CHLU: THE CAUSAL HAMILTONIAN LEARNING UNIT AS A SYMPLECTIC PRIMITIVE FOR DEEP LEARNING

Pratik Jawahar
University of Manchester
pratik.jawahar@cern.ch

Maurizio Pierini
CERN

ABSTRACT

Current deep learning primitives dealing with temporal dynamics suffer from a fundamental dichotomy: they are either discrete and unstable (LSTMs) (Pascanu et al., 2013), leading to exploding or vanishing gradients; or they are continuous and dissipative (Neural ODEs) (Dupont et al., 2019), which destroy information over time to ensure stability. We propose the **Causal Hamiltonian Learning Unit** (pronounced: *clue*), a novel Physics-grounded computational learning primitive. By enforcing a Relativistic Hamiltonian structure and utilizing symplectic integration, a CHLU strictly conserves phase-space volume, as an attempt to solve the memory-stability trade-off. We show that the CHLU is designed for infinite-horizon stability, as well as controllable noise filtering. We then demonstrate a CHLU's generative ability using the MNIST dataset as a proof-of-principle.

1 INTRODUCTION

The modeling of physical systems and long-term dependencies remains an open challenge in deep learning. Deep learning architectures, while empirically successful, fail to learn implicit conservation laws that govern physical reality (Vafa et al., 2025). Discrete units (eg. RNNs (Salehinejad et al., 2018), Transformers (Vaswani et al., 2017)) often lack interpretable state dynamics, while continuous-depth models like Neural ODEs (NODEs) (Chen et al., 2019) typically model dissipative systems, making them unsuitable for long-term information preservation. Hamiltonian Neural Networks (Greydanus et al., 2019) have emerged to enforce energy conservation in traditional Neural Networks, but they are predominantly designed for *simulation* rather than *general inference* on high-dimensional data.

We introduce the **Causal Hamiltonian Learning Unit (CHLU)**{pronounced: *clue*}, a computational learning primitive that treats energy conservation not as a target to be learned, but as a structural prior to moderate latent evolution. A CHLU integrates relativistic mechanics (Hawking & Israel, 2010) with symplectic geometry (Siegel, 2014) and redefines information propagation as an evolution of the unit's internal state. This work is an early stage description of a CHLU and a comprehensive study is left for future work. Our contributions are two-fold:

Relativistic Kinetic Governor: We introduce a configurable speed limit c, that ensures bounded velocity updates inherently, acting as a structural constraint on kinetic stability, which we intend to utilize in future work on networks of multiple CHLUs. In contrast, symplectic integrators are incorporated in RNNs (Chen et al., 2020; Erichson et al., 2021; Salem, 2022), specifically to solve vanishing gradients. Other works like, LorentzNet (Gong et al., 2022) have explored relativistic symmetries in high-energy physics data, while a CHLU's relativistic definitions target temporal stability of state updates. We reserve the extension to a relativistic attention mechanism for future work.

Action Propagation: We use a thermodynamic modification of the Wake-Sleep algorithm (Hinton et al., 1995), optimizing the *action* difference between "clamped-wake" and "free-sleep" trajectories. Standard Energy-Based Models (EBMs), like Helmholtz Machines (Dayan et al., 1995) and Restricted Boltzmann Machines (Smolensky, 1986), use Contrastive Divergence (Carreira-Perpinan & Hinton, 2005) or Equilibrium Propagation (Scellier & Bengio, 2017) to derive gradients by relaxing a network to a static fixed point. A CHLU's action propagation is a generalization of these

methods. On the other hand, unlike Denoising Hamiltonian Networks (Deng et al., 2025) which combine dynamics with diffusion masking, a CHLU treats generation as thermodynamic relaxation on a learned potential energy surface.

2 METHODOLOGY: THE GEOMETRY OF A CHLU

2.1 THE SEPARABLE HAMILTONIAN ENGINE

The core of a CHLU is a dynamical system defined by a learnable Hamiltonian function $\mathcal{H}(q, p)$. Unlike standard Euclidean Hamiltonians, we utilize a **Relativistic Hamiltonian** to strictly impose causal evolution bounds and ensure energy stability.

The latent state $\mathbf{z} = (\mathbf{q}, \mathbf{p})$, represents generalized positions and momenta. The Hamiltonian $\mathcal{H}$ is:

$$\mathcal{H}(q,p) = \underbrace{T(\mathbf{p})}_{\text{Relativistic Kinetic Governor}} + \underbrace{V_\theta(q)}_{\text{Learnable Potential Energy}} + \underbrace{\alpha\|q\|^2}_{\text{Global Confinement Potential}} \tag{1}$$

where $V_\theta(q)$ is a learnable non-linear potential energy function parameterized by a neural network. The final term is a weak quadratic confinement potential parameterized by α, which acts as a regularizer when V_θ is flat (analogous to a weak, local gravitational potential).

The state $z = (q, p)$ evolves according to Hamilton's equations:

$$\frac{dq}{dt} = \frac{\partial \mathcal{H}}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial \mathcal{H}}{\partial q} \tag{2}$$

2.2 RELATIVISTIC KINETIC GOVERNOR

In Newtonian mechanics, the kinetic energy is typically $T(\mathbf{p}) = \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}$, which allows for unbounded velocities. To prevent such kinetic instabilities we replace this with a Relativistic Kinetic Governor. This component is parameterized by a learnable diagonal, positive-definite mass matrix $\mathbf{M}$, and the hyperparameters, rest mass m_0 and the speed of causality c. Thus, $T(\mathbf{p})$ is defined as:

$$T(\mathbf{p}) = \sqrt{c^2\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} + m_0^2c^4} \tag{3}$$

This formulation ensures that the velocity $\dot{\mathbf{q}} = \nabla_p T$ saturates at the speed limit c (shaped by the geometry of $\mathbf{M}$) as momentum increases (see App.B), thereby enforcing strictly bounded velocity updates, preventing the kinetic explosions common in unconstrained recurrent architectures.

2.3 SYMPLECTIC INTEGRATION VIA VELOCITY VERLET

To ensure the system conserves energy over infinite horizons (or dissipates it controllably), we embed a **Dissipative Velocity Verlet** integrator (Martys & Mountain, 1999) directly into the forward pass. This allows us to switch between conservative dynamics ($\gamma = 0$) and dissipative convergence ($\gamma > 0$), where γ is a friction-parameter incorporated to the step update as:

$$\mathbf{p}_{t+0.5} = \mathbf{p}_t - \frac{\epsilon}{2}\nabla V_\theta(\mathbf{q}_t) \tag{4}$$

$$\mathbf{q}_{t+1} = \mathbf{q}_t + \epsilon\nabla T(\mathbf{p}_{t+0.5}) \tag{5}$$

$$\mathbf{p}^*_{t+1} = \mathbf{p}_{t+0.5} - \frac{\epsilon}{2}\nabla V_\theta(\mathbf{q}_{t+1}) \tag{6}$$

$$\mathbf{p}_{t+1} = (1-\gamma)\mathbf{p}^*_{t+1} \quad \text{(Dissipative Step)} \tag{7}$$

3 TRAINING DYNAMICS: HAMILTONIAN CONTRASTIVE DIVERGENCE

A CHLU's Energy-based Wake-Sleep algorithm consists of:

- **Wake Phase (Supervised):** The system is tasked with minimizing the Mean Squared Error (MSE) between the prediction (integrator step) and the target. To enforce training stability, we apply a regularization term $\mathcal{L}_{reg}$ penalizing the Lyapunov exponents of the Jacobian.

- **Sleep Phase (Unsupervised):** The system evolves freely from a replay buffer. We update weights to raise the energy of hallucinations (unless they match the data distribution).

Algorithm 1 CHLU Wake-Sleep Training (Single Step)

Require: Data trajectory $\mathbf{z}_{target}$, Initial state $\mathbf{z}_0$, Learning rate η
Wake Phase (Clamped Evolution)
$\mathbf{z}_{wake} \leftarrow \text{VelocityVerlet}(\mathbf{z}_0, \nabla\mathcal{H}, \gamma = 0)$
$\mathcal{L}_{wake} \leftarrow \text{MSE}(\mathbf{z}_{wake}, \mathbf{z}_{target}) + \lambda\mathcal{L}_{reg}(\text{Lyapunov})$
Sleep Phase (Free Evolution)
$\mathbf{z}_{sleep} \leftarrow \text{Sample}(\text{ReplayBuffer})$
$\mathbf{z}_{hallucination} \leftarrow \text{VelocityVerlet}(\mathbf{z}_{sleep}, \nabla\mathcal{H}, \gamma = 0)$
Action Update
$\Delta\theta \propto -\nabla_\theta\mathcal{H}(\mathbf{z}_{wake}) + \nabla_\theta\mathcal{H}(\mathbf{z}_{hallucination})$
$\theta \leftarrow \theta - \eta\Delta\theta$
$\text{ReplayBuffer.add}(\mathbf{z}_{hallucination})$

The final weight update is proportional to the difference of these two signals:

$$\Delta\theta \propto -\nabla_\theta\mathcal{H}(z_{wake}) + \nabla_\theta\mathcal{H}(z_{sleep}) \tag{8}$$

This contrastive signal ensures that the system learns to differentiate between physical signals (low energy) and noise (high energy). This training mechanism is currently an empirical choice.

Hamiltonian Annealing: While the core unit is conservative, we employ the controllable friction coefficient γ during inference. Setting $\gamma > 0$ drains entropy from the system, collapsing the state into the nearest stable attractor (ground state) defined by V_θ.

Generative Sampling via Langevin Dynamics: To transition from deterministic inference to generative modeling, we couple the Hamiltonian system with stochastic Langevin Dynamics (Brooks et al., 2011). During inference, we modify the momentum update as:

$$d\mathbf{p} = -\nabla V_\theta(\mathbf{q})dt - \gamma\mathbf{p}dt + \sqrt{2\gamma k_B\mathcal{T}}d\mathbf{W} \tag{9}$$

where γ is the friction coefficient, $\mathcal{T}$ is a temperature hyperparameter, and $d\mathbf{W}$ is a Wiener process (standard Brownian motion). This turns the CHLU into a sampler for the Boltzmann distribution $P(\mathbf{q}, \mathbf{p}) \propto \exp(-\mathcal{H}/k_B\mathcal{T})$. By annealing the temperature $\mathcal{T}$, the system settles into low-energy modes of the learned potential $V_\theta(\mathbf{q})$, effectively "crystallizing" noise into structured data samples.

4 Scope of Experiments and Results

For this work we provide two experiments that demonstrate specific inductive biases of a CHLU, followed by an `MNIST` (Yann, 2010) generation experiment. We choose an LSTM (Hochreiter & Schmidhuber, 1997) and a NODE as baselines not to compare performance, but rather to contrast the inductive biases of each algorithm and their learned topologies. We do not tune hyperparameters for any model. See appendices for further plots and discussions. We keep performance comparison experiments and studies on failure modes of a CHLU for a larger, comprehensive future study.

Experiment I: Long-Horizon Stability (Lemniscate Tracing) We train the three models to learn the trajectory of a Lemniscate, a self-intersecting orbit. Each model is trained on 3 cycles and asked to infer the shape for 50 cycles.

The LSTM learns the general shape but over time, numerical errors accumulate positively, pushing the trajectory into a high-energy limit cycle. The NODE accurately captures short-term dynamics with smooth curves, However, the dissipative nature of NODEs causes the trajectory to spiral inward, eventually collapsing to the origin. The CHLU preserves the shadow Hamiltonian exactly. Despite the learned trajectory being imperfect, the error is bounded. The orbit remains closed and stable indefinitely, demonstrating that symplectic constraints are necessary for long-term topological fidelity.

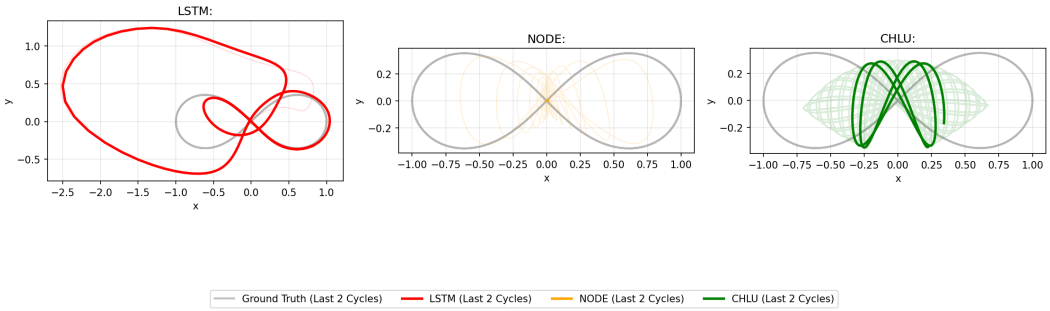

Figure 1: Lemniscate tracing experiment. The transparent lines show inference cycles $0 - 48$ and the solid lines show the final 2 cycles, for LSTM, NODE and CHLU (left to right)

Experiment II: Kinetic Safety (The Perturbed Sine Wave) The models are trained on 100 sine wave trajectories of length $T = 1000$, with frequencies $\omega \sim U(0.5, 2.0)$, where each algorithm outputs the states (q, p). At inference, the initial states are perturbed and the subsequent trajectory is predicted.

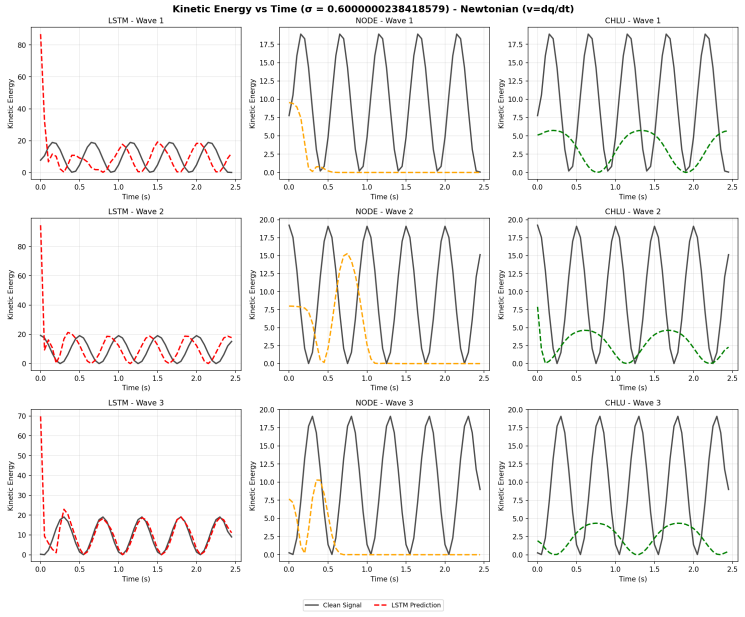

Figure 2: The LSTM, NODE and CHLU (left to right), each predicts the state q given the same initial perturbed states, from which we calculate the KE and overlay it on the expected KE.

The LSTM predicts an initial velocity spike to correct for the noise instantly. In physical systems, such instant, near-infinite acceleration is non-physical and represents a catastrophic failure of the model's internal physics engine. The NODE collapses the wave completely as a trivial solution to MSE reduction. The CHLU, by virtue of the dissipative verlet step, smoothly saturates the velocity at c. The perturbation is converted into a phase shift rather than a magnitude divergence. This demonstrates that imposing a causal speed limit is a robust defense against initialization instabilities.

Experiment III: Thermodynamic Generation We train a CHLU on 10k `MNIST` images. Then we use Langevin Dynamics during the generative phase to produce digits from a starting point that is the centroid of 2k held out images with added Gaussian noise.

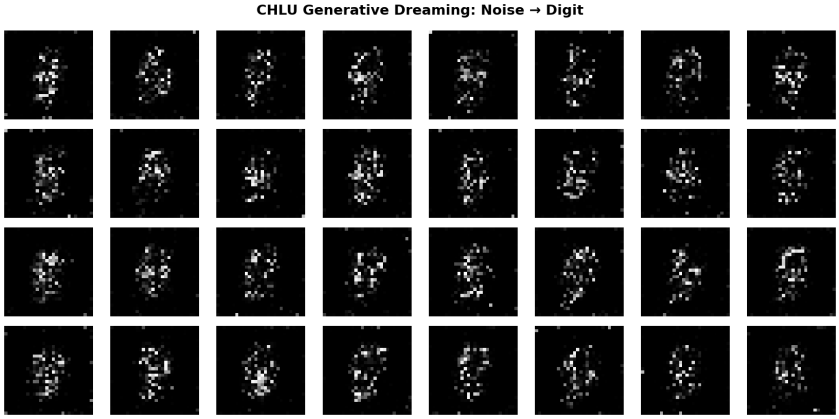

Figure 3: `MNIST` digits generated from the centroid of the test set with added noise.

By treating data points as low-energy wells in $V_\theta(q)$, the CHLU generates samples by annealing the input noise. We observe distinct digit modes, although some digits $(3, 5, 8, 9)$ are generated more often than others. See Appendices for further results from different generative modes.

5 Conclusion: From Learning Dynamics to Enforcing Geometry

Current approaches to temporal modeling suffer from a fundamental trade-off: discrete recurrences (LSTMs) offer expressivity at the cost of stability, while continuous approximations (Neural ODEs) achieve smoothness at the cost of dissipation. We have introduced a CHLU to resolve this dichotomy not by learning better approximation functions, but by enforcing a stricter geometric reality. Our experiments demonstrate that this inductive bias yields distinct, verifiable advantages: topological fidelity, kinetic safety and thermodynamic generation. We envision this primitive serving as a unit in deep, physically consistent networks targeting systems that learn better world models. See Appendices for discussions on future directions. Our software package is available at (Jawahar, 2026).

References

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011. ISBN 1-4200-7942-5.

Miguel A. Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pp. 33–40. PMLR, 2005. ISBN 2640-3498.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations, December 2019. URL `http://arxiv.org/abs/1806.07366`. arXiv:1806.07366 [cs].

Zhengdao Chen, Jianyu Zhang, Martin Arjovsky, and Léon Bottou. Symplectic Recurrent Neural Networks, April 2020. URL `http://arxiv.org/abs/1909.13334`. arXiv:1909.13334 [cs].

Peter Dayan, Geoffrey E. Hinton, Radford M. Neal, and Richard S. Zemel. The Helmholtz Machine. *Neural Computation*, 7(5):889–904, September 1995. ISSN 0899-7667. doi: 10.1162/neco.1995.7.5.889. URL `https://doi.org/10.1162/neco.1995.7.5.889`.

Congyue Deng, Brandon Y. Feng, Cecilia Garraffo, Alan Garbarz, Robin Walters, William T. Freeman, Leonidas Guibas, and Kaiming He. Denoising Hamiltonian Network for Physical Reasoning, March 2025. URL `http://arxiv.org/abs/2503.07596`. arXiv:2503.07596 [cs].

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented Neural ODEs, October 2019. URL http://arxiv.org/abs/1904.01681. arXiv:1904.01681 [stat].

N. Benjamin Erichson, Omri Azencot, Alejandro Queiruga, Liam Hodgkinson, and Michael W. Mahoney. Lipschitz Recurrent Neural Networks, April 2021. URL http://arxiv.org/abs/2006.12070. arXiv:2006.12070 [cs].

G. W. Gibbons and S. N. Solodukhin. The Geometry of Small Causal Diamonds. *Physics Letters B*, 649(4):317–324, June 2007. ISSN 03702693. doi: 10.1016/j.physletb.2007.03.068. URL http://arxiv.org/abs/hep-th/0703098. arXiv:hep-th/0703098.

Shiqi Gong, Qi Meng, Jue Zhang, Huilin Qu, Congqiao Li, Sitian Qian, Weitao Du, Zhi-Ming Ma, and Tie-Yan Liu. An efficient Lorentz equivariant graph neural network for jet tagging. *Journal of High Energy Physics*, 2022(7):30, July 2022. ISSN 1029-8479. doi: 10.1007/JHEP07(2022)030. URL https://doi.org/10.1007/JHEP07(2022)030.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian Neural Networks, September 2019. URL http://arxiv.org/abs/1906.01563. arXiv:1906.01563 [cs].

Stephen Hawking and W. Israel. *General Relativity: an Einstein Centenary Survey*. March 2010. URL https://ui.adsabs.harvard.edu/abs/2010grae.book.....H. Publication Title: General Relativity: an Einstein Centenary Survey ADS Bibcode: 2010grae.book.....H.

Geoffrey E. Hinton, Peter Dayan, Brendan J. Frey, and Radford M. Neal. The "Wake-Sleep" Algorithm for Unsupervised Neural Networks. *Science*, 268(5214):1158–1161, May 1995. doi: 10.1126/science.7761831. URL https://www.science.org/doi/abs/10.1126/science.7761831.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997. ISBN: 0899-7667.

Pratik Jawahar. PRAkTIKal24/CHLU: CHLU: Causal Hamiltonian Learning Unit, March 2026. URL https://zenodo.org/records/18834889.

Nicos S. Martys and Raymond D. Mountain. Velocity Verlet algorithm for dissipative-particle-dynamics-based models of suspensions. *Physical Review E*, 59(3):3733, 1999.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training Recurrent Neural Networks, February 2013. URL http://arxiv.org/abs/1211.5063. arXiv:1211.5063 [cs].

Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent Advances in Recurrent Neural Networks, February 2018. URL http://arxiv.org/abs/1801.01078. arXiv:1801.01078 [cs].

Fathi M. Salem. Recurrent Neural Networks (RNN). In Fathi M. Salem (ed.), *Recurrent Neural Networks: From Simple to Gated Architectures*, pp. 43–67. Springer International Publishing, Cham, 2022. ISBN 978-3-030-89929-5. doi: 10.1007/978-3-030-89929-5_3. URL https://doi.org/10.1007/978-3-030-89929-5_3.

Benjamin Scellier and Yoshua Bengio. Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation, March 2017. URL http://arxiv.org/abs/1602.05179. arXiv:1602.05179 [cs].

Carl Ludwig Siegel. *Symplectic Geometry*. Elsevier, May 2014. ISBN 978-1-4832-6935-1. Google-Books-ID: ObjSBQAAQBAJ.

Paul Smolensky. Information Processing in Dynamical Systems: Foundations of Harmony Theory. February 1986. URL https://apps.dtic.mil/sti/html/tr/ADA620727/. Number: CUCS32186.

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models, December 2025. URL http://arxiv.org/abs/2507.06952. arXiv:2507.06952 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

LeCun Yann. MNIST handwritten digit database. *ATT Labs.*, 2010.

Kuai Yu, Crystal Su, Xiang Liu, Judah Goldfeder, Mingyuan Shao, and Hod Lipson. LyTimeT: Towards Robust and Interpretable State-Variable Discovery. *arXiv preprint arXiv:2510.19716*, 2025.

A Extended Related Work

A.1 Theoretical Foundations: Causal Geometries

Our architectural constraints draw on the holographic principle and the geometry of causal diamonds (Gibbons & Solodukhin, 2007). Unlike standard recurrent units which allow arbitrary information flow, a CHLU restricts latent updates to a strict, learnable light-cone parametrized by c. This ensures that the learned dependencies respect causality, stabilizing the learning of long-range temporal dependencies. We argue that this approach with ad-hoc non-causal data masks is as important as an inherently non-causal algorithm with causal data masks (for eg. transformers trained on causally masked natural language data).

A.2 Advanced Stability: Comparisons to LyTimeT

Recent works in robust state discovery, such as LyTimeT (Yu et al., 2025), utilize Lyapunov regularization terms in the loss function to constrain otherwise unstable latent dynamics. In contrast, CHLU decouples stability from optimization. The Relativistic Kinetic Governor guarantees global boundedness (BIBO stability) by architectural construction, rendering the system immune to divergence. We utilize Lyapunov regularization strictly for topological sculpting of the potential energy surface $V_\theta(q)$ to ensure that valid data modes form deep, asymptotic attractors, rather than to prevent non-physical explosions.

A.3 Generative Lineage: RBMs and Helmholtz Machines

A CHLU shares the thermodynamic DNA of Restricted Boltzmann Machines (RBMs) and Helmholtz Machines, utilizing a "Wake-Sleep" contrastive learning objective. However, where RBMs rely on stochastic Gibbs sampling over static bipartite graphs, a CHLU employs deterministic Symplectic Integration over a continuous phase space. Furthermore, unlike Helmholtz machines which require separate recognition and generative weights, a CHLU unifies inference and generation into a single reversible Hamiltonian operator. Because Hamiltonian dynamics are time-reversible, a CHLU's "recognition" model is strictly the time-reversal of the "generative" model, ensuring geometric consistency between the latent prior and the data likelihood.

B Derivation of Relativistic Gradients

The canonical momentum update requires the gradient of the relativistic kinetic energy $T(\mathbf{p}) = \sqrt{c^2\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} + m_0^2c^4}$. The gradient with respect to momentum $\mathbf{p}$ is derived as:

$$\frac{\partial T}{\partial \mathbf{p}} = \frac{1}{2\sqrt{c^2\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} + m_0^2c^4}} \cdot (2c^2\mathbf{M}^{-1}\mathbf{p}) = \frac{c^2\mathbf{M}^{-1}\mathbf{p}}{\sqrt{c^2\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} + m_0^2c^4}} \tag{10}$$

As $|\mathbf{p}| \to \infty$, the term $m_0^2c^4$ becomes negligible compared to the momentum term. The velocity saturates as:

$$\lim_{|\mathbf{p}|\to\infty} \nabla_p T \approx \frac{c^2\mathbf{M}^{-1}\mathbf{p}}{\sqrt{c^2\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}}} = c\frac{\mathbf{M}^{-1}\mathbf{p}}{\sqrt{\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}}} \tag{11}$$

This confirms that the velocity is bounded by c within the geometry defined by $\mathbf{M}$.

C Further Experimental Results

The full list of hyerparameters used for the paper are provided as default configuration options with our software package released with this work.

C.1 Lemniscate Experiment

We probed the learned potential energy, $V_\theta(q)$ to find interpretable plots, Fig.4, that show the stable attractor we intended to learn.

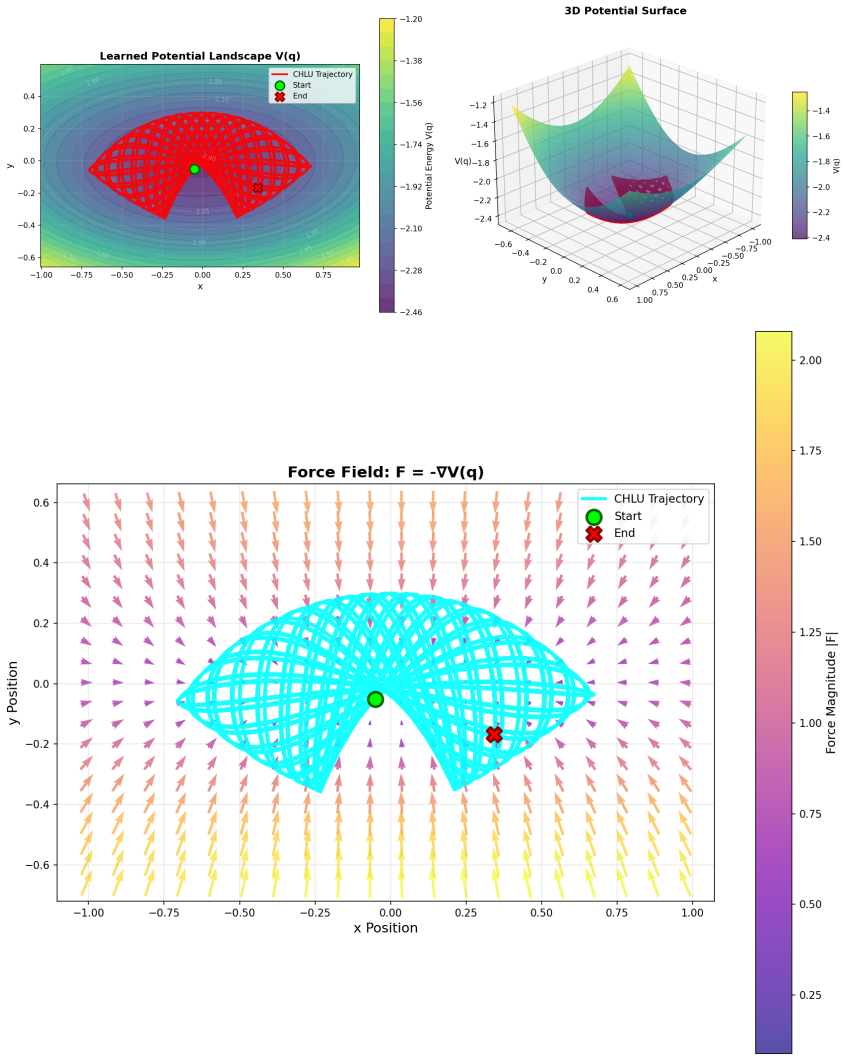

Figure 4: The learned lemniscate potential energy plotted as $2D$, $3D$ heatmaps (top left, right) as well as a force field (bottom). The output trajectory is overlayed on all 3 for visual comparison.

C.2 Perturbed Sine Wave

We showed the kinetic energy calculated from the output state q for all 3 models in Fig.2. Here we show the same plot on the top 3 rows of Fig.5. The bottom 3 rows, however, are the kinetic energies for the same inference steps, but this time calculated using the output state p. Although the LSTM produces near perfect kinetic energy based on p, the LSTM fails to learn the fundamental correlation between $\dot{q}$ and p, and as a result produces 2 different initial kinetic energies for a single system. This further shows that LSTMs are handicapped by design when the goal is to learn physically consistent internal models since its only objective is to minimize next-step prediction error. The CHLU on the other hand collapses to a phase-shifted, magnitude conserved sine wave, since it is inherently biased to conserve learned energy states in the form of stable attractors. We discuss this as a failure mode in App.D. It is important to note that we are not comparing direct performance here, but showcasing the inductive biases of our model in contrast to the baselines.

The phase space plots for the produced sine waves themselves are shown in Fig.6. The LSTM locks onto the perfect loop over time, but the initial few steps from the perturbed conditions show that the model has discontinuous phase space trajectories that are non-physical. The CHLU locks onto the

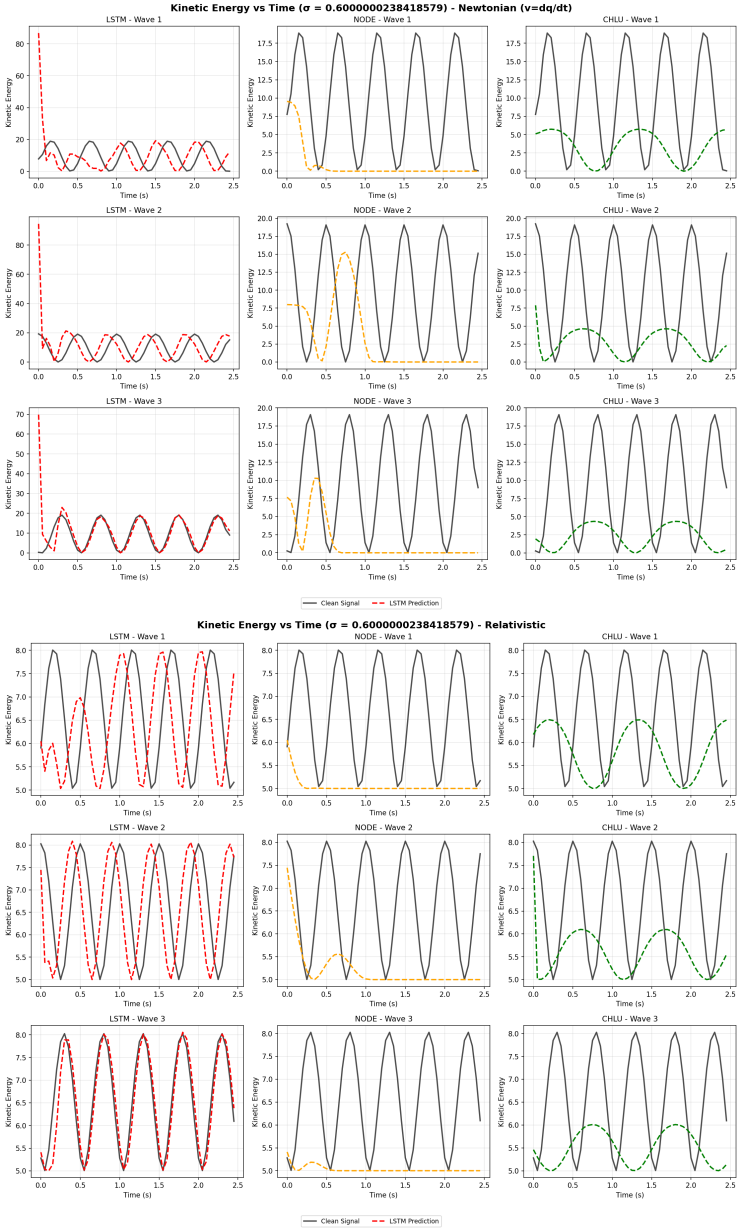

Figure 5: Kinetic energies calculated from the output state q (top 3 rows), compared to the kinetic energies calculated from output state p (bottom 3 rows) for the same 3 sine waves with the same inital perturbed conditions for all models (left to right: LSTM, NODE, CHLU).

nearest sine wave trajectory based on the initial noise dissipation that includes partial information dissipation as well (discussed in App.D).

C.3 MNIST Generation

Fig.3 showed the generation results from the thermal-annealed mode (using and annealing both langevin temperature $\mathcal{T}$, and friction γ). The plot shown included a $tanh()$ function applied to the 1000th generation step for better visualization. In Fig.7 we show the evolution of the first 5 digits through steps $200 - 1000$, every 200 steps.

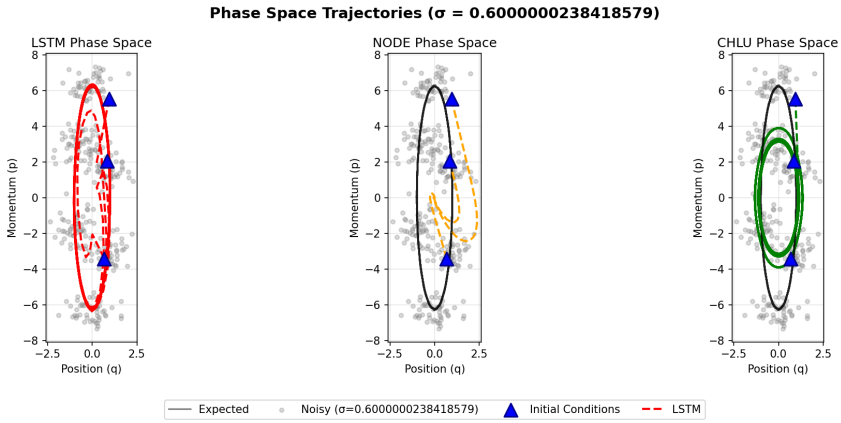

Figure 6: Phase space plots corresponding to the kinetic energy plots in Figs.2,5, with the blue triangles representing the perturbed initial states and the grey ellipse showing the expected trajectory (left to right: LSTM, NODE, CHLU).

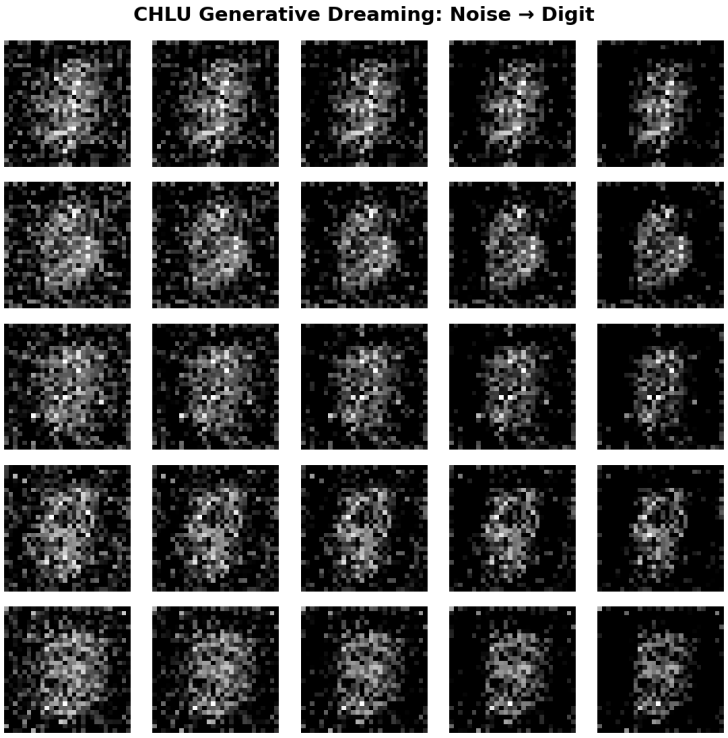

Figure 7: MNIST evolution from step $200 - 1000$ from left to right, at intervals of 200 steps, for 5 digits (rows).

Fig.8 shows a generation following non-thermal, deterministic generation without friction ($\mathcal{T}$ and γ set to 0). Here the system settles into denoised digits faster, but the generation seems to drive all output pixels to the lowest values showing that the most stable state learned by the non-thermal deterministic values is a collapsed low energy representation where all pixels are set to their lowest values. In comparison the stochastic langevin mode causes the system to settle in non-collapsed states showing distinct digits.

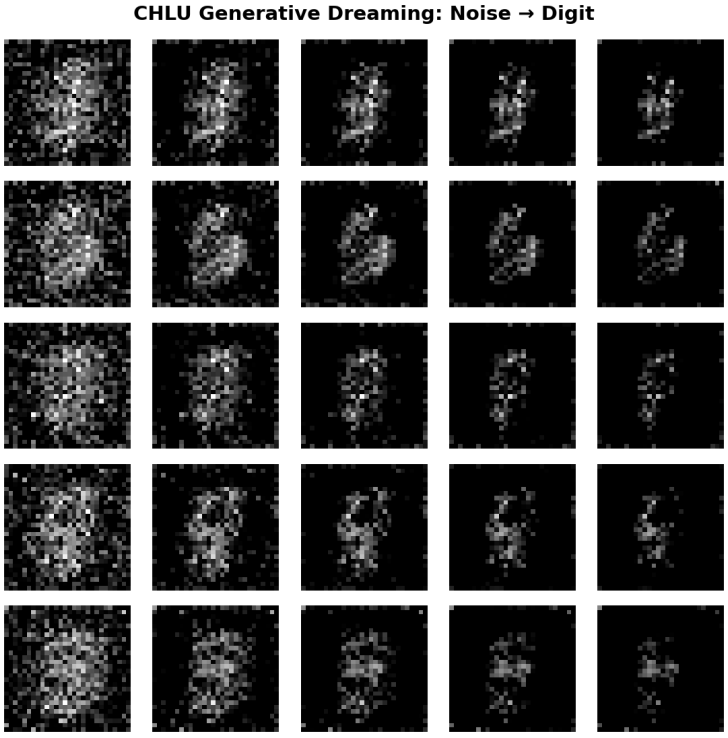

Figure 8: MNIST evolution from step $200 - 1000$ from left to right, at intervals of 200 steps, for 5 digits (rows), in the non-thermal conservative mode.

D LIMITATIONS AND FAILURE MODES:

It is crucial to note that a CHLU's strength, conservation, is also its primary limitation.

1. **Hyper-Stability:** Because a CHLU is strictly conservative, it lacks a natural mechanism to "forget" noise. In purely dissipative systems (e.g., a damped pendulum), a conservative CHLU will oscillate forever unless the friction parameter is used, thereby giving up long-horizon stability. However, we argue that a solution to the memory-stability tradeoff is a combination of a Dissipative CHLU and a Symplectic RNN, giving design flexibility from both ends.
2. **Stiffness:** The Velocity Verlet integrator, while symplectic, can become unstable if the learned potential V_θ becomes extremely stiff (very high curvature), requiring adaptive time-stepping which breaks symplecticity.

E FUTURE DIRECTIONS

Future work will explore "Deep Symplectic Networks," stacking/connecting multiple CHLUs to model complex systems, including a comprehensive study of performance, biases and limitations.

We identify four key areas for future expansion:

- **Lorentz Boosting (Lateral Energy):** A proposed mechanism to resolve ambiguity by applying a "Lorentz Boost" to the latent reference frame as a form of global attention. This goal is to mechanically cluster semantically related features without altering weights during inference.
- **Wormholes:** Implementing sparse non-local connections penalized by energy cost (rather than physical changes to space-time topology) to allow "tunneling" across the causal diamond, facilitating long-range information transfer without $O(N^2)$ complexity. This allows

for ad-hoc breaking of CHLU's internal conservation laws to provide sparse-non local attention, potentially a key requirement to reach comparable performance to modern deep learning algorithms on fundamnetally non-causal tasks.

- **Connected CHLU Netowrks:** Connecting multiple CHLUs to allow each unit to propagate learned information at scale.
- **Hierarchical Causality:** Stacking CHLUs where the "Output Shell" of one unit becomes the "Input Shell" (Boundary Condition) of the next, allowing for multi-scale physical modeling.
- **Shell Jumping:** An Adaptive Computation Time mechanism where the network projects the state to higher-dimensional manifolds ("Shells") when uncertainty is high.

ACKNOWLEDGMENTS

PJ's and MP's work was carried out as part of the SMARTHEP MSCA-ITN. SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086.