
Image-Encoded Financial Time Series as a Testbed for Self-Supervised Learning Paradigms

Ian Hajra

Department of Computer Science
Department of Applied Mathematics
Brown University
Providence, RI
ihajra@brown.edu

Abstract

We conduct a systematic paradigm-level evaluation of self-supervised learning (SSL) on financial time series encoded as images. Financial data poses a distinctive stress test for SSL: the signal-to-noise ratio is extremely low, statistical structure shifts across market regimes, and there is no semantic object for an encoder to latch onto. We convert price and return series into images via three complementary encodings (Gramian Angular Field and Markov Transition Field (GAF+MTF), candlestick renders, and a returns heatmap designed for multi-asset panels) and evaluate six SSL objectives spanning every major paradigm: contrastive learning (SimCLR), self-distillation (DINO), redundancy reduction (VICReg), masked reconstruction (MAE), and latent prediction (I-JEPA, LeJEPA). Representation quality is assessed via linear and k -nearest-neighbour probes on two economically motivated tasks: GICS sector classification on S&P 500 equities and Fama–French factor-recovery on 30-industry portfolios. Methods are ranked across all encoding and probe combinations, with lower average rank indicating more consistent performance. We find that LeJEPA achieves the best average rank across all encoding and probe combinations (2.3 versus 2.7 for a fully supervised baseline), and that the best-performing paradigm is encoding-dependent: LeJEPA leads on GAF+MTF and candlestick while SimCLR dominates the heatmap linear probe. Crucially, augmentation design must be matched to encoding structure, as no single augmentation strategy transfers across all encodings, and window size selection is non-trivial, with optimal horizons ranging from one month (candlestick) to six months (GAF+MTF). Our results demonstrate that self-supervised pre-training on image-encoded financial data can match or exceed supervised performance without any labels, and that the JEPA family of latent-prediction objectives is particularly well-suited to noisy, non-stationary financial imagery. Code: <https://github.com/ianhajra/senior-thesis>

1 Introduction

Self-supervised learning has reshaped representation learning in vision and language, enabling models to learn rich, transferable representations from unlabelled data that rival or exceed supervised alternatives at scale. Yet the financial domain has largely been bypassed by this revolution. Financial time series present a distinctive combination of properties that make them an important and underexplored testbed: an extremely low signal-to-noise ratio, pronounced non-stationarity across market regimes, the absence of any semantic object analogous to the ImageNet categories that anchor vision SSL benchmarks, and a small effective sample size despite long price histories. It is not obvious that SSL methods developed for natural images can extract any meaningful structure from such data.

We investigate this question through the lens of image-encoded financial time series. Rather than operating on raw one-dimensional sequences, we encode price and return series as two-dimensional images (Gramian Angular Fields and Markov Transition Fields (GAF+MTF), standard candlestick charts, and a returns heatmap for multi-asset panels) and apply the full modern vision SSL toolbox to these representations. This image-encoding pathway is our central methodological premise: it makes the entire ecosystem of vision SSL architectures, objectives, and evaluation protocols directly applicable to financial data without bespoke sequence modelling.

The primary research questions are:

1. **Data capacity:** which rolling window size best captures market structure, and does the answer depend on the encoding?
2. **Augmentations:** do standard vision augmentations transfer to financial imagery, or are domain-specific alternatives necessary?
3. **SSL paradigm:** which pre-training objective best learns representations that reflect known economic structure?

We do not attempt to predict asset returns, which would be inconsistent with the weak-form Efficient Market Hypothesis Fama [1970]. Instead, we ask whether learned representations recover known sector and factor structure, specifically higher-level market dynamics that reflect broad economic organisation rather than individual price movements. Concretely, we evaluate representation quality via two economically motivated probe tasks: 11-class GICS sector classification on S&P 500 individual equities, and Fama–French factor-recovery (5-class quintile classification) on 30-industry portfolio returns. Both probes are applied post hoc to frozen encoder features, with labels introduced only at probe time and never during SSL pre-training.

Our findings establish three principal conclusions. First, optimal window size and augmentation design are both encoding-dependent: no single configuration generalises across all three image representations. Second, across all encodings and probe combinations, the JEPA family of latent-prediction objectives achieves the strongest and most consistent results, with LeJEPA attaining the best average rank overall and narrowly outperforming a fully supervised ViT-Tiny baseline. Third, self-supervised pre-training on image-encoded financial data can match or exceed supervised performance without any labels, a result that holds most strongly on the returns heatmap, where every SSL method approaches or exceeds the supervised ceiling, and holds least strongly on candlestick images, where fine-grained local patterns benefit from direct label supervision.

2 Background and Related Work

2.1 Self-Supervised Learning

Modern SSL methods fall into five broad paradigms. *Contrastive learning* (SimCLR Chen et al. [2020], MoCo He et al. [2020]) pulls together representations of augmented views of the same instance while pushing apart views from different instances, using the InfoNCE Oord et al. [2018] loss; learned representations are therefore strongly dependent on the choice of augmentation pipeline. *Self-distillation* (BYOL Grill et al. [2020], DINO Caron et al. [2021]) trains an online student to match a slowly-updated EMA teacher, removing the need for explicit negatives via an asymmetric architecture and stop-gradient. *Redundancy reduction* (Barlow Twins Zbontar et al. [2021], VICReg Bardes et al. [2022]) instead operates on the statistics of the embedding space, pushing the cross-correlation matrix of two views toward the identity to simultaneously enforce invariance and decorrelate dimensions. *Masked autoencoding* (MAE He et al. [2022]) masks a large fraction of input patches and trains a lightweight decoder to reconstruct them in pixel space, forcing the encoder to build holistic representations from partial observations. *Joint-Embedding Predictive Architectures* (JEPAs) depart from all of the above by training a predictor to forecast the *latent representation* of masked regions from visible context, avoiding pixel-space reconstruction entirely; the prediction target is an abstract embedding rather than raw observations, so no encoder capacity is wasted on modelling irrelevant low-level noise. I-JEPA Assran et al. [2023] instantiates this principle for images; V-JEPA Bardes et al. [2024] extends it to video; LeJEPA Balestrierio and LeCun [2025] provides theoretical grounding by proving that an isotropic Gaussian is the provably optimal embedding distribution under downstream risk minimisation and introduces SIGReg, a linear-complexity regulariser that enforces this geometry

without stop-gradient or asymmetric teacher–student heuristics. The JEPA family is motivated by LeCun’s LeCun [2022] proposal for a non-generative, energy-based cognitive architecture operating entirely in latent space. Detailed descriptions of each paradigm are provided in Appendix A.

2.2 Vision Transformers

Vision Transformers (ViTs) Dosovitskiy et al. [2021] treat an image as a sequence of fixed-size patch tokens processed by a standard transformer encoder whose self-attention mechanism imposes no spatial inductive bias. ViT-Tiny is adopted as the shared backbone across all methods in this work to ensure that differences in downstream probe performance are attributable solely to the pre-training objective rather than to architectural capacity. Scaling experiments to larger backbone capacities are left to future work; see §6.

2.3 Encoding Financial Time Series as Images

A key premise of this work is that financial data can be encoded as images, enabling direct application of modern vision SSL methods without bespoke sequence modelling.

2.3.1 Gramian Angular Fields and Markov Transition Fields

Gramian Angular Fields (GAFs) encode a univariate time series as a square image by normalising the series to $[-1, 1]$, mapping each value $\tilde{x}_i$ to a polar angle $\phi_i = \arccos(\tilde{x}_i)$, and encoding pairwise temporal correlations as $\text{GASF}_{ij} = \cos(\phi_i + \phi_j)$ and $\text{GADF}_{ij} = \sin(\phi_i - \phi_j)$. The polar angle mapping preserves temporal ordering by construction, distinguishing GAFs from orderless statistical encodings. Markov Transition Fields (MTFs) complement GAFs by encoding regime-transition dynamics: the series is quantised into Q equal-frequency bins, a $Q \times Q$ empirical transition probability matrix $\mathbf{W}$ is computed, and the MTF image is defined entry-wise as $\text{MTF}_{ij} = W_{q(x_i), q(x_j)}$. GAFs and MTFs were introduced by Wang & Oates Wang and Oates [2015] as complementary representations designed to be used together; together they provide both the angular correlation structure and the quantised regime-transition dynamics of the series in a single multi-channel image.

2.3.2 Candlestick Rendering

Candlestick images rasterise the open, high, low, close, and volume (OHLCV) record of each bar into the two-dimensional format familiar to human chartists, preserving the visual grammar (ex. body height, wick length, color) through which technical analysts identify actionable patterns. This encoding is included as a strong naive baseline grounded in the technical analysis literature.

2.3.3 Returns Heatmap

The returns heatmap is a design contribution of the present work, motivated by the factor structure of equity returns established by Fama & French Fama and French [1993]. It arranges a panel of assets as a two-dimensional image in which rows correspond to assets, columns to time steps, and pixel intensity to the log return of each asset at each period. Each column is cross-sectionally z -scored at each time step to remove market-wide level shifts, exposing relative comovement between assets as spatial texture. The 30 industry portfolios are interpolated along the row axis to fill the 224×224 image height. This design is motivated by the observation that cross-sectional factor loadings manifest as horizontal banding in the resulting image, making latent factor structure directly legible as spatial texture and allowing an SSL encoder to recover factor exposure from visual pattern rather than explicit factor assignment. The z -scoring and rank-normalised intensity encoding are deliberate choices to separate cross-sectional variation from market-level noise.

2.4 SSL for Time Series and Finance

The dominant paradigm for self-supervised representation learning on time series operates on raw one-dimensional sequences, using augmentations such as jittering, scaling, and temporal cropping to construct positive pairs; representative methods include TS2Vec Yue et al. [2022] and TS-TCC Eldele et al. [2021]. These approaches inherit the augmentation-design bottleneck of their vision counterparts and do not benefit from the rapid architectural advances (such as masked autoencoders, JEPA-family objectives) developed for 2D image data. In finance specifically, Dolphin et al. Dolphin et al. [2024]

learn asset embeddings via contrastive learning and evaluate on hedging tasks, while temporal representation approaches Hwang et al. [2024] have highlighted concept drift as the central challenge for transfer. Concurrent work has extended JEPA objectives to 1D time series directly Girgis et al. [2024], Ennadir et al. [2025], He et al. [2026]. The contribution of the present work is an image-encoding pathway that opens the full modern vision SSL toolbox to financial sequence data, enabling a systematic paradigm-level comparison not possible within the 1D SSL literature. Detailed descriptions of related time-series SSL methods are provided in Appendix B.

2.5 Financial Datasets and Asset Pricing

The three-factor model of Fama & French Fama and French [1993] decomposes excess returns into market, size, and value components; the five-factor extension Fama and French [2015] adds profitability and investment factors. The Kenneth French Data Library provides canonical, reproducible factor returns and pre-formed portfolio series at daily frequency. Fama’s Fama [1970] weak-form efficiency hypothesis holds that current prices fully reflect all past price information; cross-sectional factor structure, however, is consistent with efficient markets as compensation for systematic risk. It is precisely this persistent cross-sectional signal, not time-series predictability, that our SSL representations are designed to recover.

2.6 Evaluating Self-Supervised Representations

A linear probe trains a single linear layer on top of frozen encoder features, measuring linear decodability without backbone fine-tuning. kNN probing assigns class membership by majority vote among the k nearest neighbours in feature space, providing a purely geometric evaluation free from optimisation artefacts; together, these constitute the standard two-tier protocol for SSL representation quality Kolesnikov et al. [2019]. RankMe Garrido et al. [2023] estimates representation quality without labels by computing the effective rank of the empirical feature covariance matrix via the entropy of its normalised singular value spectrum. High effective rank indicates that the encoder distributes information across many dimensions and has not undergone dimensional collapse; low effective rank signals representational degeneration. RankMe is integrated as an online training callback to provide a collapse-detection signal that complements the labelled probe evaluations.

3 Methods

The experimental design pairs two complementary financial datasets with a battery of image encodings and six SSL pre-training objectives spanning every major paradigm. The Fama–French 30-industry portfolios provide a multi-asset, portfolio-level testbed whose natural factor structure gives rise to an interpretable representation-quality probe; the S&P 500 individual equity series provide an asset-level testbed whose GICS sector labels constitute a second, independently motivated probe task. Together the two datasets enable generalisation claims that neither alone could support. All images are pre-rendered to disk before any training begins; training jobs consume standard image datasets with no domain-specific encoding at runtime.

3.1 Unified Training Infrastructure

All experiments are implemented within `stable-pretraining` Balestrieri et al. [2025], a unified SSL training library that provides a common forward-function interface for SimCLR, DINO, VICReg, MAE, I-JEPA, and LeJEPA, together with a shared run registry and standardised logging. The library couples each pre-training run with `OnlineProbe`, `OnlinekNN`, and `RankMe` callbacks that evaluate linear decodability, neighbourhood geometry, and effective rank on a held-out validation set at every checkpoint without interrupting training. Adopting a single training infrastructure is the core methodological control of this work: because all methods share the same data pipeline, augmentation schedule, optimiser configuration, backbone, and evaluation protocol, observed differences in downstream probe performance are attributable to the pre-training objective alone.

3.2 Datasets

3.2.1 Fama–French 30-Industry Portfolios

We use the value-weighted returns of the Fama–French 30-industry portfolio series, sourced from the Kenneth French Data Library at daily frequency, spanning July 1963 through December 2023, yielding approximately 15,100 trading days across 30 industry portfolios after standard preprocessing. The multi-asset structure of this dataset motivates the returns heatmap encoding described in §3.3.2.

3.2.2 S&P 500 Individual Equities

Daily OHLCV data for S&P 500 constituent equities are sourced via `yfinance` over the period January 2013 through December 2023, yielding approximately 2,500 trading days per ticker. The 2013–2023 window encompasses multiple distinct market regimes including a sustained bull market, the COVID-19 shock of March 2020, and the 2022 rate-hiking cycle. Sector labels are obtained from the GICS classification system via the `yfinance.info` field, yielding 11 sector classes used exclusively as probe targets at evaluation time. This dataset is subject to survivorship bias, as only current index constituents are included; see §6. Dataset statistics and sector distributions are provided in Tables 3–6 of Appendix D.

3.2.3 Pre-Rendering and Dataset Construction

All rolling window images are computed offline and saved to disk prior to any training run. For the S&P 500, rolling windows are applied to each ticker independently; the full pool of resulting images is randomly subsampled to 45,000 images. For the Fama–French dataset, rolling windows are applied to the full 30-asset panel with a stride of 2 trading days, yielding approximately 5,000 images. A metadata file records the ticker or dataset identifier, window date range, GICS sector label or factor quintile labels, and split assignment for each image.

3.2.4 Train / Validation / Test Split

Both datasets are partitioned by a strictly chronological temporal split into train (70%), validation (10%), and test (20%) sets. Temporal ordering is preserved within each split to prevent look-ahead leakage. SSL pre-training is conducted exclusively on the training split. Probe labels are introduced only at probe-training time using the validation set; final evaluation is reported on the held-out test set. For the Fama–French factor-recovery probe, quintile bin boundaries are computed on the training split and applied without recomputation to validation and test, preventing distributional leakage across the temporal boundary.

3.3 Time-Series-to-Image Encodings

Each dataset is encoded into images of resolution 224×224 pixels using a rolling window applied along the time axis. All encodings are computed at the window sizes identified by the ablation in §4.1.

3.3.1 Encodings for S&P 500

GAF+MTF. Following Wang & Oates Wang and Oates [2015], the series is normalised to $[-1, 1]$ and mapped to polar angles $\phi_i = \arccos(\tilde{x}_i)$. Channel 1 encodes the Gramian Angular Summation Field $\text{GASF}_{ij} = \cos(\phi_i + \phi_j)$; channel 2 encodes the Gramian Angular Difference Field $\text{GADF}_{ij} = \sin(\phi_i - \phi_j)$; channel 3 encodes the Markov Transition Field $\text{MTF}_{ij} = W_{q(x_i), q(x_j)}$ with $Q = 8$ equal-frequency bins. The three-channel stacking maps naturally to an RGB image for ViT input.

Candlestick. The OHLCV record of each bar is rasterised at 224×224 using the standard candlestick convention: body height encodes the open-to-close range, with green (red) bodies indicating close above (below) open; upper and lower wicks encode the high and low; a volume bar is rendered at the bottom.

3.3.2 Encodings for Fama–French

Returns Heatmap. Rows correspond to the 30 industry portfolios, columns to time steps within the rolling window, and pixel intensity to the log return of portfolio i at time j . Each column is cross-sectionally z -scored to remove market-wide level shifts, exposing relative comovement as spatial texture. Industry portfolios are interpolated along the row axis to fill the 224×224 image height. Cross-sectional comovement and factor loadings manifest as horizontal banding in this representation, allowing an SSL encoder to recover factor exposure from visual pattern rather than explicit factor assignment.

3.4 SSL Pre-Training Methods

All methods share a ViT-Tiny backbone trained from random initialisation. The shared optimiser is AdamW with base learning rate 1.5×10^{-4} (linear scaling from 10^{-4} at batch size 256), weight decay 0.05, batch size 256, and a cosine learning rate schedule with a 10-epoch linear warmup. All method-specific hyperparameters follow each method’s original paper defaults. One method per paradigm is included to ensure that observed differences are attributable to the pre-training objective rather than architectural capacity; two JEPA variants are included because the JEPA family is the primary theoretical focus of this work. A supervised ViT-Tiny baseline trained from scratch on the probe labels is included as a reference ceiling; all SSL results are reported relative to this baseline. Table 1 summarises all methods; full hyperparameter tables are provided in Appendix F.

Table 1: Summary of methods. Full training and method-specific hyperparameters are reported in the corresponding tables referenced in the final column.

Method	Paradigm	Objective	Epochs	Hyperparameters
SimCLR	Contrastive	InfoNCE	100	See Table 10
VICReg	Redundancy-reduc.	Var./Inv./Cov.	100	See Table 11
DINO	Self-distillation	Cross-entropy	100	See Table 12
MAE	Masked recon.	Pixel MSE	100	See Table 13
I-JEPA	JEPA	Latent MSE	100	See Table 14
LeJEPA	JEPA + SIGReg	Latent MSE + SIGReg	100	See Table 15
Supervised	Baseline	Cross-entropy	100	See Table 16

3.5 Augmentation Pipeline

Standard vision augmentations encode natural-image priors that are harmful for financial imagery: horizontal flipping destroys temporal ordering in GAF and MTF images; color jitter corrupts semantically meaningful pixel intensities such as the red/green body convention in candlestick renders; and random resized cropping destroys the global pairwise correlation structure that GAF and MTF encodings are designed to preserve. We therefore define a curated set of domain-compatible candidate augmentations and select the best composition via greedy forward selection on SimCLR, the paradigm most sensitive to augmentation design. The winning augmentation configuration is applied uniformly across all methods.

3.5.1 Candidate Augmentations

Eight candidate augmentations are considered: four drawn from the standard SimCLR pipeline (Random Resized Crop, Horizontal Flip, Color Jitter, Random Grayscale) and four designed specifically for financial imagery. *Gaussian Blur* attenuates microstructure noise artefacts without destroying global temporal structure. *Magnitude Scaling* multiplies pixel intensities by a scalar drawn from $[0.8, 1.2]$, encoding the intuition that the same pattern at different volatility regimes should yield similar representations. *Gaussian Noise Injection* adds i.i.d. Gaussian noise with $\sigma = 0.05$, motivated by the microstructure noise present in financial data. *Temporal Masking* zeros out a randomly positioned contiguous block of time columns spanning 10–20% of image width, improving robustness to missing data.

3.5.2 Augmentation Selection Protocol

Greedy forward selection is performed on SimCLR trained for 100 epochs. At each step, the remaining candidate augmentation that yields the highest balanced accuracy on the validation linear probe is added; selection terminates when adding any further augmentation fails to improve validation probe accuracy. The final augmentation compositions identified per encoding are: *GAF+MTF*: Random Resized Crop; *Candlestick*: Color Jitter, Gaussian Blur, Gaussian Noise, Random Resized Crop; *Heatmap*: Random Resized Crop, Temporal Masking, Color Jitter, Gaussian Noise. These configurations are applied uniformly to all pre-training methods.

3.6 Window Size Ablation

Rolling window size controls the temporal resolution of each encoded image and is treated as an encoding-specific hyperparameter selected via the validation linear probe. Four candidate sizes are evaluated: 20 days (≈ 1 month), 63 days (1 quarter), 126 days (6 months), and 252 days (1 year). Ablation is conducted on SimCLR (100 epochs) on both datasets using the validation linear probe as the selection criterion. The selected window sizes are: 126 trading days for GAF+MTF, 20 trading days for candlestick, and 20 trading days for the returns heatmap. Full ablation results are reported in §4.1.

3.7 Evaluation Protocol

All evaluation metrics are computed on the held-out test split using frozen encoder features; no backbone fine-tuning is performed at any stage.

3.7.1 Downstream Probe Tasks

Fama–French: Factor-Recovery Probe. For each rolling window in the Fama–French dataset, the cumulative return of each of the five Fama–French factors (Mkt-RF, SMB, HML, RMW, CMA) over that window is computed and assigned to one of five quintile bins determined from the training split. This yields a five-class label per factor per window; the probe task is five-class classification evaluated independently per factor and reported as balanced accuracy *averaged across all five factors*, providing a single summary score per method and encoding. The cumulative-return labelling scheme is preferred over within-window beta estimation because beta estimation on short windows introduces substantial sampling noise.

S&P 500: Sector Classification. Each rolling window is assigned the GICS sector label of the underlying equity, yielding an 11-class classification task. Balanced accuracy is used because the overrepresentation of the Information Technology sector in the S&P 500 makes raw accuracy a misleading summary statistic.

3.7.2 Linear Probe and kNN

The linear probe trains a single linear layer on top of frozen encoder features for 100 epochs using AdamW with learning rate 10^{-3} , and reports balanced accuracy on the test set. The kNN probe assigns class membership by majority vote over the $k = 20$ nearest neighbours in cosine-distance feature space. Both are integrated as `OnlineProbe` and `OnlinekNN` callbacks evaluated at every checkpoint on the validation set during pre-training.

3.7.3 Collapse Diagnostics via RankMe

RankMe Garrido et al. [2023] is logged at every checkpoint as an online collapse-detection signal. A persistently low RankMe score indicates that the encoder’s embedding space has degenerated onto a low-dimensional manifold; comparing RankMe trajectories alongside probe accuracy curves disambiguates cases where one configuration achieves higher probe accuracy by collapsing less rather than by encoding more discriminative structure.

4 Experiments and Results

We report results from three studies conducted in sequence: (1) a window-size ablation to identify the optimal temporal horizon per encoding; (2) an augmentation study to identify domain-compatible transforms; and (3) a paradigm-level SSL comparison using the configurations selected in studies 1 and 2.

4.1 Study 1: Window Size

Setup. SimCLR with a ViT-Tiny backbone is trained for 100 epochs at each of four window sizes $\in \{20, 63, 126, 252\}$ trading days for all three encodings, with all other hyperparameters held fixed (see Table 8 in Appendix F).

Results. Probe accuracy and RankMe curves are shown in Figures 4, 5, and 6. For GAF+MTF, validation linear probe accuracy increases monotonically with window size up to 126 days and plateaus at 252 days, with the 126-day window yielding the best tradeoff between probe performance and representational diversity as measured by RankMe. For candlestick, the 20-day window dominates on both linear and kNN probes; longer windows compress more bars into the same image resolution, blurring the local patterns that carry discriminative information. For the returns heatmap, performance is close across the 20-day and 126-day windows; 20 trading days is selected after comparing both probe metrics and RankMe trajectories.

Collapse diagnostics. No representational collapse is detected on GAF+MTF or candlestick across any window size. The heatmap encoding exhibits partial collapse at some window sizes (RankMe scores decline in early training before recovering) yet probe accuracy continues to separate across window sizes.

Selected configurations. GAF+MTF: 126 trading days (6 months). Candlestick: 20 trading days (1 month). Returns heatmap: 20 trading days (1 month).

4.2 Study 2: Augmentations

Setup. Using the window sizes selected in Study 1, each of the eight candidate augmentations is evaluated individually on SimCLR (100 epochs), and augmentations are added greedily until no further gain is observed on the validation linear probe.

Results. Probe curves for the augmentation study are shown in Figures 10, 11, and 12. For GAF+MTF, spatial augmentation alone suffices: Random Resized Crop is the single best transform and no additional augmentation provides consistent improvement, consistent with GAF+MTF encoding global temporal correlation structure that is robust to mild spatial perturbation. For candlestick, a richer mix is beneficial (Color Jitter, Gaussian Blur), Gaussian Noise, and Random Resized Crop are all selected. This reflects the fact that candlestick renders contain both visual (color) and magnitude (volatility) information that respond to different invariance pressures. For the heatmap, the broadest set is selected (Random Resized Crop, Temporal Masking, Color Jitter, Gaussian Noise), suggesting that multivariate cross-sectional structure benefits from both spatial perturbations and finance-specific transforms; notably, Temporal Masking is uniquely helpful for heatmap encodings and harmful for the single-asset encodings where temporal ordering is critical.

Collapse diagnostics. Several augmentations trigger partial collapse on GAF+MTF and candlestick (most notably Horizontal Flip, which destroys temporal ordering in both encodings), confirmed by declining RankMe trajectories. The selected augmentation configurations exhibit no collapse for any encoding.

Key finding. No single augmentation strategy transfers across all three encodings; the optimal choice is tightly coupled to the structure of the encoding.

4.3 Study 3: SSL Paradigm Comparison

Setup. All six SSL methods and the supervised baseline are trained for 100 epochs using the encoding-specific window sizes and augmentation configurations identified in Studies 1 and 2.

Results. Training curves are shown in Figures 13, 14, and 15. Summary balanced accuracy at the end of training is reported in Table 2.

Table 2: SSL comparison: balanced accuracy (%) via linear probe (Lin.) and kNN at the end of 100 epochs. **Bold:** highest value per column. Underline: second highest. Avg. Rank: mean rank across all six probe columns (1 = best; lower is better); Random is excluded from ranking. Random baselines are $1/11 \approx 9.1\%$ for GAF+MTF and Candlestick and $1/5 = 20.0\%$ for Heatmap. Linear probe accuracy is the running average of the last 10 epochs to reduce noise. Heatmap results are balanced accuracy averaged across all five Fama–French factors.

Method	GAF+MTF		Candlestick		Heatmap		Avg. Rank
	Lin.	kNN	Lin.	kNN	Lin.	kNN	
SimCLR	<u>10.3</u>	9.3	10.0	9.6	22.5	19.5	3.3
VICReg	9.4	9.0	9.6	9.6	<u>21.3</u>	18.3	4.7
DINO	9.2	9.2	9.5	<u>10.1</u>	20.0	19.4	5.0
MAE	9.3	9.3	9.2	9.3	19.9	<u>20.8</u>	4.8
I-JEPA	9.2	9.3	9.3	9.3	19.6	21.1	5.2
LeJEPA	10.4	<u>9.4</u>	<u>10.7</u>	10.0	21.2	20.1	2.3
Supervised	9.9	9.6	11.3	11.3	20.8	19.2	<u>2.7</u>
Random	9.1	9.1	9.1	9.1	20.0	20.0	—

GAF+MTF. LeJEPA leads the linear probe (10.4%), with SimCLR close behind (10.3%); under kNN, supervised dominates (9.6%) and LeJEPA ranks second (9.4%). I-JEPA and MAE exhibit low and flat RankMe scores consistent with representational collapse, suggesting their latent spaces lack the diversity needed to separate financial categories on this encoding.

Candlestick. Supervised is the clear leader on both probes (11.3%); among SSL methods, LeJEPA is the strongest on the linear probe (10.7%) and DINO on kNN (10.1%). The larger gap between supervised and SSL on the kNN probe indicates that local neighbourhood structure benefits more from label supervision here, possibly because candlestick patterns are local, short-horizon signals more precisely aligned with sector categories.

Heatmap. SimCLR leads the linear probe (22.5%) with VICReg second (21.3%); under kNN, I-JEPA edges ahead (21.1%) followed by MAE (20.8%). The strong performance of contrastive and redundancy-reduction methods on this encoding is consistent with the heatmap capturing cross-sectional comovement that is well-served by the global view invariances these objectives enforce. All SSL methods approach or exceed the supervised baseline (20.8% linear, 19.2% kNN).

Overall. LeJEPA achieves the best average rank (2.3) across all encoding-probe combinations, narrowly outperforming even the supervised baseline (2.7). The best-performing paradigm is encoding-dependent: LeJEPA leads on GAF+MTF and candlestick, while SimCLR dominates the heatmap linear probe. This interaction indicates that paradigm choice and encoding design are not independent decisions.

Collapse diagnostics. I-JEPA and MAE show evidence of representational collapse (low, flat RankMe) on GAF+MTF and candlestick, yet still outperform the random baseline. LeJEPA maintains high RankMe throughout training on all encodings, consistent with SIGReg’s theoretical guarantee of isotropic Gaussian embedding geometry.

5 Discussion

Do SSL representations capture market structure? Yes, across all three encodings and both probe tasks, every SSL method outperforms the chance-level random baseline, confirming that self-supervised pre-training on image-encoded financial data extracts economically meaningful structure despite the low signal-to-noise ratio. Gains are modest on GAF+MTF and candlestick (roughly 0.1–1.3 percentage points above random for SSL methods) but substantially larger on the heatmap (1–2.5 percentage points), suggesting that cross-sectional factor structure is more accessible to SSL objectives than individual-asset temporal patterns.

Which paradigm wins? Latent prediction (LeJEPA) achieves the best average rank across all encoding-probe combinations and is the only SSL method to outperform the supervised baseline in aggregate. This is consistent with the theoretical motivation of the JEPA family: by predicting in latent space rather than pixel space, LeJEPA is not penalised for the irreducible microstructure noise present in financial imagery. SIGReg’s isotropic Gaussian regularisation further prevents the representational collapse that afflicts I-JEPA and MAE, yielding diverse and discriminative embeddings throughout training. On the heatmap, however, SimCLR is the strongest method: cross-sectional comovement is a global view-level signal that contrastive objectives are explicitly designed to capture, and this advantage outweighs LeJEPA’s noise-robustness on that encoding. Paradigm choice should therefore be made jointly with encoding choice, not independently.

How close is SSL to supervised? Surprisingly close. LeJEPA’s average rank (2.3) surpasses the supervised baseline (2.7), though the gap widens on candlestick where supervised holds a clear advantage on both probes (11.3% vs. 10.7% linear, 11.3% vs. 10.1% kNN). The pattern is consistent with supervised learning retaining an edge when the probe target is tightly coupled to fine-grained local patterns, but SSL being competitive or superior when the relevant signal is more diffuse and global.

The role of augmentation design. No single augmentation strategy transfers across all encodings. Horizontal flipping, harmless or helpful on natural images, systematically degrades performance on GAF+MTF and candlestick by inverting temporal ordering. Domain-specific augmentations (Temporal Masking, Magnitude Scaling) provide complementary benefits for the heatmap and candlestick encodings. Future work should design augmentation curricula jointly with the encoding rather than importing natural-image pipelines wholesale.

Representational collapse. I-JEPA and MAE exhibit collapse on the single-asset encodings. This may reflect the combination of small backbone capacity (ViT-Tiny), limited dataset size, and low signal-to-noise ratio. LeJEPA’s SIGReg addresses this directly, reinforcing the practical value of theoretically motivated regularisation in data-scarce, high-noise regimes.

Practical recommendations. For a practitioner encoding individual equity series with no label budget, LeJEPA on a 126-day GAF+MTF window is the most robust choice and is the only configuration that consistently matches or exceeds supervised performance on the single-asset task. For multi-asset factor-structured data such as industry portfolio panels, SimCLR on a 20-day heatmap provides the strongest linear-probe performance and requires no specialised objective beyond the standard contrastive loss. In both settings, augmentation pipelines must be designed for the encoding geometry; the standard SimCLR natural-image pipeline should not be used without modification.

6 Limitations and Future Work

Survivorship Bias. The S&P 500 dataset comprises only current index constituents, excluding equities that were delisted, acquired, or removed over the study period. Future work should replicate experiments on a point-in-time constituent database.

Structural Breaks. Rolling windows that span major market dislocations produce encoded images whose statistical properties differ markedly from the bulk of the training distribution. Future work should evaluate probe performance conditioned on structural break periods to assess regime-robustness directly.

Single Temporal Split. All results are reported on a single chronological split, providing no variance estimate for probe accuracy. Differences in balanced accuracy below approximately two percentage points should be treated as inconclusive. A rolling-origin evaluation protocol with multiple non-overlapping test windows would provide confidence intervals over representation quality estimates.

Backbone Capacity. All methods are evaluated with a ViT-Tiny backbone for compute tractability. Paradigm-level differences may be attenuated or reversed at larger backbone capacities (ViT-S, ViT-B), particularly for JEPA objectives that theoretically benefit from high-capacity predictors Assran et al. [2023].

Method Coverage. Several prominent SSL methods are not evaluated, including BarlowTwins, MoCo, and data2vec. The present selection covers one representative per major paradigm; extending the comparison would strengthen paradigm-level conclusions.

GAF+MTF Channel Stacking. The three-channel GASF+GADF+MTF stacking is a fixed design choice following Wang & Oates Wang and Oates [2015]; the relative contribution of each channel is not ablated. Future work could decompose encoding contributions via single-channel ablations or learned channel weighting.

Scope of Financial Data. The present work is restricted to equity markets. Extension to fixed-income, foreign exchange, and commodity time series would test whether the findings generalise across asset classes with structurally different return distributions.

World Model Downstream Task. The evaluation protocol is limited to static representation quality via linear and kNN probes. A natural extension is to use pretrained encoders as inputs to a latent world model LeCun [2022], Zhou et al. [2024], asking whether JEPA-pretrained features improve rollout quality in a financial sequential decision-making setting.

7 Conclusion

We have presented a systematic paradigm-level evaluation of self-supervised learning on financial time series encoded as images. Our three main findings are as follows.

Window size and augmentation design are encoding-dependent. There is no universal temporal horizon or augmentation strategy that transfers across all three encodings. Optimal window sizes range from one month for candlestick and heatmap encodings to six months for GAF+MTF, and augmentation composition must be tailored to encoding geometry: horizontal flipping destroys temporal ordering and is actively harmful for GAF+MTF and candlestick, while Temporal Masking is uniquely beneficial for the heatmap encoding. Encoding and augmentation pipeline should be treated as jointly optimised design decisions.

LeJEPA is the best overall SSL paradigm for financial imagery, with important encoding-level variation. Across all three encodings and both probe tasks, LeJEPA achieves the best average rank (2.3), narrowly outperforming even the fully supervised baseline (2.7). Its advantage is most pronounced on the single-asset encodings (GAF+MTF and candlestick), where SIGReg prevents the representational collapse that afflicts I-JEPA and MAE, and pixel-space reconstruction is penalised for modelling microstructure noise. On the heatmap encoding, however, SimCLR is the strongest method: the global cross-sectional comovement captured by the heatmap aligns naturally with contrastive invariance objectives, and practitioners should prefer SimCLR in that setting.

Self-supervised pre-training can match or exceed supervised performance on financial imagery without any labels. Every SSL method outperforms the chance-level random baseline across all encodings and probe tasks. The SSL-supervised gap is smallest on the heatmap encoding, where all SSL methods approach or exceed the supervised baseline, and largest on the candlestick kNN probe. These results suggest that self-supervised pre-training on image-encoded financial data is a viable and practically attractive alternative to supervised learning in settings where labels are scarce, delayed, or expensive to obtain.

Ethics and Scope. This work is intended solely as an empirical benchmark of self-supervised representation learning methods on financial imagery. No part of this work is intended for real-world trading, investment decision-making, or any live financial application. All datasets used are publicly available research resources; results should be interpreted as measures of representational structure recovery, not as evidence of return predictability.

References

- Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023. URL <https://arxiv.org/abs/2301.08243>.
- Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. URL <https://arxiv.org/abs/2511.08544>.
- Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023. URL <https://arxiv.org/abs/2304.12210>.
- Randall Balestriero, Hugues Van Assel, Sami BuGhanem, and Lucas Maes. stable-pretraining-v1: Foundation model research made simple. *arXiv preprint arXiv:2511.19484*, 2025. URL <https://arxiv.org/abs/2511.19484>.
- Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022. URL <https://arxiv.org/abs/2105.04906>.
- Adrien Bardes, Quentin Garrido, Jean Ponce, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. URL <https://arxiv.org/abs/2404.08471>.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. URL <https://arxiv.org/abs/2104.14294>.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. URL <https://arxiv.org/abs/2002.05709>.
- Rian Dolphin, Barry Smyth, and Ruihai Dong. Contrastive learning of asset embeddings from financial time series. In *ICAIF*, 2024. URL <https://arxiv.org/abs/2407.18645>.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *ICLR*, 2021. URL <https://arxiv.org/abs/2010.11929>.
- Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *ICCV*, 2021. URL <https://arxiv.org/abs/2104.14548>.
- Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *IJCAI*, 2021. URL <https://arxiv.org/abs/2106.14112>.
- Sofiane Ennadir, Siavash Golkar, and Leopoldo Sarra. Joint embeddings go temporal. *arXiv preprint arXiv:2509.25449*, 2025. URL <https://arxiv.org/abs/2509.25449>.
- Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417, 1970.
- Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

- Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.
- Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *ICML*, pages 10929–10974, 2023. URL <https://arxiv.org/abs/2210.02885>.
- Abanoub M. Girgis, Alvaro Valcarce, and Mehdi Bennis. Time-series JEPa for predictive remote control under capacity-limited networks. *arXiv preprint arXiv:2406.04853*, 2024. URL <https://arxiv.org/abs/2406.04853>.
- Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. URL <https://arxiv.org/abs/2006.07733>.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. URL <https://arxiv.org/abs/1911.05722>.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. URL <https://arxiv.org/abs/2111.06377>.
- Yanan He, Yunshi Wen, Xin Wang, and Tengfei Ma. Mts-jepa: Multi-resolution joint-embedding predictive architecture for time-series anomaly prediction, 2026. URL <https://arxiv.org/abs/2602.04643>.
- Yoontae Hwang, Stefan Zohren, and Yongjae Lee. Temporal representation learning for stock similarities and its applications in investment management. 2024. URL <https://arxiv.org/abs/2407.13751>.
- Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *CVPR*, 2019.
- Yann LeCun. A path towards autonomous machine intelligence. Technical report, 2022. URL <https://openreview.net/forum?id=BZ5a1r-kVsf>.
- Lucas Maes, Ian K Hajra, Arnesh Batra, Hugues Van Assel, Damien Scieur, and Randall Balestriero. Relational representation learning. In *UniReps: 3rd Edition of the Workshop on Unifying Representations in Neural Models*, 2025. URL <https://openreview.net/forum?id=NkbiQOn5SI>.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *NeurIPS*, 2018. URL <https://arxiv.org/abs/1807.03748>.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Ghosh, Lexi Clark, Hugo Touvron, Ishan Misra, Julien Mairal, Piotr Bojanowski, Armand Joulin, and Herve Jegou. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.
- Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.
- Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *ICLR*, 2021. URL <https://arxiv.org/abs/2106.00750>.
- Zhiguang Wang and Tim Oates. Imaging time-series to improve classification and imputation. In *IJCAI*, pages 3939–3945, 2015. URL <https://arxiv.org/abs/1506.00327>.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. In *AAAI*, 2022. URL <https://arxiv.org/abs/2106.10466>.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021. URL <https://arxiv.org/abs/2103.03230>.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. URL <https://arxiv.org/abs/2411.04983>.

A Detailed SSL Background

This appendix provides detailed descriptions of each SSL paradigm surveyed in §2.1, following the taxonomy of Balestrieri et al. Balestrieri et al. [2023]. The five paradigms differ primarily in how they define the self-supervised signal and how they prevent representational collapse; understanding these differences is essential context for interpreting the paradigm-level results in §4.3.

A.1 Contrastive Learning

Contrastive self-supervised learning trains an encoder by pulling together representations of *positive pairs* (two views of the same instance generated by stochastic augmentation) while pushing apart *negative pairs* drawn from the rest of the batch, enforcing an invariance objective over the augmentation distribution via the InfoNCE Oord et al. [2018] loss, which lower-bounds mutual information between the two views. SimCLR Chen et al. [2020] demonstrated that composing aggressive augmentations, a non-linear projection head, and large batches (≥ 4096) are jointly sufficient to obtain strong linear-evaluation accuracy. MoCo He et al. [2020] decouples representation quality from batch size by maintaining a momentum-updated encoder and a FIFO queue of negatives. NNCLR Dwibedi et al. [2021] further relaxes the assumption that positive pairs must originate from the same image, substituting one view with its nearest neighbour from a memory bank. Because all three methods inherit an implicit prior that natural-image distortions preserve semantic content, augmentation design remains an Achilles heel of contrastive SSL when transferred to non-natural-image domains.

A.2 Self-Distillation

Self-distillation methods train an online *student* encoder to match the representations produced by a *teacher* network updated as an exponential moving average (EMA) of the student. A stop-gradient operator on the teacher branch breaks symmetric gradient flow that would otherwise cause representational collapse, removing the need for explicit negative pairs. BYOL Grill et al. [2020] instantiates this framework with an additional predictor MLP appended to the student branch. DINO Caron et al. [2021] extends self-distillation to a multi-crop, cross-entropy-based objective and shows that a ViT trained under this regime develops spatially coherent attention maps that segment foreground objects without any segmentation label. DINOv2 Oquab et al. [2023] reintroduced a curated large-scale dataset pipeline and register tokens to stabilise patch-level features; DINOv3 Siméoni et al. [2025] further scales dataset and model size and introduces Gram anchoring to prevent dense feature degradation.

A.3 Redundancy Reduction

Redundancy-reduction methods operate directly on the statistics of the embedding space, pushing the cross-correlation matrix of two augmented views’ representations toward the identity. Barlow Twins Zbontar et al. [2021] penalises diagonal entries deviating from one and off-diagonal entries deviating from zero. VICReg Bardes et al. [2022] decomposes the objective into three explicit terms: *Variance* (prevents dimensional collapse), *invariance* (MSE between the two views’ embeddings), and *covariance* (penalises off-diagonal covariance), making each regularisation pressure independently inspectable.

A.4 Masked Reconstruction

Masked autoencoding methods randomly mask a large fraction of input patches and train a lightweight decoder to reconstruct the missing content in pixel space. MAE He et al. [2022] demonstrated that this simple recipe applied to ViTs with 75% random patch masking yields representations competitive with supervised pre-training. In financial time-series domains, however, pixel-space reconstruction conflates the learning objective with the modelling of microstructure noise, motivating the JEPA family’s shift toward latent-space prediction.

A.5 Joint-Embedding Predictive Architectures (JEPA)

JEPAs train a predictor to forecast the latent representation of a masked or withheld region from the representations of visible context, discarding the requirement to model irrelevant low-level detail. Masking-based views are domain-agnostic—the only inductive bias is spatial or temporal locality—directly resolving the augmentation-design bottleneck that afflicts contrastive methods in non-natural-image domains. The JEPA family is motivated by LeCun’s LeCun [2022] proposal for a non-generative, energy-based cognitive architecture in which a hierarchy of world models (H-JEPA) operate entirely in latent space. I-JEPA Assran et al. [2023] instantiates this principle for static images by training a context encoder to predict EMA-teacher representations of multiple target blocks. V-JEPA Bardes et al. [2024] extends the framework to video via spatiotemporal tube masking. LeJEPA Balestrieri and LeCun [2025] provides theoretical grounding by proving that an isotropic Gaussian is the provably optimal embedding distribution under downstream risk minimisation, and introduces SIGReg—a linear-complexity regulariser that enforces this geometry—eliminating the stop-gradient and asymmetric teacher–student heuristics that prior JEPA variants inherited from self-distillation.

B Detailed Time-Series SSL Background

B.1 General Time-Series Representation Learning

TS2Vec Yue et al. [2022] learns hierarchical timestamp-level representations by applying contextual contrasting across temporal scales. TS-TCC Eldele et al. [2021] combines temporal contrasting with contextual contrasting between weakly and strongly augmented views. TNC Tonekaboni et al. [2021] constructs positive pairs by exploiting the quasi-stationarity of local temporal neighbourhoods.

B.2 JEPA for Time Series

Girgis et al. Girgis et al. [2024] propose TS-JEPA for predictive remote control, encoding high-dimensional sensory data into low-dimensional semantic embeddings and predicting future embeddings from current ones within latent space. Ennadir et al. Ennadir et al. [2025] independently propose an adaptation of JEPA for general time-series representation learning, validating on classification and forecasting benchmarks. Concurrent work has further extended the framework to anomaly prediction in industrial multivariate series via multi-resolution latent prediction He et al. [2026]. The present work is distinguished from all of these in that we apply JEPA objectives to *image-encoded* financial time series rather than raw 1D sequences, enabling the full vision SSL backbone ecosystem and a systematic cross-paradigm comparison under a controlled experimental infrastructure.

C A Note on SimCLR’s Apparent Strength

SimCLR performs notably well in several settings—most prominently leading the heatmap linear probe (22.5%) with a substantial margin—and achieves a competitive average rank of 3.3. Two factors jointly explain this result.

Reason 1: the experimental pipeline was optimised for SimCLR. Studies 1 and 2 both use SimCLR as the reference method for hyperparameter search. Window sizes and augmentation compositions are identified via SimCLR’s validation linear probe and then applied uniformly to all methods in Study 3. Methods that are less sensitive to augmentation choice, such as masking-based objectives that derive their self-supervised signal from masking rather than augmentation, may

therefore be operating suboptimally relative to their potential. Readers should interpret SimCLR’s performance as an upper bound on its true cross-paradigm-comparable rank, and the performance of masking-based methods as a lower bound on their potential under augmentation-agnostic conditions.

Reason 2: SimCLR is intrinsically robust to input noise. SimCLR’s contrastive objective confers strong robustness to input noise Maes et al. [2025]: the InfoNCE loss trains the encoder to agree on representations across two independently corrupted views, implicitly forcing invariance to the additive microstructure noise and rendering artefacts that characterise all three financial encodings. On the heatmap in particular, where cross-sectional factor comovement is a global, low-frequency pattern superimposed on substantial point-to-point return noise, this noise-robustness provides a direct inductive bias toward the discriminative signal.

Disentangling these two contributions—for example by selecting window sizes and augmentations via a held-out reference method—is a natural direction for future work.

D Datasets

This appendix documents the two datasets used throughout the paper, including data provenance, preprocessing, split boundaries, sector and factor label distributions, and encoded image counts at each candidate window size. All statistics reported here refer to the data as consumed by the SSL pre-training and probe-evaluation pipelines; raw data are not modified beyond the standard preprocessing described in §3.2.

D.1 S&P 500 Individual Equities

Table 3 summarises the row counts and ticker coverage across the three temporal splits. Thirteen tickers were excluded due to missing GICS sector labels in the `yfinance.info` field; the remaining 483 tickers span all 11 GICS sectors. Table 4 reports the per-sector ticker counts, showing the expected overrepresentation of the Technology sector (75 tickers, 15.5% of the universe) that motivates the use of balanced accuracy as the probe metric throughout. Table 5 reports encoded image counts at each of the four candidate window sizes after applying the stride-5 rolling window and the 45,000-image subsampling cap on the training split. Image counts decrease with window size because longer windows require more leading history and therefore cannot be formed at the start of each ticker’s series.

Table 3: S&P 500 dataset statistics. Data sourced via `yfinance`; 13 tickers skipped due to missing GICS sector labels. Split boundaries are fixed calendar dates chosen to place the COVID-19 shock and the 2022 rate-hiking cycle entirely within the validation and test sets respectively, preventing these regime-shift events from contaminating the pre-training distribution.

Split	Date Range	Tickers	Rows
Train	2013-01-01 — 2019-12-31	483	851,046
Val	2020-01-01 — 2021-12-31	483	243,915
Test	2022-01-01 — 2023-12-31	483	241,983
Total	2013-01-01 — 2023-12-31	483	1,336,944

Table 4: S&P 500 sector distribution (unique tickers). The Technology sector accounts for over 15% of the universe, making balanced accuracy the appropriate probe metric; raw accuracy would be dominated by this sector and would not reflect representation quality across the full sector space.

Sector	Tickers
Technology	75
Industrials	68
Financial Services	64
Healthcare	60
Consumer Cyclical	55
Consumer Defensive	37
Real Estate	32
Utilities	31
Basic Materials	22
Energy	20
Communication Services	19
Total	483

Table 5: S&P 500 encoded image counts per window size. Windows are collected with a stride of 5 trading days; subsampling caps of 45,000 (train), 12,897 (val), and 12,795 (test) images are applied uniformly across all window sizes and both encodings (GAF+MTF and candlestick). The modest decline in raw image counts at longer window sizes reflects the loss of leading observations at the start of each ticker’s series.

Window	Train		Val		Test	
	Raw	Sampled	Raw	Sampled	Raw	Sampled
20	160,159	45,000	46,651	12,897	46,451	12,795
63	155,902	45,000	42,358	12,897	42,122	12,795
126	150,226	45,000	36,157	12,897	36,362	12,795
252	138,487	45,000	24,232	12,897	23,899	12,795
Tickers (train / val / test)				473–467 / 477 / 479–481		

D.2 Fama–French 30-Industry Portfolios

Table 6 reports the temporal split boundaries and trading day counts for the Fama–French dataset. The training split spans nearly six decades (July 1963 through December 2019), providing a large and diverse pre-training corpus that encompasses multiple full business cycles, regime changes, and volatility episodes. The validation and test splits each cover two calendar years; both fall entirely after the training cutoff, enforcing strict temporal separation. Because the heatmap encoding uses a rolling window of 20 trading days applied with a stride of 2, the full training split yields approximately 7,100 heatmap images, with around 250 images in each of the validation and test splits.

Table 6: Fama–French 30-industry portfolio data splits. Daily value-weighted returns are sourced from the Kenneth French Data Library. The training split covers nearly six decades of market history, encompassing multiple business cycles and volatility regimes. The validation split spans the COVID-19 shock period and its aftermath; the test split covers the 2022 rate-hiking cycle, providing two qualitatively distinct out-of-sample evaluation regimes.

Split	Date Range	Trading Days
Train	1963-07-01 — 2019-12-31	14,223
Val	2020-01-02 — 2021-12-31	505
Test	2022-01-03 — 2023-12-29	501
Total	1963-07-01 — 2023-12-29	15,229

E Ablations

This appendix provides a detailed visual record of the two ablation studies described in §4.1 and §4.2. We first show representative encoded images at each candidate window size to build intuition for how temporal resolution affects the visual structure of each encoding, then display representative augmented images to illustrate how each candidate transform interacts with each encoding geometry, and finally present the full training curves that were used to make the configuration selections reported in the main text.

E.1 Window Size Ablation

The rolling window size is the primary dataset-level hyperparameter: it determines how many trading days of price or return history are compressed into a single 224×224 image. Shorter windows capture local, high-frequency dynamics but contain fewer observations; longer windows expose richer medium- and long-run structure at the cost of compressing more bars into the same resolution. Because the optimal balance depends on the encoding geometry, we treat window size as an encoding-specific hyperparameter selected via the validation linear probe.

Figures 1, 2, and 3 show representative images at all four candidate window sizes for each encoding, providing visual intuition for how increasing temporal horizon changes the spatial structure that an SSL encoder must process.

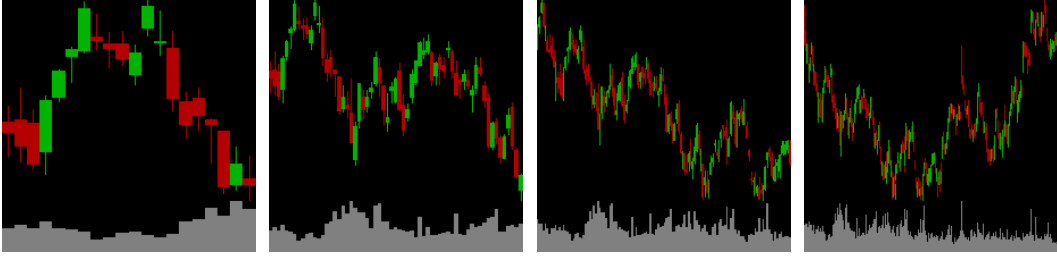


Figure 1: Candlestick encodings at window sizes of 20, 63, 126, and 252 trading days (left to right), rendered as 224×224 RGB images. Individual bars become progressively narrower as window size increases, until at 252 days each bar occupies fewer than one pixel of width and fine-grained local patterns—body height, wick asymmetry, color sequences—are effectively destroyed by aliasing. This visual degradation is consistent with the probe results, which favour the 20-day window where individual candlestick features remain legible.

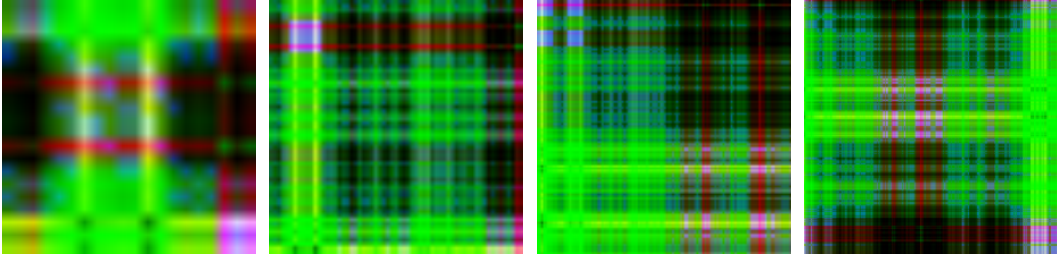


Figure 2: GAF+MTF encodings at window sizes of 20, 63, 126, and 252 trading days (left to right). Channel 1 (red): GASf; channel 2 (green): GADF; channel 3 (blue): MTF with $Q=8$ equal-frequency bins. Unlike candlestick images, GAF+MTF encodings do not suffer from resolution aliasing at longer windows because each pairwise interaction (ϕ_i, ϕ_j) maps to a distinct image pixel regardless of series length. Instead, increasing window size smooths the pairwise correlation structure, producing images with more regular, globally coherent texture—consistent with the probe results, which favour 126 days where this structure is richest without being oversmoothed.

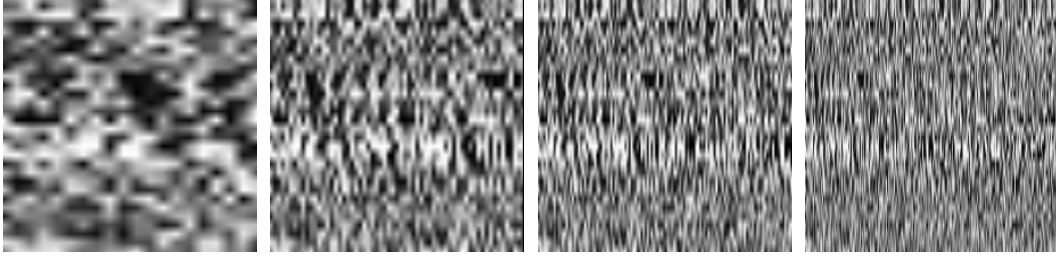


Figure 3: Returns heatmap encodings of the Fama–French 30-industry portfolio panel at window sizes of 20, 63, 126, and 252 trading days (left to right). Pixel intensity encodes the rank of each industry’s cross-sectionally z -scored log return; brighter pixels indicate relatively higher returns. Longer windows expose more of the persistent cross-sectional comovement structure, making horizontal banding—the visual signature of common factor loadings—more prominent. Performance on the 20-day and 126-day windows was close on both probe metrics; 20 trading days was selected after also accounting for RankMe trajectories showing more stable effective rank at the shorter horizon.

The full training dynamics underlying these selections are presented in Figures 4–6. Each figure shows four panels—online linear probe balanced accuracy, kNN probe balanced accuracy, SSL contrastive loss, and RankMe effective-rank score—all computed on the held-out validation set at every checkpoint. The selected window size for each encoding is the configuration that achieves the highest validation linear probe accuracy while maintaining a healthy RankMe trajectory throughout training.

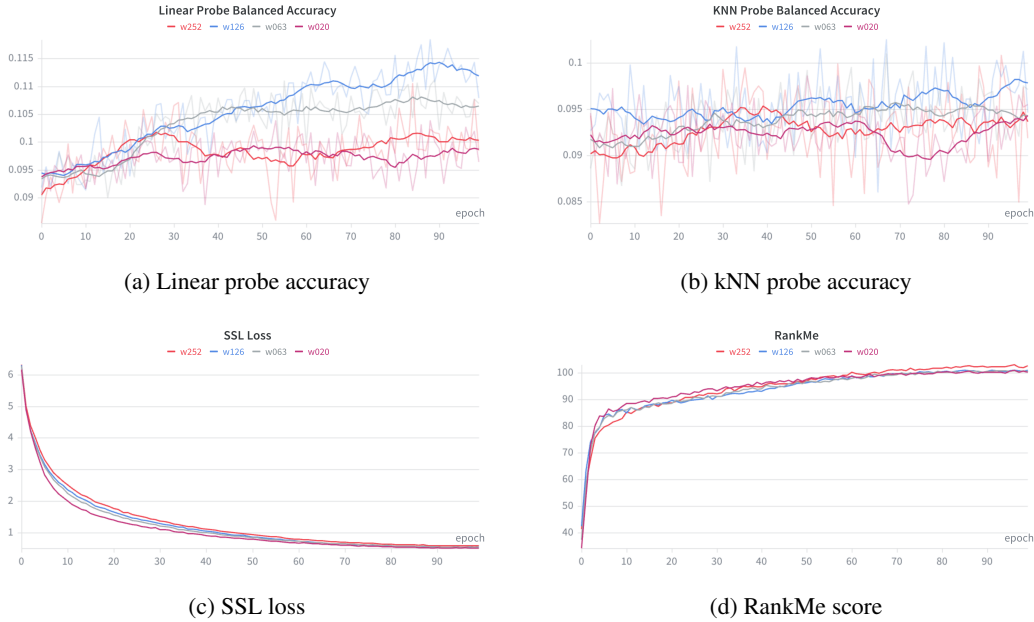


Figure 4: Window size ablation training curves for SimCLR on the GAF+MTF encoding (hyperparameters in Table 8). Linear probe accuracy increases monotonically from 20 to 126 days and plateaus at 252 days, while RankMe remains healthy across all window sizes. **Selected window: 126 trading days.**

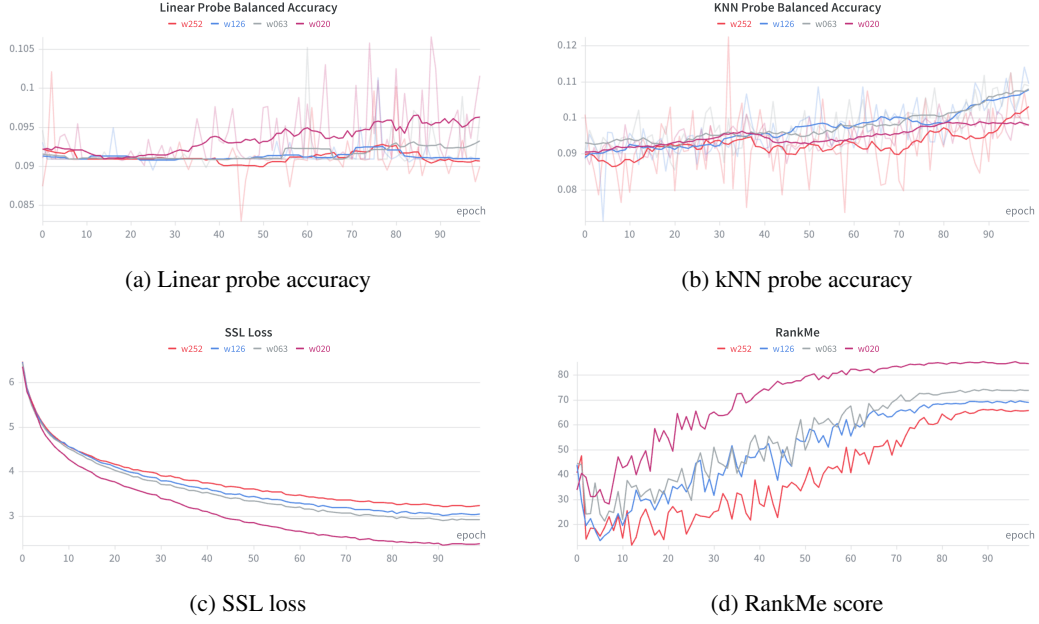


Figure 5: Window size ablation training curves for SimCLR on the candlestick encoding (hyperparameters in Table 8). Both linear and kNN probe accuracy decline monotonically beyond 20 days, consistent with the visual aliasing of individual candlestick bars at longer windows observed in Figure 1. **Selected window: 20 trading days.**

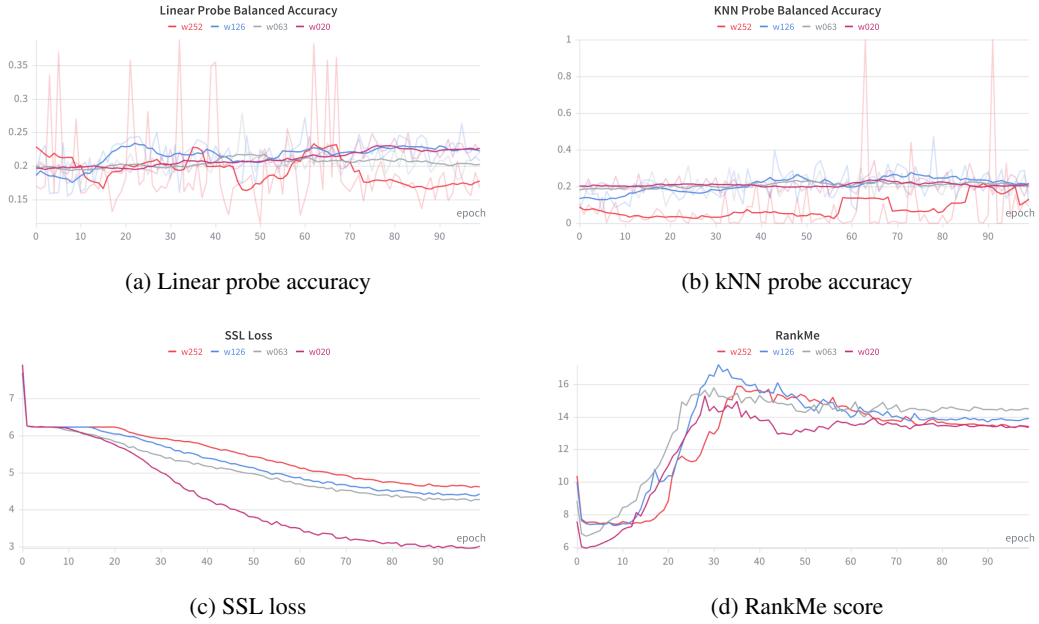


Figure 6: Window size ablation training curves for SimCLR on the returns heatmap encoding (hyperparameters in Table 8). Performance at 20 and 126 days is close on both probe metrics; the 20-day window is selected on the basis of more stable RankMe trajectories, which show a partial collapse-and-recovery pattern at longer horizons. **Selected window: 20 trading days.**

E.2 Augmentation Ablation

In natural image SSL, a canonical pipeline of horizontal flipping, random cropping, color jitter, and Gaussian blur has emerged as a near-universal standard. Financial imagery violates several of these priors: a horizontal flip reverses the temporal ordering encoded in GAF and MTF images, and color jitter corrupts the semantically critical red/green convention in candlestick renders. We evaluated eight candidate augmentations via greedy forward selection on SimCLR, using the window sizes identified in §E.1.

Figures 7, 8, and 9 show representative images of each augmentation applied to all three encodings. Hyperparameters are intentionally exaggerated relative to training values to make each transform visually distinguishable; training-time hyperparameters are given in Table 7. Several augmentations that are benign or helpful on natural images are visually destructive here: Horizontal Flip reverses temporal ordering in GAF+MTF images, and Color Jitter corrupts both the magnitude structure of GAF+MTF and the directional red/green convention of candlestick renders.

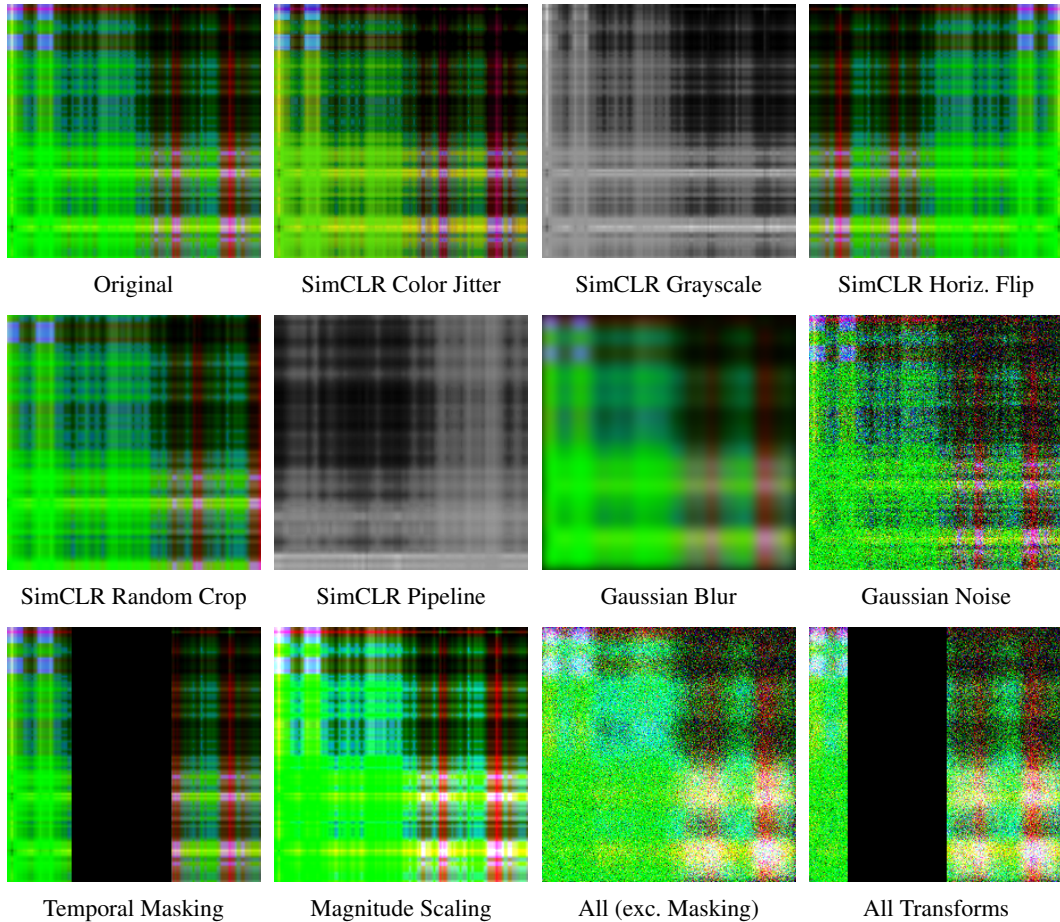


Figure 7: Data augmentation pipeline applied to GAF+MTF encodings (window size: 126 trading days). Each panel shows a single augmentation in isolation, except *SimCLR Pipeline* (full SimCLR composition), *All (exc. Masking)*, and *All Transforms*. Horizontal Flip inverts the temporal axis of both GASF and GADF, producing a structurally invalid image; Color Jitter corrupts the magnitude-encoded channel structure. Random Resized Crop preserves the global correlation texture of GAF+MTF while introducing spatial invariance, making it the single most effective augmentation for this encoding. **Selected augmentation: Random Resized Crop.**

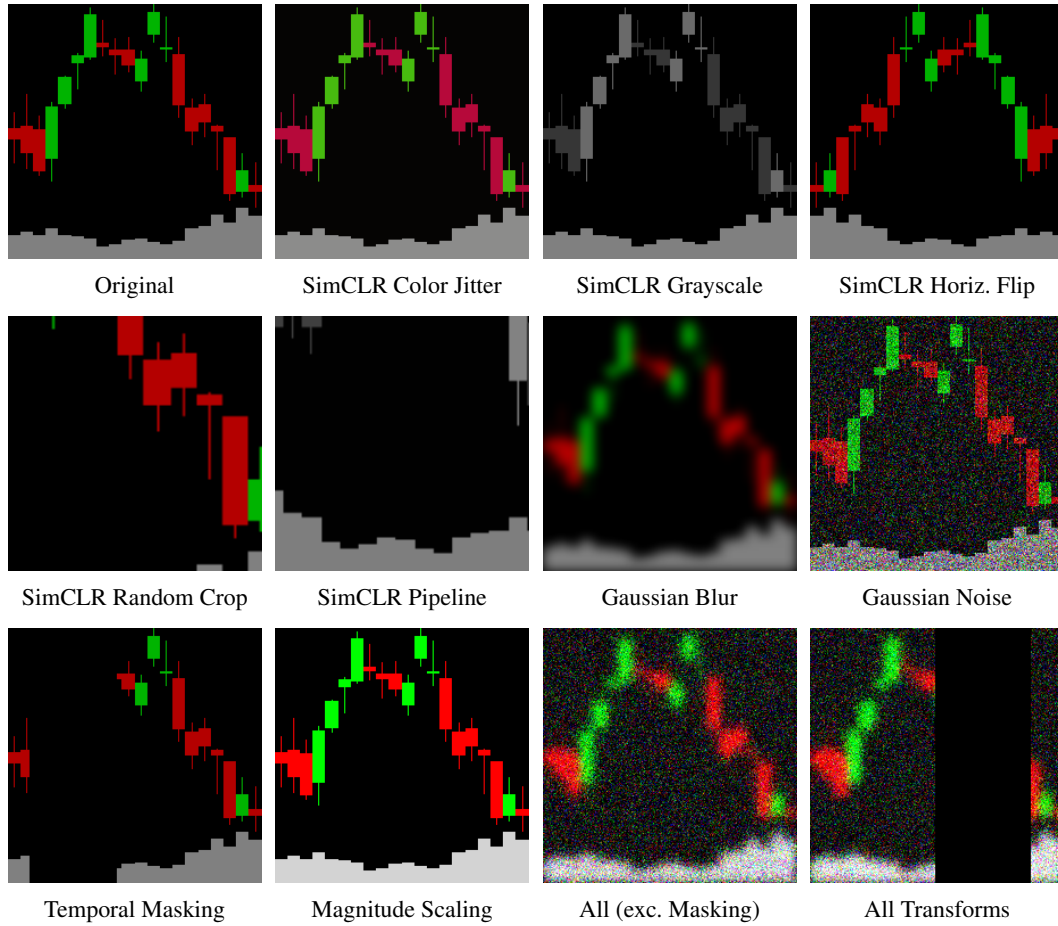


Figure 8: Data augmentation pipeline applied to candlestick encodings (window size: 20 trading days). Horizontal Flip reverses the time axis of the chart, rendering it economically meaningless. Color Jitter perturbs the red/green body coloring, which encodes the sign of each bar’s return; mild jitter is nonetheless selected because it introduces useful color invariance without fully destroying the directional signal. Gaussian Blur attenuates microstructure noise in individual bars; Gaussian Noise adds back controlled perturbations that improve robustness to rendering artefacts. **Selected augmentations: Color Jitter, Gaussian Blur, Gaussian Noise, Random Resized Crop.**

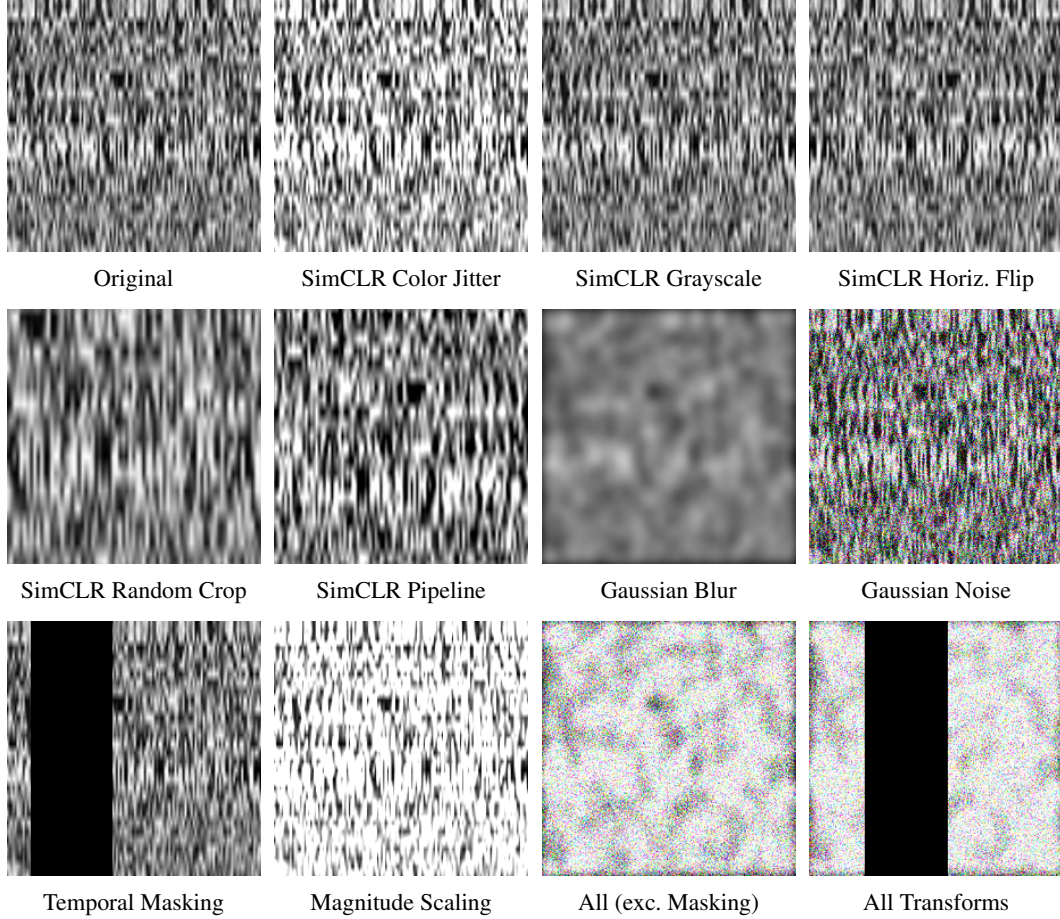


Figure 9: Data augmentation pipeline applied to heatmap encodings (window size: 20 trading days). Color Jitter and Random Grayscale are particularly disruptive: pixel intensity encodes rank-normalised cross-sectional returns, so color perturbations directly corrupt the factor signal. Temporal Masking (visible as horizontal bands of zeroed-out columns) is uniquely beneficial for the heatmap because it forces the encoder to infer factor loadings from partial time observations, improving robustness to missing data and short windows. Temporal Masking is harmful for the single-asset encodings, where zeroing out columns disrupts the temporal ordering that GAF+MTF and candlestick are designed to preserve. **Selected augmentations: Random Resized Crop, Temporal Masking, Color Jitter, Gaussian Noise.**

Table 7 reports the parameter values used to generate the augmentation visualisation figures above. All transforms are applied with $p=1.0$ and with parameter values that are intentionally exaggerated relative to training to make each transform visually distinguishable. Training-time parameter values are given in Table 9 in Appendix F.

Table 7: Hyperparameters used to generate the augmentation visualisation figures (Figures 7–9). All transforms are applied with $p = 1.0$. Parameter values are *intentionally exaggerated* relative to their training-time counterparts to make each transform visually distinguishable; see Table 9 for actual training values.

Augmentation	Parameter	Demo value	Type
SimCLR Random Resized Crop	scale range	(0.2, 1.0)	domain
	aspect ratio range	(0.75, 1.33)	domain
SimCLR Horizontal Flip	—	—	domain
SimCLR Color Jitter	brightness	0.4	domain
	contrast	0.4	domain
	saturation	0.2	domain
	hue	0.1	domain
SimCLR Random Grayscale	—	—	domain
Gaussian Blur	σ range	(4.0, 6.0)	financial
Magnitude Scaling	scale range	(1.6, 2.0)	financial
Gaussian Noise	σ	0.25	financial
Temporal Masking	mask ratio range	(0.30, 0.40)	financial

The full training dynamics from the augmentation greedy forward selection are shown in Figures 10–12. As with the window-size ablation, each figure shows four panels computed on the held-out validation set at every checkpoint.

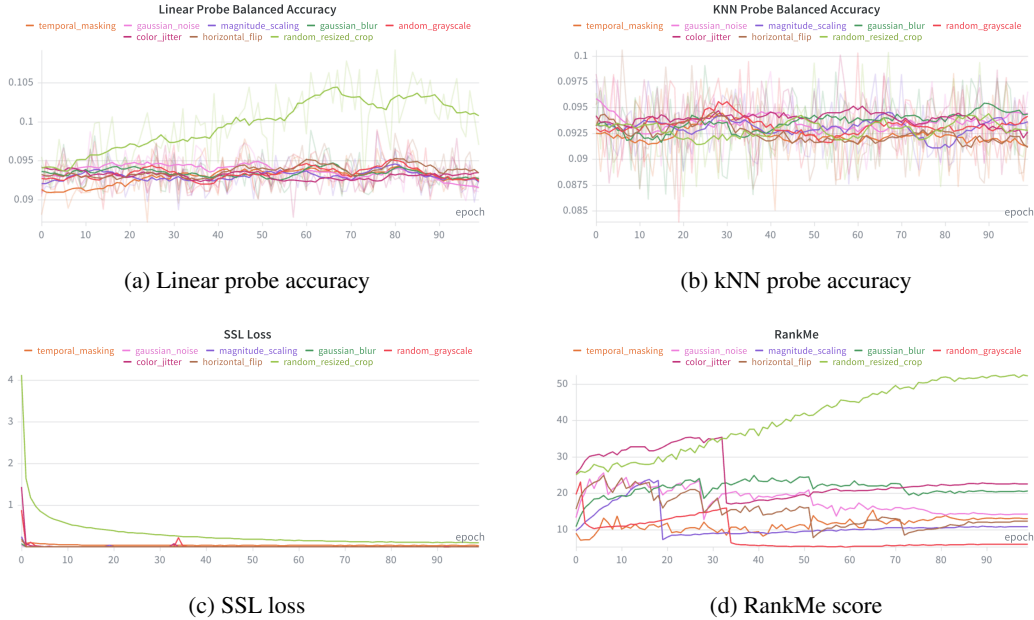


Figure 10: Augmentation ablation training curves for SimCLR on the GAF+MTF encoding, window size 126 days (hyperparameters in Table 9). Horizontal Flip causes a pronounced drop in both probe metrics and RankMe, confirming that inverting the temporal axis is destructive for this encoding. Random Resized Crop alone matches or exceeds all multi-augmentation compositions. **Selected augmentation: Random Resized Crop.**

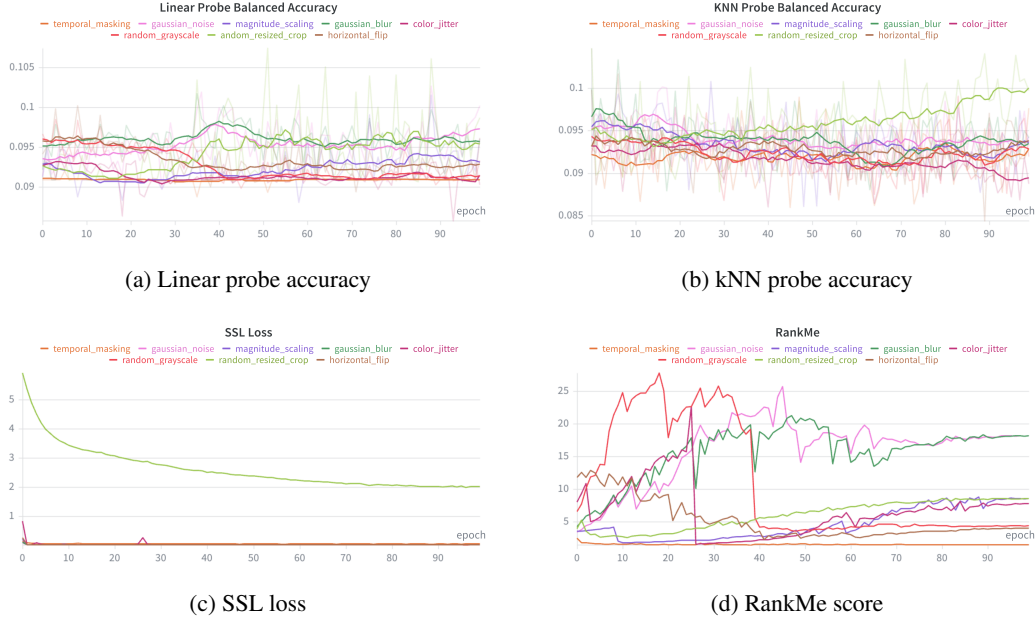


Figure 11: Augmentation ablation training curves for SimCLR on the candlestick encoding, window size 20 days (hyperparameters in Table 9). A richer augmentation composition is beneficial here: Color Jitter, Gaussian Blur, Gaussian Noise, and Random Resized Crop are all selected by greedy forward selection, reflecting the greater diversity of visual and magnitude signals present in candlestick renders. **Selected augmentations: Color Jitter, Gaussian Blur, Gaussian Noise, Random Resized Crop.**

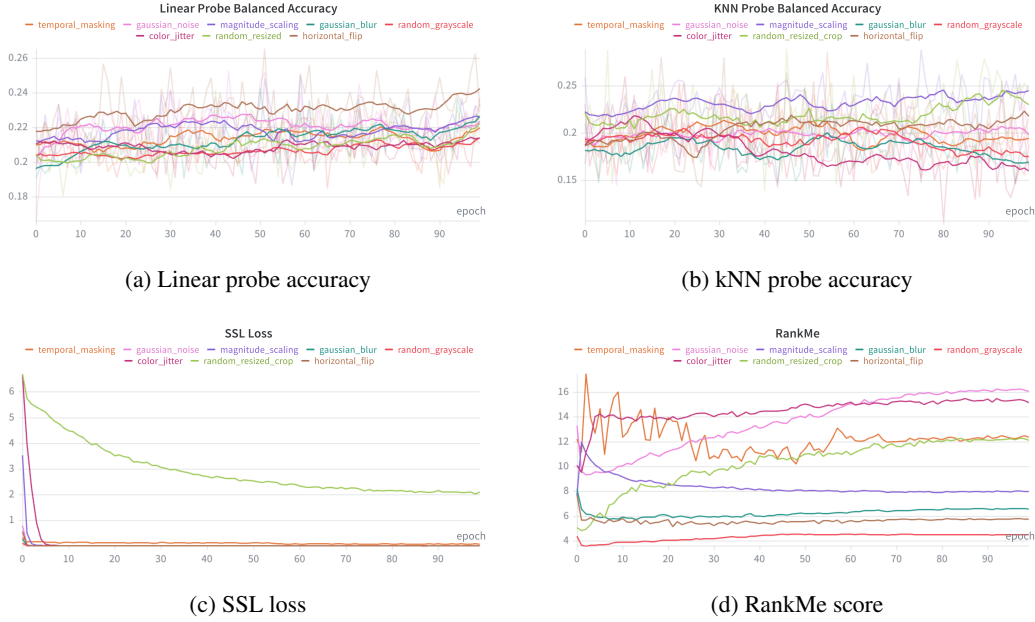


Figure 12: Augmentation ablation training curves for SimCLR on the heatmap encoding, window size 20 days (hyperparameters in Table 9). The broadest augmentation set is selected here; notably, Temporal Masking provides a unique improvement not observed for any other encoding. **Selected augmentations: Random Resized Crop, Temporal Masking, Color Jitter, Gaussian Noise.**

E.3 SSL Paradigm Comparison Training Curves

The following figures show the full training dynamics for all six SSL methods and the supervised baseline on each encoding, using the window sizes and augmentation configurations selected in the ablations above. These curves correspond to the final balanced accuracy values summarised in Table 2 in the main text. RankMe trajectories are particularly informative here: a flat or declining RankMe curve indicates representational collapse and should be read alongside the probe accuracy curves to distinguish genuine representational improvement from artefactual gains arising from low-dimensional collapse.

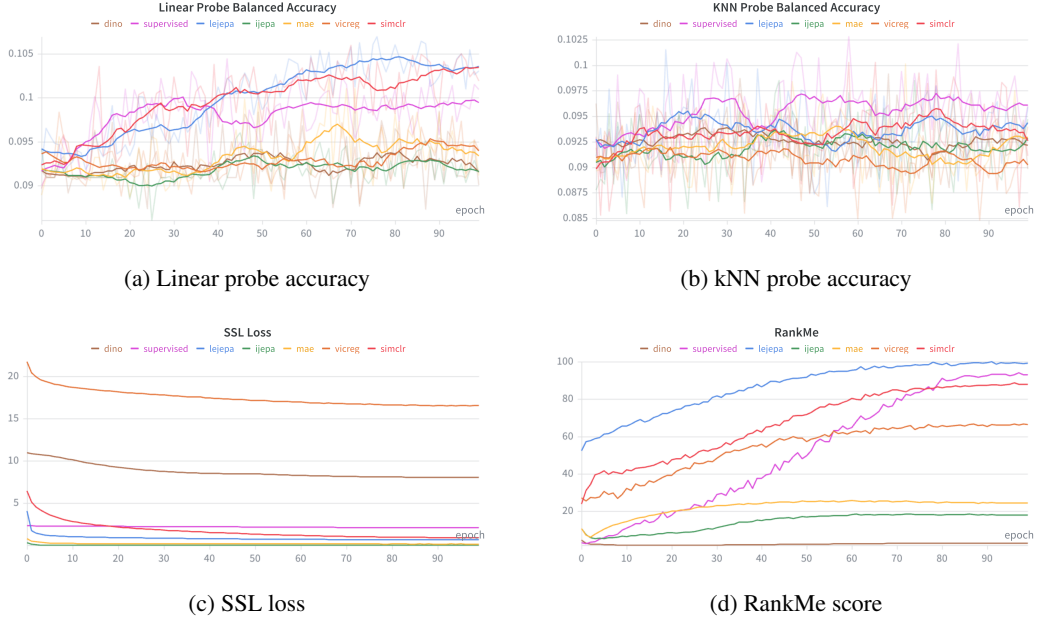


Figure 13: SSL paradigm comparison training curves on the GAF+MTF encoding (window: 126 days; augmentation: Random Resized Crop). LeJEPa leads the linear probe (10.4%) with SimCLR close behind (10.3%). I-JEPA and MAE both exhibit low, flat RankMe scores throughout training, consistent with representational collapse; their probe accuracy nonetheless remains above the random baseline, suggesting the collapsed representations still encode some coarse sector structure. LeJEPa’s RankMe remains high throughout, consistent with SIGReg maintaining the isotropic Gaussian geometry guaranteed by theory.

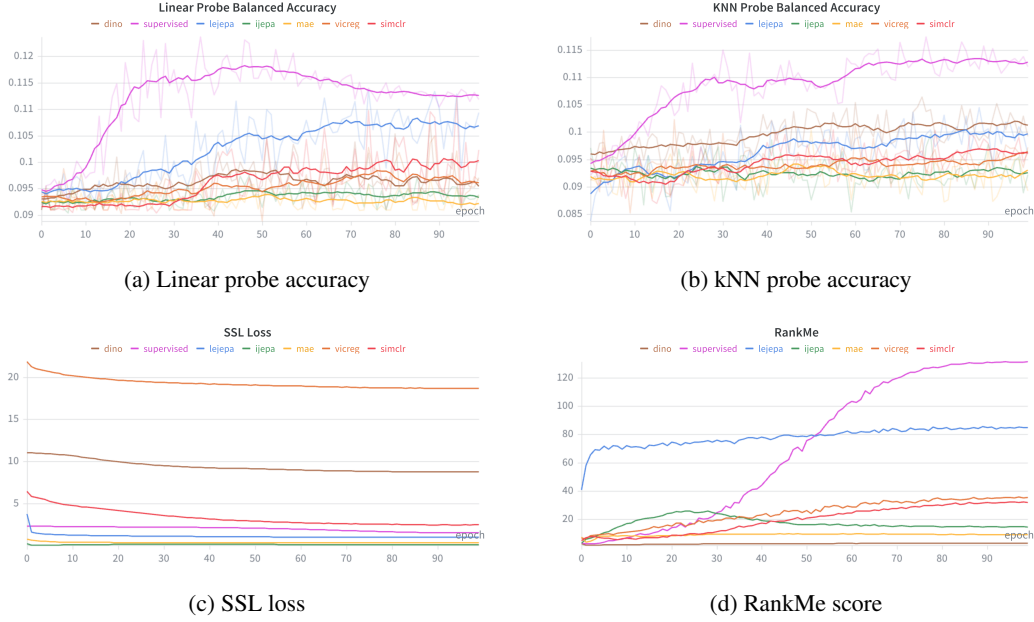


Figure 14: SSL paradigm comparison training curves on the candlestick encoding (window: 20 days; augmentation: Color Jitter + Gaussian Blur + Gaussian Noise + Random Resized Crop). The supervised baseline is the clear leader on both probes (11.3%), with a larger gap to SSL than observed on GAF+MTF or heatmap. This pattern is consistent with candlestick patterns being fine-grained, local signals whose precise spatial structure benefits from direct label supervision. Among SSL methods, LeJEPA leads on the linear probe (10.7%) and DINO on kNN (10.1%).

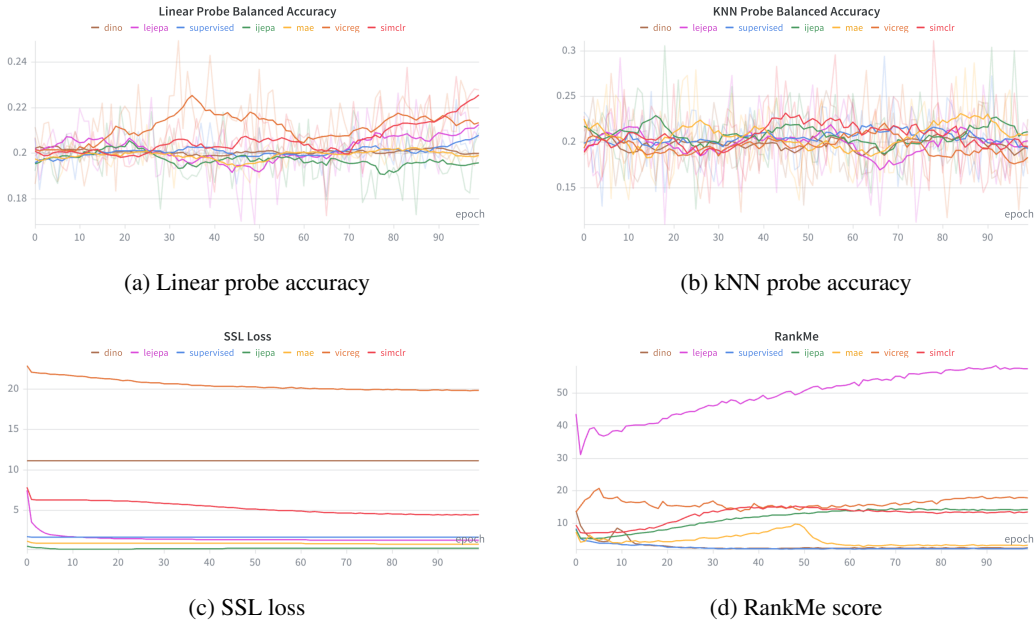


Figure 15: SSL paradigm comparison training curves on the heatmap encoding (window: 20 days; augmentation: Random Resized Crop + Temporal Masking + Color Jitter + Gaussian Noise). SimCLR leads the linear probe (22.5%) with VICReg second (21.3%); under kNN, I-JEPA edges ahead (21.1%) followed by MAE (20.8%). All SSL methods approach or exceed the supervised baseline (20.8% linear, 19.2% kNN), consistent with cross-sectional factor structure being globally diffuse and therefore well-captured by unsupervised objectives.

F Hyperparameters

This appendix reports the complete hyperparameter configurations for all experimental runs. Tables 8 and 9 cover the window-size and augmentation ablation studies respectively. Tables 10–16 cover the six SSL methods and the supervised baseline used in the paradigm comparison.

Every method shares a common substrate: a ViT-Tiny backbone trained from random initialisation for 100 epochs with AdamW, cosine annealing with 10-epoch linear warmup, and an identical evaluation protocol (linear probe, kNN, and RankMe at every checkpoint). Method-specific entries are confined to the loss function, projector or decoder architecture, and objective-specific coefficients. MAE and I-JEPA use AdamW betas (0.9, 0.95) following their respective original papers, as the reduced second-moment decay stabilises training under masked prediction objectives; all other methods use the standard (0.9, 0.999). This shared substrate is the core methodological control of the paper: because every run uses the same backbone, optimiser, schedule, data pipeline, and evaluation protocol, differences in downstream probe performance are attributable to the pre-training objective alone.

Table 8: Hyperparameters for the window size ablation runs (Study 1). SimCLR is used as the reference method; the standard SimCLR augmentation pipeline is applied during this study and is superseded by the encoding-specific pipelines identified in Study 2.

Setting	Value
<i>Training</i>	
Epochs	100
Batch size	256
Precision	16-bit mixed
Seed	42
<i>Optimiser</i>	
Optimiser	AdamW
Learning rate	1.5×10^{-4}
Weight decay	0.05
LR schedule	Cosine annealing with linear warmup
Warmup epochs	Library default
<i>SSL Method (SimCLR)</i>	
Loss	InfoNCE (NT-Xent)
Temperature τ	0.07
Projector	Linear(192, 512) $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ Linear(512, 512) $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ Linear(512, 128)
<i>Backbone</i>	
Architecture	ViT-Tiny (vit_tiny_patch16_224)
Pretrained	No (random initialization)
Embedding dim	192
<i>Augmentations (training views)</i>	
Training views	2 (identical pipeline applied independently)
Random resized crop	224×224
color jitter	Brightness 0.4, contrast 0.4, saturation 0.2, hue 0.1 ($p=0.8$)
Random grayscale	$p=0.2$
Normalisation	ImageNet mean and standard deviation
Horizontal flip	None
Validation transform	RGB conversion $\rightarrow$ ImageNet normalisation
<i>Evaluation</i>	
Linear probe	Linear(192, C), AdamW, cross-entropy, balanced accuracy
kNN probe	$k=20$, cosine distance, queue length 10,000, balanced accuracy
RankMe	Queue length 2,048
<i>Candidate window sizes</i>	
S&P 500	20, 63, 126, 252 trading days
Fama–French	20, 63, 126, 252 trading days
<i>Probe targets</i>	
S&P 500	GICS sector label ($C=11$)
Fama–French	Factor quintile per factor, averaged ($C=5$)

Table 9: Hyperparameters for the augmentation ablation runs (Study 2). Each candidate augmentation is evaluated in isolation on SimCLR; the greedy forward selection then adds augmentations one at a time until validation linear probe accuracy stops improving. Window sizes are fixed to the values selected in Study 1.

Setting	Value
<i>Training</i>	
Epochs	100
Batch size	256
Precision	16-bit mixed
Seed	42
<i>Optimiser</i>	
Optimiser	AdamW
Learning rate	1.5×10^{-4}
Weight decay	0.05
LR schedule	Cosine annealing with linear warmup
Warmup epochs	Library default
<i>SSL Method (SimCLR)</i>	
Loss	InfoNCE (NT-Xent)
Temperature τ	0.07
Projector	Linear(192, 512) $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ Linear(512, 512) $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ Linear(512, 128)
<i>Backbone</i>	
Architecture	ViT-Tiny (vit_tiny_patch16_224)
Pretrained	No (random initialization)
Embedding dim	192
<i>Augmentations (training views)</i>	
Training views	2 (single ablated transform applied independently)
Neutral resize	224×224 (used when crop is not the ablated transform)
Normalisation	ImageNet mean and standard deviation
Validation transform	RGB conversion $\rightarrow$ ImageNet normalisation
<i>Candidate transforms</i>	
Random resized crop	224×224
Horizontal flip	$p=0.5$
color jitter	Brightness 0.4, contrast 0.4, saturation 0.2, hue 0.1 ($p=0.8$)
Random grayscale	$p=0.2$
Gaussian blur	$\sigma \in [0.1, 2.0]$, $p=1.0$
Magnitude scaling	Scale $\in [0.8, 1.2]$, $p=0.5$
Gaussian noise	$\sigma=0.05$, $p=0.5$
Temporal masking	Mask ratio $\in [0.10, 0.20]$, $p=0.5$
<i>Evaluation</i>	
Linear probe	Linear(192, C), AdamW, cross-entropy, balanced accuracy
kNN probe	$k=20$, cosine distance, queue length 10,000, balanced accuracy
RankMe	Queue length 2,048
<i>Fixed window size (encoding-dependent)</i>	
S&P 500 $\times$ GAF+MTF	126 trading days
S&P 500 $\times$ candlestick	20 trading days
Fama–French $\times$ heatmap	20 trading days
<i>Probe targets</i>	
S&P 500	GICS sector label ($C=11$)
Fama–French	Factor quintile per factor, averaged ($C=5$)

Table 10: Hyperparameters for SimCLR pre-training (Study 3).

Setting	Value
<i>Training</i>	
Epochs	100
Batch size	256
Precision	16-bit mixed
Seed	42
Training views	2 (contrastive pair)
<i>Optimiser</i>	
Optimiser	AdamW
Learning rate	1.5×10^{-4}
Weight decay	0.05
Betas	(0.9, 0.999)
LR schedule	Cosine annealing with linear warmup
<i>SSL Method</i>	
Loss	InfoNCE (NT-Xent), temperature $\tau = 0.07$
Projector	Linear(192, 512) $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ Linear(512, 512) $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ Linear(512, 128)
<i>Backbone</i>	
Architecture	ViT-Tiny (vit_tiny_patch16_224)
Pretrained	No (random initialisation)
Embedding dim	192
<i>Evaluation</i>	
Linear probe	Linear(192, C), cross-entropy, balanced accuracy
kNN probe	$k=20$, cosine distance, queue length 10,000
RankMe	Queue length 2,048
<i>Probe Targets</i>	
S&P 500	GICS sector label ($C=11$)
Fama–French	Factor quintile per factor, averaged ($C=5$)
<i>Fixed Window Size</i>	
S&P 500 $\times$ GAF+MTF	126 trading days
S&P 500 $\times$ candlestick	20 trading days
Fama–French $\times$ heatmap	20 trading days

Table 11: Hyperparameters for VICReg pre-training (Study 3).

Setting	Value
<i>Training</i>	
Epochs	100
Batch size	256
Precision	16-bit mixed
Seed	42
Training views	2 (contrastive pair)
<i>Optimiser</i>	
Optimiser	AdamW
Learning rate	1.5×10^{-4}
Weight decay	0.05
Betas	(0.9, 0.999)
LR schedule	Cosine annealing with linear warmup
<i>SSL Method</i>	
Loss	VICReg: invariance coeff = 25.0, variance coeff = 25.0, covariance coeff = 1.0
Projector	Linear(192, 512) → BN → ReLU → Linear(512, 512) → BN → ReLU → Linear(512, 512)
<i>Backbone</i>	
Architecture	ViT-Tiny (vit_tiny_patch16_224)
Pretrained	No (random initialisation)
Embedding dim	192
<i>Evaluation</i>	
Linear probe	Linear(192, C), cross-entropy, balanced accuracy
kNN probe	$k=20$, cosine distance, queue length 10,000
RankMe	Queue length 2,048
<i>Probe Targets</i>	
S&P 500	GICS sector label ($C=11$)
Fama–French	Factor quintile per factor, averaged ($C=5$)
<i>Fixed Window Size</i>	
S&P 500 × GAF+MTF	126 trading days
S&P 500 × candlestick	20 trading days
Fama–French × heatmap	20 trading days

Table 12: Hyperparameters for DINO pre-training (Study 3).

Setting	Value
<i>Training</i>	
Epochs	100
Batch size	256
Precision	16-bit mixed
Seed	42
<i>Optimiser</i>	
Optimiser	AdamW
Learning rate	1.5×10^{-4}
Weight decay	0.05
Betas	(0.9, 0.999)
LR schedule	Cosine annealing with linear warmup
<i>SSL Method</i>	
Loss	Cross-entropy (student–teacher)
Teacher momentum	0.996
<i>Backbone</i>	
Architecture	ViT-Tiny (vit_tiny_patch16_224)
Pretrained	No (random initialisation)
Embedding dim	192
<i>Evaluation</i>	
Linear probe	Linear(192, C), cross-entropy, balanced accuracy
kNN probe	$k=20$, cosine distance, queue length 10,000
RankMe	Queue length 2,048
<i>Probe Targets</i>	
S&P 500	GICS sector label ($C=11$)
Fama–French	Factor quintile per factor, averaged ($C=5$)
<i>Fixed Window Size</i>	
S&P 500 $\times$ GAF+MTF	126 trading days
S&P 500 $\times$ candlestick	20 trading days
Fama–French $\times$ heatmap	20 trading days

Table 13: Hyperparameters for MAE pre-training (Study 3). AdamW betas (0.9, 0.95) follow the original MAE paper; the reduced second-moment decay stabilises the masked reconstruction objective.

Setting	Value
<i>Training</i>	
Epochs	100
Batch size	256
Precision	16-bit mixed
Seed	42
Training views	1 (single view; patch masking applied internally)
<i>Optimiser</i>	
Optimiser	AdamW
Learning rate	1.5×10^{-4}
Weight decay	0.05
Betas	(0.9, 0.95)
LR schedule	Cosine annealing with linear warmup
<i>SSL Method</i>	
Loss	MSE (normalised pixel loss, norm_pix_loss=True)
Mask ratio	0.75 (random patch masking, block size = 1)
Decoder	embed_dim= 256, depth= 4, num_heads= 4
Projector	None (MAE native architecture)
<i>Backbone</i>	
Architecture	ViT-Tiny (vit_tiny_patch16_224)
Pretrained	No (random initialisation)
Embedding dim	192
<i>Evaluation</i>	
Linear probe	Linear(192, C), cross-entropy, balanced accuracy
kNN probe	$k=20$, cosine distance, queue length 10,000
RankMe	Queue length 2,048
<i>Probe Targets</i>	
S&P 500	GICS sector label ($C=11$)
Fama–French	Factor quintile per factor, averaged ($C=5$)
<i>Fixed Window Size</i>	
S&P 500 $\times$ GAF+MTF	126 trading days
S&P 500 $\times$ candlestick	20 trading days
Fama–French $\times$ heatmap	20 trading days

Table 14: Hyperparameters for I-JEPA pre-training (Study 3). AdamW betas (0.9, 0.95) follow the original I-JEPA paper. The EMA decay schedule increases from 0.996 to 1.0 over training, progressively slowing the teacher update and stabilising the prediction target.

Setting	Value
<i>Training</i>	
Epochs	100
Batch size	256
Precision	16-bit mixed
Seed	42
Training views	1 (single view; block masking applied internally)
<i>Optimiser</i>	
Optimiser	AdamW
Learning rate	1.5×10^{-4}
Weight decay	0.05
Betas	(0.9, 0.95)
LR schedule	Cosine annealing with linear warmup
<i>SSL Method</i>	
Loss	ℓ_2 in latent feature space (JEPA prediction)
Predictor	embed_dim= 96, depth= 6
Num targets	4
Target scale	(0.15, 0.2); target aspect ratio (0.75, 1.5)
Context scale	(0.85, 1.0)
EMA decay	Start = 0.996, end = 1.0
Projector	None
<i>Backbone</i>	
Architecture	ViT-Tiny (vit_tiny_patch16_224)
Pretrained	No (random initialisation)
Embedding dim	192
<i>Evaluation</i>	
Linear probe	Linear(192, C), cross-entropy, balanced accuracy
kNN probe	$k=20$, cosine distance, queue length 10,000
RankMe	Queue length 2,048
<i>Probe Targets</i>	
S&P 500	GICS sector label ($C=11$)
Fama–French	Factor quintile per factor, averaged ($C=5$)
<i>Fixed Window Size</i>	
S&P 500 $\times$ GAF+MTF	126 trading days
S&P 500 $\times$ candlestick	20 trading days
Fama–French $\times$ heatmap	20 trading days

Table 15: Hyperparameters for LeJEPA pre-training (Study 3). LeJEPA uses a multi-crop view schedule (2 global + 6 local views) inherited from the DINO multi-crop protocol, combined with the SIGReg regulariser that enforces isotropic Gaussian embedding geometry without requiring a stop-gradient or asymmetric teacher–student architecture.

Setting	Value
<i>Training</i>	
Epochs	100
Batch size	256
Precision	16-bit mixed
Seed	42
Training views	2 global (224×224, crop scale 0.3–1.0) + 6 local (96×96, crop scale 0.05–0.3)
<i>Optimiser</i>	
Optimiser	AdamW
Learning rate	1.5×10^{-4}
Weight decay	0.05
Betas	(0.9, 0.999)
LR schedule	Cosine annealing with linear warmup
<i>SSL Method</i>	
Loss	SIGReg (Epps–Pulley goodness-of-fit) + multi-view invariance
SIGReg λ	0.02
SIGReg <code>n_slices</code>	1024
SIGReg <code>n_points</code>	17
Projector	None (encoder/projector/SIGReg head bundled internally)
<i>Backbone</i>	
Architecture	ViT-Tiny (<code>vit_tiny_patch16_224</code>)
Pretrained	No (random initialisation)
Embedding dim	192
<i>Evaluation</i>	
Linear probe	Linear(192, C), cross-entropy, balanced accuracy
kNN probe	$k=20$, cosine distance, queue length 10,000
RankMe	Queue length 2,048
<i>Probe Targets</i>	
S&P 500	GICS sector label ($C=11$)
Fama–French	Factor quintile per factor, averaged ($C=5$)
<i>Fixed Window Size</i>	
S&P 500 × GAF+MTF	126 trading days
S&P 500 × candlestick	20 trading days
Fama–French × heatmap	20 trading days

Table 16: Hyperparameters for the supervised baseline (Study 3). The supervised model is trained end-to-end on the probe labels using the same backbone, optimiser, and schedule as the SSL methods, providing a fair reference ceiling for the probe evaluation. For the Fama–French dataset, the market factor (Mkt-RF) quintile is used as the primary training label; the resulting features are evaluated on all five factors at test time.

Setting	Value
<i>Training</i>	
Epochs	100
Batch size	256
Precision	16-bit mixed
Seed	42
Training views	1 (single view)
<i>Optimiser</i>	
Optimiser	AdamW
Learning rate	1.5×10^{-4}
Weight decay	0.05
Betas	(0.9, 0.999)
LR schedule	Cosine annealing with linear warmup
<i>Supervised Method</i>	
Loss	Cross-entropy on primary label (GICS sector for S&P 500; market factor for Fama–French)
Head	Linear(192, C) — no projector
<i>Backbone</i>	
Architecture	ViT-Tiny (vit_tiny_patch16_224)
Pretrained	No (random initialisation)
Embedding dim	192
<i>Evaluation</i>	
Linear probe	Linear(192, C), cross-entropy, balanced accuracy
kNN probe	$k=20$, cosine distance, queue length 10,000
RankMe	Queue length 2,048
<i>Probe Targets</i>	
S&P 500	GICS sector label ($C=11$)
Fama–French	Factor quintile per factor, averaged ($C=5$)
<i>Fixed Window Size</i>	
S&P 500 $\times$ GAF+MTF	126 trading days
S&P 500 $\times$ candlestick	20 trading days
Fama–French $\times$ heatmap	20 trading days