Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Enhancing Commodity Factor Strategies with Deep Learning: Evidence from Basis-Momentum

Abstract

This article examines how deep learning can enhance traditional factor-based investment strategies in the commodity futures market. Building on well-established signals such as momentum and carry, we revisit the basis-momentum factor—a hybrid measure of trend and term structure—and evaluate its robustness using daily data from 19 commodities over the period 2014 to 2025. Our results confirm that basis-momentum remains a highly effective ranking signal, outperforming momentum and carry alone.

We then apply two deep learning models—Long Short-Term Memory (LSTM) and Transformer networks—to dynamically rank commodities based on both raw returns and factor-informed inputs. We find that AI models trained on economically meaningful features significantly outperform those trained on raw data, both in returns and ranking accuracy. The combination of financial intuition and machine learning produces more adaptive and interpretable strategies. Our findings highlight the value of integrating domain knowledge into AI-driven portfolio design and offer a practical framework for applying intelligent ranking to commodity investments.

KEYWORDS

Asset Ranking, Basis-Momentum, Commodity Futures, Cross-Sectional Return Prediction, Deep Learning, Factor Investing, Sequence Models (LSTM, Transformer)

1 | INTRODUCTION

Commodity futures markets offer a compelling environment for systematic investing, supported by well-documented return premia such as momentum and carry. Academic studies over the past two decades have consistently shown that investors can earn significant long-short returns by ranking commodities based on historical price trends or term structure characteristics. In particular, the basis—measured as the difference between near and far-month futures prices—serves as a powerful proxy for the term structure and has been associated with the roll yield in backwardated markets.

Recent research has gone a step further by proposing composite factors that blend price momentum with basis information. One notable innovation is the basis-momentum (BMOM) factor proposed by (Boons & Prado, 2019), which captures shifts in the futures curve and has demonstrated superior performance over traditional single-dimensional factors. However, existing studies have not sufficiently explored the robustness of BMOM across different market regimes or its enhancement through advanced machine learning techniques.

In parallel, recent advancements in machine learning, particularly sequence models like Long Short-Term Memory (LSTM) networks and Transformers, have shown increasing promise in financial prediction and portfolio construction. These models effectively capture dynamic temporal dependencies and learn complex asset ranking structures. Nevertheless, the key challenge remains balancing predictive accuracy with interpretability, especially when addressing audiences who prioritize economically justified signals.

This article addresses these gaps by introducing a factor-informed, AI-enhanced portfolio construction framework specifically tailored for commodity futures markets. Unlike existing studies such as (Hu & Ni, 2024), which apply deep learning to hedging strategies with limited emphasis on cross-sectional dynamics, or (Wang & Zhang, 2024), which focus on macro-driven return prediction using black-box machine learning, our approach centers on interpretable asset ranking using domain-informed factors such as momentum, carry and BMOM. In particular, we embed the economically grounded BMOM into LSTM and Transformer architectures to preserve financial intuition while enhancing predictive accuracy. We empirically validate the continued efficacy of momentum, carry and BMOM factors using daily data from 2014 to 2025. We then demonstrate that LSTM and Transformer models trained on factor-informed inputs significantly outperform static sorting strategies, not only in terms of return spreads and Sharpe ratios but also in their predictive alignment with actual return rankings.

The remainder of the article proceeds as follows. We first review relevant literature, highlighting existing commodity factor strategies and the use of machine learning in portfolio construction. We then describe our dataset, factor definitions, and the architecture of the proposed LSTM and Transformer models. Subsequently, we outline our evaluation methodology and present empirical results. Finally, we discuss the economic implications, model interpretability, and key takeaways for practitioners and researchers.

2 | RELATED LITERATURE

2.1 | Commodity Factor Strategies

Systematic return premia in commodity futures have been extensively documented. (Gorton & Rouwenhorst, 2006) demonstrated that commodity futures portfolios offer equity-like returns with diversification benefits, laying the groundwork for cross-sectional factor-based strategies. Momentum and carry, represented by futures basis, are among the most persistent strategies. For instance, (Miffre & Rallis, 2007) reported significant annual returns for momentum strategies, while (Szymanowska, De Roon, Nijman, & Van Den Goorbergh, 2014) found robust carry-driven return spreads.

Expanding on simple momentum and carry strategies, researchers introduced dynamic and composite signals. (Erb & Harvey, 2006) suggested multiple metrics for timing market entries and exits, while (Bianchi, Drew, & Fan, 2016) provided behavioral explanations for momentum's effectiveness. (Kim & Kang, 2014) proposed a dynamic-slope strategy exploiting changes in the futures term structure, demonstrating significant excess returns during periods of steepening backwardation.

More recently, (Boons & Prado, 2019) introduced the basis-momentum (BMOM) factor, which combines slope and curvature information from the futures curve, significantly outperforming conventional momentum and carry. However, despite its documented efficacy, existing literature has not adequately examined how BMOM's predictive power varies across different market conditions or how machine learning techniques could further enhance its performance.

2.2 | Machine Learning in Financial Markets

Machine learning has become prominent in financial forecasting and asset selection due to its capability to model complex patterns. Early neural network applications often faced instability and overfitting issues, as highlighted by (López de Prado, 2018), who recommended rigorous validation to avoid spurious findings. Our research adheres strictly to such rigorous out-of-sample evaluation procedures.

LSTM networks have shown particular efficacy in capturing temporal dependencies in asset returns. (Fischer & Krauss, 2018) and subsequent studies (Binoy & Jos, 2022) and (Islam & Hossain, 2021) demonstrated LSTM's superior predictive capabilities in financial markets. More recently, Transformers, leveraging self-attention mechanisms, have further enhanced time-series prediction. (Kampotha, Wang, & Li, 2024) and (Gezici & Sefer, 2024) confirmed Transformers' predictive advantages in various financial contexts, although highlighting their sensitivity to data size and quality.

Asset ranking via machine learning, beyond direct forecasting, has gained traction. (Alzaman, 2025) successfully integrated LSTM within reinforcement learning frameworks for stock ranking. However, these methods, especially those trained solely on raw price inputs, often lack interpretability and clear economic grounding.

2.3 | Integrating Factors and Deep Learning

While powerful, machine learning models risk losing interpretability and economic meaning when trained solely on raw data. (Lo & Singh, 2023) emphasized the need for transparent mechanisms in deep learning, presenting tools to clarify model predictions related to financial risk premia. Factor-informed learning, where economically meaningful inputs guide model training, has recently gained attention. (Wei, Cao, & Dong, 2024) showed improved generalization and transparency across asset classes through factor-based neural networks.

Our research uniquely extends these insights into the commodity futures context. Contrary to (Hu & Ni, 2024), who applied deep learning primarily to hedging strategies, or (Wang & Zhang, 2024), who used primarily macroeconomic features, we explicitly integrate the economically robust factors such as momentum, basis and BMOM within LSTM and Transformer models. This approach retains the interpretability and economic rationale of traditional factors, combining them effectively with deep learning's predictive strengths.

In doing so, our work addresses the literature gap identified by (Lo & Singh, 2023) by clearly demonstrating how AI models can economically interpret and utilize structured financial signals. Our study thus significantly advances the existing literature by merging robust economic theory with state-of-the-art machine learning, providing a model framework that is both interpretable and practically applicable.

3 | METHODOLOGY

3.1 | Data Description

Our commodity universe comprises 19 highly liquid futures contracts traded on the Chicago Mercantile Exchange (CME) Group, Intercontinental Exchange (ICE) Futures U.S., and London Metal Exchange (LME), as detailed in Table I. The selection includes a diversified set of assets across energy (e.g., crude oil, natural gas), metals (e.g., gold, copper), agriculture (e.g., corn, soybeans), and soft commodities (e.g., coffee, sugar), ensuring both sectoral diversity and market representativeness. These contracts were chosen based on the availability of continuous daily price series, consistent trading volume, and relevance in institutional portfolios. We use daily settlement prices for 19 actively traded commodity futures contracts listed on the CME, ICE, and LME. Our sample spans from January 2014 through February 2025.

For each commodity, we construct continuous return series using the front-month contract, rolling to the next liquid contract approximately five trading days prior to expiration. To compute the basis and basis-momentum factor, we additionally collect data on second-nearest contracts. All return series are adjusted for roll yield and cleaned for extreme observations beyond ±20% to mitigate the impact of erroneous ticks or illiquidity-induced spikes.

TABLE I **Commodity Futures Universe**

Commodity	Exchange	Expiry Months
Crude Oil	CME Group	Every month
Natural Gas	CME Group	Every month
Corn	CME Group	Mar, May, Jul, Sep, Dec
Soybeans	CME Group	Jan, Mar, May, Jul, Aug, Sep, Nov
Live Cattle	CME Group	Feb, Apr, Jun, Aug, Oct, Dec
Gold	CME Group	Feb, Apr, Jun, Aug, Oct, Dec
Heating Oil	CME Group	Every month
Gasoline	CME Group	Every month
Wheat	CME Group	Mar, May, Jul, Sep, Dec
Lean Hog	CME Group	Feb, Apr, May, Jun, Jul, Aug, Oct, Dec
Silver	CME Group	Mar, May, Jul, Sep, Dec
Sugar	ICE Futures U.S.	Mar, May, Jul, Oct
Cotton	ICE Futures U.S.	Mar, May, Jul, Oct, Dec
Cocoa	ICE Futures U.S.	Mar, May, Jul, Sep, Dec
Coffee	ICE Futures U.S.	Mar, May, Jul, Sep, Dec
Orange Juice	ICE Futures U.S.	Jan, Mar, May, Jul, Sep, Nov
Aluminum	LME	Every month
Copper	LME	Every month
Nickel	LME	Every month

Note: This exhibit lists the 19 commodity futures included in our sample, specifying the primary exchange and standard expiration months for each contract. The contracts are selected based on high liquidity, trading activity, and uninterrupted daily prices for return computation. CME = Chicago Mercantile Exchange; ICE = Intercontinental Exchange; LME = London Metal Exchange.

3.2 | Factor Construction

We consider three widely studied cross-sectional signals to rank commodities:

- **Momentum**: Defined as the 12-month cumulative return of the front-month contract, computed monthly.

$$MOM_{i,t} = \prod_{s=t-11}^{t} (1 + R_{i,s}) - 1, \tag{1}$$

where denotes the monthly return of commodity at month . A higher indicates that the commodity has demonstrated strong performance over the prior year.

- **Basis (-Carry)**: Computed as the percentage price difference between the second-nearest and front-month contracts.

$$Basis_{i,t} = \frac{P_{i,t}^{(2nd)} - P_{i,t}^{(1st)}}{P_{i,t}^{(1st)}}, \tag{2}$$

where $P_{i,t}^{(1st)}$ and $P_{i,t}^{(2nd)}$ are the prices of the first and second nearby contracts, respectively. A negative basis implies backwardation. For consistency with prior literature, we define the carry factor as the negative of the basis value such that higher values correspond to higher expected returns. That is, we use $Carry = -Basis$ as the sorting signal.

- **Basis-Momentum (BMOM)**: Following (Boons & Prado, 2019), this signal is the difference in 12-month returns between the nearest and second-nearest contracts.

$$BMOM_{i,t} = MOM_{i,t}^{(1)} - MOM_{i,t}^{(2)}, \tag{3}$$

where $MOM_{i,t}^{(1)}$ and $MOM_{i,t}^{(2)}$ are the 12-month returns of the first and second nearby contracts, respectively. A high $BMOM_{i,t}$ indicates that the front-month contract has outperformed the second-nearest, suggesting tightening market conditions (i.e., increasing backwardation).

Using price information, (Boons & Prado, 2019) decomposes the $BMOM_{i,t}$ as:

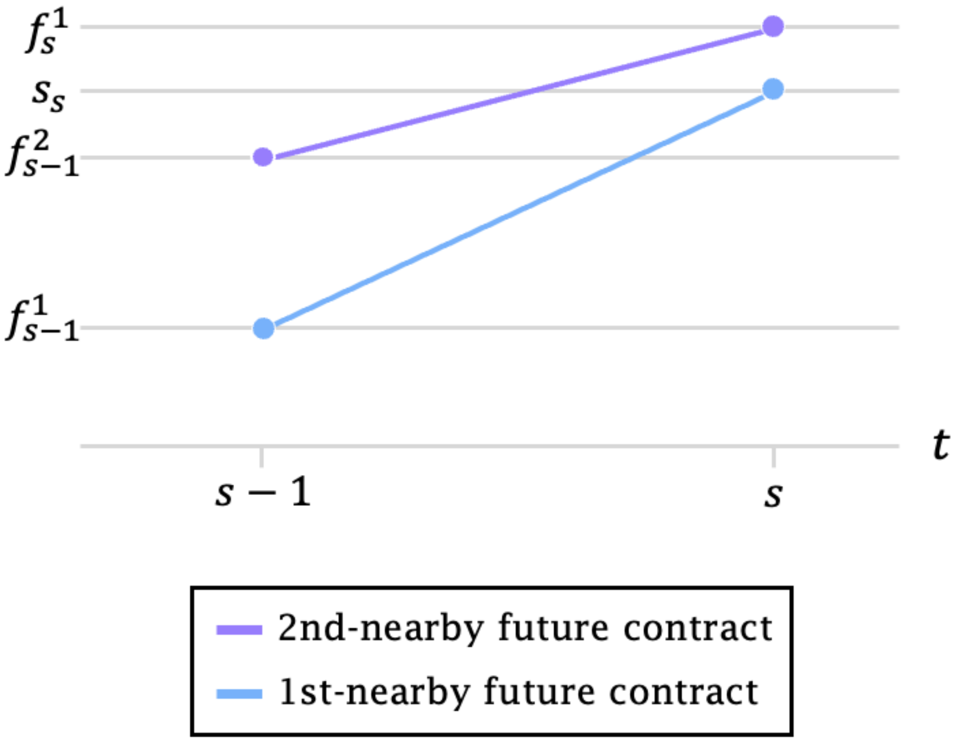
$$\sum_{s=t-11}^{t} (s_s - f_{s-1}^{1}) - \sum_{s=t-11}^{t} (f_s^{1} - f_{s-1}^{2}), \tag{4}$$

which compares the slope of the first-nearby contract price with that of the second. A positive $BMOM_{i,t}$ implies a flattening or inversion of the term structure, indicating backwardation; a negative value corresponds to contango. This intuition is visually illustrated in Figure1, which shows how the relative slopes of front and deferred contracts evolve over time.

3.3 | Deep Learning Model for Asset Ranking

To dynamically rank commodities based on cross-sectional return signals, we implement two sequence learning architectures: Long Short-Term Memory (LSTM) networks and Transformer encoders. These models are not employed for direct price forecasting; instead, they infer relative performance rankings to inform long–short portfolio construction. Importantly, our basis-momentum (BMOM) factor captures intricate slope and curvature dynamics within the commodity futures term structure, inherently characterized by medium-to-long-term temporal dependencies. Effectively modeling these complex, regime-sensitive dynamics necessitates deep learning architectures capable of selectively encoding and managing past information.

Modeling Philosophy. Unlike traditional factor-based sorters that apply static thresholds to predefined signals, our deep learning models are explicitly trained to recognize evolving interdependencies among economically grounded features such as momentum, carry, and BMOM. The LSTM's gating mechanism, comprising forget, input, and output gates, uniquely facilitates the selective retention and discarding of historical information. This capability is particularly suited to capturing BMOM's inherent temporal dependencies, allowing the model to effectively manage noise preceding structural regime shifts and enhance signal robustness. Economically, the BMOM factor encodes the dynamic slope and curvature of the futures term structure, FIGURE 1 . **Illustration of Basis-Momentum Signal Construction**



Note: This figure visualizes the computation of the basis-momentum (BMOM) factor, defined as the difference in 12-month cumulative returns between the nearest and second-nearest futures contracts. It compares the price paths of these two contracts at two points in time, s–1 and s, and reflects the slope differentials embedded in the term structure. Equation (4) corresponds to the decomposition proposed by (Boons & Prado, 2019), where a steeper rise in the first-nearby contract relative to the second-nearby indicates tightening market conditions (i.e., flattening contango).

which are influenced by inventory shocks, convenience yields, and changing speculative demand. These structural features unfold over varying horizons—some gradually, others with abrupt shifts—making BMOM signals ideal for deep architectures with complementary strengths. LSTM networks, with their gated memory, are well-suited to learning medium-term temporal persistence associated with evolving backwardation or contango. In parallel, the Transformer's self-attention mechanism excels at detecting localized inflection points and regime shifts, such as sudden curve inversions, by emphasizing relevant time-step interactions. Together, these architectures allow our model to internalize both persistent and abrupt changes in market structure, yielding a ranking system that adapts to diverse commodity environments.

LSTM Model. The LSTM network processes a 12-month rolling input window composed of daily engineered features across 19 commodities. We use two stacked LSTM layers with hidden dimensions between 32 and 64, depending on feature complexity. The model outputs a cross-sectional score for each commodity at time $t+1$, which is later used to determine its relative ranking. The LSTM's gating mechanism enables the model to capture temporally persistent structures embedded in factor dynamics while filtering out high-frequency noise.

Transformer Model. The Transformer encoder, originally introduced by (Vaswani et al., 2017), uses self-attention to model temporal dependencies without recurrence. Our implementation includes 3 stacked attention layers with 8 heads and a feed-forward dimension of 64, along with positional encodings. This architecture enables flexible weighting of historical observations and captures localized shifts in return structures. The compact design helps prevent overfitting, considering the relatively small size of our commodity dataset.

Labeling and Portfolio Mapping. Both models are trained using a ternary classification scheme. At each time step, commodities are labeled as +1 (long), –1 (short), or 0 (neutral), based on their realized next-day return rankings. Specifically, the top 4 performers are labeled +1, the bottom 4 as –1, and the remainder as 0. These predicted labels are directly used to construct long–short portfolios for the subsequent trading day. The economic relevance of this scheme is evaluated using return spreads between the long and short baskets, as well as the Pearson rank correlation between model-implied and realized return rankings. This approach ensures that model performance is directly aligned with cross-sectional investment outcomes.

Implementation Summary. For each day t, the model receives a window of past feature data and outputs a score for each asset. Commodities are ranked by these scores, and the top and bottom 4 are selected for long and short positions on day $t+1$. This process is repeated sequentially across the test period. All model evaluations are conducted strictly out-of-sample, with no overlap between training and test periods. A schematic of this process is presented in Figure 2.

3.4 | Training and Testing Procedure

The dataset is divided into two periods: a training/validation period from January 2014 to March 2021, and a test period from April 2021 to February 2025. To prevent look-ahead bias and ensure that model evaluation reflects a realistic investment setting, we clearly separate training, validation, and test periods. Models are trained exclusively on data from January 2014 through March 2021. All hyperparameter tuning is conducted within this range using walk-forward validation. The out-of-sample test period, spanning April 2021 to February 2025, remains entirely unseen until final evaluation. During training, hyperparameters are optimized using a rolling validation window (e.g., 2019–2021) to ensure generalization. We adopt the Optuna framework for hyperparameter optimization, employing a Bayesian search algorithm over a rolling validation period. The validation set is updated in a walk-forward fashion to simulate realistic deployment conditions. To mitigate overfitting, we implement early stopping based on validation loss, apply dropout regularization (rates tuned between 0.1 and 0.3), and cap model complexity by limiting the number of LSTM/Transformer layers and hidden dimensions. After validation, the final model parameters are fixed and evaluated on the out-of-sample test set (2021–2025) without further adjustment, simulating real-time performance.

3.5 | Portfolio Construction and Rebalancing Frequency

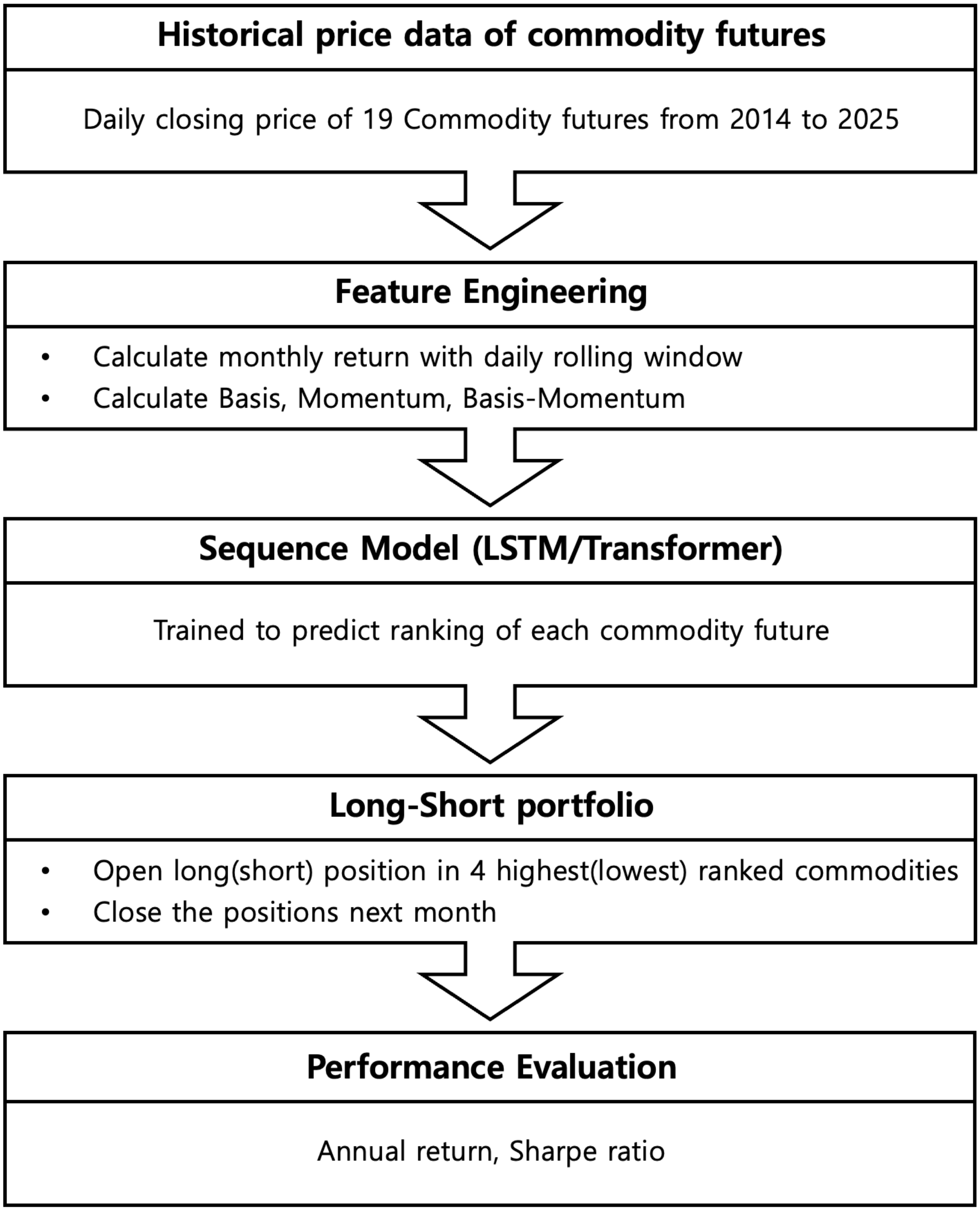
For traditional factor portfolios, we use monthly rebalancing based on the latest factor signals. For machine learning–based strategies, rebalancing is also conducted monthly using model-generated rankings with daily rolling window to ensure comparability. In practice, we observe that ML-based rankings are relatively stable over short horizons. To empirically validate this observation, we calculate the overlap ratio of long and short positions across consecutive days. On average, over 70% of positions remain unchanged, indicating that ML-based portfolios do not exhibit excessive churn despite daily ranking updates. This stability mitigates concerns regarding trading costs and allows for implementation flexibility (e.g., daily or weekly rebalancing) without significantly impairing performance.

3.6 | Evaluation Metrics and Assumptions

- **Annualized Return**: Mean of monthly returns × 12.
- **Annualized Volatility**: Standard deviation of monthly returns $\times \sqrt{12}$.
- **Sharpe Ratio**: Return divided by volatility.
- **Rank Correlation**: Pearson correlation between predicted and actual next-period return ranks.

All return figures represent excess returns of fully collateralized long–short futures positions. Trading costs are not explicitly modeled, but we assume liquid markets with minimal slippage for large institutional trades. A separate analysis of turnover and transaction costs is left for future research.

FIGURE 2 . **Deep Learning–Based Commodity Ranking and Portfolio Formation**


Historical price data of commodity futures
Daily closing price of 19 Commodity futures from 2014 to 2025
Feature Engineering
• Calculate monthly return with daily rolling window
• Calculate Basis, Momentum, Basis-Momentum
Sequence Model (LSTM/Transformer)
Trained to predict ranking of each commodity future
Long-Short portfolio
• Open long(short) position in 4 highest(lowest) ranked commodities
• Close the positions next month
Performance Evaluation
Annual return, Sharpe ratio

Note: This figure summarizes the machine learning pipeline for constructing AI-enhanced commodity portfolios. Daily closing prices of 19 commodity futures are first processed through feature engineering steps to derive momentum, basis, and basis-momentum signals using rolling return windows. These engineered features serve as inputs to sequence models (LSTM or Transformer), which are trained to predict the cross-sectional ranking of commodities for the subsequent period. Based on the model-implied scores, long-short portfolios are formed by taking positions in the top and bottom four ranked assets each month. Portfolio performance is evaluated on an out-of-sample test set using annualized return and Sharpe ratio, under the assumption of frictionless execution.

4 | EMPIRICAL RESULTS

4.1 | Performance of Traditional Factor-Based Strategies

The table II presents the performance of long–short portfolios based on three cross-sectional factors— basis-momentum (BMOM), carry, and momentum—using both nearby contract returns (Panel A) and spreading returns between nearby and deferred contracts (Panel B). Among the three, the BMOM factor demonstrates the strongest performance, delivering an annualized return of 17.21% with a Sharpe ratio of 0.80 in the nearby return specification. This outperforms both the momentum strategy (4.76%, Sharpe 0.17) and the carry strategy (0.90%, Sharpe 0.03), whose return spreads are modest and statistically insignificant.

When considering spreading returns (Panel B), BMOM again leads with a 2.86% return and a Sharpe ratio of 0.45, while both carry (–2.48%) and momentum (–1.10%) produce negative spreads with Sharpe ratios below zero. The consistently superior performance of BMOM across both specifications affirms its robustness as a composite signal that captures both slope and curvature effects in the futures curve, aligning with the findings of (Boons & Prado, 2019).

TABLE II **Performance of Traditional Commodity Factors (2014–2025)**

	Basis-Momentum				**Carry**	**Momentum**
	High4	Mid	Low4	High4–Low4	High4–Low4	High4–Low4
Panel A: Nearby Returns						
Average Return (%)	12.75	4.70	-4.50	**17.21**	0.90	4.76
t-statistic	(3.06)	(1.24)	(-0.98)	(3.58)	(0.17)	(0.80)
Sharpe Ratio	0.69	0.28	-0.21	**0.80**	0.03	0.17
Panel B: Spreading Returns						
Average Return (%)	3.90	1.20	1.03	**2.86**	-2.48	-1.10
t-statistic	(3.45)	(3.48)	(1.04)	(1.93)	(-1.71)	(-0.82)
Sharpe Ratio	0.80	0.69	0.24	**0.45**	-0.42	-0.21

Note: The table reports performance statistics of long–short commodity portfolios formed using three cross-sectional signals—Basis-Momentum (BMOM), Carry (Basis), and Momentum—over the 2014–2025 sample period. Panel A uses nearby (front-month) futures returns, while Panel B uses the return spread between the nearest and second-nearest contracts to isolate term structure effects. Each portfolio is constructed by going long in the top 4 and short in the bottom 4 ranked commodities based on each signal. Reported statistics include average return, t-statistic, and Sharpe ratio of the High4–Low4 portfolios.

4.2 | Performance of AI-Enhanced Sorting Models

The table III presents the performance of AI-driven long–short portfolios constructed using LSTM and Transformer models trained on various input features over the full test period. Across all feature specifications, deep learning models significantly outperform traditional static sorts in both average return and Sharpe ratio.

Among the LSTM-based models, the version trained with the basis-momentum (BMOM) factor achieves the strongest performance, delivering an annualized High4–Low4 return of 69.36% with a Sharpe ratio of 2.55. Momentum-based LSTM models follow closely, with a return of 67.68% and Sharpe of 3.02. The Transformer model with BMOM input also performs robustly, yielding a 51.84% spread and a Sharpe ratio of 2.21, though generally trailing the LSTM models.

In addition to return and risk-adjusted performance, we report maximum drawdown (MDD) to assess downside risk. LSTM-based models trained on momentum and BMOM features show significantly lower drawdowns (–17.01% and –19.92%, respectively) compared to their static sorting counterparts (–35.83% and –22.74%). Transformer-based models also exhibit improved drawdown profiles, especially under momentum input (–14.67%). These results suggest that deep learning models not only enhance return generation but also provide better capital preservation.

The results underscore two key findings. First, integrating economically meaningful factor inputs—especially BMOM—enhances the predictive capacity of sequence models. Second, LSTM architectures consistently outperform Transformer counterparts across feature sets. This advantage may stem from LSTM's gated memory mechanism, which better captures medium-term return persistence while suppressing short-term volatility noise.

These outcomes support the use of factor-informed deep learning frameworks for dynamic asset ranking, with particular strength in the LSTM + BMOM configuration.

TABLE III **Full-Period Performance of AI-Based Ranking Strategies (2022.3–2025.2)**

Model	Input	High4	Low4	High4–Low4	Sharpe Ratio	MDD
Static Sorting	Momentum	28.80	-3.24	31.92	1.09	-35.83
	Carry	19.92	-5.28	25.20	1.12	-27.74
	Basis-Momentum (BMOM)	30.84	-7.56	**38.40**	**1.52**	**-22.74**
LSTM	Raw Returns	32.76	-25.32	57.95	2.62	**-14.25**
	Momentum	39.00	-28.80	67.68	**3.02**	-17.01
	Carry	16.32	-21.36	37.68	1.45	-26.70
	BMOM	37.68	-31.56	**69.36**	2.55	-19.92
Transformer	Raw Returns	25.20	-22.32	47.52	1.88	-18.43
	Momentum	30.72	-19.08	49.80	**2.36**	**-14.67**
	Carry	13.20	-22.44	35.64	1.55	-23.61
	BMOM	32.88	-18.96	**51.84**	2.21	-15.85

Note: The table reports performance statistics for long–short portfolios constructed using AI models (LSTM and Transformer) trained on different input features over the full test period (March 2022 to February 2025). Inputs include raw returns, momentum, carry (–basis), and basis-momentum (BMOM). Models output daily rankings used to form High4–Low4 portfolios, where the top 4 assets are held long and the bottom 4 short. Reported figures are annualized returns, Sharpe ratios based on monthly returns, and maximum drawdown (MDD) as a measure of downside risk. LSTM models trained on momentum and BMOM achieve the highest Sharpe ratios (3.02 and 2.55, respectively), while Transformer models show improved drawdown profiles.

4.3 | Robustness Over Subperiods

To evaluate the temporal robustness and regime adaptability of the proposed models, we divide the full test sample (March 2022 – February 2025) into three rolling 12-month subperiods, corresponding roughly to: (i) early post-COVID volatility (2022.3–2023.2), (ii) stabilization and mild normalization (2023.3–2024.1), and (iii) late-cycle commodity deceleration amid global tightening (2024.2–2025.2). The table IV presents the results.

Across all subperiods, LSTM models consistently outperform static factor sorts and Transformer -based models. In the most volatile period (2022.3–2023.2), the LSTM model trained on momentum signals delivers the strongest performance, achieving a High4–Low4 return of 64.70% and a Sharpe ratio of 2.77. This suggests that momentum-based temporal signals were particularly effective during high-variance regimes immediately following post-COVID shocks. Significantly, this model also achieves a relatively low MDD of –17.01%, indicating resilience against downside risk during turbulent markets.

Meanwhile, the LSTM-BMOM model also performs well, with a 44.30% return and Sharpe ratio of 1.60, slightly lower but still superior to static BMOM (1.78) and the Transformer-BMOM counterpart (1.91). These results highlight that while BMOM offers a more stable and structure-aware signal, raw momentum may react more strongly under turbulent conditions. The corresponding MDD for LSTM-BMOM (–18.52%) is also well-contained compared to static BMOM (–18.27%), reinforcing its capital-preserving nature.

During the second subperiod (2023.3–2024.1), which saw partial rebalancing of global supply chains and decelerating inflation, the LSTM-BMOM model reached its highest performance, delivering a 66.19% and a Sharpe ratio of 3.44. This suggests that in moderately volatile, momentum-driven regimes, the model can fully capitalize on latent signal structures in term spreads. The model's MDD in this phase (–11.67%) further supports its robustness under normalized conditions.

In the final subperiod (2024.2–2025.2), as the commodity cycle entered a cooling phase with policy-driven tightening and lower dispersion, the LSTM-BMOM model remained effective (65.60%, Sharpe 2.27), albeit with reduced relative gains. The Transformer models also retained reasonable performance, though consistently trailing the LSTM versions in return stability and Sharpe ratio. Interestingly, the MDD of LSTM-BMOM (–19.92%) remained lower than that of static models, reinforcing its value as a drawdown-conscious strategy.

These findings affirm that factor-informed deep learning models, especially LSTM architectures with BMOM input, are not merely overfitting to short-term trends but generalize effectively across distinct commodity market regimes. Importantly, these models also offer superior downside protection—capturing returns while containing drawdowns—a trait highly valued in institutional portfolio design.

TABLE IV **Subperiod Performance of AI-Based Ranking Strategies**

Model	Input	High4	Low4	High4–Low4	Sharpe Ratio	MDD
Panel A: Subperiod (2022.3–2023.2)						
Static Sorting	Momentum	-11.83	-2.50	-9.28	-0.32	-35.83
	Carry	-15.43	-16.51	1.07	0.05	-21.78
	Basis–Momentum (BMOM)	18.40	-35.42	**53.56**	**1.78**	**-18.27**
LSTM	Raw Returns	20.27	-33.59	53.86	2.67	**-13.64**
	Momentum	22.13	-42.57	**64.70**	**2.77**	-17.01
	Carry	-0.21	-25.58	25.37	0.92	-26.70
	BMOM	14.85	-29.45	44.30	1.60	-18.52
Transformer	Raw Returns	6.39	-30.50	36.89	1.50	-15.40
	Momentum	18.66	-28.54	**47.20**	**1.97**	**-13.49**
	Carry	-0.36	-30.11	29.75	1.45	-19.20
	BMOM	13.53	-30.23	43.76	1.91	-13.96
Panel B: Subperiod (2023.3–2024.1)						
Static Sorting	Momentum	29.23	-20.85	**49.85**	2.22	-11.57
	Carry	17.27	-30.64	47.69	**2.81**	**-7.39**
	Basis–Momentum (BMOM)	18.41	-13.03	31.29	1.65	-13.25
LSTM	Raw Returns	23.05	-26.47	49.51	2.46	-11.22
	Momentum	33.89	-22.48	56.36	2.99	-12.44
	Carry	9.72	-23.83	33.55	2.08	**-10.06**
	BMOM	30.74	-35.45	**66.19**	**3.44**	-11.67
Transformer	Raw Returns	24.38	-24.98	49.36	2.54	-11.40
	Momentum	31.30	-16.49	47.79	**2.89**	-11.18
	Carry	6.72	-14.57	21.29	0.91	-23.61
	BMOM	29.22	-21.03	**50.25**	2.85	**-8.35**
Panel C: Subperiod (2024.2–2025.2)						
Static Sorting	Momentum	52.03	1.06	50.97	1.60	-21.57
	Carry	38.03	-7.33	45.36	1.71	-21.33
	Basis–Momentum (BMOM)	44.53	-6.62	**51.15**	**1.86**	**-20.72**
LSTM	Raw Returns	48.43	-2.08	50.51	2.08	-14.25
	Momentum	46.23	-10.20	56.43	2.54	**-11.78**
	Carry	31.89	-1.07	32.96	1.12	-23.40
	BMOM	51.59	-14.01	**65.60**	**2.27**	-19.92
Transformer	Raw Returns	32.51	-0.44	32.95	1.25	-18.43
	Momentum	32.13	-4.23	36.36	1.74	-14.67
	Carry	29.47	-17.96	**47.43**	**2.02**	**-11.85**
	BMOM	41.04	6.21	34.83	1.46	-15.85

Note: The table reports model performance across approximately 12-month subperiods within the full test window (March 2022 to February 2025). Each subperiod corresponds to distinct commodity market regimes, ranging from post-COVID volatility to stabilization and eventual tightening. Reported figures reflect annualized High4–Low4 returns, Sharpe ratios, and maximum drawdown (MDD) for each model–input pair. Models are re-evaluated in each subperiod to assess adaptability to changing macroeconomic and market conditions. LSTM-Momentum achieves the highest Sharpe ratio (2.77) in the most volatile regime (2022.3–2023.2), while LSTM-BMOM consistently performs well across periods, especially in more stable phases. Transformer models generally exhibit lower drawdowns, notably Transformer-BMOM with –8.35% in 2023.3–2024.1.

4.4 | Ranking Accuracy Analysis

To assess how effectively each model captures the cross-sectional ordering of asset returns—a critical aspect of ranking-based portfolio construction—we compute the Pearson rank correlation between model-implied rankings and realized return rankings over the next period. The table V presents these correlations across different model–input configurations.

Among all configurations, the LSTM model trained on basis-momentum (BMOM) features exhibits the highest alignment, with a Pearson rank correlation of 0.2015. This is followed by the Transformer-BMOM model at 0.1742. In contrast, traditional static sorting based on raw returns or simple momentum yields correlations below 0.11, suggesting limited predictive structure.

These results confirm that deep learning models, especially those incorporating BMOM, generate rankings more consistent with future return realizations. This alignment indicates that performance gains from AI-based methods are not merely driven by extreme predictions but by improved ordering consistency across the asset cross-section. Significantly, the structural advantage of BMOM—capturing both slope and curvature of the term structure—appears to facilitate more economically informed return forecasting within sequential models.

TABLE V **Rank Alignment Between Model Predictions and Realized Returns**

Model	Input Feature	Pearson Rank Correlation
Static Sorting	Raw Returns	0.0858
	Momentum	0.0771
	Carry	**0.1061**
	BMOM	0.0985
LSTM	Raw Returns	0.1700
	Momentum	0.1721
	Carry	0.1061
	BMOM	**0.2015**
Transformer	Raw Returns	0.1534
	Momentum	0.1419
	Carry	0.1021
	BMOM	**0.1742**

Note: The table reports the Pearson rank correlation between model-implied commodity rankings and realized return rankings over the next month. Higher values indicate closer alignment between model predictions and actual return behavior. Results are presented for static sorting, LSTM, and Transformer models using various input features. Among all configurations, the LSTM model trained on basis-momentum (BMOM) features exhibits the highest rank correlation (0.2015), followed by the Transformer–BMOM model (0.1742).

5 | DISCUSSION: ECONOMIC INSIGHTS FROM FACTOR-INFORMED AI

The empirical results yield several important insights for researchers and practitioners in systematic commodity investing, especially at the intersection of financial theory and AI-driven implementation.

First, the evidence reaffirms the persistent relevance of classical factor strategies. The basis-momentum (BMOM) signal, originally proposed by (Boons & Prado, 2019), continues to outperform traditional momentum and carry across volatile and structurally diverse periods. This supports the notion that term structure information—capturing the relative slopes between nearby and deferred futures—is an economically grounded and persistent source of cross-sectional predictability.

Second, deep learning models—specifically LSTM and Transformer architectures—substantially improve cross-sectional ranking performance relative to static sorting methods. These models are trained not just to predict returns, but to learn rank-ordered signals aligned with realistic portfolio construction. Our ternary labeling scheme (Long = +1, Short = –1, Neutral = 0) embeds this structure into training, and contributes to more actionable model outputs.

Third, the models' predictive performance is enhanced when they are trained on economically meaningful features rather than raw return sequences. LSTM models using BMOM inputs, for example, achieve the highest Pearson rank correlation with realized returns (0.2015), surpassing all other configurations. This shows that feature design grounded in financial intuition—not just data mining—improves both signal clarity and generalization. Furthermore, regularization techniques such as dropout, early stopping, and validation-based hyperparameter tuning help mitigate overfitting, ensuring robustness across market regimes.

Fourth, empirical results reveal that deep learning models—particularly those trained on BMOM — exhibit context-sensitive signal amplification. For instance, during the 2023.3–2024.1 period, which coincided with a rebalancing of global supply chains and a partial easing of inflationary pressures, the LSTM-BMOM model achieved its highest Sharpe ratio (3.44). This suggests that the model identified periods of tightening term structures—often associated with backwardation—as regimes where momentum signals carried stronger predictive power. In contrast, during the volatile post-COVID adjustment phase (2022.3–2023.2), the model downweighted BMOM, instead amplifying raw momentum, consistent with a heightened preference for directional persistence. These regime-specific signal shifts indicate that the models are not merely detecting momentum, but learning conditional structures in the futures curve—akin to nonlinear factor timing mechanisms grounded in market microstructure.

Fifth, while our reported returns are gross of transaction costs, we recognize that institutional implementation requires attention to turnover, slippage, and execution costs. Despite daily ranking monitoring, the portfolios exhibit relatively low churn, with over 70% overlap in long and short positions across consecutive days. Our internal turnover estimates suggest that monthly portfolio turnover averages between 25% and 35% depending on model-input combinations, indicating efficient rebalancing behavior. Given the high liquidity of the selected futures contracts, particularly in energy and metal sectors, modest round-trip transaction costs are unlikely to offset the observed performance gains. Nevertheless, explicitly modeling transaction costs or liquidity constraints in the training objective remains an important area for future research.

Sixth, subperiod analysis reveals that LSTM models trained on momentum outperform BMOM variants in highly volatile regimes (e.g., 2022.3–2023.2), whereas BMOM regains dominance in more stable environments (e.g., 2023.3–2024.1). This suggests that model effectiveness is not uniform, but context-sensitive—highlighting the importance of regime awareness in model interpretation.

Finally, the rank-correlation analysis (Exhibit 6) provides interpretability into model behavior. The fact that deep learning models exhibit stronger alignment between predicted and realized return rankings confirms that their superior portfolio performance stems not from extreme bets, but from improved asset ordering across the cross-section. This result validates the use of rank-based evaluation as a diagnostic complement to return-based metrics.

Taken together, these findings demonstrate that AI can meaningfully enhance—but not replace— economic reasoning in systematic investing. Effective integration requires careful attention to model architecture, training procedures, and especially the economic structure of input features. The strongest results emerge when financial domain knowledge informs the design of learning systems.

While this study focuses on the commodity futures market, the proposed framework is applicable to a broad range of asset classes where cross-sectional factors drive return dispersion. For example, in equity markets, combining valuation-based signals (e.g., earnings yield, book-to-market) with temporal momentum may yield benefits similar to basis-momentum in commodities. Likewise, in currency markets, term-structure analogues such as interest rate differentials and forward premiums can be integrated with sequence models to learn regime-aware carry strategies. The model architecture and ranking framework remain invariant; only the input features and asset universe need adaptation. This suggests that deep learning–based ranking systems—when grounded in domain-specific financial intuition—can provide a flexible foundation for cross-asset portfolio design and tactical allocation.

6 | CONCLUSION

This article examined the intersection of factor investing and deep learning within the commodity futures market. Our results confirm that classic factors, particularly basis-momentum, continue to deliver economically meaningful return spreads. These findings hold across multiple market environments and demonstrate the enduring relevance of cross-sectional signals rooted in futures term structure.

Building on this foundation, we integrated two deep learning architectures, LSTM and Transformer, into the asset ranking process. When trained on economically motivated inputs, these models consistently outperform traditional sorting rules. The improvements stem not only from enhanced return predictability, but also from the models' ability to adaptively filter signals based on underlying market conditions.

The key contribution of this study lies in showing that financial theory and artificial intelligence are not opposing paradigms, but complementary ones. Factor-informed models benefit from the interpretability and economic grounding of traditional signals, while leveraging the pattern-recognition capabilities of modern AI architectures.

For practitioners, this suggests that incorporating well-established factors as inputs into sequence learning models can yield robust and adaptive portfolio strategies. For researchers, it opens the door to further exploration of hybrid approaches that combine financial insight with machine learning flexibility.

Future work may extend this framework by incorporating alternative features such as macroeconomic indicators, inventory shocks, or inter-commodity relationships, and by evaluating reinforcement learning or graph-based models. Additionally, integrating transaction costs and liquidity constraints into training objectives will be crucial for real-world deployment.

These findings also suggest that deep sequence models—when guided by economically motivated inputs—can serve as a generalizable blueprint for ranking-based portfolio construction beyond commodity markets. In particular, asset classes characterized by return dispersion and temporally persistent signals, such as global equities or currencies, may benefit from similar architectures that combine financial intuition with adaptive learning.

In sum, the basis-momentum factor emerges as a particularly powerful tool—both as a traditional ranking signal and as a guiding input for AI-enhanced portfolio construction. When grounded in economic structure, artificial intelligence can amplify the edge provided by market fundamentals.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Alzaman, C. (2025). Optimizing portfolio selection through stock ranking and matching: A reinforcement learning approach. , *269*, 126430. doi: 10.1016/j.eswa.2023.126430

Bianchi, R. J., Drew, M. E., & Fan, J. H. (2016). Commodities momentum: A behavioral perspective. , *72*, 133–150. doi: 10.1016/j.jbankfin.2016.07.015

Binoy, S. J., & Jos, J. (2022). Financial market forecasting using macro-economic variables and rnn. In *2022 2nd international conference on advance computing and innovative technologies in engineering (icacite)* (pp. 1366–1371).

Boons, M., & Prado, M. P. (2019). Basis-momentum. , *74*(1), 239–279. doi: 10.1111/jofi.12738

Erb, C. B., & Harvey, C. R. (2006). The strategic and tactical value of commodity futures. , *62*(2), 69–97. doi: 10.2469/faj.v62.n2.4084

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. , *270*(2), 654–669. doi: 10.1016/j.ejor.2017.11.054

Gezici, A. H. B., & Sefer, E. (2024). Deep transformer-based asset price and direction prediction. , *12*, 24164–24178. doi: 10.1109/ACCESS.2024.3382662

Gorton, G., & Rouwenhorst, K. G. (2006). Facts and fantasies about commodity futures. , *62*(2), 47–68. doi: 10.2469/faj.v62.n2.4083

Hu, Y., & Ni, J. (2024). A deep learning-based financial hedging approach for the effective management of commodity risks. *Journal of Futures Markets*, *44*(6), 879–900. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/fut.22497` doi: 10.1002/fut.22497

Islam, M. S., & Hossain, E. (2021). Foreign exchange currency rate prediction using a gru-lstm hybrid network. , *3*, 100009. doi: 10.1016/j.sofcom.2021.100009

Kampotha, P., Wang, H., & Li, C. (2024). Transformer-based models for commodity trading price forecasting. In *2024 7th international conference on data science and information technology (dsit)* (pp. 1–5).

Kim, S.-H., & Kang, H.-G. (2014). A new strategy using term-structure dynamics of commodity futures. , *11*(3), 282–288. doi: 10.1016/j.frl.2013.11.007

Lo, A. W., & Singh, M. (2023). Deep-learning models for forecasting financial risk premia and their interpretations. , *23*(6), 917–929. doi: 10.1080/14697688.2023.2184433

López de Prado, M. (2018). *Advances in financial machine learning*. Wiley.

Miffre, J., & Rallis, G. (2007). Momentum strategies in commodity futures markets. , *31*(6), 1863–1886. doi: 10.1016/j.jbankfin.2006.12.007

Szymanowska, M., De Roon, F., Nijman, T., & Van Den Goorbergh, R. (2014). An anatomy of commodity futures risk premia. , *69*(1), 453–482. doi: 10.1111/jofi.12096

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. , *30*.

Wang, S., & Zhang, T. (2024). Predictability of commodity futures returns with machine learning models. *Journal of Futures Markets*, *44*(2), 302–322. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/fut.22477` doi: 10.1002/fut.22477

Wei, P., Cao, Y., & Dong, Y. (2024). *Deep latent factor models in asset pricing*. Working paper, available at SSRN: `https://ssrn.com/abstract=5084287`.