity factor model (FFCPS), Fama-French 5-factor model (FF5), Fama-French 6-factor model (FF6), Stambaugh-Yuan 4-factor model (SY), and Hou-Xue-Zhang 4-factor model (Q4). The first two columns present the alphas for the overall sample. The remaining four columns present alphas for subsamples excluding microcap firms in the smallest decile and quintile, respectively. Returns and Sharpe ratios are annualized. q_n symbolizes the n^{th} NYSE size percentile. "*," "**," and "***" denote significance at the 10%, 5% and 1% level, respectively.

	AP Performance				AP Excess Alpha					
	(1)	(2)	(3)		(4)	(5)	(6)	(7)	(8)	(9)
Firms	All	$> q_{10}$	$> q_{20}$	Factor Models	All $\alpha(\%)$	R^2	$> q_{10}$ $\alpha(\%)$	R^2	$> q_{20}$ $\alpha(\%)$	R^2
Return(%)	26.61	24.82	24.01	CAPM	22.12***	0.047	18.95***	0.150	18.20***	0.168
Std.Dev.(%)	15.35	15.68	14.59	FFC	23.80***	0.243	19.76***	0.433	18.92***	0.487
Sharpe	1.73	1.58	1.65	FFC+PS	24.35***	0.243	18.65***	0.436	17.27***	0.493
Skewness	2.26	3.71	3.01	FF5	26.56***	0.358	22.26***	0.489	21.53***	0.548
Kurtosis	11.98	28.57	17.14	FF6	27.64***	0.381	23.01***	0.500	22.08***	0.555
Turnover	0.19	0.22	0.23	SY	30.14***	0.227	24.26***	0.384	22.95***	0.429
MDD	0.09	0.06	0.06	Q4	28.54***	0.366	23.88***	0.480	22.99***	0.530

Practical Performance and Mass Customization



Panel Trees as Interpretable & Effective Greedy Search

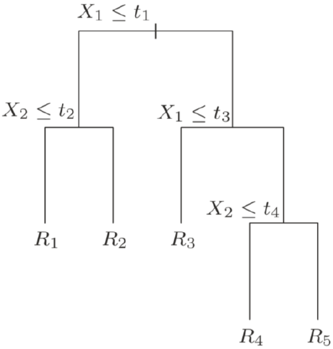
- Interpretable and not-so-large AI models (not supervised/unsupervised).
- Human intelligence and *The Art of War*: divide and conquer.
- Conventional trees not designed for panel data, economic objectives, and interpretability comes at the cost of overfitting.
- Panel trees: panel data analytics with global split criteria using economic goals instead of local split criteria based on statistical error minimization.

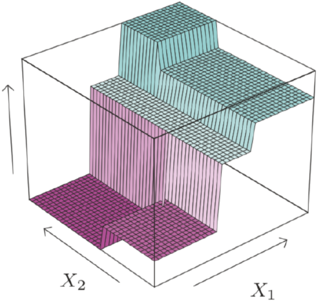
Panel Trees as Interpretable & Effective Greedy Search

- Interpretable and not-so-large AI models (not supervised/unsupervised).
- Human intelligence and *The Art of War*: divide and conquer.
- Conventional trees not designed for panel data, economic objectives, and interpretability comes at the cost of overfitting.
- Panel trees: panel data analytics with global split criteria using economic goals instead of local split criteria based on statistical error minimization.
- Creating test assets and latent factor models:
  - "**Growing the Efficient Frontier on Panel Trees**" (Cong, Feng, He, & He, 2021); JFE forthcoming.
- Uncommon factors and macro regimes for asset pricing:
  - "**Sparse Modeling Under Grouped Heterogeneity with an Application to Asset Pricing**" (Cong, Feng, He, & Li, 2022).
- Heterogeneous predictability and trading:
  - "**Mosaics of Predictability**" (Cong, Feng, He, & Wang, 2024).

Traditional Regression Trees: Intuition

Hierarchical: use less and less data → overfit.

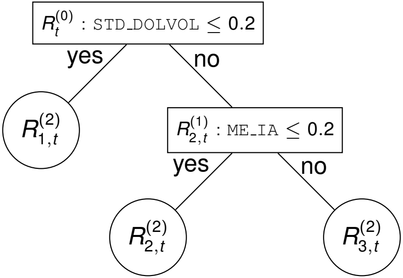




Growing the Efficient Frontier on Panel Trees

Lin William Cong, Cornell University
Gavin Feng, City University of Hong Kong
Jingyu He, City University of Hong Kong
Xin He, University of Science and Technology of China

P-Tree Test Assets and Factor Models



- The second split gives us three leaf basis portfolios and a updated SDF:

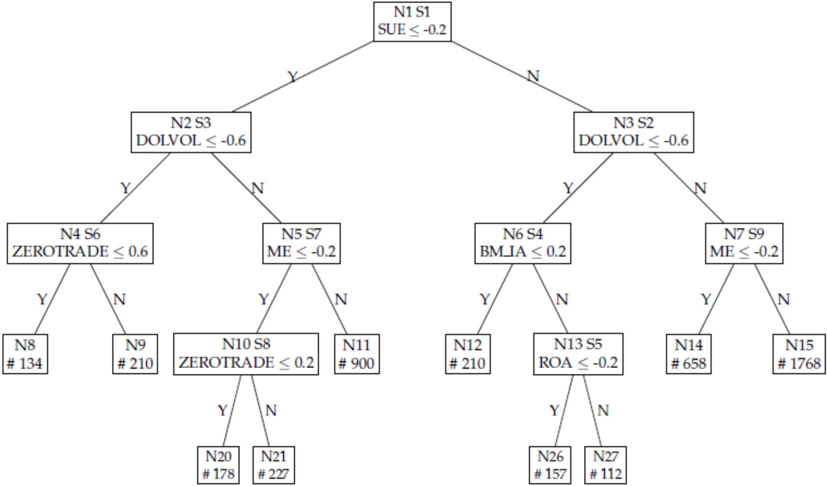
$$f_t^{(2)} = \widehat{\Sigma}_2^{-1} \widehat{\mu}_2 R_t^{(2)} = w_{21} R_{1,t}^{(2)} + w_{22} R_{2,t}^{(2)} + w_{23} R_{3,t}^{(2)},$$

- For the second split, the algorithm searches over all leaf nodes, characteristics, and breakpoints.
- The split criterion is calculated based on the entire cross section, thus P-Tree and its SDF are global.

Data for U.S. Equities

- 1981-2020 monthly observation for US equities
- Returns and lag-one-month characteristics
- Standardize the characteristics in the cross-section into Uniform $[-1, 1]$
- 61 characteristics in 6 categories: momentum, value-versus-growth, investment, profitability, intangibles, and frictions
- Periods 1981-2000 and 2001-2020 as training and test samples.

A P-Tree Grown on U.S. Equities



A Grown P-Tree

	SR	α_{CAPM}	α_{FF5}	α_{Q5}	α_{RP5}	α_{IP5}	SR	α_{CAPM}	α_{FF5}	α_{Q5}	α_{RP5}	α_{IP5}
	Panel B1: 20 Years In-Sample (1981-2000)						Panel B2: 20 Years Out-of-Sample (2001-2020)					
P-Tree1	7.12	1.85	1.88	1.71	1.64	1.24	3.24	1.34	1.31	1.23	1.30	0.91
P-Tree1-5	12.14	0.80	0.79	0.77	0.75	0.71	2.98	0.52	0.47	0.49	0.48	0.32
P-Tree1-10	19.16	0.64	0.65	0.62	0.61	0.57	2.57	0.36	0.29	0.33	0.34	0.19
P-Tree1-15	27.74	0.53	0.53	0.52	0.52	0.49	2.48	0.30	0.23	0.27	0.27	0.16
P-Tree1-20	41.12	0.47	0.47	0.46	0.46	0.45	2.57	0.26	0.20	0.24	0.25	0.15
	Panel C1: 20 Years In-Sample (2001-2020)						Panel C2: 20 Years Out-of-Sample (1981-2000)					
P-Tree1	5.82	1.50	1.46	1.50	1.37	1.36	4.35	1.50	1.59	1.35	1.43	1.10
P-Tree1-5	9.61	0.66	0.65	0.66	0.64	0.62	4.11	0.57	0.71	0.51	0.60	0.30
P-Tree1-10	14.24	0.46	0.46	0.46	0.45	0.45	4.07	0.40	0.49	0.35	0.44	0.23
P-Tree1-15	20.52	0.38	0.38	0.38	0.37	0.37	3.98	0.36	0.45	0.31	0.40	0.18
P-Tree1-20	26.95	0.34	0.34	0.34	0.33	0.33	3.82	0.31	0.40	0.26	0.35	0.14

Uncommon Factors and Asset Heterogeneity in the Cross Section and Time Series

Lin William Cong, Nanyang Technological University
Gavin Feng, City University of Hong Kong
Jingyu He, City University of Hong Kong
Junye Li, Fudan University

Empirical Asset Pricing Involving Factor Models

- Factor models for explaining cross-sectional return dynamics.
  - Well-known (risk/observable) factors: Market, Size, Value, Momentum, $\cdots$
- Search for ‘true/universal (factor) model not rejected by AP tests.
  - Test assets matter.
  - E.g., FF explains 5×5 ME-B/M portfolios, but now small-growth.
- Challenges/bottlenecks for current approaches:
  - Fishing for more missing factors?
  - Factor selection in a factor zoo?
  - Estimation of unconditional vs. conditional models?
  - Choices for test assets? Weak factors?
  - Some assets are simply mispriced?

(Futile?) Search for a Universal Model

- Einstein's quest for the "Theory of Everything."
- Sparsity for tractability and interpretation, but not necessarily universal models.
- Standard factor modeling in empirical asset pricing:

$$\begin{aligned} r_{1,t} &= \alpha_{1,t} + \beta_{1,1,t} f_{1,t} + \cdots + \beta_{1,k,t} f_{k,t} + \epsilon_{1,t} \\ &\vdots \\ r_{n,t} &= \alpha_{n,t} + \beta_{n,1,t} f_{1,t} + \cdots + \beta_{n,k,t} f_{k,t} + \epsilon_{n,t} \end{aligned}$$

  - LHS observations/assets are heterogeneous; grouped heterogeneity.
  - Burden all on RHS model estimation and selection.
  - Sorting/test asset construction for common factors & cross-cluster spread.
- *Joint* observation clustering and heterogeneous model selection:
  - Model selection on RHS: observations following one common factor model.
  - Observation clustering on LHS: split the panel allowing uncommon factors.
  - Data-driven, economically guided tree method preserving interpretability and guarding against overfitting.

Factor Selection in BCM Leaf Clusters

Prob.	N8	N18	N19		N48	N49	N50	N51	N52	N53	N54	N55
	Panel A: I(SVAR $\leq$ -0.2)I(ME $\leq$ 0.2)				Panel C: I(SVAR $>$ -0.2)I(SVAR $\leq$ 0.6)							
$>$ 0.95	MKTRF IVOL BETA	MKTRF IVOL BETA BAB SUE UMD STR	MKTRF IVOL BETA STR HML RMW SMB UMD		MKTRF STR IVOL	MKTRF STR HMLM	MKTRF IVOL	MKTRF IVOL ROE STR BETA SMB	MKTRF STR BETA SMB UMD HML RMW IA	MKTRF SMB BAB	BETA STR IVOL IMD	BETA STR IVOL UMD MKTRF BAB
$>$ 0.9	LTR	LTR						SUE				LTR
$>$ 0.7	BAB SUE					BETA BAB SMB	UMD	UMD		BETA	SMB	
In-Sample α	-0.36	0.36	0.37		-0.07	0.20	0.41	0.03	0.21	0.07	0.91	-0.10
Out-of-Sample α	0.45	0.33	0.14		1.12	0.18	0.38	0.08	0.08	-0.17	-0.13	-0.09
Prob.	N20	N21	N22	N23			N56	N57	N29	N30	N62	N63
	Panel B: I(SVAR $\leq$ -0.2)I(ME $>$ 0.2)						Panel D: I(SVAR $>$ -0.2)I(SVAR $>$ 0.6)					
$>$ 0.95	MKTRF IVOL HML SMB UMD	MKTRF HML SMB STR BETA BAB ROE HMLM	MKTRF SMB UMD	MKTRF SMB HMLM LIQ RMW ROE STR				UMD MKTRF SMB	MKTRF IVOL	MKTRF IVOL STR	MKTRF IVOL SMB	IVOL SMB STR QMJ
$>$ 0.9												
$>$ 0.7			BAB	REG						NI	NI	MKTRF
In-Sample α	0.07	-0.22	0.12	-0.24			1.61	-1.05	0.25	1.20	-0.88	-0.85
Out-of-Sample α	0.12	0.14	-0.02	-0.10			1.68	0.02	0.38	0.38	-0.14	0.27

Common and Uncommon Factors

- Common factors:
  - MKTRF has been selected with a probability of 100% in most leaf clusters except for N54, N56, and N63.
  - Size (SMB) is also important, selected with probability $>$ 70% in 13 out of 21 clusters.
  - STR and BETA are important, $>$ 70% in 12 and 10 clusters.
  - Value (HML) is $>$ 70% only for 4 out of 21 clusters.
- Odd-ball out: What makes N56 so special? No factor is selected!
  - Stocks with highest SVAR and lowest EP, MOM, and ME, basically all junk stocks.
- Factor usefulness and grouped heterogeneity in disguise.
- Evaluate Economic Fitness by Leaf Clusters: Agnostic performs slightly better; N21, N30, N54 have $R^2 > 24\%$ OOS.

Mosaics of Predictability

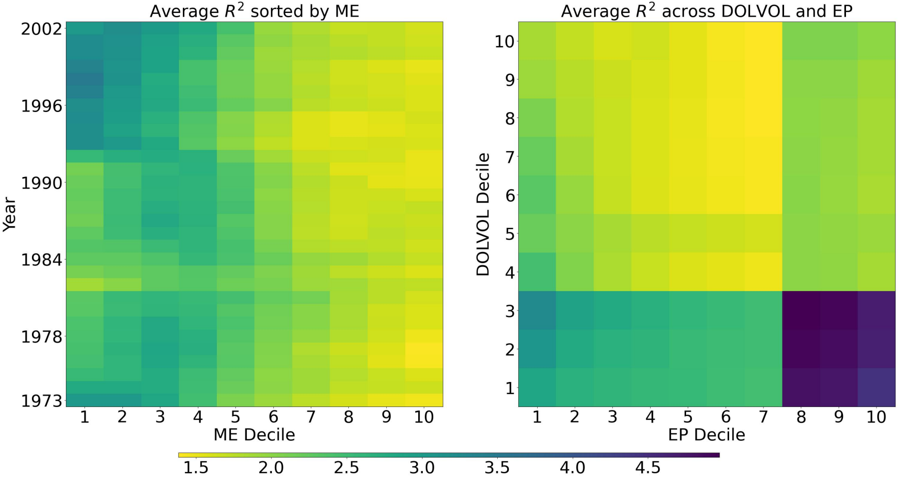
Lin William Cong, Nanyang Technological University
Gavin Feng, City University of Hong Kong
Jingyu He, City University of Hong Kong
Yuanzhi Wang, City University of Hong Kong

Goal-Oriented Clusters of Differential Return Predictability

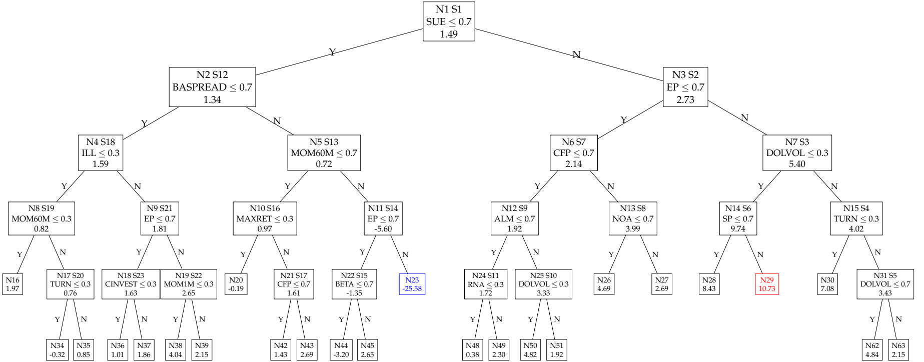
- Return predictability taken as attribute of predictors or models; but is unobervable, heterogeneous, and time-varying.
- However, asset return predictability is:
  - Unobservable, and not well-defined, e.g., anomaly average return, predictor significance, out-of-sample R^2, forecast-implied portfolios.
  - Potentially heterogeneous across assets, e.g., small-cap, distressed stocks are more predictable (Avramov, Cheng, and Metzker, 2023).
  - Potentially time-varying (e.g., Henkel et al., 2011).
- Goal-oriented tree-based clustering and heterogeneity modeling:
  - Measure predictability as signal-to-noise ratio $R^2_{i,t} = 1 - \frac{\text{Var}(\epsilon_{i,t})}{\text{Var}(r_{i,t})}$.
  - Grouped heterogeneity strikes the balance: Hard to estimate $R^2_{i,t}$ directly for each i and t; running a panel model induces a single R^2 but ignores heterogeneity.
  - Data-driven asset groups maximizing differences in predictability.
  - Interpretable based on firm characteristics and aggregate predictors.

Goal-Oriented Clusters of Differential Return Predictability

Return predictability taken as attribute of predictors or models; but is unobervable, heterogeneous, and time-varying.



Cross-Sectional Tree-based Clusters



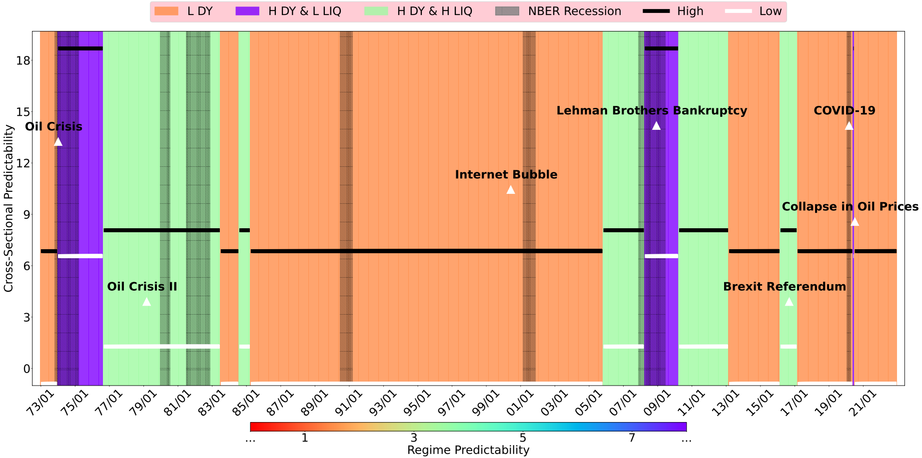
Highly Predictable: N29 (10.73%):

$\mathbb{1}\{\text{SUE} > 0.7\}\mathbb{1}\{\text{EP} > 0.7\}\mathbb{1}\{\text{DOLVOL} \leq 0.3\}\mathbb{1}\{\text{SP} > 0.7\}$

Less Predictable: N23 (-25.58%):

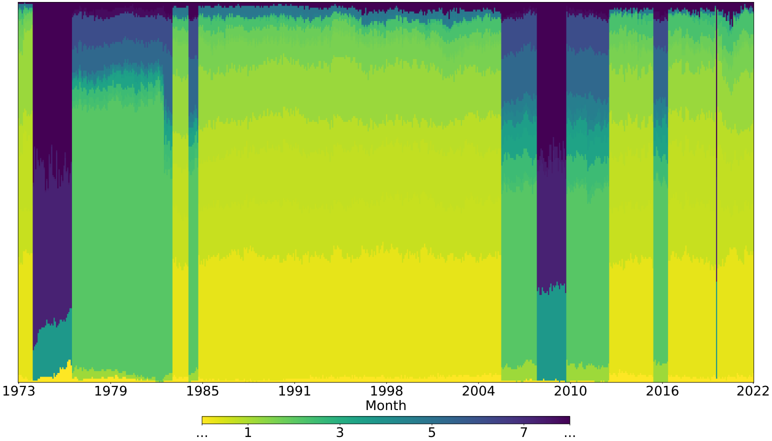
$\mathbb{1}\{\text{SUE} \leq 0.7\}\mathbb{1}\{\text{BASPREAD} > 0.7\}\mathbb{1}\{\text{MOM60M} > 0.7\}\mathbb{1}\{\text{EP} > 0.7\}$

Time-Varying Predictability under Market Regimes



- S&P 500 dividend yield (DY) and Pastor-Stambaugh liquidity (LIQ).
- Time-series partitions display larger predictability heterogeneity (color bar).
- Further cross-sections enlarge the gaps (y-axis).
- Numerous events trigger regime changes (e.g., Oil Crisis, COVID-19).

Mosaics of Predictability by Calendar Months



- S&P 500 dividend yield and Pastor-Stambaugh liquidity predicts regime switches.
- Some regimes correspond to events (Oil Crisis, COVID-19, etc.).

Model Disagreement Anomaly & Predictability-Based Investment:

Heterogeneity disagreement and mispricing $R^2_{\text{CMG},j} = R^2_{\text{C},j} - R^2_{\text{G},j}$
Large and significant OOS Alphas, dominated by long leg.

	1973 - 2002 (in-sample)			2003 - 2022 (out-of-sample)		
	T3	B3	T3 - B3	T3	B3	T3 - B3
		Panel A: Performance				
Avg (%)	2.75	0.40	2.35	2.10	1.07	1.03
Ann. SR	1.56	0.25	2.33	1.18	0.52	0.75
		Panel B: Unexplained monthly alphas (%)				
CAPM	2.35***	-0.01	2.35***	1.18***	0.03	1.15***
FF3	1.94***	-0.32**	2.26***	1.22***	0.07	1.14***
FF5	1.89***	-0.37***	2.26***	1.31***	0.11	1.20***
FF5+MOM+IVOL	2.09***	-0.24*	2.33***	1.39***	0.19	1.20***
Q5	2.19***	-0.05	2.24***	1.43***	0.09	1.34***
BS6	1.84***	-0.32**	2.15***	1.37***	0.10	1.26***
DHS3	2.68***	0.17	2.51***	1.44***	0.26	1.19***
SY4	2.05***	-0.26*	2.31***	1.89***	0.46	1.43***

Bridging Structured Knowledge and Data:

A Unified Framework with Finance Applications

Cao, Chen, Cong, & Shi (2025)
Cao, Chen, Cong, Gan, & Shi (2026)

Why A Knowledge-Informed Deep Learning Framework?

ML success fostered a dominant narrative, by Anderson (2008):

— “This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. [...] With enough data, the numbers speak for themselves.”

However,

- **Purely data-driven AI** is flexible, but expensive and fragile (in e.g., finance):
  - low signal-to-noise ratios, non-stationarity, and regime shifts;
  - spurious patterns can look profitable in-sample but decay out-of-sample.

Why A Knowledge-Informed Deep Learning Framework?

ML success fostered a dominant narrative, by Anderson (2008):

— “This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. [...] With enough data, the numbers speak for themselves.”

However,

- **Purely data-driven AI** is flexible, but expensive and fragile (in e.g., finance):
  - low signal-to-noise ratios, non-stationarity, and regime shifts;
  - spurious patterns can look profitable in-sample but decay out-of-sample.
- **Structural models/economic theories** interpretable, but often rigid:
  - tractability requires simplified assumptions;
  - calibration is often separate from prediction;
  - misspecification becomes costly in high-dimensional markets.

Why A Knowledge-Informed Deep Learning Framework?

ML success fostered a dominant narrative, by Anderson (2008):

— “This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. [...] With enough data, the numbers speak for themselves.”

However,

- **Purely data-driven AI** is flexible, but expensive and fragile (in e.g., finance):
  - low signal-to-noise ratios, non-stationarity, and regime shifts;
  - spurious patterns can look profitable in-sample but decay out-of-sample.
- **Structural models/economic theories** interpretable, but often rigid:
  - tractability requires simplified assumptions;
  - calibration is often separate from prediction;
  - misspecification becomes costly in high-dimensional markets.
- **SKINNs:** make theory a differentiable, learnable module inside DL.

Not data versus theory; theory becomes part of the learning system.

The SKINNS Framework

Structured-**K**nowledge-**I**nformed **N**eural **N**etworks (SKINNs): better reconcile the dichotomy, particularly for finance and beyond:

- SKINNs integrate data learning with structural parameters identification
- Diverse formats of SK to embed domain-specific theories
- Scalable to complex, (ultra-) high-dimensional SK

We demonstrate the power of SKINNs in various finance applications:

- Option pricing → (10%-15% better) OOS pricing & hedging performance
- Asset pricing → (29%-42% better) OOS Sharpe ratio, implementable portfolios
- Realized vol forecasting → (˜20% better) OOS forecasting performance

The SKINNs Framework

Consider a data-driven function approximator (NN component):

$$
\begin{aligned}
f(\mathbf{X} \subseteq \mathbb{R}^d; \theta) &: \mathbb{R}^d \mapsto \mathbb{R}, \\
f^{(l)} &= h\left(\mathbf{b}^{(l-1)} + \mathbf{W}^{(l-1)} f^{(l-1)}\right), \forall\, l \in \mathbb{N}^+,\ 1 \leq l \leq L, \\
f^{(0)} &= \mathbf{X}
\end{aligned}
$$

One typically learn the NN parameters $\theta = \{\mathbf{b}^{(l)}, \mathbf{W}^{(l)}\}_{l=1}^{L}$ by minimizing a data loss:

$$
\mathcal{L}_{\text{Data}}(\theta; \mathcal{D}_{\text{Obs}} \subseteq \{\mathbf{X}_{\text{Obs}}, \mathbf{y}_{\text{Obs}}\}) = \frac{1}{N_{\text{Obs}}} \| f_\theta(\mathbf{X}_{\text{Obs}}) - \mathbf{y}_{\text{Obs}} \|_2^2
$$

- **No** awareness of principled scientific structures (no structured knowledge)!
- Result in a local optimal with spurious patterns → can't generalize well

The SKINNs Framework

SKINNs embed structured knowledge into the above learning process (SK component):

Definition (Structured Knowledge Representation)

Structured Knowledge is a representation function

$$g(\mathbf{X}^{\mathrm{SK}} \subseteq \mathbb{R}^{d_{\mathrm{SK}}}; \phi \subseteq \mathbb{R}^{d_\phi}) : \mathbb{R}^{d_{\mathrm{SK}}} \mapsto \mathbb{R},$$

which is C^1-smooth w.r.t. $\mathbf{X}^{\mathrm{SK}}, \phi$, or Lipschitz continuous at least, such that ϕ is Gradient Descent learnable. Moreover, $\mathbf{X}^{\mathrm{SK}} \subseteq \mathbf{X}$, $d_{\mathrm{SK}} \ll d$, $d_\phi \ll d_\theta$.

- The function space of $g_\phi(\mathbf{X}^{\mathrm{SK}})$ is given by theories → known SK representation
- Its structural parameters $\phi \in \mathbf{\Phi}$ are unknown → undetermined SK parameters

The SKINNs Framework

Force the data-driven function approximator $f_\theta(\mathbf{X})$ to be SK aware (via a SK loss):

$$\mathcal{L}_{\text{SK}}(\theta, \phi; \mathbf{X}_{\text{Colloc}}) \equiv \frac{1}{N_{\text{Colloc}}} \| f_\theta(\mathbf{X}_{\text{Colloc}}) - g_\phi(\mathbf{X}^{\text{SK}}_{\text{Colloc}}) \|_2^2$$

SKINNs inform the gradient dynamics of training $f_\theta(\mathbf{X})$ though a composite loss:

$$\mathcal{L}_{\text{Total}}(\theta, \phi; \mathbf{X}_{\text{Obs}} \cup \mathbf{X}^{\text{SK}}_{\text{Colloc}}, \mathbf{y}_{\text{Obs}}) = \lambda_{Data} \underbrace{\mathcal{L}_{\text{Data}}(\theta; \mathbf{X}_{\text{Obs}}, \mathbf{y}_{\text{Obs}})}_{\text{Data fidelity}} + \lambda_{\text{SK}} \underbrace{\mathcal{L}_{\text{SK}}(\theta, \phi; \mathbf{X}^{\text{SK}}_{\text{Colloc}})}_{\text{Theory consistency}}$$

We find the optimal θ, ϕ simultaneously by solving the optimization with (S)GD:

$$(\theta^*, \phi^*) =: \underset{\theta \in \boldsymbol{\Theta}, \phi \in \boldsymbol{\Phi}}{\arg\min} \; \mathcal{L}_{\text{Total}}(\theta, \phi; \mathbf{X}_{\text{Obs}} \cup \mathbf{X}^{\text{SK}}_{\text{Colloc}}, \mathbf{y}_{\text{Obs}})$$

The SKINNs Framework

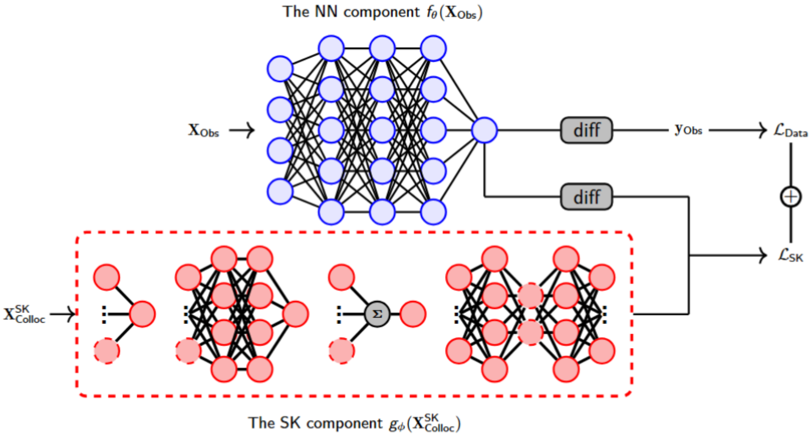
SKINNs work as a generalizability-improved **predictor** & an econometric **estimator**:

- A “white box” to govern a “black box”
- The nuisance θ and the structural ϕ are jointly identifiable (proved in our paper, Section 2.4)
- The estimator-side of SKINNs presents reduced estimation variance than NLS

Our SK representation allows various forms of “theories”:

- Symbolic equations: learn equation parameters
  $\hookrightarrow$ Example: Black-Scholes-Merton model (option), ARMA (realized vol)
- Distributions: learn probabilities
  $\hookrightarrow$ Example: SDF, martingale pricing law
- Downstream optimization objectives: learn optimal decision variables
  $\hookrightarrow$ Example: mean-variance utility maximization

The SKINNs Framework



SK Representation Examples

- Black-Scholes-Merton as the SK rep:

$$g^{\text{BSM}}(\mathbf{X}^{\text{SK}};\phi\in\{\sigma\}) = S_t\Phi\left(d_1(\sigma)\right) - Ke^{-r(T-t)}\Phi\left(d_2(\sigma)\right),$$

$$d_1(\phi\in\{\sigma\}) = \frac{1}{\sigma\sqrt{T-t}}\left(\log(S_t/K) + (r+\frac{\sigma^2}{2})(T-t)\right),$$

$$d_2(\phi\in\{\sigma\}) = d_1(\sigma) - \sigma\sqrt{(T-t)}$$

- ARMA(1, 1) as the SK rep:

$$y_t = c + \alpha\, y_{t-1} + \varepsilon_t + \beta\,\varepsilon_{t-1}, \quad \varepsilon_t \sim \mathcal{N}(0,\sigma^2)$$

$$g^{\text{ARMA}}\left(\mathbf{X}^{\text{SK}} = y_{t-1};\phi\in\{c,\alpha,\beta\}\right) = \mathbb{E}[y_t|\mathcal{F}_t] = c + \alpha\, y_{t-1} + \beta\,\varepsilon_{t-1}$$

- Unknown Distribution as the SK rep:

$$g_\phi(\mathbf{X}^{\text{SK}};\phi\in\{\mathbb{Q}(z^{[i]})\}_i^\infty) = e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}[\underbrace{u(z,\mathbf{X}^{\text{SK}})}_{\text{Payoff function}}|\mathcal{F}_t]$$

$$= e^{-r(T-t)}\sum_{i=1}^{\infty} u(z^{[i]},\mathbf{X}^{[i],\text{SK}})\,\mathbb{Q}(z^{[i]})$$

SK Representation Examples

- Downstream Mean-Variance Optimization as the SK rep:

$$\mathbf{w}^* := \arg\min_{\mathbf{w}} \; - \mathbf{w}^\top \hat{\mathbf{r}} + \eta \, \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}$$
$$s.t \;\; \mathbf{w}^\top \mathbf{1} = \mathbf{1}$$
$$\mathbf{l} \leq \mathbf{w} \leq \mathbf{u}$$

The MVO is typically solved given a learned asset-pricing model $\hat{\mathbf{r}} = f_\theta(\mathbf{X})$
$\hookrightarrow$ Two-stage optimization; no interaction with each other

- Turn such a two-stage optimization into an SK component:

$$g^{\text{MVO}}(f_\theta, \phi \in \{\mathbf{w}\}) = \mathbf{w}^\top f_\theta + \eta \, \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}$$
$$s.t \;\; \mathbf{w}^\top \mathbf{1} = \mathbf{1}$$
$$\mathbf{l} \leq \mathbf{w} \leq \mathbf{u}$$

Similar concept used in Jensen et al. (2026) and Wang et al. (2026)

Empirical Application to Option Pricing

We use comprehensive S&P 500 index option data:

- Daily S&P 500 index call options (European style)
- Covering the period from January 4, 1996, to December 31, 2022
- Options with maturities between 7 and 365 calendar days

Model training schedule:

- Train on three-month option panels
- Test over the next two consecutive months (first month: shorter prediction horizon; second month: longer prediction horizon)
- 317 training & test subsamples in total

What Counts as Structured Knowledge?

Structured knowledge	**Examples in finance**	**What SKINNs learn**
Parametric formulas	BSM, Heston, ARMA	volatility, persistence, jump, or state parameters
Simulation / surrogate models	non-affine stochastic volatility, complex SDEs	latent parameters inside a differentiable surrogate
Distributional restrictions	no-arbitrage, martingale pricing, SDF restrictions	risk-neutral probabilities or pricing kernels
Optimization objectives	mean-variance utility, portfolio constraints	decision variables and predictive functions jointly

- The word “theory” is used broadly: analytical models, simulations, constraints, objectives, or previously learned knowledge.
- The only practical requirement: the structured module should be differentiable enough for gradient-based learning.

Approach	**Useful idea**	**Main limitation in finance**
PINNs	penalize PDE residuals	require differential equations and observable state variables
Transfer learning	pre-train on a source model, then fine-tune	source model often fixed; structural parameters calibrated separately
Model-free shape constraints	enforce monotonicity, convexity, bounds	helpful but limited economic interpretation
SKINNs	jointly train f_θ and g_ϕ	theory and data update each other in one optimization loop

SKINNs are less about solving a PDE; more about learning with economic structure.

Main Empirical Message: Structure Helps Generalization

- **Short horizons:** flexible NNs are competitive when the test period is close to the training period.
- **Longer horizons:** pure data-driven patterns decay as the data-generating process shifts.
- **SKINNs improve out-of-sample pricing:**
  - roughly **10–15% lower pricing RMSE** relative to leading NN benchmarks;
  - gains are larger when the market is volatile.
- **SKINNs also improve hedging:**
  - all variants significantly outperform benchmark NNs in Delta-hedging tests;
  - simple structured modules can be especially effective for hedging.

The value of structure is mostly out-of-sample and under stress.

When Does Structured Knowledge Matter Most?

- The paper studies relative performance against a plain NN:

$$\Delta RMSE_t^{Model} = RMSE_t^{Model} - RMSE_t^{NN}.$$

  A more negative value means the model outperforms the NN more.

When Does Structured Knowledge Matter Most?

- The paper studies relative performance against a plain NN:

$$\Delta RMSE_t^{Model} = RMSE_t^{Model} - RMSE_t^{NN}.$$

  A more negative value means the model outperforms the NN more.

- In low-volatility periods:
  - option surfaces are smoother;
  - patterns are easier for data-driven models to learn;
  - differences across models are modest.

When Does Structured Knowledge Matter Most?

- The paper studies relative performance against a plain NN:

$$\Delta RMSE_t^{Model} = RMSE_t^{Model} - RMSE_t^{NN}.$$

  A more negative value means the model outperforms the NN more.

- In low-volatility periods:
  - option surfaces are smoother;
  - patterns are easier for data-driven models to learn;
  - differences across models are modest.

- In high-volatility periods:
  - SKINNs' gains become statistically and economically larger;
  - highest-volatility quintile: pricing-error reductions are about 3–4 times larger than boundary-constrained NNs.

Stable economic regularities are most valuable when historical correlations break.

Estimator Side: Learning Economic Objects

- SKINNs are not only predictors; they also estimate latent structural parameters ϕ.
- Examples from option pricing:
  - BSM-based SKINN learns a volatility parameter that tracks implied volatility;
  - Heston-type SKINNs learn state variables with smoother dynamics than stand-alone calibration;
  - non-parametric martingale pricing learns high-dimensional risk-neutral probabilities.
- Why this matters:
  - better interpretability than a black-box NN;
  - latent parameters can be monitored as risk indicators;
  - parameter stability can be compared across regimes and models.

The learned “prior” is not just a regularizer;
it can be an economic object.

Beyond Options: Back to Asset Pricing and Investment

- The same framework applies whenever prediction is followed by an economic decision.
- Standard workflow:

$$\text{predict returns} \quad \Longrightarrow \quad \text{optimize portfolio.}$$

  This is often suboptimal because the prediction model does not know the downstream objective.

- SKINNs can embed the portfolio objective directly:

$$\begin{aligned} \min_{w} \quad & -w^\top f_\theta(X) + \eta w^\top \Sigma w \\ \text{s.t.} \quad & w^\top \mathbf{1} = 1, \qquad \ell \le w \le u. \end{aligned}$$

- Interpretation:
  - $f_\theta(X)$ learns return relationships;
  - w becomes a learnable decision variable;
  - the model is trained to be useful for the portfolio objective, not just for MSE.

Takeaways for AI in Finance

- **Data alone is not almighty.**
  - In finance, distribution shifts and low signal-to-noise make pure pattern recognition fragile.
- **Theory alone is not enough.**
  - Structural models are interpretable, but often too restrictive or misspecified.
- **SKINNs offer a practical middle ground.**
  - flexible prediction;
  - structured regularization;
  - interpretable latent parameters;
  - direct integration with portfolio objectives and constraints.

Prediction + theory + objective
in one differentiable learning loop.

GenAI and SKINNs for Asset Pricing and Investment

- Generative modeling and GenAI for finance goes beyond language models and are **just starting to be explored**.
- **General-objective optimization at scale**

  Data-driven generation/search over a larger solution and strategy space; **Deep RL** and **Panel Trees** are two major search algorithms.
- **Environment modeling and simulation**

  Generative models can learn states, dynamics, agents, and feedback, to train better strategies and conduct counterfactual/stress-test simulations.
- **Structured knowledge in data-driven generative models**

  Theories, constraints, objectives, and economic mechanisms can discipline generation.

Search broadly, simulate richly, and discipline with knowledge.