

**PROCEEDINGS OF THE
33RD MINISYMPOSIUM**

OF THE

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE
AND SYSTEMS ENGINEERING**

BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS

FEBRUARY 9–10, 2026

BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS



**BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
DEPARTMENT OF ARTIFICIAL INTELLIGENCE
AND SYSTEMS ENGINEERING**

© 2026 Department of Artificial Intelligence and Systems Engineering,
Budapest University of Technology and Economics.
For personal use only – unauthorized copying is prohibited.

ISBN 978-615-112-044-6

Head of Department:
László Gönczy

General Chairs¹:
Kristóf Marussy
Oszkár Semeráth

Homepage of the Conference:
<http://minisy.mit.bme.hu/>

Sponsored by:
Schnell László Foundation

¹The editors would like to thank the reviewers for their contribution: Péter Antal, Balázs Bank, András Gézi, Gábor Hullám, Imre Kocsis, Zsolt Kollár, Tamás Kováczházy, István Majzik, Kristóf Marussy, György Orosz, Oszkár Semeráth, László Sujbert, and András Vörös.

Contents

Nada Akel, Oszkár Semeráth, and László Gönczy: Towards Automated Data Invariants Synthesis and Validation	1
Kawthar Alghourani and Kristof Csorba: Toward Reliable 3D Understanding of Steel Fibers in CT: A Vision for Hybrid Manual–Automated Annotation, Tracking, and Validation	5
Ákos Dúcz: Computer-search-based Proof Methods in Geometry and Combinatorics	11
Bertalan Lichter and Dániel Hadházi: Towards Differentiable Machining: A Neural-Differential Framework for Physics-Aware Toolpath Optimization	15
Márk Marosi and Péter Antal: Positional Regularities in Canonical SMILES: Implications for Molecular Machine Learning	21
Mátyás Antal, András Gézsi, and Péter Antal: Benchmarking LLMs as Sequential Diagnosticians: An Interactive Bayesian Environment	27
Mohaisen Mohammed, Dániel Hadházi, and Gábor Hullám: Beyond Dice: Strategy-Dependent Acquisition in Active Learning for JSRT Lung Segmentation	33
Benedek Pour, Huba Németh, and Tamás Dabóczy: Model-Based Artificial Intelligence Approach for Speed and Position Estimation of PMSM	39
Gábor Révy, Dániel Hadházi, and Gábor Hullám: Towards Semiautomatic Pulmonary Artery-Vein Separation in CT Images	44
Ronald Rádai and Tamás Kovácsházy: Beyond Ranging: Real-time Wireless Communication Utilising Wi-Fi Fine Timing Measurement	48
Casi Setianingsih and Bálint Hartmann: Time Segmented Deep Forecasting of Power Grid Frequency Deviations	54
Balázs Ádám Toldi, András Földvári, András Pataricza, and Zsigmond Papp: AI-Supported Performance Analysis	60
Mihály Vetró and Gábor Hullám: Detecting Discrete Temporal States of Functional Connectivity in Electroencephalography	65
Pál Weisz and György Orosz: Towards Efficient Separability of Models in Automated Worst-Case Circuit Analysis	71

Towards Automated Data Invariants Synthesis and Validation

Nada Akel , Oszkár Semeráth , László Gönczy 

Budapest University of Technology and Economics

Department of Artificial Intelligence and Systems Engineering

Budapest, Hungary

Email: nada.ake1@edu.bme.hu, {[semerath.oszkar](mailto:semerath.oszkar@vik.bme.hu), [gönczy.laszlo](mailto:gönczy.laszlo@vik.bme.hu)}

Abstract—In data space environments, multiple partners share and process data via components that transform heterogeneous data from different sources. Although these transformations are necessary to enable interoperability and analytics, any error in their implementation introduces unintended or illegitimate changes that violate data invariants, thereby affecting data integrity and trust. This paper proposes an approach for deriving expected output data invariants from predefined input data model invariants and partial transformation specifications. Based on the derived invariants, we introduce a validation framework that verifies transformed data and detects violations of invariants caused by faulty or unauthorized transformations.

Index Terms—Data spaces, Data transformation, Data invariants, Data integrity.

I. INTRODUCTION

Context. Data transformation is a fundamental operation in data spaces, where data is processed by a set of components operating on heterogeneous inputs with varying structures, formats, and quality levels. As data flows through the data space, components (often developed by different organizations) apply transformations to the data as part of their processing, such as schema standardization, value normalization, data cleansing, and aggregation. However, these transformations can introduce inconsistencies, errors, or even unauthorized modifications, affecting the integrity and quality of shared data and raising a fundamental challenge to building trust in data space environments [1]. This highlights the need for an approach that verifies the preservation of data invariants during transformations and prevents unintended or illegitimate data changes.

Problem. Data transformations are a core functional requirement of data space components; however, there is currently no systematic methodology to verify whether these transformations preserve predefined data invariants. Consequently, unintended or illegitimate data changes occurring during the transformation process may remain undetected, significantly undermining data integrity and trust within data space environments. Furthermore, these transformations are often documented using informal natural language descriptions and simple input-output examples. This lack of formalization makes it challenging to establish a clear traceability link between the input and output models directly from the provided specification, preventing the automated derivation of verification constraints.

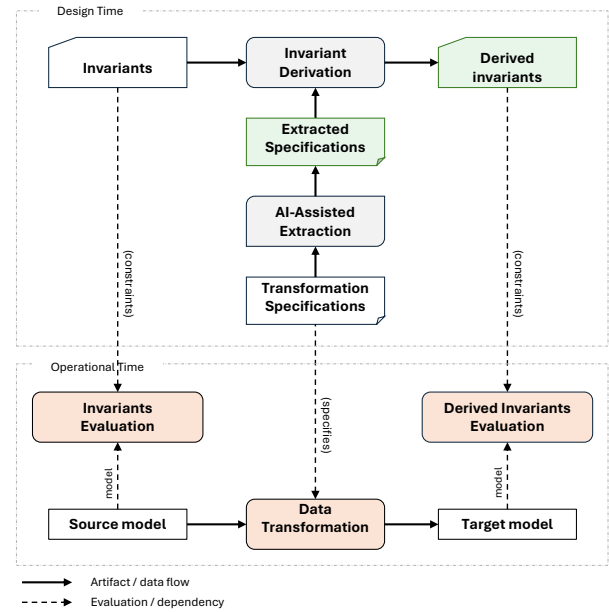


Fig. 1. Overview of the proposed approach

Objectives. The goal of this paper is to introduce an approach for verifying that data transformations executed within data space components preserve predefined data invariants. In addition, we propose a method to derive output invariants from input data model invariants and partial transformation specifications. Our primary focus will be on data structured with both relations and attributes, particularly attributed graph data such as Experience API statements [2], RDF triples [3], and Neo4j graphs [4].

Contribution. We propose a data transformation verification approach that derives output model invariants from predefined input model invariants and partial transformation specifications. To support this process, we employ an AI-assisted method to extract the detailed transformation specifications. Furthermore, we present a validation framework for data space components that evaluates transformed data against these derived invariants, enabling the systematic detection of unintended or illegitimate data modifications.

II. PROPOSED APPROACH & RESEARCH QUESTIONS

A. General Setup

In this section, we present a general overview of the proposed approach (illustrated in Figure 1). Our approach operates on data within data spaces, which we collectively refer to as *models*, and we denote the set of all possible models by $\mathcal{M}$. A data transformation is defined as a function ($t : \mathcal{M} \rightarrow \mathcal{M}$) that takes a *source model* ($S \in \mathcal{M}$) and produces a target model ($T = t(S)$, $T \in \mathcal{M}$).

Domain-specific requirements and constraints are often imposed on the model [5], expressed as invariants ($I \in \mathcal{I}$). An invariant is a function ($I : \mathcal{M} \rightarrow \{true, false\}$) that determines whether a model is valid (if $I(M) = true$) or invalid. In general, a data veracity component [6] ensures the continuous evaluation of these requirements, allowing us to assume that the source model of each transformation is valid.

B. Extraction of Transformation Specifications

In practice, the data transformations performed by a specific data space component are often not fully formalized. Instead, they are described using natural-language specifications and supported by example input–output models. Our approach uses an AI-assisted method to extract and formalize data transformation specifications, employing large language models (LLMs) to extract formal specifications from the provided descriptions. The LLM produces a structured intermediate representation that captures the transformation semantics, attribute mappings, and traceability links between the source and target models. This intermediate representation serves as input to the invariant-derivation step, which applies the transformation rules to synthesize derived invariants.

C. Derivation of Invariants

Our approach aims to exploit this traceability to track the impact of an invariant through transformations, enabling us to derive novel invariants in the data space that can be directly evaluated in the target model T . Therefore, an invariant transformation is defined as

$$t' : \mathcal{I} \times \mathcal{S} \rightarrow \mathcal{I},$$

where $I' = t'(I, s)$ is the derived invariant. Such an invariant transformation is correct if it over-approximates the original invariant:

$$\forall M \in \mathcal{M} : I(S) \Rightarrow I'(T).$$

In other words, if a correctly derived invariant rejects a target model as invalid, then either the transformation is incorrect, the source model is invalid, or the extracted specification contains an error.

D. Research Questions

The key question of our vision paper is the following:

- How to formulate traceability links for data transformation that cover a large set of transformation rules typical in data space ecosystems.
- How to transform data space invariants to derived invariants that can find errors in data transformations?

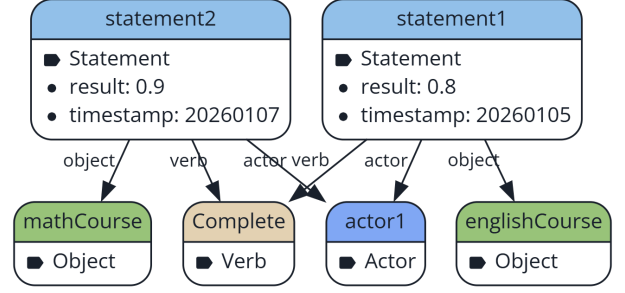


Fig. 2. Example of xAPI input data with predefined model invariants.

III. EXPECTED CONTRIBUTIONS BY EXAMPLE

To illustrate our contribution, we introduce a simple running example based on xAPI as an input data model. The input data captures students' course completion events with their achieved scores. Figure 2 shows an example input model. Suppose that the following invariants hold for all input statements.

Data invariants:

- $I1$: Every statement contains exactly one actor, one verb, and one object.
- $I2$: A student cannot complete the same course twice with the same timestamp (No duplicate events).
- $I3$: Score values must satisfy $0 \leq result \leq 1$.

We will use the Shapes Constraint Language (SHACL) to formalize data invariants, providing a basis for deriving and verifying invariants across data transformations.

We consider a learning analytics scenario in which xAPI statements are collected from a learning management system and processed by a data space component. This component transforms data from one format to another and aggregates xAPI traces to generate learning analytics summaries.

The transformation is not fully formalized; instead, it is described in natural language and illustrated with input–output examples, as shown below.

- Transformation description: For each learner and each course, the component aggregates completion statements and computes the learner's average score. The output dataset contains one record per student course pair.
- Input data format: xAPI.
- Output data format. The output structure is defined as a CSV table consisting of the following columns: studentId, courseId, avgScore, and completionCount.

Our contribution is a systematic approach to derive expected output invariants. By combining predefined input invariants with partial transformation specifications, scenario descriptions, the output data format, and AI-assisted specification extraction, we derive the invariants that are expected to hold for the transformed data.

Derived invariants:

- $I1'$: $completionCount \geq 1$.
- $I2'$: $0 \leq avgScore \leq 1$, since the averages of the values in $[0,1]$ must also be in $[0,1]$.

- *I3'*: No student–course pair appears more than once in the output.
- *I4'*: Aggregation must not introduce students or courses that do not appear in the input data.

Verification step: After the transformation is executed, the output dataset will be validated against the derived output invariants, expressed as SHACL constraints. Example validation checks that detect invariant violations include:

- If $avgScore \geq 1$.
- If $completionCount = 0$.
- If two rows share the same (studentId, courseId).
- If a student appears in the output but not in the input.

Any such violation indicates that the transformation has introduced unintended changes, even if the output may still conform to its schema. The example demonstrates how predefined input invariants and partial transformation specifications with output structures can be combined to derive invariants on transformed data.

IV. EVALUATION PLAN

To evaluate our approach, we will address the following research questions.

- RQ1** How efficient is our technique in establishing traceability links in transformations typical in data spaces?
- RQ2** How efficient are our derived invariants in detecting errors in typical transformation rules?

To answer **RQ1**, we will assess the coverage of the proposed traceability approach by measuring the proportion of commonly used data transformations in data space components that can be represented through traceability links. For this purpose, a representative set of transformation operations, such as filtering, projection, normalization, and aggregation, will be considered.

Measured value: percentage of the successfully detected traceability links.

Selected data: A representative set of data transformation primitives executed in data space components will be collected.

Measurement setup: The proposed approach will be applied to generate traceability links, which are then compared against the intended transformation of the executed operation. A transformation is considered covered if the traceability link successfully maps the many-to-one relationship between the source records and the summarized output attribute.

An experimental evaluation is conducted to assess the error detection capability of the derived invariants **RQ2**. The measured value and the associated measurement setup are specified below.

Research area	References
View modeling	[8], [9]
Access control & Partial modeling	[10]–[13]
Incremental view maintenance	[14], [15]
Data integrity protection	[16]
Schema validation	[17]–[19]

TABLE I
TABLE OF RELATED WORK

Measured value: The empirical probability that a randomly selected transformation and an injected bug can be detected by the proposed technique.

Selected data: We will collect a set of Python data transformation scripts and adopt a mutation testing setup.

Measurement setup: Using mutmut [7], we introduce mutants into the transformation logic. A mutant is considered killed if the derived invariants detect the injected fault; otherwise, it is classified as surviving.

V. RELATED WORK

Data integrity and data transformation are a relevant topic in several loosely connected research lines in the literature. This section aims to summarize the related invariant validation techniques from various related topics focusing on *attributed graph data*. Table I collects and characterizes related publications and approaches.

A. Modeling and Model Query Technologies

Model validation and model transformation are central challenges in model-based engineering. Modern modeling environments built on established modeling technologies [20] rely on model query languages [21] to validate and transform attributed models.

Among these techniques, view modeling [9] plays an important role. Similar to data transformation, its goal is to extract and emphasize problem-specific aspects of a model, producing an abstract, materialized view of the original source model. View modeling approaches typically depend on model queries [8] and can be formally analyzed using various query-lifting techniques [12]. These techniques enable precise traceability between source and target models through a method known as predicate lifting. Similar technologies are also introduced for materialized views for larger databases [14], [15].

Model-based access-control techniques [10] pursue a related but distinct objective: they construct tailored fragments of the overall model that are visible and/or editable for specific users. While view modeling focuses on creating an abstraction of the complete model, access-control techniques aim to expose only a selected subset of it. Partial modeling approaches such as [12] support the analysis of such model fragments and allow the evaluation of invariants under open-world semantics.

Model-based techniques introduce specification-extraction methods [13] that derive executable model queries and transformations directly from natural-language specifications. A

key observation of the paper is that current large language models struggle to consistently generate complex queries correctly. To improve reliability, this paper proposes the generation of traceability links rather than attempting to produce full transformations end-to-end.

B. Data Integrity Validation Techniques

Existing work addresses data integrity in distributed data sharing through hash-based approaches [16]. However, these solutions primarily focus on data integrity during exchange. They do not explicitly address integrity preservation across data transformations performed by data space components.

C. Schema Validation with SHACL

Shapes Constraint Language (SHACL) [22] is a W3C recommendation language for the validation of RDF graphs against a set of constraints. These constraints can be used to describe a graph's properties, detect potential errors in the data, or provide assurances of data quality [17] [18]. Recent work has proposed comprehensive benchmarks for evaluating SHACL engines, focusing on correctness, measuring performance, and evaluating the completeness and comprehensiveness of the generated validation reports using real-world datasets [19]. As such, SHACL provides a suitable foundation for formalizing data invariants for graph-based data.

VI. CURRENT STATUS

In our previous work, we proposed a method for testing data transformations by generating diverse input datasets as test cases [23]. In this paper, we extend this line of research by introducing a new black-box approach for transformation testing. Currently, the research problem and core research questions have been defined, and the fundamental concepts underlying the proposed approach have been established.

The next step is to investigate the traceability of model transformations. Subsequently, implement an automated translator that derives output data invariants from input data invariants based on the defined transformation traceability links.

Future work will also include the development of a prototype validation framework in which the proposed approach is integrated and evaluated within a data space environment.

VII. ACKNOWLEDGEMENTS

This work has been partially supported by the European Union in the frame of the project European Data-space for Growth and Education – Skills, short: EDGE Skills (101123471)

REFERENCES

- [1] A. Hutterer and B. Krumay, "Integrating heterogeneous data in dataspace—a systematic mapping study," 2022.
- [2] IEEE, "9274.1.1 xAPI Base Standard Overview," accessed: 2025-10-07. [Online]. Available: <https://opensource.ieee.org/xapi/xapi-base-standard-documentation/-/blob/main/9274.1.1%20xAPI%20Base%20Standard%20Overview.md>
- [3] F. Manola, E. Miller, B. McBride *et al.*, "RDF primer," *W3C recommendation*, vol. 10, no. 1-107, p. 6, 2004.
- [4] J. J. Miller, "Graph database applications and concepts with Neo4j," in *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, vol. 2324, no. 36, 2013, pp. 141–147.
- [5] M. Altendörfer, T. M. Guggenberger, and F. Möller, "A Design Theory for Data Quality Tools in Data Ecosystems: Findings from Three Industry Cases," *Data & Knowledge Engineering*, vol. 153, p. 102333, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0169023X24000570>
- [6] B. Z. Péter, N. Akel, L. Gönczy, and I. Kocsis, "Data veracity assurance in data spaces," in *International Congress on Blockchain and Applications*. Springer, 2025, pp. 232–242.
- [7] M. Gligoric, V. Jagannath, and D. Marinov, "MuTMuT: Efficient Exploration for Mutation Testing of Multithreaded Code," in *2010 third international conference on software testing, verification and validation*. IEEE, 2010, pp. 55–64.
- [8] C. Debreceeni, Á. Horváth, Á. Hegedüs, Z. Ujhelyi, I. Ráth, and D. Varró, "Query-driven incremental synchronization of view models," in *Proceedings of the 2nd Workshop on View-Based, Aspect-Oriented and Orthographic Software Modelling*, 2014, pp. 31–38.
- [9] F. Madiot and M. Paganelli, "Eclipse Sirius Demonstration." *P&D@MoDELS*, vol. 1554, pp. 9–11, 2015.
- [10] C. Debreceeni, G. Bergmann, I. Ráth, and D. Varró, "Enforcing fine-grained access control for secure collaborative modelling using bidirectional transformations," *Software & Systems Modeling*, vol. 18, no. 3, pp. 1737–1769, 2019.
- [11] O. Semeráth, A. S. Nagy, and D. Varró, "A graph solver for the automated generation of consistent domain-specific models," in *Proceedings of the 40th International Conference on Software Engineering*, ser. ICSE '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 969–980.
- [12] O. Semeráth, C. Debreceeni, Á. Horváth, and D. Varró, "Incremental backward change propagation of view models by logic solvers," in *Proceedings of the ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems*, 2016, pp. 306–316.
- [13] J. A. H. López, M. Földiák, and D. Varró, "TEXT2VQL: teaching a model query language to open-source language models with chatgpt," in *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems*, 2024, pp. 13–24.
- [14] M. Budiu, L. Ryzhyk, G. Zellweger, B. Pfaff, L. Suresh, S. Kassing, A. Gyawali, M. Budiu, T. Chajed, F. McSherry *et al.*, "Dbasp: automatic incremental view maintenance for rich query languages," *The VLDB Journal*, vol. 34, no. 4, p. 39, 2025.
- [15] F. McSherry, "Materialize: a platform for building scalable event based systems," in *Proceedings of the 16th ACM International Conference on Distributed and Event-Based Systems*, 2022, pp. 3–3.
- [16] N. Fotiou, G. Xylomenos, and Y. Thomas, "Data integrity protection for data spaces," in *Proceedings of the 17th European Workshop on Systems Security*, 2024, pp. 44–50.
- [17] P. Pareti and G. Konstantinidis, "A review of SHACL: from data validation to schema reasoning for RDF graphs," *Reasoning Web International Summer School*, pp. 115–144, 2021.
- [18] C. Cortés, L. Ehrlinger, L. Etcheverry, and F. Naumann, "Is SHACL Suitable for Data Quality Assessment?" *arXiv preprint arXiv:2507.22305*, 2025.
- [19] E. Martínez-Sarmiento, E. Ruckhaus, J. Toledo, D. Dona, and O. Corcho, "ERA-SHACL-Benchmark: A real-world benchmark to assess the performance and quality of in-memory SHACL engines," 2025.
- [20] D. Steinberg, F. Budinsky, E. Merks, and M. Paternostro, *EMF: eclipse modeling framework*. Pearson Education, 2008.
- [21] D. Varró, G. Bergmann, Á. Hegedüs, Á. Horváth, I. Ráth, and Z. Ujhelyi, "Road to a reactive and incremental model transformation platform: three generations of the VIATRA framework," *Software & Systems Modeling*, vol. 15, no. 3, pp. 609–629, 2016.
- [22] H. Knublauch and D. Kontokostas, "Shapes Constraint Language (SHACL)," World Wide Web Consortium (W3C), W3C Recommendation. [Online]. Available: <https://www.w3.org/TR/2017/REC-shacl-20170720/>
- [23] N. Akel, O. Semeráth, and L. Gönczy, "Synthetic data generation for trustworthy data space services," in *Proceedings of the 41th ACM/SIGAPP Symposium on Applied Computing*, 2026.

Toward Reliable 3D Understanding of Steel Fibers in CT:

A Vision for Hybrid Manual–Automated Annotation, Tracking, and Validation

Kawthar Alghourani , Kristof Csorba
Budapest University of Technology and Economics
Department of Automation and Applied Informatics
Budapest, Hungary

Email: kawthar.alghourani@edu.bme.hu , csorba.kristof@vik.bme.hu

Abstract—Steel-fiber-reinforced concrete (SFRC) relies on the distribution, orientation, and continuity of embedded fibers, yet these characteristics are difficult to assess non-destructively after hardening. X-ray computed tomography (CT) provides volumetric imaging of internal structures, but converting CT volumes into reliable fiber-level descriptions remains challenging due to imaging noise, partial volume effects, and frequent fiber interactions (contact, overlap, and apparent merging). This vision paper outlines a research agenda centered on practical and reproducible annotation of steel fibers in volumetric CT data. We argue for a hybrid workflow in which automated slice-based detection and inter-slice association provide scalable first-pass annotations, while targeted manual inspection supplies contextual corrections and supports trustworthy validation. The paper identifies core technical gaps (robust separation of touching fibers, stable identity tracking across slices, and uncertainty-aware validation), proposes a modular methodology to address them, and describes an evaluation plan grounded in selective manual annotation, consistency checks, and sensitivity analysis. The overarching goal is to enable transparent CT-based SFRC assessment that is feasible for laboratory-sized datasets while producing fiber-level descriptors suitable for downstream mechanical interpretation and quality assessment.

I. PROBLEM STATEMENT & MOTIVATION

Steel-fiber-reinforced concrete (SFRC) is increasingly used in structural applications where enhanced toughness, crack control, and post-cracking performance are required. By bridging cracks and redistributing stresses after matrix cracking, steel fibers can significantly improve ductility and energy absorption in concrete structures. However, the effectiveness of fiber reinforcement depends not only on the fiber volume fraction, but also on the spatial distribution, orientation, continuity, and interaction of fibers within the hardened composite [3]. Fiber clustering, preferential alignment, or interruptions in fiber continuity can lead to uneven stress transfer and reduced mechanical performance, making reliable characterization of the internal fiber network a central challenge in both research and engineering practice.

Once concrete has hardened, assessing the internal arrangement of steel fibers becomes particularly difficult. Conventional destructive techniques, such as cutting and polishing

followed by visual inspection, provide only limited and often biased information due to sparse sampling and the inherently three-dimensional nature of fiber networks. Moreover, destructive methods prevent repeated measurements and are unsuitable for studying the same specimen under varying conditions. These limitations motivate the use of non-destructive evaluation techniques capable of capturing volumetric information.

X-ray computed tomography (CT) is a well-established non-destructive imaging modality that enables three-dimensional visualization of internal structures in a wide range of materials [1], [2]. In materials science, CT has been successfully applied to investigate pores, cracks, inclusions, and reinforcement elements without physically altering the specimen. For SFRC, CT provides direct volumetric access to embedded steel fibers, offering the potential to extract fiber-level descriptors such as counts, orientation distributions, spatial density, and connectivity. Such descriptors are essential for linking microstructural characteristics to macroscopic mechanical behavior and for supporting quality assessment in laboratory and industrial settings.

Despite its potential, CT does not directly yield interpretable fiber objects. Instead, it produces grayscale voxel volumes in which steel fibers must be separated from the cementitious matrix and interpreted as individual entities. This interpretation step is non-trivial in realistic laboratory scans. Image noise, reconstruction artifacts, and partial-volume effects can blur fiber boundaries and cause fragmented or unstable appearances across slices [7]. In addition, fibers frequently touch or overlap, leading to merged components that obscure individual identities. Depending on fiber orientation and field of view, a single physical fiber may only be partially visible across a subset of slices, further complicating consistent three-dimensional reconstruction.

To address these challenges, many practical CT analysis pipelines rely on classical image-processing techniques, including intensity thresholding, morphological filtering, and connected-component analysis [4], [5]. These methods are attractive because they are transparent, interpretable, and com-

putationally efficient. Slice-based processing remains particularly common in materials research, as it allows intermediate results to be inspected and parameters to be adjusted in a controlled and reproducible manner [10]. However, while such approaches scale well to large datasets, they exhibit systematic failure modes in dense or ambiguous regions, where fiber interactions and imaging artifacts dominate.

More recently, learning-based segmentation methods have demonstrated strong performance in general computer vision tasks [8], [9]. Nevertheless, applying these techniques to SFRC CT data introduces practical difficulties. Training reliable models requires sufficiently large and representative annotated datasets, while defining ground truth becomes ambiguous in regions where fibers touch or partially overlap. Furthermore, learning-based approaches may be sensitive to scanner settings, reconstruction parameters, and material composition, which can limit their robustness and interpretability in laboratory-scale research workflows.

These observations suggest that neither fully manual nor fully automated approaches are sufficient on their own for reliable CT-based fiber analysis. Manual annotation benefits from expert judgment and contextual interpretation across slices, but it is time-consuming, subjective, and difficult to scale. Fully automated pipelines offer scalability and reproducibility, yet struggle to handle ambiguity and to communicate uncertainty. This tension motivates a hybrid perspective, in which automated methods provide scalable first-pass annotations, while targeted manual inspection is used to resolve ambiguous cases and to support trustworthy validation.

In this vision paper, we argue that reliable three-dimensional understanding of steel fibers in SFRC requires such a hybrid manual–automated annotation strategy.

II. RELATED WORK

X-ray computed tomography (CT) has become an essential non-destructive technique for analyzing internal structures of heterogeneous materials. In cementitious composites, CT enables three-dimensional visualization of aggregates, pores, cracks, and reinforcing elements without destructive sectioning. Landis and Keane [5] provided an overview of X-ray microtomography applications in material characterization, emphasizing its capability for volumetric analysis and quantitative phase identification. The ability to reconstruct internal reinforcement geometry has made CT particularly valuable for the study of steel fiber-reinforced concrete (SFRC).

A. CT-Based Characterization of Steel Fibers

The mechanical performance of SFRC strongly depends on fiber orientation, spatial distribution, and fiber–fiber interactions. Early CT-based investigations focused on extracting global descriptors such as orientation tensors and fiber volume fraction. Babout *et al.* [6] employed volumetric CT reconstruction to estimate fiber orientation statistics and evaluate dispersion patterns within concrete specimens. Similarly, Brisard *et al.* [7] quantified orientation distributions derived from CT

segmentation and investigated the relationship between fiber alignment and structural properties.

More recent work has extended these approaches to high-resolution CT datasets. Zhang *et al.* [1] performed detailed three-dimensional characterization of steel fibers in ultra-high-performance concrete, focusing on geometric reconstruction and orientation analysis. Hernández *et al.* [3] further developed quantitative methods to evaluate fiber orientation and distribution metrics from segmented CT volumes.

Although these studies demonstrate the feasibility of extracting volumetric descriptors, their primary focus lies in global statistical analysis rather than explicit tracking of individual fibers across slices. In many cases, fiber continuity is implicitly assumed once segmentation is performed, and merge or split events are not analyzed separately.

B. Segmentation Strategies in Cementitious CT Imaging

Accurate segmentation constitutes a critical step in CT-based analysis. Traditional segmentation methods rely on intensity thresholding, region-growing, and morphological operations. In multi-phase cementitious materials, Zhou *et al.* [8] combined watershed algorithms with machine learning classifiers to separate phases in CT images. Such hybrid approaches aim to balance interpretability and automation.

With the increasing availability of annotated datasets, deep learning methods have gained attention. Wang *et al.* [2] proposed a convolutional neural network (CNN)-based framework for automatic steel fiber segmentation in CT images of concrete, demonstrating improved pixel-level detection accuracy compared to classical methods. Similarly, Zhang *et al.* [9] applied CNN architectures to automate segmentation of fiber-reinforced concrete, reporting enhanced robustness under varying imaging conditions.

Li *et al.* [4] introduced a hybrid methodology combining image processing with machine learning techniques to improve fiber detection in composite materials. Their approach illustrates the trend toward integrating data-driven models with classical processing pipelines.

Despite these advances, segmentation-oriented studies primarily evaluate performance using pixel-level metrics such as precision, recall, or intersection-over-union. Less emphasis is placed on identity continuity across slices or on structured modeling of ambiguous configurations arising in dense fiber regions.

C. Volumetric Analysis and 3D Reconstruction

Three-dimensional analysis of CT data requires consistent association of segmented regions across consecutive slices. Schladitz *et al.* [10] discussed 3D image analysis techniques for fiber-reinforced concrete, highlighting the importance of volumetric reconstruction and spatial quantification. Their work emphasizes extraction of geometric descriptors from reconstructed volumes but assumes stable segmentation and continuity.

In practical CT datasets of SFRC, however, several challenges complicate identity preservation: partial volume effects,

Hybrid Framework for Reliable 3D Understanding of Steel Fibers in CT Data

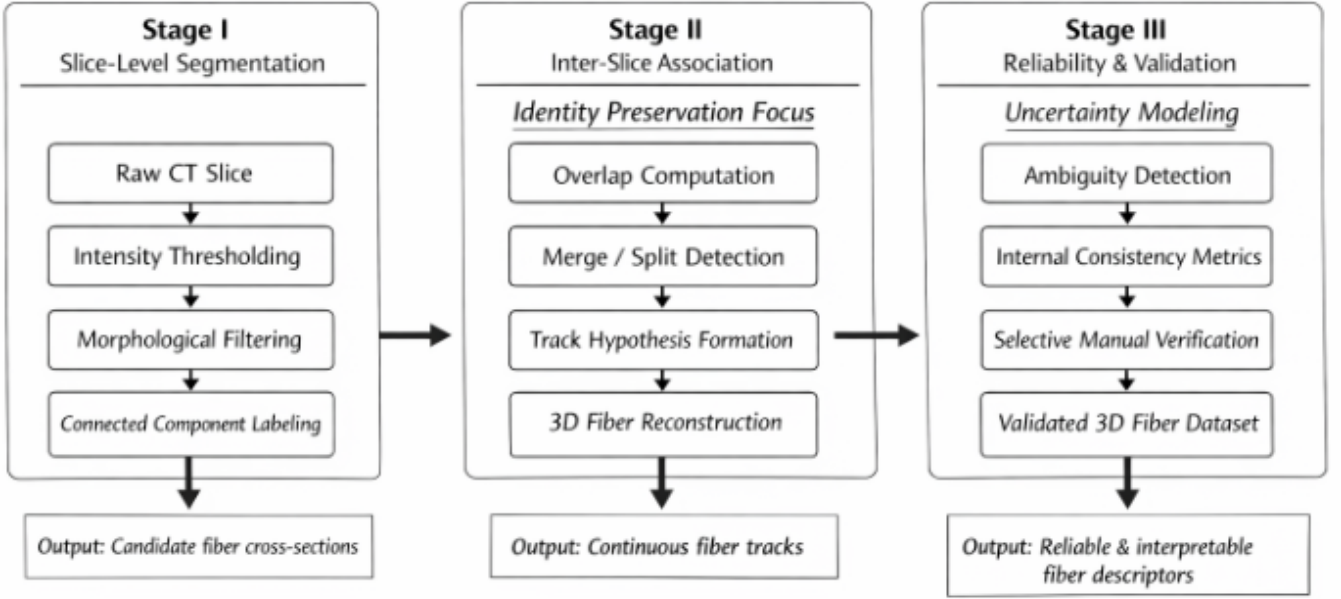


Fig. 1. Proposed hybrid framework for reliable 3D understanding of steel fibers in CT data. The workflow progresses from slice-level segmentation to inter-slice identity preservation and uncertainty-aware validation.

contact between adjacent fibers, and noise-induced irregularities in cross-sectional shapes. These factors may produce merge or split events that complicate deterministic 3D reconstruction. While existing literature addresses segmentation and statistical quantification extensively, systematic treatment of slice-to-slice identity ambiguity has received comparatively limited attention.

D. Limitations of Existing Approaches

The reviewed studies collectively demonstrate substantial progress in CT-based characterization and automated segmentation of steel fibers. Nevertheless, several limitations can be identified:

- Fiber continuity is often assumed rather than explicitly validated during inter-slice association.
- Merge and split events in dense regions are rarely classified or quantified as separate phenomena.
- Evaluation frameworks primarily focus on segmentation accuracy rather than tracking stability.
- Manual annotation is typically used to generate ground truth datasets rather than integrated as a structured validation component.

As a result, reliability assessment of individual fiber identity across the entire CT volume remains underexplored.

E. Positioning of the Present Work

In contrast to segmentation-centric approaches, the present study emphasizes identity-preserving tracking and ambiguity-aware modeling. Rather than maximizing pixel-wise accuracy

alone, the proposed framework integrates overlap-based association, geometric continuity constraints, and explicit detection of merge and split events. Furthermore, a hybrid validation strategy is introduced, in which targeted manual review is guided by automatically detected ambiguity indicators.

By focusing on slice-consistent identity propagation and structured uncertainty modeling, this work complements existing segmentation and statistical characterization studies and aims to enhance reliability in three-dimensional fiber analysis of SFRC.

III. PROPOSED APPROACH

The objective of the proposed research is to develop a structured, reliability-oriented framework for three-dimensional characterization of steel fibers in X-ray CT volumes of steel fiber-reinforced concrete (SFRC). While prior work has focused primarily on segmentation accuracy or global orientation statistics, the present approach emphasizes identity preservation, ambiguity detection, and structured validation. The methodology is designed as a staged pipeline that progressively transforms raw CT slices into validated 3D fiber representations. The overall architecture is illustrated in Fig. 1.

A. Data Characteristics and Preprocessing Assumptions

The framework operates on grayscale CT volumes acquired from hardened SFRC specimens. Each volume consists of a stack of equally spaced axial slices with fixed in-plane resolution and slice thickness. Steel fibers exhibit significantly higher attenuation coefficients than the surrounding cement matrix and aggregates, enabling intensity-based discrimination.

However, several imaging challenges must be considered:

- Partial volume effects at fiber boundaries
- Beam hardening artifacts
- Noise-induced intensity fluctuations
- Fiber contact and overlap in dense regions

These factors motivate a conservative segmentation strategy and the explicit modeling of ambiguity in later stages.

B. Slice-Level Segmentation

The first stage of the pipeline extracts candidate fiber cross-sections independently for each slice. The segmentation process consists of the following sub-steps:

1) *Intensity Thresholding*: A global or semi-adaptive threshold is applied to separate high-density steel regions from the lower-density matrix. Threshold selection is guided by histogram inspection and validated on representative slices. The goal is not maximal recall, but stable detection of high-confidence fiber cores.

2) *Morphological Filtering*: Binary masks resulting from thresholding are refined using morphological operators. Opening operations suppress isolated noise components, while selective closing mitigates minor discontinuities within elongated regions. Minimum area constraints are imposed to eliminate small spurious detections that are unlikely to represent physical fiber cross-sections.

These operations preserve elongated morphology while reducing segmentation artifacts.

3) *Connected Component Labeling*: Following refinement, connected-component analysis assigns unique labels to contiguous pixel groups. Each labeled component is treated as a candidate fiber cross-section and characterized by geometric descriptors, including:

- Area
- Centroid coordinates
- Bounding box
- Equivalent diameter

At this stage, the output consists of a structured set of candidate regions per slice.

C. Inter-Slice Association

Slice-level components must be linked to reconstruct continuous fibers in 3D. The second stage establishes correspondence across consecutive slices.

1) *Spatial Overlap Computation*: For each labeled component in slice i , spatial overlap with components in slice $i+1$ is computed. Overlap ratios are evaluated relative to both regions to ensure symmetric consistency.

A one-to-one overlap indicates a strong candidate for identity propagation. Multiple overlaps (one-to-many or many-to-one) signal potential merge or split events.

2) *Geometric Continuity Constraints*: To prevent implausible associations, additional constraints are introduced:

- Maximum allowable centroid displacement between slices
- Consistency of projected fiber orientation

- Gradual cross-sectional area variation

Associations violating these constraints are rejected to preserve physical plausibility.

3) *Track Hypothesis Formation*: Identity labels are propagated iteratively through the slice stack to form preliminary 3D tracks. Each track is represented as an ordered sequence of cross-sectional regions.

Track termination occurs when no valid correspondence is detected in subsequent slices.

D. Ambiguity Modeling and Event Classification

Dense fiber regions often produce merge and split configurations that cannot be resolved deterministically. Instead of forcing arbitrary decisions, the framework explicitly models ambiguity.

1) *Merge Events*: A merge event occurs when multiple tracks converge into a single component in a subsequent slice. These events are flagged and recorded.

2) *Split Events*: A split event is identified when a single track corresponds to multiple components in the next slice.

3) *Boundary-Induced Fragmentation*: Tracks that terminate near specimen boundaries are distinguished from true discontinuities.

Each event type is annotated within the track representation, allowing downstream reliability assessment.

E. Reliability Metrics

To quantify stability and confidence, several internal metrics are computed:

- Track persistence length (number of consecutive slices)
- Merge/split frequency per track
- Variance of centroid displacement
- Cross-sectional area consistency

Tracks satisfying predefined stability thresholds are classified as high-confidence fibers, while ambiguous segments are marked for review.

F. Selective Manual Validation

Manual annotation is incorporated as a targeted validation mechanism rather than exhaustive labeling.

Representative slices are selected using stratified sampling based on:

- Fiber density
- Proximity to specimen boundaries
- Detected ambiguity events

Manual review focuses on validating identity propagation in ambiguous regions and correcting mis-associations if necessary.

This hybrid strategy balances automation efficiency with expert reliability.

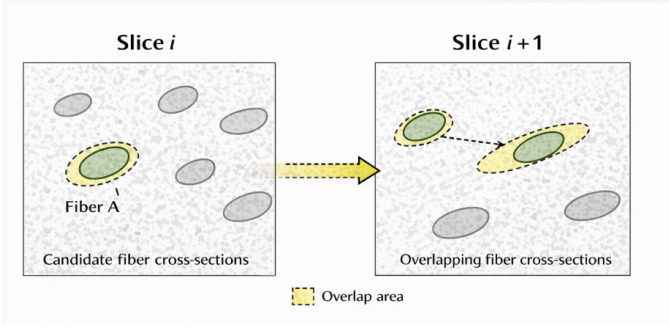


Fig. 2. Illustration of slice-to-slice overlap-based association between consecutive CT slices.

G. Output Representation and Engineering Relevance

The final output consists of validated 3D fiber tracks with associated confidence indicators. From these tracks, engineering descriptors are derived:

- Fiber length distribution
- Orientation vectors
- Spatial dispersion metrics
- Contact frequency estimation

Unlike purely segmentation-based approaches, the proposed method prioritizes traceable identity and uncertainty transparency, providing a structured foundation for reliable structural interpretation and future machine learning integration.

IV. EXPECTED CONTRIBUTIONS

We distinguish contributions that can be achieved incrementally and validated within realistic PhD timelines.

A. Contributions Achieved / In Progress

- A reproducible slice-based pipeline for first-pass steel fiber candidate detection.
- An overlap-driven inter-slice association baseline for building 3D track hypotheses.
- A structured comparison of manual versus automated annotation behavior, emphasizing typical disagreement scenarios (touching fibers, fragmentation, partial visibility).

B. Planned Contributions

- **Hybrid annotation protocol:** guidelines and tooling for targeted manual review that integrates with automated tracking (not a separate manual workflow).
- **Uncertainty-aware tracking:** track-level ambiguity detection and confidence scoring that highlights when automated identity assignment is unreliable.
- **Benchmark slice set and reporting:** a small, shareable annotated slice subset and standardized reporting of parameters, counts, and sensitivity.
- **Fiber descriptors for downstream use:** orientation and spatial distribution measures derived from tracked fibers, enabling correlation studies with mechanical testing.

C. Preliminary Feasibility Observations

Preliminary experiments conducted on laboratory-scale CT volumes indicate that overlap-based association produces stable track hypotheses in low-to-medium density regions. Ambiguity events are primarily observed in dense intersection zones and boundary-adjacent slices, supporting the need for structured uncertainty modeling and selective manual validation. These observations confirm the practical feasibility of the proposed staged framework.

V. EVALUATION PLAN

Evaluation must reflect the reality that exhaustive ground truth is impractical. We therefore propose a multi-layer validation strategy.

A. Internal Consistency Metrics

To evaluate tracking stability and identity preservation, several internal consistency metrics are defined:

- **Track Persistence Length:** number of consecutive slices in which a fiber track remains continuous.
- **Merge/Split Frequency:** number of non-bijective overlap events encountered per track.
- **Centroid Drift:** inter-slice displacement of fiber centroids relative to expected geometric continuity.
- **Area Stability:** variance of cross-sectional area along a reconstructed track.

Tracks satisfying predefined stability thresholds are classified as high-confidence fibers, while segments exhibiting excessive instability are marked as ambiguity-prone.

B. Selective Manual Reference

Manual annotations on selected slices provide a reference for:

- detection completeness in easy regions (clear contrast, low density),
- behavior in hard regions (high density, intersections),
- comparison of counts and identity consistency over short slice windows.

C. Consistency and Sanity Checks

We will validate internal consistency by measuring:

- track continuity (frequency of breaks and reappearances),
- distribution plausibility (e.g., excessive long-span outliers),
- spatial stability (centroid drift limits for plausible tracks),
- parameter sensitivity (robustness across reasonable threshold and filtering ranges).

D. Ablation and Sensitivity Analysis

Ablation studies will isolate the effect of pipeline components:

- threshold choice (global vs adaptive),
- morphological operators and structuring element size,
- minimum area filtering,
- association criteria (overlap only vs overlap + geometry).

TABLE I
PLANNED TIMELINE FOR THE PROPOSED RESEARCH AGENDA.

Period	Milestones
0–2 months	Finalize hybrid annotation protocol; define representative slice sets; consolidate reporting format for parameters and outputs.
2–5 months	Improve inter-slice association with ambiguity detection; implement track confidence indicators; run sensitivity studies on multiple volumes.
5–8 months	Derive fiber descriptors (orientation, spatial density); connect descriptors to mechanical interpretation; prepare internal evaluation report.
8–12 months	Prepare journal/conference submission(s); methodology + evaluation; release reproducible code and example outputs (where permitted).

E. Region-Based Evaluation Strategy

To assess reliability under varying structural conditions, the evaluation distinguishes between “easy” and “hard” regions within the CT volume.

Easy regions are defined as areas with low fiber density, clear contrast between steel and matrix, and minimal contact between adjacent fibers. In these regions, segmentation and identity propagation are expected to remain stable.

Hard regions include high-density fiber clusters, intersection zones, boundary-adjacent slices, and areas affected by noise or partial volume effects. These regions are more likely to produce merge or split events and tracking ambiguity.

Performance metrics are computed separately for both categories to assess robustness under realistic structural variability.

The goal is not just to maximize a single metric but to identify which components contribute most to stable, interpretable results.

VI. CURRENT STATUS & TIMELINE

Table I outlines a feasible plan aligned with a PhD timeline, emphasizing incremental deliverables and early validation.

VII. RISKS AND MITIGATION STRATEGIES

Several technical risks may affect the reliability of the proposed framework.

A. Segmentation Instability

CT noise, partial volume effects, and threshold sensitivity may lead to fragmented or merged fiber cross-sections, particularly near boundaries.

Mitigation: Threshold values are selected conservatively using histogram analysis, and morphological filtering is applied to regularize segmented regions. Sensitivity analysis can be conducted to assess robustness.

B. Merge and Split Events in Dense Regions

In high-density fiber zones, adjacent fibers may appear as merged regions across slices, leading to ambiguous identity propagation.

Mitigation: Instead of enforcing deterministic separation, the framework explicitly detects and records merge and split

events through overlap analysis. Ambiguous cases are flagged for targeted review.

C. Identity Propagation Errors

Incorrect slice-to-slice associations may occur if geometric continuity assumptions are violated.

Mitigation: Geometric constraints, including centroid displacement limits and cross-sectional area stability checks, are used to reject implausible correspondences.

D. Manual Validation Bias

Selective manual verification may introduce subjectivity.

Mitigation: Manual review is guided by predefined ambiguity flags and region stratification criteria to ensure consistency and reproducibility.

VIII. CONCLUSION

This vision paper argues that reliable CT-based understanding of steel fibers in SFRC requires a practical hybrid annotation strategy rather than purely manual or purely automated workflows. By combining transparent slice-based detection, interpretable inter-slice association, and targeted manual validation guided by uncertainty indicators, the proposed agenda aims to produce trustworthy fiber-level descriptors at realistic effort levels. The expected outcome is a reproducible methodology that supports both research studies and practical assessment tasks, enabling improved understanding of how fiber distribution and continuity relate to structural performance.

ACKNOWLEDGMENT

The author thanks colleagues and supervisors for discussions on CT-based analysis and annotation workflows for steel-fiber-reinforced concrete.

REFERENCES

- [1] Y. Zhang, H. Li, and Z. Wang, “Three-dimensional characterization of steel fibers in ultra-high-performance concrete using X-ray CT,” *Construction and Building Materials*, vol. 246, p. 118455, 2020.
- [2] X. Wang, Y. Liu, and Z. Chen, “Deep learning-based segmentation of steel fibers in CT images of concrete,” *Automation in Construction*, vol. 130, p. 103839, 2021.
- [3] J. Hernández, A. Martínez, and M. Pérez, “Quantitative analysis of fiber orientation and distribution in fiber-reinforced concrete using X-ray CT,” *Materials*, vol. 15, no. 7, p. 2345, 2022.
- [4] H. Li, Y. Zhou, and K. Zhang, “Hybrid image processing and machine learning approach for fiber detection in composite materials using CT,” *Composite Structures*, vol. 310, p. 116600, 2023.
- [5] E. N. Landis and D. T. Keane, “X-ray microtomography,” *Materials Characterization*, vol. 61, no. 12, pp. 1305–1316, 2013.
- [6] L. Babout, E. Maire, and J. Adrien, “Investigation of steel fiber orientation in concrete using X-ray computed tomography,” *Construction and Building Materials*, vol. 165, pp. 800–812, 2018.
- [7] S. Brisard, E. Maire, and L. Babout, “Quantification of fiber orientation in reinforced concrete from X-ray computed tomography,” *Cement and Concrete Research*, vol. 109, pp. 1–10, 2018.
- [8] Z. Zhou, L. Wu, and J. Yang, “Multi-phase segmentation of CT images of cementitious materials using random forest and watershed,” *Cement and Concrete Composites*, vol. 104, p. 103348, 2019.
- [9] Y. Zhang, L. Xu, and H. Li, “Automated segmentation of fiber-reinforced concrete in X-ray CT images using convolutional neural networks,” *Automation in Construction*, vol. 124, p. 103567, 2021.
- [10] K. Schladitz, A. Nowacka, and M. Pahn, “3D image analysis of fiber-reinforced concrete using computed tomography,” *Materials Characterization*, vol. 188, p. 111689, 2022.

Computer-search-based Proof Methods in Geometry and Combinatorics

Ákos Dúcz

Department of Computer Science and Information Theory

Budapest University of Technology and Economics

Budapest, Hungary

akos.ducz@gmail.com

Abstract—Recent new methods in geometry and combinatorics rely on finding and proving the existence of witness graphs, which are geometrical structures with desirable parameters. This approach, combined with large-scale computer search methods, can provide useful bounds on a large number of geometric or combinatorial problems that would otherwise be out of reach to prove by hand.

In this project, we apply this framework to a number of questions in pure mathematics, such as the double-cap conjecture or high-dimensional sphere packing.

I. PROBLEM STATEMENT

A recent method in geometry is the use of witness graphs as combinatorial objects, in order to provide upper bounds on the maximal measure or cardinality of certain sets. A famous example of this is the Erdős-Moser conjecture, which asks about the maximal measure of a unit-distance avoiding subset in the plane. This conjecture was recently resolved by Gergely Ambrus, Adrián Csizsárik, Máté Matolcsi, Dániel Varga, and Pál Zsámboki [1], who found a 23-vertex unit-distance witness graph as part of their proof. This witness is a relatively large object with special properties, therefore finding it by hand is completely unfeasible.

This method uses a Monte Carlo Tree Search (MCTS) algorithm to find large graphs with desirable parameters. The values of these parameters are calculated as solutions to certain linear programming problems. Solving these LPs is a hard computational task, as their sizes are exponential in the size of the witness, therefore only approximate solutions are found during the MCTS. Verification of both the witnesses and the LP solutions is one of the main challenges in our research.

The framework of [1] can be applied in many other interesting problems, even some that are not inherently geometric in nature (see the Proposed Research Directions section). We intend to employ this framework to some well-known problems, while also continuously improving our computer search and witness verification methods. Our goal is to build a robust framework that can tackle a large number of related mathematical conjectures, without the need to find long and convoluted proofs by hand.

II. OVERVIEW OF OUR APPROACH

Generally, our framework consists of a heuristically guided search algorithm, which iteratively builds mathematical structures, and evaluates them based on some desired property. It is helpful if this property is monotonous with regards to taking substructures, as this eliminates the need for backtracking during the search.

This general model is applicable to a large number of search problems, but more concretely, we look for certain types of graphs whose corresponding LP problems have small maximal solutions. For example, we use MCTS to find unit-distance graphs in $\mathbb{R}^2$ that have small fractional chromatic numbers, in order to obtain good upper bounds on the Erdős-Moser problem.

Given a certain problem, our method can be loosely broken down into the following steps:

- Restate the problem in terms of an upper bound on the density of some special set (for example, unit-distance avoiding sets in $\mathbb{R}^2$).
- Identify relevant constraints and find the corresponding LP problem.
- Find a witness with the desired properties that minimizes the LP problem (here we use MCTS and approximate LP solutions).
- Verify that the witness graph has the desired properties and prove that the solution of the LP is correct (for example, prove that the graph you found is truly a unit-distance graph).
- With some additional standard probabilistic techniques, obtain an upper bound on the original problem in terms of the LP's solution.

The exact details of this process depend heavily on the problem in question, but this is the general outline. Many problem classes admit this form, including:

- Maximal sphere packings in dimension n .
- Several problems in additive number theory, where an estimate on the sizes of special subsets is required.
- Other Erdős-Moser type questions, including the double-cap conjecture (a similar question on the sphere).

To see how a witness graph may provide an upper bound on the measure of special sets, consider the Erdős-Moser problem:

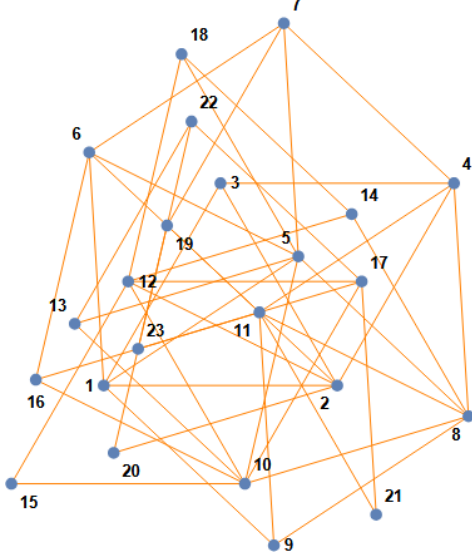


Fig. 1. The unit distance graph G_{23} used in [1]

Suppose we have some set $S \subseteq \mathbb{R}^2$ with density μ which avoids unit distances. Consider a witness graph G with vertices from $\mathbb{R}^2$ such that two vertices x and y of G are connected if and only if $|x - y| = 1$ (this is called a unit-distance graph in the plane). Let $IR(G)$ denote the independence ratio of G , that is $IR(G) = \alpha(G)/|V(G)|$. Now consider randomly “throwing” G onto the plane (more rigorous formalization is omitted here for the sake of clarity).

Taking the expectation $\mathbb{E}[|V(G)' \cap S|]$, where $V(G)'$ is the transformed set of vertices, by the linearity of expectation we find that it is equal to $\mu|V(G)|$. On the other hand, the set $V(G)' \cap S$ is an independent set of G , since S avoids unit distances. Therefore $\mu|V(G)| = \mathbb{E}[|V(G) \cap S|] \leq \alpha(G)$, and from this it follows that $\mu \leq IR(G)$ by definition.

The independence ratio is a relatively weak parameter to use compared to more sophisticated ones, but it demonstrates the power of this method: it is sufficient to find a finite witness object in order to bound the size of all possible sets S .

III. PROPOSED RESEARCH DIRECTIONS

A. General application of the framework

Our planned research involves the formalization of a broader class of problems within this framework. This includes generalizing existing solutions to higher-dimensional and more abstract spaces, as well as identifying entirely new domains for the application of this methodology. This holds the potential to yield bounds on problems that have remained analytically intractable due to the inherent complexity of their underlying constructions.

Current efforts are directed toward the application of this framework to the problem of high-dimensional sphere packings, a domain with critical implications for algebraic coding theory [6]. Conventionally, a sphere packing in $\mathbb{R}^n$ is characterized by a discrete set of central points subject to a minimum

pairwise distance constraint that ensures no two spheres overlap. However, this problem is not directly formalizable in our current framework, as the measure of any discrete point set is zero, rendering our techniques inapplicable. To address this, we propose substituting discrete centers with dense subsets in $\mathbb{R}^n$, thereby enabling the formulation of a more complex set of constraints for the underlying linear program.

We plan to apply this framework to various high-dimensional problems in geometry and information theory, focusing on configurations where inherent complexity makes traditional analytical bounding techniques difficult.

The planned research extends the framework to several classes of extremal problems in geometry and combinatorics. Specifically, we intend to investigate:

- **Independence density** and forbidden geometric configurations [2]
- **General Packing Problems:** Application to n -dimensional sphere packings and related packing problems in general metric spaces to establish new density bounds.
- **Additive Combinatorics and Group Theory:** Tiling problems within group-theoretic contexts and the study of sets free of arithmetic progressions (AP-free sets).
- **Pattern-Avoidance in Point Sets:** Formalization of configuration-free point sets in $\mathbb{R}^n$, such as triangle-free sets or those avoiding more complex fixed geometric patterns.
- **Measurable Chromatic Numbers:** The measurable variant of the notoriously difficult Hadwiger-Nelson problem. [4]
- **Quantum Information Theory:** The construction and existence of Mutually Unbiased Bases, leveraging the ability of the system to handle complex geometric constraints. [7]

The Hadwiger-Nelson problem asks for the chromatic number of the plane: that is, the minimal number of colors used to color $\mathbb{R}^2$, such that no two points at unit distance have the same color. A variation of this problem asks for this minimal number in case our color classes are measurable. The current best bound on this is at least 5 colors, and we hope that using our methods this lower bound can be improved to 6.

The mutually unbiased bases problem (MUB) asks for the largest number of mutually unbiased bases in the d -dimensional Hilbert space $\mathbb{C}^d$. Two orthonormal bases X, Y in $\mathbb{C}^d$ are MUB, if $\langle x_j, y_k \rangle = 1/\sqrt{d}$ for every j, k . This is a question about maximal sets of vectors with prescribed distance constraints, which also fits our framework.

B. Improving the search algorithm

Our search algorithm utilizes MCTS to navigate the large search space of potential witnesses, directed by approximate LP evaluations and heuristics. This algorithm constructs witness graphs incrementally, prioritizing the exploration of branches that show the highest potential to yield good candidates while dynamically optimizing the computational budget allocated to each section of the search tree.

Determining which vertices to add and which paths to explore is a complex optimization problem. To address this, we emphasize the development of refined heuristics that improve accuracy while remaining efficient to compute. For the problem classes considered, this involves finding good estimates for the fractional chromatic number, and its geometric counterpart, the Geometric Fractional Chromatic Number or *GFCN*.

At present, these values are derived by solving linear programs of exponential size relative to the witness graph. To improve search efficiency, it is critical to develop robust heuristics that can provide fast, reliable approximations for these parameters. This challenge may be addressed by training neural networks to serve as fast-executing heuristics for graph-theoretical constants. We are currently investigating models based on the transformer architecture as a way to learn the mapping between witness graph structures and the values of their associated LPs.

Beyond the scope of this proposal, improvement on machine learning methods for approximations of large scale linear programming problems would lead to applications to fields such as molecule modeling or crystallography.

C. Leveraging the overlapping constraint structure of the search tree

The search process may be further accelerated by considering the overlapping constraints between the different linear programs at each node of the search tree. Since the witness graphs are constructed incrementally, each child node inherits the majority of its constraints and variables from its parent while introducing new ones related to the new vertex. This high degree of overlap represents a significant computational opportunity that our current algorithm does not exploit. We plan to explore techniques to leverage this structure, such as employing dual simplex warm-starts to accelerate re-optimization or adopting column-generation methods to handle the incremental addition of variables more efficiently.

D. Efficient vertex representations

The choice of vertex representation is a central challenge in our framework, as it must balance two competing needs: support the execution of rigorous formal verification methods to ensure that all mathematical constraints are definitively satisfied, while remaining flexible enough for efficient modification during the search process.

Assuming we are working in $\mathbb{R}^n$, our approach employs a two-stage representation strategy. During the MCTS search phase, witness vertices are represented as floating-point vectors to maintain high computational throughput. When a candidate witness achieves a promising upper bound, its vertices are translated into a symbolic representation, which allows for the exact formal verification of the properties of the witness graph ensuring that numerical inaccuracies do not compromise the validity of the proof. This is critical, for example, when addressing problems with exact distance constraints, such as unit-distance graphs. Finally, the resulting LP solutions are

validated through interval arithmetic methods, while some analytical bounds may be done by hand.

The conversion from floating-point to symbolic coordinates often yields very complex expression trees that are difficult to simplify or manipulate algebraically. To mitigate this, the search phase may be conducted directly within a suitable field extension of the rationals, such as $\mathbb{Q}[\sqrt{5}]$. This approach simplifies the verification process by maintaining exact coordinate representations throughout the search, thereby eliminating the need for a separate symbolic translation phase.

Identifying the ideal vertex representation is a critical design decision because the representation itself restricts the variety of vertices available to the algorithm, thereby bounding the potential search space.

E. Generalizing to new structures

While the current implementation is built around finding graphs, the underlying methodology is applicable to a broad range of mathematical structures. This MCTS-based framework may be used to find counterexamples to conjectures in fields outside of graph theory, such as group theory, coding theory or Ramsey-related fields.

IV. CURRENT STATUS

As previously noted, a landmark achievement of this methodology is the resolution of the Erdős-Moser conjecture by [1]. This problem dates back to the 1960s, originating from the collaborative work of Paul Erdős and Leo Moser. Erdős specifically conjectured that any set in the plane avoiding unit distances must have an upper density of at most $1/4$. Progress on the conjecture continued for several decades as various studies incrementally lowered the upper bound. By 2018, the bound stood at roughly 0.2544 following work by Máté Matolcsi et al. The final resolution came in 2022, when Gergely Ambrus, Adrián Csizsárik, Máté Matolcsi, Dániel Varga, and Pál Zsámboi [1], utilized the methods described here, combined with Fourier analytical arguments. Another result obtained in a similar way is proving that the fractional chromatic number of the plane is at least 4. This result, established by Máté Matolcsi, Imre Z. Ruzsa, Dániel Varga and Pál Zsámboi, also relied on the discovery of a specific witness graph through an automated search process [5].

Our ongoing research focuses on extending this methodology to geometric problems governed by distance-based constraints. We have already established improved (currently unpublished) upper bounds for the double-cap problem in three dimensions. Furthermore, as previously noted, the development of a suitable generalization to discrete point-sets is also underway. This extension will enable us to investigate high-dimensional packing problems using the method of automated witness discovery.

Beyond geometric applications, these techniques extend to additive number theory and group theory. A notable recent example is the work of Vsevolod Lev, Máté Matolcsi, Péter Pál Pach, and Dániel Varga [3], which establishes the maximal density of Kravitz sets using a similar framework.

V. TECHNICAL RISKS

As much of the verification is not possible to do by hand (due to the extremely large number of constraints that need to be checked), we have to resort to verifying our witnesses by computer. This makes our proofs somewhat less rigorous, as it leaves open the possibility of software bugs or developer oversights invalidating our results. The complete codebase currently contains over 30,000 lines of Python code, and is not open source. Even though most of the code (such as the search implementation) is not publicly available, a short program is provided with each publication that can be used to verify the witness used in the proof. This allows for multiple independent parties to re-verify and possibly find errors in our proofs.

ACKNOWLEDGMENT

I would like to thank Dániel Varga for the opportunity to contribute to this project, and for his invaluable guidance and insights throughout my work. I also thank Viktória Nemkin and Anna Gujgiczner for their extensive support and assistance in the preparation of this paper.

REFERENCES

- [1] Gergely Ambrus, Adrián Csiszárík, Máté Matolcsi, Dániel Varga, and Pál Zsámboki, "The density of planar sets avoiding unit distances." *Mathematical Programming*, vol. 207, no. 1, pp. 303–327, 2024. doi: 10.1007/s10107-023-02012-9.
- [2] Davi Castro-Silva, "Geometrical sets with forbidden configurations," *Forum of Mathematics, Sigma*, vol. 11, p. e44, 2023. doi:10.1017/fms.2023.43.
- [3] Vsevolod Lev, Máté Matolcsi, Péter Pál Pach, Dániel Varga, "On the density of Kravitz sets." *arXiv preprint*, 2025. doi: 10.48550/arXiv.2510.16522.
- [4] Kenneth J. Falconer, "The realization of distances in measurable subsets covering $\mathbb{R}^n$." *Journal of Combinatorial Theory, Series A*, vol. 31, no. 2, pp. 184–189, 1981. doi: 10.1016/0097-3165(81)90014-5.
- [5] Máté Matolcsi, Imre Z. Ruzsa, Dániel Varga, Pál Zsámboki, "The fractional chromatic number of the plane is at least 4." *arXiv preprint*, 2023. doi: 10.48550/arXiv.2311.10069.
- [6] John H. Conway and Neil J. A. Sloane, *Sphere Packings, Lattices and Groups*, vol. 290. New York, NY, USA: Springer, 2013. doi: 10.1007/978-1-4757-2016-0.
- [7] Máté Matolcsi, Ákos K. Matszangosz, Dániel Varga, Mihály Weiner, "Triplets of mutually unbiased bases." *arXiv preprint*, 2025. doi: 10.48550/arXiv.2503.14752.

Towards Differentiable Machining: A Neural-Differential Framework for Physics-Aware Toolpath Optimization

Bertalan Lichter , Dániel Hadházi 

Budapest University of Technology and Economics

Dept. of Artificial Intelligence and Systems Engineering

Budapest, Hungary

lichterbertalan@gmail.com, hadhazi@mit.bme.hu

Abstract—Toolpath synthesis in Computer-Aided Manufacturing (CAM) has long been constrained by rigid geometric heuristics and predefined templates (e.g., trochoidal or zig-zag patterns). Although these approximations ensure coverage, they inherently limit the physical optimality of the cutting process, often failing to maintain constant Cutter-Workpiece Engagement (CWE) in complex geometries. This paper proposes a Neural-Differential Framework that transitions path planning from discrete template matching to continuous, physics-aware synthesis. By employing a neural surrogate to approximate the material removal dynamics, an end-to-end differentiable pipeline is constructed that optimizes spline parameters (control points and knot vectors) via gradient-based optimization algorithms. This approach unlocks global path optimization, allowing the tool trajectory to emerge as a solution to physical constraints rather than geometric rules. Furthermore, the framework leverages the massive parallelism of modern GPU architectures to perform high-fidelity physics evaluation at scale, a capability that traditional CPU-bound geometric kernels lack.

Index Terms—differentiable programming, neural surrogate, physics-aware learning, NURBS optimization, computer-aided manufacturing, toolpath synthesis, cutter-workpiece engagement, end-to-end differentiability

I. PROBLEM STATEMENT & MOTIVATION

In subtractive computer-aided manufacturing (CAM), generating the path that the cutting tool takes is still dominated by rigid geometric heuristics rather than physical optimization. While modern CNC machines are capable of executing complex free-form motions, the algorithms generating these paths remain trapped in discrete geometry. Standard strategies, such as trochoidal milling, rely on pre-defined templates [1] (e.g., circles or spirals) to remove material.

The fundamental limitation is that the Cutter-Workpiece Engagement (CWE), the critical variable determining tool life and surface quality, is calculated via the boolean subtraction of swept volumes. This operation is inherently non-differentiable, acting as a “black box” that prevents the application of modern gradient-based optimization techniques. As a result, industry relies on “well-behaved” geometric patterns that hope to satisfy physical constraints, rather than synthesizing free-form paths that are guaranteed to optimize them. While there exist solutions that generate constant CWE angle (θ_e) paths -

like VoluMill [2] or TrueMill [3] -, they sacrifice cycle-time to achieve this, and are limited to G^1 -continuous paths, which can result in large spikes in the acceleration at the transitions between path-segments.

The suboptimality of the status quo causes unnecessarily long manufacturing times, shortened lifespans of cutting tools, and inefficient material removal rates (MRR) — all contributing substantial costs for the manufacturer, which inevitably accumulate in the final price of the manufactured workpieces.

II. RELATED WORK

A. Geometric Heuristics in Path Planning

The current paradigm in high-speed roughing relies heavily on procedural templates designed to guarantee geometric coverage while bounding mechanical stress and utilization of the tools. Foundational strategies, such as direction-parallel (zig-zag) and contour-parallel (offset) milling, prioritize area clearance efficiency but often suffer from kinematic discontinuities at turn-around points or “pocket collapse” in non-convex centers, where the tool is forced into full-slotting conditions ($\theta_e = 180^\circ$).

To mitigate these engagement spikes, Otkur and Lazoglu [4] established the trochoidal milling strategy, which superimposes a circular motion onto a linear path to explicitly bound the radial depth of cut (a_e). Similarly, Voronoi-based approaches (medial axis transform) [5] attempt to maximize clearance by guiding the tool along the pocket’s skeleton, ensuring a safe distance from channel walls. More recent works, such as Combinatory Logic Synthesis (CLS) [1] have automated the selection and sequencing of these templates, effectively treating toolpath generation as a discrete assembly or connectivity problem.

Although these heuristics successfully bound the maximum engagement in ideal scenarios, they remain fundamentally constrained by their geometric rigidity. Because the underlying primitives - whether circles, offsets, or skeletons - are fixed, the path cannot locally deform to accommodate complex topological variations (e.g., varying channel widths or acute corners) without violating physical constraints or introducing

excessive air-cutting. Consequently, these methods optimize for coverage strategy rather than continuous physical stability.

B. Parametric Adaptation & The Physics Barrier

To overcome the topological rigidity of fixed templates, recent research has shifted toward parametric path adaptation, utilizing higher-degree curves to improve boundary conformance. Chang et al. [6] introduced a "Double-NURBS" strategy, which replaces the standard circular trochoid with two synchronized splines: a centerline curve and an envelope curve. By independently manipulating the control points of the envelope, the toolpath can dynamically compress to fit varying slot widths while maintaining C^2 continuity. Similarly, Zhou et al. [7] demonstrated that optimizing the aspect ratio of elliptical trochoids could reduce engagement variance by over 30% in titanium alloys compared to circular baselines. Even the most recent heuristic advancements, such as the adaptive stepover model by Sougavabar et al. [8], rely on parametric modulation of fixed trochoidal templates rather than free-form synthesis.

However, a fundamental limitation persists across these methods: they operate by tuning the parameters of a pre-defined template rather than synthesizing the path topology itself. More critically, the optimization of these parameters typically hits a "physics barrier." Because the underlying engagement function (the boolean subtraction of swept volumes) is non-differentiable, researchers are forced to rely on zeroth-order stochastic search algorithms. For instance, Jacsó et al. [9] utilized Genetic Algorithms (GA) to optimize B-Spline trochoids, but noted that the computational cost of evaluating the fitness function via discrete simulation scales poorly with path complexity. This lack of gradient information precludes efficient, high-dimensional optimization, trapping the field in a trade-off between geometric flexibility and computational tractability.

C. Physics-Aware Feed Rate Scheduling

Recognizing that geometry alone cannot guarantee process stability, recent works have integrated physical constraints directly into the control loop. Jacsó et al. [10] demonstrated the efficacy of using Artificial Neural Networks (ANNs) to optimize feed rates along a fixed path. While this approach effectively mitigates tool overload by slowing down in high-engagement zones, it treats the toolpath geometry as an immutable constraint.

A significant challenge in these data-driven approaches is the cost of acquiring experimental training data. To address this, Wang et al. [11] demonstrated that transfer learning can effectively bridge the gap between mechanistic simulations and real-world cutting dynamics, achieving high predictive fidelity with minimal physical experiments. Building on this logic, our framework extends the neural approximation from 1D feed scheduling to 2D spatial synthesis, utilizing simulation-based pre-training to ensure robust gradients.

D. Implicit Representations

The concept of optimizing toolpaths via implicit scalar fields was pioneered by Zou et al. [12], who formulated path generation as an energy minimization problem balancing iso-scallop errors and smoothness. More recently, Dutta et al. [13] applied Implicit Neural Representations (INRs) to multi-axis manufacturing, enabling differentiable collision avoidance for static geometries. Although these approaches are powerful for trajectory generation, they operate on static implicit fields representing the final part geometry. Consequently, they fail to model the time-dependent topological evolution of the In-Process Workpiece (IPW), a critical requirement for estimating cutter engagement during high-material-removal rate (MRR) roughing operations. Thus, their utility is primarily limited to surface finishing and multi-axis contouring, leaving the dynamic physics of pocketing unaddressed.

E. Differentiable Programming

In the broader context of computer science, the paradigm of differentiable programming has revolutionized control tasks in robotics and fluid dynamics [14], [15]. By constructing computational graphs that are end-to-end differentiable, these frameworks allow physical parameters to be optimized via backpropagation. However, the application of this paradigm to subtractive manufacturing has been limited by the discrete nature of material removal. There remains a critical gap in bridging differentiable implicit representations with the time-dependent dynamics of the in-process workpiece. Our framework addresses this by unifying the geometric flexibility of parametric curves with the data-driven physical awareness of neural surrogates.

III. PROPOSED APPROACH

We propose an end-to-end differentiable framework where the toolpath is treated as a continuous function $\mathbf{S}(u; \mathbf{P}, \mathbf{U})$ parameterized by its control points $\mathbf{P}$ and knot vector $\mathbf{U}$. $\mathbf{S}$ is chosen to be a Non-Uniform Rational B-Spline (NURBS) [6], [16]. This allows for great flexibility, while retaining the ability to obtain the derivatives analytically. Crucially, this construction will result in a C^2 -continuous path, avoiding spikes in acceleration. The core innovation is the replacement of the non-differentiable Boolean subtraction kernel with a *Neural Surrogate* $\mathcal{N}_\theta$, enabling gradient-based synthesis.

A. The Optimization Problem

Formally, we define path synthesis as a constrained optimization problem where the primary objective is to minimize the cycle-time. The traversal of the path will be constrained by two parameters: the maximum feed-rate ($F_{\max}$) and the maximum centripetal acceleration ($a_{\max}$). We assume a kinematic limit model where the tool velocity is only constrained by the curvature — effectively assuming instantaneous acceleration capabilities. While this will result in the underestimation of the actual cycle-time for segments with rapidly changing curvature, it optimizes the kinematic capacity which will correlate well with the actual cycle-time — especially for

smooth path segments (where this assumption mostly holds true). According to this assumption, at each point u along the path, the machine is able to achieve the maximum admissible velocity: $v_{\text{limit}}(u) = \min\left(F_{\text{max}}, \sqrt{a_{\text{max}}/\kappa(u)}\right)$, where $\kappa(u)$ refers to the curvature at the point u along the path.

With this assumption, the optimization problem can be constructed as follows:

$$\begin{aligned} & \underset{\mathbf{P}, \mathbf{U}}{\text{minimize}} \int_0^1 \|\mathbf{S}'(u)\| \max\left(\frac{1}{F_{\text{max}}}, \sqrt{\frac{|\kappa(u)|}{a_{\text{max}}}}\right) du \\ & \text{subject to:} \\ & g_{\text{geo}}(\mathbf{S}(u)) \geq r_{\text{tool}} \quad \forall u \in [0, 1] \\ & \max_{u \in [0, 1]} (\mathcal{N}_{\theta}(\Phi(u))) \leq \theta_{\text{max}} \\ & \int_0^1 (1 - \cos(\mathcal{N}_{\theta}(\Phi(u)))) \\ & \quad \cdot a_p r_{\text{tool}} \|\mathbf{S}'(u)\| du = V_{\text{excess}} \end{aligned} \quad (1)$$

where $\mathbf{S}(u)$ is the toolpath, and stands for $\mathbf{S}(u; \mathbf{P}, \mathbf{U})$, where $\mathbf{P}$ are the control points and $\mathbf{U}$ are the knots of the spline. Note that in the cycle-time approximation, we divide by $v_{\text{limit}}(u)$, which can also be done by replacing the min function with a max and taking the reciprocal of each operand.

In the constraints, r_{tool} is the radius of the cutting-tool, θ_{max} is the maximum cutter-workpiece engagement angle allowed. g_{geo} is the geometric constraint function: it assigns the shortest signed distance to the boundary of the target geometry. It is implemented as a static signed distance field (SDF) [17], which is negative for points inside the target geometry (the regions which we want to keep) and positive outside.

The feature extraction function $\Phi(u)$ is an abbreviation of the following: $\Phi(\mathbf{S}(t; \mathbf{P}, \mathbf{U}), G_{\text{init}}) \quad \forall t \in [0, u]$. It depends on the spline $\mathbf{S}$ (the historical part - every point before u) and the initial geometry (G_{init}) — also implemented as a signed distance field. The function Φ extracts the local geometric state. It synthesizes a tensor that encodes features regarding both the initial geometry (G_{init}) and the evolving (historical) toolpath. This allows the surrogate to correctly identify un-machined regions even when the cutter is far from the final pocket boundaries, ensuring accurate engagement angle estimation. This function supplies a vector to the neural surrogate $\mathcal{N}_{\theta}$, which maps its output to a single scalar: the engagement angle (θ_e) at that point.

Finally, we denote the axial depth of cut with a_p , the volume of material that is to be removed with V_{excess} ($V_{\text{excess}} = V_{\text{initial}} - V_{\text{final workpiece}}$). The last constraint approximates the amount of volume removed by following the path $\mathbf{S}$, and enforces it to equal the amount of volume defined by the initial and target geometries. Note that while we define this as an equality, it would be sufficient to define it as an inequality, where the amount of volume removed is at least the excess volume V_{excess} , as removing more would result in the violation of the first constraint.

To solve this via unconstrained gradient descent, we formulate the scalar loss function $J(\mathbf{P}, \mathbf{U})$ using the penalty method, where each λ represents the penalty coefficients for the constraint violations.

$$\begin{aligned} J(\mathbf{P}, \mathbf{U}) &= \mathcal{J}_{\text{len}} \\ &+ \lambda_{\text{geo}} \mathcal{L}_{\text{geo}} \\ &+ \lambda_{\text{phys}} \mathcal{L}_{\text{phys}} \\ &+ \lambda_{\text{vol}} \mathcal{L}_{\text{vol}} \\ &= \int_0^1 \|\mathbf{S}'(u)\| \max\left(\frac{1}{F_{\text{max}}}, \sqrt{\frac{|\kappa(u)|}{a_{\text{max}}}}\right) du \\ &+ \lambda_{\text{geo}} \int_0^1 (r_{\text{tool}} - g_{\text{geo}}(\mathbf{S}(u)))^+ du \\ &+ \lambda_{\text{phys}} \int_0^1 (\mathcal{N}_{\theta}(\Phi(u)) - \theta_{\text{max}})^+ du \\ &+ \lambda_{\text{vol}} \left(V_{\text{excess}} - \int_0^1 a_p r_{\text{tool}} (1 - \cos(\mathcal{N}_{\theta}(\Phi(u)))) \right. \\ &\quad \left. \cdot \|\mathbf{S}'(u)\| du \right)^+ \end{aligned} \quad (2)$$

where the operator $(\cdot)^+$ denotes the positive part function. In standard formulations, $(x)^+ = \max(0, x)$, which however yields a vanishing gradient for $x < 0$, as well as only having a subgradient at $x = 0$.

To ensure consistent non-vanishing gradients and improve numerical stability near the boundary interface, we replace the hard maximum with a smooth approximation, the softplus function [18]:

$$\rho(x) = \frac{1}{\beta} \ln(1 + e^{\beta x}) \quad (3)$$

where β controls the sharpness of the transition. This smooth penalty ensures that the optimizer receives continuous feedback even as the constraints approach satisfaction, preventing oscillation and facilitating convergence to a robust solution.

Similarly, we substitute the hard maximum in the kinematic objective with the log-sum-exp (LSE) function to maintain C^∞ continuity across the loss landscape. This modification eliminates gradient discontinuities at the crossover point between feed-limited and acceleration-limited regimes, ensuring the optimizer can smoothly negotiate the trade-off without instability.

$$\text{LSE}(x, y) = \frac{1}{\beta} \log(e^{\beta x} + e^{\beta y}) \quad (4)$$

where β serves the same purpose as with softplus — in fact, softplus is a special case of LSE: $\rho(x) = \text{LSE}(x, 0)$.

With this substitution, we arrive at our final formalization:

$$\begin{aligned}
J(\mathbf{P}, \mathbf{U}) = & \int_0^1 \|\mathbf{S}'(u)\| \text{LSE} \left(\frac{1}{F_{\max}}, \sqrt{\frac{|\kappa(u)|}{a_{\max}}} \right) du \\
& + \lambda_{\text{geo}} \int_0^1 \rho(r_{\text{tool}} - g_{\text{geo}}(\mathbf{S}(u))) du \\
& + \lambda_{\text{phys}} \int_0^1 \rho(\mathcal{N}_{\theta}(\Phi(u)) - \theta_{\max}) du \\
& + \lambda_{\text{vol}} \rho \left(V_{\text{excess}} - \int_0^1 a_p r_{\text{tool}} (1 - \cos(\mathcal{N}_{\theta}(\Phi(u)))) \right. \\
& \quad \left. \cdot \|\mathbf{S}'(u)\| du \right)
\end{aligned} \tag{5}$$

- 1) **Primary Objective** ($\mathcal{J}_{\text{len}}$): The total cycle-time (assuming infinite tangential acceleration). By depending on both $\mathbf{P}$ and $\mathbf{U}$, the optimizer minimizes cycle time not only by geometric shortening (moving control points) but by re-parameterizing the curve to be kinematically efficient. This acts as a global "tension" force, pulling the spline taut in regions where physical constraints allow for higher feed rates. Note that the calculation of $\kappa(u)$ can be done in closed-form, enabling fast calculation without the need of local numerical approximation.
- 2) **Geometric Constraint** ($\mathcal{L}_{\text{geo}}$): A smooth exterior penalty function, derived from the Signed Distance Field (SDF) of the target geometry, penalizing the violation of the boundary condition, traditionally formulated as $\mathbf{S}(u; \mathbf{P}, \mathbf{U}) \in \Omega_{\text{safe}}$. While traditional boolean checks yield a binary result (prohibiting differentiability), we formulate $\rho(g_{\text{geo}}(\cdot))$ as a continuous penalty defined over the positive component of the SDF. By applying the softplus activation ($\rho(\cdot)$), we transform the "hard" wall into a steep, differentiable potential field. This ensures non-vanishing gradients even at the boundary interface, enabling the optimizer to actively guide the toolpath away from potential gouging configurations via back-propagation along the surface normal ∇g_{geo} .
- 3) **Physical Constraint** ($\mathcal{L}_{\text{phys}}$): A penalty applied to each point, at which the estimated engagement-angle θ_e exceeds the allowed maximum. This estimation is done by the neural surrogate $\mathcal{N}_{\theta}$. This term represents the core novelty of our approach, replacing discrete simulation with a differentiable estimator. The surrogate receives the feature map $\Phi(u)$, which encodes the differential between the initial workpiece geometry (G_{init}) and the dynamic toolpath history, calculated from the spline $\mathbf{S}$. Crucially, because $\mathcal{N}_{\theta}$ is differentiable with respect to its inputs, resulting in gradients flowing through the implicit history calculation back to the spline parameters $\mathbf{P}$ and $\mathbf{U}$, allowing the path to deform in a way that ensures constant engagement.
- 4) **Volume Constraint** ($\mathcal{L}_{\text{vol}}$): With only the other two constraints, the optimal path would be one with zero length and in an area far from the initial shape. We

need to motivate the path to cover the entire area which we want to remove. To achieve this, we calculate the amount of volume removed by following the path. This derivation [10] assumes zero curvature at the point of engagement. Although this approximation will result in some numerical error, it should serve as a sufficient proxy. Additionally, we only penalize paths which do not remove enough material, and do not penalize removing too much. This is because paths that remove too much will inevitably violate the geometric constraint (2), and thus will be penalized there.

B. Initialization

The loss-landscape of the loss function (equation 5) is highly non-convex, which could lead the optimizer to get stuck in unacceptable local optima. Although this problem cannot be entirely eliminated, a good initialization would certainly help the optimizer. We plan to use traditional patterns to initialize the toolpath, which can be further optimized using the previously defined approach. In general, we can state that by using first-order optimizers, only local optima can be achieved. To mitigate this, techniques such as multiple initialization could be employed.

IV. EXPECTED CONTRIBUTIONS

This project aims to bridge the gap between deep learning and material science. We anticipate the following contributions.

A. Theoretical Contributions

We provide the mathematical formulation of toolpath synthesis as a constrained functional optimization problem with respect to the spline parameters ($\mathbf{P}, \mathbf{U}$), proving that toolpaths can be treated as learnable entities rather than static geometric templates. This article establishes the bases of such a framework; however, we will likely refine it further in light of empirical results.

We expect to theoretically demonstrate that the non-differentiable Boolean subtraction of swept volumes can be approximated by a differentiable neural surrogate, establishing a gradient path from physical forces back to geometric parameters. We will formulate the model's architecture, based on geometric invariances, as well as the corresponding feature mapping function Φ .

B. Methodological Contributions

We propose a reusable architectural framework for end-to-end toolpath optimization that integrates standard deep learning libraries (PyTorch [19]) with parametric curve modeling, enabling the use of modern optimizers (e.g., Adam [20], L-BFGS [21]) in CAM.

C. Empirical Objectives

We hypothesize that this framework will yield several quantifiable improvements over state-of-the-art heuristics.

First, a statistically significant reduction in cycle-time. By being able to adaptively deform with high degrees of freedom,

the toolpath can bend and shorten to reduce the length, all the while keeping curvature as high as possible to avoid having to slow down — the latter of which many current tools mainly disregard.

Furthermore, a reduction in sudden spikes in acceleration — caused by current tools only yielding G^1 -continuous paths.

Finally, the optimization is highly parallelisable; thus, we can exploit the wide availability of GPUs and find paths within seconds.

D. Artifacts

We plan to make available a reference implementation of the optimization-loop, implemented in PyTorch [19]. It would convert target geometries and parameters, such as the radius of the cutting-tool (r_{tool}), and generate G-code [22], [23], which can be supplied directly to the machine.

V. EVALUATION PLAN

To validate the theoretical claims and practical efficacy of the Neural-Differential Framework, we propose a multi-phase evaluation strategy.

A. Surrogate Validation

First, we will train the neural surrogate model and rigorously evaluate its predictive fidelity against physical simulations of ground-truth. Beyond traditional accuracy metrics (MSE, R^2), we will analyse the gradient cosine similarity between the surrogate’s back-propagated gradients and finite-difference gradients from the physics kernel. This ensures that the model learns the correct underlying dynamics for optimization, not just a scalar mapping. The test set will explicitly include pathological edge cases — such as full-slotting ($\theta_e = 180^\circ$), corner singularity, and zero-engagement transitions — to verify the model’s robustness at the boundaries of the physical domain.

B. Numerical Benchmarking

We will employ numerical experiments to show the performance of the proposed approach relative to the selected baseline. Specifically, we will simulate the machining: first on simpler geometries and then on progressively complex ones. To ensure an industry-standard baseline, tools like VoluMill [2] will be used as reference. During the simulation, several parameters would be observed, including cutter-workpiece engagement, cutting-tool load, and cycle-time. These, along with the optimization time, would serve as the principal validation method.

If both time and resources permit, we plan to validate our approach by manufacturing real workpieces, thereby confirming whether the proposed method is able to be used in an industry setting, under noisy conditions. During these experiments, similar measures would be collected as during the simulations, allowing for a more complete picture.

VI. CURRENT STATUS & TIMELINE

A. Current Status

The theoretical foundations of the neural-differential framework have been laid out (as seen in section III). We have also formally defined the feature function Φ .

B. Timeline

- *February 2026 to April 2026:* We plan to write a proof-of-concept implementation of the neural surrogate model $\mathcal{N}_\theta$, alongside the feature function Φ , in order to validate the approach. We aim to achieve this by the end of March.
- *May 2026 to September 2026:* Our next step will be to expand the simulation results, experiment with different parameterizations, and evaluate several benchmarks. We are planning on starting with straight slotting problems, then moving onto curvilinear slots before eventually tackling general cases.
- *October 2026 to December 2026:* All remaining simulation-based experiments will be completed, and if possible, real experiments will be carried out to further solidify our findings. Once the results are demonstrated to be statistically better from those of the baseline methods, a paper will be written for later publishing.

VII. RISKS & MITIGATION

While in the loss-function, we account for several constraints, these are inevitably “soft-constraints”. Thus, the optimizer could choose to violate one if the reduction in the main loss term would be significant enough. In the context of CNC machining, this is unacceptable (primarily for the geometric constraint, which does not allow the tool to cut into the final workpiece). To counter this, the use of techniques, such as annealing or the interior-point method is vital. For instance, the optimization might start out with λ_{geo} having a lower value, but over several epochs, it would get increased. A similar case could be made for increasing the β parameter of the softplus and log-sum-exp functions to over time sharpen the transition between the penalized and the non-penalized regimes.

The success of the proposed framework is dependent on the success of the neural surrogate model, both in accuracy and in the smoothness of its gradients. This can be aided by regularization and the construction of both the model $\mathcal{N}_\theta$ itself and the feature extraction function Φ in a way that geometric invariances are structurally present, simplifying the problem. Additionally, such an approach will always be prone to local optima, which is why good initialization is crucial. There are several other techniques which could also improve the quality of the found solutions, like annealing [24]. Furthermore, high-quality training data is critical as well: it needs to contain edge-cases, different initial geometries and different splines (histories) for the model to effectively learn the underlying problem with the least possible bias.

While simulation provides a safe environment, physical experiments introduce unmodeled dynamics such as tool deflection, run-out, and material inhomogeneity. A discrepancy between the surrogate’s engagement model and real-world

cutting physics could lead to suboptimal or unsafe toolpaths. This would necessitate a high degree of expert supervision throughout the real experiments. We will also start with high-stiffness materials, like machinable wax, before progressing to metals.

ACKNOWLEDGEMENTS

The authors would like to thank Attila Böröndi for introducing the problem domain of trochoidal milling and providing the initial domain-specific insights that motivated this research direction. Furthermore, we thank Dr. Ádám Jacsó for his invaluable insights in refining the loss function.

REFERENCES

- [1] T. Schäfer, J. A. Bergmann, R. G. Carballo, J. Rehof, and P. Wiederkehr, "A synthesis-based tool path planning approach for machining operations," *Procedia CIRP*, vol. 104, pp. 918–923, 2021, 54th CIRP CMS 2021 - Towards Digitalized Manufacturing 4.0. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2212827121010520>
- [2] E. C. Sherbrooke and G. Coleman, "Non-concentric milling," Patent US 10 579 040 B2, Mar., 2020.
- [3] G. Coleman, A. Diehl, and R. B. Patterson, "Engagement milling," Patent US 7 451 013 B2, 2008.
- [4] M. Otkur and I. Lazoglu, "Trochoidal milling," *International Journal of Machine Tools and Manufacture*, vol. 47, no. 9, pp. 1324–1332, 2007.
- [5] M. Held, "A fast incremental algorithm for computing the voronoi diagram of a planar shape," in *Communicating with Virtual Worlds*. Springer, 1993, pp. 318–329.
- [6] C.-H. Chang, M. Huang, and H.-T. Yau, "A double-nurbs approach to the generation of trochoidal tool path," *The International Journal of Advanced Manufacturing Technology*, vol. 125, no. 3, pp. 1757–1776, 2023.
- [7] X. Zhou, J. Zhou, Q. Qi, C. Zhang, and D. Zhang, "Effects of toolpath parameters on engagement angle and cutting force in ellipse-based trochoidal milling of titanium alloy ti-6al-4v," *Applied Sciences*, vol. 13, no. 11, p. 6550, May 2023. [Online]. Available: <http://dx.doi.org/10.3390/app13116550>
- [8] M. A. Sougavabar, K. Kelam, and I. Lazoglu, "A new trochoidal milling strategy for high-performance cnc machining," *Journal of Manufacturing Processes*, vol. 157, pp. 747–762, 2026.
- [9] A. Jacso, B. S. Sikarwar, R. K. Phanden, R. K. Singh, J. Ramkumar, and G. N. Sahu, "Optimisation of tool path shape in trochoidal milling using b-spline curves," *The International Journal of Advanced Manufacturing Technology*, vol. 121, no. 5–6, p. 3801–3816, Jun. 2022. [Online]. Available: <http://dx.doi.org/10.1007/s00170-022-09527-z>
- [10] A. Jacso, T. Szalay, B. S. Sikarwar, R. K. Phanden, R. K. Singh, and J. Ramkumar, "Investigation of conventional and ann-based feed rate scheduling methods in trochoidal milling with cutting force and acceleration constraints," *The International Journal of Advanced Manufacturing Technology*, vol. 127, no. 1–2, p. 487–506, May 2023. [Online]. Available: <http://dx.doi.org/10.1007/s00170-023-11506-x>
- [11] J. Wang, B. Zou, M. Liu, Y. Li, H. Ding, and K. Xue, "Milling force prediction model based on transfer learning and neural network," *Journal of Intelligent Manufacturing*, vol. 32, no. 4, p. 947–956, Jun. 2020. [Online]. Available: <http://dx.doi.org/10.1007/s10845-020-01595-w>
- [12] Q. Zou, J. Zhang, B. Deng, and J. Zhao, "Iso-level tool path planning for free-form surfaces," *Computer-Aided Design*, vol. 53, pp. 117–125, 2014.
- [13] N. Dutta, T. Zhang, T. Liu, Y. Chen, and C. C. L. Wang, "Implicit neural field-based process planning for multi-axis manufacturing: Direct control over collision avoidance and toolpath geometry," 2025. [Online]. Available: <https://arxiv.org/abs/2511.17578>
- [14] Y. Hu, L. Anderson, T.-M. Li, Q. Sun, N. Carr, J. Ragan-Kelley, and F. Durand, "DiffTaichi: Differentiable programming for physical simulation," *arXiv preprint arXiv:1910.00935*, 2019.
- [15] Z. K. Lawal, H. Yassin, D. T. C. Lai, and A. Che Idris, "Physics-informed neural network (pinn) evolution and beyond: A systematic literature review and bibliometric analysis," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 140, 2022.
- [16] H. Qin, "Dynamic non-uniform rational b-splines," Ph.D. dissertation, Ph. D. thesis, University of Toronto, 1995.
- [17] J. Gomes and O. Faugeras, "Reconciling distance functions and level sets," *Journal of Visual Communication and Image Representation*, vol. 11, no. 2, pp. 209–223, 2000.
- [18] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, "Incorporating second-order functional knowledge for better option pricing," *Advances in neural information processing systems*, vol. 13, 2000.
- [19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
- [20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [21] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1, pp. 503–528, 1989.
- [22] *ISO 6983-1:2009 Automation systems and integration — Numerical control of machines — Program format and definitions of address words — Part 1: Data format for positioning, line motion and contouring control systems*, International Organization for Standardization Std., 2009, ed. 2.
- [23] Y. Yusof, "Iso 14649 (step-nc): new standards for cnc machining," *International Journal of Integrated Engineering*, vol. 2, no. 1, 2010.
- [24] H. Pan, X. Niu, R. Li, Y. Dou, and H. Jiang, "Annealed gradient descent for deep learning," *Neurocomputing*, vol. 380, pp. 201–211, 2020.

Positional Regularities in Canonical SMILES: Implications for Molecular Machine Learning

Márk Marosi, Péter Antal

Department of Artificial Intelligence and Systems Engineering
Budapest University of Technology and Economics
Budapest, Hungary
{marosi, antal}@mit.bme.hu

Abstract—Canonical SMILES representations are ubiquitous in molecular machine learning, yet the algorithmic rules that produce them introduce systematic positional patterns unrelated to molecular chemistry. We present a comprehensive statistical analysis demonstrating that canonical SMILES contain significant positional artifacts: 29% higher mutual information $I(\text{position}; \text{token})$ compared to random SMILES, dramatically elevated position-topology correlations (up to $17.5\times$ higher for ring membership), and algorithm-specific fingerprints enabling 65% classification accuracy for identifying which canonicalizer produced a given string. These findings suggest that models trained exclusively on canonical SMILES may learn positional shortcuts rather than genuine chemical representations, with implications for generalization and robustness. We propose random SMILES augmentation and representation-invariant evaluation as standard practice for molecular ML.

Index Terms—SMILES representation, molecular machine learning, canonicalization, information theory, shortcut learning

I. PROBLEM STATEMENT & MOTIVATION

SMILES (Simplified Molecular Input Line Entry System) has become the dominant text representation for molecules in cheminformatics and machine learning [1]. Originally developed for chemical database storage and retrieval, SMILES encodes molecular graphs as linear strings by performing a depth-first traversal of the molecular structure. A single molecule can be represented by many valid SMILES strings, with the number growing combinatorially with molecular size. For a molecule with n atoms, the number of valid SMILES can exceed $n!$ in pathological cases.

To ensure reproducibility and enable exact matching in databases, canonicalization algorithms produce a unique, deterministic SMILES for each molecule. These algorithms employ sophisticated rules to determine atom ordering: Morgan algorithm variants for symmetry breaking [2], atom priority based on atomic number, connectivity, and stereochemistry, and deterministic tie-breaking procedures. Major cheminformatics toolkits, including RDKit [3], Indigo [4], and OpenBabel [5], each implement distinct canonicalization schemes.

We hypothesize that these canonicalization rules introduce systematic positional patterns that are algorithmic artifacts rather than reflections of underlying chemistry. If this hypothesis holds, machine learning models trained on canonical SMILES may learn these positional shortcuts rather than genuine chemical representations, a phenomenon known as

shortcut learning [6]. Such models would appear to perform well on in-distribution test sets but fail when:

- Molecules are represented with different canonicalization schemes
- Data augmentation introduces random SMILES variants
- Models are deployed on representations from different sources

This research addresses three key questions:

- 1) What positional regularities exist in canonical versus random SMILES?
- 2) How do these patterns relate to molecular graph topology?
- 3) Can we distinguish between different canonicalization algorithms based solely on positional features?

Understanding these questions is critical for the molecular ML community. Recent years have seen explosive growth in transformer-based molecular models including ChemBERTa [7], MolBERT [8], and chemical reaction predictors [9]. These models process SMILES as sequential tokens and may be particularly susceptible to learning positional shortcuts if such patterns exist in training data. In our concurrent mechanistic interpretability work [10], we found that molecular transformers develop specialized circuits for SMILES syntax (ring closures, branch balancing) and encode chemical constraints like valence capacity in their residual streams, making it increasingly important to understand what statistical patterns exist in training data that models might exploit.

II. RELATED WORK

This section contextualizes our findings within the broader landscape of molecular machine learning. We review existing literature on SMILES representations, the phenomenon of shortcut learning in deep neural networks, and the application of information-theoretic measures to chemical data.

A. SMILES Representations in Machine Learning

The SMILES representation [1], [11] has been widely adopted for molecular machine learning due to its simplicity and compatibility with sequence models. Recurrent neural networks [12] and transformers [9] have achieved state-of-the-art results on tasks including property prediction, reaction prediction, and molecular generation.

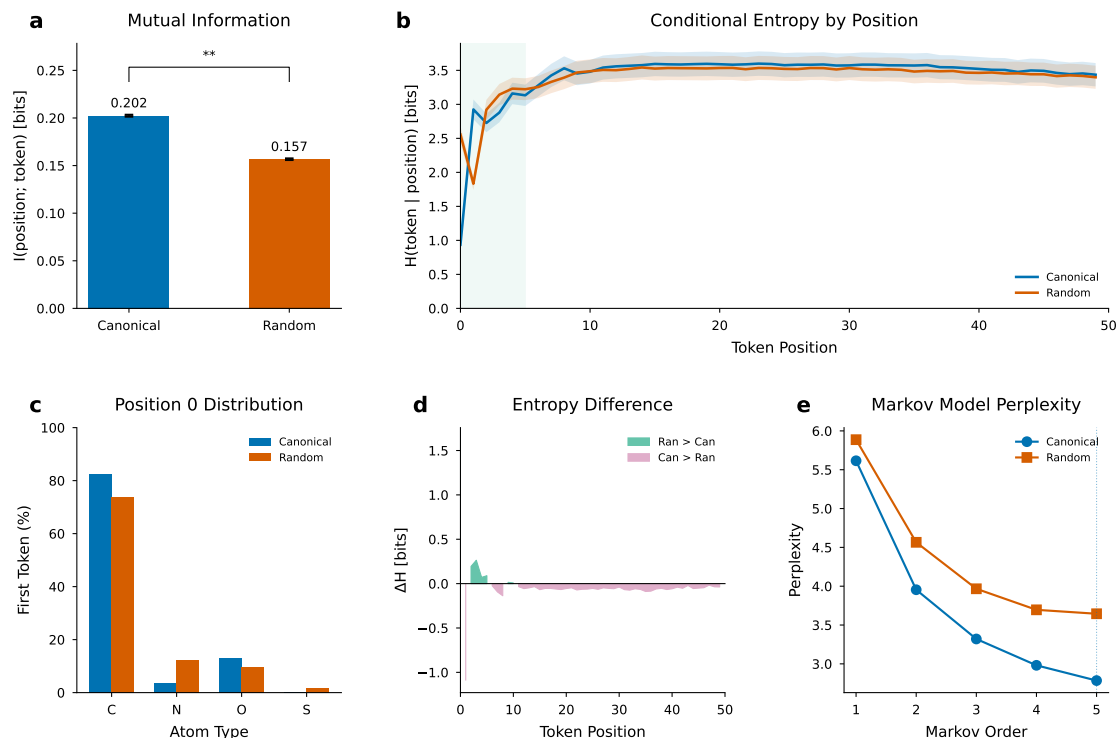


Fig. 1. Information-theoretic quantification of positional bias in canonical SMILES. (a) Mutual information metrics show canonical SMILES (0.202 bits) possess 29% higher positional predictability than random SMILES (0.157 bits). (b) Conditional entropy is markedly lower for canonical strings in early positions, indicating high predictability. (c) The first token distribution is heavily skewed, with 82.5% of canonical SMILES starting with carbon compared to 73.8% for random. (d) Positional entropy differences (ΔH) confirm sustained structural bias. (e) Lower perplexity in canonical SMILES across all Markov orders (1–5) indicates they are significantly easier to predict syntactically.

Several works have noted that SMILES-based models can be sensitive to representation choices. Bjerrum [13] demonstrated that training on multiple random SMILES per molecule improves model robustness and achieves better generalization on molecular property prediction. Tetko et al. [14] showed that SMILES enumeration during training leads to more invariant representations in retrosynthesis prediction, with augmented models outperforming those trained on canonical SMILES alone.

Alternative representations have been proposed to address the limitations of SMILES. SELFIES (Self-Referencing Embedded Strings) [15] guarantees syntactic validity, while molecular graphs with graph neural networks [16] avoid sequential encoding entirely. However, SMILES remains dominant due to its simplicity and the availability of large pre-trained models.

B. Shortcut Learning in Deep Learning

Shortcut learning occurs when models exploit spurious correlations in training data rather than learning the intended task [6]. In natural language processing, models have been shown to exploit positional patterns [17], annotation artifacts [18], and lexical cues [19] that do not generalize. Similar phenomena have been observed in computer vision, where texture rather than shape can dominate predictions [20].

In molecular ML, our concurrent mechanistic analysis has revealed that transformers develop attention heads implementing pointer algorithms for SMILES syntax parsing and encode valence constraints as linear features [10]. However, the statistical structure of canonical SMILES that enables or encourages such pattern learning remains poorly characterized. The present work provides the first systematic characterization of exploitable positional patterns in SMILES representations, complementing our mechanistic studies by characterizing the data-side patterns that models may learn to exploit.

C. Information Theory in Molecular Representations

Information-theoretic analysis has been applied to understand molecular representations. Entropy measures have been used to characterize molecular diversity [21] and evaluate generative models [22]. However, positional mutual information analysis, measuring how much position reveals about token identity, has not been applied to SMILES representations. Our work fills this methodological gap.

III. METHODOLOGY

To rigorously quantify positional artifacts, we employ a matched-pair experimental design comparing canonical and random SMILES. The following subsections detail our dataset selection, SMILES generation protocols, tokenization strategy,

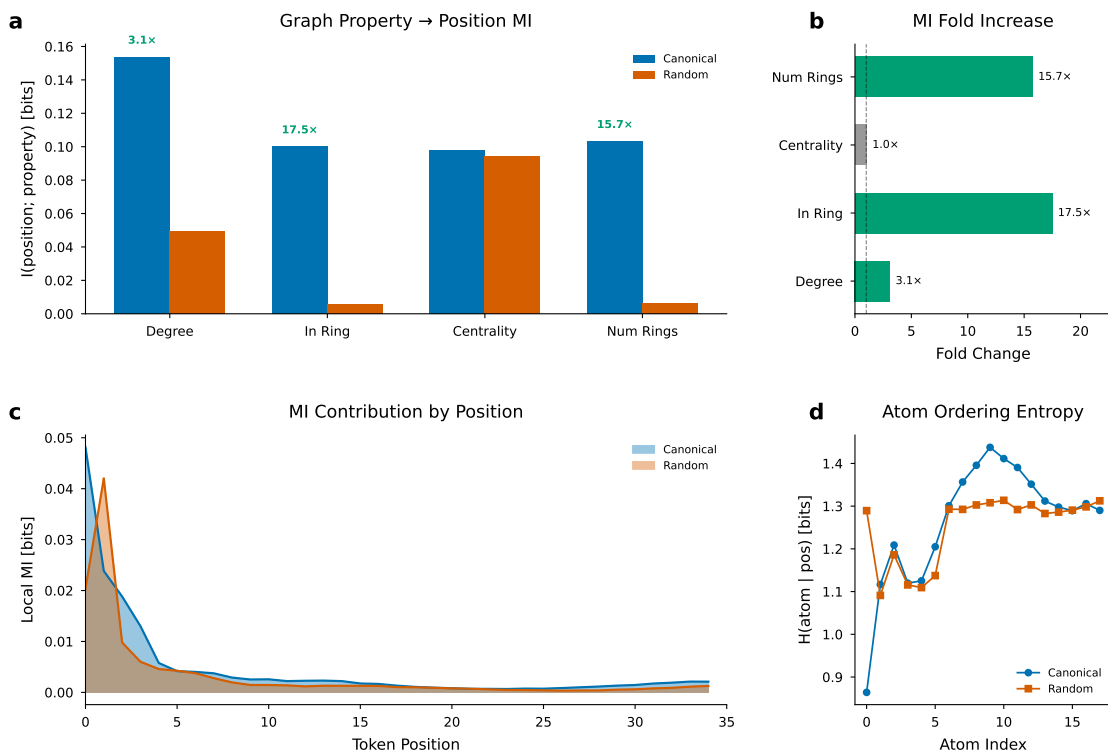


Fig. 2. Positional leakage of graph topology. (a) Mutual information between token position and specific graph properties, showing canonical SMILES (blue) heavily encode topological data. (b) The fold increase in mutual information is substantial for ring-related features, with ring membership ("In Ring") showing a 17.5 $\times$ increase in predictability based on position alone. (c) Local MI contributions show topological leakage is concentrated in the early parts of the sequence. (d) Comparative atom ordering entropy.

and the specific information-theoretic and classification metrics used for analysis.

A. Datasets

We analyze molecules from two established benchmarks:

- **ZINC-250K** [23]: Drug-like molecules filtered for synthetic accessibility and commercial availability. We use a subset of 50,000 molecules.
- **ChEMBL** [24]: Bioactive molecules with annotated activities against biological targets. We use 10,000 molecules for cross-dataset validation.

All molecules are parsed and validated with RDKit, excluding those with parsing errors. The 95th percentile sequence length is 55 tokens, and the mean molecular weight is 350 Da.

B. SMILES Generation

For each molecule, we generate matched pairs:

- **Canonical SMILES:** `Chem.MolToSmiles(mol, canonical=True)` using RDKit
- **Random SMILES:** `Chem.MolToSmiles(mol, canonical=False, doRandom=True)` with 5 independent draws per molecule using different random seeds

This matched-sample design ensures that observed differences are attributable to representation choice rather than molecular properties. For algorithm comparison, we additionally generate canonical SMILES using Indigo and OpenBabel.

C. Tokenization

SMILES strings are tokenized using a regex-based tokenizer that preserves chemically meaningful units. The tokenizer recognizes bracketed atoms (e.g., `[NH3+]`, `[O-]`), two-letter elements (Br, Cl), single-letter atoms, ring closure digits, bond symbols, and stereochemistry markers (`@`, `/`, `\`) as individual tokens. This ensures that chemically meaningful substrings are not split across token boundaries.

D. Information-Theoretic Measures

We compute mutual information between position X and token Y :

$$I(X; Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

Higher MI indicates stronger position-token dependencies. We also compute conditional entropy $H(Y|X)$, which measures remaining uncertainty about tokens given position. To enable



Fig. 3. Algorithmic fingerprinting of SMILES strings. (a) Classification accuracy for identifying the source canonicalizer (RDKit, Indigo, or OpenBabel) using only positional features; GradBoost and XGBoost achieve $\approx 65\%$ accuracy against a 33% baseline. (b) Confusion matrix shows OpenBabel is the most distinct algorithm (75% recall), while RDKit and Indigo share more similarities. (c) Feature importance analysis reveals that ring position statistics (μ , σ) and branch depth are the most discriminative positional artifacts. (d) First token distributions vary by algorithm, with OpenBabel showing a distinct preference profile.

comparison across distributions with different entropy, we report normalized mutual information (NMI):

$$\text{NMI}(X; Y) = \frac{I(X; Y)}{H(Y)} \quad (2)$$

where $H(Y)$ is the marginal entropy of the token distribution. NMI ranges from 0 (position reveals nothing about token identity) to 1 (position fully determines the token). Bootstrap confidence intervals (1,000 resamples, 95% level) quantify statistical uncertainty.

For graph properties, we compute $I(\text{position}; P)$ where $P \in \{\text{degree, in-ring, centrality, ring-count}\}$. These properties are extracted per atom using RDKit’s graph analysis functions.

E. Algorithm Forensics

To test whether positional patterns are algorithm-specific, we train classifiers to identify which canonicalizer (RDKit, Indigo, OpenBabel) produced a SMILES string. We extract 81 positional features including:

- First token identity (one-hot, 12 dimensions)
- Token types at positions 1–10
- Normalized positions of first ring/branch/double bond
- Statistics of aromatic, ring, and branch positions
- Structural complexity metrics (max branch depth, token entropy)

We evaluate Logistic Regression, Random Forest, Gradient Boosting, SVM (RBF kernel), and XGBoost using 10-fold stratified cross-validation.

IV. RESULTS

Our analysis reveals systematic statistical differences between canonical and random SMILES representations. We present evidence across three dimensions: pure positional predictability, the leakage of graph topology into sequence position, and the ability to fingerprint specific canonicalization algorithms.

A. Positional Mutual Information

Canonical SMILES exhibit 29% higher mutual information $I(\text{position}; \text{token})$ than random SMILES (0.202 vs. 0.157 bits, $p < 0.01$, $z = 84.8$). Figure 1 illustrates these findings.

Importantly, the marginal token entropy $H(\text{token})$ is nearly identical between representations (3.43 bits for both), as expected since both encode the same molecules with the same atoms. The difference lies not in total information content but in how predictably that information is distributed across positions. Higher MI means that knowing a token’s position reduces uncertainty about its identity more in canonical than in random SMILES, creating exploitable positional shortcuts.

The first token shows the most dramatic bias: 82.5% of canonical SMILES start with carbon versus 73.8% for random

(χ^2 test, $p < 10^{-50}$). This reflects canonicalization rules that prioritize atoms by atomic number and connectivity, systematically selecting carbon atoms with specific connectivity patterns as starting points. Notably, nitrogen appears $3.3\times$ more frequently at position 0 in random SMILES (12.1% vs. 3.7%), indicating that canonicalization actively suppresses certain atoms from the starting position.

The conditional entropy at position 0 is 64% lower for canonical SMILES (0.93 vs. 2.56 bits), indicating high predictability. N-gram analysis confirms that canonical SMILES are more predictable at all orders, with 4-gram perplexity 16% lower than random (3.32 vs. 3.97). The optimal Markov order is 5 for canonical and 4 for random representations, suggesting that canonical SMILES contain longer-range positional dependencies. Table I summarizes these metrics.

TABLE I
INFORMATION-THEORETIC COMPARISON

Metric	Canonical	Random	Diff.
$I(\text{pos}; \text{token})$	0.202 bits	0.157 bits	+29%
First token = C	82.5%	73.8%	+9%
4-gram perplexity	3.32	3.97	-16%
$H(\text{token} \text{pos}_0)$	0.93 bits	2.56 bits	-64%
NMI (normalized)	0.056	0.044	+27%

B. Graph-Topology Encoding

Position encodes molecular graph structure far more strongly in canonical SMILES. Figure 2 and Table II show mutual information between position and graph properties.

Ring membership shows the most striking difference: $I(\text{position}; \text{in_ring})$ is $17.5\times$ higher in canonical SMILES (0.100 vs. 0.006 bits). This means a model could learn to predict whether an atom is in a ring based primarily on its position, without understanding chemistry. Degree (connectivity) shows $3.1\times$ higher MI, and ring count shows $15.7\times$ higher MI.

Only closeness centrality shows comparable MI across representations ($1.0\times$). This null result is informative: centrality is a global graph property requiring knowledge of all pairwise distances, whereas degree and ring membership are local properties computable during the depth-first traversal that canonicalization performs. The algorithms encode what they compute, not abstract graph-theoretic quantities. These dramatic differences have implications for how models process SMILES, since in our concurrent work we identified attention heads in molecular transformers that implement pointer algorithms for matching ring digits [10], and the systematic position-ring correlations we observe may facilitate learning such circuits.

TABLE II
GRAPH PROPERTY-POSITION MUTUAL INFORMATION

Property	Canonical	Random	Fold Δ
Degree	0.154 bits	0.049 bits	$3.1\times$
In-ring	0.100 bits	0.006 bits	$17.5\times$
Num. rings	0.103 bits	0.007 bits	$15.7\times$
Centrality	0.098 bits	0.094 bits	$1.0\times$

C. Algorithm Forensics

Using only positional features, we achieve 65% accuracy identifying which canonicalizer produced a SMILES (random baseline: 33%). Figure 3 shows results and Table III summarizes classifier performance.

OpenBabel is most distinctive (75% per-class accuracy), while RDKit and Indigo show 26–27% cross-misclassification, reflecting their similar Morgan-based approaches. The most discriminative features are ring position statistics (`mean_ring_pos`, `std_ring_pos`), indicating different algorithms traverse ring systems differently.

Critically, canonicalizers rarely agree: only 0.2% of molecules receive identical SMILES from all three algorithms. Pairwise agreement is also low (RDKit-Indigo: 2.8%, RDKit-OpenBabel: 1.2%, Indigo-OpenBabel: 0.5%). This confirms positional patterns are algorithm-specific artifacts rather than chemistry-derived.

TABLE III
ALGORITHM IDENTIFICATION RESULTS

Classifier	Accuracy	Rel. Gain [†]
Gradient Boosting	64.8% $\pm$ 1.8%	+94%
XGBoost	64.6% $\pm$ 1.4%	+94%
Logistic Regression	62.8% $\pm$ 1.3%	+88%
SVM (RBF)	61.8% $\pm$ 1.4%	+85%
Random Forest	55.7% $\pm$ 1.3%	+67%
Random Baseline	33.3%	—

[†]Relative gain: $(\text{acc} - 33.3\%)/33.3\%$

D. Cross-Dataset Validation

Results generalize across datasets. MI for canonical SMILES is 0.202 bits on ZINC versus 0.175 bits on ChEMBL, with the canonical-random gap persisting in both (ZINC: +29%, ChEMBL: +24%). The slightly lower absolute MI on ChEMBL likely reflects its greater structural diversity. Critically, both datasets show the same qualitative pattern: canonicalization introduces exploitable positional regularities regardless of the underlying chemical space.

V. DISCUSSION

Our findings establish that canonical SMILES contain systematic positional patterns that are algorithmic artifacts rather than reflections of molecular chemistry. The 29% higher mutual information, $17.5\times$ higher position-topology correlation for ring membership, and 65% algorithm identification accuracy from positional features alone provide strong evidence that these patterns exist and are exploitable.

These results complement our concurrent mechanistic interpretability work on molecular transformers [10], which identified specialized attention heads for SMILES syntax (ring digit matching, branch balancing) and linear representations of valence capacity. The present data-side analysis suggests that the systematic positional patterns in canonical SMILES may facilitate the development of such circuits, as consistent statistical regularities in training data provide learnable signals.

VI. CONTRIBUTIONS

Completed contributions: (1) A comprehensive statistical framework quantifying positional patterns in SMILES, including information-theoretic measures with bootstrap confidence intervals. (2) Novel graph-sequence information theory revealing that ring membership is predictable from position in canonical SMILES ($17.5\times$ MI increase). (3) Algorithm forensics demonstrating 65% accuracy identifying canonicalizers from positional features alone.

VII. RISKS & MITIGATION

Technical risks: Models may be robust to representation changes, showing minimal degradation on random SMILES.

Scope limitations: Analysis covers drug-like molecules; patterns may differ for polymers, materials, or natural products. Validation on QM9 [25] and PubChem [26] would extend scope.

Ethical considerations: Findings have implications for deployed molecular ML systems. Limitations should be communicated clearly and best practices recommended.

VIII. CONCLUSION

This work provides the first comprehensive characterization of positional artifacts in canonical SMILES representations. We demonstrate that canonical SMILES have 29% higher position-token mutual information, encode molecular topology at predictable positions ($17.5\times$ higher MI for ring membership), and carry algorithm-specific fingerprints enabling 65% classification accuracy.

The ability to identify which algorithm produced a SMILES string using only positional features is perhaps the most striking evidence that these patterns are algorithmic artifacts, not chemistry. For molecular ML practitioners, we recommend random SMILES augmentation during training and cross-representation evaluation as standard practice to ensure models learn genuine chemical structure rather than positional shortcuts.

REFERENCES

- [1] D. Weininger, "SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988.
- [2] H. L. Morgan, "The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service," *Journal of Chemical Documentation*, vol. 5, no. 2, pp. 107–113, 1965.
- [3] RDKit Contributors, "RDKit: Open-source cheminformatics software," 2026, accessed: 2026-01-07. [Online]. Available: <https://www.rdkit.org>
- [4] EPAM Systems, "Indigo toolkit," 2026, accessed: 2026-01-07. [Online]. Available: <https://lifescience.opensource.epam.com/indigo/>
- [5] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox," *Journal of Cheminformatics*, vol. 3, no. 1, p. 33, 2011.
- [6] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.
- [7] S. Chithrananda, G. Grand, and B. Ramsundar, "ChEMBERTa: Large-scale self-supervised pretraining for molecular property prediction," arXiv preprint, 2020. [Online]. Available: <https://arxiv.org/abs/2010.09885>
- [8] B. Fabian, T. Edlich, H. Gaspar, M. H. S. Segler, J. Meyers, M. Fiscato, and M. Ahmed, "Molecular representation learning with language models and domain-relevant auxiliary tasks," *CoRR*, vol. abs/2011.13230, 2020. [Online]. Available: <https://arxiv.org/abs/2011.13230>
- [9] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction," *ACS Central Science*, vol. 5, no. 9, pp. 1572–1583, 2019.
- [10] K. Váradi, M. Marosi, and P. Antal, "Circuits, features, and heuristics in molecular transformers," arXiv preprint, 2025. [Online]. Available: <https://arxiv.org/abs/2512.09757>
- [11] D. Weininger, A. Weininger, and J. L. Weininger, "SMILES. 2. algorithm for generation of unique SMILES notation," *Journal of Chemical Information and Computer Sciences*, vol. 29, no. 2, pp. 97–101, 1989.
- [12] M. H. S. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks," *ACS Central Science*, vol. 4, no. 1, pp. 120–131, 2018.
- [13] E. J. Bjerrum, "SMILES enumeration as data augmentation for neural network modeling of molecules," arXiv preprint, 2017. [Online]. Available: <https://arxiv.org/abs/1703.07076>
- [14] I. V. Tetko, P. Karpov, R. Van Deursen, and G. Godin, "State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis," *Nature Communications*, vol. 11, p. 5575, 2020.
- [15] M. Krenn, F. Hase, A. K. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation," *Machine Learning: Science and Technology*, vol. 1, no. 4, p. 045024, 2020.
- [16] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 1263–1272. [Online]. Available: <https://proceedings.mlr.press/v70/gilmer17a.html>
- [17] T. Niven and H.-Y. Kao, "Probing neural network comprehension of natural language arguments," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 4658–4664. [Online]. Available: <https://aclanthology.org/P19-1459/>
- [18] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith, "Annotation artifacts in natural language inference data," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 2 (Short Papers)*, 2018, pp. 107–112. [Online]. Available: <https://aclanthology.org/N18-2017/>
- [19] R. T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 3428–3448. [Online]. Available: <https://aclanthology.org/P19-1334/>
- [20] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: <https://openreview.net/forum?id=Bygh9j09KX>
- [21] G. M. Maggiora, M. Vogt, D. Stumpfe, and J. Bajorath, "Molecular similarity in medicinal chemistry," *Journal of Medicinal Chemistry*, vol. 57, no. 8, pp. 3186–3204, 2014.
- [22] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik, and A. Zhavoronkov, "Molecular sets (MOSES): A benchmarking platform for molecular generation models," *Frontiers in Pharmacology*, vol. 11, p. 565644, 2020.
- [23] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "ZINC: A free tool to discover chemistry for biology," *Journal of Chemical Information and Modeling*, vol. 52, no. 7, pp. 1757–1768, 2012.
- [24] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit, and A. R. Leach, "The ChEMBL database in 2017," *Nucleic Acids Research*, vol. 45, no. D1, pp. D945–D954, 2017.
- [25] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," *Scientific Data*, vol. 1, p. 140022, 2014.
- [26] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton, "Pubchem 2019 update: improved access to chemical data," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1102–D1109, 2019.

Benchmarking LLMs as Sequential Diagnosticians: An Interactive Bayesian Environment

Mátyás Antal, András Gézsi, Péter Antal

Budapest University of Technology and Economics

Department of Artificial Intelligence and Systems Engineering

Budapest, Hungary

Email: {antalm, gezsi, antal}@mit.bme.hu

Abstract—Real-world clinical diagnostics is usually a sequential decision problem, including an information-gathering policy with options to stop by performing a diagnosis or rejecting to do so. Modern AI assistants for clinical decision support include an agent architecture choosing which information to acquire next (tests, inspections, sensors) and when to stop and commit, trading acquisition costs against often highly asymmetric misdiagnosis losses. Large Language Models (LLMs) are increasingly used as interactive agents, yet common evaluations emphasize final-answer accuracy and rarely test whether agents behave decision-theoretically under uncertainty—especially in stopping and risk-taking, where safety-relevant failures concentrate.

We propose a controlled benchmarking setting in which a Bayesian network (BN) is available as a probabilistic world model. Treating the BN as a generative simulator enables unlimited grounded episodes, while posterior inference and Value of Information (VOI) provide transparent normative baselines. We introduce **BNENV**, an interactive environment that wraps BNs into cost-aware sequential diagnosis with grouped observations (test panels), prerequisite constraints, and both vector and text interfaces. We outline benchmarks designed to stress stopping behavior and tail risk, and we propose evaluation metrics centered on expected utility, regret, catastrophic error rates, and calibration of stopping decisions. We summarize prototype status and discuss risks and mitigations for making this benchmark informative and reproducible.

Index Terms—Bayesian Networks, Value of Information, Sequential Decision Making, LLM Agents, Benchmarking, Diagnosis, POMDP

I. INTRODUCTION

Many diagnostic workflows—medicine, industrial troubleshooting, predictive maintenance, and safety monitoring—are inherently sequential. An agent must decide which measurement to take next and when to stop collecting evidence and whether commit to a diagnosis or action. These choices are governed by explicit test costs and complex loss functions, e.g. in binary problems by asymmetric losses: a rare false negative can dominate expected harm, while false positives may be tolerable. Formally, the problem is a sequential decision process over a belief state, closely related to Partially Observable Markov Decision Process (POMDP) planning [4].

LLMs are increasingly deployed as interactive assistants, but typical evaluations focus on static accuracy or the correctness of a final response. In sequential diagnosis, the most consequential failure modes are behavioral: premature closure, over-testing, and risk-taking that ignores fine-grained outcomes and

utility structure. Two agents with similar (averaged) accuracy can differ dramatically in expected utility/loss if loss heavily penalizes rare catastrophic errors. This motivates benchmarks that measure utility and stopping behavior directly rather than treating “more accurate” as synonymous with “safer.”

This paper proposes a favorable setting for rigorous evaluation: a Bayesian network (BN) is available as a probabilistic, causal generative world model [1]–[3]. The BN supports (i) unlimited grounded episodes by sampling cases and revealing outcomes through controlled tests, (ii) normative baselines via posterior inference and VOI analysis [9], [10], and (iii) systematic control of difficulty through structure, priors, costs, utilities, and prerequisite constraints. The goal is not to replace BN inference with LLMs; it is to benchmark LLM hidden/latent *policies* against decision-theoretic references in a transparent simulator.

a) Contributions: We present: (1) **BNENV**, an interactive BN-grounded diagnosis environment with grouped observations and prerequisite constraints; (2) dual interfaces (vector and controlled text) that enable comparisons across classical policies, Reinforcement Learning (RL) agents, LLM agents and human participants; (3) a benchmark agenda with task families targeting stopping and tail-risk failures; and (4) an evaluation protocol centered on expected utility, regret versus VOI baselines, catastrophic error rates, and stopping calibration.

b) Research questions: The benchmark prioritizes decision behavior over domain knowledge:

- **RQ1 (Utility alignment):** Do LLM agents choose tests and diagnoses that maximize expected utility under explicit, asymmetric payoffs?
- **RQ2 (Stopping):** Do LLMs stop in a manner consistent with VOI, or do they exhibit premature closure or excessive testing?
- **RQ3 (Risk sensitivity):** How do policies respond to changes in false-negative penalties, prevalence, and test costs?
- **RQ5 (Mitigations):** Which interventions (structured prompting, posterior/VOI tools, hybrid planners) reduce catastrophic outcomes reliably?

SYSTEM
You are a sequential diagnosis agent. Goal: Diagnose 'Pathology' (Benign/Malignant). Costs: Demography (\$0), Ultrasound (\$300), Doppler (\$20), TumorMarker (\$200). Rewards: Correct (+\$1000/+\$500), Wrong (-\$1000/-\$450).
BNEnv
Current State: Observed Demography (Meno: Post, PillUse: 05-5year). Available: Ultrasound (\$300), TumorMarker (\$200).
LLM AGENT
Reasoning: I should request the cheaper available group first to minimize cost. TumorMarker (\$200) is cheaper than Ultrasound (\$300) and provides CA125 data. Action: request TumorMarker
BNEnv
Observation: TumorMarker observed. CA125: 65 $\approx$. Available: Ultrasound (\$300).
LLM AGENT
Reasoning: Post-menopausal status combined with elevated CA125 indicates high malignancy risk. Further testing (Ultrasound) adds significant cost (\$300) without guaranteed accuracy improvements. Action: diagnose Malignant

Fig. 1. Simplified interaction trace. The agent leverages low-cost variables (Demography, TumorMarker) to reach a high-confidence diagnosis, avoiding the expensive Ultrasound step based on cost-benefit reasoning.

II. BACKGROUND AND RELATED WORK

A. Decision-theoretic Diagnosis and Value of Information

Bayesian networks are a standard formalism for diagnostic reasoning under uncertainty [1], [3]. When measurements have explicit costs and decisions carry asymmetric losses, diagnosis becomes decision-theoretic: actions should maximize expected utility, and information gathering should be justified by VOI [9]. Decision-theoretic troubleshooting operationalizes these ideas in practical diagnostic/repair workflows [10]. Because non-myopic information gathering is computationally challenging, prior work studies approximations and efficient algorithms for VOI computation in graphical models [11], [12]. Inference itself can be intractable in general; exact inference is NP-hard and even some approximation regimes are hard [13]. Our agenda leverages BNs precisely because they enable transparent belief updates and strong baselines in controlled settings, while also allowing stress tests where exact VOI is infeasible.

B. Sequential Acquisition and Interactive LLM Agents

A related machine learning literature studies sequential feature acquisition under costs with a “stop-and-predict” action analogous to committing to a diagnosis [14], [16]. Many approaches cast acquisition as RL or POMDP-like optimization over an information-cost trade-off [15], [17]. These works reinforce a key point for evaluation: under asymmetric risks, accuracy and utility can diverge and the reward structure must be measured end-to-end.

LLMs are increasingly evaluated as interactive agents that alternate between reasoning, information gathering, and action [18]–[20]. These benchmarks illuminate grounded interaction and planning failures, but they rarely provide an explicit probabilistic world model that supports normative VOI baselines. In the medical domain, sequential benchmarks incorporating interaction and test costs have been proposed [22]. Our focus is complementary: when a BN is available, we can generate unlimited controlled episodes and compare LLM policies directly against decision-theoretic references. This is particularly relevant because stopping decisions are calibration-sensitive, and LLMs can be miscalibrated depending on elicitation and context [21].

C. IOTA Model and Earlier Work

In medical diagnosis, Bayesian approaches have long been used to reason under uncertainty and to explicitly represent insufficient or ambiguous information. In the context of ovarian cancer diagnosis, a Bayesian multilayer perceptron framework with hybrid Markov Chain Monte Carlo (MCMC) inference was developed to quantify predictive uncertainty and to support *classification with rejection*, allowing the system to defer decisions on hard cases rather than commit under high uncertainty [5].

Within the IOTA framework, expert knowledge and clinical literature were combined with data to construct Bayesian network models of ovarian tumors. These models were published as clinically grounded diagnostic representations and are used in this work as an expert-based gold standard [7].

Large clinical studies and health-economic analyses show that most ultrasound-benign adnexal masses can be safely managed conservatively, avoiding unnecessary surgery without increasing missed malignancies; for example, guideline-driven adoption of expert ultrasound assessment reduced benign ovarian surgeries by 30% at the population level, while the IOTA5 study demonstrated a malignancy rate below 0.5% under expectant management. These findings highlight that accurate, uncertainty-aware diagnostic decision-making yields substantial patient safety and system-level economic benefits [8].

III. BNENV: BN-GROUNDED INTERACTIVE DIAGNOSIS

A. Environment and Reward Model

Let $\mathcal{B}$ be a BN over discrete variables $\mathbf{X}$ with target $T \in \mathbf{X}$. Each episode begins by sampling a hidden “case” $\mathbf{x}^* \sim P(\mathbf{X})$. Observations are organized into *groups* (e.g., test panels) indexed by $g \in V_G$. Group g reveals a subset $\mathbf{X}_g \subseteq \mathbf{X}$ and is governed by a prerequisite graph $\mathcal{G} = (V_G, E_G)$: a group is available if its prerequisites have been observed.

At any step, the agent either requests an available group g or commits to a diagnosis $\hat{t}$. Requesting g reveals the realized values $\mathbf{x}_g^*$ and adds them to the evidence set E . The normative belief state is the posterior $P(T | E)$ obtained via BN inference [1], [3]. Rewards are cost-aware:

$$R(a) = \begin{cases} -C(g) & a = \text{request}(g), \\ U(t^*, \hat{t}) & a = \text{diagnose}(\hat{t}), \end{cases}$$

where $C(g)$ is the acquisition cost and U is an explicit (possibly highly asymmetric) payoff matrix. This formulation isolates the core competence of sequential diagnosis: selecting information and stopping to maximize expected utility.

B. Interfaces for LLM and Non-LLM Agents

To enable controlled comparisons across agent types, BNEnv provides two equivalent interfaces:

- **Vector interface:** a structured observation encoding that includes group status (hidden/available/observed) and the posterior over T .
- **Text interface:** a controlled natural-language (or delexicalized symbolic) description of observed evidence, currently available tests with costs, and the payoff matrix. The agent responds with and optional reasoning block and constrained actions such as `request <group>` or `diagnose <label>`.

Figure 1 illustrates an example interaction trace using the text interface. The system prompt describes the scenario, and the model interacts with the environment through the defined textual schema.

C. Preliminary Results

Fig. 2 summarizes an initial end-to-end pilot in BNEnv on a binary ovarian BN diagnosis task with grouped, prerequisite-constrained acquisitions and an explicitly asymmetric payoff matrix (Fig. 2c). We compare the baselines policies against three LLM agents: two open-weight models (Qwen3:4B, Qwen3:32B) and one proprietary model (gpt-5-nano).

The LLM agents expose two characteristic failure modes that are largely invisible to accuracy-only evaluation. First, smaller models fail to reliably follow the constrained action format: Qwen3:4B exhibits the highest invalid-action failure rate (Fig. 2f), which directly degrades both reliability and utility. Second, while larger models improve diagnostic accuracy (Fig. 2e), they tend to over-acquire information before committing, requesting additional tests even when the expected utility gain is unlikely to justify the added cost. This issue is especially exacerbated for the proprietary model, where to pressure to make models that avoid catastrophic outcomes is larger than the open-weight counterparts.

IV. BASELINES AND BENCHMARK FAMILIES

A. Normative and Ablation Baselines

A central advantage of BN-grounded tasks is the availability of transparent decision-theoretic references. We plan to benchmark against:

- **Myopic VOI:** for each available group, compute its expected utility improvement over immediate diagnosis and select the best positive-VOI test; stop when no test has positive net value [9], [10]. In sufficiently small networks, this can be equivalent to the optimal solution, when look-ahead cut-off is high enough.

- **Finite-horizon rollout VOI:** approximate non-myopic acquisition by sampling-based look-ahead (depth- k rollouts), enabling stronger baselines when myopia is insufficient [11], [12].
- **Ablations:** random, entropy-greedy, cost-greedy, and threshold policies to isolate which mechanisms drive performance.

These baselines form a “ladder” from simple heuristics to principled (approximate) planning, and they provide a meaningful regret reference for LLM agents.

B. Benchmark Families Targeting Stopping and Tail risk

Beyond toy BNs, we propose benchmark families designed to elicit specific decision-theoretic phenomena:

- **Ambiguity regimes:** many weak signals where the main challenge is when to stop, not which single observation to request.
- **Shifted priors/Conditional Probability Tables (CPTs):** controlled distribution shifts (prevalence changes, CPT perturbations) to evaluate robustness and policy transfer.
- **Large structured BNs:** regimes where exact VOI is infeasible, enabling studies of how LLM heuristics compare when normative computation is approximate [13].

Each family can be generated at scale by varying structure, costs, and utilities, enabling statistically stable evaluation of rare catastrophic outcomes.

V. EVALUATION PROTOCOL

A. Primary Metrics

The benchmark evaluates agents as sequential decision-makers under explicit utilities:

- **Expected utility:** mean episodic return under the full reward structure.
- **Regret:** return gap versus a chosen normative baseline (e.g., myopic or rollout VOI), reported with confidence intervals.
- **Catastrophic error rates:** frequency of extreme-loss outcomes (e.g., false negatives under large penalties), complemented by tail summaries when appropriate.
- **Stopping behavior:** how often agents diagnose when a positive-VOI test exists, and how often they continue testing when VOI indicates stopping.

B. Diagnostics and Reporting

To interpret failures and mitigations, we also report:

- **Calibration vs. action:** when agents provide confidence, compare it to the BN posterior and relate miscalibration to stopping errors [21].
- **Compute and latency:** wall-clock runtime and scalability for baselines and LLM agents.
- **Interface and prompt ablations:** natural language vs. delexicalized prompts to separate semantic priors from decision competence.

We enforce constrained action formats, log full trajectories, and evaluate over many BN-sampled episodes with fixed seeds where applicable to improve reproducibility.

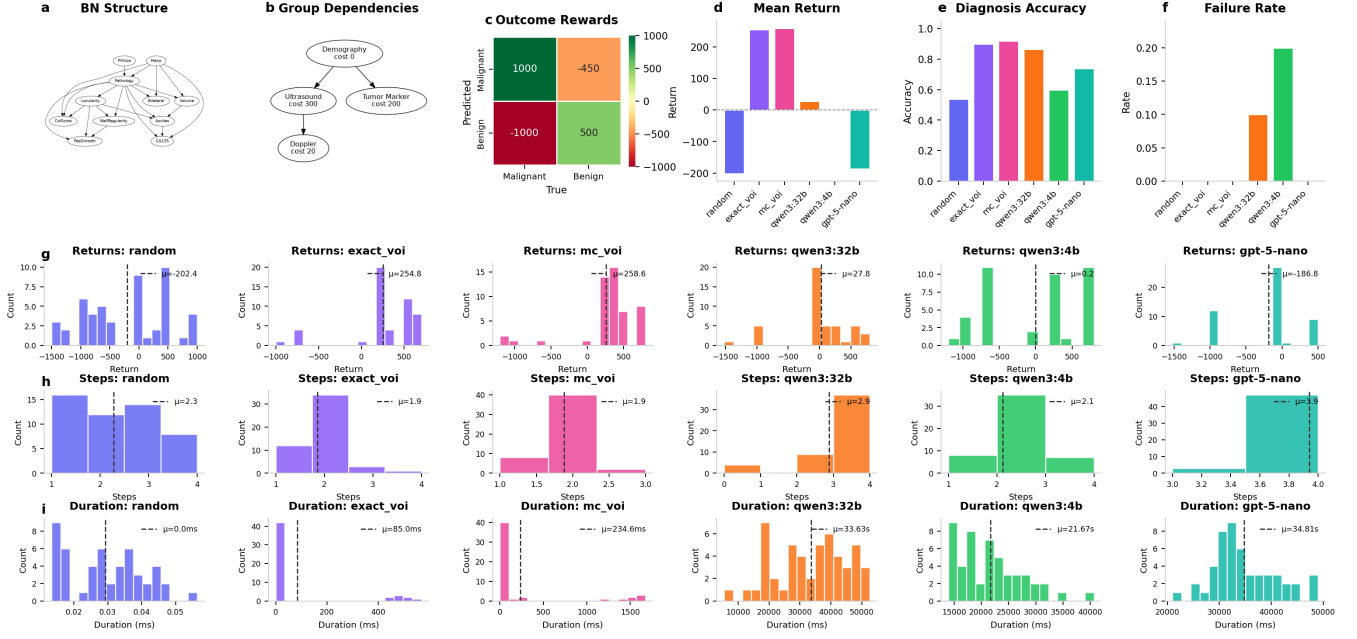


Fig. 2. *Prototype benchmark visualization and summary statistics.* (a) Bayesian network (BN) structure used as the generative world model (target: Pathology). (b) Observation-group prerequisite graph showing which test panels unlock others and their acquisition costs (e.g., Demography, TumorMarker, Ultrasound, Doppler). (c) Utility / payoff matrix for diagnostic outcomes (rows: predicted label, columns: true label), illustrating asymmetric misdiagnosis penalties. (d) Mean episodic return (expected utility) across policies. (e) Diagnostic accuracy across policies. (f) Failure rate, i.e., fraction of episodes where the agent produced an invalid (unparseable) action format under the constrained interface. (g) Return distributions (histograms) per policy with the empirical mean marked. (h) Distributions of the number of interaction steps (tests requested before stopping/diagnosis) with mean marked. (i) Wall-clock episode duration distributions with mean marked, highlighting the runtime gap between algorithmic baselines and LLM agents. The myopic VOI policy can be considered exact due to the short horizons of the environment.

VI. STATUS, ROADMAP, AND DISCUSSION

The current working prototype of BNEnv supports BN loading, episode simulation, prerequisite-constrained grouped tests, utility-based rewards, baseline VOI policies, and an LLM-facing text interaction loop. Early pilots exhibit the motivating phenomenon: utility-centric evaluation reveals failure modes that accuracy alone can mask, including bimodal return distributions driven by rare catastrophic errors.

We will extend BNEnv from the current prototype to a benchmark suite focused on ovarian diagnosis across *multiple* Bayesian networks (BNs), with richer diagnostic targets, more realistic acquisition structure, and broader model coverage. The emphasis remains decision behavior (test choice and stopping) under explicit asymmetric utilities.

a) R1: Multiple ovarian BNs over increasing complexity. Instead of a single BN, we will evaluate on a *set* of ovarian BNs spanning several complexity regimes (small/medium/large), so we can measure how policies scale from tractable settings (exact inference/VOI feasible) to harder settings where inference and VOI are approximate. Each regime will include multiple BN variants (e.g., controlled shifts in priors/CPTs) to test robustness and to obtain stable estimates of tail risk.

b) R2: Cancer subtypes beyond benign vs. malignant. We will move from a binary target to a multi-class target that includes ovarian cancer subtypes, enabling subtype-aware

utilities and explicit measurement of severe misdiagnoses in the tails. The benchmark will report both overall utility and subtype-specific catastrophic outcomes under asymmetric loss structures.

c) R3: Real-world grounded groups and costs. We will replace the current hand-crafted observation grouping and cost scheme with group definitions and acquisition costs derived from real-world motivated sources and workflows. Groups will be designed as realistic “test panels” (and constrained by prerequisites where applicable), and costs will be stored as versioned profiles separate from the BN files so we can run controlled sweeps over resource constraints and report fully reproducible settings.

d) R4: More LLMs (open and proprietary). So far, the prototype was only run with Qwen3:4B and Qwen3:32B due to limited time and funding. We will expand evaluation to a wide range of open-weight and proprietary models (e.g., GPT-family, Gemini-family), using the same constrained action interface and logging.

e) R5: Evaluation metrics emphasizing utility, regret, tail risk, and stopping. We keep expected utility as the primary metric and report regret versus a ladder of VOI-based baselines (myopic VOI $\rightarrow$ rollout/lookahead VOI where feasible). We will explicitly measure catastrophic error rates (extreme-loss outcomes), and we will characterize stopping behavior by quantifying (i) diagnoses made while positive-VOI tests remain and (ii) unnecessary continued testing when VOI

indicates stopping. For subtype targets, we additionally break down catastrophic rates and confusion patterns by subtype to localize where policies fail under asymmetric penalties.

A. Risks, Mitigations, and Ethics

Synthetic benchmark critique. BNs are controlled testbeds for decision behavior, not substitutes for clinical or field validation. We mitigate this by emphasizing delexicalized variants, robustness sweeps, and (when available) inclusion of a curated BN reflecting a real domain.

Baseline optimality and scalability. We report baselines as a ladder (myopic VOI $\rightarrow$ rollout VOI) and include compute budgets and sensitivity analyses, especially for large BNs where approximate inference is unavoidable [12], [13].

LLM variance and reproducibility. We constrain action syntax, run many episodes, log trajectories, and report uncertainty. When sampling-based decoding is used, we treat it as part of the policy and report both mean performance and tail risk.

Ethical scope. BNEnv is not a clinical decision-support system. It is an evaluation framework intended to characterize agent behavior under uncertainty; any downstream deployment requires domain-specific validation and governance.

VII. CONCLUSION

We propose benchmarking LLMs as sequential diagnosticians in a setting where a Bayesian network world model is available. BN-grounded simulation enables unlimited controlled episodes and supports normative references via posterior inference and VOI, making it possible to evaluate the safety-relevant aspects of decision behavior: utility alignment, stopping criteria, tail risk, and robustness under shifts. BNEnv provides an environment scaffold for systematic comparisons of LLM agents, VOI-based policies, and hybrid tool-augmented variants, with metrics that move beyond accuracy toward decision-theoretic reliability.

ACKNOWLEDGMENT

This research was supported by the National Research, Development, and Innovation Fund of Hungary under Grant OTKA-K139330, TKP2021-EGA-02, and the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. This research was carried out as part of E-GROUP ICT SOFTWARE Zrt.’s “IPCEI-CIS FedEU.ai – Federated Cloud-Edge AI” project, supported by the Government of Hungary and the Ministry of National Economy, under grant number NGM/4043/1/2024. The authors acknowledge the support of E-GROUP ICT SOFTWARE Zrt. for providing the research infrastructure and resources essential to this work. The research presented in this paper forms part of E-GROUP’s core research domain and contributes to the advancement of its strategic directions in federated, cloud-edge, and cutting-edge AI technologies.

REFERENCES

- [1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988, doi:<https://doi.org/10.1016/C2009-0-27609-4>.
- [2] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, UK, 2000.
- [3] F. V. Jensen and T. D. Nielsen, *Bayesian Networks and Decision Graphs*, 2nd ed. Springer, 2007, doi:<https://doi.org/10.1007/978-0-387-68282-2>.
- [4] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, “Planning and acting in partially observable stochastic domains,” *Artificial Intelligence*, vol. 101, no. 1–2, pp. 99–134, 1998, doi:[https://doi.org/10.1016/S0004-3702\(98\)00023-X](https://doi.org/10.1016/S0004-3702(98)00023-X).
- [5] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor, “Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection,” *Artificial Intelligence in Medicine*, vol. 29, no. 1–2, pp. 39–60, 2003, doi:[https://doi.org/10.1016/S0933-3657\(03\)00053-8](https://doi.org/10.1016/S0933-3657(03)00053-8).
- [6] Z. Guenfold and P. Antal, “Bayesian exploration of dependencies of laboratory tests and evaluation of test redundancy,” in *Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, Spain, 2018, pp. 1–6, doi:<https://doi.org/10.1109/PAIS.2018.8598485>.
- [7] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor, “Using literature and data to learn Bayesian networks as clinical models of ovarian tumors,” *Artificial Intelligence in Medicine*, vol. 30, no. 3, pp. 257–279, 2004, doi:<https://doi.org/10.1016/j.artmed.2003.11.007>.
- [8] W. Froyman, C. Landolfo, B. De Cock, L. Wynants, P. Sladkevicius, A. C. Testa, C. Van Holsbeke, E. Domali, R. Fruscio, E. Epstein, M. J. Dos Santos Bernardo, D. Franchi, M. J. Kudla, V. Chiappa, J. L. Alcazar, F. P. G. Leone, F. Buonomo, L. Hochberg, M. E. Coccia, S. Guerriero, N. Deo, L. Jokubkiene, J. Kaijser, A. Coosemans, I. Vergote, J. Y. Verbakel, T. Bourne, B. Van Calster, L. Valentin, and D. Timmerman, “Risk of complications in patients with conservatively managed ovarian tumours (IOTA5): a 2-year interim analysis of a multicentre, prospective, cohort study,” *Lancet Oncology*, vol. 20, no. 3, pp. 448–458, 2019, doi:[https://doi.org/10.1016/S1470-2045\(18\)30837-4](https://doi.org/10.1016/S1470-2045(18)30837-4).
- [9] R. A. Howard, “Information value theory,” *IEEE Transactions on Systems Science and Cybernetics*, vol. 2, no. 1, pp. 22–26, 1966, doi:<https://doi.org/10.1109/TSSC.1966.300074>.
- [10] D. Heckerman, J. S. Breese, and K. Rommelse, “Decision-theoretic troubleshooting,” *Communications of the ACM*, vol. 38, no. 3, pp. 49–57, 1995, doi:<https://doi.org/10.1145/203330.203341>.
- [11] D. E. Heckerman, E. J. Horvitz, and B. Middleton, “An approximate nonmyopic computation for value of information,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 3, pp. 292–298, 1993, doi:<https://doi.org/10.1109/34.204912>.
- [12] A. Krause and C. Guestrin, “Optimal nonmyopic value of information in graphical models—efficient algorithms and theoretical limits,” in *Proceedings of IJCAI*, 2005.
- [13] P. Dagum and M. Luby, “Approximating probabilistic inference in Bayesian belief networks is NP-hard,” *Artificial Intelligence*, vol. 60, no. 1, pp. 141–153, 1993, doi:[https://doi.org/10.1016/0004-3702\(93\)90036-B](https://doi.org/10.1016/0004-3702(93)90036-B).
- [14] S. Ji and L. Carin, “Cost-sensitive feature acquisition and classification,” *Pattern Recognition*, vol. 40, no. 5, pp. 1474–1485, 2007, doi:<https://doi.org/10.1016/j.patcog.2006.11.008>.
- [15] G. Contardo, L. Denoyer, and T. Artières, “Sequential cost-sensitive feature acquisition,” *arXiv preprint arXiv:1607.03691*, 2016, doi:<https://doi.org/10.48550/arXiv.1607.03691>.
- [16] H. Shim, S. J. Hwang, and E. Yang, “Joint active feature acquisition and classification with variable-size set encoding,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, doi:<https://doi.org/10.48550/arXiv.1709.05964>.
- [17] S. (et al.), “A reinforcement learning guided adaptive cost-sensitive feature acquisition framework,” *Applied Soft Computing*, 2022, doi:<https://doi.org/10.1016/j.asoc.2022.108437>.
- [18] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht, “ALFWorld: Aligning text and embodied environments for interactive learning,” in *International Conference on Learning Representations (ICLR)*, 2021, doi:<https://doi.org/10.48550/arXiv.2010.03768>.
- [19] S. Yao, H. Chen, J. Yang, and K. Narasimhan, “WebShop: Towards scalable real-world web interaction with grounded language agents,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022, doi:<https://doi.org/10.48550/arXiv.2207.01206>.

- [20] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, “ReAct: Synergizing reasoning and acting in language models,” in *International Conference on Learning Representations (ICLR)*, 2023, doi:<https://doi.org/10.48550/arXiv.2210.03629>.
- [21] S. Kadavath *et al.*, “Language models (mostly) know what they know,” *arXiv preprint arXiv:2207.05221*, 2022, doi:<https://doi.org/10.48550/arXiv.2207.05221>.
- [22] H. Nori *et al.*, “Sequential diagnosis with language models,” *arXiv preprint arXiv:2506.22405*, 2025, doi:<https://doi.org/10.48550/arXiv.2506.22405>.

Beyond Dice: Strategy-Dependent Acquisition in Active Learning for JSRT Lung Segmentation

Mohaisen Mohammed, Dániel Hadházi, Gábor Hullám
Budapest University of Technology and Economics
Department of Artificial Intelligence and Systems Engineering
Budapest, Hungary
Email: {mohaisen, hadhazi, gabor.hullam}@mit.bme.hu

Abstract—Active learning reduces annotation effort in medical image segmentation, yet strategy comparisons typically focus on global metrics such as the Dice coefficient while overlooking which images are acquired and how acquisition patterns influence performance. We present a strategy-centered analysis of active learning for lung segmentation on the JSRT dataset (247 chest X-rays), treating acquisition decisions as first-class objects of study. Three strategies are compared: Random, SDS (Segmentation Dispersion Score), and CAS (Composite Acquisition Score), using U-Net with MC-dropout. Going beyond Dice, we additionally evaluate boundary precision via HD95 and report strategy-dependent acquisition patterns, outlining a plan to establish whether acquisition diversity correlates with boundary quality and reliability.

Index Terms—active learning, medical image segmentation, JSRT dataset, uncertainty estimation, acquisition analysis

I. PROBLEM STATEMENT & MOTIVATION

Accurate lung segmentation in chest radiographs is foundational for automated diagnosis systems. Boundary errors in lung masks can lead to critical downstream failures: (i) missed nodule detection when under-segmentation excludes lesion regions, (ii) false positive pneumothorax detection when over-segmentation obscures pleural line changes, and (iii) inaccurate lung volume estimates affecting triage decisions in emergency and ICU settings. Tuberculosis is a common illness in developing countries, where subtle texture cues matter and imprecise lung segmentation can obscure them. Under-segmentation can hide apical abnormalities while over-segmentation inflates area-based thresholds, potentially delaying confirmatory testing in both cases.

Active learning promises to reduce annotation costs by intelligently selecting informative samples for expert labeling. An *acquisition strategy* is the rule that decides which unlabeled image to label next; different strategies use different notions of informativeness and therefore may select different images from the same pool. Performance is commonly judged by the Dice coefficient, a global overlap metric, yet Dice can mask clinically relevant boundary errors and tells us nothing about which *specific images are selected* or *how acquisition patterns differ between strategies*. Without examining acquisition behavior, we cannot distinguish performance gains from genuine informativeness versus initialization variance.

This work makes acquisition decisions a central object of study, asking: Do strategies select overlapping or comple-

mentary samples? Does acquisition diversity correlate with boundary quality measured by boundary-sensitive metrics such as HD95? How much early-round difference is stochastic noise versus meaningful strategy differences? These questions are essential for deploying active learning in safety-critical settings.

II. RELATED WORK

Active learning for medical image segmentation remains challenging, with no demonstrably adequate approach [1]. The difficulty lies in defining acquisition criteria that capture boundary localization failures. Uncertainty-based methods (Bayesian deep learning [2], MC-dropout [3]–[6]) dominate due to their connection to model confidence. Diversity-oriented strategies [7], [8] and hybrid approaches [9] aim to ensure broad coverage. Extension of these methods to segmentation tasks remains a research question. However, systematic validation for boundary precision and clinical safety is lacking.

Despite this diversity, few works report acquired samples or quantify inter-strategy overlap, limiting interpretability. Our work addresses this gap by reporting per-round acquisition overlap (Jaccard index, union size) alongside standard metrics, connecting strategy behavior to downstream outcomes for more clinically meaningful comparisons.

III. PROPOSED APPROACH

We conduct a controlled comparison of three active learning strategies on lung segmentation using the JSRT dataset (247 posteroanterior chest X-rays with expert-annotated lung masks) [10]. Our experimental design prioritizes reproducibility and interpretability.

A. Monte Carlo Dropout for Uncertainty Estimation

The core challenge in active learning for segmentation is identifying which unlabeled images contain boundary errors that the model cannot yet accurately predict. We employ Monte Carlo (MC) dropout [3] as a computationally efficient method for pixel-wise uncertainty quantification. By keeping dropout layers active during inference and performing multiple stochastic forward passes ($T=30$ passes), we obtain a distribution of predictions for each pixel. The variance of this

distribution serves as an uncertainty estimate: high variance indicates that the model produces inconsistent predictions across dropout configurations, suggesting unreliable segmentation in that region.

The basic idea underlying our strategies is to define heuristic measures that translate pixel-wise MC-dropout uncertainty into sample-level acquisition scores, with particular emphasis on uncertainty along predicted lung boundaries. Boundary regions are clinically critical, as errors in these areas directly affect downstream diagnostic tasks. Our heuristics aggregate spatial patterns of uncertainty to prioritize images where the model exhibits the most uncertainty in anatomically critical locations, estimating segmentation error without ground-truth annotations.

B. Acquisition Strategies

We compare three acquisition strategies with distinct selection criteria. **Random** sampling serves as a baseline, selecting samples uniformly at random without considering model uncertainty. **SDS (Segmentation Dispersion Score)** prioritizes images containing multiple spatially distinct regions of high uncertainty along predicted lung boundaries. The strategy identifies connected components of uncertain boundary pixels and combines information about how many such regions exist with the overall fraction of uncertain boundary pixels, aiming to select images where uncertainty is dispersed across multiple boundary segments, potentially indicating complex or ambiguous anatomical structures [3], [4]. **CAS (Composite Acquisition Score)** combines three complementary aspects of boundary uncertainty: the strength and spatial extent of uncertainty gradients along boundaries (capturing sharp transitions in model confidence), the degree of spatial fragmentation of uncertain regions (similar to SDS), and the smoothness characteristics of uncertain boundary segments (identifying irregular or jagged uncertain regions). These components are normalized and combined to produce a holistic assessment of boundary prediction quality.

All strategies employ the same U-Net architecture [11] with dropout layers (rate $p=0.2$) after each encoder and decoder block to enable MC-dropout uncertainty estimation. As described above, uncertainty is quantified via $T=30$ stochastic forward passes, computing pixel-wise prediction variance and aggregating it over boundary regions [3], [5].

C. Training and Deterministic Controls

All strategies share the same training pipeline so that acquisition behavior, rather than model configuration, is the primary experimental variable. To isolate stochastic effects from strategy differences, deterministic experiments fix all random seeds, weight initialization, fold assignments, and data order across strategies. Under these conditions, round-0 performance must be identical because every strategy starts from the same labeled subset and the same model state. Our results confirm this expectation (Dice 0.9680 ± 0.0007 , HD95 5.61 ± 0.65 mm at round 0), validating the deterministic controls. Any performance differences observed in later rounds

can therefore be attributed to the distinct samples acquired by each strategy rather than to initialization variance.

D. Acquisition Analysis

For each round, we systematically log selected sample IDs for each strategy, pairwise overlap metrics including intersection size and Jaccard index, union size across all strategies, and per-sample error trajectories measured by Dice and HD95 across rounds. This comprehensive acquisition-level data enables us to test whether diversity in sample selection correlates with performance improvements and helps us understand the strategic differences in data exploration patterns.

E. Metrics

We report Dice coefficient and 95th percentile Hausdorff distance (HD95) as complementary measures of segmentation quality [12], [13]. The Dice coefficient measures region overlap and is computed as $\text{Dice} = 2|P \cap G|/(|P| + |G|)$, where P is the predicted mask and G is the ground truth. While intuitive and widely used, Dice is not sensitive to localized boundary errors because it aggregates all pixels equally, which can mask small boundary deviations. A segmentation with a small but clinically critical boundary error (e.g., a 5 mm under-segmentation at the lung apex where tuberculosis lesions often appear) may still achieve a high Dice score if the overall mask area is large. This averaging behavior makes Dice insufficient for assessing boundary precision in safety-critical applications.

To address this limitation, we also report HD95, the 95th percentile of distances from predicted boundary pixels to the nearest ground-truth boundary pixel (and vice versa). HD95 is boundary-sensitive: it directly quantifies the largest systematic deviations in boundary localization while being robust to small outliers (by excluding the worst 5% of distances). A low HD95 value (e.g., ≤ 5 mm) indicates that predicted boundaries are tightly aligned with ground truth across most of the contour, whereas a high HD95 reveals localized boundary failures even when Dice is high. By reporting both metrics, we ensure that our evaluation captures both global overlap quality (Dice) and boundary localization precision (HD95), which are essential for clinical deployment.

F. Research Questions

RQ1: How can we define heuristic measures that effectively estimate segmentation error in unlabeled images by utilizing MC-dropout-based pixel-wise uncertainty? Specifically, which spatial patterns and aggregations of boundary uncertainty correlate most strongly with actual segmentation errors?

RQ2: Is segmentation error estimation via uncertainty an adequate metric to order samples when increasing the training set size? Does prioritizing samples with high estimated boundary error lead to faster improvement compared to random acquisition?

RQ3: How much do different uncertainty-based strategies select overlapping versus complementary samples, and does this acquisition diversity correlate with performance differences?

RQ4: How much early-round performance variability is due to stochastic initialization versus true strategy differences, and do these strategy differences generalize across budgets and datasets beyond JSRT?

Hypotheses: (H1) Heuristics that capture spatial dispersion and boundary-specific uncertainty patterns will more accurately estimate segmentation error than global uncertainty measures, leading to more effective sample prioritization. (H2) Strategies with lower pairwise Jaccard overlap in early rounds will exhibit faster HD95 convergence due to broader exploration of boundary error modes.

IV. EXPECTED CONTRIBUTIONS

A. Completed Work

Deterministic round-0 cross-validation showing identical performance across strategies (Dice 0.9680 ± 0.0007 , HD95 5.61 ± 0.65 mm) confirms that early differences can arise from stochastic variation (Fig. 1). Per-round acquisition overlap analysis reveals low agreement (Jaccard 0.00–0.16, mean 0.09) across strategies (Table I). Visual diagnostics for high-variance samples in non-deterministic round-0 runs (Fig. 2) and round-0 versus round-5 segmentation comparisons (Fig. 3) demonstrate boundary error reduction.

B. Planned Work

We will extend deterministic cross-validation across all rounds (0–5) and multiple budget settings (10, 15, 20, 25 samples acquired per round). We will validate findings on an additional dataset (e.g., Montgomery County or Shenzhen chest X-ray datasets) to test generalization and examine additional strategies. We will compare acquisition diversity (Jaccard/union) with HD95 reduction rate to assess their relationship. For reproducibility, we will release open-source resources: code, trained models, per-round acquisition lists (with sample IDs and random seeds), and visual overlays.

V. EVALUATION PLAN

A. Quantitative Evaluation

Performance metrics include Dice and HD95 evaluated on held-out validation and test sets, providing complementary views of segmentation quality. Acquisition metrics track intersection size, Jaccard index, and union size per round to quantify inter-strategy overlap. Statistical testing employs paired comparisons with deterministic reruns and Wilcoxon signed-rank tests [14] for non-normal distributions. Effect sizes are reported using Cohen’s d alongside p -values to assess practical significance beyond statistical significance.

B. Qualitative Evaluation

TP/FP/FN overlays visualize error modes (e.g., apical under-segmentation, basal over-segmentation). Per-case Dice and HD95 trajectories highlight cases with the largest inter-strategy disagreement. Cross-fold consistency checks assess stability. All results include 95% confidence intervals and effect size estimates to support informed interpretation beyond binary significance testing.

VI. CURRENT STATUS & TIMELINE

A. Current Status

Deterministic round-0 cross-validation is complete, and all strategies achieve identical performance with fixed initialization and fixed fold partitioning. Non-deterministic round-0 analysis uses different random initializations (seeds not synchronized across runs) while keeping the same initial labeled set and folds, and reveals large case-specific performance gaps (e.g., JPCNN068: Random Dice 0.97 vs. SDS 0.93 vs. CAS 0.91), highlighting stochastic sensitivity. Round-0 versus round-5 overlays show consistent boundary tightening across strategies (Fig. 3). Low-budget convergence analysis (Table II) shows all strategies reach similar performance by round 5 (Dice ≈ 0.975). Acquisition overlap remains low across all rounds (Table I), indicating fundamentally different exploration patterns.

B. Timeline

Jan–Feb 2026: Complete deterministic cross-validation across rounds 1–5; analyze acquisition-performance correlations across budget settings. **Mar–Apr 2026:** Extend experiments to second dataset; analyze cross-dataset generalization of acquisition patterns and error modes. **May–Jun 2026:** Finalize statistical analysis; draft journal paper; prepare open-source release.

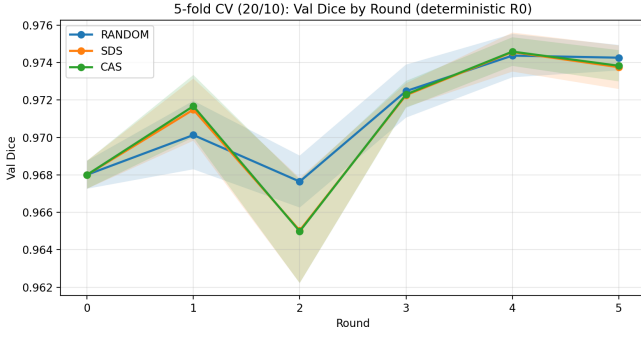
Compute resources: Institutional GPU cluster access secured; low-budget prioritization ensures feasibility within timeline.

VII. DISCUSSION & RISKS

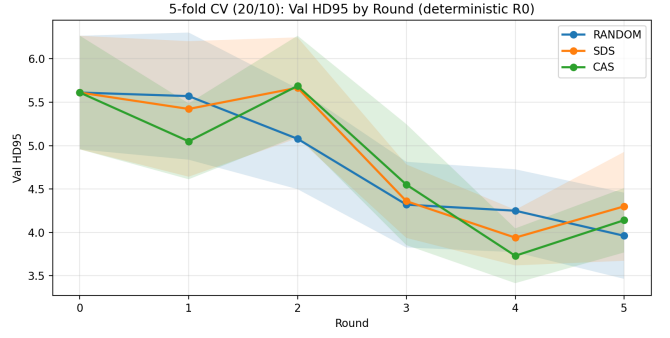
The minimal acquisition overlap (Jaccard 0.00–0.16) observed across strategies constitutes a key finding with several important implications. Different active learning strategies utilize fundamentally different subsets of the data space despite operating on the same unlabeled pool. This suggests that “uncertainty” is not a singular, strategy-independent property of an image but rather is shaped by the mathematical formulation of each acquisition function. The low inter-strategy agreement demonstrates that each strategy’s notion of informativeness differs substantially, leading to different data exploration patterns.

This finding has direct implications for ensemble-based active learning approaches. If different uncertainty-based strategies systematically select disjoint samples, then combining multiple acquisition functions may yield more comprehensive data coverage than relying on a single strategy. The high union sizes (60–67 unique samples out of 75 total slots per round) confirm that over 80% of each round’s acquired budget explores different regions of the feature space depending on the strategy employed. This diversity in acquisition behavior warrants further investigation to determine whether it correlates with improved robustness and generalization, particularly in safety-critical medical applications where both average performance and worst-case reliability matter.

The deterministic controls implemented in our experiments provide strong evidence that early-round performance differences can arise purely from stochastic variation in weight



(a) Validation Dice by round (deterministic round-0).



(b) Validation HD95 by round (deterministic round-0).

Fig. 1: Deterministic round-0 cross-validation curves. Identical round-0 metrics demonstrate early differences arise from stochastic variation. SDS shows temporary HD95 spike at round 2.

initialization rather than genuine strategy effectiveness. Our round-0 results show identical performance across all strategies (Dice 0.9680 ± 0.0007), validating that differences observed in non-deterministic runs reflect initialization variance. This underscores the importance of controlling for stochasticity when comparing active learning strategies and suggests that reported performance gains in the literature may sometimes be confounded by random effects.

A. Risks & Mitigation

Clinical risk from boundary errors is mitigated through HD95-focused evaluation, qualitative overlays showing true/false positives and negatives, and per-case error tracking that identifies worst-case failures. Stochastic variance is controlled via deterministic reruns with synchronized random seeds and cross-seed analysis to separate genuine strategy effects from initialization artifacts. Dataset bias is addressed through planned cross-dataset validation on Montgomery County and Shenzhen chest X-ray datasets [15], ensuring that findings generalize beyond JSRT. Computational constraints are managed through institutional GPU cluster access, post-hoc evaluation strategies, and prioritization of low-budget settings that reflect realistic annotation scenarios.

Future work will investigate the necessity of modifying acquisition strategies based on dataset characteristics, clinical context, and budget constraints to ensure optimal performance across different scenarios. If acquisition overlap does not correlate with performance improvements, we will report this as a negative result that clarifies current assumptions in active learning practice and informs future method design. Such transparency is essential for advancing the field beyond optimistic bias in published results.

TABLE I: Acquisition overlap per round. Strategies select 25 samples (rounds 0–4) or 14 (round 5). Overlap counts samples selected by all; Union counts the number of different samples; Jaccard indices measure pairwise agreement (R=Random, S=SDS, C=CAS). Union sizes 60–67 (of 75 total) show over 80% acquired budget is unique.

Round	Overlap	Union	Jac. R–S	Jac. R–C	Jac. S–C
0	1	63	0.163	0.064	0.064
1	0	62	0.064	0.087	0.136
2	1	62	0.087	0.111	0.111
3	0	67	0.064	0.020	0.087
4	3	60	0.163	0.136	0.111
5	0	39	0.000	0.000	0.120

TABLE II: Round-5 performance on JSRT (initial 20 labels, budget 10/round). All strategies converge to similar performance.

Strategy	Val Dice	Test Dice
Random	0.9774	0.9771
SDS	0.9753	0.9746
CAS	0.9755	0.9734

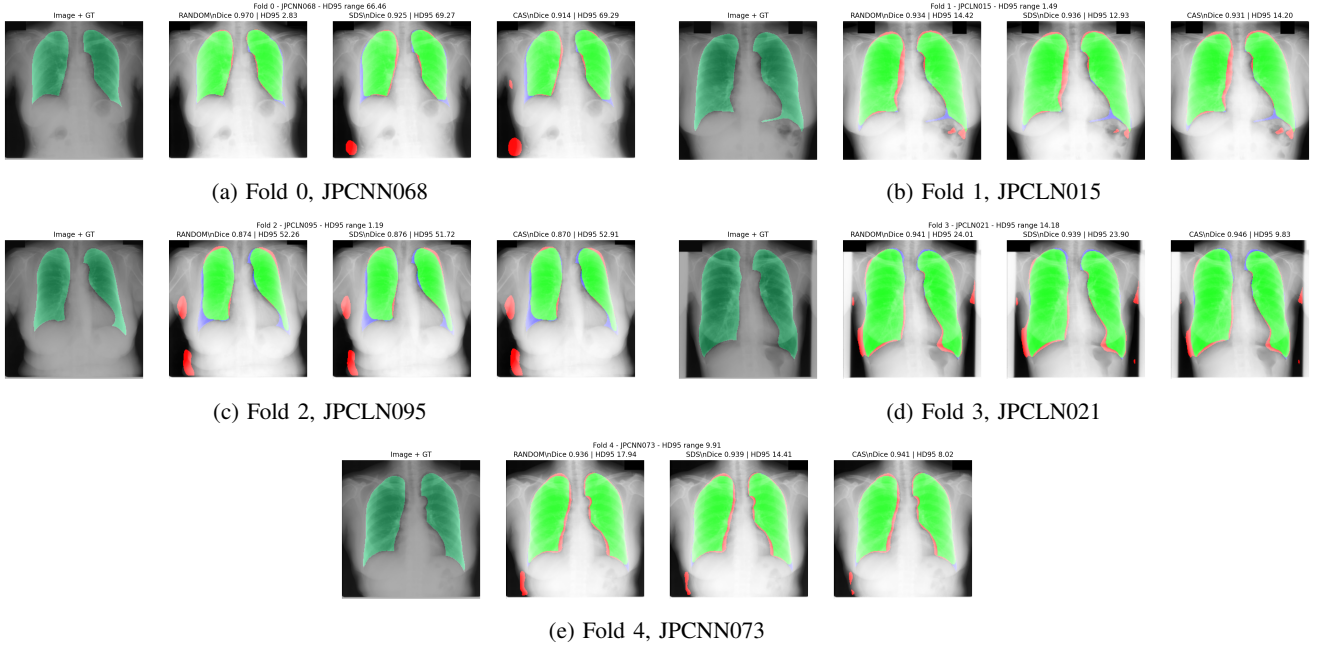


Fig. 2: Non-deterministic round-0 segmentation overlays (TP green, FP red, FN blue) for high-variance samples, demonstrating importance of controlling stochastic initialization.

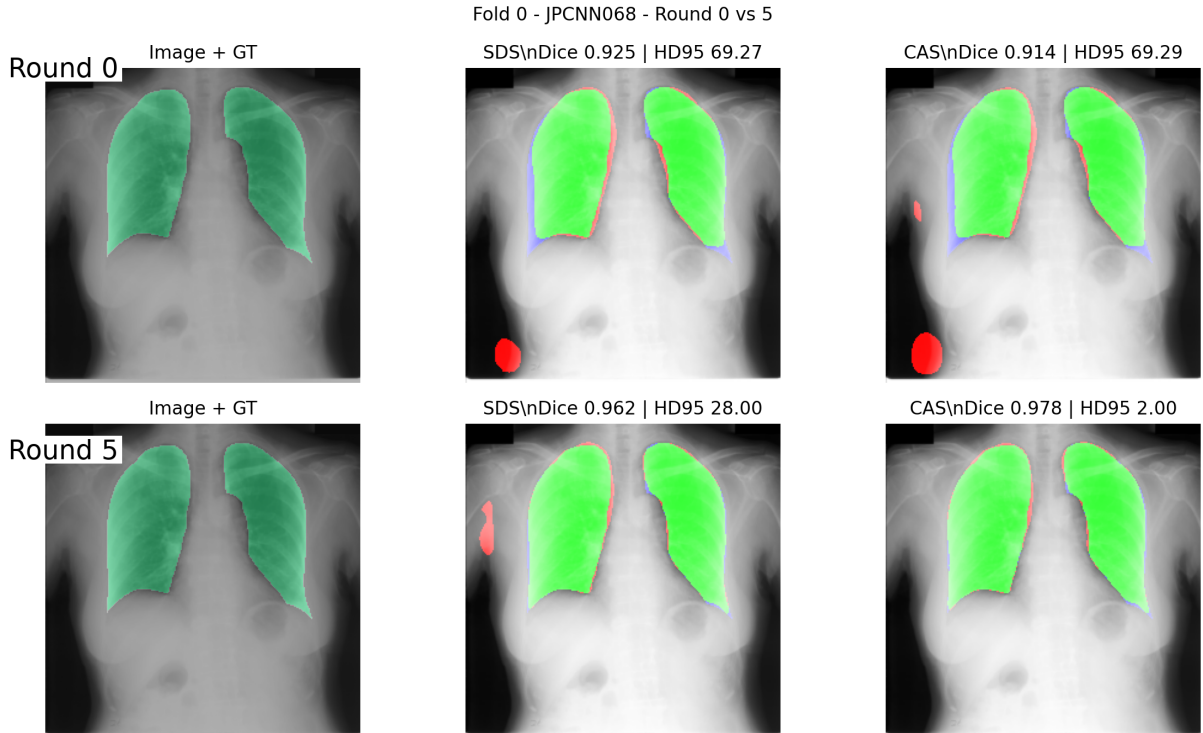


Fig. 3: Round-0 vs. round-5 segmentation comparisons (JPCNN068, Fold 0). Upper row: round 0; lower row: round 5. Columns: Image+GT, SDS, CAS. Round 5 shows boundary error reduction.

REFERENCES

- [1] H. Wang, Q. Jin, S. Li, S. Liu, M. Wang, and Z. Song, "A comprehensive survey on deep active learning in medical image analysis," *Medical Image Analysis*, vol. 95, p. 103201, Jul. 2024. [Online]. Available: <http://dx.doi.org/10.1016/j.media.2024.103201>
- [2] B. Ghoshal, S. Swift, and A. Tucker, *Bayesian Deep Active Learning for Medical Image Analysis*. Springer International Publishing, 2021, pp. 36–42. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-77211-6_4
- [3] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning (ICML)*. PMLR, 2016, pp. 1050–1059.
- [4] V. Nath, D. Yang, B. A. Landman, D. Xu, and H. R. Roth, "Diminishing uncertainty within the training pool: Active learning for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2534–2547, Oct. 2021. [Online]. Available: <http://dx.doi.org/10.1109/TMI.2020.3048055>
- [5] S. Ma, H. Wu, A. Lawlor, and R. Dong, "Breaking the barrier: Selective uncertainty-based active learning for medical image segmentation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2024, pp. 1531–1535. [Online]. Available: <http://dx.doi.org/10.1109/ICASSP48485.2024.10446026>
- [6] G. Thakur, "Investigate uncertainty estimation in transfer learning based active learning for medical image segmentation," in *2024 3rd International Conference for Innovation in Technology (INOCON)*. IEEE, Mar. 2024, pp. 1–7. [Online]. Available: <http://dx.doi.org/10.1109/INOCON60754.2024.10511579>
- [7] M. Gaillochet, C. Desrosiers, and H. Lombaert, "Active learning for medical image segmentation with stochastic batches," *Medical Image Analysis*, vol. 90, p. 102958, Dec. 2023. [Online]. Available: <http://dx.doi.org/10.1016/j.media.2023.102958>
- [8] A. Choi, J. Joo, F. Scalzo, Y. Sun, A. Vepa, and Z. Yang, "Integrating deep metric learning with coreset for active learning in 3d segmentation," in *Advances in Neural Information Processing Systems 37*, ser. NeurIPS 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024, pp. 71 643–71 671. [Online]. Available: <http://dx.doi.org/10.52202/079017-2288>
- [9] X. Li, M. Xia, J. Jiao, S. Zhou, C. Chang, Y. Wang, and Y. Guo, "Hal-ia: A hybrid active learning framework using interactive annotation for medical image segmentation," *Medical Image Analysis*, vol. 88, p. 102862, Aug. 2023. [Online]. Available: <http://dx.doi.org/10.1016/j.media.2023.102862>
- [10] M. Chavan, V. Varadarajan, S. Gite, and K. Kotecha, "Deep neural network for lung image segmentation on chest x-ray," *Technologies*, vol. 10, no. 5, p. 105, Sep. 2022. [Online]. Available: <http://dx.doi.org/10.3390/technologies10050105>
- [11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.
- [12] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel, "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, p. 102035, Jul. 2021. [Online]. Available: <http://dx.doi.org/10.1016/j.media.2021.102035>
- [13] Y. Kim, D. Kang, Y. Mok, S. Kwon, and J. Paik, "Bidirectional meta-kronecker factored optimizer and hausdorff distance loss for few-shot medical image segmentation," *Scientific Reports*, vol. 13, no. 1, May 2023. [Online]. Available: <http://dx.doi.org/10.1038/s41598-023-35276-4>
- [14] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
- [15] S. Jaeger, A. Karargyris, S. Candemir, J. Siegelman, L. R. Folio, S. Antani, and G. Thoma, "Automatic tuberculosis screening using chest radiographs," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 233–245, 2014.

Model-Based Artificial Intelligence Approach for Speed and Position Estimation of PMSM

Benedek Pour¹, Huba Németh², Tamás Dabóczi¹

¹ *Department of Artificial Intelligence and Systems Engineering (MIT)
Faculty of Electrical Engineering and Informatics (VIK)
Budapest University of Technology and Economics (BME),
Budapest, Hungary
[pour, daboczi]@mit.bme.hu*

² *Knorr-Bremse Research and Development Center, Budapest, Hungary
huba.nemeth@knorr-bremse.com*

Abstract—This document summarizes the possible methods for improving sensorless position estimation of permanent magnet synchronous motors using both traditional model-based and artificial intelligence methods, as well as the research vision.

Keywords—permanent magnet synchronous motor, sensorless, position estimation, sliding mode observer, model-based signal processing, artificial intelligence, neural network

I. PROBLEM FORMULATION

Thanks to significant technological advances in semiconductor-based power electronics over the past few decades, electric drives have become a key component of energy-efficient equipment. Among the various types of electric rotating machines, permanent magnet synchronous motors (PMSM) are currently the most widely used. They can be operated with very good efficiency thanks to their relatively high energy density. However, to achieve adequate efficiency, it is important to keep a 90° angle between the electromagnetic field generated by the voltage applied to the stator of the motor and that generated by the rotor's permanent magnet, throughout the operation. The PMSM is a three-phase rotating machine. To ensure sufficient circular symmetry, the phases are shifted by 120° relative to each other. In this way, three sinusoidal voltages are required to create the rotating electromagnetic field, with phases also shifted by 120° relative to each other. The drive created in this way is a difficult-to-control, time- and speed-dependent system with a high degree of non-linearity. In the early 1980s, the widespread use of microcontrollers made it possible to use complex AC drives. As a result, field-oriented control (FOC) was developed, which simplified the three-phase system into a two-coordinate time-invariant linear problem using two transformations. The Clarke transformation projects the three a-b-c phases, which are offset by 120° relative to each other, into a rectangular stationary α - β coordinate system. The Park transformation projects the resulting fixed-axis rectangular coordinate system onto a rotating rectangular d-q coordinate system utilizing the exact angle of the rotor. The coordinate systems described are shown in Figure 1.

With the help of the two transformations, the three-phase description can be split into active ("direct") and reactive ("quadrature") components, where the d direction corresponds to the direction of the rotor's permanent magnetic flux, and the

q direction is the component perpendicular to it. It follows that the current output in the reactive direction is responsible for accelerating the rotor, while the current output in the active direction primarily plays a role in field weakening [1].

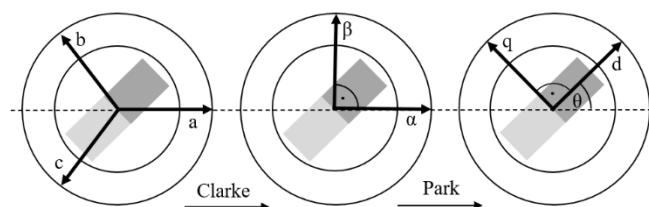


Fig. 1. FOC coordinate systems

Knowing the angle of the rotor is essential for precise drive control. This can be easily measured using absolute angle sensors that exploit various physical phenomena. However, in many applications this solution is not feasible due to its high cost. Nevertheless, some parameters of the PMSM are dependent on the rotor position, so the angle information can be determined by measuring these quantities and using known physical laws. This principle is called sensorless measurement technology. The term "sensorless" refers to the fact that the quantity to be observed is not measured directly with a sensor, but rather estimated from additional effects, provided that the relationships between these effects are known analytically. This, of course, requires the observation of additional (usually several) physical processes with sensors.

In the field of sensorless PMSM drives, we distinguish between two separate operating modes. One is the stationary standstill position and the speed range below 10% of the nominal speed. In this case, the typical drive method is an open-loop starting, where the rotating magnetic field is controlled directly [2]. This is possible in applications that allow less efficient starting. Another solution is to utilize the angle dependency of the coil inductance for position estimation and providing field-oriented control based on that. Solutions used to measure inductance basically apply additional test signal injection [3]. The next range of operation is the higher speed region where the rotor angular velocity is greater than 10% of the nominal value. In this case, the angular dependence of the electromotive force generated by the motor can be used for position estimation [4].

PMSM drives basically implement physical relationships describing a simple motor model. These lack the description of a number of significant effects, such as "cogging torque", the imperfection of the mechanical components of the system, and the temperature, position and torque dependence of the parameters describing the electric machine. In practice, the linearized models described above can be used to achieve adequate efficiency over most of the operating range. However, at working points outside of this range, it is necessary to introduce completely unique methods that take the specific system into account. Optimization requires lengthy engineering effort and many iterations. In addition, in many cases, the entire system may change during its lifetime, and the pre-tuned control may no longer give the desired results. These effects justify the need to refine known methods and to introduce various adaptive algorithms.

II. TRADITIONAL MODEL-BASED APPROACH

This research examines the possibility of improving sensorless speed and position estimation in the range above 10% of the nominal speed of the PMSM. Traditional methods exploit the position dependence of the electromotive force generated by the electric machine. The exact relationship is contained in the system of differential equations describing the behavior of the PMSM. The phase voltages in a surface-mounted magnetic PMSM in a coordinate system fixed to the stator can be written as follows:

$$\begin{bmatrix} v_\alpha \\ v_\beta \end{bmatrix} = \begin{bmatrix} R + pL_s & 0 \\ 0 & R + pL_s \end{bmatrix} \begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} + \begin{bmatrix} e_\alpha \\ e_\beta \end{bmatrix} \quad (1)$$

$$\begin{bmatrix} e_\alpha \\ e_\beta \end{bmatrix} = \omega_r \lambda_{PM} \begin{bmatrix} -\sin \theta_r \\ \cos \theta_r \end{bmatrix} \quad (2)$$

where v_α and v_β are stator voltages in the stationary α - β frame, i_α and i_β are stator currents in the stationary α - β frame, e_α and e_β are voltages derived from back electromotive forces in the stationary α - β frame, R is phase resistance, L_s is phase inductance, p is differential operator, λ_{PM} is permanent magnet flux linkage, ω_r is rotor angular velocity, and θ_r is rotor position relative to the stationary frame.

The rotor position can be determined by equation (3):

$$\hat{\theta}_r = -\tan^{-1} \left(\frac{\hat{e}_\alpha}{\hat{e}_\beta} \right) \quad (3)$$

where $\hat{\theta}_r$ is the estimated rotor position, $\hat{e}_\alpha$ and $\hat{e}_\beta$ are the estimated back EMF signals of SPMSM calculated from equation (1) [4].

This calculation alone has very limited applicability in determining position. During dynamic control of the system, the estimation error varies significantly in each operating phase, and the direct use of measurements results an estimation with high variance, because measurement noise is greatly amplified. Filtering the calculated position may be considered, but the desired accuracy can only be achieved at the cost of significant delay. For this reason, several well-known robust estimators have been developed for the problem.

The most used solution is the sliding mode observer (SMO). The mathematical model of the PMSM described in equation (1) can be rewritten to replace the voltages derived

from the generated electromotive force with a switching function:

$$\begin{cases} L_s \frac{di_\alpha}{dt} = -R\hat{i}_\alpha + u_\alpha - k \cdot \text{sgn}(\hat{i}_\alpha - i_\alpha) \\ L_s \frac{di_\beta}{dt} = -R\hat{i}_\beta + u_\beta - k \cdot \text{sgn}(\hat{i}_\beta - i_\beta) \end{cases} \quad (4)$$

where k is the constant observer gain and $\text{sgn}(x)$ is the sign function. Thus, the equation system of the SMO is determined. The method uses the physical laws of PMSM to estimate the currents transformed into an orthogonal stationary coordinate system with the help of a mathematical model. The voltage generated by the electric machine can be calculated as a function of the error signal utilizing the estimated and measured currents. The classic SMO architecture is shown on Fig. 2.

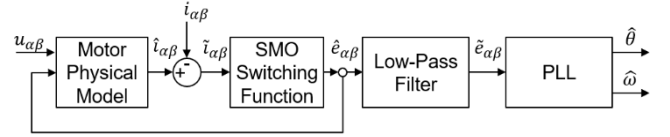


Fig. 2. Conventional SMO architecture

The described switching function plays a direct role in the robustness and fast convergence of the method [5]. The output of the SMO switching function is a pulse width modulated representation of the generated electromotive voltage. Using low pass filter the two sinusoidal signals described by (2) can be produced. We know that these are corresponding angular functions of the rotor position, which can be estimated from these signals using a phase-locked loop (PLL). The phase detector estimates the angular velocity from an error signal generated by a trigonometric identity using a PI structure. From this, the rotor position can be derived by integrator, which is attached back to the input of the trigonometric identity to determine the error signal. This closed-loop structure provides a robust and accurate estimate of angular velocity and position. The basic architecture is shown in "Fig. 3".

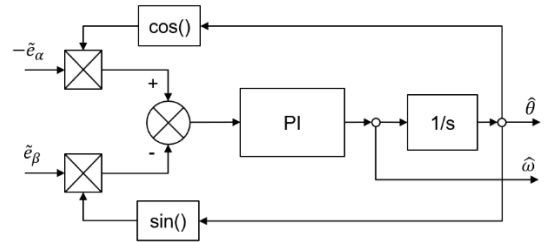


Fig. 3. Basic PLL architecture

The described solution calculates the position and speed signals based on a mathematical model that describes the physical laws of PMSM and estimates them based on the difference between the measured and estimated currents. However, SMO and similar model-based approaches do not include the distortions observed in many real systems, so their estimates are inaccurate in certain operating ranges, which can significantly reduce the achievable stability and efficiency.

III. STATE OF THE ART HYBRID SOLUTIONS

Numerous research results have been published to improve the position estimation algorithms described above, supplementing the traditional model-based approach with

artificial intelligence methods that have become increasingly widespread in recent decades. The following section presents a few published architectures that supplement mathematical models describing physical relationships with neural networks (NN), enabling the universal, data-based modeling of unknown distorting effects.

The solution resulting in the most general model is to replace the complete traditional SMO estimator described in Section 2 with a neural network. Since the goal is to model a dynamic system, a recurrent neural network (RNN) with memory was selected. This is a black-box approach that can estimate the signals corresponding to the generated voltage from the phase voltages and currents. That is, according to equation (2), the sinusoidal functions of the position information. From these signals, the position and angular velocity of the rotor can be determined using PLL [6]. A block diagram of the solution is shown in Figure 3.

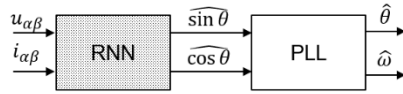


Fig. 4. Blackbox approach with model-based output

The following method retains the model-based constraints, replacing the low-pass filter and PLL from the traditional SMO architecture with RNN. The rotor position and speed are thus determined by a neural network, whose inputs are the measured phase currents and voltages, as well as the estimated generated voltages produced by the switching function of the SMO architecture [7]. The block diagram of the solution is shown in Figure 4.

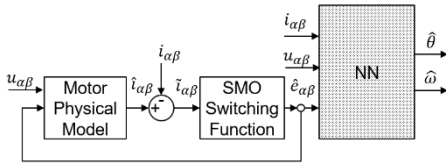


Fig. 5. Blackbox approach with model-based input

The above described two methods [5, 6] train the neural networks offline using simulation or real measurement data. The network can achieve proven good results in the operating and dynamic ranges covered during training. Therefore, training the method on a large amount of data is essential.

The next solution aims to correct the estimated outputs of conventional SMO using a neural network. This application integrates a feedforward neural network that contains significantly fewer neurons than the RNNs used in algorithms [5, 6]. The error function used for online learning consists of the difference between its estimated position and the reference position. This hybrid solution is therefore essentially designed to adaptively correct the mechanical uncertainty of servo applications that are not described by physical relationships [8]. A block diagram of the solution is shown in Figure 5.

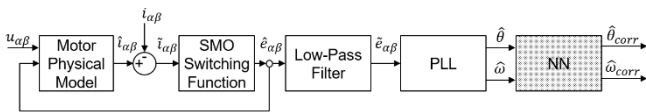


Fig. 6. Model-based approach with correction NN

Depending on the implementation, the switching functions of traditional SMO contain constant parameters that are tuned during development. However, it can be observed that with a

fixed set of parameters, the position and velocity estimation errors differ in different operating ranges. The next two solutions [9, 10] are intended to overcome this limitation. The feedforward neural network used in the study [9] determines the value of the switching function parameter based on the square mean of the current error signal. The network is trained offline based on data generated in simulation. This method integrates the determination of unknown physical relationships of a model parameter using artificial intelligence, thereby improving the efficiency of the classical structure. A block diagram of the solution is shown in Fig. 6.

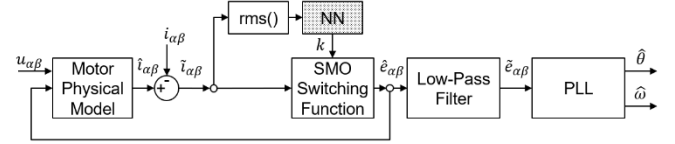


Fig. 7. Model-based description improved by NN

The method used in the research [10] introduces not only the phase current error but also the absolute value of the estimated angular velocity into the neural network that determines the parameters of the switching function. This creates a mathematical model that describes more complex relationships than the previous method. Training is online by minimizing the quadratic cost function according to expression (5), where k_1 , k_2 are the outputs of the net, k'_1 , k'_2 are the current values of parameters and w_1 , w_2 are weighting factors.

$$\min F = w_1(k_1 - k'_1) + w_2(k_2 - k'_2) \quad (5)$$

This allows the hybrid algorithm to reduce the chattering phenomenon in an adaptive way according to changes in the real system, thus increasing the accuracy of the angle and speed estimation over a wide operating range. A block diagram of the solution is shown in Fig. 7.

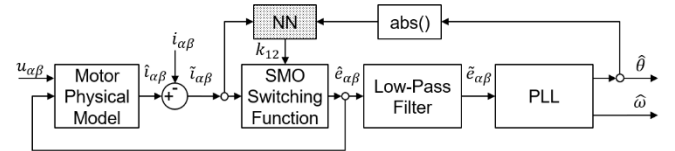


Fig. 8. Adaptive NN aided SMO

IV. PROPOSED APPROACH

The aim of our research is – similar to the methods described in the previous chapter – to combine the advantages of the traditional model-based approach and artificial intelligence methods to estimate the position and angular velocity of the PMSM rotor as accurately as possible.

It is known that significant improvements can be achieved by selecting the correct switching function [5, 8] and parameterizing it appropriately [9, 10]. One solution is a properly selected filter architecture that can produce an estimate of the generated voltage from the switched signal without delay. It is also known that the PLL at the end of a traditional SMO-based position estimator gives better results if the estimated error signal arriving at its input is normalized [11]. Based on these previous results, a non-linear switching function will be developed with the generated electromotive voltage on its output. It will be adaptively tuned by machine learning.

The literature describes an adaptive estimator that can determine motor parameters online based on the Lyapunov stability criterion applied to the conventional SMO method [12]. At the same time, it is possible to estimate the motor parameters quickly using artificial intelligence methods [13]. The present research will develop a model containing the well-known physical relationships of PMSM, whose exact structure or parameters will be adapted online by machine learning.

The development of the researched method is expected to yield numerous results, the knowledge of which is essential for hybrid model building.

- One of the key questions is how the known model-based approach, in this case SMO, can be supplemented with a relatively small neural network so that it is able to learn the hard-to-define properties and physical relationships of the real system. During the present research a specific architecture will be developed, which will increase the accuracy and robustness of the model-based method with artificial intelligence addition.
- When using artificial intelligence methods, choosing the right error function is an important task. In addition to defining the basic error, several stability criteria can also be introduced. These can be used to increase the convergence speed on the one hand, and on the other hand, to avoid the learning process becoming stuck at a local minimum. During the present research to train the algorithm an appropriate error function will be determined.
- Since the algorithm to be developed is an adaptive estimator, it is expected to be capable of fast learning during runtime. The appropriate teaching method also contributes to increasing the convergence speed and avoiding local minima. Part of the research is the selection of the appropriate teaching method required by the developed solution, which guarantees the rapid convergence required for the adaptive system.
- The SMO switching function block can be considered as a serial compensator that generates the appropriate output from the difference between the estimated and measured phase currents. If such a system is defined using a neural network, the vanishing gradient problem arises during weight training in a steady state. The basic idea of training the parameters of a serial compensator as a neural network has already been researched [14]. In addition, the present research will address the solutions to the problems arising during the teaching of the hybrid model.

V. EVALUATION PLAN

The stated expectation is that the hybrid model under investigation should be capable of approximating the nonlinear real system more accurately than the methods known to date. There are basically two main approaches for validating the developed algorithm.

A. Validation on Simulation

One possible solution is to model the real system and investigate the researched algorithm in a simulation environment. The biggest advantage of this method is that the investigated method can be validated quickly and conveniently over a wide operating range. However, it must

be considered that the real nonlinearity and distortion, which hybrid modeling is basically intended to describe, has to be modeled mathematically. Therefore, approximation of the real behavior is quite limited and difficult. For this reason, only the accuracy and robustness of basic operating structures can be demonstrated via simulation validation.

B. Experimental Validation

Another possible solution for validation is to drive and measure a real system. In this case, we need a complete electric machine drive and a digital controller that processes the measurement data and angle sensor information. The advantage of this method is that the hybrid model under investigation can be validated to determine real nonlinearity and distortion. The disadvantage is that during validation, only limited operating structures can be implemented depending on the drive chain and test equipment. If the rotor position and speed estimation algorithm under investigation is not part of the drive chain, validation can also be performed offline by saving the measurement data. However, if it is part of the drive chain, it is necessary to implement it in embedded environment and to run it in the embedded digital controller.

During validation, it is necessary to prove that the developed algorithm is fundamentally correct and that the hybrid model can approximate the deviations arising from the nonlinearity of the system within a reasonable adjustment time. Furthermore, the architecture of the integrated neural network, the cost function, the training method, and the training algorithm must be correctly selected so that the artificial intelligence method be operational and provide the expected added benefit.

VI. CURRENT STATUS AND TIMELINE

The research conducted so far includes discovery of traditional model-based rotor position estimates of electric motors and exploration of previous results of speed estimation. It also includes extensive literature research on the hybrid use of traditional model-based descriptions and artificial intelligence methods.

During the research, a MATLAB Simulink simulation environment was created in which an electric machine controlled by a cascade PI structure can be investigated in closed loop with several different sensorless traditional SMO position estimators. PMSM was modeled using classical motor equations containing physical relationships. Additive noise can be superimposed to the signals. The complex distortions characteristic for a real system should be implemented according to the needs of the research.

In addition, a framework implemented in Python has been created that can run any traditionally implemented model-based description and algorithm containing a neural network on simulation or real measurement data. During optimization, it is possible to use well-known error functions and learning algorithms, and unique solutions can also be defined.

The next step is to define the precise hybrid model, to resolve the expected contributions and to validate the model in simulation.

The research is primarily conducted within the framework of Knorr-Bremse's ongoing projects. For this reason, validation on the actual system depends on the company's infrastructure and its current availability. For the time being, the sensorless electric machine drive implemented in the

MATLAB Simulink environment drives a high-voltage 5 kW motor and the load motor of the test bench consisting of the same equipment. Currently, a rotor position sensor cannot be integrated into the digital controller, so the hybrid model used in the research cannot be validated exactly. It is expected that an electric machine drive containing a position sensor will be available soon that allows the validation on the real system.

ACKNOWLEDGMENT

T. Dabóczy acknowledges the financial support of National Laboratory for Renewable Energy (Project no. RRF-2.3.1-21-2022-00009), implemented with the support provided by the Recovery and Resilience Facility of the European Union within the framework of Programme Széchenyi Plan Plus.

REFERENCES

- [1] M. B. B. Sharifian, T. Herizchi, K. G. Firouzjah, "Field Oriented Control of Permanent Magnet Synchronous Motor Using Predictive Space Vector Modulation", IEEE Symposium on Industrial Electronics and Applications, 2009, [10.1109/ISIEA.2009.5356385](https://doi.org/10.1109/ISIEA.2009.5356385)
- [2] M. Wang, Y. Xu, J. Zou, H. Lan, "An Optimized I-F Startup Method for BEMF-based Sensorless Control of SPMSM", IEEE Transportation Electrification Conference and Expo, 2017, [10.1109/TEEC-AP.2017.8080874](https://doi.org/10.1109/TEEC-AP.2017.8080874)
- [3] R. Ni, K. Lu, F. Blaabjerg, D. Xu, "A Comparative Study on Pulse Sinusoidal High Frequency Voltage Injection and INFORM Methods for PMSM Position Sensorless Control", IEEE, 2016, [10.1109/IECON.2016.7793021](https://doi.org/10.1109/IECON.2016.7793021)
- [4] K. Lee, J.-I. Ha, "Evaluation of Back-EMF Estimators for Sensorless Control of Permanent Magnet Synchronous Motors", Journal of Power Electronics, 2012, [10.6113/JPE.2012.12.4.604](https://doi.org/10.6113/JPE.2012.12.4.604)
- [5] Z. Qiao, T. Shi, Y. Wang, Y. Yan, C. Xia, "New Sliding-Mode Observer for Position Sensorless Control of Permanent-Magnet Synchronous Motor", IEEE Transactions on Industrial Electronics, 2013, [10.1109/TIE.2012.2206359](https://doi.org/10.1109/TIE.2012.2206359)
- [6] D. S. Putra, S.-C. Chen, H.-H. Kong, F. Cheng, "Design and Implementation of Machine-Learning Observer for Sensorless PMSM Drive Control", MDPI Applied Sciences, 2022, [10.3390/app12062963](https://doi.org/10.3390/app12062963)
- [7] Q. Gao, C. Zong, "Rotor position estimation method of PMSM based on recurrent neural network", International Conference on Power and Energy System Engineering, 2021, [10.1016/j.egyr.2021.11.091](https://doi.org/10.1016/j.egyr.2021.11.091)
- [8] M.-S. Wang, T.-M. Tsai, "Sliding Mode and Neural Network Control of Sensorless PMSM Controlled System for Power Consumption and Performance Improvement", MDPI Energies, 2017, [10.3390/en10111780](https://doi.org/10.3390/en10111780)
- [9] Y. Huang, Y. Xie, W. Han, "A Dynamic Self-Adjusting System for Permanent Magnet Synchronous Motors Using an Improved Super-Twisting Sliding Mode Observer", MDPI Sensors, 2025, [10.3390/s25123623](https://doi.org/10.3390/s25123623)
- [10] G. Wang, H. Zhang, "A Second-order Sliding Mode Observer Optimized by Neural Network for Speed and Position Estimation of PMSMs", Springer, Journal of Electrical Engineering & Technology, 2022, [10.1007/s42835-021-00892-5](https://doi.org/10.1007/s42835-021-00892-5)
- [11] Q. An, Q. An, X. Liu, J. Zhang, K. Bi, "Improved Sliding Mode Observer for Position Sensorless Control of Permanent Magnet Synchronous Motor", IEEE Transportation Electrification Conference and Expo, 2018, [10.1109/TEEC-AP.2018.8433265](https://doi.org/10.1109/TEEC-AP.2018.8433265)
- [12] Y. Liu, J. Fang, K. Tan, B. Huang, W. He, "Sliding Mode Observer with Adaptive Parameter Estimation for Sensorless Control of IPMSM", MDPI Energies, 2020, [10.3390/en13225991](https://doi.org/10.3390/en13225991)
- [13] C.-J. Kwon, S.-J. Kima, W.-Y. Hanb, W.-K. Minc, "Sensorless control for permanent magnet synchronous motor using a neural network based adaptive estimator", ICMIT 2005, Control Systems and Robotics, 2005, [10.1117/12.664658](https://doi.org/10.1117/12.664658)
- [14] A. Zribi, M. Chtourou, M. Djemel, "A New PID Neural Network Controller Design for Nonlinear Processes", arXiv, Systems and Control, 2015, [10.1142/S0218126618500652](https://doi.org/10.1142/S0218126618500652)

Towards Semiautomatic Pulmonary Artery-Vein Separation in CT Images

Gábor Révy, Dániel Hadházi, Gábor Hullám

Department of Artificial Intelligence and Systems Engineering

Budapest University of Technology and Economics

Budapest, Hungary

Email: {revy, hadhazi, gabor.hullam}@mit.bme.hu

Abstract—Accurate pulmonary vessel segmentation, and the separation into arterial and venous networks, guides physicians during surgical planning, particularly in thoracic procedures. Despite significant advancements, the automation of this task is still a work in progress. Many studies show promising results, but they often rely on high-quality imaging or complex computational resources, which limits their applicability in clinical practice.

This vision paper outlines a research direction towards semi-automatic pulmonary artery-vein separation in CT images. Our aim is to explore solutions that remain effective under challenging imaging conditions, while maintaining low computational requirements. These challenges include low-quality CT scans with large slice thickness, without contrast agent or imprecise timing of contrast injection, and the absence of filtering algorithms during reconstruction resulting in streaking artifacts. This work discusses key design principles, assumptions, and methodological ideas that could enable robust and practical pulmonary vessel separation.

A central aspect of our research is ensuring that physicians can easily trust and adopt the tool in clinical settings. A system that is both effective and intuitive is critical for widespread usage. Physicians will only use such a program if they trust its reliability, understand how it works, and find it simple to interact with. As such, special emphasis is placed on achieving a semiautomatic workflow with minimal user intervention—requiring only a few clicks for correction. The method will be designed to run on standard computing configurations, with an open-source implementation to support the work of practicing physicians.

Index Terms—semiautomatic, medical image processing, machine learning-based methods, classical image processing algorithms, hybrid approach, pulmonary vessel segmentation, artery, vein

I. INTRODUCTION

In recent decades, surgical planning systems, alongside CADe (computer-aided detection) and CADx (computer-aided diagnosis) systems, have been key areas of research. These systems frequently rely on the integration of the latter two to automate as many steps as possible. By accurately identifying arterial and venous pulmonary vessels, physicians can better assess the vascular structure of the lungs, plan surgeries more effectively, and minimize risks. The resulting surgical models, guide interventions, and therefore can improve outcomes.

Pulmonary vessel segmentation in CT imaging involves classifying each voxel to determine whether it belongs to the vascular network within the lungs. Separation means the classification of these vessels into arterial (A) and venous

(V) classes, assigning each segmented vessel to the correct network. However, both tasks present several challenges.

A. Challenges

One of the primary obstacles is the absence of contrast agents in many clinical CT scans, which makes it difficult to clearly delineate the vessel boundaries, especially near the hilar region where vessels enter the lungs. Even when contrast agents are used, improper or lack of filtering during reconstruction can lead to streaking artifacts, which degrade the quality of the scans and introduce “artificial boundaries” between vessels, further complicating segmentation.

Additionally, a key characteristic of a separation algorithm is the range of slice thickness within which the algorithm is able to function reliably. Many approaches rely on the principle of “traceability” [1], where vessel segments across slices are connected in 3D, ensuring that segments of the same vessel are in the same connected component, while those of different vessels remain separate. However, CT scans with large slice spacing (e.g., greater than ~ 2 mm) often fail to meet this requirement. In these cases, when two vessels are in close proximity along the superior-inferior axis, the segmentation algorithm might erroneously connect them into a single 3D component. As a result, algorithms [2] relying on “traceability” may mislabel vessels, causing leakage of one vessel type (A/V) into the other during the separation process. For example, voxel-based marching algorithms, which propagate labels from seed points, might inadvertently cross from one vessel to another, leading to incorrect labeling without incorporating additional contextual information.

II. RELATED WORK

Pulmonary A/V segmentation and separation is a long-standing issue, and we ourselves have already attempted various approaches to address it. These approaches mainly included classical image processing algorithms. The most promising one was a fuzzy distance transform-based [3] algorithm [4]. We proposed [5] some improvements to that algorithm including a correction method to make it operate in a more semiautomatic manner (enabling greater user interaction) and easier to correct its results. The weakness of this approach is that it is still prone to streaking artifacts and the partial volume effect.

Building upon the results of our previous work, it became clear that machine learning techniques are essential for a more robust and accurate operation. While classical image processing algorithms offered some success, they were still limited by inherent issues. These limitations underscored the need for machine learning-based methods to address the complexities of pulmonary vessel segmentation.

Advancements in machine learning, particularly convolutional neural networks (CNNs), have demonstrated significant potential in overcoming these challenges. By leveraging large datasets, CNNs can effectively capture complex patterns in imaging data, offering superior accuracy and robustness compared to traditional methods. Our ongoing experiments indicated that machine learning models are less prone to streaking artifacts, which could significantly improve segmentation precision.

We would like to highlight a research by Chu *et al.* [6] which appears promising from the perspective of our research. They developed a framework called HiPaS for pulmonary artery-vein segmentation. It achieved 91.8% dice score and 98.0% sensitivity for artery-vein segmentation on non-contrast CT. HiPaS utilizes an explicit spatial normalization module, and an iterative, multi-resolution segmentation scheme that uses lower-level vessel segmentations as priors for higher-order A/V labeling. The spatial normalization is performed via an Inter-and-Intra Slice Super-Resolution (I2SR) module, which resamples CT scans into a normalized size with $1mm$ slice thickness. The HiPaS framework also introduces a cascading multistage segmentation strategy called the Saliency-Transmission Segmentation (STS) module inspired by the hierarchical nature of manual vessel annotation. Specifically, the vascular tree is decomposed into multiple levels corresponding to different vessel branch orders, and each level is segmented in a sequential manner. At each stage, the predicted probability map of lower-order vessels is propagated as prior information and combined with the original CT volume to guide the segmentation of higher-order vessels. This progressive operation enables the modeling of structural dependencies across different vessel scales. A dataset comprising 1073 CT volumes was generated for their study, including 315 CTPA and 758 non-contrast CT scans. A subset [7] of this dataset was made publicly available consisting of 250 CT scans.

The dataset was analyzed thoroughly and some experimental trainings were also performed using this dataset. An example CT slice is shown in Figure 2. By first examining the dataset, one can conclude that there is a clear oversegmentation of the vessels. This appears slightly more pronounced where vessels of different type visually overlap due to the reconstruction. Here, the oversegmentation perpendicular to the vessel is even greater at the point of overlap.

A great advantage of the dataset is that overlapping labeling was allowed for the medical annotators. This means that in case of visually overlapping vessels both labels can be present. This allows for training a system that is able to handle such cases. The downside also stems from this: there are many

errors in manual labeling (typically a few per scan), and the labeling along the course of the vessel typically gets misaligned at the overlapping part. Figure 1 shows an example for this: the overlapping part is shown This is hard to detect algorithmically.

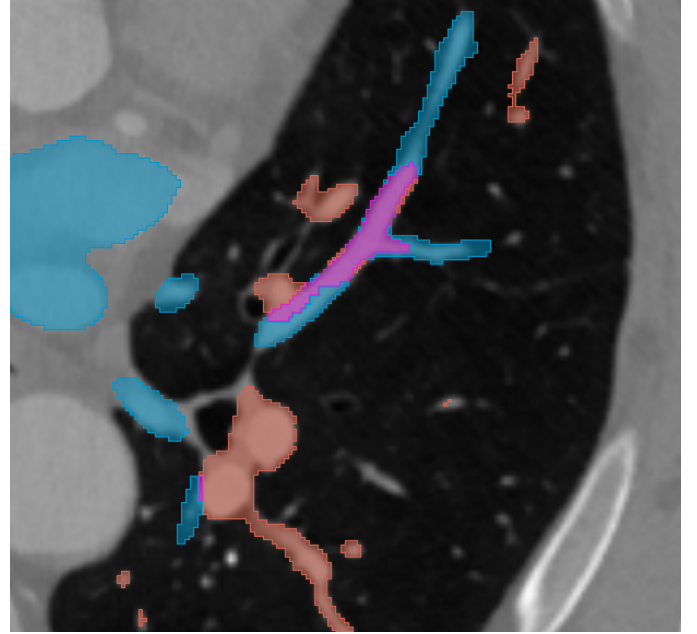


Fig. 1. Error in the manual labeling in the HiPaS dataset. A long section of a vessel is marked both as artery and vein (marked with purple).

III. SET RESEARCH DIRECTION

Our aim is to integrate machine learning-based methods with semiautomatic tools, so that a balance between automation and user intervention can be maintained, allowing for real-time corrections. This hybrid approach not only enhances the accuracy of our previous results but also provides a more user-friendly experience. Preliminary results show promising results, suggesting that introducing machine learning could be a game-changer, facilitating more reliable yet still modifiable pulmonary A/V segmentation and separation.

While machine learning-based approaches have shown promising results in the realm of pulmonary A/V segmentation, we argue that a fully machine learning-driven solution may not always be the most optimal for every aspect of the whole process. Therefore, we propose a hybrid approach where the image processing components are enhanced using machine learning, but the logical framework for vessel segmentation and separation remains algorithmic in nature. The core idea behind the hybrid approach is to combine the strengths of both worlds: leveraging machine learning techniques for complex image feature extraction, such as detecting vessels and segmenting the regions of interest (RoI), while using traditional algorithmic methods for logic processing, and connecting local results (results in a RoI). This enables us to maintain greater control over specific segmentation steps, particularly when more robust, deterministic logic is required

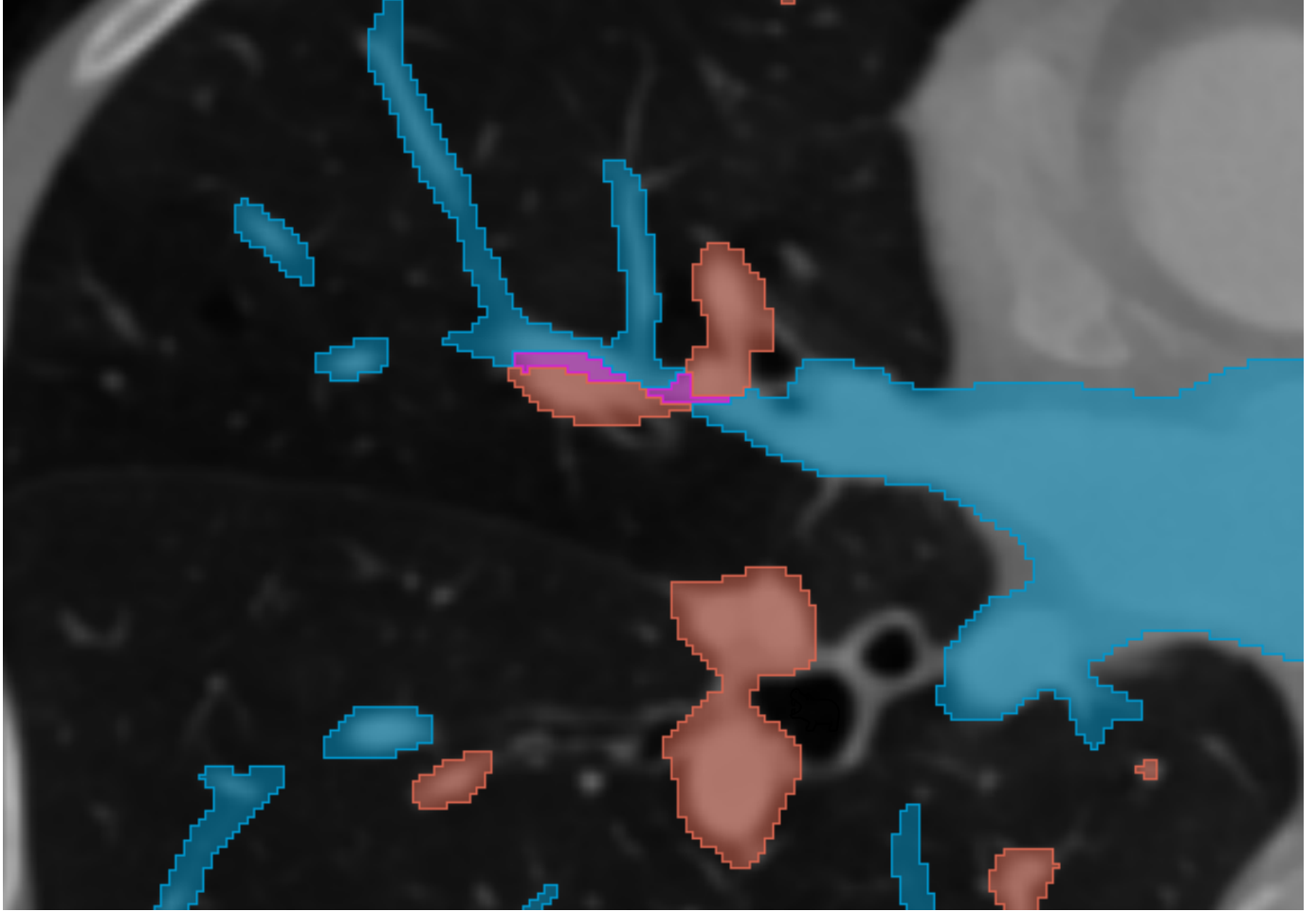


Fig. 2. A labeled CT slice from the HiPaS dataset. Red and blue denote artery and vein, respectively.

to handle edge cases or complex anatomical structures. Furthermore, the algorithmic nature ensures that the segmentation process remains somewhat interpretable.

The idea arises from the difficulty in establishing a constraint for a neural network, whereby a vessel is associated with a single label, which remains fixed throughout its course. This could be handled by a separate algorithm.

Convolutional neural networks are good at pattern matching. The A/V vessel structures near the hilum have a very characteristic shape. A 3D convolutional neural network is expected to be able to label them properly, since the A/V separation is local. This means that it can be performed by simply looking at the structure and no tracing back to the heart is required thus enforcing the continuity constraint is not expected to be necessary. Our preliminary experiments mainly focused on the automatic A/V labeling of the vessels near the hilum. Initial results show that the network is able to separate the vessels quite robustly. Figure 3 shows an example prediction for a thoracic CT scan. Here, two separate 3D convolutional neural networks were trained for the hilar regions of the two lungs. The two regions are easily distinguishable in the images.

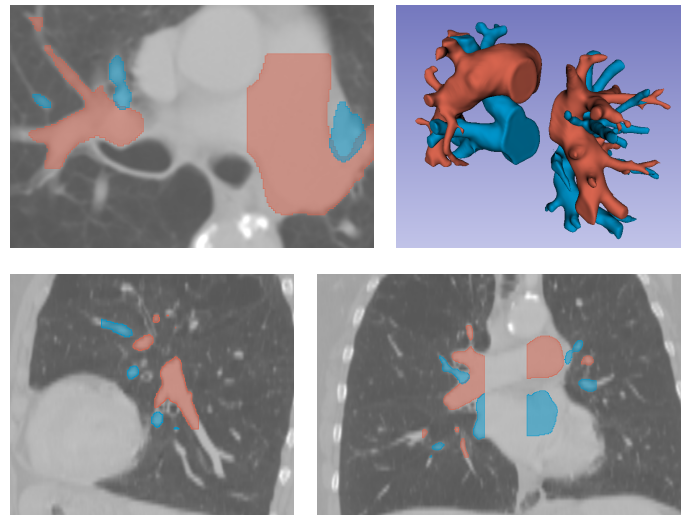


Fig. 3. Initial results the vessel segmentation near the hilum. The views are presented in the following order: axial, 3D, sagittal, coronal.

The next step would be the propagation of these labels deeper inside the lungs. This is inspired by the actual practice of manual segmentation: during preoperative design clinicians label the vessels by tracing them back to the hilum. One of the main concepts of our proposed algorithm is to segment the pulmonary vessels and build a graph from that segmentation. Based on the visual information and possibly incorporating a classifier neural network, the labeling can be propagated along the edges of the graph. The challenge (and basically the whole task) is to differentiate between the branching of one vessel, and the convergence and subsequent divergence of different, closely running vessels which appear the same in a graph created from the binary segmentation of the vascular network. There are several possibilities for incorporating machine learning at this step. The edge labeling can be approached both as a classification or a segmentation task. A CNN could extract features which are then assigned to edges. Then, a spin glass classifier [8] could be utilized to classify the edges. Approaching it as a segmentation task, another idea is to create a CNN-based region growing algorithm that only grows the segmented region along the same vessel.

The hybrid approach has the potential to provide a more flexible and robust solution, where the strengths of machine learning are utilized without fully relying on it for every step. Furthermore, it allows us to keep the process transparent and adaptable, ensuring that clinicians or researchers can adjust the results when necessary, while still benefiting from the automation and efficiency that machine learning provides.

IV. EVALUATION PLAN

To validate our approach, we will conduct a comprehensive evaluation using a combination of metrics and clinical evaluation.

- 1) We will compare the segmentation accuracy of our hybrid approach against existing methods using commonly used metrics that include Dice similarity coefficient (DSC), sensitivity, and precision.
- 2) We also plan to create a graph-based metric that considers the length of the correctly labeled vessels instead of the volume of them. This would reflect the accuracy of the approach better.
- 3) During the examination of the dataset, we observed that there are numerous labeling errors, which will need to be taken into account during the evaluation.
- 4) Furthermore, we will conduct a clinical evaluation involving radiologists or clinicians to assess the usability and clinical relevance of our segmentation results. The feedback from clinicians will be used to guide further improvements and to validate the effectiveness of the hybrid approach in improving diagnostic workflows and its potential for practical use in clinical settings.

V. CURRENT STATUS AND TIMELINE

Current status We have completed the preliminary design of the hybrid approach, integrating a 3D convolutional neural network for initial vessel segmentation in the hilar region.

Preliminary experiments have shown that the CNN can robustly segment the vessels near the hilum, achieving promising results in terms of accuracy and efficiency. We have also developed a basic region-growing and a graph-based method to propagate the segmented labels to deeper regions of the lungs.

By the end of March The development of a GUI for the algorithm is in progress in the form of a Slicer 3D¹ [9] plugin that ensures the system is practical for real-world clinical use. The goal is to make the interface intuitive while allowing for manual adjustments when necessary. In parallel with this, the algorithms are being optimized to ensure they run at an acceptable speed on an average PC, including memory and computational optimizations, as well as leveraging parallel processing where possible. Parallel to this, the semi-automatic correction is being improved. These improvements are iteratively adjusted based on the user interface testing.

By the end of May Once the system is functional and user-friendly, we will perform the evaluation using the beforementioned metrics to ensure the approach meets both performance and usability standards. The evaluation will be performed on a public dataset such as the AirRC dataset [10].

By the end of June In parallel, we will start preparing a journal article, which will be finalized based on the results.

REFERENCES

- [1] Y. Kitamura, Y. Li, W. Ito, and H. Ishikawa, "Data-dependent higher-order clique selection for artery-vein segmentation by energy minimization," *International Journal of Computer Vision*, vol. 117, no. 2, pp. 142–158, 2016.
- [2] E. M. Van Rikxoort and B. Van Ginneken, "Automated segmentation of pulmonary structures in thoracic computed tomography scans: a review," *Physics in Medicine & Biology*, vol. 58, no. 17, p. R187, 2013.
- [3] P. K. Saha, F. W. Wehrli, and B. R. Gomberg, "Fuzzy distance transform: theory, algorithms, and applications," *Computer Vision and Image Understanding*, vol. 86, no. 3, pp. 171–190, 2002.
- [4] P. K. Saha, Z. Gao, S. K. Alford, M. Sonka, and E. A. Hoffman, "Topomorphologic separation of fused isointensity objects via multiscale opening: Separating arteries and veins in 3-d pulmonary ct," *IEEE transactions on medical imaging*, vol. 29, no. 3, pp. 840–851, 2010.
- [5] G. Révy and A. Bodnár, "Towards pulmonary vessel separation," *31th Minisymposium of the Department of Measurement and Information Systems*, pp. 13–18, 2024.
- [6] Y. Chu, G. Luo, L. Zhou, S. Cao, G. Ma, X. Meng, J. Zhou, C. Yang, D. Xie, D. Mu *et al.*, "Deep learning-driven pulmonary artery and vein segmentation reveals demography-associated vasculature anatomical differences," *Nature Communications*, vol. 16, no. 1, p. 2262, 2025.
- [7] Y. Chu, "Hipas artery-vein segmentation on non-contrast ct," Feb. 2025. [Online]. Available: <https://doi.org/10.5281/zenodo.14879605>
- [8] J. Staal, S. N. Kalitzin, and M. A. Viergever, "A trained spin-glass model for grouping of image primitives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 7, pp. 1172–1182, 2005.
- [9] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka *et al.*, "3d slicer as an image computing platform for the quantitative imaging network," *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1323–1341, 2012.
- [10] J. Liu, Z. Zhang, B. Niu, S. Kang, J. Ren, L. Wang, and K. Xu, "a custom annotated dataset for segmentation of pulmonary veins, arteries, and airways," *Scientific Data*, vol. 12, no. 1, p. 1806, 2025.

¹<https://www.slicer.org/>

Beyond Ranging: Real-time Wireless Communication Utilising Wi-Fi Fine Timing Measurement

Ronald Rádai, Tamás Kovácsházy

Budapest University of Technology and Economics,
Department of Artificial Intelligence and Systems Engineering,
Budapest, Hungary

Email: {radair, khazy}@mit.bme.hu

Abstract—Precise clock synchronization is a fundamental requirement in real-time distributed embedded systems, yet low-cost microcontrollers typically rely on free-running crystal oscillators with non-negligible frequency drift. This paper explores a novel use of Wi-Fi Fine Timing Measurement (FTM) to estimate and compensate for the relative frequency drift between the quartz oscillators of two ESP32-S3 microcontrollers.

The proposed approach employs a minimalistic setup in which one node operates as a standard Wi-Fi FTM Responder, while the other acts as an FTM Initiator continuously issuing ranging requests. For each FTM exchange, the initiator records the timestamp quadruple (T_1 , T_2 , T_3 , T_4), which is subsequently logged via a serial interface. By post-processing these timestamps with appropriate estimation algorithms, the relative frequency offset between the two oscillators can be derived, enabling the characterization of long-term drift behavior.

Beyond offline analysis, this work outlines a forward-looking vision in which drift estimation is performed directly on the initiator node and used to actively correct oscillator frequency via a Digitally Controlled Crystal Oscillator (DCXO) integrated with a modified ESP32 platform. Such closed-loop compensation has the potential to significantly improve clock alignment and long-term synchronization accuracy without external time references. The presented concept opens new possibilities for low-cost, wireless clock synchronization in distributed microcontroller-based systems using commodity Wi-Fi hardware.

Index Terms—Wi-Fi FTM, 802.11mc, clock synchronization, oscillator drift, ESP32-S3, DCXO, embedded systems.

I. INTRODUCTION

Time alignment is a core requirement for real-time distributed embedded systems that rely on coherent sampling, sensor fusion, cooperative actuation, or deterministic communication. In many low-cost deployments, nodes maintain time using uncompensated quartz oscillators. Even when two devices are aligned at a single instant, their clocks diverge over time due to a relative frequency offset caused by manufacturing tolerance, temperature, aging, and power-supply sensitivity. Typical errors of only a few parts per million (ppm) already yield microseconds of divergence over minutes, which can be critical in real-time applications.

Wired communication solutions, such as Ethernet, support precise clock (time and frequency) synchronization support by hardware timestamping and by implementing the IEEE 1588

Precision Time Protocol [1] or some of its profiles such as IEEE 802.1AS [2]. However, widely used wireless solutions, such as Bluetooth, Bluetooth Low Energy (BLE), Wi-Fi, and Ultra-Wideband (UWB), do not directly support hardware timestamping for clock synchronization purposes, delaying the deployment of real-time distributed systems in a wireless domain for years due to low accuracy and precision [3]. On the other hand, these technologies advertise various Time of Flight (ToF) based technologies for ranging and localization, which require substantially different approach to timestamping than time synchronization, so the direct utilization of ToF timestamp is an open scientific and practical implementation issue in clock synchronization applications. This paper investigates the application of Wi-Fi FTM for clock synchronization purposes.

Wi-Fi Fine Timing Measurement (FTM) was standardized primarily for ToF-based indoor ranging in the IEEE 802.11mc working group and currently defined in the IEEE 802.11-2024 [4] standard. However, the builtin mechanism also exposes precise two-way hardware timestamps. This suggests a broader use: treating FTM as a *wireless timing sensor* that continuously observes the relationship between two clocks. Unfortunately, not only these timestamps cannot be used directly for clock synchronization, but practical experience shows that Wi-Fi FTM implementations lack consistency, transparency, so their detailed measurement-based scientific evaluation for clock synchronization purposes is required. A similar approach is investigated by [5] in the ns-3 network simulator without practical experimentation with implementations. In this vision paper, we describe a working measurement setup on the ESP32-S3 platform (Wi-Fi enabled microcontrollers supporting FTM) and show how the observed timestamp discontinuities (most likely due to hardware and software implementation issues) require robust processing. We then outline how drift estimation could be moved on-device and used to discipline an oscillator via a digitally controlled crystal oscillator (DCXO).

A. Contributions

This paper contributes the following new scientific results:

- A minimal two-node ESP32-S3 measurement setup that continuously records FTM timestamps (T_1 , T_2 , T_3 , T_4).

- An overflow-/reset-tolerant correction method that enables long-horizon analysis despite free-running timer discontinuities.
- A forward-looking architecture for online drift estimation and DCXO-based closed-loop correction, enabling improved long-term synchronization without external time references.

II. BACKGROUND: FTM AS A CLOCK OBSERVATION

Wi-Fi Fine Timing Measurement (FTM) is a wireless ranging technique standardized by the IEEE (introduced in IEEE 802.11mc and incorporated into later IEEE 802.11 revisions) that enables devices to estimate distance over standard Wi-Fi links, often reaching near sub-meter accuracy under favorable radio conditions. The method relies on an exchange of precisely timestamped frames between two devices — typically an initiating station (the *Initiator*) and an FTM-capable access point or peer (the *Responder*)—from which the signal round-trip time (RTT) is computed. Since radio waves propagate at approximately the speed of light, an RTT measurement with nanosecond-level timestamp resolution can be converted into a propagation-delay estimate and, consequently, a distance estimate (often exposed as “Wi-Fi RTT” on some platforms).

In practice, accurate FTM ranging requires tight timing control and calibration. Therefore, both endpoints must support the IEEE 802.11 FTM procedure in hardware/firmware and make the relevant timestamps available through the driver and operating-system APIs. Performance also depends on environmental and implementation factors such as multipath, line-of-sight availability, channel bandwidth, interference, and device chipset behavior; as a result, real-world accuracy can vary from sub-meter to several meters [6] [7].

A single FTM exchange yields a quadruple of timestamps:

- T_1 : Time of departure of FTM frame from FTM Responder,
- T_2 : Time of arrival of FTM frame at FTM Initiator,
- T_3 : Time of departure of ACK from FTM Initiator,
- T_4 : Time of arrival of ACK at FTM Responder.

From these, a commonly used RTT estimator is:

$$\text{RTT} = (T_4 - T_1) - (T_3 - T_2). \quad (1)$$

Whereas ranging uses RTT to infer the propagation delay, our focus is on how repeated exchanges can be interpreted as observations of the *time-base relationship* between the two devices. Over sufficiently long intervals, the slope of the inferred clock-offset trajectory corresponds to the relative frequency offset (clock drift) between the Initiator and the Responder. However, the resolution, the accuracy and precision of the timestamps, and the details of access depend on implementation. The introduction of Wi-Fi IEEE 802.11az: Enhancements for Positioning (often referred to as Next Generation Positioning, NGP), which is now incorporated also into the generic IEEE 802.11-2024 [4] standard. Initial performance evaluation of the IEEE 802.11az (NGP) to IEEE 802.11mc (FTM) can be found in [8]. Due to the unavailability of NGP compatible devices this study is restricted to IEEE 802.11mc (FTM).



Fig. 1. ESP32-S3 module used during the measurements (photograph by the authors).

III. SYSTEM OVERVIEW AND DATA COLLECTION

No hardware description of Wi-Fi functionality including FTM, such as registers and their functions, is made available to the ESP32-S3 (similar to any other known Wi-Fi hardware, unfortunately). The software layers of the FTM implementation on the ESP32-S3 platform is available as a closed-source software library (as all other Wi-Fi firmware components), only a binary library, header files (Application Programming Interface, API), and documentation describing the API are made available by the manufacturer. Therefore, we are forced to employ a black-box approach towards understanding the operation of FTM by analyzing the data made available by the API. For this, a measurement system is constructed utilizing two ESP32-S3 modules. Fig. 1 shows the ESP32-S3 module used in the experimental setup.

A. Minimalistic roles of modules

We implemented two firmware roles for the two modules:

- **Responder node:** an ESP32-S3 running a standard Wi-Fi FTM Responder. It does not execute additional application logic; it only replies to incoming FTM requests.

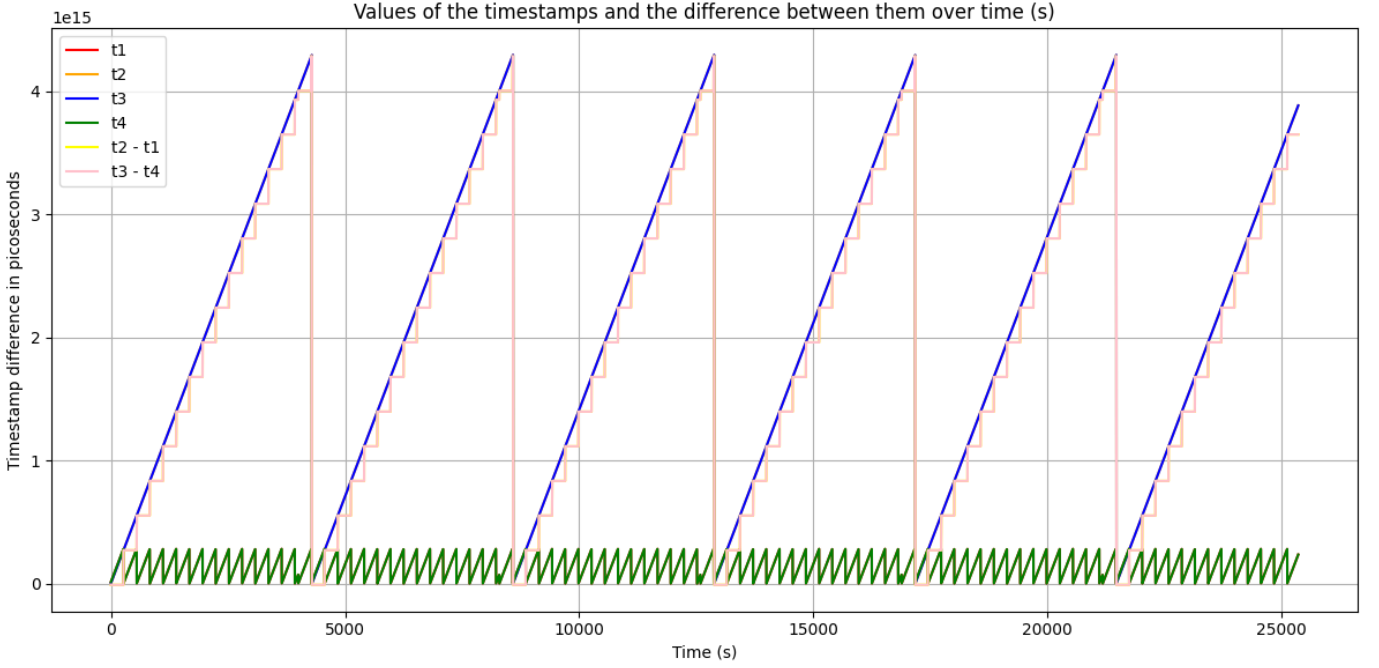


Fig. 2. Raw FTM timestamp traces and the differences between them.

- **Initiator node:** an ESP32-S3 running a Wi-Fi FTM Initiator that continuously initiates FTM sessions toward the responder. For each session it records (T_1, T_2, T_3, T_4) and logs the values via serial output for offline analysis.

This intentionally reduces confounding effects from application workload and makes timestamp behavior primarily attributable to the Wi-Fi subsystem and the underlying clocks.

B. Log format and post-processing

Each logged record contains: a tag to identify the measurement data, local time of the Initiator node with microseconds accuracy, an index of the measurement data, the RSSI (Received Signal Strength Indicator) value at the measurement and the four timestamps. Logs are parsed into a time series and processed to: (i) calculate RTT for each measurement point (ii) detect discontinuities in the data and parse it to segments, (iii) correct timestamps within segments to be comparable with each other, and (iv) estimate drift from the slope of offset observables.

IV. PRACTICAL CHALLENGE: FREE-RUNNING TIMER DISCONTINUITIES

During the experiments we found that the free-running, nominally nanosecond-resolution 64-bit timers on the two nodes do not “reset” or overflow at the same experiment time. In the recorded data, this appears as non-monotonic behavior: a timestamp occasionally decreases relative to the previous sample.

This breaks naive long-window calculations (e.g., last minus first), because a single discontinuity can dominate the computed difference. Therefore, drift estimation must be robust

to such events: either by explicitly segmenting the data or by applying a correction that neutralizes discontinuities before differencing.

Fig. 2 shows an example, where raw timestamps exhibit periodic drops.

V. OVERFLOW-TOLERANT CORRECTION AND DRIFT EXTRACTION

A. Correction values (two clock domains)

Since T_1 and T_4 share the responder clock and T_2 and T_3 share the initiator clock, we maintain two iterative correction values:

- c_i : correction baseline for initiator timestamps,
- c_r : correction baseline for responder timestamps.

Both start at zero. If any timestamp in a domain decreases compared to the previous sample, we treat it as a discontinuity and update the corresponding baseline.

B. Algorithm

Algorithm 1 summarizes the procedure. The goal is not to reconstruct an absolute timer epoch, but to produce corrected timestamps that evolve smoothly *within segments*, enabling reliable slope estimation.

Unfortunately, with this approach, we could not compensate the timestamps because as we investigated further we found that the free-running timers were reset at an undeterministic point (or it could be a deterministic point but it is not trivial to detect what this point it is).

Instead, we used a different approach. We broke the datasets to segments with the algorithm summarized in Algorithm 2. After the separation, we could compensate the datasets from

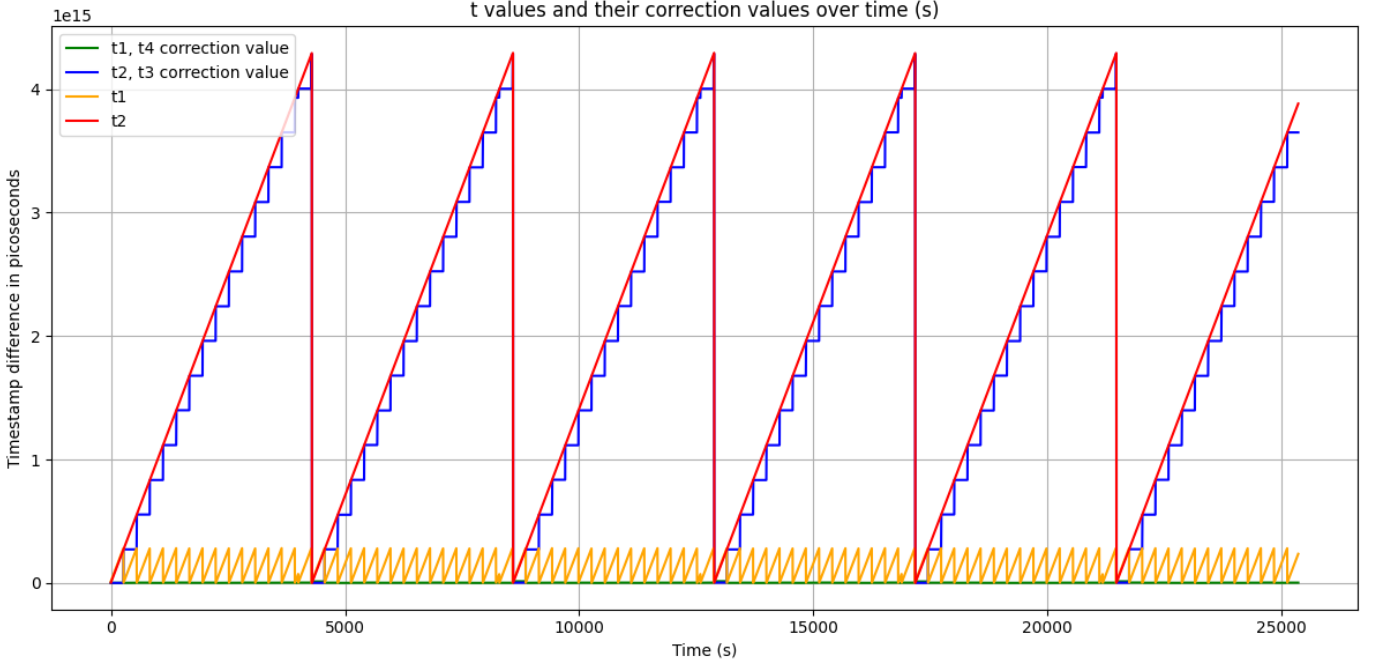


Fig. 3. Raw FTM timestamp traces and the derived correction values for discontinuity handling.

Algorithm 1 Discontinuity-tolerant correction for FTM timestamp streams

```

1: Input: sequences  $T_1, T_2, T_3, T_4$  for samples  $k = 0 \dots N-1$ 
2:  $c_i \leftarrow 0; c_r \leftarrow 0$ 
3: for  $k \leftarrow 1$  to  $N-1$  do
4:   if  $T_1(k) < T_1(k-1)$  or  $T_4(k) < T_4(k-1)$  then
5:      $c_r \leftarrow T_1(k-1)$   $\triangleright$  responder discontinuity
6:   end if
7:   if  $T_2(k) < T_2(k-1)$  or  $T_3(k) < T_3(k-1)$  then
8:      $c_i \leftarrow T_2(k-1)$   $\triangleright$  initiator discontinuity
9:   end if
10:   $\tilde{T}_1(k) \leftarrow T_1(k) - c_r; \tilde{T}_4(k) \leftarrow T_4(k) - c_r$ 
11:   $\tilde{T}_2(k) \leftarrow T_2(k) - c_i; \tilde{T}_3(k) \leftarrow T_3(k) - c_i$ 
12: end for
13: Output: corrected sequences  $\tilde{T}_1, \tilde{T}_2, \tilde{T}_3, \tilde{T}_4$ 

```

the *cut_points* we found. Each measured data should be compensated with the previous *cut_points* timestamp, because the timestamps are only consistent in these smaller intervals.

C. Offset observables

Using corrected timestamps, we compute approximate one-way RTT as:

$$\widehat{\text{owRTT}} = \frac{T_4(k) - T_1(k) - (T_3(k) - T_2(k))}{2} \quad (2)$$

under a symmetry assumption that is common in two-way timing.

Algorithm 2 Discontinuity-tolerant segmentation of FTM timestamp streams

```

1: Input: sequences  $T_1, T_2, T_3, T_4$  for samples  $k = 0 \dots N-1$ 
2:  $\text{cut\_points}[0..N]; cp_i \leftarrow 0$ 
3: for  $k \leftarrow 1$  to  $N-1$  do
4:   if  $T_2(k) < T_2(k-1)$  or  $T_4(k) < T_4(k-1)$  then  $\triangleright$ 
     Any hardware timer reset
5:      $\text{cut\_points}(cp_i) \leftarrow k$ 
6:      $cp_i \leftarrow cp_i + 1$ 
7:   end if
8: end for
9: Output: cut_point array which contains the points where
     any of the timers reseted and  $cp_i$  containing the number
     of the cut_points found

```

We then define two complementary offset observables:

$$d_{21}(k) = (\tilde{T}_2(k) - \tilde{T}_1(k)) - \widehat{\text{owRTT}}(k), \quad (3)$$

$$d_{43}(k) = (\tilde{T}_4(k) - \tilde{T}_3(k)) - \widehat{\text{owRTT}}(k). \quad (4)$$

Within each segment, d_{21} and d_{43} evolve approximately linearly with time; their slope corresponds to a relative frequency mismatch between the two oscillators. An example of visualizing d_{21} and d_{43} can be seen on Fig. 4 (where d_{43} is drawn as d_{34} to prevent drawing multiple segments over each other).

D. Drift estimation

For each continuous segment s , we fit a line $d(t) \approx \alpha_s t + \beta_s$ where β_s is equal to 0 because we are always take the

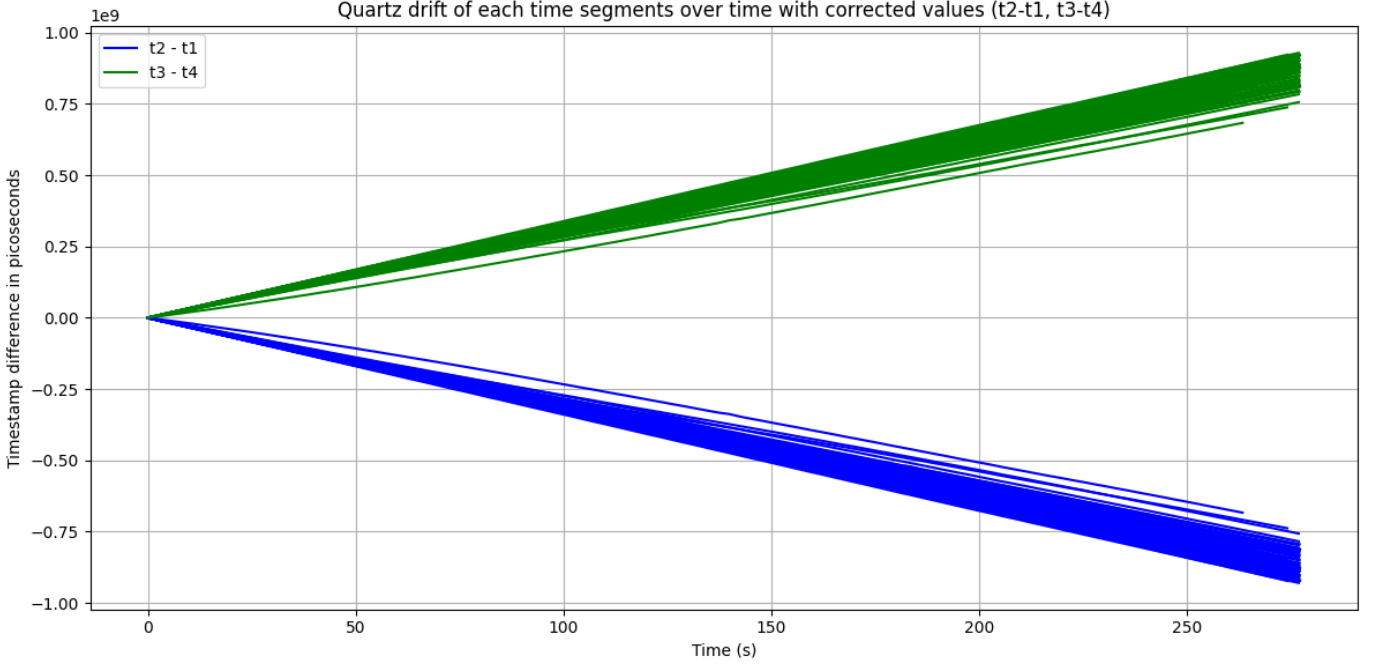


Fig. 4. Segmented drift trajectories using corrected values.

first element of the segment as starting point. The relative frequency offset is:

$$\epsilon_{\text{ppm}} \approx \frac{\alpha_s}{\text{segment_time} \cdot 10^6}. \quad (5)$$

A robust implementation may average slopes across segments and reject outliers caused by transient channel effects. Fig. 4 and Fig. 5 illustrates the corrected and segmented offset trajectories, enabling per-segment drift estimation.

VI. PRELIMINARY MEASUREMENT

In our measurements with two ESP32-S3 nodes, we observed a relative oscillator mismatch of approximately 3.2ppm. This magnitude lies within the typical tolerance range of low-cost quartz oscillators for wireless communication systems and is sufficient to cause significant clock divergence over time in synchronization-sensitive systems.

The key practical insight is that discontinuity handling is essential: naive “first minus last” approaches can be invalidated by timer epoch mismatches, while the proposed correction enables stable slope extraction within segments.

VII. VISION: ON-DEVICE ESTIMATION AND DCXO COMPENSATION

The long-term objective is to transition from offline analysis to real-time drift compensation. We envision an initiator node that continuously estimates drift online and drives a DCXO (integrated with, or externally connected to, a modified ESP32 platform) to correct frequency error. The DCXO is more expensive than the low-cost oscillator; however, real-time distributed systems are more cost tolerant than low cost

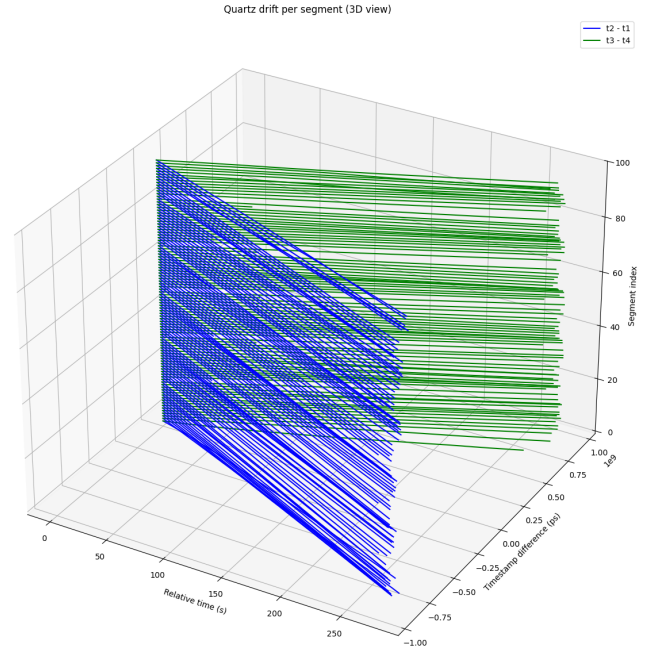


Fig. 5. Segmented drift trajectories using corrected values in 3D

wireless sensors nodes for which the ESP32-S3 module is designed for.

Such a closed-loop system could:

- improve long-term clock alignment without GNSS/PTP infrastructure,
- reduce re-synchronization traffic by disciplining fre-

- quency rather than repeatedly resetting time,
- improve the accuracy of distributed timestamping for sensing and control.

A. Open research questions

Several questions remain open:

- **Delay asymmetry and bias:** how MAC scheduling and channel conditions bias $\widehat{\text{owRTT}}$, T_1, T_2, T_3, T_4 values and the inferred slopes.
- **Estimator design:** which online methods (e.g., recursive least squares, Kalman filtering, robust regression) best balance compute cost and accuracy.
- **Control stability:** required DCXO resolution/range and loop bandwidth for stable disciplining under noisy measurements and disturbances such as temperature changes.
- **Scaling beyond two nodes:** contention effects when multiple initiators share a responder or when using mesh-like topologies.

VIII. CONCLUSIONS

Wi-Fi FTM is typically viewed as a ranging mechanism. This paper proposes a broader interpretation: FTM can act as a low-cost, standardized wireless instrument for observing and estimating relative quartz oscillator drift between microcontroller nodes. Using a minimal two-ESP32-S3 setup, we collected T_1 – T_4 timestamps and identified a key obstacle—timer discontinuities that occur at different times across devices. We presented a correction strategy that enables segment-wise drift estimation and reported preliminary measurements indicating approximately 3.2 ppm frequency mismatch. Finally,

we outlined a vision for on-device drift estimation and DCXO-based closed-loop frequency correction to improve long-term synchronization without external references.

REFERENCES

- [1] IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems, in IEEE Std 1588-2019 (Revision of IEEE Std 1588-2008), vol., no., pp.1-499, 16 June 2020, doi: 10.1109/IEEEESTD.2020.9120376.
- [2] IEEE Standard for Local and Metropolitan Area Networks—Timing and Synchronization for Time-Sensitive Applications, in IEEE Std 802.1AS-2020 (Revision of IEEE Std 802.1AS-2011), vol., no., pp.1-421, 19 June 2020, doi: 10.1109/IEEEESTD.2020.9121845.
- [3] M. Schappacher, M. Gharsalli, A. Sikora, K. Naik, D. Welte and S. Hadian “Practical IEEE 802.1AS-Based TSN Time Synchronization over Wi-Fi for Bare-Metal Devices”, 2024 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Guwahati, India, 2024, pp. 1-6, doi: 10.1109/ANTS63515.2024.10898348.
- [4] IEEE Standard for Information Technology—Telecommunications and Information Exchange between Systems Local and Metropolitan Area Networks—Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, in IEEE Std 802.11-2024 (Revision of IEEE Std 802.11-2020), vol., no., pp.1-5956, 28 April 2025, doi: 10.1109/IEEEESTD.2025.10979691.
- [5] S. H. Mohan and R. C. Sofia, “Fine Time Measurement based Time Synchronization for Multi-AP Wireless Industrial Environments,” 2023 19th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Montreal, QC, Canada, 2023, pp. 399-404, doi: 10.1109/WiMob58348.2023.10187788.
- [6] T. Kovácsázy; M. Rusvai “Performance Evaluation of Wi-Fi FTM Indoor Positioning for Embedded Applications”, Global Power, Energy and Communication Conference (GPECOM), <https://doi.org/10.1109/GPECOM61896.2024.10582641>, 05 July 2024
- [7] A. Kulkarni A. Lim “Preliminary Study on Indoor Localization using Smartphone-Based IEEE 802.11mc”, CoNEXT '19 Companion, <https://doi.org/10.1145/3360468.3368171>, 09 Dec. 2019
- [8] G. Rajendran, K. Sharma, V. Chetlapalli, J. Parekh “Performance comparison of 802.11mc and 802.11az Wi-Fi Fine Time Measurement protocols”, arXiv:2511.17935, <https://arxiv.org/abs/2511.17935>, 22 Nov. 2025

Time Segmented Deep Forecasting of Power Grid Frequency Deviations

Casi Setianingsih, Bálint Hartmann
Budapest University of Technology and Economics
Department of Electric Power Engineering
Budapest, Hungary

Email: setianingsihcasi@edu.bme.hu, hartmann.balint@vik.bme.hu

Abstract— Frequency deviations reflect short-term power imbalance and contain information about system-wide synchronization and time-delayed coupling. We analyze about one month of synchronized transmission-grid frequency measurements from 17 monitoring points using lagged Pearson cross-correlation for lags 1–10, reported for two aggregations: an all-month baseline and a weekday–weekend split. Weekends show higher coherence than weekdays across all evaluated lags (e.g., at lag 1 the global mean correlation is 0.258 on weekends versus 0.174 on weekdays). From the lag sweep, we extract peak correlation and best lag to summarize an empirical delay structure; most pairs peak at small lags, while some corridors require larger lags (≈ 8 –10 samples) for maximum alignment. To connect these coupling patterns to forecasting design, we run a controlled univariate benchmark on a reference station (24,083,128 samples in Feb 2025, 0.1 s sampling) using LSTM, TCN, and PatchTST under the same time split, train-only preprocessing, and early stopping, with window lengths $k \in \{1, 3, 10, 20\}$. Increasing the window from $k=1$ to $k=20$ improves accuracy for all models and reduces architecture sensitivity; at $k=20$ the three models converge to nearly the same nRMSE ($\approx 11.5\%$).

Keywords—power grid frequency, lagged cross correlation, LSTM, PatchTST, Temporal Convolutional Networks.

I. INTRODUCTION

Power grid frequency deviations indicate short-term power imbalance and support operational monitoring and stability assessment [1]–[3]. Wide-area monitoring relies on time-synchronized measurements to observe system dynamics and to support disturbance-aware decisions beyond conventional supervisory signals [1]–[3].

Robust frequency analytics from wide-area measurements remain challenging. Synchronized streams can be incomplete across stations and days, and availability gaps can bias correlation analysis or reduce usable training data [4]–[6]. In addition, empirical studies report non-Gaussian frequency statistics and regime-dependent correlation structure, so modeling and evaluation should match operating regimes [7]. Finally, dynamical heterogeneity implies that station responses can differ substantially; overly uniform assumptions may hide spatio-temporal effects in measured dynamics [8].

These observations motivate correlation analysis beyond a single lag. Lagged dependencies can reflect effective delays that are consistent with finite-speed electromechanical disturbance propagation, linking measurement-based coupling to time-delayed system response [9]. Our earlier 24-hour study showed that lag design can materially affect short-horizon frequency prediction; here we extend the analysis to an approximately one-month dataset and quantify how lagged coupling and dominant delays vary under an all-month baseline and a weekday–weekend regime split [10].

To connect grid-level lag patterns to predictive modeling, we run a controlled deep-forecasting benchmark with a fixed training and evaluation protocol. We evaluate multiple architectures under the same window settings ($k=1, 3, 10, 20$) to isolate the effect of window length from model choice [10], [11]. The benchmark compares LSTM, PatchTST, and TCN using the same time-based split, preprocessing, and early stopping, and includes training-setting checks to separate the impact of larger k from longer training [10], [11]. These architectures are widely used sequence baselines, including temporal convolution models and patch-based transformers for time-series prediction [11], [12]. We also cite a ConvLSTM-based frequency forecasting study to position our baselines within the broader grid-frequency literature [13].

II. DATA DESCRIPTION

This study uses synchronized grid frequency measurements from the Hungarian transmission system operator (MAVIR) over about one month from 17 measurement points (Station A to Station Q). Availability is checked per segment because the synchronized set is not always complete across days, which can affect wide-area analytics and learning datasets if ignored [14], [15]. To process the high-density stream efficiently, XML records are converted into time-based shards and stored as compact binary arrays to reduce parsing and I/O overhead, consistent with scalable data-processing practice [16], [17]. For the univariate forecasting benchmark, we select a baseline station with zero missing rate and non-flat variability; Station E is used, yielding 24,083,128 samples over Feb 2025 with a 0.1 s sampling step.

III. METHODS

This section summarizes the forecasting formulation, the lagged-correlation component, and the model implementations used in the univariate benchmark. Three sequence models are evaluated under the same split and preprocessing pipeline: an LSTM regressor, a temporal convolutional network (TCN) [18], and PatchTST [19], all implemented in the same PyTorch framework.

A. Forecasting Setup and Window Length

Forecasting is posed as one-step-ahead regression on the frequency deviation series Δf at the baseline station (Section II). At time index i , the input is a context window containing the most recent k samples, and the target is the next sample. The parameter k is treated as an experimental factor and evaluated at $k=1, 3, 10$, and 20. In the implementation, the command-line argument ‘lags’ maps to the context window as: window=1 when lags=0, otherwise window=lags. Thus, the effective context length equals the specified lag window.

B. Lagged Correlation Features

Lagged Pearson cross-correlation is computed for each station pair for $k=1$ to $k=10$, where k is a shift of k samples.

Because the sampling interval of the synchronized stream can vary by dataset, we report lags in samples and interpret them as effective delays. For each pair, we extract (i) the peak correlation over the lag sweep and (ii) the corresponding best lag, which is standard in correlation-based delay estimation [20]. Correlation summaries are reported for two aggregations only: (1) an all-month baseline and (2) a weekday-weekend split. The all-month baseline provides a robust estimate of lag-aligned coupling that is less sensitive to short-lived fluctuations or sparse segments. The weekday-weekend split is a minimal but operationally meaningful regime contrast because demand and dispatch conditions systematically differ between working days and rest days (e.g., office and school loads), which is reflected in consumption profiles and frequency-quality statistics [21]–[23].

C. Data Split and Preprocessing

All benchmark runs use the same chronological split with two fixed cut points, producing training, validation, and test periods. The split is identical for all models and all window settings. Preprocessing is fit on the training period only. First, extreme Δf outliers are clipped using training quantiles (0.0001 and 0.9999), and the resulting thresholds are applied unchanged to validation and test. Second, the clipped series is min-max scaled using parameters computed from the training period only, then applied unchanged to validation and test. This train-only fitting rule follows standard time-series practice and reduces information leakage through preprocessing or model selection [24]. Fig. 1 summarizes the split and sample counts used in all runs.

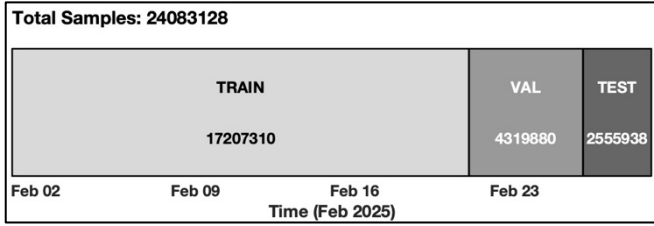


Fig. 1. Time-based split of the univariate benchmark dataset (Station E) showing total samples and the train/validation/test partitions over Feb. 2025.

D. Model Architectures

All models predict a single next step value from a fixed length context window and differ only in the sequence encoder.

a) LSTM regressor

This study uses a stacked LSTM as a recurrent baseline, followed by a linear regression head. LSTM gating supports learning long-range dependencies via an explicit memory mechanism [25]. The model takes inputs of shape $[B, k, 1]$, returns hidden outputs for all time steps, selects the final output vector, and maps it to a scalar prediction with a linear layer. Default settings are `hidden_size=64` and `num_layers=2`.

b) TCN regressor

The TCN model follows a residual temporal convolution design with dilated convolutions that expand the receptive field while remaining fully parallelizable over the time axis [18]. The implementation uses a stack of Temporal Block modules. Each block applies two 1D convolutions with ReLU and dropout, uses causal padding followed by a chomp operation to preserve causality, and adds a residual connection with an optional 1 by 1 convolution for channel matching.

Dilation is set as 2^i per block. Default settings are `channels=(32,32,32)`, `kernel size ksz=3`, and `dropout 0.1`. The final hidden state is taken at the last time index and passed through a linear head to produce the next step prediction.

c) PatchTST regressor

PatchTST is implemented as a transformer based forecaster that tokenizes the context window into patches along the time axis, reducing token length and enabling attention based learning over patch representations [19]. In the univariate setting, the input has shape $[B, 1, k]$, where B is the batch size and k is the context window length in samples. The time axis is patchified using a patch length (`patch_len`) and sliding step (`stride`) via an unfold operation, producing $n_{patches}$ patch tokens given by

$$n_{patches} = \left\lfloor \frac{k - patch_len}{stride} \right\rfloor + 1 \quad (1)$$

Each patch is linearly projected to d_{model} (token embedding dimension), a learnable positional embedding is added, and the resulting token sequence is processed by a TransformerEncoder with n_{layers} (number of encoder layers) and n_{heads} (number of attention heads). The regression head predicts the next step value from the final token representation. Default settings are `dmodel=64`, `nheads=4`, `nlayers=2`, and `dropout 0.1`.

E. Training and Evaluation Protocol

All architectures are trained with the same objective and evaluation setup. Mean squared error is minimized using Adam with a fixed learning rate $lr=1 \times 10^{-3}$ [26]. Training runs for a fixed epoch budget with early stopping monitored on validation MSE using a patience of 2, and the best validation checkpoint is used for test evaluation [27]. Performance is reported using nRMSE on inverse scaled predictions in the clipped Δf domain, using the same split and preprocessing for all architectures and all k values.

IV. RESULTS AND ANALYSIS

A. Operating Regimes: Weekday vs Weekend Analysis

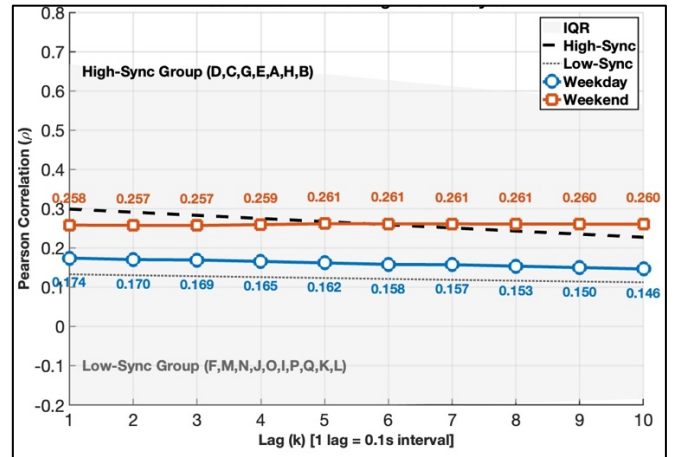


Fig. 2. Mean lagged correlation for weekday vs. weekend regimes, shown for the High-Sync and Low-Sync station groups.

Grid synchronization differs systematically between operating regimes. Fig. 2 shows higher coherence on weekends than on weekdays across all evaluated lags. At lag 1, the global mean correlation is 0.258 on weekends versus 0.174 on weekdays. A plausible explanation is that weekday

operation introduces additional variability through demand-side fluctuations and operational actions that reduce alignment across monitoring points [28], [29].

A persistent feature is the gap between the High-Sync and Low-Sync station groups. The High-Sync Group (D, C, G, E, A, H, B) shows strong mutual alignment, consistent with tight coupling through dominant transmission paths and closely aligned measurements for wide-area monitoring. The Low-Sync Group (F, M, N, J, O, I, P, Q, K, L) exhibits lower correlation to the main reference behavior, suggesting stronger influence from local dynamics or regional variations rather than a single shared trend. Importantly, this separation remains visible on both weekdays and weekends: operating regimes shift overall correlation levels, but the relative ordering of consistently high- vs low-synchronization behavior is stable. This is compatible with evidence that dynamic responses are shaped by the physical interconnection structure [30] and the wide-area monitoring context [31].

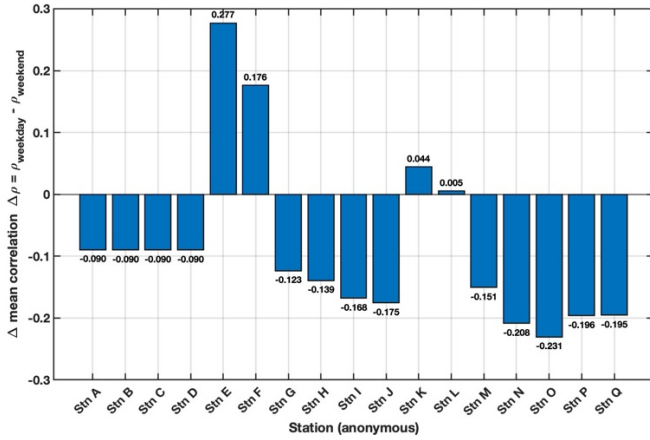


Fig. 3. Per-station weekday-weekend regime shift in mean cross-correlation at lag 1 ($\Delta\bar{\rho} = \bar{\rho}_{wd} - \bar{\rho}_{we}$); negative values indicate stronger average coupling on weekends.

Define the station-level mean coupling at lag 1 as

$$\bar{\rho}_i^{(r)}(1) = \frac{1}{N-1} \sum_{j=1, j \neq i}^N \rho_{ij}^{(r)}(1) \quad (2)$$

where i and j index stations, N is the number of stations, $\rho_{ij}^{(r)}(1)$ is the Pearson cross-correlation between stations i and j at lag $k=1$, computed separately on the weekday subset ($r = wd$) and the weekend subset ($r = we$); diagonal terms ($i = j$) are excluded. The regime shift is $\Delta\bar{\rho}_i = \bar{\rho}_i^{(wd)}(1) - \bar{\rho}_i^{(we)}(1)$, with $\Delta\bar{\rho}_i < 0$ indicating stronger weekend coupling and $\Delta\bar{\rho}_i > 0$ indicating stronger weekday coupling. This lagged cross-correlation-based alignment practice [20]. As shown in Fig. 3, most stations exhibit negative $\Delta\bar{\rho}$, confirming that the weekday-weekend coherence gap is broadly distributed across the monitoring set. The strongest weekend-favored shifts occur at Station O (-0.231) and Station N (-0.208), followed by Station P (-0.196) and Station Q (-0.195), while Station L (+0.005) and Station K (+0.044) show near-invariant behavior. In contrast, Station E (+0.277) and Station F (+0.176) exhibit higher mean coupling on weekdays, highlighting heterogeneous station-level sensitivity despite the overall increase in weekend coherence. Overall, this station-level pattern is consistent with the negative off-

diagonal global mean $\Delta\rho$ (weekday minus weekend) at lag 1 (global mean $\Delta\rho \approx -0.0849$).

B. Spatial Interaction and Delay Structure: All Month Analysis

Spatial interaction is visualized using four reference sources (Station D, Station M, Station A, and Station K) and a fixed station ordering on the horizontal axis (used only for visualization). The peak-correlation map reveals clear coupling heterogeneity. Several interactions remain strongly aligned, while Station F shows counter-swing behavior; for example, its peak correlation drops to -0.83 relative to Station M, indicating pair-specific dynamics that can oppose the dominant system trend. Marker colors encode the best lag, showing that the lag of maximum alignment varies by pair. Small-delay behavior appears for close pairs such as Station D-Station C, which peak at lag 1. Larger effective delays occur toward the south-east end (e.g., involving Station K and Station L), where peak alignment requires lags 8–10. These maps provide a month-long baseline delay structure referenced in IV-D. Fig. 4 provides the corresponding peak-correlation and best-lag map for these four references, making the coupling heterogeneity and the pair-dependent delay structure visually explicit. A fixed station ordering is used only to arrange the station-pair results in the figures for readability. This ordering is a visualization choice and does not affect the computed correlations or best-lag values.

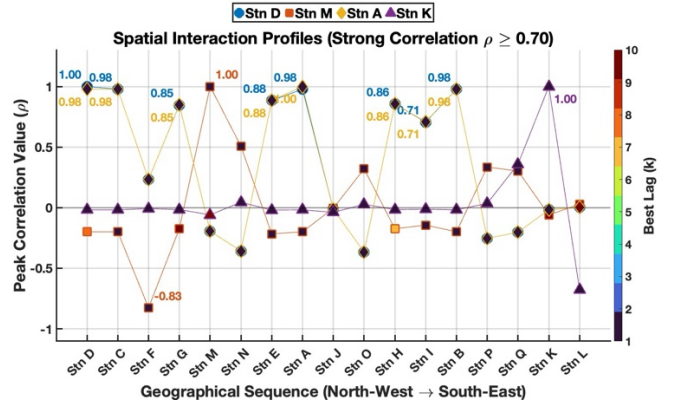


Fig. 4. All-month peak correlation and best-lag (k) profiles across the station sequence for the selected reference stations. Stations are shown in a fixed index order for visualization; the order is not used in the correlation computation.

C. Forecasting Performance Results

This section evaluates the predictive performance of LSTM, PatchTST, and TCN models using a univariate baseline from Station E.

TABLE I. NRMSE (%) AT EPOCH 100 ACROSS CONTEXT WINDOWS k AND NET CHANGE FROM $k = 1$ TO $k = 20$.

Model	$k = 1$ (%)	$k = 3$ (%)	$k = 10$ (%)	$k = 20$ (%)	Δ nRMSE ($k=20 - k=1$) (%)
LSTM	11.75	11.74	11.62	11.54	-0.21 (↓)
PatchTST	11.73	11.71	11.63	11.55	-0.18 (↓)
TCN	11.75	11.71	11.61	11.56	-0.19 (↓)

Performance is measured using nRMSE for each context window length k under the same split and preprocessing. Table I reports final (epoch 100) nRMSE values and shows a

consistent decrease in error as k increases, indicating that longer input history provides additional predictive information. The net gain from expanding the window from $k=1$ to $k=20$ is largest for LSTM and smaller for PatchTST and TCN, while results converge at $k=20$ within a narrow band (11.54–11.56%). Fig. 5 visualizes the same trend and highlights reduced architecture sensitivity once the context window is sufficiently long.

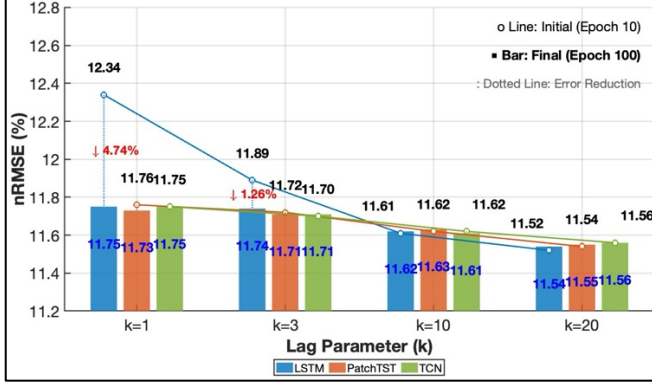


Fig. 5. nRMSE of LSTM, PatchTST, and TCN for context window lengths $k \in \{1, 3, 10, 20\}$.

1. Impact of Context Window Length:

The results show that context window length is a primary factor in improving forecasting accuracy. For all models, the nRMSE decreases as k increases from 1 to 20. For example, in the initial baseline (Epoch 10), the LSTM model starts with an nRMSE of 12.34% at $k=1$ and drops to 11.52% when the window is expanded to $k=20$. This aligns with broader time-series forecasting practice that the amount of relevant historical context available to the model can strongly influence predictive accuracy [32][33].

2. Model Comparison and Architecture Sensitivity:

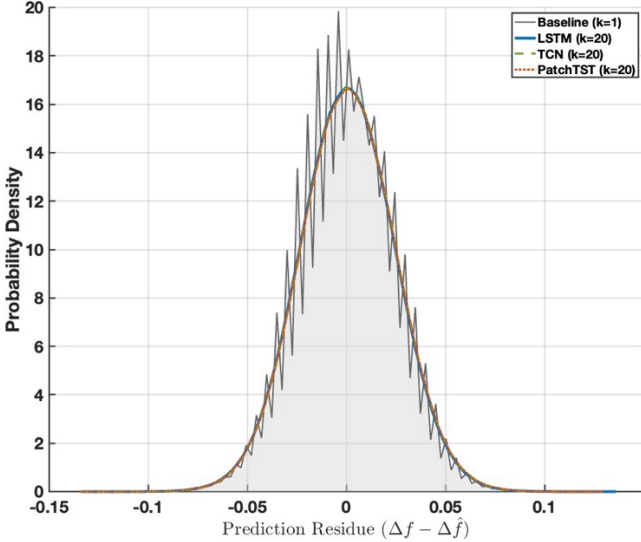


Fig. 6. Probability density functions of prediction residuals at station E comparing the $k=1$ baseline with $k=20$ optimized models.

At short windows, the convolutional and transformer-based models are more efficient than the recurrent baseline. At $k=1$, both TCN (11.75%) and PatchTST (11.76%) outperform LSTM (12.34%). As the window grows to $k=20$, all three architectures converge, with final nRMSE between 11.54% and 11.56%. This supports the view that capturing time-delayed effects depends strongly on whether the input

window spans the relevant delays; once sufficient delayed information is included, remaining differences between architectures can be small in practice [32], [34].

Residual distributions provide an additional view of the window-length effect. Fig. 6 compares prediction residuals at Station E (≈ 24 million samples, 100% availability). Moving from $k=1$ (0.1 s history) to $k=20$ (2.0 s history) narrows the residual distribution and reduces error variance, producing a distribution more concentrated around zero. This indicates higher precision when the input window contains more delay-relevant history.

At $k=20$, the residual distributions of LSTM, TCN, and PatchTST largely overlap. This indicates that, under the current one-step setup and preprocessing, the three architectures extract similar predictive information once sufficient history is provided. The residual peak near zero also suggests limited systematic bias (no consistent over- or under-prediction). Together, these observations support a practical predictability limit under this experimental setup: additional architectural changes or longer history yield diminishing changes in the residual shape at this station.

3. Gains from Extended Training:

Comparing epoch 10 to epoch 100 shows that longer training mainly benefits short windows. The largest improvement occurs for LSTM at $k=1$, where nRMSE drops by 4.74% (12.34% to 11.75%), with a smaller gain of 1.26% at $k=3$. For $k=10$ and $k=20$, additional training brings little further improvement, suggesting that performance saturates early once the window already contains sufficient history. This diminishing-return behavior is consistent with empirical observations in time-series training [33], [34].

D. Discussion: Delay Propagation and Predictability Ceiling

The key novelty is the empirical link between measurement-based lagged coupling and the context requirements of deep forecasting. Best-lag analysis shows that some station pairs require larger shifts (up to ≈ 8 –10 samples) to achieve maximum alignment. This helps explain why very short context ($k=1$) yields the highest errors: the model input cannot span delayed interactions implied by the coupling structure. When the window is expanded to $k=20$, it covers the observed lag range and the three architectures converge within about 0.3% nRMSE. This suggests that lag/window design is a physically grounded equalizer: once the input contains the relevant delay information, architecture choice becomes secondary to the information content of the time series [34].

The weekday–weekend coherence gap is consistent with weekly seasonality in electricity demand, where consumption patterns differ between workdays and weekends [35]. Combined with the best-lag maps, this supports the interpretation that window design is physically motivated: a sufficiently long input window can internalize delay effects and reduce sensitivity to architecture choice. Fig. 7 further illustrates that some residual spikes occur at the same times across LSTM, TCN, and PatchTST, indicating shared hard-to-predict events rather than model-specific failures.

Aligned prediction failures across models suggest that remaining errors are driven by components not captured by the available history (e.g., stochastic disturbances or unobserved exogenous actions), rather than by a specific algorithm. Since $k=20$ already spans the dominant delay range

observed in the coupling analysis, synchronized residual spikes likely correspond to events with weak precursors in the measured signal. Across runs, all three architectures consistently approach an nRMSE around 11.5%, which we interpret as an information-saturation point for this station under the current one-step setup.

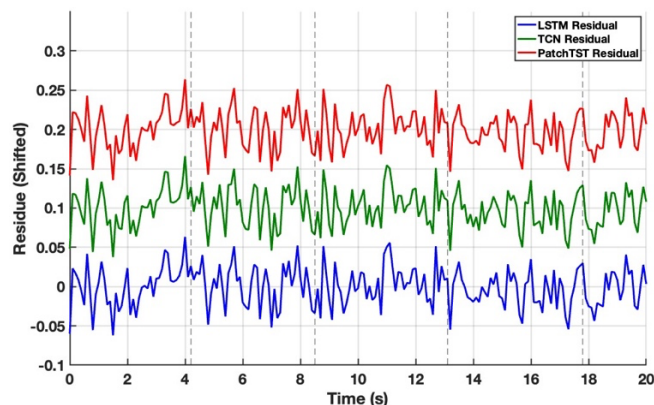


Fig. 7. Synchronized temporal residual analysis at station E comparing LSTM, TCN, and PatchTST architectures at $k=20$.

V. CONCLUSION

This study finds that frequency synchronization patterns are stable across regimes, even though average correlation is higher on weekends (0.258) than on weekdays (0.174). The station groups that remain consistently high- vs low-synchronization are largely unchanged, suggesting that relative coupling structure is shaped by the underlying interconnection. Best-lag maps reveal effective delays up to about 8–10 samples for some station pairs. In the univariate benchmark, using a forecasting window that spans these delays ($k=20$) reduces error and leads LSTM, TCN, and PatchTST to converge near 11.5% nRMSE. Overall, the results support window/lag design as a key factor for short-horizon forecasting and as a practical bridge between measurement-based coupling and model performance.

ACKNOWLEDGMENT

The authors acknowledge the financial support of the Stipendium Hungaricum Scholarship. Bálint Hartmann acknowledges the support of the Bolyai János Research Scholarship of the Hungarian Academy of Sciences (BO/131/23), and the National Research Excellence Programme of the National Research, Development and Innovation Office (ADVANCED 150405).

REFERENCES

- [1] B. Appasani and D. K. Mohanta, "A review on synchrophasor communication system: communication technologies, standards and applications," *Protection and Control of Modern Power Systems*, vol. 3, art. no. 37, 2018, doi: <https://doi.org/10.1186/s41601-018-0110-4>
- [2] B. Appasani, A. V. Jha, S. K. Mishra, and A. N. Ghazali, "Communication infrastructure for situational awareness enhancement in WAMS with optimal PMU placement," *Protection and Control of Modern Power Systems*, vol. 6, art. no. 9, 2021, doi: <https://doi.org/10.1186/s41601-021-00189-9>
- [3] D. S. Kumar, N. S. N. Manjula, and K. P. Vittal, "Micro-synchrophasor based special protection scheme for distribution system automation in a smart city," *Protection and Control of Modern Power Systems*, vol. 5, art. no. 7, 2020, doi: <https://doi.org/10.1186/s41601-020-0153-1>
- [4] A. G. Phadke and T. Bi, "Phasor measurement units, WAMS, and their applications in protection and control of power systems," *Journal of Modern Power Systems and Clean Energy*, vol. 6, pp. 619–629, 2018, doi: <https://doi.org/10.1007/s40565-018-0423-3> (Springer)
- [5] A. Sundararajan, T. Khan, A. Moghadas, and A. I. Sarwat, "Survey on synchrophasor data quality and cybersecurity challenges, and

- evaluation of their interdependencies," *Journal of Modern Power Systems and Clean Energy*, vol. 7, pp. 449–467, 2019, doi: <https://doi.org/10.1007/s40565-018-0473-6> (Springer)
- [6] C. Huang, F. F. Li, D. Zhou, J. Guo, Z. Pan, Y. Liu, and Y. Liu, "Data quality issues for synchrophasor applications Part I: a review," *Journal of Modern Power Systems and Clean Energy*, vol. 4, pp. 342–352, 2016, doi: <https://doi.org/10.1007/s40565-016-0217-4> (Springer)
- [7] X. Wen, M. Anvari, L. Rydin Gorjão, G. C. Yalcin, V. Hagenmeyer, and B. Schäfer, "Nonstandard power grid frequency statistics across continents," *Scientific Reports*, 2025, doi: <https://doi.org/10.1038/s41598-025-25334-4>
- [8] B. Hartmann *et al.*, "Dynamical heterogeneity and universality of power-grids," *Sustainable Energy, Grids and Networks*, 2024, doi: <https://doi.org/10.1016/j.segan.2024.101491>
- [9] Y. Xu, F. Wen, G. Ledwich, and Y. Xue, "Electromechanical wave in power systems: theory and applications," *Journal of Modern Power Systems and Clean Energy*, vol. 2, pp. 163–172, 2014, doi: <https://doi.org/10.1007/s40565-014-0054-2> (Springer)
- [10] C. Setianingsih and B. Hartmann, "Lag-Enhanced LSTM for Power Grid Frequency Prediction," in *Proc. Int. Youth Conf. on Energy (IYCE)*, 2025, doi: <https://doi.org/10.1109/IYCE66046.2025.11155029>
- [11] T. Limouni, R. Yaagoubi, K. Bouziane, K. Guissi, and E. H. Baali, "Accurate one step and multistep forecasting of very short-term PV power using LSTM-TCN model," *Renewable Energy*, 2023, doi: <https://doi.org/10.1016/j.renene.2023.01.118>
- [12] M. Gong, Y. Wang, J. Huang, H. Cui, S. Jing, and F. Zhang, "A novel wind power prediction model based on PatchTST and temporal convolutional network," *Environmental Progress & Sustainable Energy*, 2025, doi: <https://doi.org/10.1002/ep.14584>
- [13] A. Sathe and W. R. Yang, "Grid Frequency Forecasting in University Campuses using Convolutional LSTM," *arXiv*, 2023, doi: <https://doi.org/10.48550/arXiv.2310.16071>
- [14] C. Huang, F. F. Li, L. Zhan, Y. Xu, Q. Hu, D. Zhou, and Y. Liu, "Data quality issues for synchrophasor applications Part II: problem formulation and potential solutions," *Journal of Modern Power Systems and Clean Energy*, vol. 4, pp. 353–361, 2016, doi: <https://doi.org/10.1007/s40565-016-0213-8> (Springer)
- [15] S. Biswas, T. Yin, S. A. R. Naqvi, J. D. Follum, A. Varghese, T. Ahmad, and P. V. Etingov, "An Open Access Repository of Synchrophasor Data Quality Examples: Curation and Example Applications," *IEEE Access*, 2025, doi: <https://doi.org/10.1109/ACCESS.2025.3560602>
- [16] Y. Zhang, T. Huang, and E. F. Bompard, "Big data analytics in smart grids: a review," *Energy Informatics*, vol. 1, art. no. 8, 2018, doi: <https://doi.org/10.1186/s42162-018-0007-5>
- [17] L. Belcastro *et al.*, "Programming big data analysis: principles and solutions," *Journal of Big Data*, 2022, doi: <https://doi.org/10.1186/s40537-021-00555-2>
- [18] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv*, 2018, doi: <https://doi.org/10.48550/arXiv.1803.01271>
- [19] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," *arXiv*, 2022, doi: <https://doi.org/10.48550/arXiv.2211.14730>
- [20] M. Sun, Z. Niu, X. Zhu, and Z. Huang, "High-Precision Time Delay Estimation Algorithm Based on Generalized Quadratic Cross-Correlation," *Mathematics*, vol. 13, no. 15, art. 2397, 2025, doi: <https://doi.org/10.3390/math13152397>
- [21] Y. Mohammadi, B. Polajžer, R. C. Leborgne, and D. Khodadad, "Quantifying power system frequency quality and extracting typical patterns within short time scales below one hour," *Sustainable Energy, Grids and Networks*, vol. 38, art. 101359, 2024, doi: <https://doi.org/10.1016/j.segan.2024.101359>
- [22] F. Møller Andersen, P. A. Gunkel, C. Bjerregaard, and H. K. Jacobsen, "Changes in hourly electricity consumption profiles and price elasticities in Denmark 2019–2022," *Energy*, vol. 308, art. 132798, 2024, doi: <https://doi.org/10.1016/j.energy.2024.132798>
- [23] A. Jornaz and V. A. Samaranyake, "A Multi-Step Approach to Modeling the 24-hour Daily Profiles of Electricity Load using Daily Splines," *Energies*, vol. 12, no. 21, art. 4169, 2019, doi: <https://doi.org/10.3390/en12214169>

- [24] X. Yang *et al.*, “Research on information leakage in time series prediction based on empirical mode decomposition,” *Scientific Reports*, 2024, doi: <https://doi.org/10.1038/s41598-024-80018-9>
- [25] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: <https://doi.org/10.1162/neco.1997.9.8.1735>
- [26] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv*, 2014, doi: <https://doi.org/10.48550/arXiv.1412.6980>
- [27] L. Prechelt, “Early Stopping – But When?,” in *Neural Networks: Tricks of the Trade*, G. B. Orr and K.-R. Müller, Eds. Berlin, Heidelberg: Springer, pp. 55–69, 2002, doi: https://doi.org/10.1007/3-540-49430-8_3
- [28] F. Hao, “Impact of electric vehicle charging demand on clean energy regional power grid control,” *Energy Informatics*, vol. 8, art. no. 83, 2025, doi: <https://doi.org/10.1186/s42162-025-00538-0> (Springer)
- [29] X. Ding, J. Xu, Y. Sun, *et al.*, “A demand side controller of electrolytic aluminum industrial microgrids considering wind power fluctuations,” *Protection and Control of Modern Power Systems*, vol. 7, art. no. 49, 2022, doi: <https://doi.org/10.1186/s41601-022-00270-x> (Springer)
- [30] F. Bai, Y. Liu, Y. Liu, *et al.*, “A measurement-based approach for power system instability early warning,” *Protection and Control of Modern Power Systems*, vol. 1, art. no. 4, 2016, doi: <https://doi.org/10.1186/s41601-016-0014-0> (Springer)
- [31] M. U. Usman and M. O. Faruque, “Applications of synchrophasor technologies in power systems,” *Journal of Modern Power Systems and Clean Energy*, vol. 7, pp. 211–226, 2019, doi: <https://doi.org/10.1007/s40565-018-0455-8> (Springer)
- [32] M. Alruqimi and L. Di Persio, “Multistep Brent oil price forecasting with a multi-aspect meta-heuristic optimization and ensemble deep learning model,” *Energy Informatics*, vol. 7, art. no. 130, 2024, doi: <https://doi.org/10.1186/s42162-024-00421-4> (Springer)
- [33] H. Dhungana, “A machine learning approach for wind turbine power forecasting for maintenance planning,” *Energy Informatics*, vol. 8, art. no. 2, 2025, doi: <https://doi.org/10.1186/s42162-024-00459-4> (Springer)
- [34] J. Wibbeke, D. Alves, and S. Rohjans, “Estimating time-delayed variables using transformer-based soft sensors,” *Energy Informatics*, vol. 6, art. no. 16, 2023, doi: <https://doi.org/10.1186/s42162-023-00274-3> (Springer)
- [35] Groß, A. Lenders, F. Schwenker, and D. A. Braun, “Comparison of short-term electrical load forecasting methods for different building types,” *Energy Informatics*, vol. 4, art. no. 13, 2021, doi: <https://doi.org/10.1186/s42162-021-00172-6> (SpringerLink)

AI-Supported Performance Analysis

Balázs Ádám Toldi, András Földvári, András Pataricza and Zsigmond Papp

Budapest University of Technology and Economics

Department of Artificial Intelligence and Systems Engineering

balazs.toldi@edu.bme.hu, foldvari.andras@vik.bme.hu

pataricza.andras@vik.bme.hu, zsigmond.papp@ericsson.com

Abstract—The dependability of complex, configurable, distributed IT systems – particularly modern software-defined infrastructures – is increasingly difficult to guarantee due to the explosion of configuration variability and component interactions. Exhaustive testing is often impossible in these environments, yet critical infrastructure such as mobile networks, demands strict adherence to extra-functional requirements, such as timeliness, availability, and resilience.

To address this challenge, we propose an approach embedded in empirical model building, utilizing data from benchmark experiments and expert knowledge to characterize system behavior. This paper presents a novel framework for AI-supported performance analysis that leverages these empirical qualitative models to predict and optimize system dependability.

By employing our Mynerva stack, the proposed methodology integrates Large Language Models (LLMs) with qualitative abstraction and Explainable AI (xAI) to synthesize engineering models from limited measurement data, source code, and documentation. We propose a validation methodology within industrial 5G pilot environments and distributed ledger systems to demonstrate the framework’s potential to improve resource efficiency and service reliability.

Index Terms—Dependability Analysis, Qualitative Modeling, Large Language Models, xAI, 5G Infrastructure, Empirical Modeling

I. INTRODUCTION

Ensuring the dependability of distributed IT systems is critical, particularly in 5G telecommunications, cloud computing, edge environments, and blockchain systems [1]–[3]. As these systems grow in complexity, ensuring extra-functional properties – such as timeliness, availability, and resilience – becomes increasingly difficult [4], [5]. The primary challenge is the degree of configuration variability; in many industrial environments, the number of possible parameter combinations makes exhaustive testing and traditional verification for each individual configuration infeasible [6], [7].

Furthermore, the interference between heterogeneous implementation techniques for components often leads to unpredictable emergent behavior. While traditional performance analysis relies on extensive benchmarking and manual modeling, these methods struggle to scale alongside rapidly evolving software-defined infrastructures. There is a growing need for empirical qualitative modeling techniques that can predict and optimize system dependability by abstracting measurement data on a limited subset of candidate configurations [8], [9].

In this paper, we propose the architectural vision for a framework for AI-supported performance analysis that integrates LLMs with formal qualitative abstraction. Our ap-

proach leverages the *Mynerva stack*, a modular environment designed for LLM-assisted engineering. While this underlying infrastructure has been implemented and validated in parallel contexts, the core of our proposed work focuses on applying its synthesis capabilities to dependability analysis. By combining these capabilities with the accuracy of xAI-driven verification, our framework generates explainable engineering models from source code, documentation, and measurement data.

The key contributions of this work are as follows:

- A novel integration of qualitative modeling with xAI for the dependability analysis of highly parameterized systems.
- The introduction of the Mynerva stack, which enables local, verifiable LLM inference for model generation.
- A methodology that will iteratively refine models using counterexample learning, validated within industrial 5G pilot environments.

The remainder of this paper is organized as follows: Section III details the architectural components and the qualitative abstraction process. Section IV presents possible case studies involving industrial validation with real-world 5G data and blockchain systems, and Section V concludes the paper with a discussion on future research directions.

II. RELATED WORK

The challenge of ensuring dependability in highly configurable systems has been addressed through various approaches, ranging from combinatorial testing to model-driven engineering. This section reviews the relevant literature in three key areas: the application of LLMs in software engineering, configuration tuning and performance prediction, and the use of explainable AI for verification.

A. Configuration Tuning and Performance Prediction

Traditional approaches to configuration tuning often rely on black-box optimization techniques such as Bayesian Optimization or genetic algorithms [10]–[12]. While effective for finding optimal parameters, these methods are often computationally expensive and lack interpretability. In the context of 5G, high-fidelity digital twins have been proposed for performance prediction [13]. However, these simulation-based models inevitably diverge from the physical network over time, requiring continuous, data-driven calibration to maintain accuracy. Our proposed qualitative abstraction offers a middle ground, reducing the computational overhead while providing

interpretable, discrete state models. The goal is not to replace existing optimization techniques but to complement them by providing a more interpretable and computationally efficient framework for understanding the system’s behavior under different configurations.

B. Qualitative Modeling

A fundamental challenge in configuration testing is that while vast regions of the parameter space may exhibit similar system behavior, specific combinations can trigger distinct, emergent effects – often driven by resource saturation or race conditions. This observation justifies the treatment of configuration groups as single, typical behavioral clusters [14]–[16]

Consistent application of this principle serves two critical purposes. First, it significantly enhances the *interpretability* of intensive measurement results. By abstracting raw metrics into qualitative states, complex system behaviors become comprehensible to domain experts (e.g., network operators) who may lack specialized modeling expertise, facilitating better human validation. Second, this abstraction reduces the continuous state space into a discrete domain, thereby enabling the application of *discrete formal methods* for rigorous verification, which would otherwise be computationally intractable.

C. xAI for Verification

As AI becomes more prevalent in critical infrastructure, the need for explainability has grown. Recent works [17], [18] highlight the risks of “black-box” AI in safety-critical systems. While xAI techniques have been applied to image recognition and natural language processing [19], their application to formal dependability models remains limited [20], [21]. Our work addresses this gap by using xAI not just for explanation, but as a verification step to validate the structural integrity of the engineering models generated by the LLM.

D. LLMs for Model Synthesis

The integration of LLMs into the synthesis pipeline serves a dual purpose. First, regarding *interpretability*, the LLM bridges the gap between the qualitative abstractions and the human operator, effectively reducing the cognitive load required to understand the model’s logic [22], [23].

Second, for routine engineering tasks, the LLM acts as an abstraction layer that *encapsulates the complexity* of the underlying toolset. By providing a high-level, virtual functional interface, the LLM orchestrates the entire lifecycle – encompassing workflow execution, data management, and result interpretation – thereby masking the technical intricacies of the supporting infrastructure from the domain expert [24], [25].

III. METHODOLOGY

The core challenge of the proposed framework is to bridge the gap between static engineering artifacts and dynamic runtime behavior, thereby enabling the reliable prediction of emergent behaviors in highly configurable systems.

A. Overview

The envisioned workflow (see Figure 1) is designed to operate as a closed-loop system that iteratively refines system models based on empirical data, ensuring proper faithfulness.

The pipeline starts with four main inputs: *Source Code and Documentation* provide the structural definitions of the system; *Modeling Objectives* define the specific extra-functional requirements (e.g., latency constraints, thermal limits) acting as the target loss function; *Engineering Model* expresses the initial assumptions and design, while *Measurement Data* (such as latency logs and resource usage) offers a quantitative view of runtime behavior. These inputs are processed by the Mynerva Stack, which utilizes LLMs to synthesize an initial Qualitative Model (QM) that addresses the state-space explosion common in configurable systems via abstraction techniques. This discrete model allows for Formal Analysis and xAI Verification. The system will then generate Predictions regarding critical system states, which are validated against real-world Industrial Validation environments. Discrepancies between prediction and reality trigger a *Refinement* loop, utilizing counterexample learning to update the underlying qualitative model.

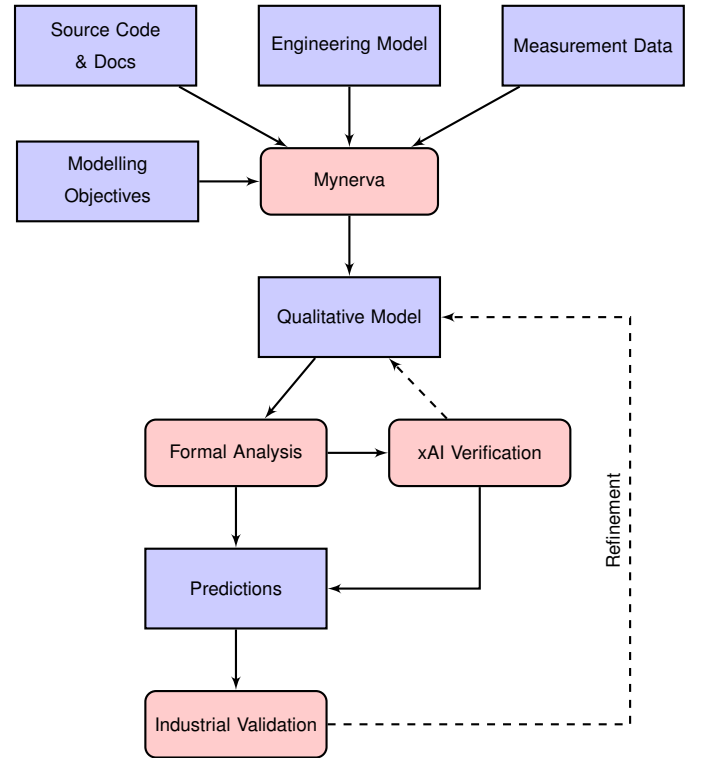


Fig. 1. Overview of the proposed workflow

B. The Mynerva Interactive Integration Fabric

The Mynerva stack serves as the backbone of our AI-supported performance analysis framework. It is designed as a modular, extensible environment for LLM-powered research workflows, consisting of three layers:

- **Inference Layer:** Handles interaction with large language models using llama.cpp for local AI model inference to conserve sensitive data. This layer abstracts the complexities of model management and hardware compatibility.
- **MCP Layer:** This layer is a collection of Model Context Protocol (MCP) servers that provide tools to the LLM for specific tasks in a standardized way. The MCP servers include Neo4j for graph database operations, Context7 for contextual documentation retrieval, and Jupyter for executing Python code.
- **Tools Layer:** This layer orchestrates the various MCP servers and provides the user with a configurable interface for adding new MCP servers. It acts as a bridge between the LLM and the underlying tools, allowing the tools to be served in a unified manner.
- **Presentation Layer:** Exposes LLM capabilities through user-friendly interfaces. OpenWebUI offers an extensible chat interface with custom filters and code execution, while n8n provides LLM-integrated process automation for structured workflows. Both clients access the unified tool and inference layers, allowing the use of both local and cloud-based models depending on task requirements.

The stack is deployed via modular Docker Compose configurations with profile-based component activation, allowing independent control of each layer based on hardware constraints and research objectives.

The Mynerva Interactive Integration Fabric is fully operational and currently supports parallel research efforts, providing the necessary local LLM inference, and workflow/agent creation capabilities.

1) Use Cases and Extensibility: While this paper focuses on dependability analysis, the Mynerva stack is designed as a general-purpose framework for LLM-assisted engineering. Its modular architecture supports a wide range of applications by swapping the underlying tool definitions and reasoning strategies.

a) Symbolic Logic Reasoning and Validation: In parallel research, the stack is utilized to enhance the reliability of critical systems through symbolic logic. This application combines the generative flexibility of LLMs with the reasoning capabilities of Answer Set Programming (ASP). Acting as a complementary layer to standard guardrails, the framework transforms natural language interactions into logical representations. By leveraging ASP’s non-monotonic reasoning, Mynerva can perform redundancy-based validation and knowledge base auditing, effectively detecting logical inconsistencies and constraint violations in both user inputs and LLM outputs.

b) Business Process Automation: Furthermore, the stack’s integration with the n8n engine allows for the automation of complex administrative workflows. In this context, the LLM acts as a reasoning agent that orchestrates multi-step business processes, providing a simple way to automate tasks such as incident management, change requests, and compliance checks, while allowing human oversight and intervention when necessary.

c) Automated data-processing: Jupyter has been the standard for interactive data analysis and visualization in the scientific community for years. It is a powerful tool for exploratory data analysis, allowing users to write and execute code in a flexible, interactive environment.

The stack’s tool layer is equipped with Jupyter and a corresponding Jupyter MCP server, which allows the LLM to execute arbitrary Python code. This enables a wide range of data processing and analysis tasks, such as log parsing, statistical analysis, and data visualization.

d) Graph-based Knowledge Representation: The Neo4j MCP server allows the LLM to interact with a graph database, enabling the construction of complex knowledge graphs that represent the relationships between user-defined entities and components. This functionality is currently handled by Lightrag [26].

C. Qualitative Abstraction and Modeling

Central to this framework is a unified analysis model that combines static source code with its corresponding runtime behavior. As depicted in the workflow, the system uses an MDSD-Based Modeling approach. In our proposed pipeline, the LLM engine will analyze the source code to extract parameter dependencies to identify how specific configuration variables causally relate to system performance.

However, exact quantitative modeling is often computationally infeasible for large-scale distributed systems. To resolve this, we apply qualitative abstraction. Continuous measurement data is mapped onto discretized states (e.g., *High*, *Medium*, *Low* CPU usage). This abstraction allows the framework to reason about the system’s “health” in symbolic terms rather than raw metrics.

D. Formal Analysis and Verification

To ensure the reliability of the AI-generated models, the framework integrates discrete formal methods with xAI. Once the qualitative model is established, formal analysis will be performed to detect potentially invalid states or unreachable transitions. A critical component of this phase is xAI Verification. Since LLMs can hallucinate relationships, the xAI module will force the model to generate explanations for its structural decisions, which are then cross-referenced against the formal constraints. This enables critical case detection, allowing the system to predict failures or violations of extra-functional properties (such as timeliness or availability) before they occur in the production environment.

Finally, the industrial validation phase compares these critical predictions against live data from the pilot environment. If a prediction fails (a false positive or negative), the system uses counterexample-guided learning. This creates a feedback mechanism that isolates the specific configuration parameters responsible for the error, and iteratively refines the qualitative model to better align with empirical reality.

IV. CASE STUDIES

To evaluate the generality and effectiveness of the proposed framework, we selected two distinct industrial-grade

distributed systems: a 5G Core Network implementation and a permissioned blockchain framework. These domains were chosen due to their high configuration complexity and the critical need for adherence to extra-functional properties such as latency, throughput, and fault tolerance.

A. 5G Core Network: Open5GS

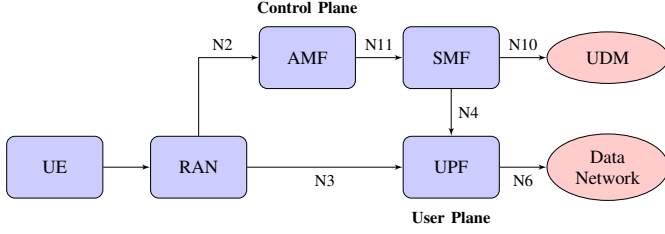


Fig. 2. Simplified 5G architecture.

The first case study focuses on *Open5GS*, an open-source implementation of the 5G Core (5GC) and Evolved Packet Core (EPC). 5G networks represent a prime candidate for empirical qualitative modeling due to the strict separation of the Control Plane (CP) and User Plane (UP), and the heavy reliance on service-based architectures. A simplified architecture of Open5GS is depicted in Figure 2, highlighting key components such as the Access and Mobility Management Function (AMF), Session Management Function (SMF), and User Plane Function (UPF), and visualizing the flow from the User Equipment (UE) through the Radio Access Network (RAN) to the data network.

1) *Experimental Setup*: The test environment consists of a containerized deployment of Open5GS, including all the key network functions such as the AMF, SMF, UPF, and others. We have configured UERANSIM to simulate UE and RAN behavior, and will create a traffic generator to produce realistic workloads.

2) *Modeling Objectives*: The primary objective is to predict the impact of control plane configurations on end-to-end latency, throughput, and session establishment success rates. The Mynerva stack ingests the C-language source code of Open5GS alongside documentation on 3GPP standards. The LLM-driven engine extracts the dependencies between configuration parameters (e.g., heartbeat timers, buffer sizes) and system states. By applying qualitative abstraction, the framework identifies critical state transitions - such as "Congestion" or "Handover Failure" - that emerge from specific parameter combinations, allowing for pre-deployment optimization without exhaustive field testing.

B. Permissioned Blockchain: Hyperledger Fabric

The second case study applies the framework to *Hyperledger Fabric*, a modular blockchain platform intended for enterprise applications. The transaction flow is illustrated in Figure 3. Unlike public blockchains, Fabric's performance is highly sensitive to deployment configurations, including batch sizes, block timeout settings, and endorsement policies.

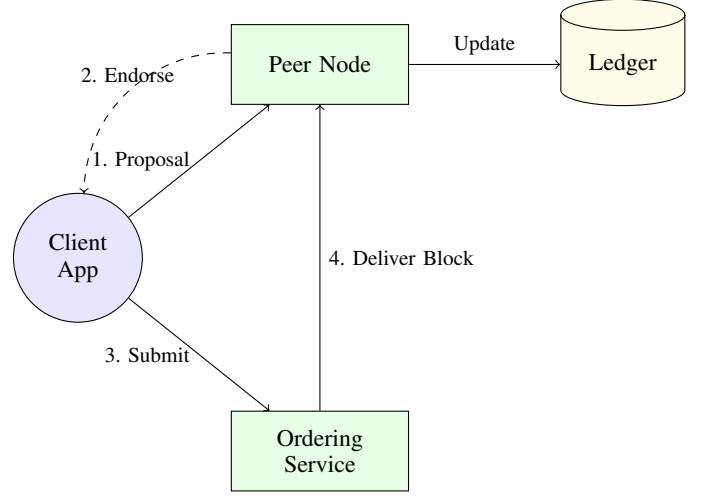


Fig. 3. Hyperledger Fabric transaction flow.

1) *Experimental Setup*: We will deploy a standard Fabric network consisting of multiple peer nodes and an ordering service based on the Raft consensus protocol. The *Hyperledger Caliper* benchmark tool will be used to inject synthetic transaction loads at varying rates to stress the system.

2) *Modeling Objectives*: The goal is to model the trade-off between *Throughput* (Transactions Per Second) and *Confirmation Latency*. The framework analyzes the system's configuration files (e.g., `configtx.yaml`) and source documentation to construct a causal model of the consensus mechanism. Using xAI verification, we will isolate the specific configuration parameters that lead to "bottleneck" states in the ordering service. This will allow system administrators to automatically tune the network for maximum throughput while maintaining bounded latency, a critical requirement for financial and supply chain use cases.

V. CONCLUSION

In this paper, we introduced the vision of a framework for the dependability analysis of complex, configurable distributed systems. By integrating Large Language Models with qualitative abstraction, we addressed the state-space explosion problem that renders traditional exhaustive configuration testing infeasible in industrial environments. The implemented Mynerva stack provides a modular environment for engineering with LLMs. We will use this toolset to generate engineering models from various sources, then utilize xAI for explainability and formal consistency. The workflow forms a closed feedback loop, where predictions are validated against real-world data, and discrepancies trigger iterative refinement of the underlying models through counterexample learning.

Our preliminary validation will be on Open5GS and Hyperledger Fabric and will demonstrate that the framework can effectively identify critical parameter dependencies and predict emergent behaviors in both telecommunications and distributed ledger contexts.

ACKNOWLEDGMENT



Project No. 2025-2.1.2-EKÖP-KDP-2025-00005 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the EKÖP_KDP-25-1-BME-27 funding scheme.

Work reported in this paper has received funding from the European Union's Horizon Europe Research and Innovation Program through the DOSS project, Grant Number 101120270.

REFERENCES

- [1] A. Avritzer, F. Di Giandomenico, A. Remke, and M. Riedl, *Assessing Dependability and Resilience in Critical Infrastructures: Challenges and Opportunities*. Berlin, Heidelberg: Springer, 2012, p. 41–63. [Online]. Available: https://doi.org/10.1007/978-3-642-29032-9_3
- [2] V. Netes, “Modern network technologies and dependability,” in *2020 International Scientific and Technical Conference Modern Computer Network Technologies (MoNeTeC)*, Oct. 2020, p. 1–10. [Online]. Available: <https://doi.org/10.1109/MoNeTeC49726.2020.9258328>
- [3] S. D. Pasham, “Fault-tolerant distributed computing for real-time applications in critical systems,” *The Computertech*, p. 1–29, Feb. 2020.
- [4] C. A. Ardagna and N. Bena, “Non-functional certification of modern distributed systems: A research manifesto,” in *2023 IEEE International Conference on Software Services Engineering (SSE)*, Jul. 2023, p. 71–79. [Online]. Available: <https://doi.org/10.1109/SSE60056.2023.00020>
- [5] T. Frank, M. Merz, K. Eckert, T. Hadlich, B. Vogel-Heuser, A. Fay, and C. Diedrich, “Dealing with non-functional requirements in distributed control systems engineering,” in *ETFA2011*, Sep. 2011, p. 1–4. [Online]. Available: <https://doi.org/10.1109/ETFA.2011.6059132>
- [6] J. Meinicke, C.-P. Wong, C. Kästner, T. Thüm, and G. Saake, “On essential configuration complexity: measuring interactions in highly-configurable systems,” in *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering*, ser. ASE '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 483–494. [Online]. Available: <https://doi.org/10.1145/2970276.2970322>
- [7] M. B. Cohen, J. Snyder, and G. Rothermel, “Testing across configurations: implications for combinatorial testing,” *SIGSOFT Softw. Eng. Notes*, vol. 31, no. 6, p. 1–9, 2006. [Online]. Available: <https://doi.org/10.1145/1218776.1218785>
- [8] A. Földvári and A. Pataricza, “Semi-automated model extraction from observations for dependability analysis,” in *2021 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, Oct. 2021, p. 1–6. [Online]. Available: <https://doi.org/10.1109/ISSREW53611.2021.00050>
- [9] —, “Model-driven, logic-supported exploratory dependability analysis,” in *2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, Jun. 2025, p. 27–34. [Online]. Available: <https://doi.org/10.1109/DSN-W65791.2025.00035>
- [10] Z. Bei, Z. Yu, H. Zhang, W. Xiong, C. Xu, L. Eeckhout, and S. Feng, “RFHOC: A random-forest approach to auto-tuning hadoop's configuration,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 5, p. 1470–1483, May 2016. [Online]. Available: <https://doi.org/10.1109/TPDS.2015.2449299>
- [11] O. Alipourfard, H. H. Liu, J. Chen, S. Venkataraman, M. Yu, and M. Zhang, “Cherrypick: adaptively unearthing the best cloud configurations for big data analytics,” in *Proceedings of the 14th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'17. USA: USENIX Association, Mar. 2017, p. 469–482.
- [12] X. Zhang, Z. Chang, Y. Li, H. Wu, J. Tan, F. Li, and B. Cui, “Facilitating database tuning with hyper-parameter optimization: a comprehensive experimental evaluation,” *Proc. VLDB Endow.*, vol. 15, no. 9, p. 1808–1821, May 2022. [Online]. Available: <https://doi.org/10.14778/3538598.3538604>
- [13] M. D. Pham, “Ai-driven digital twin for traffic prediction and calibration in 5G NR networks,” 2025. [Online]. Available: <https://aaltodoc.aalto.fi/server/api/core/bitstreams/b9de1d74-e367-4f74-a017-bb47c5af2f94/content>
- [14] L. Ferretti, G. Ansaloni, and L. Pozzi, “Cluster-based heuristic for high level synthesis design space exploration,” *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 1, p. 35–43, Jan. 2021. [Online]. Available: <https://doi.org/10.1109/TETC.2018.2794068>
- [15] E. Pruyt and T. Islam, “On generating and exploring the behavior space of complex models,” *System Dynamics Review*, vol. 31, no. 4, p. 220–249, 2015. [Online]. Available: <https://doi.org/10.1002/sdr.1544>
- [16] C. Yilmaz, M. Cohen, and A. Porter, “Covering arrays for efficient fault characterization in complex configuration spaces,” *IEEE Transactions on Software Engineering*, vol. 32, no. 1, p. 20–34, Jan. 2006. [Online]. Available: <https://doi.org/10.1109/TSE.2006.8>
- [17] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, “Interpreting black-box models: A review on explainable artificial intelligence,” *Cognitive Computation*, vol. 16, no. 1, p. 45–74, Jan. 2024. [Online]. Available: <https://doi.org/10.1007/s12559-023-10179-8>
- [18] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, “A review of trustworthy and explainable artificial intelligence (xai),” *IEEE Access*, vol. 11, p. 78994–79015, 2023. [Online]. Available: <https://doi.org/10.1109/ACCESS.2023.3294569>
- [19] A. Hanif, X. Zhang, and S. Wood, “A survey on explainable artificial intelligence techniques and challenges,” in *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*, Oct. 2021, p. 81–89. [Online]. Available: <https://doi.org/10.1109/EDOCW52865.2021.00036>
- [20] N. Khan, M. Nauman, A. S. Almadhor, N. Akhtar, A. Alghuried, and A. Alhudhaif, “Guaranteeing correctness in black-box machine learning: A fusion of explainable ai and formal methods for healthcare decision-making,” *IEEE Access*, vol. 12, p. 90299–90316, 2024. [Online]. Available: <https://doi.org/10.1109/ACCESS.2024.3420415>
- [21] J. Wiggerthale and C. Reich, “Explainable machine learning in critical decision systems: Ensuring safe application and correctness,” *AI*, vol. 5, no. 4, p. 2864–2896, Dec. 2024. [Online]. Available: <https://doi.org/10.3390/ai5040138>
- [22] C. Jeong, “Design and evaluation methods for llm-based explainable ai (xai)-based human-ai collaboration systems.”
- [23] A. Creswell, M. Shanahan, and I. Higgins, “Selection-inference: Exploiting large language models for interpretable logical reasoning,” no. arXiv:2205.09712, May 2022, arXiv:2205.09712 [cs]. [Online]. Available: <http://arxiv.org/abs/2205.09712>
- [24] E. P. Nittala, “Leveraging large language models for natural language interface in erp systems: A case study in user productivity and cognitive load,” *International Journal of Emerging Trends in Computer Science and Information Technology*, vol. 5, no. 4, p. 125–131, Dec. 2024. [Online]. Available: <https://doi.org/10.63282/3050-9246.IJETSIT-V5I4P113>
- [25] D. Luo, C. Feng, Y. Nong, and Y. Shen, “Autom3l: An automated multimodal machine learning framework with large language models,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, Oct. 2024, p. 8586–8594. [Online]. Available: <https://doi.org/10.1145/3664647.3680665>
- [26] Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang, “Lightrag: Simple and fast retrieval-augmented generation,” 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2410.05779>

Detecting Discrete Temporal States of Functional Connectivity in Electroencephalography

Mihály Vetró , Gábor Hullám 

Budapest University of Technology and Economics
Department of Artificial Intelligence and Systems Engineering
Budapest, Hungary
Email: {vetro, gabor.hullam}@mit.bme.hu

Abstract—It is commonly accepted, that the resting-state activity of the human brain - as captured by electroencephalography (EEG) - can be segmented into a sequence of just a few discrete states, called microstates, each of them being defined by a unique spatial pattern of scalp potentials. Recently, it has also come to light, that functional connectivity (FC) networks within the brain are dynamically changing, even in resting-state recordings, contradicting the assumption of temporal stationarity applied by most FC-based studies. In this paper, it is theorized, that functional connectivity networks, specifically the dynamic functional connectivity (dFC) time-series can be represented by a sequence of discrete connectivity states, at least to the same extent, as raw EEG recordings can be described using a sequence of microstates. If this assumption is proven to be true, it can enable a new class of functional brain network descriptors, that complement existing approaches, which aim to better understand and predict the development of psychiatric and neurodegenerative disorders in a clinical setting. Notably, these conditions include Alzheimer's Disease (AD) and schizophrenia, which are already associated with altered brain network activity and microstate characteristics, respectively.

Index Terms—electroencephalography, microstate analysis, dynamic functional connectivity, Alzheimer's Disease, dementia

I. INTRODUCTION

Most current research that examine resting-state functional connectivity (FC) within the human brain implicitly assume that the measured FC metrics are stationary over time, however, it was shown at least a decade ago, that this assumption is false for both haemodynamic [1] and electroencephalographic (EEG) [2] brain activity measurements. Consequently, examining the time-varying properties of functional connectivity networks, commonly referred to as dynamic functional connectivity (dFC) is relevant in almost all cases, including otherwise static investigations, that target time-averaged connectivity values, to at least complement or ensure the validity of the main results. The dFC time-series can be processed using a range of techniques, depending on the desired results and on the nature of the investigation. Most of these approaches can be sorted into one of two main categories: (1) expert systems, that utilize domain knowledge about the neurophysiological phenomena that govern functional brain networks, or empirical knowledge gained from observing the measured network dynamics, and (2) machine learning methods, which rely on little to no

domain knowledge, and are usually able to recognize more complex interactions, given sufficient data. While the second approach is promising, especially considering the continuously growing set of available neurophysiological data and the advanced capabilities of modern neural network architectures (from which graph neural networks and transformers being the most relevant), for most cases, the available domain-specific datasets are still limited and carry a large amount of variance. Mainly because of this reason, and the fact that neural networks still lack the interpretability necessary for widespread clinical adoption, our research focuses on characterizing the dFC time-series by creating well-defined features based on the inherent physiological and empirical properties observed within the data. Specifically, our method in development aims to segment the measured dFC into a series of alternating discrete states (in a similar way to how EEG data is segmented during microstate analysis [3]). The assumption is that the resulting sequence and the parameters derived from it are adequate descriptors of the temporal dynamics within the functional brain network. Such an approach could provide researchers and clinicians alike with a powerful tool to identify biomarkers for disorders, which are known to affect brain network dynamics, such as Alzheimer's Disease (AD) and schizophrenia.

A. Problem Statement

The main challenge that must be overcome in order to develop a method that can provide a reliable discrete segmentation over dFC time-series data is to define what constitutes as a connectivity state. We have devised two criteria for such states, partly based on the principles of microstate analysis:

- 1) **Convergence:** There must be a finite number of states, and the states defined for individual subjects should converge, given sufficient data.
- 2) **Discreteness:** The dynamics of the characterized time-series should only be transient near state transitions, and must be stable or quasi-stable for the majority of a state's duration.

The first criterion is necessary, because if the identified states are not consistent, i.e., cannot be matched between subjects, then that inhibits most types of group-based analysis. The second criterion determines whether the resulting segmentation can be considered discrete, or if it should be handled as a

This ongoing research is supported by the Josef Heim Scholarship of the László Schnell Foundation.

low-dimensional, but continuous representation of the original time-series, similarly to an emerging trend [4] of considering the EEG microstates as basis vectors, instead of the original discreteness assumption.

After the above mentioned two necessary criteria are evaluated, an algorithm should be devised to detect such states, and then to provide a segmentation over the dFC time-series based on this detection. Both of these tasks are highly dependent on how a state is defined, as the state definition can limit the applicable similarity measures between states and observations, as well as the mathematical space within which the operations should be conducted. Finally, descriptive parameters must be defined that can characterize the dynamic properties of the resulting segmentation.

B. Limits and Restrictions

In order to ensure that our research is feasible, and can be conducted within a reasonable time frame, we must limit the scope of the investigation, at least initially. It must be noted, that the limitations presented here could be relaxed in the future, in light of upcoming results.

The first major limit to our work, is that we **mainly focus on EEG measurements**, for two main reasons: (1) they offer superior temporal resolution compared to most imaging methods, thus enabling functional connectivity to be measured over a vast range of time scales and frequency bands, while also being capable of detecting rapid changes in connectivity values, (2) and the available resting-state data is more abundant, compared to other modalities, namely functional MRI (fMRI) and magnetoencephalography (MEG). The main drawback of EEG is its lower spatial resolution, which can be somewhat mitigated at the cost of limiting the available data, by examining combined EEG-fMRI measurements, if deemed necessary.

The second major limitation that we applied, is that we **use phase synchronization measures** to estimate functional connectivity. Overall, the exact mechanics of synchronization between functional brain areas is not yet understood, and there are multiple types of available connectivity metrics to be considered [5], namely: the correlation of haemodynamic activity (in fMRI and near-infrared spectroscopy), electrophysiological amplitude correlation, and phase synchronization metrics. The latter was chosen for its capability to measure synchronization over a wide range of frequency bands. However, it must be noted, that correlation-based metrics could highlight a different mechanism of cooperation between functional brain areas, while combined metrics, such as phase-amplitude coupling (PAC) can measure communication between different spatial scales, therefore both of these are prime candidates for future work.

The final limitation that we introduced is to measure dynamic functional connectivity (dFC) **using a sliding-window approach**. There are a few other available methods to estimate dFC, including some more abstract approaches, such as the Time-Varying Dynamic Network (TVDN) [6], and classical ones, like time-frequency based wavelet coherence. However,

the latter of these would require multiple samples from the same brain activity, to ensure stability, which would only be feasible in the presence of reproducible external stimuli, which resting-state measurements lack. Meanwhile, the abstract approaches lack clear interpretability, and the inherent simplicity of the sliding-window method.

II. RELATED WORK

In the literature examining functional networks of the human brain, the notion that the dynamic changes of these networks can have high significance besides static, time-averaged connectivity measurements, only started to gain traction around a decade ago. In this section, a brief summary is provided for the methodology and clinical applications of the dynamic functional connectivity representation in EEG, MEG and fMRI, then a few existing approaches that aim to detect discrete states within the functional brain network are introduced.

A. General Approaches for dFC Processing

A large number of studies emerged in the past years which examine the dynamics of the functional brain network directly, through a wide range of indicators, from temporal variance to metrics rooted in graph theory. Despite its lacking temporal resolution, a large number of these direct approaches were using fMRI, from measuring temporal instability of brain networks in schizophrenia [7], through default mode network (DMN) variability in patients with major depressive disorder (MDD) [8], to extracting temporal features from different spatial levels of network interactions to aid the detection of mild cognitive impairment (MCI) [9]. On the other hand, EEG-based methods seem to often utilize microstate-segmentation as a basis for dynamic functional connectivity estimation (thus replacing the straightforward sliding-window approach). These approaches range from simply aggregating connectivity maps over microstate classes to characterize different types of mental workload [10], to characterizing spatial- and temporal variability to aid the early recognition of Parkinson's Disease (PD) [11]. Besides these, sliding-window approaches are still the most common for EEG, from examining the variance of functional connectivity in patients having both PD and MCI [12], through epileptic seizure prediction using dynamic graph pattern recognition [13], to analysing dynamic functional connectivity alterations with ageing using multi-scale entropy [14].

B. Discrete State Detection in dFC

The majority of methods that aim to detect alternating discrete states within the dynamic functional connectivity time-series utilize fMRI measurements and simple clustering methods, specifically k-means clustering [15]–[17]. EEG and MEG-based methods for state detection on the other hand are less common, and are more varied. As an example, the Time-Varying Dynamic Network (TVDN) method, which identifies discrete states using partial eigendecomposition [6] was used to characterize AD [18]. Direct dFC state detection techniques, which are based on EEG measurements and sliding-window

dFC estimation are particularly rare, which is indicated by the fact, that after thorough investigation, only one study was found, which fit the criteria of using resting-state EEG, a sliding-window approach for dFC estimation, and aiming to detect discrete dFC states. This study, conducted by Chen et al. [19], focused on a microstate-like segmentation of the dFC time-series, using k-means clustering, to find patterns specific to major depressive disorder (MDD) and schizophrenia.

C. Research Gap

The detection of dFC states in resting-state EEG recordings remains a sparsely explored area within the field of dFC methodology. Furthermore, although the research conducted by Chen et al. [19] investigates the consistency of the detected dFC states, the similarity between the detected states were compared between the whole group and its own subgroups (as shown by figure 2/B of the cited study). Consequently, the task described in section I-A, with the restrictions proposed in section I-B, needs further investigation, and can be considered a research gap.

III. PROPOSED APPROACH

In order to solve the problem of alternating discrete state segmentation of the dFC time-series, we propose to build a BIDS (Brain Imaging Data Structure) pipeline, the necessary steps for which are discussed in this section. For such a solution, the dFC time-series must be estimated from the raw EEG data, then dFC states have to be defined, after which consistency and discreteness of these states must be evaluated, then finally, a set of descriptors have to be defined, which characterize the resulting discrete segmentation.

A. Computing Dynamic Connectivity

The weighted Phase-Lag Index (wPLI) will be utilized initially as a connectivity metric, which measures the consistency of instantaneous phase differences between two signals, and is weighted using the imaginary part of the Cross-Spectral Density (CSD). This latter attribute enables this method to eliminate zero-lag artifacts from volume conduction [20]. Using this metric, phase synchronization between electrodes (corresponding to different cortical areas) will be measured in five distinct frequency bands: Delta (1-4 Hz), Theta (4-8 Hz), Alpha (8-14 Hz), Beta (14-30 Hz) and Gamma (30-45 Hz). Finally, to estimate the dFC time-series, a sliding window approach will be utilized, with a shift of 40 ms, that ensures a constant 25 Hz sampling rate across all bands, and a window length defined for every band by the following function (in seconds):

$$WL(f_l, n_c) = \frac{2n_c}{f_l} \quad (1)$$

Where f_l is the lower end of the frequency band, and n_c is the number of cycles within the Morlet wavelet, used to estimate the CSD. As the number of cycles in the wavelet was chosen to be constant $n_c = 3$, the defined window length for the above-mentioned frequency bands will be 6, 1.5, 0.75, 0.43

and 0.2 seconds respectively. This approach ensures a good temporal resolution, while still having adequate reliability, by exceeding the minimal requirements set by the Shannon-Nyquist theorem, given high enough sampling rates in the raw EEG data.

B. The dFC State Definition

The dFC time-series is a sequence of connectivity graphs, over multiple frequency bands. Therefore, the most straightforward, and easy-to-interpret state definition is a set of fixed connectivity values, forming a graph structure with fixed edge weights. Because of its simplicity, most dFC studies (across different modalities) apply this state definition, then estimate a predetermined number of states as cluster centroids, using the k-means algorithm, with the dFC time points being handled as independent samples. In light of its widespread adoption, we have selected this approach as our initial method of choice, despite the multiple false assumptions that it requires, namely: the individual time points are almost certainly not independent; it is unlikely that individual states can be solely described by a single centroid; and Euclidean space is not necessarily optimal for clustering of this type.

To improve this state definition and identification, we also propose considering the individual states as probability distributions, which then can be used to calculate emission probabilities while fitting a Hidden Markov Model (HMM) to the dFC data. This approach introduces a large degree of customizability, as the emissions can have a wide range of distributions, while also dismissing the independence assumption made by k-means clustering, as the individual samples are no longer considered independent, because of the underlying Markov Chain. The main drawback of this method however, is a smaller degree of interpretability, as probability distributions are generally more complex, compared to cluster centroids.

An other possible way to define individual states, also at the cost of worse interpretability, is to consider them inherently dynamic, and use that property during state definition. An example for this would be to define the dFC states as partial differential equations (PDE), which would have similar characteristics as the dynamic states proposed in the TVDN method [6], but over the dFC time-series instead of the raw EEG data.

C. Spatially Informed Graph Distance

One major issue with using k-means clustering to identify dFC centroids is that it relies on Euclidean distance over the vectorized connectivity values, which almost completely ignores topography and spatial distribution. The former drawback can be somewhat mitigated by the inclusion of correlation distance, which solely takes into account the topographic distribution of the connectivity values. However, it still does not consider the relative spatial position of the edges in the connectivity network, therefore ignoring the fact that spatially distinct changes in connectivity can signify more prevalent changes in brain network dynamics, compared to spatially similar differences (i.e. when two such connectivity edges

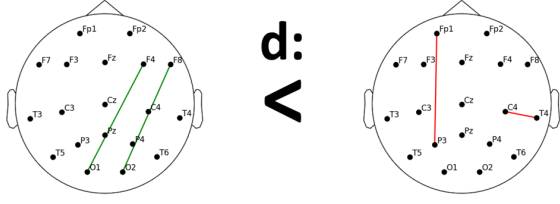


Fig. 1: A mock example for the main property of the spatially informed distance function: if two connectivity graphs differ in edges which connect similar brain areas (green edges), then the magnitude of that difference (d) should be lower compared to when they differ in edges that connect physically distinct brain areas (red edges).

change significantly, which connect a pair of similar brain areas). This conundrum is further illustrated in figure 1.

We theorize, that a distance function, which takes not only topography, but relative physical position into account, could lead to more informative cluster centroids, that represent significantly different brain networks from a functional standpoint. To create such a distance function, we propose an approach called Spatially Informed Graph Distance (SIGD). The core idea of the current state of this function between connectivity graphs G_1 and G_2 , assuming both graphs have the same topography and physical structure, therefore they only differ in connectivity values (edge weights), is the following:

$$d(G_1, G_2) = \sum_{i,j} \Delta w_i \Delta w_j \kappa(\delta_{ij}) \quad (2)$$

where $\Delta w_i = |w_i^{G_1} - w_i^{G_2}|$ is the element-wise absolute difference between the two connectivity vectors, δ_{ij} is the physical distance between connectivity edges i and j , and κ is a custom nonnegative kernel function. The main goal of the kernel function κ is to suppress spatially similar edges (i.e. edges that are close to each other), and is initially defined as:

$$\kappa(\delta_{ij}) = 1 - \exp\left(-\frac{\delta_{ij}^2}{2\sigma}\right) \quad (3)$$

with σ as a hyperparameter to control spatial smoothing. Finally, as the spatial distance between edges (as opposed to electrodes) is not trivial, δ_{ij} was defined as the mean distance between the closest non-connected node-pair, and the "other" non-connected pair. For clarification, refer to functions d_{min} and d_{max} in one of our earlier work regarding nonparametric statistical testing for connectivity graphs [21]. To ease computation, the distance function can be written in matrix form:

$$d(G_1, G_2) = \Delta w^T K \Delta w \quad (4)$$

assuming Δw is a column vector, and $K_{ij} = \kappa(\delta_{ij})$ is a precomputed kernel matrix.

The SIGD dissimilarity metric, in its current form, is strictly nonnegative (because of the absolute difference in Δw and the nonnegativity of K), and in the edge-case that two graphs only

differ in a single connectivity value, their dissimilarity will be 0, given that $\delta_{ii} = 0$, which is a common property for distance functions. In order to mitigate this latter effect, SIGD could be applied in conjunction with other dissimilarity measures, the simplest of which being absolute distance.

D. Evaluation Plan

After defining and identifying dFC states, the dFC time-series can be segmented into discrete sequences, where each time point belongs to a distinct state. To describe this resulting sequence, four main parameters are proposed, mostly rooted in classic microstate analysis:

- 1) **Time coverage:** the overall proportion of time covered by the given dFC state.
- 2) **Occurrence:** the average number of occurrences (per second) of the given dFC state.
- 3) **Mean duration:** the mean duration (in time) of the given dFC state, when it occurs.
- 4) **Mean correlation:** the mean correlation of data points belonging to the given dFC state with its defining centroid (only applicable if such a centroid is present).

The distribution of these parameters can also be computed over a dFC representation, possibly providing additional insight into connectivity dynamics. By measuring the discriminative and statistical power of these parameters while comparing different subject groups, **we can evaluate the usefulness of the dFC state segmentation** for the diagnosis and categorization of various neurodegenerative and psychological disorders, most notably Alzheimer's Disease, dementia, schizophrenia, and their subtypes.

IV. CURRENT STATUS

The applicability of k-means clustering over an Euclidean space and vectorized connectivity values, was tested as a form of preliminary analysis, using control subjects of the the AHEPA dataset [22]. This dataset contains resting-state EEG recordings from 29 healthy control subjects, and additional recordings from other subjects having Alzheimer's Disease and Mild Cognitive Impairment (MCI).

A. Validating dFC State Criteria

In order to examine the *Convergence* and *Discreteness* criteria for dFC states, two separate approaches were used, one for each requirement.

To examine if the identified states converge, the subjects were randomly divided into disjunct groups, then within-group centroids were estimated using k-means clustering, and finally, the similarity of group centroids were measured. In theory, if the convergence criterion holds, then the between-group similarity of centroids should increase as the size of the groups increases, suggesting that the centroids converge towards a common set of topographies. Pairing the centroids to be compared is an issue without a clear solution, therefore two pairing techniques were employed: (a) pairing them based on similarity, via the greedy bipartite matching algorithm, where the most similar pair gets selected first, then the second most

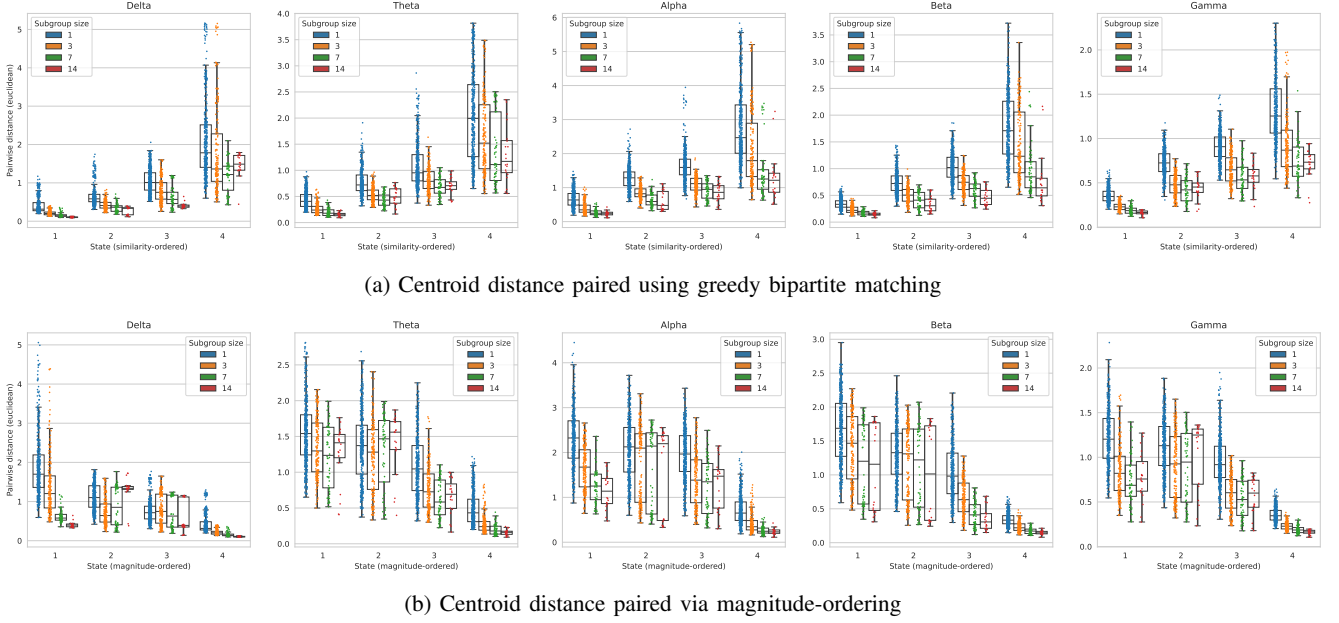


Fig. 2: Pairwise distance between dFC centroids of different groups, within five distinct frequency bins. The centroids were determined using k -means clustering in Euclidean space, $k = 4$. The effect of group size over the distances presumably indicate convergence in most cases, except for the middle two centroids in figure 2b. It has to be noted, that the magnitude of the middle two centroids were close to each other, therefore the confidence of the magnitude ordering should be considered low.

similar, and so on, until all centroids are paired, and (b) pairing based on magnitude (average connectivity), where the centroids of both groups were ordered based on magnitude, and centroids with the same index on both sides were compared. The corresponding results for both pairing approaches are presented in figure 2, using k -means clustering with $k = 4$ in an Euclidean space. As these results show, the centroids seem to converge as the group size increases, suggesting that the first criterion holds for dFC states defined using k -means clustering. This investigation was also conducted for correlation distance (using the k -medoids algorithm instead of k -means), and to higher values of k , with both leading to the same conclusion.

Investigating the second criterion (discreteness) carries a higher degree of complexity, as the definition of stable and transient dynamics is not universal. Initially, we defined transient sections in the Euclidean space as ones that traverse through a large distance while following an approximately linear trajectory. To measure this phenomenon, the linear “speed” of the dFC time-series was estimated by taking the average Euclidean distance of each point from their neighbours within a given time window. The window size was inversely proportional to the frequency, over which spectral connectivity was measured, to account for dynamics over different timescales. This linearity was matched against the distance (in time) from the nearest state-change, and the results plotted in a histogram can be observed in figure 3. As the “L”-shaped pattern on these results suggest, linearity is highest near the state-change points. There is also a large cluster in every band where the distance from the nearest change-

point and linearity is both low, which is probably caused by a large number of short states (typically under 150 milliseconds, depending on the frequency band), suggesting the need for temporal smoothing over the dFC state segmentation. Overall, the results indicate that the second criterion holds, but only for states of sufficient length.

B. Predictive Performance

During a “proof-of-concept” analysis, the extracted parameters of the state segmentation distinguished the CN group of the AHEPA dataset from the AD and FTD groups with high statistical power, after compensating for multiple comparisons, which was required by the relatively large amount of computed parameters (4 parameters, multiplied by the number of dFC states and frequency bands). Furthermore, the predictive performance of simple machine learning techniques (particularly, Random Forest and Support Vector classifiers) surpassed those which were fitted to time-averaged functional connectivity data, suggesting that the temporal dynamics of resting-state functional connectivity carry additional information connected to Alzheimer’s Disease and frontotemporal dementia.

V. FUTURE WORK AND TIMELINE

After completing the preliminary study, we plan to test the applicability of the SIGD metric, in terms of identifying potential biomarkers for early AD and dementia, and further improve the method, if necessary. Additionally, we plan to further validate the discriminative power of the extracted parameters (based on the state segmentation) with regards to

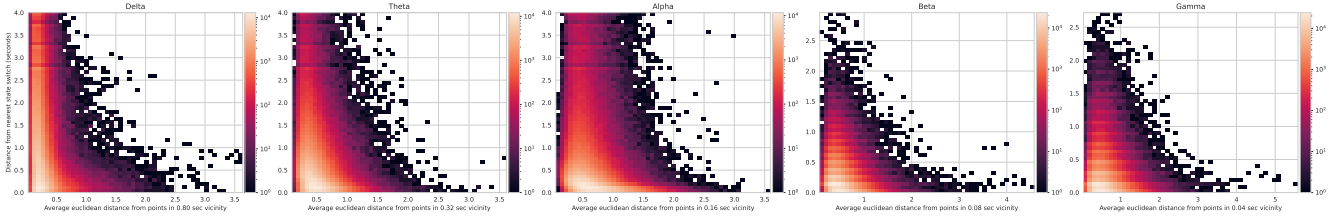


Fig. 3: Histograms of the average Euclidean distance from the neighboring instances within a time window (linearity) and distance in time from the nearest state-change, in each frequency band. The color scale is logarithmic.

AD and MCI disease progression on multiple datasets. Finally, we plan to investigate the possibilities for different dFC state definitions, and the application of HMMs for time-resolved state detection and segmentation.

We expect to finish the majority of the outlined work according to the following schedule:

Month 1: Evaluation of the SIGD metric, developing a HMM-based approach for state detection.

Month 2: Evaluating the performance metrics.

Months 3-4: Investigating alternative dFC state definitions, while balancing interpretability and expressive power.

Months 5-6: Additional evaluation and publishing.

REFERENCES

- [1] D. A. Handwerker, V. Roopchansingh, J. Gonzalez-Castillo, and P. A. Bandettini, "Periodic changes in fmri connectivity," *Neuroimage*, vol. 63, no. 3, pp. 1712–1719, 2012.
- [2] R. F. Betzel, M. A. Erickson, M. Abell, B. F. O'Donnell, W. P. Hetrick, and O. Sporns, "Synchronization dynamics and evidence for a repertoire of network states in resting eeg," *Frontiers in computational neuroscience*, vol. 6, p. 74, 2012.
- [3] A. Khanna, A. Pascual-Leone, C. M. Michel, and F. Farzan, "Microstates in resting-state eeg: current status and future directions," *Neuroscience & Biobehavioral Reviews*, vol. 49, pp. 105–113, 2015.
- [4] A. Mishra, B. Englitz, and M. X. Cohen, "Eeg microstates as a continuous phenomenon," *Neuroimage*, vol. 208, p. 116454, 2020.
- [5] A. M. Bastos and J.-M. Schoffelen, "A tutorial review of functional connectivity analysis methods and their interpretational pitfalls," *Frontiers in systems neuroscience*, vol. 9, p. 175, 2016.
- [6] F. Jiang, H. Jin, Y. Gao, X. Xie, J. Cummings, A. Raj, and S. Nagarajan, "Time-varying dynamic network model for dynamic resting state functional connectivity in fmri and meg imaging," *NeuroImage*, vol. 254, p. 119131, 2022.
- [7] E. Damaraju, E. A. Allen, A. Belger, J. M. Ford, S. McEwen, D. Mathalon, B. Mueller, G. Pearlson, S. Potkin, A. Preda, *et al.*, "Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia," *NeuroImage: Clinical*, vol. 5, pp. 298–308, 2014.
- [8] M. Demirtaş, C. Tornador, C. Falcón, M. López-Solà, R. Hernández-Ribas, J. Pujol, J. M. Menchon, P. Ritter, N. Cardoner, C. Soriano-Mas, *et al.*, "Dynamic functional connectivity reveals altered variability in functional connectivity among patients with major depressive disorder," *Human brain mapping*, vol. 37, no. 8, pp. 2918–2930, 2016.
- [9] X. Chen, H. Zhang, Y. Gao, C.-Y. Wee, G. Li, D. Shen, and A. D. N. Initiative, "High-order resting-state functional connectivity network for mci classification," *Human brain mapping*, vol. 37, no. 9, pp. 3282–3296, 2016.
- [10] K. Guan, Z. Zhang, X. Chai, Z. Tian, T. Liu, and H. Niu, "Eeg based dynamic functional connectivity analysis in mental workload tasks with different types of information," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 632–642, 2022.
- [11] C. Chu, Z. Zhang, J. Wang, Z. Li, X. Shen, X. Han, L. Bai, C. Liu, and X. Zhu, "Temporal and spatial variability of dynamic microstate brain network in early parkinson's disease," *npj Parkinson's Disease*, vol. 9, no. 1, p. 57, 2023.
- [12] G. Yi, L. Wang, C. Chu, C. Liu, X. Zhu, X. Shen, Z. Li, F. Wang, M. Yang, and J. Wang, "Analysis of complexity and dynamic functional connectivity based on resting-state eeg in early parkinson's disease patients with mild cognitive impairment," *Cognitive neurodynamics*, vol. 16, no. 2, pp. 309–323, 2022.
- [13] P. Detti, G. Z. M. de Lara, R. Bruni, M. Pranzo, F. Sarnari, and G. Vatti, "A patient-specific approach for short-term epileptic seizures prediction through the analysis of eeg synchronization," *IEEE transactions on biomedical engineering*, vol. 66, no. 6, pp. 1494–1504, 2018.
- [14] S. Nobukawa, M. Kikuchi, and T. Takahashi, "Changes in functional connectivity dynamics with aging: a dynamical phase synchronization approach," *Neuroimage*, vol. 188, pp. 357–368, 2019.
- [15] H. A. Marusak, V. D. Calhoun, S. Brown, L. M. Crespo, K. Sala-Hamrick, I. H. Gotlib, and M. E. Thomason, "Dynamic functional connectivity of neurocognitive networks in children," *Human brain mapping*, vol. 38, no. 1, pp. 97–108, 2017.
- [16] J. Li, D. Zhang, A. Liang, B. Liang, Z. Wang, Y. Cai, M. Gao, Z. Gao, S. Chang, B. Jiao, *et al.*, "High transition frequencies of dynamic functional connectivity states in the creative brain," *Scientific reports*, vol. 7, no. 1, p. 46072, 2017.
- [17] S. Zhou, G. Zou, J. Xu, Z. Su, H. Zhu, Q. Zou, and J.-H. Gao, "Dynamic functional connectivity states characterize nrem sleep and wakefulness," *Human Brain Mapping*, vol. 40, no. 18, pp. 5256–5268, 2019.
- [18] H. Jin, K. G. Ranasinghe, P. Prabhu, C. Dale, Y. Gao, K. Kudo, K. Vossel, A. Raj, S. S. Nagarajan, and F. Jiang, "Dynamic functional connectivity meg features of alzheimer's disease," *NeuroImage*, vol. 281, p. 120358, 2023.
- [19] H. Chen, Y. Lei, R. Li, X. Xia, N. Cui, X. Chen, J. Liu, H. Tang, J. Zhou, Y. Huang, *et al.*, "Resting-state eeg dynamic functional connectivity distinguishes non-psychotic major depression, psychotic major depression and schizophrenia," *Molecular Psychiatry*, vol. 29, no. 4, pp. 1088–1098, 2024.
- [20] M. Vinck, R. Oostenveld, M. Van Wingerden, F. Battaglia, and C. M. Pennartz, "An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias," *Neuroimage*, vol. 55, no. 4, pp. 1548–1565, 2011.
- [21] M. Vetró and G. Hullám, "Nonparametric statistical testing of functional connectivity in eeg data," *Proceedings of the 31th Minisymposium of the Department of Measurement and Information Systems*, vol. 31, pp. 25–30, 2024.
- [22] A. Miltiadous, K. D. Tzamoura, T. Afrantou, P. Ioannidis, N. Grigoriadis, D. G. Tsalikakis, P. Angelidis, M. G. Tsipouras, E. Glavas, N. Giannakeas, *et al.*, "A dataset of scalp eeg recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine eeg," *Data*, vol. 8, no. 6, p. 95, 2023.

Towards Efficient Separability of Models in Automated Worst-Case Circuit Analysis

Pál Weisz, György Orosz

Budapest University of Technology and Economics

Department of Artificial Intelligence and Systems Engineering

Budapest, Hungary

Email: {weisz, orosz}@mit.bme.hu

Abstract—During worst-case circuit analysis (WCA), the primary task is to identify extreme values, and it is crucial to perform WCA efficiently. In the case of circuit analysis tasks, the size and complexity (number of parameters) of circuit designs are increasing, which increases the time required for simulations or calculations. As a possible solution for complexity reduction, we set the efficient separation/decomposition of circuit models as our research goal.

The decomposition reduces the WCA task to low-dimensional subproblems, making the task less computationally demanding. Decomposition aims to group the parameters into subsets such that the parameters in disjoint subsets do not influence each other's effect on the output.

In this study, we investigate the potential advantages of model decomposition. As a possible method for decomposition, the combination of Monte Carlo simulation and sensitivity analysis is used.

Index Terms—worst-case analysis, circuit model, separability, parameter independence, decomposition, sensitivity analysis, Monte Carlo simulation

I. PROBLEM STATEMENT & MOTIVATION

Safety-critical electronic systems require rigorous testing to ensure that they function properly within the component tolerances and under extreme environmental conditions. Worst-case analysis (WCA) is a crucial component of this process, as it evaluates the behavior of circuits under adverse combinations of uncertainties [1] [2]. However, as circuit designs have grown in size and complexity, the computational time of WCA has increased significantly, and large systems often require tens of hours of analysis, despite the need for efficient analysis processes in industrial environments.

Existing WCA approaches rely mainly on monolithic circuit models and use various circuit simulation techniques and tools (e.g., SPICE). These methods, although efficient for small models, are poorly scalable; the main obstacle to SPICE-based simulations is the parallelization of the solving algorithms [3]. A solution could be to break down a complex model into several small subtasks, allowing for faster computation through parallelization.

The relevance of circuit failure analysis is further enhanced by the increasing role of various generative artificial intelligence-based methods (e.g., ChipGPT) [4] in circuit design, which require efficient execution of analysis tasks due to the lack of human control.

In relation to this aspect, explainability can be a crucial consideration in many cases, as safety-critical circuits require verification of the analysis results. Explaining the partial results in a way that is understandable and logical to humans can help justify the credibility of the entire analysis.

The fundamental objective is to decompose complex circuit models into subcircuits that have fewer parameters. There are several advantages of having identified separate subcircuits:

- The behavior of each subcircuit can be observed independently, which supports explainability: it is possible to identify which parts affect the operation of the whole circuit and how.
- Due to the reduced complexity, simulations and further calculations can be done more efficiently.

A simple example of how decomposition can speed up the WCA process can be illustrated with the Extreme Value Analysis (EVA) method [2]. EVA assumes that a minimum or maximum exists at the extreme values of the parameters, so it requires 2^N evaluation of the circuit to be analyzed to find the worst-case solution. If we partition the N parameters into two subsets $N = N_1 + N_2$, the required evaluation number is $2^{N_1} + 2^{N_2}$. For example, if $N = 8$ we have $2^8 = 256$ evaluations, and if $N_1 = 4, N_2 = 4$ we have only $2^4 + 2^4 = 32$ evaluations.

This paper argues that efficient model separability is a key component for scalable, automated worst-case circuit analysis.

II. RELATED WORK

Several approaches to decomposing circuit models have been reported in the literature. While these methods generally pursue objectives similar to those of this work, many of them adopt a different perspective or emphasize alternative aspects of the problem.

In [6], the authors introduce a method for analyzing large analog circuits by decomposing them into smaller subsystems and employing Determinant Decision Diagrams (DDD) to represent symbolic expressions and accelerate circuit analysis compactly. An improved variant of this approach, employing an s-domain simulation technique for linear analog and interconnect circuits, has also been reported [7]. In some instances, it is possible to create an analytical representation of a circuit model [8]. If this option is available, excluding the simulation step from the process is possible.

Prior work [9] proposes circuit partitioning to compute the DC operating point in SPICE. The approach accelerates the computation by handling smaller Jacobian matrices and employs techniques such as virtual probes to partition circuits without sacrificing accuracy.

There are similar graph-based approaches in the literature that focus primarily on decomposing the network representing the circuit as a graph, while ensuring that the boundary conditions defined by the model (e.g., coupling) are not violated.

The key idea of the finite-difference-time-domain (FDTD) Diakoptics method [10] is to realize modular computation of large circuits by dividing the circuit into segments and representing passive linear segments by their impulse response (or impulse response matrix).

As presented in [11], circuit decomposition is carried out using system sensitivities and Kron reduction, with a distributed algorithm used to efficiently manage the resulting subproblems while limiting data sharing.

The partitioning framework presented in [12] combines connectivity-based clustering with an enhanced tabu search to minimize edge cuts in circuit partitioning. The methodology includes graph coarsening, connectivity-aware initial partitioning, followed by uncoarsening and refinement.

WCA plays a significant role in Very Large Scale Integration (VLSI) circuit design. Prior research has demonstrated the use of Support Vector Machine-based behavioral modeling techniques to decompose large-scale circuits into smaller, manageable components [13].

A partially symbolic sensitivity analysis method is presented in [14]. The proposed method requires knowledge of the circuit model structure, as it introduces an auxiliary circuit to aid sensitivity analysis. Also, the authors agree that a large number of parameters would make the size of the circuit that could be handled prohibitive.

Various sensitivity analysis methods have been comprehensively reviewed in [15], while similar investigations can be found in [16]. The state-of-the-art methods for sensitivity analysis, according to these articles, include output variance decomposition-based methods, such as Sobol indices, Jansen (Winding Stairs) method, FAST and eFAST (which are also related to spectral analysis), Morris (One-Factor-at-a-Time, OFAT) screening, and statistical regression-based metamodeling. A more recent study evaluates the practical performance of these methods [17].

III. CORE VISION

A. Mathematical Formulation of Separability

We assume that the worst-case analysis problem can be described by a function $f(\mathbf{x})$, so we can analyze the function to see how the system can be decomposed into smaller independent models.

We consider the real-valued output of the system, which is influenced by multiple real-valued parameters.

Let us further assume that the relationship between the output and these parameters can be described by a multivariate function:

$$\mathbf{x} = [x_1, x_2, \dots, x_n] \in \mathbb{R}^n \quad (1)$$

$$f(x_1, x_2, \dots, x_n) = f(\mathbf{x}) = y \quad f: \mathbb{R}^n \mapsto \mathbb{R} \quad (2)$$

These conditions are not strict constraints on electrical networks, but they limit the generality of the problem.

Each parameter has a bounded interval:

$$x_i \in X_i = [x_{i_{min}}, x_{i_{max}}]. \quad (3)$$

The multivariate function f establishes a link between the output and the input. For the worst-case problem, the task is to find the minimum and/or maximum values of $f(\mathbf{x})$.

This is a well-known problem in mathematics, and there are several algorithms for solving it. Due to the constraint on the parameters that they are bounded, the problem in question can be solved, e.g., using constrained optimization algorithms.

However, the complexity of the problem increases proportionally to the number of parameters. Therefore, the goal is to decompose the original optimization problem into several independently solvable problems, and the solution to the original problem can be derived from the partial results.

The aim is then to define a disjoint binary grouping of the parameters, for which it is true that none of the members of the groups are needed simultaneously to evaluate the subproblems. Formally:

$$f(\mathbf{x}) = \mathcal{P}\{g_1(\mathbf{x}^{[1]}), g_2(\mathbf{x}^{[2]}), \dots, g_k(\mathbf{x}^{[k]})\} \quad (4)$$

$$\mathbf{x}^{[1]} \cup \mathbf{x}^{[2]} \cup \dots \cup \mathbf{x}^{[k]} = \mathbf{x} \quad (5)$$

$$\mathbf{x}^{[1]} \cap \mathbf{x}^{[2]} \cap \dots \cap \mathbf{x}^{[k]} = \emptyset \quad (6)$$

In equation (4), the operator $\mathcal{P}$ connects the subproblems $g_1 \dots g_k$, which are functions of $\mathbf{x}^{[1]}, \dots, \mathbf{x}^{[k]}$ disjoint sets of parameters.

Since worst-case circuit analysis focuses on extreme values, all functions considered in this work are assumed to attain finite minima and maxima over the bounded parameter domain $X = X_1 \times \dots \times X_n$. Accordingly, both the function $f(\mathbf{x})$ and the subproblems $g_i(\mathbf{x}^{[i]})$ are analyzed primarily in terms of their extreme values.

Separability with respect to any extreme value searching in this context means that, e.g., if $\mathbf{x}_M$ is the global maximum of $f(\cdot)$:

$$\mathbf{x}_M = \arg \max_{\mathbf{x}} f(\mathbf{x}) \quad (7)$$

and $\mathbf{x}_M^{[i]}$ is the maximum of $g_i(\cdot)$

$$\mathbf{x}_M^{[i]} = \arg \max_{\mathbf{x}^{[i]}} g_i(\mathbf{x}^{[i]}) \quad (8)$$

then the maximum of $f(\cdot)$ is the union of maxima of $g_i(\cdot)$:

$$\mathbf{x}_M = \mathbf{x}_M^{[1]} \cup \mathbf{x}_M^{[2]} \cup \dots \cup \mathbf{x}_M^{[k]}. \quad (9)$$

Without the loss of generality, it is possible to perform extreme value searching for the minimum of $f(\cdot)$ similarly.

Based on our efforts in this area of research, we can assume that in most of the cases $\mathcal{P} = "+"$, i.e.:

$$f(\mathbf{x}) = \mathcal{P}\{g_1(\mathbf{x}^{[1]}), g_2(\mathbf{x}^{[2]})\} = g_1(\mathbf{x}^{[1]}) + g_2(\mathbf{x}^{[2]}) \quad (10)$$

A similar approach was reported in the literature [18].

In many practical WCA scenarios, the objective function can be expressed or transformed into an additive form through monotonic transformations. While the suggested mathematical formulation does not exclude other composition operators $\mathcal{P}$, the additive case is emphasized here because it guarantees that extrema of the function f can be reconstructed from the extrema of independent subproblems. The implications and limitations of this assumption are further discussed in Section VI.

B. Theoretical Background

There are different specific solutions that address certain model optimization and circuit simulation tasks. However, to the best of the authors' knowledge, there is no related work in the field of worst-case circuit analysis regarding model separability problems based on sensitivity analysis.

When it comes to model separation, it can encompass several concepts, so it is essential to define what model separation means in this context. During worst-case circuit analysis, the model itself is an electrical model: having different components with numerous parameters, and these components are connected to each other forming a graph-like structure often represented by schematic diagrams. Performing a worst-case analysis also means that different tasks may determine the aspects of parameter dependencies differently.

Our notion of dependency goes beyond structural constraints: depending on the task, different parameters may or may not be functionally dependent, affecting separability. The goal is not necessarily the actual separability of the circuits, but rather the separability for a given task.

The proposed research direction differs from and presents something new compared to the methods found in the literature in that it optimizes circuit separation based on sensitivity analysis.

The linear error propagation method is based on a first-order Taylor series approximation of the functions. It is basically used for tolerance-type uncertainty of parameters. The basic idea is that the parameters $\mathbf{x}$ of an analysis problem have an operating point $\mathbf{x}_0$, for which we know both the bounded tolerance intervals and the function relation $f(\mathbf{x})$ describing the output. We replace the function at its operating point by a linear term, and then examine, parameter by parameter x_i , how much its error causes the output to change.

$$f(\mathbf{x}_0) + \Delta f(\mathbf{x}_i) \approx f(\mathbf{x}_0) + \frac{\partial f(\mathbf{x}_0)}{\partial x_i} \cdot \Delta x_i \quad (11)$$

Sensitivity analysis reveals how changes in each parameter change the output value. For each parameter, we can define sensitivity $S^{(i)}$:

$$S^{(1)} = \frac{\partial f(\mathbf{x}_0)}{\partial x_1}, S^{(2)} = \frac{\partial f(\mathbf{x}_0)}{\partial x_2}, \dots, S^{(i)} = \frac{\partial f(\mathbf{x}_0)}{\partial x_i} \quad (12)$$

In the initial stage of the research, the relationships among parameters are evaluated using a method based on classical sensitivity analysis to verify the effectiveness of the proposed approach. Alternative approaches, including other derivative-based methods, may be explored in future work.

Monte Carlo simulation is widely used in worst-case analysis to assess the impact of component tolerances and process variations via statistical sampling. Unlike analytical or symbolic methods that rely on symbolic system equations or linearization and often yield overly conservative worst-case estimates, Monte Carlo analysis captures nonlinear effects and parameter interactions.

When combined with derivative-based sensitivity analysis, Monte Carlo analysis can serve as a practical approach to improving the efficiency of circuit separation.

Monte Carlo simulations are well-suited for this purpose, as they enable thorough exploration of the parameter space and facilitate the identification of operating conditions that affect circuit behavior in different ways, such as operational amplifiers with or without slew-rate limitation, saturation of circuit components, etc.

Although Monte Carlo simulation explores the full parameter space, in the proposed approach it is used only once as an exploratory preprocessing step. Its purpose is not to directly determine worst-case values, but to identify parameter dependencies via sensitivity behavior.

C. Proposed Approach

The core idea is based on classic derivative-based sensitivity analysis, with a slightly different approach addressing the identification of the parameters' mutual dependence.

We make the following assumptions:

- The function $f(\mathbf{x})$ is not necessarily available in analytical or symbolic form.
- The evaluation of the circuit could be computationally expensive.
- We already have results from a Monte Carlo simulation.

It is assumed that the function $f(\mathbf{x})$ does not admit a closed-form analytical or symbolic representation; otherwise, the separability process would be considerably simplified. Instead, the system behavior is accessible only through numerical evaluations obtained from, e.g., Monte Carlo simulations. While separability determination is more straightforward when an analytical expression is available, the proposed approach is also applicable to scenarios in which only simulation-based models (e.g. SPICE) are available.

The second assumption is used, because if the evaluation of the circuit model is fast, the significance of the decomposition could decrease.

The third assumption is used, because our algorithm is based on a Monte Carlo simulation. Monte Carlo simulations are often used in industrial and scientific projects, so they can be used for other analysis purposes besides independence analysis.

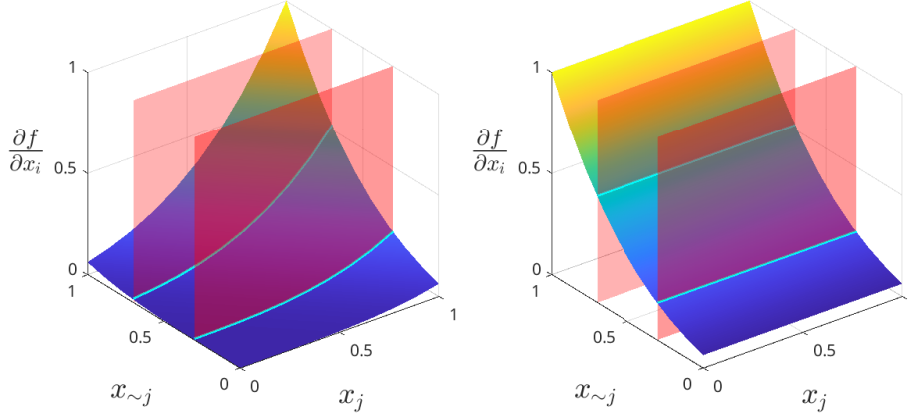


Fig. 1. Illustration of the proposed approach in a 2-dimensional parameter space for cases when x_i and x_j are dependent and when they are independent.

We outline a potential approach that satisfies the assumptions outlined previously. The steps would be as follows:

- Formulation of the optimization problem.
- Execution of an initial Monte Carlo simulation on the circuit model.
- Analysis of parameter dependencies based on the simulation results and sensitivity analysis.
- Model decomposition.
- Detailed analysis of the decomposed circuit models using parallel execution of simplified simulation tasks.

The core idea behind the separation method is to observe the effect of one parameter on the sensitivity of the output to another parameter:

$$f_j^{(i)}(\mathbf{x}) = \left. \frac{\partial f(\mathbf{x})}{\partial x_i} \right|_{x_{\sim j} = \text{const.}} \quad (13)$$

$$x_{\sim j} = [x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n] \in \mathbb{R}^{n-1} \quad (14)$$

This heuristic is employed to assess the influence of parameters on one another. If sensitivity remains unchanged with respect to another parameter, the parameters are considered independent. This heuristic approach is also substantiated by the literature. [19]

The essence of the method is the following. Select points from the Monte Carlo simulation, along one parameter at a time (x_j), and examine the output sensitivity of x_i by fixing the values of all parameters other than x_j . This is essentially a one-factor-at-a-time screening, which is a standard method in sensitivity analysis. [15] [16]

Continue by using numerical gradient estimation methods for the sensitivity analysis, relying on the points obtained as a result of the Monte Carlo simulation.

Since the function $f(\mathbf{x})$ is not available in analytical form, sensitivities are estimated numerically based on simulation data. In this work, derivative-based sensitivities are approximated using finite-difference schemes evaluated at selected Monte Carlo samples.

For parameter x_i , the local sensitivity is approximated as

$$\frac{\partial f}{\partial x_i} \approx \frac{f(x_i + \Delta x_i) - f(x_i)}{\Delta x_i} \quad (15)$$

where the values of $f(x_i + \Delta x_i)$ and $f(x_i)$ are coming from the simulation data.

After completion of this step, eq. (13) is evaluated once for the specified parameters along x_j . As shown in Fig. 1., the red plane (an affine subspace in higher dimensions) intersects with the surface resulting in $f_j^{(i)}(\mathbf{x})$ as the intersection line.

The values of the initially fixed parameters ($x_{\sim j}$) should also be selected based on curated sampling methods, and multiple sweeps are to be performed along the parameter of interest to ensure thorough coverage of the parameter space.

Using this method, sensitivity trajectories described by eq. (13) along all parameters are examined iteratively, resulting in a data set that captures the output sensitivity with respect to all the parameters as a function of the whole system parameter set, at least within the examined tolerance interval. Up to this point, only a single Monte Carlo simulation would need to be run.

The resulting gradient estimates are treated as stochastic quantities due to sampling noise, which motivates the use of statistical tests when assessing parameter independence based on the obtained data. Such methods involve examining the different features and distributions of the resulting $f_j^{(i)}(\mathbf{x})$ functions.

After reliably identifying the parameter dependencies and forming disjoint groups of dependent parameters, the model separation can be performed, allowing further, more efficient simulations.

IV. EVALUATION PLAN & METHODOLOGY

The methodology used to carry out the research can be structured as follows:

Reviewing existing literature and obtaining the necessary background knowledge, based on the conceptual design of the

decomposition method, statistical methods are employed to identify the independent parts.

To this end, further examination of the related work is necessary in the following areas:

- sampling methods for Monte Carlo simulations;
- gradient estimation based on a finite number of samples, when the function is not necessarily known analytically, but only at specific points where the value of the function is known;
- statistical methods for independence analysis.

Algorithm development: the core of the research work is the integration of the above ideas into a single method suitable for performing the decomposition task. This involves testing several methods to see which one best fits the task at hand.

Implementing and verifying methods based on our previous work in this area:

- implement the designed algorithms in MATLAB, generate the data needed to run them using LTspice or MATLAB,
- The circuit samples used for testing are collected from our previous research work and are publicly available [20].

V. EXPECTED CONTRIBUTIONS

By solving the current research problem, we can address problems that require human intuition, such as breaking down an existing circuit into smaller functional blocks. In this way, complex analysis tasks can be decomposed into smaller sub-tasks, thus reducing the resulting complexity of the analysis task.

For circuit analysis tasks, the execution time of simulations or calculations can be critical, especially if the evaluation is not to be performed at a single operating point. However, instead, the analysis must be conducted under different parameters and environmental factors. If the analysis can be performed on smaller subcircuits, the number of parameter combinations will be significantly reduced, thus reducing the analysis time. Decomposition of circuits can help obtain meaningful, explainable analysis results by avoiding the need to understand/process the behavior of a complex circuit as a whole, but rather by investigating the behavior of smaller parts of the circuit.

Decomposition has traditionally been a fundamental method in engineering problem-solving, and its automated solution can speed up processes that rely on human intuition.

The development of analysis methods will ultimately support the verification of devices containing electrical circuits at the design stage, thus enabling faster device development cycles and safer devices.

This research is expected to rigorously establish the validity of the proposed approach through both formal mathematical analysis and practical experimentation. The comprehensive development and systematic verification of the approach are planned for the next stage, paving the way for its potential deployment and broader impact.

VI. RESEARCH CHALLENGES & OPEN QUESTIONS

During the course of this research, several assumptions introduced earlier still require thorough examination. To realize this vision, the mathematical characterization of separability must be established, the limitations and potential approaches of Monte Carlo sampling methods must be investigated, and efficient techniques for gradient estimation in high-dimensional spaces must be developed.

We assumed in III-A that in most cases, $\mathcal{P}$ represents the addition operator. Furthermore, we also assume that separability is possible if the analysis task can be transformed in such a way that it can be expressed as a sum of functions.

For instance, when the transfer function of a linear system is expressed as a product of rational fractions, the logarithm of this function can be equivalently represented as the sum of the logarithms of its individual factors. It should be noted, however, that all factors in the product must be strictly positive; otherwise, the logarithm becomes complex-valued, in which case the mapping can no longer be regarded as monotonic.

So far, we cannot provide a mathematical formulation to determine whether a particular problem is separable or not.

The large parameter space also poses a significant challenge, as Monte Carlo analysis requires evaluating the function at multiple points; however, as the dimensionality increases, these points become increasingly sparse. This sparsity is particularly problematic for gradient estimation, which typically relies on nearby points to compute numerical gradients. As the space becomes sparser, these points lie farther apart and consequently provide a poor characterization of the gradient in the region under investigation.

Inaccurate gradient estimation may distort the characteristics of $f_j^{(i)}(\mathbf{x})$ in (13), and potentially its distribution, thereby affecting the decision of whether the parameters should be classified as independent. Future work will address this by modeling estimation noise and using statistically grounded thresholding based on sensitivity significance rather than strictly zero or fixed thresholds.

VII. SUMMARY

This vision paper proposes a promising direction for improving worst-case circuit analysis through a novel model decomposition approach based on sensitivity analysis and Monte Carlo simulation.

Examining changes in parameter sensitivities may reveal parameter dependencies and enable effective circuit partitioning. This approach could reduce computational complexity and improve the explainability of large-scale circuit analysis.

Future research may yield automated, scalable, and more interpretable analysis techniques that support faster design cycles and enhanced safety as circuit complexity increases.

ACKNOWLEDGMENT

This research was conducted with the support of the Ministry of Culture and Innovation and the National Research, Development and Innovation Fund within the framework of the Doctoral Excellence Scholarship Program (DKÖP-26).

This work was supported in part by the RESCHIP4EU project of HADEA (ref. No: 101158828-RESCHIP4EU).

REFERENCES

- [1] B. Johanson, D. Russell, W. Swavely, Reliability Analysis Center, "Worst case circuit analysis application guidelines," Reliability Analysis Center, 1993
- [2] B. A. Lenertz, "Electrical design worst-case circuit analysis: guidelines and draft standard (REV A)," AEROSPACE REPORT NO. TOR-2013-00297, The Aerospace Corporation, June 3, 2013.
- [3] Chen, X., Wang, Y. & Yang, H. "Parallel sparse direct solver for integrated circuit simulation," Springer International Publishing, 2017, 3, doi: 10.1007/978-3-319-53429-9
- [4] Chang, K., Wang, Y., Ren, H., Wang, M., Liang, S., Han, Y., Li, H. & Li, X. "ChipGPT: How far are we from natural language hardware design," (2025), <https://arxiv.org/abs/2305.14019>, doi: 10.48550/arXiv.2305.14019
- [5] Charras-Garrido, M. & Lezaud, P. "Extreme Value Analysis: an Introduction," *Journal De La Société Française De Statistique*. 154, 66-97 (2013)
- [6] X. Tan and C.-J. R. Shi, "Hierarchical symbolic analysts of large analog circuits with determinant decision diagrams," *1998 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1998, vol. 6, pp. 318–321 vol.6., doi: 10.1109/ISCAS.1998.705275
- [7] Tan, S. A general hierarchical circuit modeling and simulation algorithm. *IEEE Transactions On Computer-Aided Design Of Integrated Circuits And Systems*. 24, 418-434 (2005), doi: 10.1109/TCAD.2004.842815
- [8] B. Bak, P. Weisz, Gy. Orosz, "Worst-Case Circuit Analysis using Schematic Conversion and Decomposition," *2024 25th International Carpathian Control Conference (ICCC)*, Krynica Zdrój, Poland, 22-24 May 2024, doi: 10.1109/ICCC62069.2024.10569978
- [9] Bizzarri, F., Brambilla, A. & Storti Gajani, G. "FastSpice circuit partitioning to compute DC operating points preserving Spice-like simulators accuracy," *Simulation Modelling Practice And Theory*. 81 pp. 51-63 (2018), doi: 10.1016/j.simpat.2017.12.001
- [10] Huang, T., Houshmand, B. & Itoh, T. "The FDTD diakoptics method," *1993 IEEE MTT-S International Microwave Symposium Digest*. pp. 1435-1438 vol.3 (1993), doi: 10.1109/MWSYM.1993.276728
- [11] Alkhraijah, M., Sigler, D. & Molzahn, D. "A Decomposition Method for Solving Sensitivity-Based Distributed Optimal Power Flow," (2025), <https://arxiv.org/abs/2512.22419>, doi: 10.48550/arXiv.2512.22419
- [12] L. Yin, H. Hu, and C. Li, "Research on Circuit Partitioning Algorithm Based on Partition Connectivity Clustering and Tabu Search," *Technologies*, vol. 13, no. 2, 2025, doi: 10.3390/technologies13020081
- [13] Alt, S., Marek-Sadowska, M. & Wang, "L. Circuit Partitioning for Behavioral Full Chip Simulation Modeling of Analog and Mixed Signal Circuits," *International Journal Of Modeling And Optimization*. 4 pp. 74-80 (2014,1), doi: 10.7763/IJMO.2014.V4.350
- [14] M. Iordache, L. DUMITRIU, and D. Niculae, "On the sensitivity analysis of analog circuits," *Annals of the University of Craiova, Electrical Engineering Series*, vol. 32, Jan. 2008.
- [15] E. Ryan, O. Wild, A. Voulgarakis, and L. Lee, "Fast sensitivity analysis methods for computationally expensive models with multi-dimensional output," *Geoscientific Model Development*, vol. 11, pp. 3131–3146, Aug. 2018, doi: 10.5194/gmd-11-3131-2018.
- [16] I. Kovacs, M. D. Topa, A. Buzo, M. Rafail, and G. Pelz, "Comparison of sensitivity analysis methods in high-dimensional verification spaces", 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:31574889>
- [17] D. Francom and A. Nachtsheim, "A review and comparison of different sensitivity analysis techniques in practice", 2025. doi: 10.48550/arXiv.2506.11471
- [18] Banos, R., Gil, C., Montoya, M. & Ortega, J. "A parallel evolutionary algorithm for circuit partitioning," *Eleventh Euromicro Conference On Parallel, Distributed And Network-Based Processing, 2003. Proceedings*. pp. 365-371 (2003), doi: 10.1109/EMPDP.2003.1183612
- [19] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge, UK: Cambridge University Press, 2004. [Online]. Available: https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf
- [20] P. Weisz and Gy. Orosz, "Framework for Automated Worst-Case Analysis," in *Proceedings of the 32nd Minisymposium*, Budapest, Hungary, 2025, pp. 78-83., doi: 10.3311/MINISY2025-005