A Conservation Law Perspective on Explainability in Spiking Neural Networks

Sylvester Kaczmarek
Imperial College London, London, UK
research@sylvesterkaczmarek.com

Abstract

Explainability in event-driven dynamical models such as Spiking Neural Networks (SNNs) is often obtained by adapting methods designed for static networks, while the explanation itself is implicitly defined over a spatio-temporal computation graph. We give a formal viewpoint that treats Layer-wise Relevance Propagation (LRP) in SNNs as a discrete conservative transport scheme, where relevance is the transported quantity and the LRP redistribution rule defines a flux on graph edges. Under mild conditions, the backward pass satisfies a discrete continuity equation on the unrolled graph. This connects LRP's conservation axiom to a standard notion from numerical conservation laws and yields a small set of implementation-level verification checks. The contribution is primarily conceptual and formal, with an illustrative empirical example.

1 Introduction

Spiking Neural Networks (SNNs) are dynamical systems whose states evolve in continuous time and are driven by discrete events (Maass, 1997). This temporal structure complicates explainability, since many attribution methods were designed for feed-forward computation graphs that do not include explicit state dynamics. In practice, LRP-style methods have been adapted to SNNs and used to produce spatio-temporal attributions (Kim & Panda, 2021; Sun et al., 2024), yet the justification for relevance conservation in a leaky, event-driven setting is rarely stated as an explicit balance law.

This paper connects LRP in SNNs to conservation laws from mathematical physics and numerical methods (LeVeque, 1992). The key observation is that an SNN, once unrolled in time, is a directed acyclic computation graph whose nodes include both spatial interactions (synapses) and temporal interactions (state updates). LRP assigns a scalar relevance to each node and redistributes it to predecessors. That redistribution can be viewed as a flux that satisfies a discrete continuity equation on the unrolled graph.

Our contribution is primarily formal. We make the conservation structure explicit on the unrolled graph and derive transparent accounting checks for implementations. We do not claim a benchmark-scale empirical study in this tiny-paper format.

Contributions.

- We define a spatio-temporal graph for unrolled SNN dynamics and an associated notion of relevance density and flux.
- We show that a broad class of LRP rules induces a locally conservative discrete balance law on this graph, with an explicit residual term when stabilizers are used.
- We give explicit handling of leak, constant terms, and spike-reset updates as graph computations, yielding simple accounting checks for explanation correctness.

2 Background and related work

2.1 SNN dynamics as an unrolled computation graph

Consider a directed network of LIF neurons. For neuron i, the membrane potential $V_i(t)$ obeys (Maass, 1997):

$$\tau_m \frac{dV_i(t)}{dt} = -(V_i(t) - V_{\text{rest}}) + I_i(t), \tag{1}$$

where τ_m is the membrane time constant, V_{rest} is the resting potential, and $I_i(t)$ is synaptic input. In discrete time with step Δt, a forward Euler update gives:

$$V_i(t) = aV_i(t-1) + bI_i(t) + (1-a)V_{\text{rest}}, \quad a = 1 - \frac{\Delta t}{\tau_m}, \quad b = \frac{\Delta t}{\tau_m}. \tag{2}$$

Spikes are events defined by a threshold on a pre-reset potential, and a reset map then updates state. Unrolling equation 2 over time yields a directed acyclic graph over variables $\{V_i(t)\}$ plus auxiliary variables for spikes and currents.

2.2 LRP and conservation

LRP attributes a model output to intermediate nodes and ultimately to inputs by redistributing relevance backward through the computation graph (Bach et al., 2015; Montavon et al., 2019). For a node u with relevance R_u, and predecessors $v \in \text{pred}(u)$, a common family of rules takes the form

$$R_{v \leftarrow u} = R_u \frac{z_{v \to u}}{\sum_{v' \in \text{pred}(u)} z_{v' \to u} + s_u}, \tag{3}$$

where $z_{v \to u}$ are forward contributions and s_u is a stabilizer or correction term. When $s_u = 0$ and the denominator is nonzero, the rule is exactly conservative at node u in exact arithmetic: $\sum_v R_{v \leftarrow u} = R_u$.

LRP has been extended to temporal models such as recurrent networks (Arras et al., 2017) and applied to spiking and event-driven settings (Kim & Panda, 2021; Sun et al., 2024). Separately, saliency methods benefit from sanity checks that detect explanations insensitive to data or model parameters (Adebayo et al., 2018). Our aim is to state a conservation-law interpretation that makes such checks natural for SNN explanations.

3 LRP as a discrete conservation law on a spatio-temporal graph

3.1 Relevance density, flux, and graph divergence

Definition 1 (Spatio-temporal graph). *Let $G = (\mathcal{V}, \mathcal{E})$ be the unrolled computation graph of an SNN over times $t = 0, \ldots, T$. Nodes $u \in \mathcal{V}$ correspond to scalar variables such as $V_i(t)$, $I_i(t)$, spike indicators, and constants (bias-like nodes). Directed edges $(v \to u) \in \mathcal{E}$ represent functional dependence in the forward computation, for example, $V_i(t-1) \to V_i(t)$ and $I_i(t) \to V_i(t)$ induced by equation 2.*

We interpret node relevance R_u as a *relevance density* on $\mathcal{V}$. Each redistribution $R_{v \leftarrow u}$ defines a *relevance flux* on the reverse edge $(u \to v)$:

$$J_{u \to v} := R_{v \leftarrow u}. \tag{4}$$

To state a discrete continuity equation, we need a discrete divergence. For a node u, define the net outflux and influx:

$$(\nabla \cdot J)_u := \sum_{v \in \text{pred}(u)} J_{u \to v} - \sum_{w \in \text{succ}(u)} J_{w \to u}. \tag{5}$$

This should be read as a graph-theoretic bookkeeping operator, analogous to a finite-volume divergence (LeVeque, 1992), rather than as a continuum differential operator. The first sum collects flux leaving u toward its predecessors in backward propagation, and the second sum collects flux entering u from successors.

3.2 LOCAL BALANCE LAW INDUCED BY LRP

The next proposition is mathematically modest but operationally useful, since it makes the local residual explicit at each node.

Proposition 1 (Local balance with explicit residual)**.** *For any node u with relevance R_u redistributed according to equation 3, define the* node residual

$$\varepsilon_u := R_u - \sum_{v \in \mathrm{pred}(u)} R_{v \leftarrow u}. \tag{6}$$

Then the backward redistribution satisfies the local balance law

$$\sum_{v \in \mathrm{pred}(u)} J_{u \to v} = R_u - \varepsilon_u. \tag{7}$$

In particular, if $s_u = 0$ and the denominator in equation 3 is nonzero, then $\varepsilon_u = 0$ in exact arithmetic.

Proof sketch. Substitute equation 3 into the sum over predecessors. The residual equation 6 is the difference between R_u and the sum of redistributed relevance, yielding equation 7. □

Remark 1 (Stabilizers as sinks)**.** *A common stabilizer is $s_u = \epsilon \operatorname{sign}(\sum_{v'} z_{v' \to u})$ with $\epsilon > 0$. In that case, ε_u is generally nonzero and can be interpreted as a controlled sink of relevance at u. One can make the scheme strictly conservative by adding an explicit sink node that receives ε_u.*

Failure modes and scope. The conservation statements above require the denominator in equation 3 to be well-defined. In practice, hard zeros can arise at inactive gating nodes or from cancellation of signed contributions. A robust implementation therefore needs an explicit convention: use a stabilizer s_u, route unreassigned relevance to an explicit sink node, or define a skip rule for inactive branches. Inhibitory contributions and signed LRP variants can be handled by applying the same bookkeeping separately to positive and negative channels, but we do not develop that extension here.

3.3 TEMPORAL FLUX FROM THE LIF UPDATE

We now make the spatio-temporal structure explicit for the LIF Euler update equation 2. Consider a node $u \equiv V_i(t)$ with predecessors $v_1 \equiv V_i(t-1)$, $v_2 \equiv I_i(t)$, and $v_3 \equiv V_{\text{rest}}$ (a constant node). Let forward contributions be

$$z_{v_1 \to u} = aV_i(t-1), \quad z_{v_2 \to u} = bI_i(t), \quad z_{v_3 \to u} = (1-a)V_{\text{rest}}. \tag{8}$$

Applying equation 3 to u yields fluxes $J_{u \to v_1}$ (temporal), $J_{u \to v_2}$ (toward inputs at time t), and $J_{u \to v_3}$ (toward the constant term). This realizes a discrete transport step of relevance backward along the temporal edge $V_i(t) \to V_i(t-1)$ with a conservative split over the affine update terms.

3.4 A DISCRETE SPATIO-TEMPORAL CONTINUITY EQUATION

Group nodes by time index and focus on state nodes $\{V_i(t)\}$. Let $R_i(t)$ denote relevance associated with $V_i(t)$. The backward propagation across time and synapses induces net temporal and spatial fluxes. The resulting balance can be stated as a discrete continuity equation:

$$\frac{R_i(t) - R_i(t-1)}{\Delta t} + (\nabla_j \cdot J^{\text{spatial}})_i(t) + (\nabla_\tau \cdot J^{\text{temporal}})_i(t) = -\frac{\varepsilon_i(t)}{\Delta t}, \tag{9}$$

where $\varepsilon_i(t)$ aggregates residual terms from the local redistributions involving $V_i(t)$ and where graph divergences are defined as in equation 5 restricted to spatial (synaptic) and temporal (state-update) edges. With $s_u = 0$ for all involved nodes, the right-hand side vanishes in exact arithmetic, yielding a conservative discrete continuity equation on the unrolled graph.

Derivation note. Eq. equation 9 is obtained by summing the node-wise balance in equation 7 over state nodes at time t, separating predecessor edges into spatial and temporal sets, and rearranging terms into divergence form. Boundary terms at $t = 0$ and at the output layer are absorbed into the initialization and output relevance assignments. Accordingly, equation 9 should be read as a discrete bookkeeping identity on the unrolled graph, rather than as a claim about a continuum limit or a deeper PDE result.

4 Handling spikes and reset as graph updates

Spike emission and reset introduce discontinuities in time, but they are still computations on the unrolled graph.

Let $V_i^{\text{pre}}(t)$ be the pre-reset potential and define a spike indicator $S_i(t) = \mathbb{I}[V_i^{\text{pre}}(t) \geq V_{th}]$. A simple reset map is

$$V_i(t) = (1 - S_i(t))V_i^{\text{pre}}(t) + S_i(t)\, V_{\text{reset}}, \tag{10}$$

with constant V_{reset}. Once $S_i(t)$ is treated as a realized forward variable on the unrolled graph, equation 10 is an affine computation on $(V_i^{\text{pre}}(t), S_i(t), V_{\text{reset}})$, so the same redistribution logic used for equation 2 applies to the reset step. This avoids requiring surrogate gradients for relevance redistribution itself. Surrogate gradients remain useful for training (Neftci et al., 2019). At the thresholding node that generates $S_i(t)$, however, hard zeros or degeneracies may require a stabilizer or explicit sink convention; the conservation accounting then records the resulting residual rather than hiding it.

5 Verification checks implied by the conservation viewpoint

The conservation-law interpretation provides simple, implementation-level checks:

Local conservation residual. For each node u, compute ε_u from equation 6. When using strictly conservative redistributions ($s_u = 0$), ε_u should be zero up to numerical precision. When using stabilizers, ε_u should match the chosen sink behavior.

Global accounting. Let R_{out} be the relevance assigned at the output. Let R_{in} be the sum of relevances at the input nodes plus any explicit sink nodes. A correct implementation should satisfy $R_{\text{out}} \approx R_{\text{in}}$ with a discrepancy explainable by accumulated residuals.

Sanity checks. Randomization tests from Adebayo et al. (2018) can be applied to spatio-temporal explanations: randomize weights or labels and verify that attribution patterns change accordingly. These tests complement conservation checks, since a conservative explanation can still be uninformative.

6 Illustrative analysis

This example is illustrative rather than benchmark-scale. We use an explanation for a detected anomaly from a cislunar robotics simulation context (Izzo et al., 2022). The scenario involves a simulated micrometeoroid impact that manifests as a short transient spike in one accelerometer axis while other sensor readings remain nominal. The SNN flags the event as anomalous.

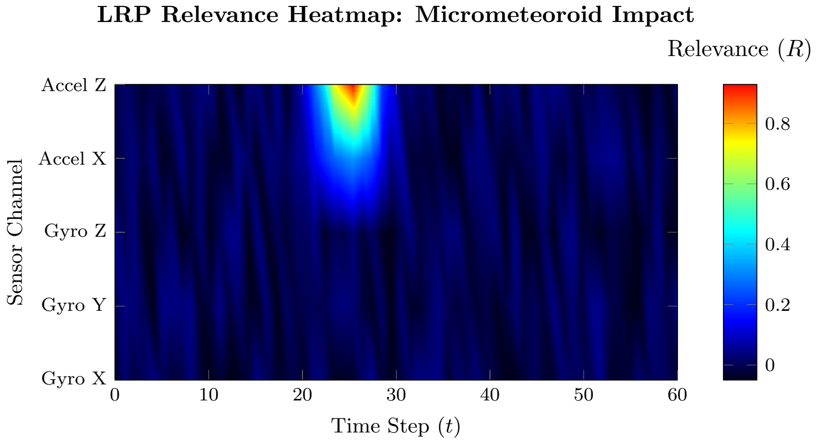

Figure 1: LRP relevance heatmap for a simulated micrometeoroid impact. The heatmap shows relevance density distributed across sensor channels and time. Relevance concentrates on Accel Z around the event time, consistent with backward transport of relevance to the causal input segment.

Figure 1 visualizes relevance over channels and time. In the conservation-law view, the anomaly score initializes relevance at the output node, and the backward pass transports this relevance through temporal and spatial edges until it reaches the input segment that most strongly contributed to the output. Along the way, the local residuals in equation 6 provide explicit accounting for any stabilizer-induced sinks.

7 Discussion and conclusion

We presented a conservation-law viewpoint on explainability in SNNs by interpreting LRP as a discrete transport scheme on the unrolled spatio-temporal computation graph. This yields a precise notion of relevance flux and divergence on a graph, a discrete continuity equation with an explicit residual term, and practical accounting checks for implementations, including conservation residuals and randomization-based sanity checks.

The same viewpoint suggests extensions. Different redistribution rules correspond to different numerical fluxes, while stability choices such as stabilizers correspond to controlled sinks. These choices can be analyzed using standard tools from numerical conservation laws (LeVeque, 1992), with the goal of producing explanations whose behavior is predictable under time discretization, resets, and event sparsity.

This viewpoint is complementary to completeness-style attribution methods such as Integrated Gradients or DeepLIFT, but here the emphasis is on local conservation and residual accounting on the unrolled spatio-temporal graph. We do not provide benchmark-scale comparisons against alternative attribution methods, a systematic analysis of stabilizer sensitivity, or a broad study across SNN datasets and neuron models. Those are natural next steps, but the present paper's main contribution is to make the conservation bookkeeping explicit and testable on the unrolled spatio-temporal graph.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL `https://papers.nips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html`.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 159–168, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5221. URL https://aclanthology.org/W17-5221/.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015. doi: 10.1371/journal.pone.0130140. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140.

Dario Izzo, Alexander Hadjiivanov, Dominik Dold, Gabriele Meoni, and Emmanuel Blazquez. Neuromorphic computing and sensing in space. *arXiv preprint arXiv:2212.05236*, 2022. doi: 10.48550/arXiv.2212.05236. URL https://arxiv.org/abs/2212.05236.

Youngeun Kim and Priyadarshini Panda. Visual explanations from spiking neural networks using inter-spike intervals. *Scientific Reports*, 11:19037, 2021. doi: 10.1038/s41598-021-98448-0. URL https://www.nature.com/articles/s41598-021-98448-0.

Randall J. LeVeque. *Numerical Methods for Conservation Laws*. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, 1992. doi: 10.1007/978-3-0348-8629-1. URL https://link.springer.com/book/10.1007/978-3-0348-8629-1.

Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997. doi: 10.1016/S0893-6080(97)00011-7. URL https://www.sciencedirect.com/science/article/pii/S0893608097000117.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: An overview. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Hansen, and Klaus-Robert Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pp. 193–209. Springer, Cham, 2019. doi: 10.1007/978-3-030-28954-6_10. URL https://link.springer.com/chapter/10.1007/978-3-030-28954-6_10.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019. doi: 10.1109/MSP.2019.2931595.

Mingyuan Sun, Donghao Zhang, Zongyuan Ge, Jiaxu Wang, Jia Li, Zheng Fang, and Renjing Xu. Eventrpg: Event data augmentation with relevance propagation guidance. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=i7LCsDMcZ4.