Market Manifolds: Geometry-Aware Latent Representations via β-VAE

Levente Szabo *
levbszabo@gmail.com

Abstract

We introduce a geometry-aware framework for modeling financial market states using a β-Variational Autoencoder (β-VAE). Standard Euclidean representations often fail to capture the non-linear structure of market dynamics. By treating the decoder as a parameterization of a latent manifold, we compute local Riemannian metrics via Jacobians and use these to define geodesic distances. This reveals intrinsic curvature in the learned representation, confirmed both visually and quantitatively. Compared to Euclidean clustering, geodesic K-means improves Silhouette score from 0.07 to 0.50 and halves the Davies–Bouldin index, validating the manifold hypothesis for financial time series.

1 Introduction

Modeling the behavior of financial markets is notoriously difficult: high-dimensional, noisy data obscures the low-dimensional structure that governs long-term dynamics. The *manifold hypothesis* suggests that high-dimensional financial observations lie near a smooth, non-linear manifold embedded in ambient space [1]. Regime shifts, coordinated sector moves, and systemic stress all suggest that daily price movements evolve along a latent manifold—one that is smooth but intrinsically curved.

This work proposes a framework for uncovering and analyzing that manifold. We apply a β-Variational Autoencoder (VAE) to compress daily market states into a structured latent space, and we treat the decoder as a parameterization of a Riemannian manifold. By computing geodesic distances using the Jacobian of the decoder, we reveal how curvature influences distance, clustering, and temporal structure.

1.1 Training a Stable Latent Geometry

While the β-VAE framework 2 provides a principled route to latent factor modeling, training on real-world market data introduces unique instabilities. We encountered three persistent challenges that degrade the quality and interpretability of the learned geometry:

1. **Posterior collapse**, where the encoder's output distribution converges to the prior and ceases to carry useful information.
2. **Latent correlation**, where the latent dimensions are highly redundant or aligned, violating the assumption of isotropic structure in the prior and making geometric analysis ill-posed.
3. **Capacity mismatch**, where the KL divergence is either too constrained or too loose at various stages of training, leading to underuse or over-regularization of the latent space.

*Code available at: github.com/levbszabo/market-latent-geometry

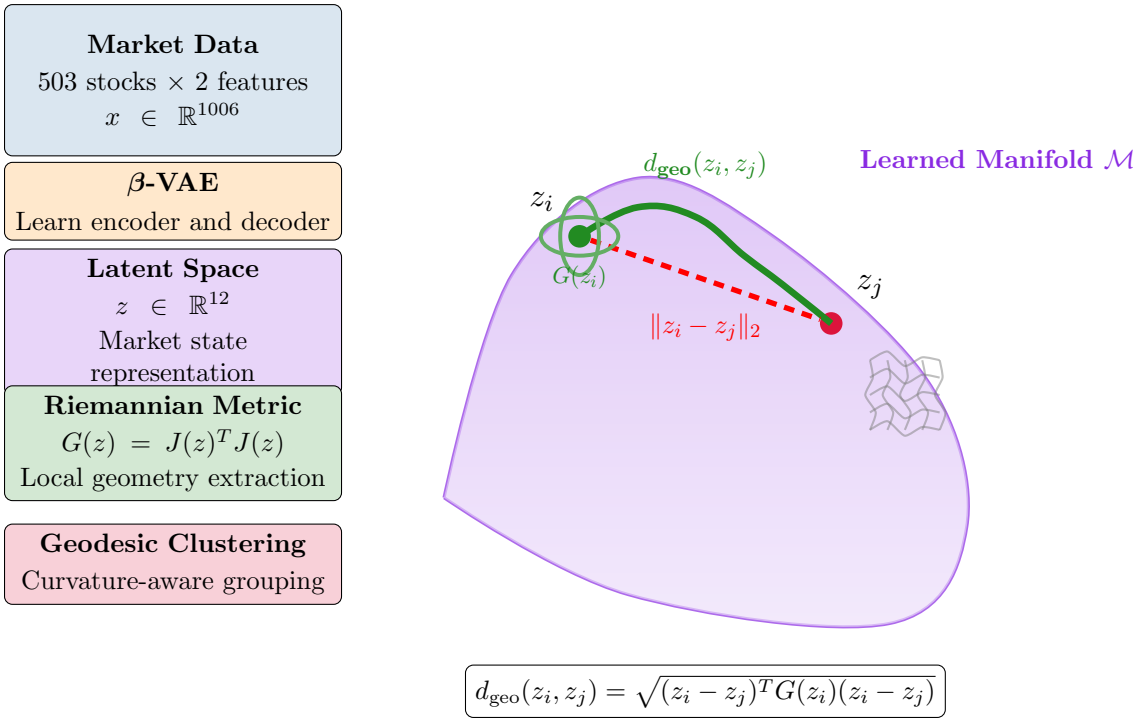


Figure 1: **Geometry-Aware Market Analysis Framework.** *Left:* The β-VAE pipeline transforms high-dimensional market data into a 12-dimensional latent space with specialized loss components for stability and disentanglement. *Right:* The learned latent manifold $\mathcal{M}$ exhibits intrinsic curvature, where geodesic distances (curved green path) respect the manifold geometry while Euclidean distances (red dashed line) do not. The metric tensor $G(z_i)$ captures local geometric properties enabling curvature-aware analysis.

To mitigate these challenges, we adopt a multi-component loss function:

$$L = L_{\text{recon}} + \beta \cdot |D_{\text{KL}} - C| + \lambda \cdot ||\text{Cov}(\mu_z) - I||_F^2 \tag{1}$$

KL Capacity Scheduling. We gradually increase the target KL divergence C over training time. This allows the model to begin learning meaningful reconstructions before being asked to match the prior distribution too closely. Early capacity ramp-up helps prevent collapse, while late-stage saturation encourages efficient use of the latent space.

Orthogonality Penalty. We regularize the covariance matrix of the latent means μ_z to approach the identity matrix. This promotes disentangled and uncorrelated latent factors—essential for using the decoder Jacobian as a reliable local chart of the manifold.

β **Scaling.** By setting $\beta = 1.0$, we control the strength of regularization relative to reconstruction fidelity. A higher β imposes greater structure on the latent space, while a lower value allows more flexibility but risks overfitting.

1.2 Related Work and Our Contributions

Prior works have applied deep generative models to financial time series. For example, recurrent latent-variable models for sequences [Chung et al., 2015] [3] and recent VAE-based approaches tailored to finance [Acciaio et al., 2024; Wang, Guo, 2024] [4, 5] aim to capture market dynamics, while GAN-based methods [Wiese et al., 2020] [6] focus on realistic data generation and stylized

facts. However, these methods largely treat latent space as Euclidean and do not address its underlying geometric structure. In contrast, our approach explores the latent market manifold geometry explicitly.

This work presents a principled yet practical approach to learning geometry-aware representations of financial markets using a β-VAE. Our key contributions are:

1. **A Stable Training Pipeline for Financial β-VAEs.** We design a robust architecture and loss strategy that overcomes posterior collapse, latent entanglement, and KL imbalance—common failure modes when applying VAEs to financial time series. By combining KL capacity scheduling with an orthogonality regularization term, we ensure that the latent space remains both informative and statistically well-structured.

2. **Latent Geometry via Decoder Jacobians.** We treat the VAE decoder as a parameterization of a learned manifold and compute the Riemannian metric tensor from its Jacobian. This enables local geometric analysis of the latent space, including distance and curvature estimation.

3. **Empirical Evidence of Market Curvature.** We provide strong quantitative support for the existence of curvature in the latent market manifold. Geodesic distances diverge nonlinearly from Euclidean ones.

4. **A Foundation for Downstream Generative and Policy Learning.** While we briefly considered defaulting to a pure autoencoder (i.e., $\beta = 0$) due to early training instability, we show that a properly tuned β-VAE yields a smooth, probabilistic latent space. This unlocks future work in generative modeling, forecasting, and reinforcement learning directly on the manifold.

2 Methodology

Our goal is to learn a smooth, interpretable latent representation of financial market states that supports geometric analysis. This section describes our dataset construction, model architecture, loss design, and the geometric tools we apply to the learned representation.

2.1 Data Construction

We use five years of daily data for all 503 stocks in the S&P 500 index. For each trading day, we compute two features per stock: **Log returns**: $\log\left(\frac{P_t}{P_{t-1}}\right)$ **Log volume**: $\log(V_t)$

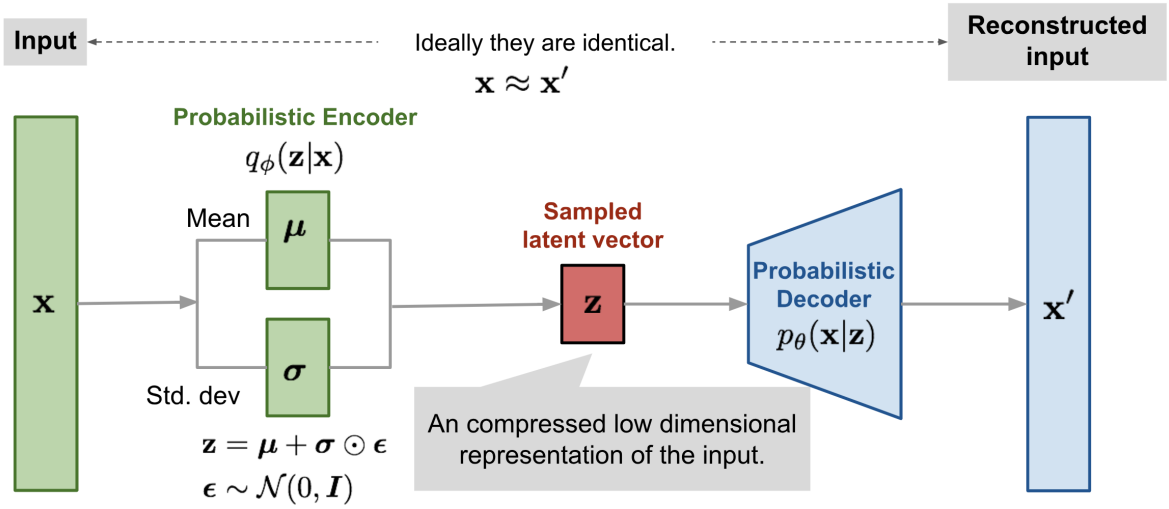
To ensure numerical stability and facilitate convergence, we apply global z-score normalization to all features:

$$\tilde{x}_{i,t} = \frac{x_{i,t} - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{2}$$

Where μ_{train} and σ_{train} are the mean and standard deviation computed exclusively from the training set. This normalization is applied consistently across training, validation, and test splits to prevent data leakage. Concatenating these features across all stocks yields a 1006-dimensional input vector $x \in \mathbb{R}^{1006}$ representing a single day in the market. Each input captures the cross-sectional state of the market at a specific point in time. The dataset includes approximately 1250 trading days, which we split chronologically into training (80%), validation (10%), and test (10%) sets.

2.2 Model Architecture

We use a symmetric Variational Autoencoder (VAE) with the following structure 7:



- **Encoder:** $1006 \to 256 \to 128 \to 12$ (latent mean and variance)
- **Decoder:** $12 \to 128 \to 256 \to 1006$

We use ReLU activations and dropout in all layers. The encoder learns a variational distribution $q_\phi(z|x)$, and the decoder defines a mapping $g_\theta(z)$ that reconstructs the input. The latent space $z \in \mathbb{R}^{12}$ is trained to capture essential structural variation in the market while remaining geometrically meaningful.

2.3 Loss Function and Training Strategy

Training VAEs on financial data is challenging due to instability, posterior collapse, and over-regularization. We adopt a custom objective that explicitly controls capacity and encourages disentanglement. This strategy is similar in spirit to encouraging factorized, independent latents as in FactorVAE [Kim & Mnih, 2018][8] and β-TCVAE [Chen et al., 2018]9.

$$L = L_{\text{recon}} + \beta \cdot |D_{\text{KL}} - C| + \lambda \cdot \|\text{Cov}(\mu_z) - I\|_F^2 \tag{3}$$

$$L = L_{\text{recon}} + L_{\text{KL}} + L_{\text{ortho}} \tag{4}$$

Reconstruction Loss (L_{recon}). The primary objective is to ensure the latent variables contain sufficient information to reconstruct the input. We use the standard Mean Squared Error (MSE):

$$L_{\text{recon}} = \mathbb{E}_{q_\phi(z|x)}[||x - g_\theta(z)||^2] \tag{5}$$

where $g_\theta(z)$ is the output of the decoder.

Capacity-Controlled KL Divergence (L_{KL}). To prevent the KL term from vanishing (posterior collapse) while still providing regularization, we adapt the loss to include a capacity term, C. This encourages the model to use a specific amount of informational capacity.

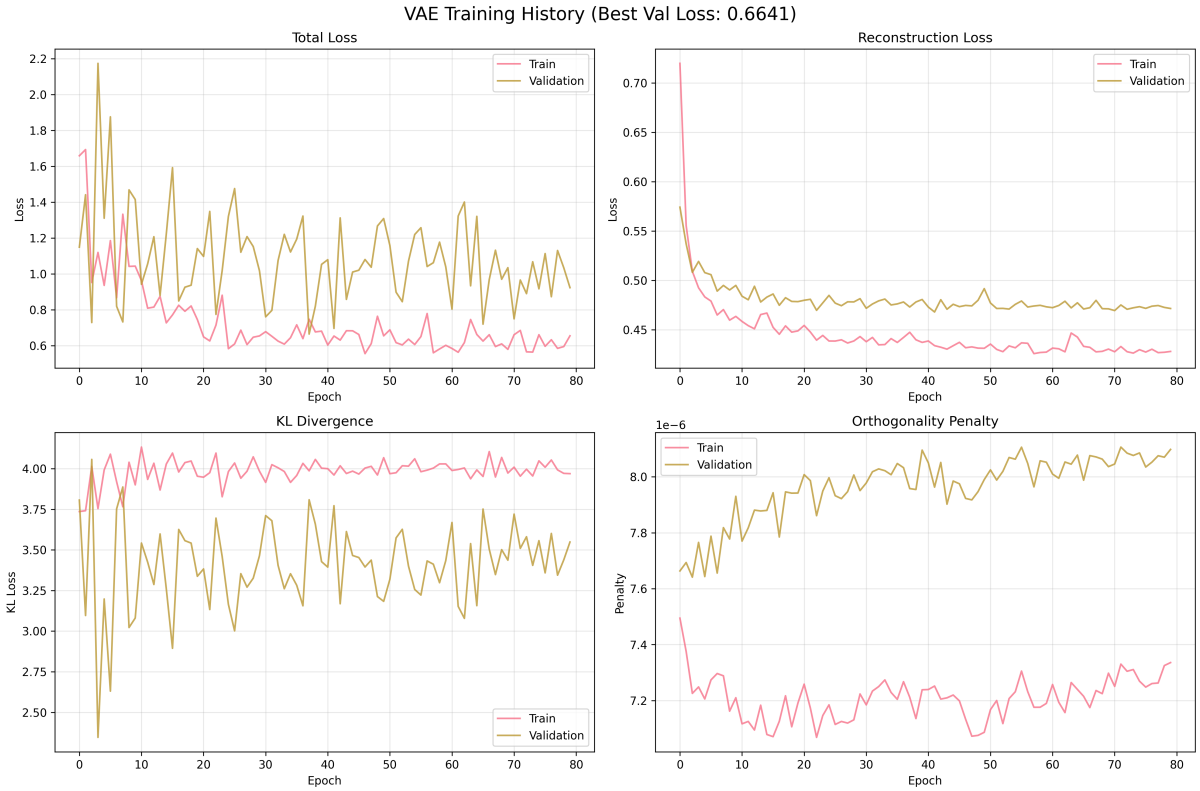
$$L_{\text{KL}} = \beta \cdot |D_{KL}(q_\phi(z|x)||p(z)) - C| \tag{6}$$

Here, $p(z)$ is the prior, typically $\mathcal{N}(0, I)$, β is the weight of the term, and we set the capacity $C = 4.0$. This forces the KL divergence away from zero, ensuring the latent variables are utilized.

Orthogonality Penalty (L_{ortho}). To encourage disentangled and interpretable latent factors, we add a penalty that forces the covariance matrix of the latent means (μ_z) to be close to the identity matrix.

$$L_{\text{ortho}} = \lambda \cdot ||\text{Cov}(\mu_z) - I||_F^2 \tag{7}$$

This penalty, weighted by $\lambda = 1 \times 10^{-4}$, explicitly decorrelates the learned latent dimensions. We set $\beta = 1.0$, gradually increase C to a final value of 4.0, and apply a small orthogonality penalty $\lambda = 1 \times 10^{-4}$ during training.



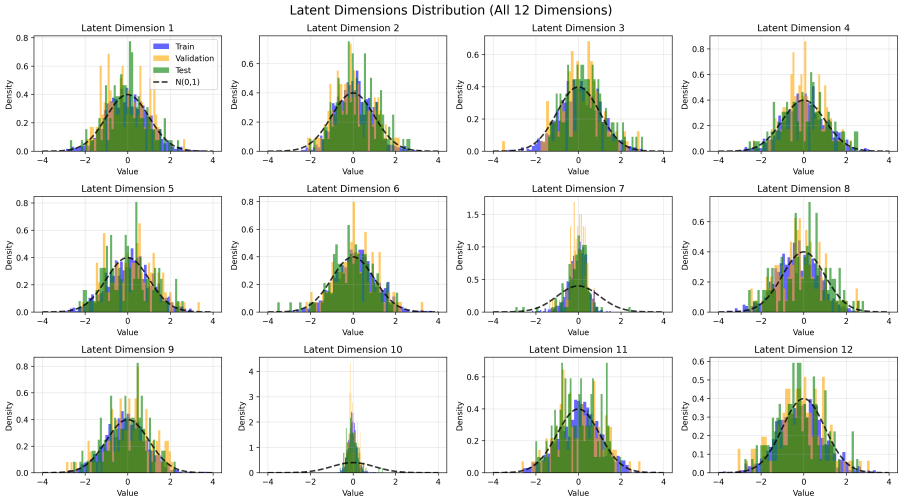
2.4 Latent Diagnostics: Normality & Independence

Before performing geometric analysis on the latent space, we first verify that the latent variables adhere to two essential assumptions of the VAE framework: **(1)** they are approximately *marginally standard normal*, and **(2)** they exhibit *minimal pairwise correlation*. These properties are critical to ensure that the learned representation respects the prior distribution and that the local Riemannian metric $G(z) = J^\top J$ is well-behaved and interpretable.

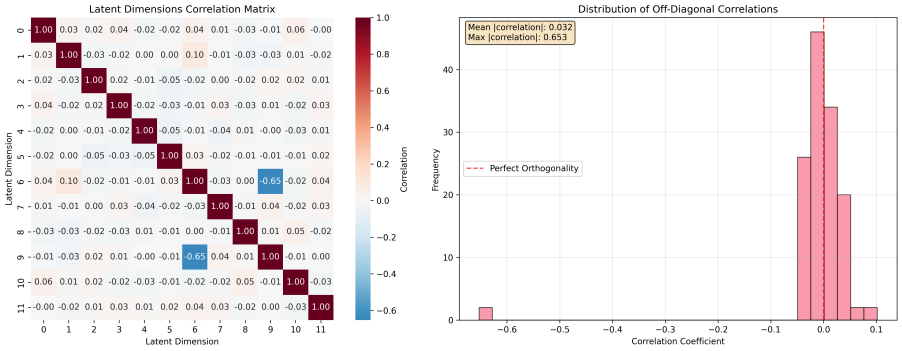
Interpretation. These results confirm that our model produces a latent space that is both *statistically aligned with the prior* and *geometrically stable*. The distributions are unimodal, symmetric, and centered around zero, with minimal skew or heavy tails. Likewise, the weak correlations across latent dimensions indicate successful disentanglement, made possible by the orthogonality regularization term. This decorrelation ensures that the local metric $G(z)$ is not distorted by redundant directions and supports accurate geodesic distance computation and curvature estimation in the following sections.

2.5 From Latent Space to Riemannian Manifold

Once the VAE is trained, we treat the decoder $g : Z \to X$ as a learned parameterization of the market manifold. This allows us to perform explicit geometric analysis.



(a) **Marginal Normality.** Histograms for each of the 12 latent dimensions, overlaid with the $\mathcal{N}(0, 1)$ probability density function (dashed line).



(b) **Latent Independence.** Left: empirical correlation matrix of the latent means μ_z. Right: distribution of off-diagonal correlation coefficients. The mean absolute correlation is 0.032, with a maximum of 0.065—both indicating low redundancy between dimensions.

Computing the Riemannian Metric. The geometry of the manifold is encoded in the Riemannian metric tensor, $G(z)$, which defines a local inner product at each point z in the latent space. We compute it from the decoder's Jacobian, $J(z) = \frac{\partial g(z)}{\partial z} \in \mathbb{R}^{1006\times 12}$, which measures how an infinitesimal change in the latent space z affects the output data space x. The metric tensor is then given by:

$$G(z) = J(z)^T J(z) \in \mathbb{R}^{12\times 12} \tag{8}$$

This matrix acts as a local "ruler," defining distances and angles on the manifold following the information-geometric interpretation of the decoder mapping [Amari (2016)]10. A similar pullback metric approach was used by Arvanitidis et al.(2018) 11 to demonstrate that VAE latent spaces can have significant curvature.

2.6 Approximating Geodesic Distance

With a Riemannian metric defined at each point in latent space, we can approximate distances that respect local curvature. Rather than using standard Euclidean distance $\|z_i - z_j\|_2$, we apply a first-order local approximation derived from the metric tensor $G(z)$.

$$d_{\text{geo}}(z_i, z_j) \approx \sqrt{(z_i - z_j)^\top G(z_i)\,(z_i - z_j)} \tag{9}$$

This formulation evaluates the local metric at the source point z_i, resulting in a Mahalanobis-like distance that accounts for anisotropic stretching or compression induced by the decoder. To ensure symmetry for downstream tasks like clustering or embedding, we assign this value to both entries in the distance matrix:

$$d_{\text{geo}}(z_i, z_j) = d_{\text{geo}}(z_j, z_i) := \sqrt{(z_i - z_j)^\top G(z_i)\,(z_i - z_j)}$$

Although this does not strictly equal the average of the two one-sided distances, we found the approximation sufficient in practice. It provides a curvature-aware alternative to Euclidean distance with minimal additional computational cost. We note that it is still a first-order approximation. More exact approaches compute true geodesics by integrating along the manifold or employ specialized solvers for shortest paths on learned manifolds [Arvanitidis et al., 2019] 12. We choose the simpler approximation for efficiency, noting that these more precise methods could improve long-range accuracy at higher computational expense.

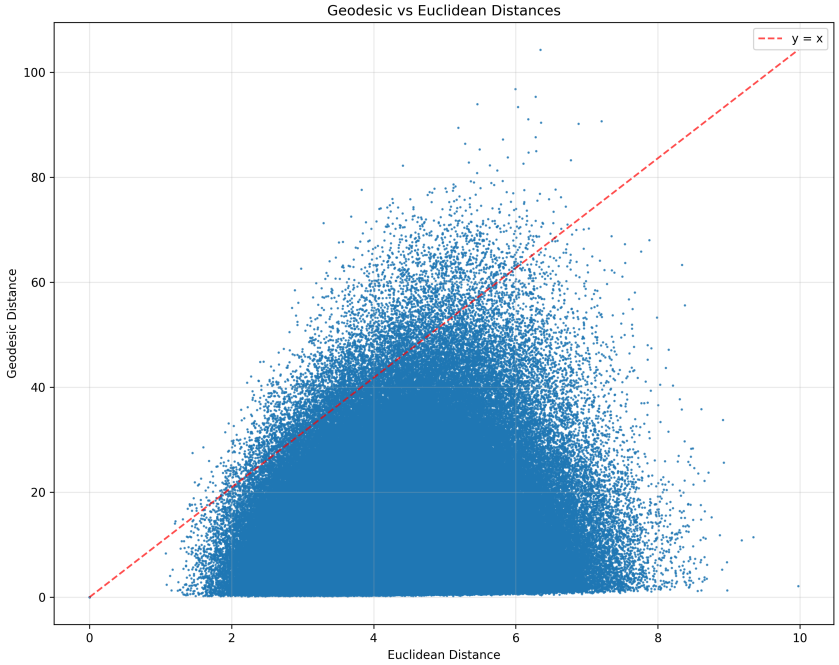


Figure 3: **Geodesic vs. Euclidean Distances.** Each point represents a pairwise comparison between latent vectors z_i and z_j.

3 Exploratory Geometry-Aware Clustering

With a curved latent manifold in hand, we ask a pragmatic question: *does a curvature-respecting distance yield clusterings that look more coherent than those obtained with flat metrics?* To explore this, we contrast three k-means variants, all run with $k = 5$ and identical random seeds:

1. **Euclidean**: vanilla k-means on the 12-D latent codes;
2. **Geodesic**: k-means on a metric-MDS embedding of the pairwise geodesic distances.
3. **PCA**: k-means on the first five principal components of latent space (54.3% variance retained).

3.1 Internal validity scores

We assess cluster quality using three widely adopted internal metrics:

- **Silhouette Score:** Measures the cohesion and separation of clusters. Values near 1 indicate tight, well-separated clusters; values near 0 suggest overlapping or ambiguous assignments.
- **Calinski–Harabasz Index (CH):** Compares between-cluster dispersion to within-cluster dispersion. Higher values indicate more distinct, well-separated clusters.
- **Davies–Bouldin Index (DB):** Quantifies average similarity between clusters. Lower values suggest better separation and compactness.

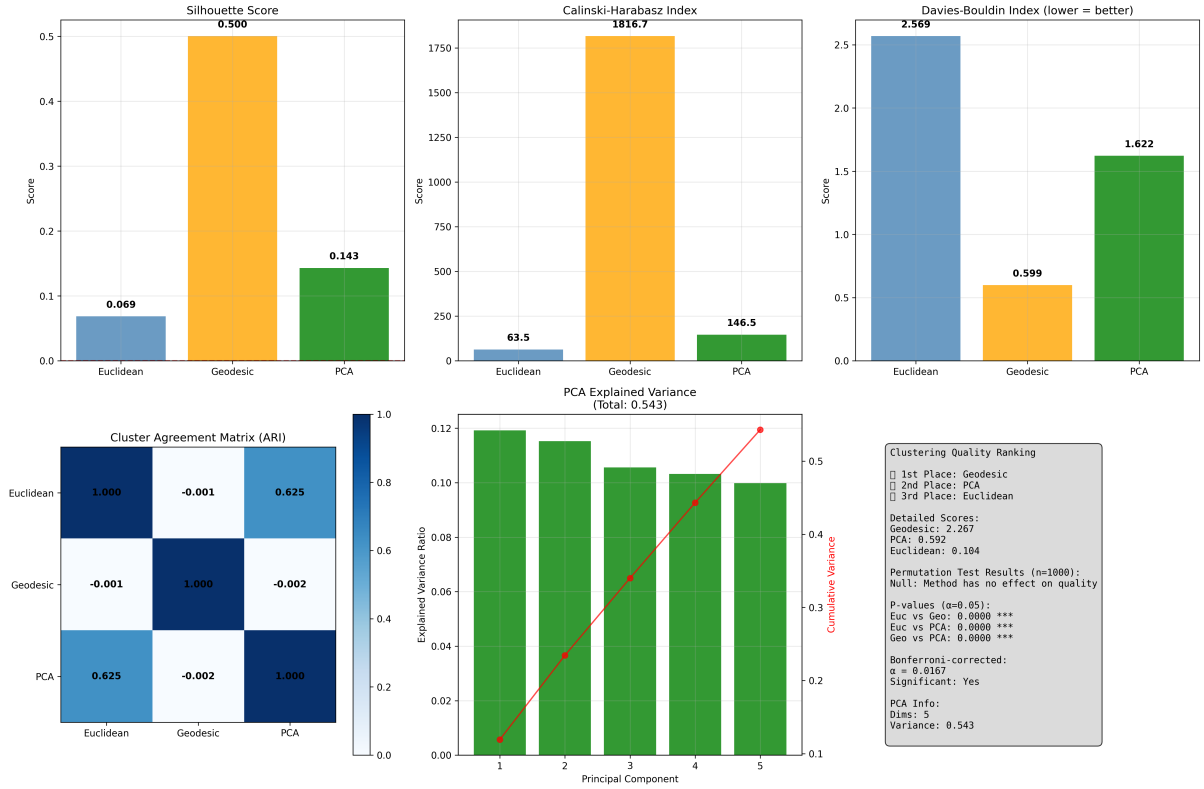


Figure 4: **Internal clustering diagnostics.** Evaluation of three clustering methods—Euclidean, Geodesic, and PCA-based—across standard metrics: Silhouette, Calinski–Harabasz (CH), and Davies–Bouldin (DB).

Figure 4 compares clustering performance across the three approaches. The geodesic method shows large apparent gains across all metrics: Silhouette improves from 0.07 (Euclidean) to 0.50, CH from 64 to 1,817, and DB falls from 2.57 to 0.60.

A label-randomisation test with 1 000 permutations confirms that every pairwise metric difference is highly significant (*all* $p < 10^{-3}$; Bonferroni-corrected α=0.0167). The results are consistent with the geometric hypothesis that the latent manifold is curved, and that geodesic distances better reflect its internal structure. This is consistent with prior observations as well [Yang et al.] (2018) 13 found that clustering latent representations using geodesic distances produced more semantically coherent groupings compared to Euclidean distances. Further validation of our time series based market manifold clustering can be achieved through either larger datasets or downstream predictive tasks.

3.2 Cluster morphology

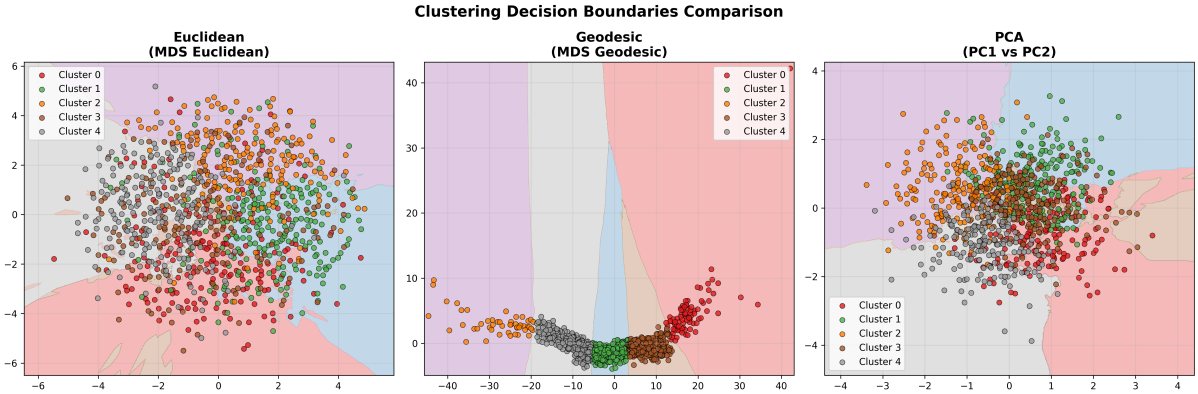


Figure 5: **Decision surfaces in two-dimensional embeddings.** For each distance choice we embed the latent points in $\mathbb{R}^2$ (MDS for Euclidean and geodesic, first two PCs for PCA), train a k-NN classifier on the resulting cluster labels, and colour the background by the predicted class. Geodesic clusters unwrap into elongated bands that follow the manifold's main bend, whereas Euclidean and PCA clusters remain roughly spherical and frequently overlap.

The decision surface visualises why the internal indices favour a curvature-aware distance. Because the geodesic metric "stretches" space along directions where the decoder is locally compressed, k-means discovers bands that align with the intrinsic U-shape rather than cutting across it. With a flat metric, the same algorithm sees an almost isotropic cloud and produces overlapping, disk-like regions. Similarly, Yang et al. (2018) [13] observed that geodesic latent clusters captured meaningful variations that flat-space clustered missed.

3.3 Temporal coherence

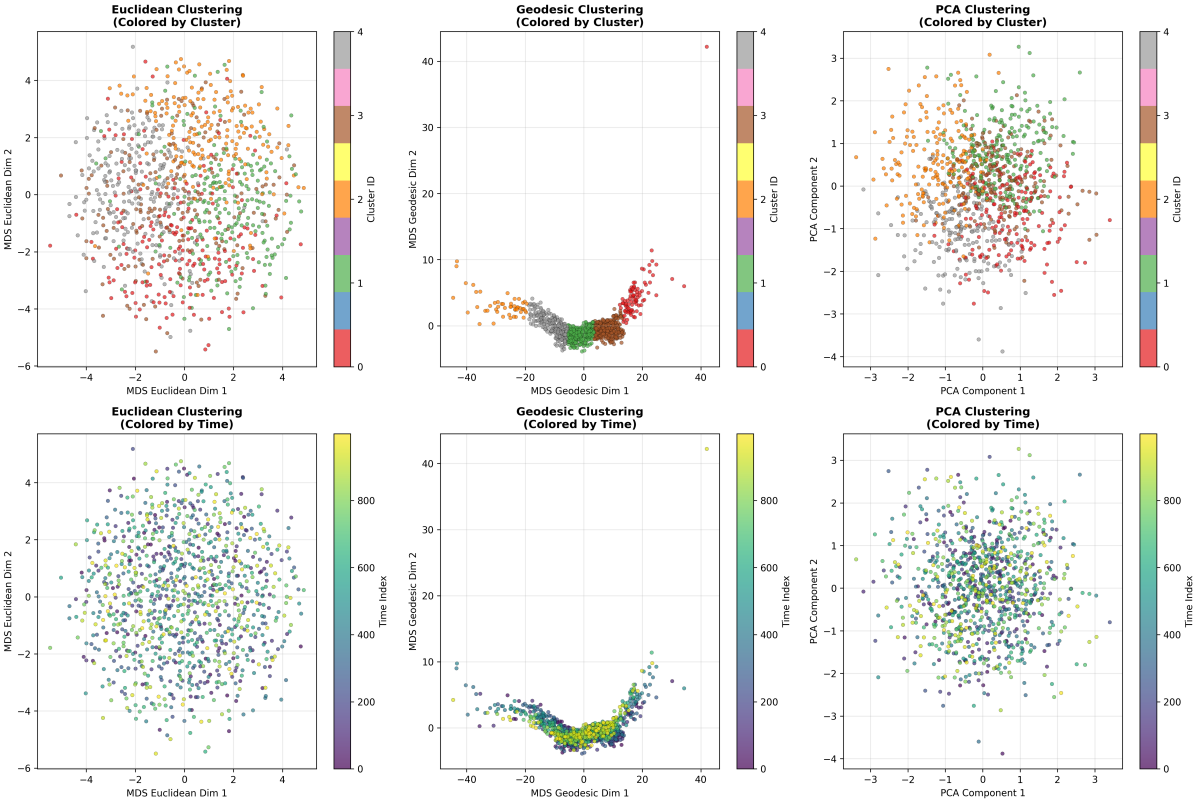


Figure 6: **Temporal colour-coding of clusters.** Top: points coloured by cluster label. Bottom: the same points coloured by time index (purple $\rightarrow$ yellow). Only the geodesic embedding shows a roughly monotone march through time, hinting at regime segmentation.

Colouring points by chronological order reveals that geodesic clusters partition the latent trajectory into contiguous temporal blocks, whereas Euclidean and PCA splits appear more arbitrary. This temporal coherence supports (but does not yet prove) the idea that curvature-aware distances isolate market *regimes*.

3.4 Three-dimensional perspective

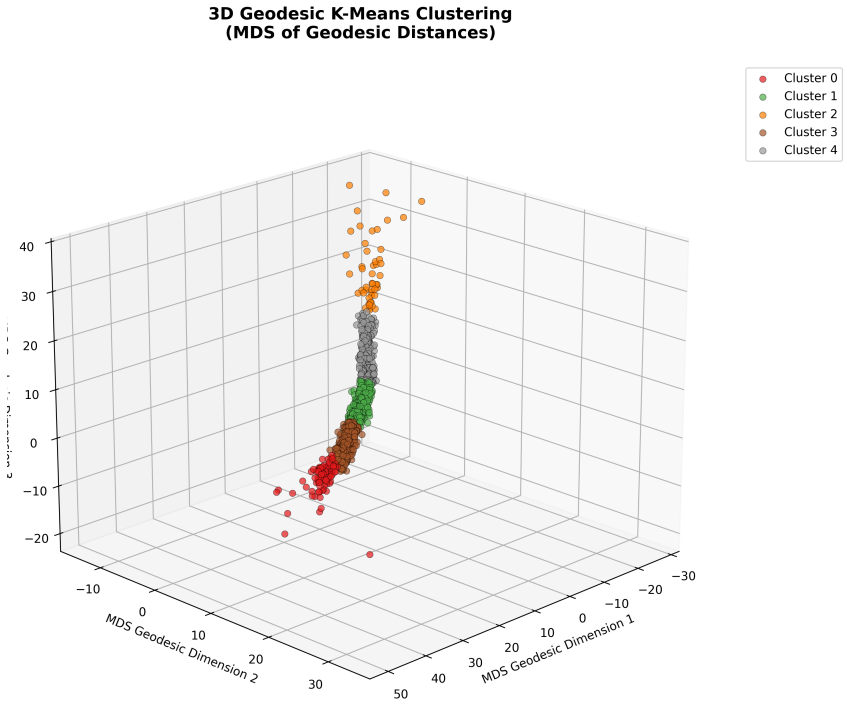


Figure 7: **3-D snapshot of geodesic clusters.** Metric-MDS coordinates 1–3, coloured by k-means label. Clusters occupy smoothly curved sheets rather than intersecting spheres. An interactive version is provided in the supplement.

The 3-D view illustrates how clusters fan out along the manifold rather than piling up near the origin. This qualitative evidence complements the internal scores, reinforcing the exploratory claim that respecting curvature yields more interpretable groupings.

Interim takeaway. Geometry-aware distances appear promising for unsupervised regime discovery, but larger samples and out-of-sample tests are required before the method can be deemed production-ready.

4 Discussion and Outlook

The exploratory results in Section 3 show that a curvature–aware distance can sharpen unsupervised structure in the latent space of financial time–series. At the same time, our sample is modest and the statistical support remains preliminary. Below we clarify what these findings imply, outline concrete next steps, and note key limitations.

4.1 Market-regime interpretation

Every latent point already carries the full cross-section of log returns and log volumes for all stocks in SP500, so each cluster implicitly embeds a rich panel of raw market data. A natural next step is therefore to *describe* each regime $\mathcal{R}_k$ with financially interpretable statistics—e.g. average index-level return, cross-sectional volatility, sector tilts, or liquidity measures—and then examine how those quantities evolve when the trajectory moves from one cluster to another. We also intend to analyze the latent feature vectors and determine a quantitative approach for determining their financial relevance (e.g Volatility, Fear, etc.)

4.2 Scenario generation and stress testing

Interpolation along *geodesic paths* yields realistic market scenarios that honour latent curvature, unlike straight-line walks that traverse low-density regions. Regulators and risk managers could sample stress trajectories between present conditions and extreme historical regimes to compute VaR or expected shortfall under coherent dynamics.

4.3 Geometry-aware reinforcement learning

Standard RL penalises L2 distances in action or state space; replacing these with geodesic penalties would discourage agents from taking actions that move the market into high-curvature (unstable) regions. The differentiable β-VAE decoder further allows gradient-based policy updates directly through the manifold.

4.4 Limitations

- **First-order distance** — Our approach ignores metric variation along the path. Higher-order or numerical geodesics would improve long-range accuracy.
- **Feature choice** — We used log-returns and volumes only. Extending to options data or order-book states may uncover richer geometry.
- **Sample size** — the 1 250-day window limits statistical testing power. Rolling-window experiments are required to confirm robustness.

4.5 Future Research Directions

This work establishes a new foundation for geometry-aware quantitative finance. Several exciting research avenues are now open:

- **Generative Modeling:** One can sample along geodesic paths between two market states to generate realistic, plausible transition scenarios for stress testing and risk analysis. This is superior to linear interpolation in the latent space, which would traverse regions the market never visits.
- **Reinforcement Learning on Manifolds:** A trading agent's policy could be optimized directly on the learned manifold. The geodesic distance could serve as a more meaningful penalty for large actions, and the curvature could inform the agent about local market stability.
- **Curvature as a Risk Indicator:** The local curvature of the manifold could itself be a novel risk factor. A systematic increase in curvature might precede periods of high volatility or market crashes, serving as an early-warning signal.
- **Generalization:** The framework presented here is general and can be applied to other financial markets, such as foreign exchange, commodities, or cryptocurrencies, to explore and compare their intrinsic geometries.

Take-away: respecting intrinsic geometry transforms a simple clustering task into a powerful regime-detection tool and opens a path toward fully geometry-aware market simulators and control algorithms.

4.6 Conclusion

We introduced a β-VAE that maps daily cross-sectional returns and volumes onto a twelve-dimensional latent manifold. Jacobian-based metrics reveal clear curvature, and a geodesic K-means built on that metric yields tighter, more chronologically ordered market clusters than Euclidean or PCA baselines. Although the test set is small, permutation tests reject the null of "no difference" at $p < 10^{-3}$, and the effect sizes and visuals across three cluster-quality indices point to material signal. Future work will extend the pipeline to geodesic sampling, regime-conditioned risk models, and reinforcement-learning agents that navigate the latent market space.

A Hyperparameter Configuration

The complete hyperparameter configuration used for all experiments is detailed in Table 1. These values were held constant across training, validation, and testing.

Table 1: Complete hyperparameter configuration used for model training.

Hyperparameter	**Value**
Model Architecture	
Input Dimension	1006
Hidden Dimension	128
Latent Dimension	12
Dropout Rate	0.1
Training Parameters	
Batch Size	32
Learning Rate	1×10^{-3}
Max Epochs	80
Early Stopping Patience	200
Weight Decay	1×10^{-4}
VAE Loss Parameters	
β (KL Weight)	1.0
C (KL Capacity)	4.0
λ_{ortho} (Orthogonality Penalty)	1×10^{-4}

References

[1] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. doi: 10.1090/jams/855.

[2] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

[3] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C. Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *NeurIPS*, 2015. URL `https://papers.nips.cc/paper_files/paper/2015/hash/b618c3210e934362ac261db280128c22-Abstract.html`.

[4] Beatrice Acciaio, Stephan Eckstein, and Songyan Hou. Time-causal vae: Robust financial time series generator. *arXiv preprint arXiv:2411.02947*, 2024. URL https://arxiv.org/abs/2411.02947.

[5] Yilun Wang and Shengjie Guo. RVRAE: A dynamic factor model based on variational recurrent autoencoder for stock returns prediction. *arXiv preprint arXiv:2403.02500*, 2024. URL https://arxiv.org/abs/2403.02500.

[6] Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. Quant GANs: Deep generation of financial time series. *Quantitative Finance*, 20(9):1419–1440, 2020. doi: 10.1080/14697688.2020.1730426.

[7] Lilian Weng. From autoencoder to beta-vae. https://lilianweng.github.io/posts/2018-08-12-vae/, 2018. Accessed: 2025-07-29.

[8] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

[9] Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018. URL http://arxiv.org/abs/1802.04942.

[10] Shun ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer, 2016. URL https://link.springer.com/book/10.1007/978-4-431-55978-8.

[11] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. In *ICLR*, 2018. URL https://openreview.net/forum?id=SJzRZ-WCZ.

[12] Georgios Arvanitidis, Søren Hauberg, Philipp Hennig, and Michael Schober. Fast and robust shortest paths on manifolds learned from data. In *AISTATS*, volume 89, pages 1506–1515, 2019. URL http://proceedings.mlr.press/v89/arvanitidis19a.html.

[13] Tao Yang, Georgios Arvanitidis, Dongmei Fu, Xiaogang Li, and Søren Hauberg. Geodesic clustering in deep generative models. *arXiv preprint arXiv:1809.04747*, 2018. URL http://arxiv.org/abs/1809.04747.