Understanding Spiking Neural Networks Through Loss Landscapes and Diffusion Dynamics

Anonymous authors
Paper under double-blind review

Abstract

Understanding the interplay between loss landscape geometry and weight dynamics is crucial for elucidating the generalization mechanisms of Spiking Neural Networks (SNNs). However, how distinct plasticity rules shape these geometric and dynamical properties remains underexplored. We present a unified framework integrating time-averaged mean squared displacement (TAMSD) with surrogate-based curvature measurements to systematically characterize three plasticity paradigms: global backpropagation (GP), local Hebbian learning (LP), and hybrid plasticity (HP). While all paradigms exhibit a transition from early super-diffusion to late sub-diffusion, they display fundamentally distinct late-stage signatures. GP sustains irregular exploration within sharp minima; LP exhibits nearly lag-independent displacement with rough trajectories and intermediate performance; crucially, HP achieves an intermediate effective displacement-scaling exponent, balancing exploration and stabilization. This dynamical behavior correlates with HP converging to significantly flatter minima, yielding superior test accuracy and robustness to input corruptions compared to GP and LP. Our results provide complementary geometric and dynamical evidence for an implicit regularization effect inherent to hybrid plasticity, where local updates modulate the optimization trajectory without explicitly penalizing curvature. These findings underscore the value of jointly analyzing approximate geometry and empirical training trajectories to uncover the mechanistic principles governing learning in SNNs.

1 Introduction

Spiking Neural Networks (SNNs) offer energy-efficient and biologically plausible computation, making them promising candidates for neuromorphic hardware (Maass, 1997; Indiveri & Liu, 2015). Although learning rules such as surrogate gradient methods (Zenke & Ganguli, 2018; Gygax & Zenke, 2025) and exact gradient approaches (Wunderlich & Pehle, 2021) have enabled the training of SNNs, a fundamental question remains open: why do certain SNN training paradigms generalize better than others, and what are the geometric and dynamical mechanisms underlying these differences? The discrete, event-driven nature of spiking dynamics violates the core assumptions of existing theories developed for artificial neural networks (ANNs), precluding their direct application to SNNs. In this study, we adapt methods from non-equilibrium statistical mechanics to investigate loss landscape and anomalous diffusion in weight dynamics across different SNN learning paradigms. Focusing on three representative learning paradigms as examples: Global Plasticity (GP), Local Plasticity (LP), and Hybrid Plasticity (HP) (Wu et al., 2022), We elucidate the mechanism by which these learning rules affect the learning process of SNNs.

1.1 Related Work

SNN training methods. One line of research has focused on developing training methods for SNNs. Surrogate gradient learning (Neftci et al., 2019; Zenke & Ganguli, 2018) enables back-propagation through time by smoothing the spike function; its theoretical properties have been recently elucidated(Gygax & Zenke, 2025). Complementary approaches include exact gradient methods (Wunderlich & Pehle, 2021), spatio-temporal backpropagation (Wu et al., 2018), biologically inspired local rules such as STDP (Bi & Poo, 1998), and eligibility propagation (e-prop) (Bellec et al., 2020), which addresses the credit assignment problem in recurrent SNNs through biologi-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

cally plausible eligibility traces. A hybrid framework combining global (backpropagation-driven) and local (Hebbian-driven) weight branches is proposed, demonstrating enhanced performance and robustness.

Loss landscape geometry and SGD dynamics. In another line of research, the loss landscape geometry of deep ANNs has been investigated as essential for understanding the performance of SGD. Connections between deep networks and spin glass models have been drawn, revealing that the number of local minima decreases with depth (Choromanska et al., 2015). From a complementary perspective using algebraic geometry, deeper architectures have been shown to have fewer stationary points and more clustered minima, providing an independent theoretical explanation for the optimisability of deep networks (Becker et al., 2020). The Hessian spectrum has been found to comprise a bulk near zero and outlier eigenvalues that govern optimisation (Sagun et al., 2018; Papyan, 2018).Visualisation and dimension-reduction studies further reveal that skip connections smooth the landscape (Li et al., 2018) and that SGD navigates through saddle points to select descent directions(Im et al., 2016). Diffusion approximations for SGD have been given rigorous error bounds (Li & Wang, 2022), and stability boundaries have been characterised by local loss surface geometry (Dexter et al., 2024).

Concurrently, random matrix analyses of trained weight matrices have revealed the emergence of heavy-tailed spectral distributions (Arous & Guionnet, 2008; Mahoney & Martin, 2019). Sharp minima have been shown to generalise poorly (Keskar et al., 2016), with further support from subsequent studies (Chen et al., 2025; Foret et al., 2021). The maximum eigenvalue $\lambda_{\max}$ has been shown to self-regulate to $\sim 1/\eta$ during training—the edge of stability phenomenon (Cohen & Koltun, 2021; Kalra et al., 2025). Saddle points, rather than local minima, have been identified as the main obstacle for high-dimensional non-convex optimisation, motivating second-order methods for escape (Dauphin et al., 2014). Theoretical analysis has further shown that SGD can efficiently escape saddle points in non-convex landscapes (Fang et al., 2019), and that the anisotropic structure of SGD noise facilitates escape from sharp minima, acting as an implicit regulariser (Zhu et al., 2019). More broadly, it has been shown that the full solution landscape—comprising all stationary points and their connecting pathways—can be systematically constructed via high-index saddle dynamics (Yin et al., 2021).

Anomalous diffusion in neural network training. Furthermore, neural network training has been analysed by modelling SGD as a process exhibiting anomalous diffusion. SGD was modelled as a stochastic diffusion process via Langevin dynamics (Welling & Teh, 2011), and this connection was further developed (Mandt et al., 2017). Heavy-tailed behaviour in the stochastic gradient noise has been reported, leading to modelling SGD as a single stochastic differential equation driven by a Lévy alpha-stable process, with follow-up studies confirming these findings (Gurbuzbalaban et al., 2021). Within this framework, it has been demonstrated that weight dynamics in deep ANNs exhibit anomalous diffusion: training begins in a super-diffusive regime and gradually transitions to subdiffusion, with the diffusion exponent correlated with generalisation performance (Chen et al., 2022). Building on this, the loss landscape has been shown to possess multifractal structure, providing a unified explanation for diverse geometric and dynamical phenomena in deep learning (Ly & Gong, 2025). Despite these advances, no analogous framework exists for SNNs, where the landscape evolves non-stationarily and noise is shaped by surrogate-gradient and spike-driven effects. Our contributions are:

Contributions. Our main contributions are summarized as follows:

- **Unified geometric-dynamical framework for SNNs.** We establish a novel analytical framework that bridges loss landscape geometry with anomalous weight diffusion in Spiking Neural Networks. By integrating time-averaged mean squared displacement (TAMSD) with surrogate-based curvature measurements, we provide a systematic characterization of the distinct optimization trajectories induced by Global (GP), Local (LP), and Hybrid (HP) plasticity.
- **Distinct dynamical signatures and diffusion regimes.** We identify a universal transition from early super-diffusion to late sub-diffusion across all paradigms. However, each paradigm exhibits unique late-stage dynamical signatures: GP sustains irregular exploration in sharp regions, LP suffers from premature stagnation with rough trajectories, while HP

achieves an intermediate diffusion scaling exponent. This balanced dynamical regime allows HP to stabilize convergence while retaining sufficient exploration capacity, offering a mechanistic explanation for its robust performance.

- **Implicit regularization via hybrid plasticity.** We suggest that HP converges to significantly flatter minima compared to GP and LP, leading to superior generalization and noise robustness. Crucially, we reveal that this flatness is not achieved through explicit curvature penalization; rather, local Hebbian updates act as an *implicit regularizer* that modulates the effective noise covariance, guiding the optimization trajectory toward broad basins without direct gradient manipulation.

2 Methodology

2.1 Learning Rules and Paradigms

We use three optimisation paradigms to train SNNs(Wu et al., 2022): global plasticity (GP), local plasticity (LP), and hybrid plasticity (HP). The SNN use LIF model and the three optimisation schemes are all based on a common core construction: the network weights of layer l are decomposed into two independent parts:

$$\mathbf{w}^l = \mathbf{w}^l_{\mathrm{GP}} + \mathbf{w}^l_{\mathrm{LP}}, \tag{1}$$

where l is the layer index, $\mathbf{w}^l$ is the total weight parameter of layer l, $\mathbf{w}^l_{\mathrm{GP}}$ is the global plasticity (GP) weight branch, and $\mathbf{w}^l_{\mathrm{LP}}$ is the local plasticity (LP) weight branch. GP denotes the global plasticity mechanism, where weights are updated by the global error gradient computed via backpropagation, optimising the overall objective function and capturing global features. LP denotes the local plasticity mechanism, where weights are updated independently by the activities of adjacent neurons (such as spike states or membrane potentials) through local learning rules, without relying on global error backpropagation, enhancing local spatiotemporal plasticity and dynamic adaptation.

2.1.1 Global Plasticity

GP follows the global error-driven learning principle: supervised errors are allocated layer-by-layer via BPTT with surrogate gradients. The local branch is completely disabled ($\mathbf{w}^l_{\mathrm{LP}} \equiv 0$), and the parameter space reduces to the feedforward weights $\{\mathbf{W}^l\}$ alone. The weight update is

$$\tau_{\mathrm{GP}} \frac{dw^l_{ij}}{dt} = -\nabla_{\mathbf{w}^l} L_{\mathrm{CE}}. \tag{2}$$

2.1.2 Local Plasticity

LP follows Hebbian learning rules, where the weight update is determined by pre- and post-synaptic neuronal properties:

$$\tau_{\mathrm{LP}} \frac{dw_{ij}}{dt} = -w_{ij}(t) + P\big(t, \mathrm{pre}_j(t), \mathrm{post}_i(t); \theta\big), \tag{3}$$

2.1.3 Hybrid Plasticity

HP integrates GP and LP into a single model. Crucially, this is *not* a joint optimisation of a combined loss $L_{\mathrm{CE}} + L_{\mathrm{hebb}}$, but rather two independent update streams: one gradient-driven and one correlation-driven. The global branch $\mathbf{w}^l_{\mathrm{GP}}$ receives gradients from L_{CE} (as in the GP update), while the local branch $\mathbf{w}^l_{\mathrm{LP}}$ is updated via the Hebbian rule. The effective weights $\mathbf{w}^l = \mathbf{w}^l_{\mathrm{GP}} + \mathbf{w}^l_{\mathrm{LP}}$ naturally integrate both signals.

Table 1: Network architectures and training hyperparameters.

Parameter	Value	Parameter	Value
Dataset	Fashion-MNIST	Batch Size	100
Architecture	$[784 \to 512 \to 256 \to 10]$	Number of Epochs	20
Surrogate Gradient Steepness (κ)	2	Time Window (T)	10
Membrane Decay (γ)	0.4	Threshold (v_{th})	0.35
Synaptic Filter (τ_w)	40	Weight Decay (ρ)	0.95
GP Learning Rate	1×10^{-3}	LP Learning Rate	5×10^{-4}

2.2 Characterizing anomalous diffusion

2.2.1 Time-averaged Mean Squared Displacement

Treating the optimisation trajectory $\mathbf{w}(t)$ as a stochastic process, the time-averaged mean squared displacement (TAMSD) is defined as

$$\Delta r^2(t_w, \tau) = \frac{1}{T} \sum_{t=t_w}^{t_w+T} \sum_{i=1}^{d} \left(w_i(t+\tau) - w_i(t)\right)^2, \tag{4}$$

where τ is the lag time, T is the length of the time interval $[t_w, t_w + T]$, and $w_i(t)$ is the value of the i-th coordinate weight at time t. t_w is the time lapse after the start of the training process (i.e., waiting time). The time variables t, t_w, and T are in units of iteration, and the unit time step corresponds to a single update step of the weights.

2.2.2 Local Diffusion Exponent

The instantaneous, lag-dependent scaling is further characterized by the logarithmic derivative

$$\beta(t_w, \tau) = \frac{\partial \ln \Delta r^2(t_w, \tau)}{\partial \ln \tau}, \tag{5}$$

which tracks how the local diffusion regime evolves across training.

2.2.3 Global Fitting and Regime Classification

The global diffusion exponent α is extracted by fitting $\Delta r^2(t_w, \tau) \sim \tau^\alpha$ over the full lag-time range:

$$\begin{cases} \alpha < 1 & \text{subdiffusive,} \\ \alpha \approx 1 & \text{normal,} \\ \alpha > 1 & \text{superdiffusive.} \end{cases} \tag{6}$$

2.3 Characterizing the Loss Landscape

2.3.1 Hessian Definition

The Hessian matrix is defined as the second derivative of the loss function with respect to synaptic weights:

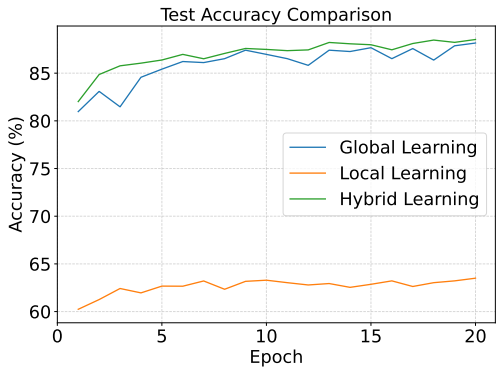
$$\mathbf{H} = \nabla^2_{\mathbf{w}} L(\mathbf{w}) \in \mathbb{R}^{N_{\text{out}}N_{\text{in}} \times N_{\text{out}}N_{\text{in}}}, \tag{7}$$

where $\mathbf{w} \in \mathbb{R}^{N_{\text{out}} \times N_{\text{in}}}$ is the weight matrix and $L(\mathbf{w})$ is the loss function defined by the corresponding learning paradigm.

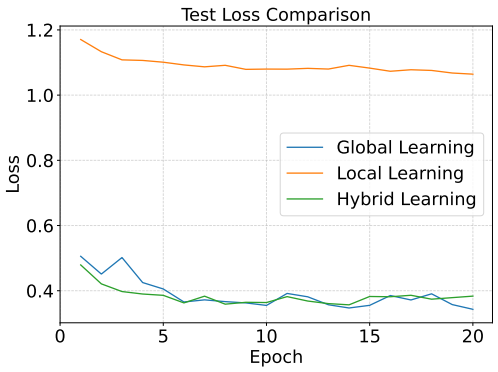
3 Results and Analysis

3.1 Training Dynamics and Generalization Performance

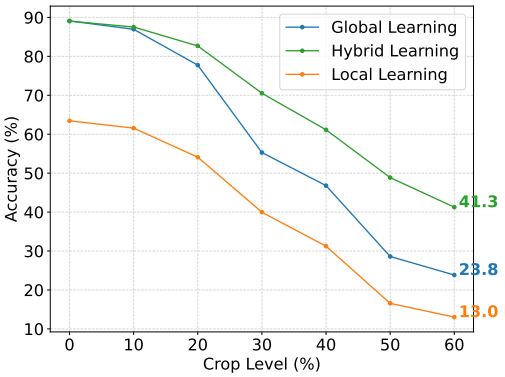
First of all, we revaluate the fully-connected Spiking Neural Networks (SNNs) on Fashion-MNIST under three learning paradigms. The network architecture and training hyperparameters are summarized in Table1. As shown in Figure1, Hybrid Plasticity (HP) and Global Plasticity (GP) achieve



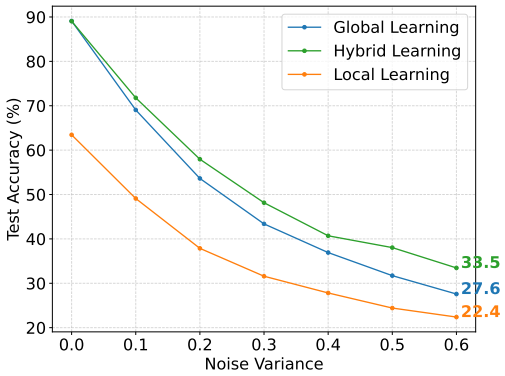
(a) Evolution of test accuracy with epoch.



(b) Evolution of test loss with epoch.



(c) Performance comparison in the cropping experiments.



(d) Performance comparison in the noise experiments.

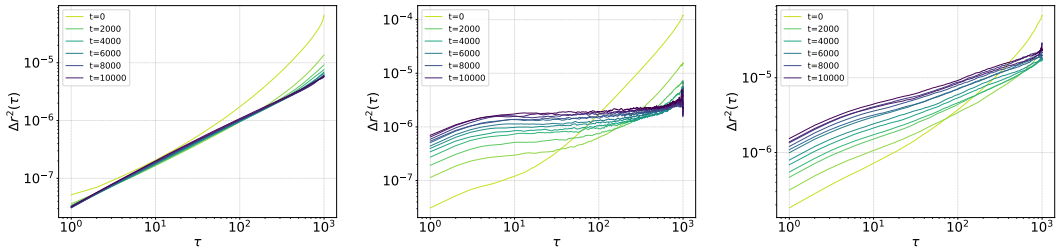
Figure 1: Training dynamics and robustness across learning paradigms on the FC-SNN. (a) Test classification accuracy as a function of training epoch. (b) Test loss as a function of training epoch. (c) Performance comparison in cropping experiments, where a fixed proportion of the peripheral input pixels is zeroed out. (d) Performance comparison in noise experiments, where additive Gaussian noise is injected into the input at varying intensity levels.

comparable accuracy, with HP being slightly higher. Moreover, HP suggests significantly superior noise robustness compared to GP and attains the lowest test loss. In contrast, GP exhibits a relatively higher final loss. Under Local Plasticity (LP), the model performs the worst. These results indicate that HP exhibits significantly superior noise robustness compared to both GP and LP, consistent with the findings of Wu et al.(Wu et al., 2022).

3.1.1 Anomalous Diffusion of Synaptic Weights

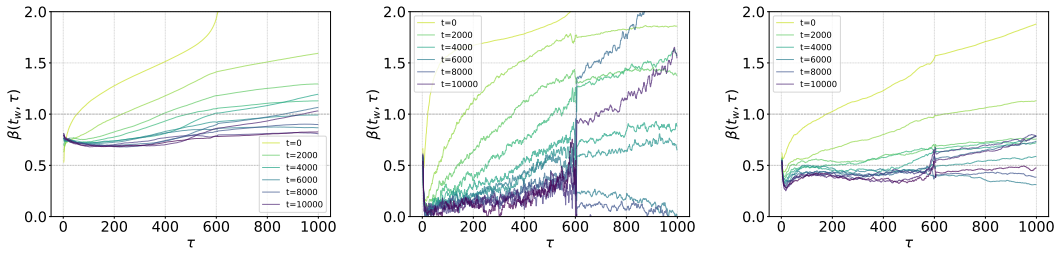
We compute the time-averaged mean squared displacement (TA-MSD) for each paradigm (Equation (4)). As shown in Figure 2, the diffusion gradually slows over time for all paradigms. Under LP, the first two layers exhibit no significant diffusion, while only the final layer shows comparatively pronounced diffusion under the influence of backpropagated errors. Under GP, the diffusion decays rapidly yet remains relatively strong at later times. To further characterise the local scaling behaviour, we extract the time-dependent diffusion exponent $\beta(t_w, \tau)$ (Figure 3), which reveals how the instantaneous scaling regime of weight dynamics changes throughout training. Under all three paradigms, β exhibits a clear transition from superdiffusive to subdiffusive behaviour, consistent with the non-stationary nature of the optimisation process.

We compute the time-averaged mean squared displacement (TA-MSD) for each paradigm and extract the diffusion exponent α via linear regression in log-log space ($\log(\text{MSD}) \sim \alpha \log(\Delta)$). Overall, all three learning paradigms exhibit an evolution from early superdiffusive ($\alpha > 1$) to late subdiffusive



(a) Global Plasticity (GP). (b) Local Plasticity (LP). (c) Hybrid Plasticity (HP).

Figure 2: Time-resolved TA-MSD of the full weight vector (all effective weights concatenated) for the four-layer FC-SNN under (a) GP, (b) LP, and (c) HP. Each panel shows the displacement $\Delta r^2(\tau)$ as a function of lag time τ at successive training stages (single representative run); darker colours indicate earlier stages and lighter colours later stages.



(a) Global Plasticity (GP). (b) Local Plasticity (LP). (c) Hybrid Plasticity (HP).

Figure 3: Evolution of the local scaling exponent $\beta(t_w, \tau)$ as a function of lag time for the full weight vector of the four-layer FC-SNN under (a) GP, (b) LP, and (c) HP (single representative run). Darker colours indicate earlier training stages and lighter colours later stages. The instantaneous scaling regime of weight dynamics changes throughout training, reflecting non-stationary optimisation behaviour.

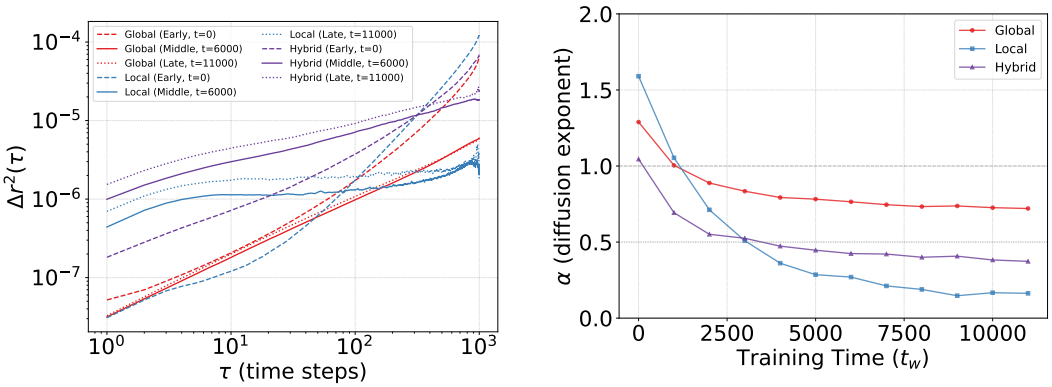
($\alpha < 1$) weight dynamics, reflecting a transition from extensive exploration to local convergence during optimisation.

Specifically, under Global Plasticity (GP) the diffusion exponent α decays most rapidly in the early stage, yet remains at a relatively high level in the late stage ($\alpha_{\text{late}} \approx 0.8$). Under Local Plasticity (LP), the diffusion exponents of the first two layers approach zero ($\alpha \approx 0.1$), effectively freezing the weights, whereas the final layer still exhibits significant diffusion under the modulation of local error signals. This reveals a hierarchical dynamical mechanism in which shallow feature extractors stabilise quickly while deep decision boundaries continue to fine-tune. For Hybrid Plasticity (HP), the decay of α is comparatively gradual, ultimately evolving to a lower level ($\alpha_{\text{late}} \approx 0.5$), reflecting its balanced characteristic of maintaining both exploration and convergence.

These regimes directly explain the convergence behaviors observed in Figure 1: all paradigms transit from early superdiffusive exploration to late subdiffusive convergence, yet with distinct layer-dependent rates. Figure 5 summarizes the extracted α values across training phases.

3.1.2 Loss Landscape Visualisation

We visualize the loss landscapes by projecting the effective weight trajectories onto 2D planes (Figure 6)(Li et al., 2018). Specifically, we employ a common projection plane defined by the average principal eigenvectors across all paradigms for direct comparison, alongside paradigm-specific PCA projections. Observations reveal that the HP paradigm exhibits the relatively flattest loss basin, populated with several shallow local minima; its training trajectory also shows noticeable roughness, reflecting local fluctuations during optimization. In contrast, the GP paradigm converges to the sharpest and relatively deep minimum along a smooth path, while the LP paradigm displays intermediate sharpness, with a basin narrower than HP but less sharp than GP.



(a) TA-MSD grows with lag time τ.

(b) Evolution of diffusion exponent α with waiting time t_w.

Figure 4: Cross-mode diffusion analysis for the four-layer FC-SNN. (a) TA-MSD grows with lag time τ. (b) Diffusion exponent α evolves with waiting time t_w. HP maintains intermediate diffusion behaviour between GP and LP.

	Early	Mid	Late
Global	1.2892	0.6975	0.7019
Local	1.5905	0.0363	0.0342
Hybrid	1.0453	0.3490	0.3245

Figure 5: Summary of diffusion exponents α across FC layers, learning modes, and training phases for the four-layer FC-SNN. Bars compare the late-stage scaling exponent attained by each paradigm in every layer.

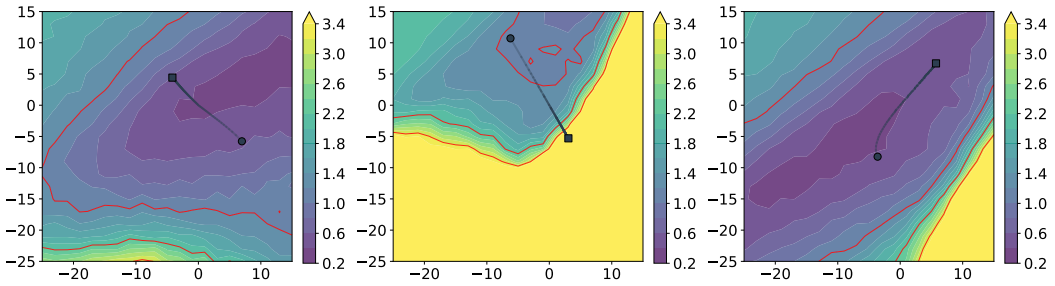
3.1.3 Surrogate-Based Curvature

To quantify the flatness of the loss landscape, we track the evolution of the largest Hessian eigenvalue throughout training (Figure 7). As training progresses, the largest eigenvalue for all three paradigms exhibits a monotonically decreasing trend, while consistently maintaining the ordering $\lambda_{\max}^{\mathrm{GP}} > \lambda_{\max}^{\mathrm{LP}} > \lambda_{\max}^{\mathrm{HP}}$. This indicates that Hybrid Plasticity (HP), through the implicit regularisation afforded by local plasticity, effectively suppresses convergence toward high-curvature regions and thereby guides the optimisation trajectory toward flatter minima with better generalisation performance.

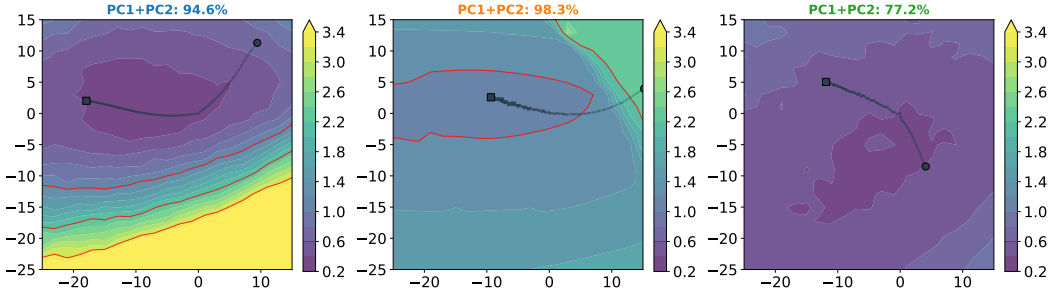
3.2 Summary of Supplementary Analyses

Four supplementary studies in the appendix complement and qualify these findings; we summarise them briefly.

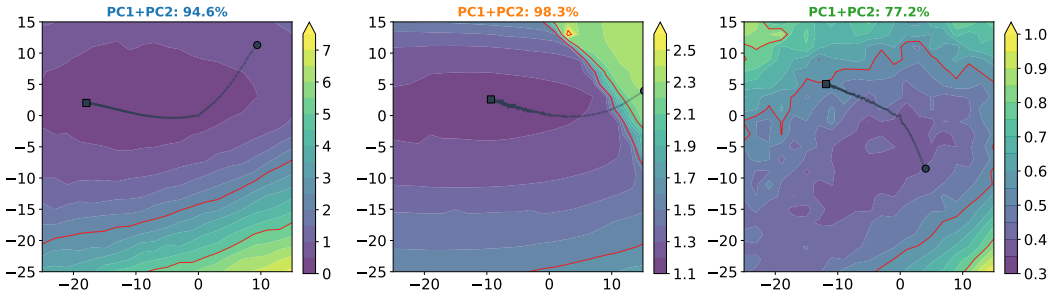
Theoretical account. Appendix B models SGD as overdamped Langevin dynamics and decomposes the MSD into an intra-basin (normal diffusion) and an inter-basin component, the latter arising from power-law curvature evolution and saddle-point escape; the resulting geometry–dynamics–generalization correspondence, with its assumptions, is given in Appendix C.



(a) Common PCA Projection



(b) Individual PCA Projection



(c) Individual PCA Projection(Details)

Figure 6: 2D loss landscapes and training trajectories of FC-SNNs under three learning paradigms. Darker colors denote lower loss. Overlaid curves show the combined effective weight trajectories at epoch 20, from initialization (circle) to the trained solution (square). (a) Projected onto the average principal eigenvectors of all paradigms for direct comparison. (b) Projected onto the top two principal components of each individual paradigm; cumulative variance ratios are shown above each panel.(c) Same as (b), but with a different color scale range.

EventProp comparison. Appendix C.2 shows that exact-gradient EventProp sustains stronger early superdiffusion ($\alpha \approx 1.33$–1.50 vs. $\alpha_{\mathrm{GP}} \approx 1.05$), indicating that surrogate-gradient smoothing alters the noise covariance governing diffusion.

Hard weight-norm constraints. Appendix D imposes the regularisation explicitly by projecting weights onto fixed-norm spheres: HP prolongs superdiffusion and sharpens the spectrum, while GP freezes without curvature growth, so the unconstrained flat-minima–subdiffusion coupling does not transfer directly to constrained weights.

Three-layer ablation. Appendix E confirms on a shallower FC-SNN that the superdiffusive-to-subdiffusive crossover and HP's flat-minimum convergence persist, with GP additionally showing progressive sharpening and late-stage diffusion re-escalation.

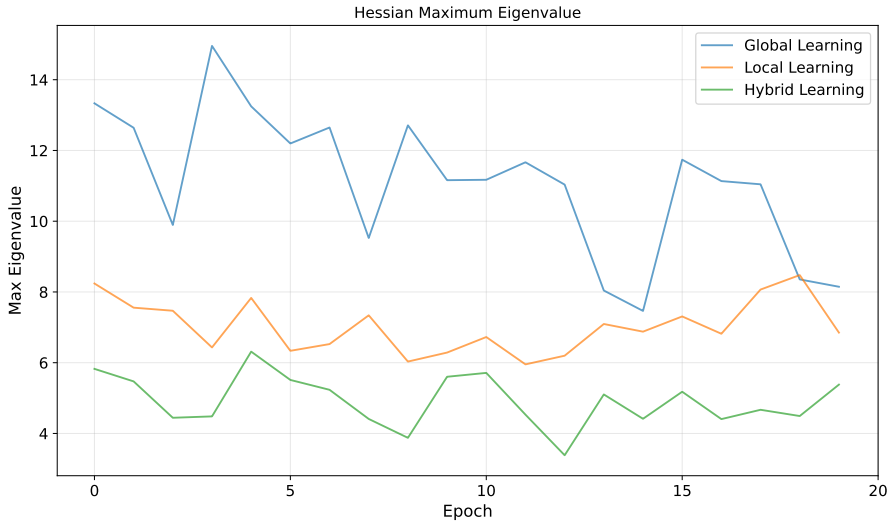


Figure 7: Evolution of the largest Hessian eigenvalue $\lambda_{\max}$ with training epoch for the four-layer FC-SNN. The spectral trajectory reveals how loss landscape curvature decreases as optimisation proceeds; HP consistently attains the lowest curvature.

4 Conclusion

In this work, we presented a comprehensive geometric-dynamical analysis of learning in Spiking Neural Networks (SNNs), systematically comparing Global (GP), Local (LP), and Hybrid (HP) plasticity paradigms. Our study proceeds by first characterizing the anomalous diffusion dynamics of weight trajectories, followed by an examination of the underlying loss landscape geometry, thereby linking stochastic optimization behavior to generalization performance.

From a dynamical perspective, we identified a universal two-phase diffusion pattern across all paradigms: an early super-diffusive phase driven by extensive exploration, transitioning to a late sub-diffusive phase as the network converges. Crucially, HP exhibits a distinct intermediate diffusion scaling exponent, balancing the irregular, persistent exploration of GP with the premature stagnation observed in LP. This balanced dynamical regime prevents both excessive noise accumulation and early freezing of weights. Correspondingly, our geometric analysis reveals that HP converges to significantly flatter minima compared to GP and LP. We attribute this to an implicit regularization effect, where local Hebbian updates modulate the effective noise covariance, guiding the optimization trajectory toward broad, stable basins without explicit curvature penalization. This synergy between controlled diffusion and flat landscape convergence explains HP's superior robustness to input corruptions.

Our framework provides a mechanistic understanding of how hybrid learning rules shape the solution space in SNNs. Future work will extend this analysis in several directions. First, we plan to evaluate exact gradient methods such as EventProp and eligibility propagation (e-prop) to determine how precise temporal credit assignment influences diffusion regimes and landscape curvature. Second, we aim to generalize our findings to diverse network architectures, including convolutional and recurrent SNNs, to assess the scalability of these geometric-dynamical principles. By bridging statistical mechanics with neuromorphic learning, this work lays the foundation for designing more robust and efficient training algorithms for spiking networks.

Acknowledgements

The code used for the experiments and analyses in this work was generated with the assistance of AI coding tools (large language models). All experimental designs, results, and interpretations were reviewed and verified by the authors.

References

Gérard Ben Arous and Alice Guionnet. The spectrum of heavy tailed random matrices. *Communications in Mathematical Physics*, 278(3):715–751, 2008.

Simon Becker, Yao Zhang, and Alpha A Lee. Geometry of energy landscapes and the optimizability of deep neural networks. *Physical review letters*, 124(10):108301, 2020.

Guillaume Bellec, Franz Scherr, Anand Subramoney, Elias Hajek, Darjan Salaj, Robert Legenstein, and Wolfgang Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, 11(1), July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17236-y.

Guo-qiang Bi and Mu-ming Poo. Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of Neuroscience*, 18(24):10464–10472, December 1998. ISSN 1529-2401. doi: 10.1523/jneurosci.18-24-10464.1998.

Guozhang Chen, Cheng Kevin Qu, and Pulin Gong. Anomalous diffusion dynamics of learning in deep neural networks. *Neural Networks*, 149:18–28, May 2022. ISSN 0893-6080. doi: 10.1016/j.neunet.2022.01.019.

Huanran Chen, Yinpeng Dong, Zeming Wei, Yao Huang, Yichi Zhang, Hang Su, and Jun Zhu. Unveiling the basin-like loss landscape in large language models, 2025.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Jeremy M. Cohen and Vladlen Koltun. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2021.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014.

Gregory Dexter, Borja Ocejo, Sathiya Keerthi, Aman Gupta, Ayan Acharya, and Rajiv Khanna. A precise characterization of SGD stability using loss surface geometry. In *International Conference on Learning Representations (ICLR)*, 2024.

Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex sgd escaping from saddle points. In *Conference on Learning Theory*, pp. 1192–1234. PMLR, 2019.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Mert Gurbuzbalaban, Asuman Ozdaglar, Umut Simsekli, and Lingjiao Yuan. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning (ICML)*, pp. 3869–3879, 2021.

Julia Gygax and Friedemann Zenke. Elucidating the theoretical underpinnings of surrogate gradient learning in spiking neural networks. *Neural Computation*, 37(5):886–925, 2025.

Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after kramers. *Reviews of modern physics*, 62(2):251, 1990.

Daniel Jiwoong Im, Michael Tao, and Kristin Branson. An empirical analysis of the optimization of deep network loss surfaces, 2016.

G. Indiveri and Shih-Chii Liu. Memory and information processing in neuromorphic systems. *Proceedings of the IEEE*, 103:1379–1397, 2015. URL `https://api.semanticscholar.org/CorpusID:13984215`.

Advait Kalra, Yuhang He, and Mengming Barkeshli. Universal sharpness dynamics in neural network training: Fixed point analysis, edge of stability, and route to chaos. In *International Conference on Learning Representations (ICLR)*, 2025.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2016.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Lei Li and Yuliang Wang. On uniform-in-time diffusion approximation for stochastic gradient descent, 2022.

Andrew Ly and Pulin Gong. Optimization on multifractal loss landscapes explains a diverse range of geometrical and dynamical properties of deep learning. *Nature Communications*, 16(1), April 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-58532-9.

Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, December 1997. ISSN 0893-6080. doi: 10.1016/s0893-6080(97)00011-7.

Michael W. Mahoney and Charles Martin. Traditional and heavy-tailed self regularization in neural network models. In *International Conference on Machine Learning (ICML)*, 2019.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Ralf Metzler and Joseph Klafter. The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports*, 339(1):1–77, 2000. doi: 10.1016/S0370-1573(00)00070-3.

Emre O Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

Vardan Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size. *arXiv preprint arXiv:1811.07062*, 2018.

Levent Sagun, Utku Evci, V. Ugur Güney, Yann N. Dauphin, and Léon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=rJO1_M0Lf`.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning (ICML)*, pp. 5827–5836, 2019.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning (ICML)*, pp. 681–688, 2011.

Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12:331, 2018.

Yujie Wu, Rong Zhao, Jun Zhu, Feng Chen, Mingkun Xu, Guoqi Li, Sen Song, Lei Deng, Guanrui Wang, Hao Zheng, Songchen Ma, Jing Pei, Youhui Zhang, Mingguo Zhao, and Luping Shi. Brain-inspired global-local learning incorporated with neuromorphic computing. *Nature Communications*, 13(1), January 2022. ISSN 2041-1723. doi: 10.1038/s41467-021-27653-2.

Timo C Wunderlich and Christian Pehle. Event-based backpropagation can compute exact gradients for spiking neural networks. *Scientific Reports*, 11(1):12829, 2021.

Jianyuan Yin, Bing Yu, and Lei Zhang. Searching the solution landscape by generalized high-index saddle dynamics. *Science China Mathematics*, 64(8):1801–1816, 2021.

Friedemann Zenke and Surya Ganguli. Superspike: Supervised learning in multilayer spiking neural networks. *Neural computation*, 30(6):1514–1541, 2018.

Zeyuan Zhu, Yu Gong, Yang Liu, et al. The anisotropic noise in SGD: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning (ICML)*, 2019.

A Supplementary Details

A.0.1 GP Hessian

Under GP, the Hessian is computed exactly via second-order automatic differentiation: the gradient of the cross-entropy loss with respect to the static synaptic weights is differentiated again via backpropagation. This method leverages the chain rule on the computation graph and does not require explicitly constructing the gradient Jacobian.

A.0.2 LP Hessian

Under LP, we consider the Hebbian loss function:

$$L_{\mathrm{LP}} = -\sum_{t=1}^{T}\sum_{l=2}^{n_l}\left(\mathbf{s}^{l-1}(t)\right)^{\top}\mathbf{w}_{\mathrm{LP}}^{l}(t)\ \tanh\!\left(\frac{\mathbf{u}^{l}(t)}{\theta}-\boldsymbol{\eta}\right), \tag{8}$$

where θ is a scaling parameter and $\boldsymbol{\eta}$ is a threshold offset vector.

A.0.3 HP Hessian

Under HP, the total weight decomposes into two independent parts (Eq. (1)), which linearly superpose during the forward pass. The Hessian strictly splits into two additive terms:

$$\mathbf{H}_{\mathrm{HP}} = \nabla^2_{\mathbf{w}_{\mathrm{LP}}} L_{\mathrm{LP}} + \nabla^2_{\mathbf{w}_{\mathrm{GP}}} L_{\mathrm{GP}}. \tag{9}$$

The largest eigenvalue $\lambda_{\max}$ is computed via power iteration.

A.1 Spiking Neuron Dynamics

A.1.1 Continuous-Time LIF Model

Following Wu et al.(Wu et al., 2022), the membrane potential evolution of neuron i in layer l is described by the continuous-time differential equation:

$$\tau_u \frac{du_i^l(t)}{dt} = -u_i^l(t) + \sum_{j=1}^{n_{l-1}} w_{ij}^l(t)\, s_j^{l-1}(t) - v_{\mathrm{th}} \sum_k \delta(t - t_k^i), \tag{10}$$

where t is the continuous time variable, l is the layer index, i is the neuron index in layer l, $u_i^l(t)$ is the membrane potential of neuron i in layer l at time t, τ_u is the membrane time constant, n_{l-1} is the total number of neurons in layer $l-1$, j is the neuron index in layer $l-1$, $w_{ij}^l(t)$ is the synaptic weight from neuron j in layer $l-1$ to neuron i in layer l, $s_j^{l-1}(t)$ is the spike input activity of neuron j in layer $l-1$ at time t, v_{th} is the firing threshold, k is the spike index, t_k^i is the exact firing time of the k-th spike of neuron i, and $\delta(\cdot)$ is the Dirac delta function.

The last term $-v_{\mathrm{th}} \sum_k \delta(t - t_k^i)$ represents the hard reset mechanism, i.e., an instantaneous drop of magnitude $-v_{\mathrm{th}}$ upon spike emission at firing times t_k^i.

Additionally, the spike output of the neuron is defined as:

$$s_i^l(t) = \Theta\big(u_i^l(t) - v_{\mathrm{th}}\big), \tag{11}$$

where $s_i^l(t)$ is the spike output state of neuron i in layer l at time t, and $\Theta(\cdot)$ is the Heaviside step function. When the membrane potential $u_i^l(t)$ reaches or exceeds the threshold v_{th}, the function outputs 1 (indicating the neuron fires a spike); otherwise it outputs 0.

A.1.2 Discrete-Time Update

Discretising Eq. (10) via forward Euler and absorbing the synaptic current time constant into the effective membrane decay coefficient yields the compact discrete-time update formula used in experiments:

$$u_i^l[t] = \left(u_i^l[t-1] - s_i^l[t-1] \cdot v_{\text{th}}\right)\gamma + I_i^l[t], \tag{12}$$

where $\gamma = \exp(-\Delta t/\tau_u)$ is the membrane decay coefficient (in our experiments $\gamma = 0.4$), and $I_i^l[t] = \sum_j w_{ij}^l[t]\, s_j^{l-1}[t]$ is the effective input current at the current timestep

B Discrete-time SGD and Gradient-noise Decomposition

Given that full-batch gradient descent is computationally expensive, SGD estimates the weight parameters $\mathbf{w} = (w_1, w_2, \ldots, w_d)$ by minimising the minibatch loss. At step t it applies

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\, \widetilde{\nabla} L_t(\mathbf{w}_t), \tag{13}$$

where $\widetilde{\nabla} L_t$ is the minibatch gradient and η is the learning rate. In a feed-forward network with $L+1$ layers (including input and output), the minibatch gradient is evaluated via back-propagation with complexity $O(n \prod_{i=1}^{L} N_{i-1} N_i)$, where n is the number of SGD iterations per epoch and N_i is the dimension of layer i.

The partial absence of the dataset generates gradient noise. Writing the minibatch gradient as the full-batch gradient plus noise,

$$\widetilde{\nabla} L_t(\mathbf{w}_t) = \nabla L(\mathbf{w}_t) + \boldsymbol{\xi}_t, \tag{14}$$

the update can be rewritten as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla L(\mathbf{w}_t) + \mathbf{U}_t, \qquad \mathbf{U}_t \triangleq -\eta \boldsymbol{\xi}_t, \tag{15}$$

which is the Euler-Maruyama discretisation of the overdamped Langevin SDE in Eq. (16). The noise covariance $\mathbf{D}(t)$ in Eq. (16) absorbs the learning rate and the variance of the minibatch noise.

To see why this approximation is valid, note that the per-step noise $\mathbf{U}_t = -\eta \boldsymbol{\xi}_t$ has mean zero. Because the minibatch gradient is an average of B sample gradients, the fluctuation $\boldsymbol{\xi}_t = \widetilde{\nabla} L_t - \nabla L$ has covariance $O(1/B)$. Hence $\mathbf{U}_t$ has covariance $O(\eta^2/B)$. In the continuous-time limit (small η, many steps), the discrete trajectory is viewed as evolving in continuous time. If one SGD step corresponds to an infinitesimal time interval $dt = \eta$, the deterministic drift contributes $-\nabla L(\mathbf{w})\, dt$, while the accumulated independent minibatch fluctuations converge, by the Central Limit Theorem, to a Gaussian process with covariance proportional to dt. Matching this covariance to $\mathbf{D}(t)\, dt$ gives $\mathbf{D}(t) = (\eta^2/B)\, \mathbf{C}(\mathbf{w}(t))$, where $\mathbf{C}(\mathbf{w}(t))$ is the sample-gradient covariance evaluated along the trajectory. In the main text we work with the eigenbasis projections $\sigma_i^2(t) = \mathbf{v}_i^\top \mathbf{D}(t)\, \mathbf{v}_i$ as effective, aggregated noise-strength time series.

B.1 Understanding Learning via Anomalous Diffusion

Mini-batch SGD can be modelled as overdamped Langevin dynamics on the loss landscape $L(\mathbf{w})$ (Metzler & Klafter, 2000), and the stochastic update reads

$$d\mathbf{w} = -\nabla L(\mathbf{w})\, dt + \mathbf{D}(t)^{1/2}\, d\mathbf{W}_t, \tag{16}$$

where $\mathbf{W}_t$ is a standard Wiener process and $\mathbf{D}(t)$ is the noise covariance matrix, with elements $D_{\mu\nu}(t)$ encoding the aggregate noise from mini-batch sampling, surrogate-gradient approximation, and event-driven spike perturbations. The weight vector $\mathbf{w}$ thus performs a random walk on $L(\mathbf{w})$, with flat regions producing slow diffusion and steep regions fast diffusion.

B.1.1 Eigenspace Decomposition and Two-component Msd

Projecting the Langevin dynamics (Eq. (16)) onto the Hessian eigen-directions $\mathbf{v}_i$, the displacement $q_i = \mathbf{v}_i^\top (\mathbf{w} - \mathbf{w}^*)$ along each eigen-direction decouples as:

$$dq_i = -\lambda_i(t)\, q_i\, dt + \sigma_i(t)\, dW_t^{(i)}, \tag{17}$$

where $\lambda_i(t)$ is the i-th Hessian eigenvalue and $\sigma_i^2(t) = \mathbf{v}_i^\top \mathbf{D}(t)\,\mathbf{v}_i$ is the projected noise strength. This decomposition maps the high-dimensional diffusion process onto a set of independent one-dimensional stochastic equations, each governed by the corresponding local curvature $\lambda_i(t)$. Retaining only the maximum-curvature direction $\lambda(t) = \lambda_{\max}^{\mathrm{HP}}(t)$ and solving gives the intra-basin theoretical MSD:

$$\Delta r^2_{\mathrm{intra}}(t) = \int_0^t \sigma(s)^2 \exp\Big(-2\int_s^t \lambda(u)\,du\Big)\,ds, \tag{18}$$

In practice, $\lambda(t)$ is obtained from the maximum Hessian eigenvalue trajectory, and $\sigma(t)$ is estimated per 1000-step window via $\sigma^2_{\mathrm{windows}} = \Delta r^2(\tau = 1)$.

Crucially, as long as $\lambda(t) > 0$, Eq. (18) always predicts *normal* diffusion ($\Delta r^2_{\mathrm{intra}} \sim \tau$ at short lag times, saturating to a plateau), regardless of how $\lambda(t)$ varies. The anomalous diffusion ($\alpha \neq 1$) observed empirically (Eq. (6)) therefore cannot arise from intra-basin dynamics alone. We attribute it to inter-basin transitions—rare but large jumps across saddle points or barrier regions in the non-quadratic loss landscape, analogous to Kramers escape events in complex energy landscapes (Hänggi et al., 1990). The total MSD thus decomposes as:

$$\Delta r^2_{\mathrm{total}}(\tau) = \underbrace{\Delta r^2_{\mathrm{intra}}(\tau)}_{\text{intra-basin (normal)}} + \underbrace{\Delta r^2_{\mathrm{inter}}(\tau)}_{\text{inter-basin (anomalous)}}, \tag{19}$$

where $\mathrm{MSD}_{\mathrm{inter}}$ captures the excess displacement from basin-to-basin hopping. This anomalous component arises from two distinct mechanisms, which we derive separately.

B.1.2 Decomposition of Inter-basin Contributions

The intra-basin MSD (Eq. (18)) always yields normal diffusion ($\alpha = 1$) as long as the local curvature remains positive. The anomalous behaviour ($\alpha \neq 1$) observed empirically—both superdiffusion ($\alpha > 1$) and subdiffusion ($\alpha < 1$)—must therefore originate from non-stationary effects that violate the quadratic-basin, positive-curvature assumptions—a situation familiar from the inherent-structure description of complex energy landscapes: (i) temporal evolution of the curvature itself, and (ii) the presence of negative-curvature directions. We derive each contribution separately.

Mechanism 1: Curvature evolution (power-law anomalous diffusion) During optimisation the loss landscape evolves non-stationarily: $\lambda(t)$ changes as the optimiser traverses the landscape. Modelling this as a power law $\lambda(t) = \lambda_0(t/t_0)^{-\gamma}$ with $\gamma \in \mathbb{R}$, the effective diffusion coefficient along the maximum-curvature direction scales as:

$$D_{\mathrm{eff}}(t) = \frac{\sigma^2}{2\lambda(t)} \sim t^{\gamma}. \tag{20}$$

When $\gamma > 0$ (curvature decay / landscape flattening), the confining force weakens and the "walls" of the basin recede, allowing the trajectory to explore an ever-expanding region—producing superdiffusion. When $\gamma < 0$ (curvature increase / landscape sharpening), the basin walls tighten over time, progressively confining the trajectory more strongly—producing subdiffusion. The TAMSD at waiting time t_w with lag time τ is approximately:

$$\Delta r^2_{\mathrm{decay}}(\tau; t_w) \approx \int_{t_w}^{t_w+\tau} \frac{\sigma^2}{\lambda(s)}\left[1 - \exp\Big(-2\int_s^{t_w+\tau}\lambda(u)\,du\Big)\right] ds. \tag{21}$$

In the regime where τ is large enough that the exponential term saturates (the intra-basin relaxation has completed), the integrand simplifies to $\sigma^2/\lambda(s) \sim s^{\gamma}$, and the integral yields:

$$\Delta r^2_{\mathrm{decay}} \sim \int_{t_w}^{t_w+\tau} s^{\gamma}\,ds \sim \tau^{1+\gamma}. \tag{22}$$

Defining $\alpha_{\mathrm{decay}} \equiv 1 + \gamma$, we obtain the power-law anomalous contribution:

$$\Delta r^2_{\mathrm{decay}}(\tau) \approx A\,\tau^{\alpha_{\mathrm{decay}}}. \tag{23}$$

The sign of γ determines the diffusion regime: $\gamma > 0$ ($\alpha_{\mathrm{decay}} > 1$, superdiffusion) corresponds to curvature decay during landscape flattening, while $\gamma < 0$ ($\alpha_{\mathrm{decay}} < 1$, subdiffusion) corresponds to curvature increase as the landscape sharpens around a narrowing basin.

Mechanism 2: saddle-point escape (exponential superdiffusion) At saddle points or in highly non-convex regions, some Hessian eigenvalues become negative ($\lambda_{\text{neg}} < 0$). Along such directions, the curvature-induced restoring force reverses sign and becomes a repelling force that pushes the weight away from the critical point. The eigen-mode dynamics become:

$$dq_i = +|\lambda_{\text{neg}}|\, q_i\, dt + \sigma_i(t)\, dW_t^{(i)} \tag{24}$$

The positive feedback term $+|\lambda_{\text{neg}}|\, q$ amplifies any displacement exponentially: once the trajectory is perturbed away from the saddle, the deterministic drift accelerates it further. Solving gives the approximate MSD along the unstable direction:

$$\langle q_i^2(\tau)\rangle \approx \frac{\sigma_i^2}{|\lambda_{\text{neg}}|}\left(e^{2|\lambda_{\text{neg}}|\tau} - 1\right). \tag{25}$$

Unlike the stable eigen-modes (whose Δr^2 saturates to a plateau σ^2/λ), the unstable mode exhibits exponential growth—the hallmark of deterministic escape from a saddle point accelerated by noise. Summing over N_{neg} unstable directions with similar $|\lambda_{\text{neg}}|$, we obtain:

$$\Delta r^2_{\text{escape}}(\tau) \approx B\left(e^{2|\lambda_{\text{neg}}|\tau} - 1\right), \quad B \approx \frac{N_{\text{neg}}\,\sigma^2}{|\lambda_{\text{neg}}|}. \tag{26}$$

Combined decomposition Combining Eqs. (23) and (26), the total inter-basin MSD is approximately:

$$\Delta r^2_{\text{inter}}(\tau) \approx \underbrace{A\,\tau^{1+\gamma}}_{\text{curvature evolution}} + \underbrace{B\left(e^{2|\lambda_{\text{neg}}|\tau} - 1\right)}_{\text{saddle-point escape}}, \tag{27}$$

where $\gamma \in \mathbb{R}$ is the curvature evolution rate defined above. The two mechanisms produce distinct signatures in the effective diffusion coefficient $D_{\text{eff}}(\tau) = \Delta r^2_{\text{inter}}(\tau)/(2\tau)$: power-law scaling $D_{\text{eff}} \sim \tau^{\gamma}$ for curvature evolution (superdiffusive when $\gamma > 0$, subdiffusive when $\gamma < 0$), and exponential growth $D_{\text{eff}} \sim e^{2|\lambda_{\text{neg}}|\tau}/\tau$ for saddle-point escape. The intra-basin component $\Delta r^2_{\text{intra}}$ serves as a baseline: by subtracting it from the empirical TAMSD, we isolate $\Delta r^2_{\text{inter}}$ and thereby quantify the degree of non-quadratic exploration. The sign of γ naturally accounts for the observed transition from early-phase superdiffusion ($\gamma > 0$, landscape flattening) to late-phase subdiffusion ($\gamma < 0$, basin sharpening) during training.

C Derivation of the Intra-Basin MSD

Consider the eigen-mode dynamics of Eq. (17) along a direction with positive Hessian eigenvalue $\lambda(t) > 0$,

$$dq(t) = -\lambda(t)q(t)\, dt + \sigma(t)\, dW_t, \tag{28}$$

with $q(0) = 0$. Multiplying by the integrating factor $I(t) = \exp\big(\int_0^t \lambda(u)\, du\big)$ gives

$$d\big(I(t)q(t)\big) = I(t)\sigma(t)\, dW_t. \tag{29}$$

Integrating from 0 to t yields

$$q(t) = \frac{1}{I(t)}\int_0^t I(s)\sigma(s)\, dW_s = \int_0^t \sigma(s)\exp\Big(-\int_s^t \lambda(u)\, du\Big)\, dW_s. \tag{30}$$

Applying the Itô isometry,

$$\langle q^2(t)\rangle = \int_0^t \sigma(s)^2 \exp\Big(-2\int_s^t \lambda(u)\, du\Big)\, ds, \tag{31}$$

which is Eq. (18). Discretising with unit time step ($\eta = 1$) and writing $\Lambda(t) = \sum_{u=0}^{t-1}\lambda(u)$ gives the equivalent discrete sum used in the implementation:

$$\Delta r^2(\tau) = \sigma^2\, e^{-2\Lambda(\tau)} \sum_{s=0}^{\tau-1} e^{2\Lambda(s)}. \tag{32}$$

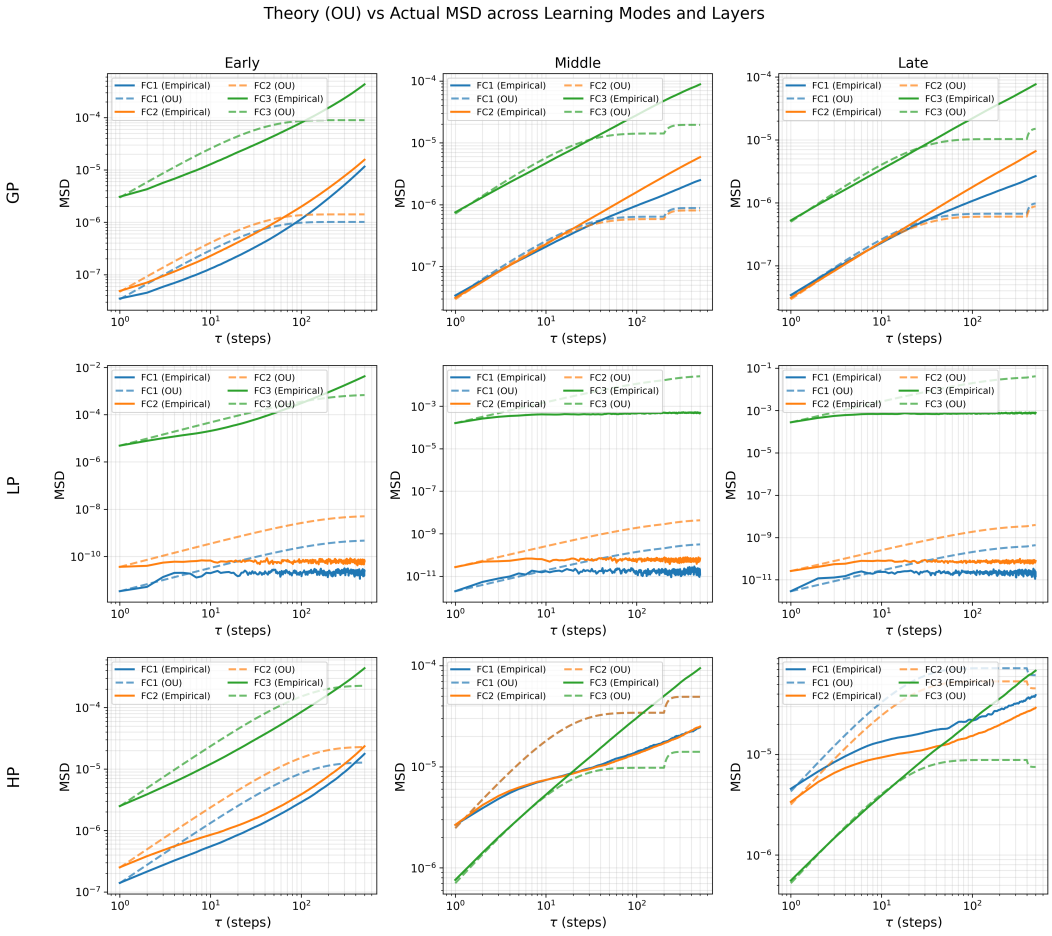


Figure 8: TA-MSD across learning paradigms (rows: GP, LP, HP) and training phases (columns: Early, Middle, Late). Solid lines denote actual TA-MSD; dashed lines denote the intra-basin prediction $\mathrm{MSD}_{\mathrm{intra}}$ (Eq. (18)) obtained from the eigenspace decomposition with time-dependent $\lambda(t) = \lambda_{\max}^{\mathrm{HP}}(t)$ and time-varying $\sigma(t)$. The intra-basin model captures short-lag behaviour but systematically underestimates long-lag diffusion, especially under GP and HP, indicating the presence of inter-basin transitions not accounted for by the quadratic-basin approximation (see Eq. (19) and discussion below).

Unstable eigen-mode (negative curvature)

For a negative eigenvalue $\lambda_{\mathrm{neg}} < 0$, Eq. (28) becomes

$$dq(t) = +|\lambda_{\mathrm{neg}}|\, q(t)\, dt + \sigma\, dW_t. \tag{33}$$

The integrating factor is now $I(t) = \exp(-|\lambda_{\mathrm{neg}}|t)$, yielding

$$q(t) = \int_0^t \sigma\, \exp\big(|\lambda_{\mathrm{neg}}|(s-t)\big)\, dW_s. \tag{34}$$

Applying the Itô isometry,

$$\langle q^2(t)\rangle = \sigma^2 \int_0^t \exp\big(2|\lambda_{\mathrm{neg}}|(s-t)\big)\, ds = \frac{\sigma^2}{|\lambda_{\mathrm{neg}}|}\left(e^{2|\lambda_{\mathrm{neg}}|t} - 1\right), \tag{35}$$

which is Eq. (25). Unlike the stable case ($\lambda > 0$), where $\langle q^2\rangle$ saturates to σ^2/λ, the unstable case exhibits exponential growth, reflecting escape from saddle points.

C.0.1 Geometry-Dynamics-Generalization Correspondence

By jointly analysing Hessian curvature, diffusion exponents, and the MSD decomposition of Eq. (27), we establish a quantitative link between landscape geometry and weight dynamics. Flat minima consistently coincide with subdiffusive weight dynamics, whereas sharp minima induce erratic, superdiffusive exploration.

The two-component decomposition provides a mechanistic interpretation of this coupling across training stages. In the *early* stage, the exponential (escape) term is expected to dominate at large τ: the highly non-convex landscape with abundant saddle points drives superdiffusion via unstable escape, and the GP paradigm, with its more erratic exploration, should exhibit the strongest such contribution. In the *late* stage, the exponential contribution is expected to diminish as the landscape becomes more convex (fewer negative eigenvalues), leaving the power-law (curvature decay) term as the dominant source of residual superdiffusion. Under HP, both contributions are expected to be balanced: the escape term should be suppressed (HP converges to flatter, more convex minima), while the decay term remains moderate, enabling sustained but controlled exploration.

This picture is consistent with the evolution of the local scaling exponent $\beta(t_w, \tau)$ (Fig. 3): the upward curvature of β at large τ in the early stage is compatible with an exponential (escape) contribution, while the gradual increase at small τ is compatible with the power-law (curvature decay) contribution. The transition from early superdiffusion to late subdiffusion corresponds to the suppression of the exponential term as the landscape convexifies.

Ultimately, this geometric-dynamical coupling explains why HP outperforms GP and LP. Local Hebbian updates reshape the effective potential, smoothing sharp ridges and inducing controlled subdiffusion that enhances basin stability without sacrificing convergence speed.

C.1 Key Assumptions and Limitations

Our theoretical framework rests on several assumptions whose validity should be carefully considered:

1. **Instantaneous eigen-direction approximation.** Eq. (17) treats the Hessian eigendirections $\mathbf{v}_i(t)$ as approximately constant over the timescale of each analysis window. This adiabatic approximation holds when the landscape geometry evolves slowly relative to the intra-basin relaxation time, which is well satisfied during the subdiffusive late-training regime but may break down during rapid basin-to-basin transitions.

2. **Local quadratic approximation.** The intra-basin MSD (Eq. (18)) assumes the loss surface is locally quadratic near the minimum. This is valid in the vicinity of well-defined minima but fails near saddle points or in highly anharmonic regions—precisely where the inter-basin mechanisms (curvature decay and saddle-point escape) become necessary.

3. **Intra/inter decomposition.** Eq. (19) assumes the cross-term between intra-basin and inter-basin displacements is negligible. This holds when intra-basin fluctuations are symmetric about the local minimum, which is satisfied in quasi-equilibrium but may introduce errors during active transitions.

4. **Power-law curvature decay.** The model $\lambda(t) = \lambda_0(t/t_0)^{-\beta}$ is a phenomenological assumption. While our empirical Hessian tracking (Fig. 7) confirms a monotonic decrease in $\lambda_{\max}$, the precise functional form may vary across architectures and tasks. The framework is flexible enough to accommodate alternative decay models by substituting the appropriate $\lambda(t)$ into Eq. (18).

5. **Gaussian noise approximation.** The Langevin model (Eq. (16)) assumes Gaussian gradient noise. Recent work (Simsekli et al., 2019) has shown that SGD noise can exhibit heavy tails. Our framework can accommodate non-Gaussian noise through the empirical noise strength $\sigma(t)$, but the closed-form expressions derived here assume Gaussianity. Extending the theory to Lévy-driven dynamics is an important direction for future work.

C.2 Comparison with Exact-Gradient EventProp

All experiments reported above employ surrogate gradients to backpropagate errors through the non-differentiable spike function. To quantify the approximation bias introduced by this smoothing, we compare against EventProp, an exact gradient-descent framework for spiking neural networks proposed by Wunderlich and Pehle(Wunderlich & Pehle, 2021). EventProp leverages an adjoint formulation to compute exact gradients with respect to spike times, avoiding the need for a smoothed surrogate $\sigma'(\cdot;\kappa)$. This provides a principled baseline for assessing how surrogate-gradient smoothing alters the optimisation landscape and weight diffusion.

We train a three-layer FC-SNN (784→512→10) on Fashion-MNIST using the EventProp reference implementation with the same hyperparameters as our GP paradigm: batch size 100, Adam optimiser with learning rate $\eta = 0.01$, and 20 epochs. Figure 9 shows the resulting training dynamics. EventProp reaches a moving-average test accuracy of approximately 90% and stabilises at a cross-entropy loss near 0.9. The optimisation trajectory exhibits rapid initial convergence followed by persistent high-variance minibatch oscillations—a consequence of exact gradients being more sensitive to precise spike-timing fluctuations than the smoothed surrogate used in GP.

To compare diffusion regimes, we compute the time-averaged mean squared displacement for the exact-gradient trajectory during the early training phase (steps 0–1000). Figure 10 displays the MSD curves for the input-to-hidden ($\mathbf{W}_1$) and hidden-to-output ($\mathbf{W}_2$) weight matrices. The fitted diffusion exponents are $\alpha_{\mathbf{W}_1} = 1.33$ and $\alpha_{\mathbf{W}_2} = 1.50$, both indicating superdiffusive exploration. Both values exceed the GP value ($\alpha_{\mathrm{GP}} \approx 1.05$) reported in Section 2.3, suggesting that exact gradients sustain stronger directional drift during early training than surrogate-gradient methods, which suppress drift by averaging over threshold-crossing variations.

The local diffusion exponent $\beta(t_w, \tau)$, shown in Figure 11, reveals additional structure. For $\mathbf{W}_2$, $\beta(t_w, \tau)$ drops sharply below normal diffusion within the first 20 lags, then oscillates before settling around 1.6. The combined weight vector $\mathbf{W}_1 \oplus \mathbf{W}_2$ converges more smoothly toward $\beta \approx 1.2$. These non-stationary dynamics contrast with the relatively stable superdiffusive regime of GP and underscore that exact-gradient optimisation explores weight space more aggressively in early phases. The implication is that surrogate gradients not only introduce approximation bias in gradient magnitude, but also fundamentally alter the stochastic noise covariance $\mathbf{D}(\mathbf{w})$ that governs Langevin diffusion. Future work will extend this comparison to recurrent architectures and investigate whether the larger early-phase diffusion of EventProp correlates with different generalisation properties.

D Hard Weight-Norm Constraints: Diffusion and Landscape Analysis

In the main text, hybrid plasticity is regularised implicitly through the local Hebbian branch. Here we examine a variant in which the constraint is imposed *directly on the weights*: after each update, the weight matrix of every layer is projected onto a sphere of fixed norm, $\|\mathbf{W}_1\| = 13.8$, $\|\mathbf{W}_2\| = 10.2$, $\|\mathbf{W}_3\| = 2.5$. We train the three-layer FC-SNN (784→512→256→10) under this hard constraint with both the GP and the HP objective (Eqs. (36) and (38)), using SGD (learning rate 10^{-2}, momentum 0.9, batch size 100, seed 111) for 30 epochs on FashionMNIST, and compare the resulting diffusion dynamics and landscape geometry. All analyses use the same TA-MSD and Hessian machinery as the main text, applied to the *combined* weight vector $\mathbf{w} = (\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3)$ of dimension $d = 535{,}552$.

D.1 Hard Weight-Norm Constrained Loss

Let $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3\}$ denote the weights of the three FC layers ($784 \to 512 \to 256 \to 10$). The hard-constraint experiment replaces the implicit Hebbian regularisation of the main text with an explicit constraint on the weight norms, formulated as the constrained optimisation

$$\min_{\mathbf{W}} \; L_{\mathrm{CE}}(\mathbf{W}) \quad \text{subject to} \quad \|\mathbf{W}_l\|_F = c_l, \;\; l = 1, 2, 3, \tag{36}$$

where L_{CE} is the cross-entropy loss, $\|\cdot_F$ is the Frobenius norm, and the target radii are $(c_1, c_2, c_3) = (13.8, 10.2, 2.5)$. Problem (36) is solved by projected SGD: after each gradient step,

$$\widetilde{\mathbf{W}}_l^{(t+1)} = \mathbf{W}_l^{(t)} - \eta \, \nabla_{\mathbf{W}_l} L_{\mathrm{CE}}\big(\mathbf{W}^{(t)}\big), \qquad \mathbf{W}_l^{(t+1)} = c_l \, \frac{\widetilde{\mathbf{W}}_l^{(t+1)}}{\big\|\widetilde{\mathbf{W}}_l^{(t+1)}\big\|_F}, \tag{37}$$

Table 2: Fitted MSD exponent α of the combined weight vector in the early and late training phases under hard weight-norm constraints.

Mode	α_{early}	α_{late}
GP	1.30	0.76
HP	1.58	0.69

so that $\|\mathbf{W}_l^{(t)}\|_F = c_l$ holds after every update.

In the HP variant the objective additionally contains a local Hebbian term,

$$L = L_{\text{CE}} - \lambda \sum_{l=1}^{3} \cos\big(\mathbf{W}_l, \mathbf{H}_l\big), \qquad \lambda = 0.7, \tag{38}$$

which maximises the cosine similarity between each weight matrix and its Hebbian correlation matrix

$$\mathbf{H}_l = \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{B}\sum_{b=1}^{B} \boldsymbol{\phi}\big(\mathbf{u}_b^l[t]\big)\,\big(\mathbf{x}_b^{l-1}[t]\big)^\top, \qquad \phi(u) = \tanh\!\left(\frac{u}{v_{\text{th}}} - \eta_0\right), \tag{39}$$

where $\mathbf{x}_b^{l-1}[t]$ is the (exponentially scaled) presynaptic activity of sample b at timestep t, $\mathbf{u}_b^l[t]$ is the postsynaptic membrane potential, $B = 100$ is the minibatch size, $T = 10$ is the simulation window, $\beta = 1$ absorbs the presynaptic gain, and $\eta_0 = 0$ is the threshold offset. Training uses SGD with learning rate $\eta = 10^{-2}$, momentum 0.9, batch size $B = 100$ on FashionMNIST (600 steps per epoch, 18,000 steps over 30 epochs), with random seed 111.

D.2 Anomalous Diffusion under Hard Constraints

Figure 12 tracks the fitted MSD exponent α of the combined weight vector as a function of training epoch, for GP (red) and HP (purple). Both modes exhibit the familiar transition from early superdiffusion to late subdiffusion (Table 2), but with markedly different trajectories: GP decays early and saturates, crossing $\alpha = 1$ within the first ≈ 5 epochs and settling at $\alpha_{\text{late}} = 0.76$, whereas HP remains superdiffusive for the first ≈ 12 epochs and then decays more deeply, reaching $\alpha_{\text{late}} = 0.69$. The Hebbian term thus both prolongs the exploratory (superdiffusive) phase and produces stronger late-stage confinement.

The local diffusion exponent $\beta(t_w, \tau)$ (Figure 13) confirms this picture for both modes. At early waiting times ($t_w = 0$), β curves upward at large lag times—an escape-like signature compatible with rare large jumps across barrier regions, as in the inter-basin mechanism of Eq. (27). At late waiting times ($t_w = 9000, 17000$), β decreases monotonically below 1 at intermediate lags for both modes, indicating confined, subdiffusive dynamics, before rising again at the largest lags where the trajectory leaves the basin. Figure 14 shows the underlying TA-MSD curves at early, middle, and late training stages; displacement growth slows markedly in the late stage for both GP and HP.

D.3 Landscape Geometry under Hard Constraints

The two modes diverge most sharply in their curvature evolution. Figure 15 shows the largest eigenvalue of the full Hessian of the combined weight vector over training. Under HP, $\lambda_{\max}$ grows monotonically from ≈ 27 to ≈ 60 (peaking at ≈ 65), whereas under GP it *decreases* slightly, from ≈ 19 to ≈ 14, fluctuating in the range 10–19 throughout. The Hebbian term therefore drives a substantial sharpening of the landscape, while the pure cross-entropy objective under the same hard constraint does not sharpen at all. Notably, GP's diffusion nonetheless slows to strong subdiffusion ($\alpha \to 0.76$) *without* any accompanying curvature growth—showing that under hard constraints the freezing of weight dynamics is not dictated by the Hessian eigen-structure.

D.4 Implications

These observations refine the geometric–dynamical correspondence of the main text in three ways. First, the hard weight-norm constraint itself is sufficient to freeze diffusion: GP slows to strong

subdiffusion ($\alpha \to 0.76$) even though its $\lambda_{\max}$ does not grow, so the late-stage confinement of the weight trajectory cannot be read off the curvature of the underlying loss—it is imposed by the projection, which removes the radial degree of freedom and confines every update to the fixed-norm manifold. Second, the Hebbian term qualitatively changes the landscape: under HP the spectrum reorganises into an isolated sharp direction together with an expanding unstable subspace ($|\lambda_{\min}|$ and the negative-eigenvalue fraction both grow), and the resulting dynamics combine a prolonged superdiffusive phase with deeper late-stage subdiffusion than GP. The flat-minima–subdiffusion coupling reported for unconstrained GP in the main text, which requires the noise covariance to co-vary with the Hessian eigen-structure, therefore does *not* carry over to either mode under hard constraints: GP freezes without sharpening, and HP sharpens through a mechanism—intra-basin confinement on the constraint manifold rather than bulk curvature growth—that is fundamentally different from ANN-style curvature–noise coupling. This suggests that theories of SNN learning built on unconstrained loss landscapes may not apply unchanged once biologically motivated constraints are imposed directly on the synaptic weights.

E Architecture Ablation: Shallower FC-SNN

To verify that the geometric–dynamical picture of the main text is not specific to the four-layer depth, we repeat the full analysis on a shallower three-layer FC-SNN (784→512→10, one hidden layer; layers are counted including the input layer, consistently with the four-layer network of the main text). All hyperparameters are identical to Table 1 except that training is run for 75 epochs instead of 20, which provides both an extended convergence phase and a longer observation window for the diffusion analysis. All diffusion and curvature quantities are computed with the same TA-MSD and Hessian machinery as the main text.

E.1 Training Dynamics

Figure 16 summarises the training dynamics. HP and GP attain comparable final test accuracy (88.4% vs. 88.3%), both clearly above LP (78.6%). The training loss separates the paradigms more strongly: HP reaches the lowest training loss (≈ 0.07 at epoch 75), followed by GP (≈ 0.16) and LP (≈ 0.53). Notably, GP exhibits intermittent sharp drops in test accuracy (e.g., to 83.2% around epoch 22), whereas the HP trajectory remains stable throughout—consistent with GP exploring a rougher, higher-curvature region of the landscape. The test loss of HP fluctuates at a higher level in late training (0.58–0.64) despite its lower training loss and equal or higher accuracy, indicating that the Hebbian branch sustains a non-vanishing local noise floor that does not degrade classification performance.

E.2 Anomalous Diffusion

The TA-MSD curves (Figure 17) confirm that the universal transition from early superdiffusion to late subdiffusion persists on the shallower architecture under all three paradigms. For the combined weight vector (Figure 18b and Table 3), all paradigms start superdiffusive ($\alpha \approx 1.1$–1.3) and end subdiffusive, with a clear separation in the late stage: GP sustains the largest exponent ($\alpha_{\text{late}} \approx 0.72$), whereas LP ($\approx 0.26$) and HP ($\approx 0.24$) both freeze into deep subdiffusion. Two architecture-dependent differences from the four-layer network are worth noting. First, under GP the combined exponent does not decay monotonically: after reaching ≈ 0.66 in mid-training it re-escalates to ≈ 0.72, mirroring the growth of $\lambda_{\max}$ under GP on this architecture (see below). Second, HP no longer attains a markedly intermediate exponent between GP and LP as in the four-layer network; its advantage on the shallower architecture is expressed through curvature (see below) and training stability rather than through a larger residual diffusion.

Figure 19 reports the local scaling exponent $\beta(t_w, \tau)$, which again evolves from superdiffusive to subdiffusive values as training proceeds under all paradigms, confirming that the non-stationary crossover behaviour of Figure 3 is preserved on the shallower network.

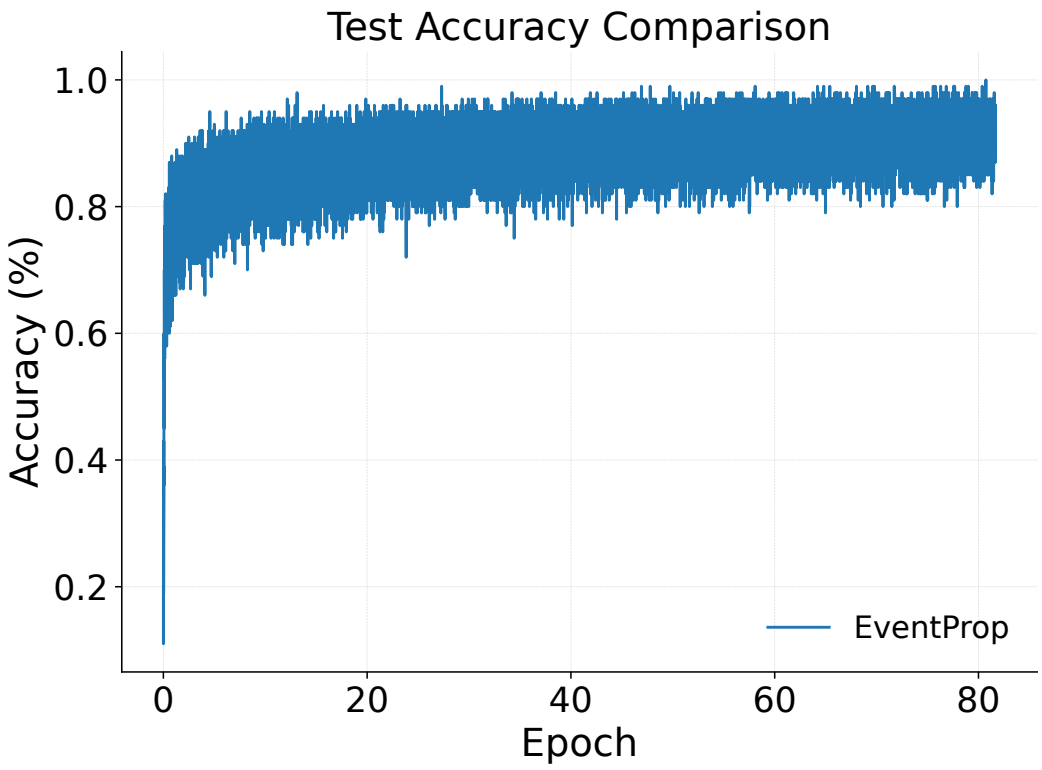
Table 3: Fitted diffusion exponent α of the combined weight vector $\mathbf{w} = (\mathbf{W}_1, \mathbf{W}_2)$ on the three-layer FC-SNN at early, middle, and late training stages.

Mode	Early	Middle	Late
GP	1.261	0.656	0.715
LP	1.286	0.353	0.256
HP	1.088	0.279	0.239

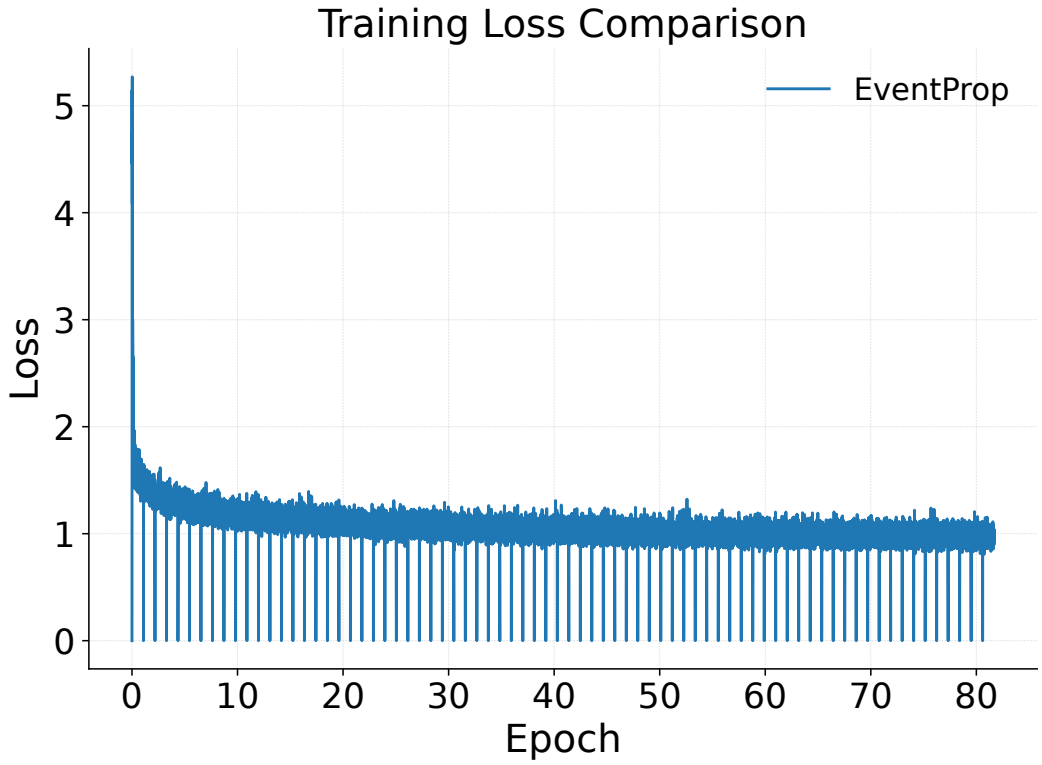
E.3 Landscape Curvature

The curvature trajectories (Figure 20) reproduce the central geometric finding: HP converges to the flattest minimum. Throughout training $\lambda_{\max}^{\mathrm{GP}}$ remains an order of magnitude larger than under the other paradigms and, in contrast to the four-layer network, *grows* over training—from ≈ 14 to ≈ 20 after a mid-training dip to ≈ 8—a progressive-sharpening regime reminiscent of the edge-of-stability phenomenon. Under LP and HP, $\lambda_{\max}$ fluctuates between ≈ 3 and 7 without a trend, with HP attaining the lowest final value (≈ 4.0 vs. ≈ 4.6 for LP). The growth of GP curvature is consistent with its late-stage re-escalation of the diffusion exponent: a sharpening landscape re-invigorates directional drift and inter-basin escape, whereas the flat, low-curvature basin found by HP confines the trajectory to controlled subdiffusion.

Overall, this ablation shows that the two main conclusions of the paper—the superdiffusive-to-subdiffusive crossover and the implicit-regularisation effect of hybrid plasticity that steers optimisation toward flat minima—are not artefacts of the four-layer depth, but generalise to a shallower SNN trained with the same paradigms, while also highlighting an architecture-dependent coupling between GP curvature growth and late-stage diffusion re-escalation.



(a) Evolution of test accuracy with training epoch



(b) Evolution of training cross-entropy loss with training epoch.

Figure 9: EventProp training dynamics on Fashion-MNIST (3-layer FC-SNN, 512 hidden units). (a) Test classification accuracy evolves with training epoch. (b) Cross-entropy loss decays over training epochs.

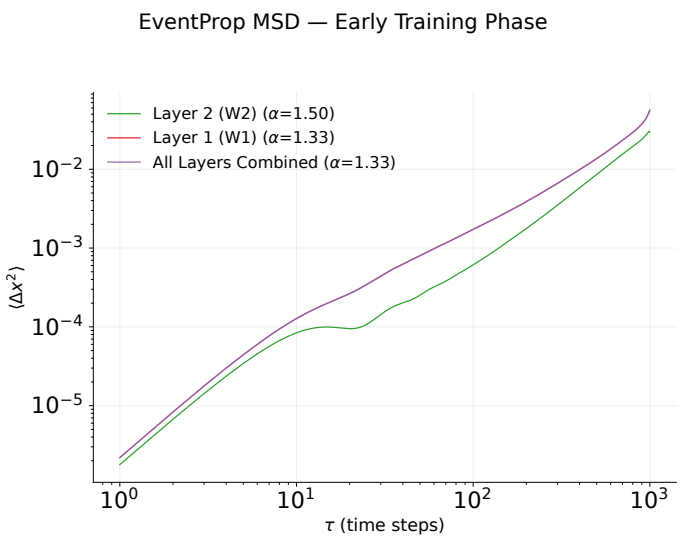


Figure 10: EventProp TA-MSD curves during early training (steps 0–1000), showing how weight displacement grows with lag time τ. Both layers exhibit superdiffusive behaviour, with the hidden-to-output matrix ($\alpha = 1.50$) showing a slightly higher exponent than the input-to-hidden matrix ($\alpha = 1.33$).

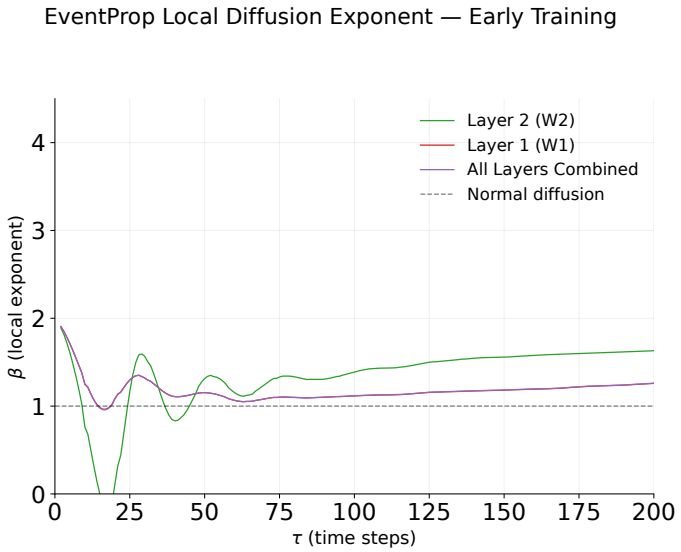


Figure 11: EventProp local diffusion exponent $\beta(t_w, \tau)$ as a function of lag time during early training. The hidden-to-output layer (W2) exhibits transient oscillations before settling to superdiffusive values, while the combined weight vector converges smoothly toward $\beta \approx 1.2$.

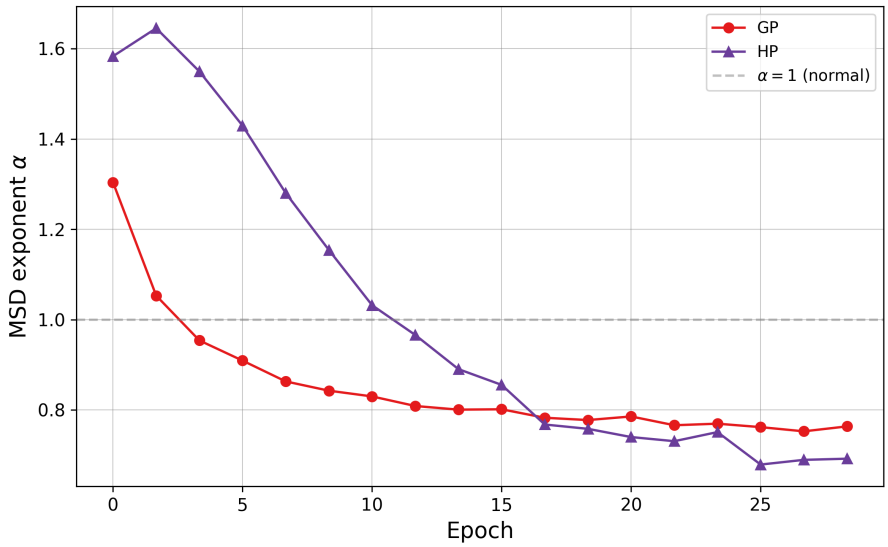


Figure 12: Evolution of the fitted MSD exponent α of the combined weight vector under hard weight-norm constraints, for GP and HP. Both modes transition from superdiffusion ($\alpha > 1$) to subdiffusion ($\alpha < 1$); HP stays superdiffusive longer and decays to a lower late-stage value.

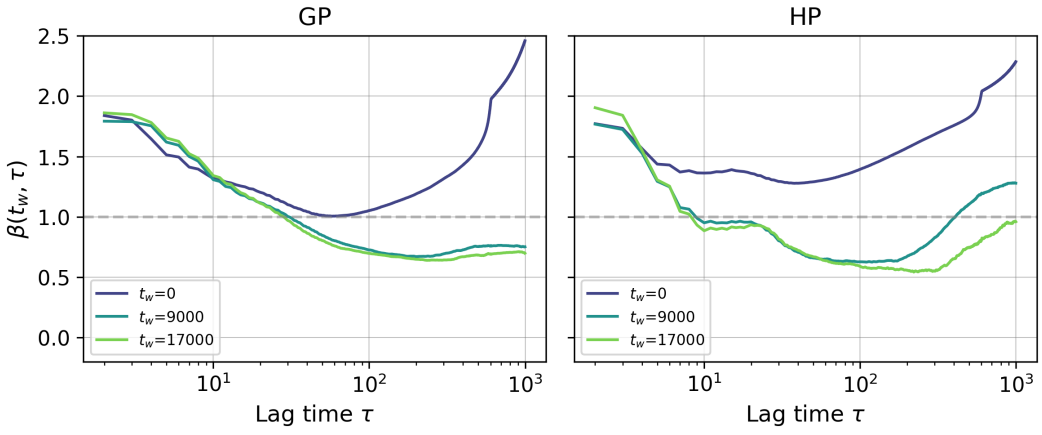


Figure 13: Local diffusion exponent $\beta(t_w, \tau)$ of the combined weight vector under hard weight-norm constraints, for GP (left) and HP (right), at waiting times $t_w = 0$, 9000, and 17000. Early waiting times display superdiffusive behaviour with upward curvature at large lags, compatible with saddle-point escape; late waiting times decay below normal diffusion ($\beta < 1$) at intermediate lags for both modes.

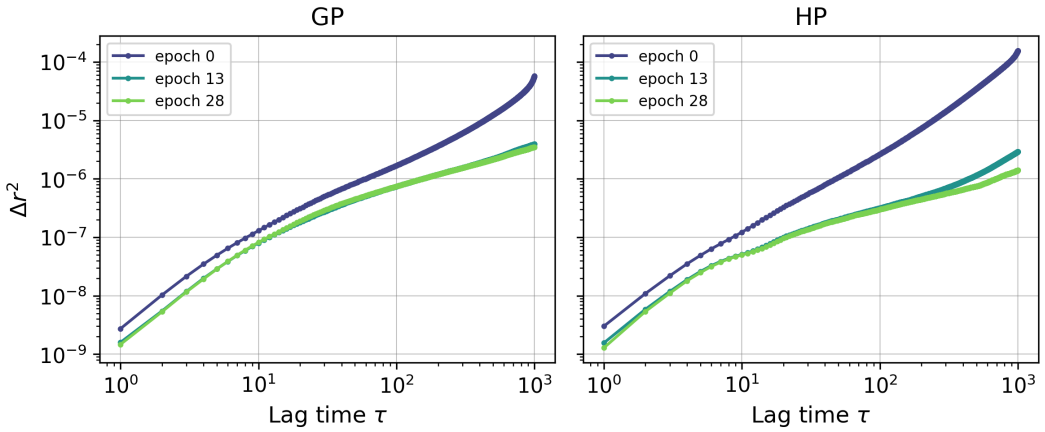


Figure 14: TA-MSD of the combined weight vector under hard weight-norm constraints as a function of lag time τ, for GP (left) and HP (right), at early, middle, and late training stages. Displacement growth slows markedly in the late stage in both modes.

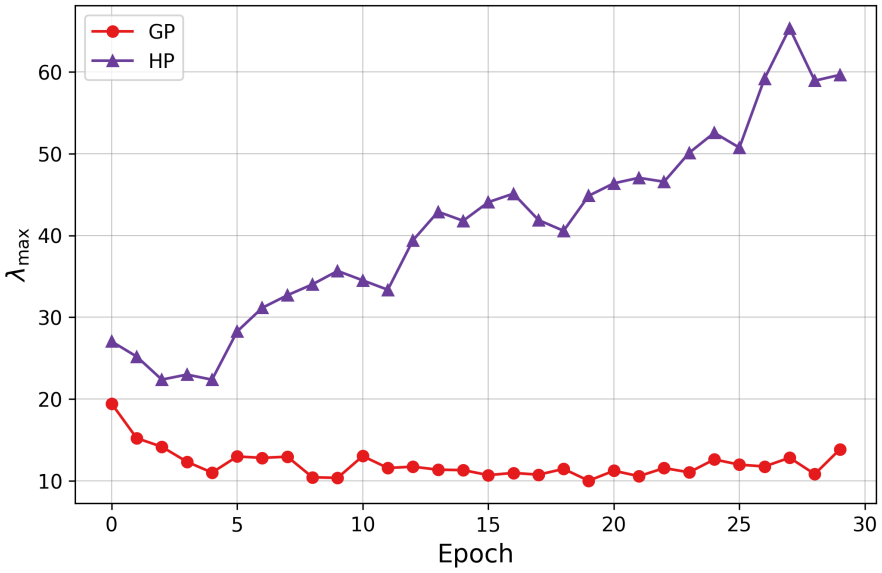


Figure 15: Evolution of the largest Hessian eigenvalue λ_{max} (full Hessian of the combined weight vector) under hard weight-norm constraints, for GP and HP. HP's λ_{max} grows monotonically by more than a factor of two, whereas GP's remains flat and even decreases slightly.

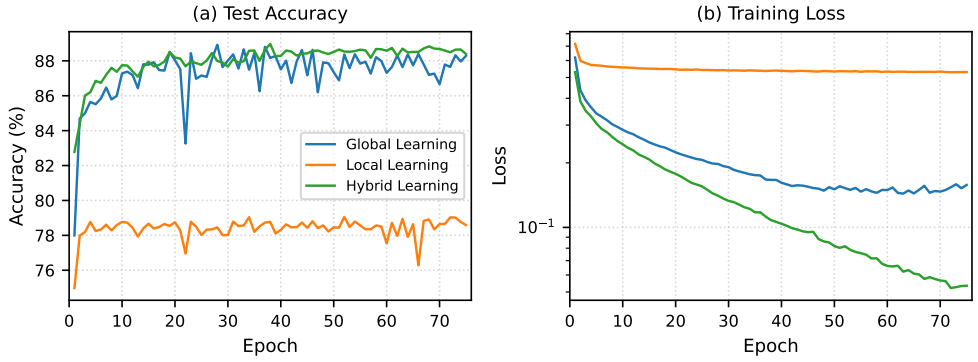
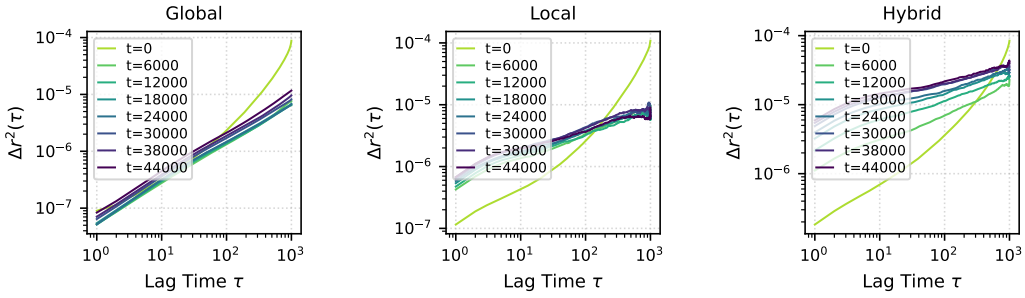
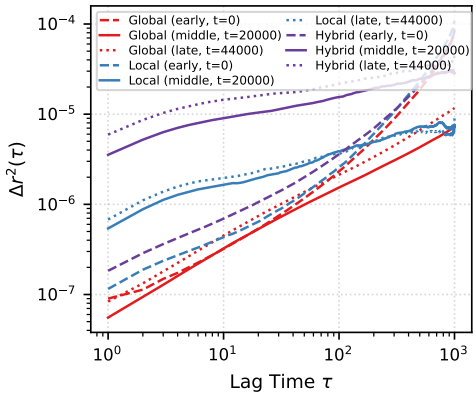


Figure 16: Training dynamics of the three-layer FC-SNN (784→512→10) under GP, LP, and HP. (a) Test classification accuracy. (b) Training loss. HP matches GP in accuracy while attaining the lowest training loss and avoiding the intermittent accuracy drops of GP; LP plateaus at a markedly lower accuracy.

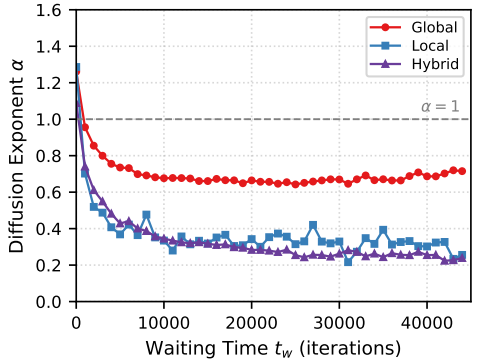


(a) Global Plasticity (GP). (b) Local Plasticity (LP). (c) Hybrid Plasticity (HP).

Figure 17: Time-resolved TA-MSD of the combined weight vector $\mathbf{w} = (\mathbf{W}_1, \mathbf{W}_2)$ for the three-layer FC-SNN under (a) GP, (b) LP, and (c) HP. Each panel shows the displacement $\Delta r^2(\tau)$ as a function of lag time τ at successive training stages (single representative run); lighter colours indicate earlier stages and darker colours later stages. Displacement growth slows with training under all paradigms, with GP alone showing a re-acceleration of the late-stage curves.

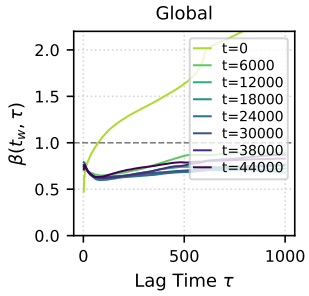


(a) TA-MSD at early, middle, and late training stages.

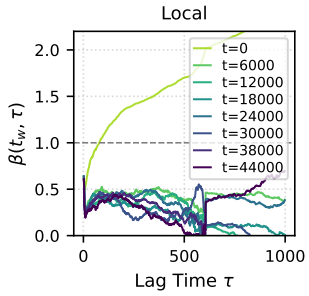


(b) Evolution of the diffusion exponent α of the combined weight vector with waiting time t_w.

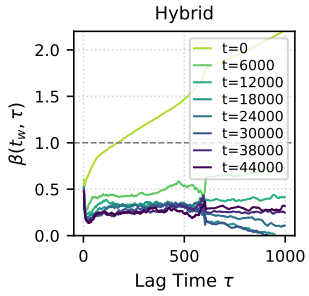
Figure 18: Cross-mode diffusion analysis for the three-layer FC-SNN. (a) TA-MSD of the combined weight vector at early, middle, and late training stages. (b) Diffusion exponent α of the combined weight vector over training: all paradigms cross from superdiffusion ($\alpha > 1$) to subdiffusion ($\alpha < 1$), and GP alone re-escalates in the late stage.



(a) Global Plasticity (GP).



(b) Local Plasticity (LP).



(c) Hybrid Plasticity (HP).

Figure 19: Local scaling exponent $\beta(t_w, \tau)$ of the combined weight vector $\mathbf{w} = (\mathbf{W}_1, \mathbf{W}_2)$ for the three-layer FC-SNN under (a) GP, (b) LP, and (c) HP. Lighter colours indicate earlier waiting times and darker colours later waiting times. All paradigms exhibit the same qualitative decay of the local exponent as training progresses.

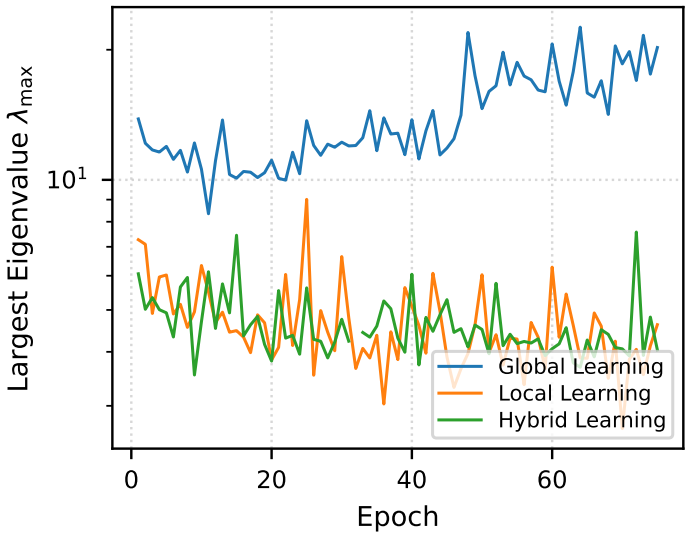


Figure 20: Evolution of the largest Hessian eigenvalue $\lambda_{\max}$ with training epoch for the three-layer FC-SNN. $\lambda_{\max}^{\mathrm{GP}}$ grows over training (progressive sharpening), whereas LP and HP remain an order of magnitude lower; HP attains the lowest final curvature.