Perceived vs. True Emergence: A Cognitive Account of Generalization in Clinical Time Series Models

Shashank Yadav
Department of Biomedical Engineering
University of Arizona
Tucson, AZ 85721
shashank@arizona.edu

Abstract

Understanding how deep learning models form high-level states is a central challenge and the cognitive science of emergence provides a promising framework for interpreting these internal processes. We investigate how a neural network can learn to perceive emergent states from complex clinical time series using an information-theoretic objective Ψ that balances temporal predictability and abstraction. We introduce a framework that distinguishes perceived emergence, a model's ability to identify emergent patterns within its training environment ($\Psi > 0$; In-Distribution), from true emergence, the persistence of these patterns under distribution shift ($\Psi > 0$; Out-Of-Distribution). We evaluate this framework by reciprocal training and verification across two large critical care datasets, MIMIC-IV and eICU, comprising 63 harmonized variables. Our experiments demonstrate that models trained with Ψ capture perceived emergence within their training environments and also exhibit true emergence across datasets, indicating robust generalization. We provide a processing account of this generalization by analyzing the internal mechanics of the learned representations and the stability of their mutual information under distributional shift, thereby contributing to a clearer understanding of how such models may achieve out-of-distribution generalization in clinical settings.

1 Introduction

Expert clinicians possess remarkable cognitive ability when observing a stream of high-dimensional patient data. They perceive abstract, high-level "gestalts" or macro-states like "developing shock", "respiratory fatigue," and "stabilizing" that are critical for making life-saving decisions. This process requires the identification of underlying emergent patterns that govern the patient's trajectory while ignoring a torrent of low-level noise and spurious correlations ("micro-features") (Schuwirth et al., 2020; Feller et al., 2023). Standard deep learning models for time-series, particularly those trained on simple predictive tasks such as forecasting or classification, often fail to develop emergence (Lim and van der Schaar, 2018). Instead, they are biased towards learning the micro-features, which makes them effective short-term forecasters but poor perceivers of the abstract, emergent states that a human expert would identify (Geirhos et al., 2020; DeGrave et al., 2021). This failure to abstract emergent states often leads to brittle models that do not generalize well to new environments.

In this work, we present a learning account of how a neural network trained on time-series can develop the cognitive-like ability to perceive such emergent states in critical care. We hypothesize that this skill is not an inherent property of the architecture, but rather a behavior that develops in response to a specific inductive bias provided by the learning objective. We use a self-supervised, information-theoretic objective, termed Ψ, which formalizes a trade-off between temporal predictability and abstraction (McSharry et al., 2024; Rosas et al., 2020). This objective function acts as a developmental pressure, forcing the model to learn a different kind of generalization, one that prioritizes the discovery

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Interpreting Cognition in Deep Learning Models.

of emergent patterns over fine-grained predictive accuracy. Additionally, we extend the current framework of emergent behavior detection by distinguishing between perceived emergence and true emergence.

- **Perceived Emergence**: A feature that is quantitatively emergent (positive verified Ψ score) within a specific data distribution (the model's "training environment"). This represents a "local rule" that the model has learned to perceive in a familiar context. A model exhibiting only perceived emergence may have simply latched onto a certain set of correlations that are specific to the training data and do not represent a fundamental property of the system.
- **True Emergence**: A feature that remains emergent even when evaluated on a novel, out-of-distribution (OOD) dataset. This represents a fundamental property that the model has successfully generalized. Achieving true emergence suggests the model has learned a robust, underlying principle of the system's behavior even when distributional shifts are present.

We use "emergence" in a specific, information-theoretic sense: a macro-state V is causally emergent when it carries predictive power about its own future beyond what is available from any single micro-feature , as captured by a positive Ψ (Rosas et al., 2020; Mediano et al., 2022). This contrasts with more general uses in AI and complex systems, where "emergence" may describe structured behavior arising across multiple scales. Our analysis therefore treats emergence as an empirically testable property of representations, closely linked to partial information decomposition. We contribute by using this framework to examine whether a model trained to perceive emergence captures dataset-specific correlations or instead identifies generalizable principles of patient dynamics. This approach enables evaluation beyond conventional performance metrics, framing emergence as a principled account of the model's generalization capabilities.

2 Methods

2.1 A Computational Model of Emergence Perception

Our methodology is centered on the Ψ objective, proposed by Rosas et al. (2020), as a computational model for learning to perceive emergent phenomena. Under the ΦID formalism, a simple *sufficient* test for causal emergence is

$$\Psi \;:=\; I(V_t; V_{t+1}) \;-\; \sum_{i=1}^{n} I(X_{i,t}; V_{t+1}) \;>\; 0, \tag{1}$$

where n is the number of input features. Equation 1 depends only on pairwise marginals and standard Shannon mutual information (Rosas et al., 2020; Mediano et al., 2022). Here we take $t' = t + 1$, though any $t' > t$ is valid.

However, the sum in (1) can *double-count* information when multiple inputs share the same signal about V_{t+1}; the redundancy is discounted by using the *Minimum Mutual Information* (MMI) measure (Barrett, 2015). This yields the adjusted sufficiency criterion:

$$\Psi_A \;:=\; I(V_t; V_{t+1}) \;-\; \sum_{i=1}^{n} I(X_{i,t}; V_{t+1}) \;+\; (n-1)\, \min_i I(X_{i,t}; V_{t+1}) \;>\; 0. \tag{2}$$

This objective can be interpreted in cognitive terms as a balance of three distinct components:

- $I(V_t; V_{t+1})$ **(Predictive Power):** This term encourages a stable, self-predictive macro-state V. A high value indicates that the internal representation is informative about its own future, reflecting a coherent, non-random temporal signal (Mediano et al., 2022; Rosas et al., 2020).
- $-\sum_{i=1}^{n} I(X_{i,t}; V_{t+1})$ **(Information Bleed Penalty):** It forces V to be an abstract summary of the whole rather than a proxy for any single micro-input X_i. It penalizes leakage of low-level details into the high-level belief, aligning with the whole-minus-parts structure of Ψ in (1).
- $(n-1)\, \min_i I(X_{i,t}; V_{t+1})$ **(Redundancy Correction):** It offsets the negative bias from double-counting shared (redundant) information across input features. Within the MMI redundancy formulation, the minimum single-source MI counterbalances the excessive subtraction that arises when many X_i carry the same signal—making Ψ_A a more robust sufficient indicator of emergence (Williams and Beer, 2010). This ensures that a truly non-emergent feature will have a Ψ score close to zero or negative, making the final score ($\Psi > 0$) a more reliable indicator of emergence.

2.2 Experimental Setup: Probing for True Emergence

To test for true emergence, we use two distinct datasets, which allows us to create a controlled test of generalization: We use the MIMIC-IV (Johnson et al., 2023) and eICU (Pollard et al., 2018) datasets from critical care. We harmonize both datasets to have the same 63 features (see Table 2) using the METRE Pipeline (Liao and Voldman, 2023). We train a feature network on one dataset and verify emergence on both. To ensure our findings are not dependent on the choice of training set, we perform a reciprocal analysis with two experiments:

1. **Experiment 1:** Train a model on the MIMIC-IV dataset and run two separate verification phases: one on the MIMIC dataset itself(a test for perceived emergence) and one on the eICU dataset (to test for true emergence).
2. **Experiment 2:** Train a second model on the eICU dataset. Then, run two verification phases: one on the eICU dataset itself (a test for perceived emergence) and one on the MIMIC-IV dataset to test for true emergence.

This reciprocal design allows us to rigorously assess whether any learned emergent feature is a universal property or an artifact of a specific training environment. For a detailed breakdown of the training and verification procedure, please see the Appendix. A key part of our methodology is the evaluation of the learned representation on out-of-distribution data using a metric we term $\Psi_{\text{A}}^{\text{OOD}}$. While the f_θ network remains frozen, the critics required to estimate the mutual information terms in $\Psi_{\text{A}}^{\text{OOD}}$ are trained on the OOD data during verification. This ensures that the verification is not biased by critics learned from the training environment and by learning new critics, we can provide an unbiased assessment of whether the representation learned by the f_θ network generalizes to a novel data distribution. This setting constitutes a genuine test of "true emergence," as neither the feature network nor the evaluation mechanism is biased towards the training data.

3 Results

3.1 Evaluation of distributional shift between MIMIC-IV and eICU

We conducted an out-of-distribution (OOD) detection experiment to evaluate the degree of distributional shift between our two datasets. Using a standard scoring-based OOD framework (Fang et al., 2024), we obtained an AUC-ROC of 0.80 (with Gaussian kernel width $\gamma = 0.08$ and $M = 512$ random Fourier features), indicating that the two datasets are moderately separable in distributional space (Figure 1). This observation is consistent with prior work that compared the feature distributions of seven critical care time series datasets and demonstrated that MIMIC-IV and eICU datasets exhibit substantial distributional shifts (Burger et al., 2024).

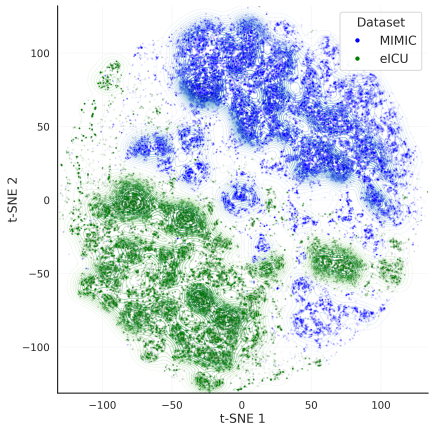
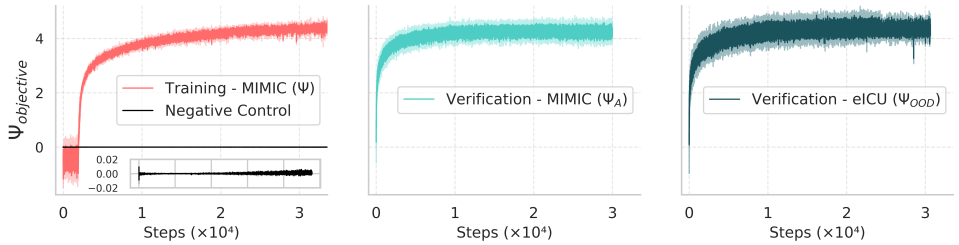


Figure 1: OOD visualization using t-sne embeddings of specific timesteps. Evaluation of mutual Out-Of-Distribution characteristic between MIMIC-IV and eICU datasets using the same feature set using kernel-PCA (Fang et al., 2024).

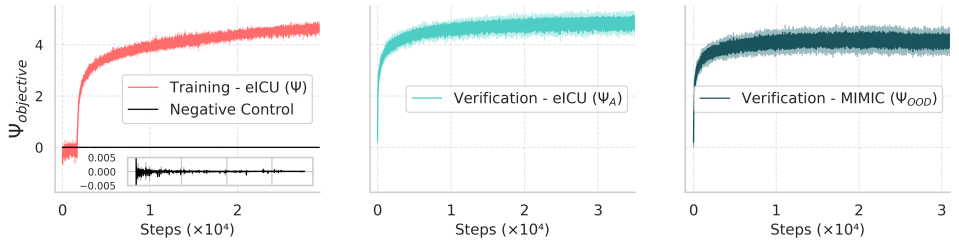
3.2 Perceived vs. True emergence

We first trained our model (f_θ) on the MIMIC-IV dataset (Training $\Psi_A = 4.45 \pm 0.07$) and confirmed that it learned a feature with a high positive verified Ψ score (Verification $\Psi_A = 4.27 \pm 0.15$). This demonstrates that it had successfully learned perceived emergence within its training environment. We then evaluated this same frozen model (f_θ) on the eICU dataset to test the model for true emergence and observed that the model is successful (Verification $\Psi_{\text{A}}^{\text{OOD}} = 4.42 \pm 0.13$) (Figure 2a). We repeated this with the reciprocal experiment. We first trained our model (f_θ) on the eICU dataset (Training $\Psi_A = 5.01 \pm 0.06$) and confirmed it learned an emergent feature with a high positive verified Ψ score (Verification $\Psi_A = 4.87 \pm 0.12$), demonstrating that it had successfully learned perceived emergence within its training environment. We then evaluated this same frozen model (f_θ) on the MIMIC-IV dataset to test for generalization and the model successfully learns true emergence (Verification $\Psi_{\text{A}}^{\text{OOD}} = 4.19 \pm 0.17$), illustrated in

Figure 2b. We conducted five independent runs per experiment and summarize performance as mean ± standard deviation (see Table 1). We also performed negative control experiments with a randomly shuffled version of MIMIC-IV and eICU where the temporal order of the data within each patient stay was destroyed. As expected, the training Ψ_A for models trained on this shuffled data was consistently near-zero or negative. This result demonstrates that the emergent feature is a genuine property of the system's dynamics and not merely a statistical artifact of the data distribution.



(a)



(b)

Figure 2: (a) Comparison of Ψ_A for training, verification on MIMIC-IV dataset and OOD verification on the eICU dataset. (b) Comparison of Ψ_A for training, verification on eICU dataset and OOD verification on the MIMIC-IV dataset.

Table 1: Verified Ψ scores from the reciprocal analysis on MIMIC-IV and eICU datasets.

Training	**Verification**	**Emergence Type**	Training ($\boldsymbol{\Psi_A}$)	**Verification ($\boldsymbol{\Psi_A}$ and $\Psi_{\mathbf{A}}^{\mathbf{OOD}}$)**
MIMIC-IV	MIMIC-IV eICU	Perceived True	4.45 ± 0.07	4.27 ± 0.15 4.42 ± 0.13
eICU	eICU MIMIC-IV	Perceived True	5.01 ± 0.06	4.87 ± 0.12 4.19 ± 0.17

3.3 A processing account of cognition: Analysis of Mutual Information

We extracted the final downward Mutual Information scores $I(X_{i,t}; V_{t+1})$ from both the ID and OOD verification runs for both experiments, which represents the contribution of each of the 63 input features to the emergent state. This allows us to probe the model's internal "reasoning" process. As shown in Figure 3 (refer Appendix), the model relies on a highly consistent set of informational cues in both environments. The Spearman's rank correlation between the feature contribution scores was high (MIMIC $\rho = 0.808$, and eICU $\rho = 0.887$), indicating that the model has learned a robust and generalizable algorithm for perceiving this emergent state.

4 Conclusion

Our results suggest that neural networks trained with the Ψ objective may capture invariant properties of patient dynamics in critical care time series that generalize across datasets. The consistency of true emergence indicates the possibility that such models are not merely tuned to dataset-specific signals but may be uncovering principles of physiological organization that extend across environments, pointing toward a pathway for models that reveal robust laws of complex systems. Such findings raise the prospect of developing models that move beyond performance metrics to identify deeper structures within patient trajectories and hold potential for a more principled understanding of emergent temporal dynamics in critical care. Future work may compare Ψ to alternative objectives such as InfoNCE (Oord et al., 2018) and VICReg (Bardes et al., 2021) using matched critics to assess which inductive biases most effectively capture true emergence.

References

Bardes, A., Ponce, J., and LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.

Barrett, A. B. (2015). Exploration of synergistic and redundant information sharing in static and dynamical gaussian systems. *Physical Review E*, 91(5):052802.

Burger, M., Sergeev, F., Londschien, M., Chopard, D., Yèche, H., Gerdes, E., Leshetkina, P., Morgenroth, A., Babür, Z., Bogojeska, J., et al. (2024). Towards foundation models for critical care time series. *arXiv preprint arXiv:2411.16346*.

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. (2021). Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619.

Fang, K., Tao, Q., Lv, K., He, M., Huang, X., and Yang, J. (2024). Kernel pca for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:134317–134344.

Feller, S., Feller, L., Bhayat, A., Feller, G., Khammissa, R. A. G., and Vally, Z. I. (2023). Situational awareness in the context of clinical practice. In *Healthcare*, volume 11, page 3098. MDPI.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Liao, W. and Voldman, J. (2023). A multidatabase extraction pipeline (metre) for facile cross validation in critical care research. *Journal of Biomedical Informatics*, 141:104356.

Lim, B. and van der Schaar, M. (2018). Disease-atlas: Navigating disease trajectories using deep learning. In *Machine Learning for Healthcare Conference*, pages 137–160. PMLR.

McSharry, D., Kaplanis, C., Rosas, F., and Mediano, P. A. (2024). Learning diverse causally emergent representations from time series data. *Advances in Neural Information Processing Systems*, 37:119547–119572.

Mediano, P. A., Rosas, F. E., Luppi, A. I., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., and Bor, D. (2022). Greater than the parts: a review of the information decomposition approach to causal emergence. *Philosophical Transactions of the Royal Society A*, 380(2227):20210246.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. (2018). The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13.

Rosas, F. E., Mediano, P. A., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., and Bor, D. (2020). Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLoS computational biology*, 16(12):e1008289.

Schuwirth, L. W., Durning, S. J., and King, S. M. (2020). Assessment of clinical reasoning: three evolutions of thought. *Diagnosis*, 7(3):191–196.

Williams, P. L. and Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

A Appendix

A.1 The Role of the Feature Network and Critic Models

The framework consists of two types of models that learn together:

- **The Feature Network (f_θ):** This is the primary model, a *SkipConnectionSupervenientFeatureNetwork* (McSharry et al., 2024), whose goal is to learn the emergent representation.
- **The Critic Models:** These are essential "helper" models that act as verification test for the emergent property. There are two types:
  1. **The Decoupled Critic:** A single *DecoupledSmileMIEstimator* (McSharry et al., 2024) that estimates the "Predictive Power" ($I(V_t; V_{t+1})$).
  2. **The Downward Critics:** A set of n *DownwardSmileMIEstimators* (one for each of the 'n' input features) that estimate the "Information Bleed" ($I(X_{i,t}; V_{t+1})$).

A.2 The Training and Verification Procedure in Detail

The distinction between the training and verification phases is critical for obtaining a robust and unbiased measure of emergence.

- **Training Phase:** In this phase, the feature network and all $n + 1$ critic models are trained simultaneously. For each batch of data, the critics are first updated to become more accurate estimators for the feature network's current output. Then, the feature network is updated to maximize the Ψ score provided by these just-updated critics. This process is repeated for many epochs, allowing both the feature network and critic models to learn together.
- **Verification Phase:** The feature network, having completed its learning, is frozen and its weights are not updated. The critics trained in the training phase are discarded and a fresh set of critic models are trained from scratch and their sole purpose is to converge to the most accurate possible estimate of the mutual information. The final, stable Ψ score from this verification run is the true, unbiased measure of the learned feature's emergence. This two-phase procedure ensures that the final score is a property of the learned feature itself, not an artifact of the co-adaptive training dynamics.

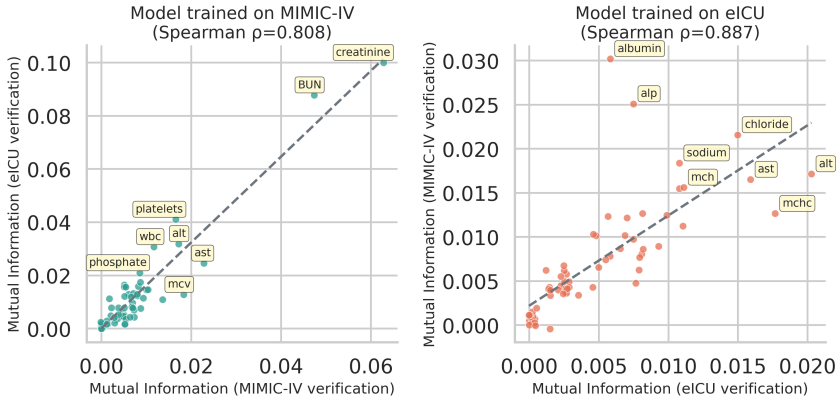


Figure 3: Comparison of feature contribution scores (downward Mutual Information) for the in-distribution vs. out-of-distribution test sets. Each point represents one of the 63 clinical variables. The high Spearman correlation indicates a stable set of learned mutual information scores, suggesting a robust internal algorithm.

A.3 Common clinical features used in the analysis across MIMIC-IV and eICU

Table 2: Common clinical features

Peripheral oxygen saturation (SpO_2)	Arterial partial pressure of CO_2 ($PaCO_2$)
Blood pH	Base excess
Bicarbonate	Total carbon dioxide (CO_2)
Hematocrit	Hemoglobin
Chloride	Temperature
Potassium	Sodium
Lactate	Glucose
Heart rate	Invasive systolic blood pressure
Invasive diastolic blood pressure	Invasive mean blood pressure
Non-invasive systolic blood pressure	Non-invasive diastolic blood pressure
Non-invasive mean blood pressure	Respiratory rate
White blood cell count (WBC)	Basophils
Eosinophils	Lymphocytes
Monocytes	Polymorphonuclear leukocytes (Neutrophils)
Band neutrophils	Troponin T
Creatine phosphokinase–MB (CK-MB)	Albumin
Total protein	Anion gap
Blood urea nitrogen (BUN)	Calcium
Creatinine	Fibrinogen
International normalized ratio (INR)	Prothrombin time (PT)
Partial thromboplastin time (PTT)	Mean corpuscular hemoglobin (MCH)
Mean corpuscular hemoglobin concentration (MCHC)	Mean corpuscular volume (MCV)
Platelet count	Red blood cell count (RBC)
Red cell distribution width (RDW)	Alanine aminotransferase (ALT)
Alkaline phosphatase (ALP)	Aspartate aminotransferase (AST)
Amylase	Bilirubin
Creatine phosphokinase (CPK)	Glasgow Coma Scale (GCS)
C-reactive protein (CRP)	Weight
Urine output	Central venous pressure (CVP)
Urine creatinine	Magnesium
Phosphate	Tidal volume (observed)
White blood cells in urine	