Neural Counterfactual Reasoning for Interacting Systems: Bridging Physics-Informed Learning and Reasoning for PHM

Amaury Wei[1], Prof. Olga Fink[2]

[1,2] *EPFL – Intelligent Maintenance and Operations Systems (IMOS) Laboratory*
CH-1015 Lausanne
Switzerland

amaury.wei@epfl.ch
olga.fink@epfl.ch

ABSTRACT

Over the past decade, advances in sensing and information technologies have enabled industries to collect large amounts of data. Yet, decision-making often remains driven by the intuition of domain experts who rely on simplistic analyses and short-term considerations. This frequently leads to suboptimal decisions that fail to account for long-term effects, particularly in complex, interconnected systems. Current data-driven strategies typically focus on immediate objectives, overlooking relational structures and longer-term impacts. There is a growing need for more transparent, generalizable models that can simulate system behavior, reason about alternative future scenarios, and extrapolate to unseen conditions—capabilities that are essential for decision-making in Prognostics and Health Management (PHM). This research aims to advance reasoning and decision support in PHM through three novel contributions: (1) a physics-informed surrogate model for simulating rigid body interactions, enabling the exploration of "what-if" scenarios, (2) an object-centric visual reasoning model for dynamics prediction in sensor-limited environments, supporting visual inspection tasks, and (3) a neuro-symbolic framework for interpretable root-cause analysis in time series, improving diagnostic transparency and providing actionable insights.

1. MOTIVATION AND PROBLEM STATEMENT

Advances in information technology have enabled industrial systems—such as wind farms, manufacturing lines, and transporation networks—to collect large amounts of multimodal data from sensors, cameras, and operational logs (Bousdekis, Lepenioti, Apostolou, & Mentzas, 2021). Yet, decision-making in PHM still heavily depends on expert intuition and rule-based systems, often leading to suboptimal decisions focused on short-term objectives (Ma, Ren, Xiang, & Turk, 2020). For example, immediate aircraft maintenance may resolve a fault but cause future disruptions in logistics (Xu, Adler, Wandelt, & Sun, 2024).

Traditional decision-making approaches struggle with complex, interconnected systems, particularly under unseen or evolving conditions. Expert reasoning is biased by prior experience, and synthesizing diverse data streams remains difficult (Sarker, 2021). Furthermore, formalizing expert knowledge into intepretable reusable rules has proven challenging, limiting scalability and transparency (Idé, 2016).

While recent data-driven methods aim to optimize decisions, they often neglect long-term implications and fail to model alternative futures. For example, predictive maintenance models may determine the optimal intervention time but cannot simulate how degradation will progress under different scenarios (Pinciroli, Baraldi, & Zio, 2023). The difficulty to extrapolate and the lack of relational modeling both limit decision support in PHM.

To address these challenges, there is a need for generalizable, interpretable models that can simulate system behavior, reason about hypothetical scenarios, and support robust decisions. These capabilities lie at the core of counterfactual reasoning (Verma, Dickerson, & Hines, 2020), which enables answering "What if?" questions such as "What if we delay maintenance?" or "What if we adjust operating conditions?". By combining simulation, forecasting, and explainability, counterfactual reasoning offers a path toward more strategic, future-proof PHM.

To address these needs, this thesis explores three complementary aspects of counterfactual reasoning in PHM. First, it introduces a physics-informed surrogate model for simulating rigid body interactions, supporting "what-if" analyses in applications such as robotic manipulation and asset handling, where accurate long-term dynamics are critical. Second, it proposes an object-centric visual reasoning framework that predicts object trajectories directly from raw video, enabling anticipatory decisions in sensor-limited environments like mobile inspection or autonomous navigation. Third, it presents a neuro-symbolic approach for interpretable root-cause analysis in time series data, improving transparency in diagnostics and supporting decision-making in complex, multi-sensor systems such as wind farms or energy networks.

2. EXPECTED CONTRIBUTIONS

2.1. Physics-Informed Surrogate Model

The first contribution of this research is the development of a physics-informed surrogate model for simulating rigid body dynamics in interacting environments. This model enables accurate long-horizon predictions of physical interactions such as collisions and group-level motion—capabilities especially valuable for PHM tasks involving robotic manipulation and simulation-based planning.

Traditional physics engines can produce accurate predictions when all environmental parameters are fully specified. However, in real-world scenarios, many factors—such as wind conditions, surface texture, or material damping—are unknown or too complex to model explicitly (Parmar, Halm, & Posa, 2021). These limitations highlight the need for data-driven surrogate models that can learn from observed behavior and generalize to unseen conditions.

Recent learning-based approaches (Sanchez-Gonzalez et al., 2020; Pfaff, Fortunato, Sanchez-Gonzalez, & Battaglia, 2021; Allen et al., 2023) have addressed this by using Graph Neural Networks (GNNs) to model interactions. Yet, these methods are limited in scalability and physical realism as they rely on simple node-to-node message passing (Battaglia et al., 2018), which struggles to capture higher-level dynamics such as momentum exchange or energy transfer.

To overcome these limitations, we propose a through a surrogate model based on higher-order topological representations. Rather than limiting interactions to pairs of nodes, we encode object surfaces and structural groupings using combinatorial complexes (Bodnar, 2022; Hajij et al., 2022), which preserve object cohesion and enable modeling of group-level dynamics. This structure also allows environmental modifications, making it well-suited for simulating counterfactual scenarios.

Building on this representation, we introduce a physics-informed message-passing framework that enforces Newtonian principles. By structuring message flows to reflect energy and momentum transfer between the physical hierarchies (e.g., from local surface-level contacts to whole-object dynamics), the model achieves both long-horizon predictive accuracy and physical consistency. These inductive biases also improve learning efficiency and generalization to out-of-distribution scenarios.

The resulting model, named HOPNet, outperforms state-of-the-art methods in simulating object motion under collisions. It enables "what-if" simulations by modifying object count, properties, or trajectories—supporting decision-making for PHM tasks in robotic asset handling. Beyond rigid body dynamics, our proposed method also offers a foundation for modeling high-level interactions in other domains, such as sensor networks or interconnected systems.

2.2. Unsupervised Vision-based Reasoning

The second contribution of this research focuses on improving predictive reasoning in vision-based systems. Many autonomous agents—such as mobile robots or inspection drones—rely on visual data, yet lack access to precise 3D geometry or physical parameters. In these settings, learning directly from video is essential to support predictive maintenance, degradation tracking, or failure anticipation.

Recent methods for unsupervised object-centric scene decomposition use attention-based autoencoders to segment scenes into discrete object "slots" (Kipf et al., 2022; Singh, Wu, & Ahn, 2022; Wu, Dvornik, Greff, Kipf, & Garg, 2023; Zadaianchuk, Seitzer, & Martius, 2023; Majellaro, Collu, Plaat, & Moerland, 2025). While promising, these approaches often fail to distinguish visually similar objects, particularly under grayscale imagery, occlusion, or non-planar motion. The resulting segmentations are unstable, leading to inaccurate or invalid dynamics predictions. Furthermore, current architectures lack inductive biases needed to respect physical principles, causing performance to degrade in realistic PHM scenarios with clutter, collisions, and long-horizon forecasting.

To address these limitations, we propose a visual reasoning framework that improves both object decomposition and predictive modeling. First, we propose to use equivariant convolutional filters (T. Cohen & Welling, 2016; T. S. Cohen, Geiger, Köhler, & Welling, 2018) to disentangle spatial location from appearance, improving object consistency and robustness to viewpoint changes. We also investigate regularization techniques, such as Kullback-Leibler divergence and feature-space separation, to encourage distinct object embeddings—even in visually ambiguous settings.

Second, we extend the model to capture temporal dynamics by integrating scene history across video frames. This enables consistent tracking of partially or fully occluded objects—critical for reasoning about collisions, interaction sequences, and downstream effects. By explicitly extracting geometric trajectories through improved decomposition, the model supports accurate prediction of rigid body evolution over time, even in unseen environments.

The resulting model enables counterfactual visual simulation, answering questions such as "What if this object were removed?" or "What if its initial position was different?", and serves as a surrogate for dynamics forecasting in visually monitored systems. We will benchmark our approach against state-of-the-art models on datasets ranging from planar motion to 3D collisions, evaluating prediction accuracy, generalization, and robustness. This contribution supports PHM tasks such as visual degradation tracking, remote inspection, and video-based anomaly forecasting in complex scenes.

2.3. Neuro-Symbolic Reasoning for Root Cause Analysis

The third contribution of this research addresses the need for interpretable diagnostics in complex industrial systems through a neuro-symbolic framework for time series analysis. While deep learning models such as transformers (Ansari et al., 2024; Gu & Dao, 2023) offer state-of-the-art performance for forecasting and classification, their predictions remain opaque, hindering trust and deployment in safety-critical PHM applications. Post-hoc explainability methods provide limited insight in time series settings, where temporal and multivariate dependencies make decisions difficult to trace.

To overcome these challenges, we propose a neuro-symbolic approach that extracts human-understandable concepts from raw time series and composes them into interpretable logical rules. Inspired by advances in concept-based reasoning (Mao, Gan, Kohli, Tenenbaum, & Wu, 2020; Koh et al., 2020), we aim to identify high-level temporal patterns—such as "time above threshold," "oscillation frequency," or "exponential decay"—and combine them using logic operators (e.g., conjunction $\wedge$, negation $\neg$) to explain anomalies or trends. These rules can support root cause analysis, condition monitoring, and transparent fault detection in PHM systems.

This framework will be implemented as an end-to-end differentiable architecture, capable of unsupervised concept discovery from multivariate inputs. To evaluate its performance, we will benchmark against existing fault detection methods using real-world datasets from wind farms and hydropower plants. Since these datasets are often sparsely labeled and heavily imbalanced, we will also consider developing synthetic, fault-annotated datasets using realistic simulators to validate rule extraction and generalization.

By linking predictive power with symbolic reasoning, our model will enable transparent, causal explanations—a critical requirement for actionable decision support in PHM. In addition, the symbolic rules generated by the model can be reused across systems or scenarios, facilitating knowledge transfer. Ultimately, this contribution bridges the gap between black-box learning and human-interpretable diagnostics, enabling more robust and trustworthy PHM solutions.

3. PROPOSED RESEARCH PLAN

3.1. Timeline

This PhD research began in January 2024. As of August 2025, the first contribution is complete and has been published in Nature Communications 16 (DOI: 10.1038/s41467-025-62250-7). Work on the second is underway, with the final contribution scheduled to begin later this year, in line with a three-year timeline (Fig. 1).

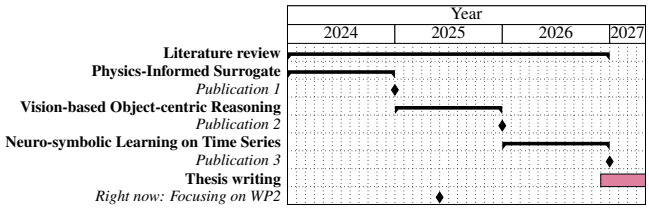

Figure 1. Research timeline

3.2. Current Progress

We now present core results from the first contribution: a physics-informed surrogate model for simulating rigid body dynamics in interacting systems. The model, HOPNet, combines topological representations with physics-aware message passing to support long-horizon, counterfactual simulations.

We introduce a physics-informed topological representation of rigid bodies using Combinatorial Complexes (CCs). Unlike standard mesh-based graphs that model only node connections, CCs allow us to define hierarchical cell types—nodes, edges, triangles, contacts, and objects—to preserve both surface consistency and object cohesion. Each cell carries features at its level (e.g., velocity for nodes, mass for objects), enabling structured, physically grounded modeling. The system evolves as a sequence of spatiotemporal complexes $\mathcal{X}^t$, with future states $\hat{\mathcal{X}}^{t+1}$ predicted autoregressively from previous ones. This representation allows accurate, long-horizon simulation of rigid body interactions, supporting counterfactual reasoning in complex PHM scenarios.

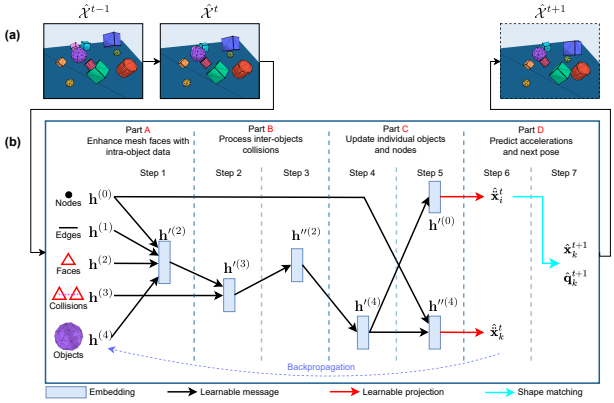

Figure 2. Overview of our method (**a**) Autoregressive roll-out approach; (**b**) Physics-informed message-passing strategy. Our sequential message-passing is inspired by Newtonian laws and tailored to process collisions.

Operating on these novel representations, we introduce a physics-informed message-passing framework that encodes Newtonian laws directly into the model (Fig. 2). Within the combinatorial complex, messages flow hierarchically between nodes, surfaces, and objects, guided by physical structure. The framework distinguishes between independent and colliding objects: in collisions, energy and momentum changes are computed at contact cells and propagated to update object states

(steps 1-4). Features are aggregated across levels to enrich representations, enabling accurate prediction of per-node and per-object accelerations (steps 5-6). These are integrated using a second-order Euler method to simulate future states. This structured approach improves physical consistency, learning efficiency, and long-horizon rollout accuracy.

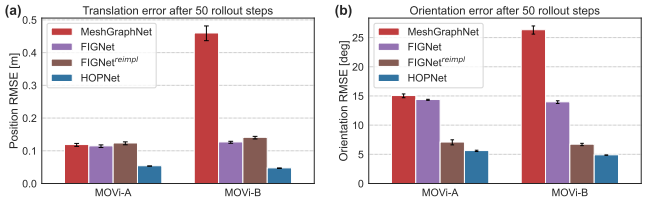

Figure 3. Autoregressive rollout performance on benchmark datasets. (**a**) Translation and (**b**) orientation errors after a rollout horizon of $T = 50$ on all dynamic objects in the scene. Error bars indicate the mean and standard deviation across three independent random seeds.

We evaluate our framework on three benchmark datasets (`MoVi-spheres`, `MoVi-A`, `MoVi-B`) involving rigid body interactions under various physical conditions (Fig. 3). Our model significantly outperforms state-of-the-art baselines MeshGraphNet (Pfaff et al., 2021) and FIGNet (Allen et al., 2023) in long-horizon prediction accuracy, achieving up to 50% longer rollouts before reaching comparable errors. This improvement enables more reliable and physically consistent simulations of dynamic systems.

Importantly, our topological representation—allowing full control over scene parameters—combined with long-term predictive accuracy enables counterfactual simulation of alternative scenarios. This makes our approach well-suited for PHM tasks such as predictive planning and decision-making in systems with multiple interacting components.

ACKNOWLEDGEMENTS

This work is supported by the Swiss National Science Foundation (SNSF) Grant Number 200021_200461.

REFERENCES

Allen, K. R., Rubanova, Y., Lopez-Guevara, T., Whitney, W., Sanchez-Gonzalez, A., Battaglia, P. W., & Pfaff, T. (2023). Learning rigid dynamics with face interaction graph networks. In *International conference on learning representations.*

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., … Wang, Y. (2024). Chronos: Learning the language of time series. *Transactions on Machine Learning Research.*

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., … Pascanu, R. (2018). *Relational inductive biases, deep learning, and graph networks.*

Bodnar, C. (2022). *Topological deep learning: Graphs, complexes, sheaves* (Doctoral dissertation, Apollo - University of Cambridge Repository). doi: 10.17863/CAM.97212

Bousdekis, A., Lepenioti, K., Apostolou, D., & Mentzas, G. (2021). A review of data-driven decision-making methods for industry 4.0 maintenance applications. *Electronics*, *10*(7), 828.

Cohen, T., & Welling, M. (2016). Group equivariant convolutional networks. In *International conference on machine learning.*

Cohen, T. S., Geiger, M., Köhler, J., & Welling, M. (2018). Spherical cnns. In *International conference on learning representations.*

Gu, A., & Dao, T. (2023). *Mamba: Linear-time sequence modeling with selective state spaces.*

Hajij, M., Zamzmi, G., Papamarkou, T., Miolane, N., Guzmán-Sáenz, A., Ramamurthy, K. N., … others (2022). *Topological deep learning: Going beyond graph data.*

Idé, T. (2016). Formalizing expert knowledge through machine learning. In *Global perspectives on service science: Japan* (pp. 157–175). Springer New York.

Kipf, T., Elsayed, G. F., Mahendran, A., Stone, A., Sabour, S., Heigold, G., … Greff, K. (2022). Conditional Object-Centric Learning from Video. In *International conference on learning representations.*

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In *International conference on machine learning.*

Ma, Z., Ren, Y., Xiang, X., & Turk, Z. (2020). Data-driven decision-making for equipment maintenance. *Automation in Construction*, *112*, 103103.

Majellaro, R., Collu, J., Plaat, A., & Moerland, T. M. (2025). Explicitly disentangled representations in object-centric learning. *Transactions on Machine Learning Research.*

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2020). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International conference on learning representations.*

Parmar, M., Halm, M., & Posa, M. (2021). Fundamental challenges in deep learning for stiff contact dynamics. In *2021 ieee/rsj international conference on intelligent robots and systems (iros).*

Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., & Battaglia, P. W. (2021). Learning mesh-based simulation with graph networks. In *International conference on learning representations.*

Pinciroli, L., Baraldi, P., & Zio, E. (2023). Maintenance optimization in industry 4.0. *Reliability Engineering & System Safety*, *234*, 109204.

Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R.,

Leskovec, J., & Battaglia, P. (2020). Learning to simulate complex physics with graph networks. In *Proceedings of the 37th international conference on machine learning.* PMLR.

Sarker, I. H. (2021). Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*, *2*(5), 377.

Singh, G., Wu, Y.-F., & Ahn, S. (2022). Simple unsupervised object-centric learning for complex and naturalistic videos. In *Advances in neural information processing systems* (Vol. 35, pp. 18181–18196).

Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. In *Workshop on ml retrospectives, surveys & meta-analyses.*

Wu, Z., Dvornik, N., Greff, K., Kipf, T., & Garg, A. (2023). Slotformer: Unsupervised visual dynamics simulation with object-centric models. In *International conference on learning representations.*

Xu, Y., Adler, N., Wandelt, S., & Sun, X. (2024). Competitive integrated airline schedule design and fleet assignment. *European Journal of Operational Research*, *314*(1).

Zadaianchuk, A., Seitzer, M., & Martius, G. (2023). Object-centric learning for real-world videos by predicting temporal feature similarities. In *Advances in neural information processing systems.*