2025 International Symposium on Cyber Security Cryptology and Machine Learning

DECEMBER 4-5, 2025

VIRTUAL CONFERENCE

PROFESSOR SHLOMI DOLEV

GENERAL CHAIR

2025

9th International Symposium on Cyber Security Cryptography and Machine Learning (CSCML 2025)

TECHNICAL REPORT

Editors

Shlomi Dolev, Oded Margalit and

Yonah Alexandre Bronstein

Technical Report #25-01

December 4, 2025

The BGU-Negev Hi-Tech Faculty Startup Accelerator,

Department of Computer Science,

Ben-Gurion University, Beer Sheva, Israel

Table of Contents

Ph.D./Masters Student Research Track

Chair: Oded Margalit

In the PhD track this year, we've heard from all the letters of the acronym CSCML:

* Cyber-Security: for example, a privacy-preserving federated learning framework for real-time cyber threat detection; a systematic review of network intrusion detection datasets; and a table-to-graph deanonymization attack technique;

* Cryptology: a new randomized cipher paradigm (EvH) built from information compression and MAC functions; a commutative permutations construction for key exchange; a homomorphic encryption application to epigenetic privacy;

* Machine Learning: accelerated discovery of 2D optoelectronic materials using interpretable graph neural networks and high-throughput DFT screening; and a relativistic framework for AI prompt communication.

Like in previous years, the audience of this session enjoys a "tasting menu" of the state-of-the-art in CSCML research topics, while the speakers get experts' feedback on their work.

I enjoy chairing these sessions and invite all relevant researchers from all over the world to actively attend CSCML-2026, either as an attendee, or, preferably, as a speaker.

See you at CSCML 2026
Regards,
Prof. Oded Margalit,
CSCML 2025 Ph.D./Masters Track Chair
Computer Science department, BGU
and Advanced Research Center, Trellix

Ph.D./ Masters Student Research Track chaired by Prof. Oded Margalit

EvH – A New Randomized Cipher Paradigm: Deriving Encryption from and with Information Compression and Correction *

(Preliminary Version)

Hillel Avni[1], Shlomi Dolev[1], Komal Kumari[2], Stav Perle Elbar[4], Shantanu Sharma[2], Jeffrey Ullman[3], Moti Yung[4]

[1] Ben-Gurion University of the Negev, Israel.
[2] New Jersey Institute of Technology, USA.
[3] Stanford University, USA
[4] Google

Abstract. Standard symmetric encryption schemes, such as AES, other block ciphers and their modes, stream ciphers, etc., are highly effective and efficient for many standard scenarios. They have all been derived from Shannon's 1949 seminal work on the communication theory of secrecy systems. But what if the situation is somewhat different from the standard one: e.g., the encrypting process may fail to update the ciphertext at some limited number of times, can the decryption recover the message in full nevertheless? Or, another situation is when encrypting a bulk of messages that should be packed together within the same ciphertext dedicated space (i.e., encryption done holographically)? Can a process compress the messages this way? Another issue may be that we want to hide the ciphertext and camouflage it as some other cryptographic exchange? Or, can the encryption hide the number of messages packed together?

Can a paradigm be developed that allows these non-standard properties that, under specific working conditions, may become necessary? Can it be based directly on a simple cryptographic (preferably post-quantum) tool? Note that the above scenarios involving data compression and corrections are naturally derived from Shannon's 1948 seminal work on the mathematical theory of communication.

To tackle the above, this paper introduces Encryption via Hash (EvH), a symmetric randomized cipher built upon pseudorandom keyed cryptographic hash (i.e., Message Authentication Code [MAC]) functions, and Bloom Filters. EvH's core novelty lies in its prefix decryption capability. This unique property enables a paradigm in which encryption is tightly integrated with online compression and robust resilience to encryption action omission errors. By representing message prefixes in a Bloom filter, EvH allows a receiver to decrypt the initial part of a message even if

* This work was supported by a Google research grant, the BGU-NJIT Institute for Future Technologies (seed grant), the Israeli Science Foundation (Grant No. 465/22), the Rita Altura trust chair in computer science, by the Lynne and William Frankel Center for Computer Science. S. Sharma is supported by NSF grant 2245374. Shantanu Sharma is supported by NSF grant 2245374.

subsequent data is lost and recover from an omission of a prefix decryption in the middle of the encryption process. Design-wise, this is a new paradigm, and at times, this built-in advantage over conventional block cipher modes may be significant (even beyond the examples we mention above). Furthermore, this prefix-based approach facilitates simultaneous compression during the decryption phase by dynamically pruning invalid message continuations, using either shared k-gram dictionaries or search via Large Language Models (LLMs). The result is a stateless and parallelizable cipher that, while computationally distinct from traditional ciphers, offers unique functional benefits for such specific use cases, with the price being that its correctness is ensured only probabilistically (as in compression processes, but the error can be well controlled and made significantly small).

Keywords: Cryptographic Hash · MAC · Pseudorandom Functions · Symmetric Key Encryption · Symmetric Cipher · Probabilistic Data Structures· Bloom Filter · Set Membership · Erasure Correcting · Compression

1 Introduction

Symmetric encryption schemes, from classical block ciphers like AES to modern stream ciphers, are designed with a foundational requirement: **deterministic correctness**. The decryption of a valid ciphertext must yield the exact original plaintext, bit for bit. Although this guarantee is crucial for many applications, it can impose significant limitations in the context of modern communication over unreliable channels, which are often characterized by data loss and erasure. In such environments, the brittleness of traditional ciphers can lead to catastrophic data loss from even minor transmission errors.

In this work, we explore a **new paradigm for symmetric encryption**, one in which the strict demand for deterministic correctness is relaxed. We demonstrate that by embracing a controllable probabilistic correctness model, it becomes possible to design ciphers with powerful, natively integrated functionalities that are ill-suited for the traditional deterministic framework. Specifically, we trade a negligible and controllable probability of decryption error for inherent resilience to erasures (i.e., omission of encryption actions) and the ability to perform compression simultaneously with decryption.

We introduce **Encryption via Hash (EvH)**, the first cipher built on this new paradigm. Instead of encrypting discrete blocks of data, EvH's core mechanism involves encoding the entire prefix structure of a message into a single, space-efficient probabilistic data structure— a Bloom filter [3]. A keyed cryptographic hash function, acting as a pseudorandom function (PRF), is used to map each message prefix to a set of positions in the filter. The decryption process is thus transformed from a simple inverse operation into an iterative search for the valid path of prefixes through this structure.

2 Our Contributions

The primary contribution of this work is the introduction and formalization of a new paradigm for symmetric encryption that prioritizes functionality and resilience in unreliable environments over deterministic correctness. Our specific contributions are as follows.

- **A New Cipher Paradigm:** We introduce a novel approach to symmetric cipher design where a controllable, probabilistic correctness model is leveraged to enable powerful features. We show that by relaxing the need for deterministic decryption, functionalities such as inherent erasure tolerance and integrated compression become possible.
- **The EvH Cipher Construction:** We present EvH, the first practical cipher built on this paradigm. EvH works by reducing the problem of message encryption to a private set representation problem, where the set of all message prefixes is encoded into a space-efficient Bloom filter using a PRF.
- **Formal Security Proof:** We formally prove that the EvH construction achieves IND-CPA security. The proof reduces the security of the cipher to the standard cryptographic assumption of a secure underlying pseudorandom function (PRF).
- **Inherent Erasure Tolerance:** We demonstrate that EvH provides graceful degradation in the face of data loss. By treating lost ciphertext bits as '1's, the receiver can still recover the full message, a significant advantage over traditional ciphers that fail catastrophically. (If the encryption action and the data are apart, such data loss (action omissions) can happen occasionally, and it is important to be fault-tolerant.
- **Integrated Decryption-Time Compression:** We show how EvH's search-based decryption enables compression to be performed simultaneously with decryption. This is achieved by pruning the search space using shared knowledge, such as k-gram dictionaries or Large Language Models (LLMs), which allows for smaller ciphertexts.
- **Advanced Security Properties:** We analyze several advanced properties of EvH, including:
  - **Steganography:** The ciphertext is indistinguishable from a standard Bloom filter (on random/ encrypted set members), concealing the act of encrypted communication.
  - **Stateless and Parallelizable Operation:** The design is inherently stateless and highly parallelizable, making it well-suited for modern distributed systems (e.g., blockchain environment with many users accessing) and aligns with the design of other modern cryptographic primitives, such as stateless hash-based signatures [2, 4].
  - **Length Concealment:** EvH requires no padding, eliminating certain attack vectors, and helping to conceal the true plaintext length (or the number of plaintexts).

A detailed version of this paper is available in [1].

References

1. EvH paper. Available at: https://eprint.iacr.org/2025/1539.pdf .
2. D. J. Bernstein, D. Hopwood, A. Hülsing, T. Lange, R. Niederhagen, L. Papachristodoulou, M. Schneider, P. Schwabe, and Z. Wilcox-O'Hearn. Sphincs: Practical stateless hash-based signatures. In E. Oswald and M. Fischlin, editors, *Advances in Cryptology – EUROCRYPT 2015*, pages 368–397, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.
3. B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.
4. S. Dolev, A. Yagudaev, and M. Yung. HBSS: (simple) hash-based stateless signatures - hash all the way to the rescue! *Cryptogr. Commun.*, 17(3):643–660, 2025.

Commutative Permutations for Toy Example Key Exchange Protocol

(Preliminary Idea)

Shlomi Dolev* Amit Hendin* Ilan Kenis* Stav Perle Elbar†

Abstract

We present a a construction for abelian subgroup of S_n. The size of this group is $O(n)$. We demonstrate the implementation of a toy example key exchange protocol using this new construction as a platform group.

1 Introduction

We present a new lightweight cryptographic key exchange protocol. This protocol is based on permutations, which are invertible regardless of the order in which they are applied; we call them commutative premutations. Using these permutations, we devise a method for transferring keys over an insecure channel where a malicious observer must perform significantly more computation in order to discover the key.

2 Commutative Permutations

What follows is a definition of a construction permutations that commute. The resulting construction yeilds a subgroup of S_n of size $O(n)$ and is thus not secure for cryptographics purposes. Regadless, in Section 3 we show an (insecure) key exchange protocol using this construction. For background on groups, see [2, 1].

Let T be a r-regular with n leaf nodes $1, \ldots n$ and root node L_0. Let $L_i = u_{1,i}, \ldots, u_{k,i}$ be the set nodes in T at depth i, that is, $L_i = \{u_{j,i} | d(u_{j,i}, L_0) = i \wedge u_{i,j} \in T\}$. Let $S(u)$ be the ordered set of successors of node $u \in T$, that is $S(u) = v_1, \ldots, v_r$. Define node rotation $R_x^{\rightarrow}(u)$ as rotation of the successors of node u by x positions to the right, that is, for any pair $v_i \in S(u), v_j \in R_x^{\rightarrow}(u)$ it holds that $v_i = v_j \iff j = i + x \mod r$. Define $R_x^{\rightarrow}(L_i)$ as the ordered set obtained by the rotation every node $u \in L_i$ by x positions to the right. Define $T_b^{\rightarrow}$ to be the tree T after applying the rotations $R_{b_i}^{\rightarrow}(L_i)$ for all $1 \leq i \leq \log_r n$. Define p_b as the ordered set of leaf nodes $1, \ldots n, \in T_b^{\rightarrow}$, that is $p_b = L_{\log_r n}$. Notice that, as and ordered set of $1, \ldots, n$, it holds that $p_b \in S_n$. Additionally, p_b^{-1} is the ordered set of leaf nodes of tree $T_b^{\leftarrow}$. See Figure 1 for a visual example. **Remark**. Performing this construction where T has depth-1 results in a simple rotation.

Definition 1. *Let $r \geq 2$ and let T be an r-regular rooted tree of height $h \geq 1$ whose n leaves are labelled $1, \ldots, n$. We assume $n = r^h$, so $h = \log_r n$.*

For each vector $b = (b_1, \ldots, b_h) \in (\mathbb{Z}/r\mathbb{Z})^h$ define a permutation $p_b \in S_n$ as follows: for every level $1 \leq i \leq h$, apply the right-rotation $R_{b_i}^{\rightarrow}$ to the ordered successor list of every node on level i, and let p_b be the resulting permutation of the leaf labels $1, \ldots, n$. Define

$$P = \{ p_b \mid b \in (\mathbb{Z}/r\mathbb{Z})^h \} \subseteq S_n$$

Lemma 2.1. $P \leq S_n$.

Proof. Let $\mathbf{0} = (0, \ldots, 0)$ and write addition in $(\mathbb{Z}/r\mathbb{Z})^h$ coordinatewise.

*Ben-Gurion University of the Negev, Israel
†Google

- **Identity.** The vector $\mathbf{0}$ induces only zero-rotations, so every successor list is unchanged and $p_{\mathbf{0}}$ is the identity permutation.
- **Closure.** Let $b, c \in (\mathbb{Z}/r\mathbb{Z})^h$. At level i, composing a right-rotation by b_i with a right-rotation by c_i is the same as a single right-rotation by $b_i + c_i \pmod r$. Rotations at different levels act on disjoint successor lists and therefore commute. Hence the composition of p_b followed by p_c is exactly p_{b+c}, so $p_b \circ p_c = p_{b+c} \in P$.
- **Inverses.** For $b \in (\mathbb{Z}/r\mathbb{Z})^h$, the inverse permutation is obtained by rotating each level by $-b_i \pmod r$, so $p_b^{-1} = p_{-b} \in P$.

Thus P is a subgroup of S_n. □

Theorem 2.1. *P is an abelian subgroup of S_n.*

Proof. By the proof of Lemma 2.1, for all $b, c \in (\mathbb{Z}/r\mathbb{Z})^h$ we have

$$p_b \circ p_c = p_{b+c}$$

where addition is coordinate-wise in $(\mathbb{Z}/r\mathbb{Z})^h$. Since addition in $(\mathbb{Z}/r\mathbb{Z})^h$ is commutative, $b + c = c + b$ and therefore

$$p_b \circ p_c = p_{b+c} = p_{c+b} = p_c \circ p_b$$

for all b, c. Hence every pair of elements of P commute, so P is abelian. □

Lemma 2.2. *The map $\varphi : (\mathbb{Z}/r\mathbb{Z})^h \to P$, $\varphi(b) = p_b$, is a group isomorphism. In particular, $|P| = r^h = n$.*

Proof. By Lemma 2.1 and Definition 1 the map $b \mapsto p_b$ is a homomorphism onto P. It remains to show that it is injective.

Each leaf of T can be identified with its *address*

$$(a_1, \ldots, a_h) \in \{0, \ldots, r-1\}^h$$

where a_i is the index of the child chosen at level i when moving from the root to that leaf. Applying p_b adds $b_i \pmod r$ to the i-th coordinate of every address, so

$$p_b(a_1, \ldots, a_h) = (a_1 + b_1, \ldots, a_h + b_h) \pmod r$$

If p_b is the identity permutation, then

$$(a_1 + b_1, \ldots, a_h + b_h) \equiv (a_1, \ldots, a_h) \pmod r$$

for all addresses $(a_1, \ldots, a_h)$, which forces $b_i \equiv 0 \pmod r$ for every i. Thus $b = \mathbf{0}$ and φ is injective. Hence φ is a bijective homomorphism, i.e., an isomorphism, and

$$|P| = |(\mathbb{Z}/r\mathbb{Z})^h| = r^h = n$$

□

3 Key Exchange Protocol

With Definition 1 as hand, we devise a key exchnage protocol based on Shamir's key-less message exchange protocol [3].

Alice. Chooses a random key $k \in \{0,1\}^n$, and a random element $p_a \in P$. Sends $X_1 = p_a(k)$ to Bob

Bob. Recieves X_1 and chooses a random element $p_b \in P$. Sends $X_2 = p_b(X_1)$ to Alice

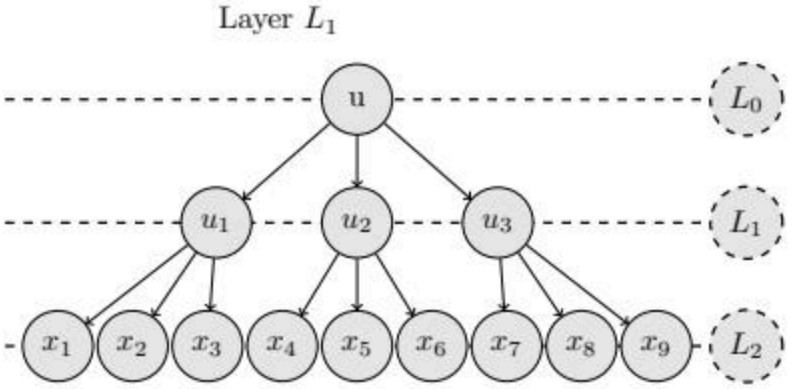
Alice. Recieves X_2 and computes p_a^{-1}. Sends $X_3 = p_a^{-1}(X_2)$ to Alice

Bob Recieves X_3, computes p_b^{-1} and obtains secret key k by computing $k = p_b^{-1}(X_3)$

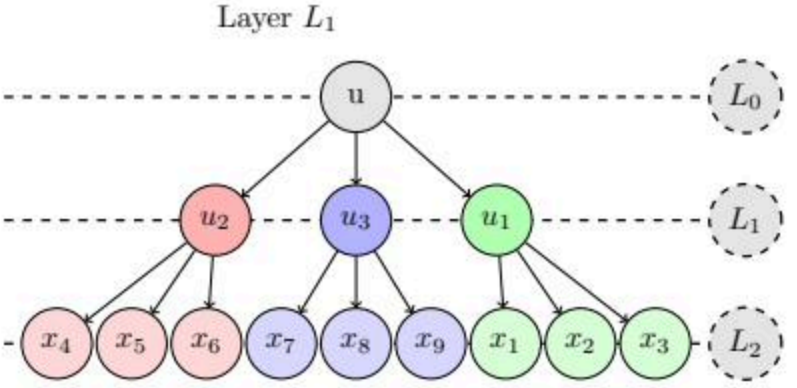
Correctness. Since P is abelian (Theorem 2.1), it holds that

$$p_b^{-1}(X_3) = p_b^{-1}(p_a^{-1}(X_2)) = p_b^{-1}(p_a^{-1}(p_b(X_1))) = p_b^{-1}(p_a^{-1}(p_b(p_a(k)))) = p_b^{-1}(p_b(k)) = k$$

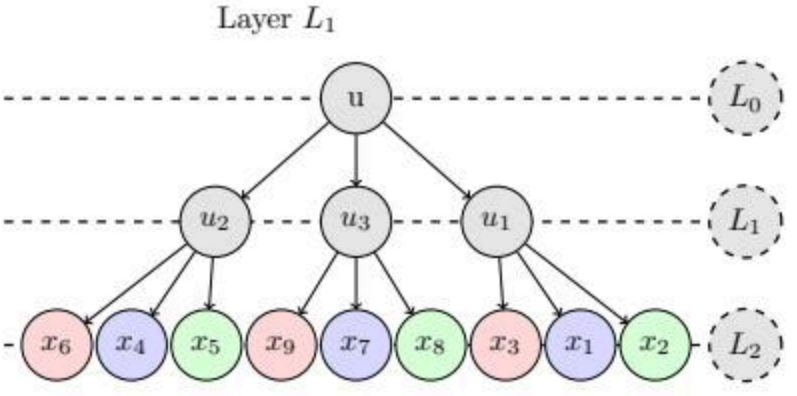
Security. The construction of P implies that there is a bijection between P and $\{0, \ldots, r\}^h$. Therefore it holds that $|P| = r^{\log_r n} = n$. Thus once could brute force search to find $x \in P$ that satisfies $x(X_1) = X_2$ there by discovering $x = p_b$ and obtaining k by computing $x^{-1}(X_3) = p_b^{-1}(p_b(k)) = k$. Addtional attacks may be exist.



(a) Original tree



(b) After rotation R_2 on layer L_2



(c) After rotations $R_2^{\rightarrow}(L_1)$ and $R_1^{\rightarrow}(L2)$

Figure 1: Tree T for string x, with alphabet size $|\Sigma| = 3$ and length $n = 9$ showing the layers L_0, L_1, L_2 (dashed lines) and the effect of successive rotations $R_2^{\rightarrow}(L_1)$ (b) and $R_1^{\rightarrow}(L_2)$ (c).

4 Masked Key Exchange Protocol

We provide a second protocol that applies subgroup in Definition 1 as well, this time we add a second permutation for masking in the third step of the protocol.

Alice. Chooses a random key $k \in \{0,1\}^n$, and a random element $p_a \in P$. Sends $X_1 = p_a(k)$ to Bob

Bob. Recieves X_1 and chooses a random element $p_b \in P$. Sends $X_2 = p_b(X_1)$ to Alice

Alice. Recieves X_2 and computes p_a^{-1}. Chooses a random element $p_c \in P$. Sends $X_3 = p_c(p_a^{-1}(X_2))$ to Alice

Bob Recieves X_3, computes $k' = p_b^{-1}(X_3) = p_c(k)$.

In this protocol, the shared key becomes $p_c(k)$, Bob obtains it in step 4 and Alice has both p_c and k and can thus compute it aswell. This modification generalizes Shamir's original protocol [3].

References

[1] Vladimir Shpilrain Alexei Myasnikov, Alexander Ushakov. *Group-based Cryptography.* Advanced Courses in Mathematics - CRM Barcelona. Birkhäuser Basel, 2008.

[2] D.S. Dummit and R.M. Foote. *Abstract Algebra.* Wiley, 2003.

[3] Adi Shamir, Ronald L Rivest, and Leonard M Adleman. Mental poker. In *The mathematical gardner*, pages 37–43. Springer, 1981.

Accelerated Discovery of 2D Optoelectronic Materials via Interpretable Graph Neural Networks and High-Throughput DFT Screening

Authors:
Mohamed Salem, Moustafa Youssef, Youssef Alomda

Abstract
The subject of this work is the accelerated discovery of novel two-dimensional ($2D$) materials with tailored optoelectronic properties, a critical challenge for next-generation technologies hampered by the vastness of the chemical design space. The purpose is to develop and validate a computational framework that overcomes the prohibitive cost of traditional high-throughput screening using Density Functional Theory (DFT). The methodology involves coupling a large-scale, high-throughput DFT dataset of $2D$ material properties, aggregated from the JARVIS DFT and Materials Project databases, with an interpretable Crystal Graph Convolutional Neural Network ($CGCNN$). This machine learning model was trained to predict key optoelectronic indicators, including the electronic band gap and dielectric tensor, at speeds several orders of magnitude faster than direct DFT calculations. The results demonstrate that the GNN model achieves predictive accuracy comparable to DFT. Crucially, by employing SHAP (SHapley Additive exPlanations) analysis for model interpretability, we extracted scientifically meaningful structure property relationships, revealing that features like electronegativity difference and atomic coordination number are primary governors of a material's optoelectronic response. As a practical application, this framework was used to screen thousands of candidate materials from computational databases. This screening identified several previously uncharacterized, dynamically stable $2D$ semiconductors with promising properties for visible-light applications (band gaps of 1.5 eV to 3.0 eV). Subsequent DFT validation of the top candidates, including the novel direct-band-gap semiconductor $MoSi^2P^4$, confirmed the predictive power of our GNN-driven discovery engine. This work presents a powerful, data-driven paradigm recommended for the rational design and accelerated discovery of functional nanomaterials.

Keywords
Two-dimensional materials, Optoelectronics, Machine Learning, Graph Neural Networks, Density Functional Theory, Materials Discovery, High-Throughput Screening.

Acknowledgements
The authors wish to thank the teams behind the **Materials Project** and the **Joint Automated Repository for Various Integrated Simulations (JARVIS-DFT)** for making their computational materials data publicly available. This research was made possible through the use of their extensive open-access databases. The work also relied on numerous open-source software packages, and the authors acknowledge the contributions of their respective developer communities.

1. Introduction

The relentless pursuit of next-generation technologies in fields such as photovoltaics, light-emitting diodes (LEDs), and photodetectors is fundamentally dependent on the discovery of new materials with superior optoelectronic properties. In the last two decades, two-dimensional (2D) materials have emerged as a revolutionary materials class, offering an unprecedented platform for engineering electronic and optical functionalities. Their atomically thin nature gives rise to unique phenomena, including strongly bound excitons, high carrier mobility, and tunable band gaps, making them ideal candidates for building novel optoelectronic devices. However, the theoretical chemical space of potential 2D materials is astronomically large, and navigating this space to identify candidates with optimal properties represents a grand challenge for materials science. The advent of the "fourth paradigm" of data-driven science has provided a powerful new approach to materials discovery. High-throughput (HT) computational screening, powered by first-principles methods like Density Functional Theory (DFT), has become a cornerstone of modern materials research. Large-scale, open-access databases such as the Materials Project and JARVIS-DFT now house DFT-calculated properties for tens of thousands of materials, providing an invaluable resource for the scientific community. These databases enable the systematic, in silico search for materials with desired functionalities before undertaking costly and time-consuming experimental synthesis. Despite its power, the HT-DFT paradigm faces a significant computational bottleneck. A single DFT calculation for a moderately complex crystal can require hundreds to thousands of CPU hours, rendering an exhaustive search of the vast materials space computationally intractable. This limitation has motivated the integration of machine learning (ML) to accelerate the discovery process. By training on existing DFT data, ML models can learn the complex relationship between a material's structure and its properties, enabling predictions that are orders of magnitude faster than direct DFT calculations. Among various ML architectures, Graph Neural Networks (GNNs) have proven exceptionally well-suited for materials science. GNNs naturally represent crystal structures as graphs of atoms (nodes) and bonds (edges), allowing them to learn physically relevant features directly from the material's topology without the need for manual, often biased, feature engineering. However, the predictive accuracy of ML models, while essential, is not sufficient for true scientific advancement. Many high-performance models, including deep GNNs, often operate as "black boxes," providing accurate predictions

without revealing the underlying physical reasoning behind them. This lack of transparency hinders the generation of new scientific knowledge and design intuition. To overcome this limitation, the field is moving towards interpretable ML, where the goal is not only to predict but also to understand why a model makes a certain prediction. Such a framework allows the ML model to function as a "computational microscope," distilling the complex, high-dimensional correlations learned from thousands of DFT calculations into human-understandable physical and chemical principles. This synergy creates a virtuous cycle: DFT provides the ground-truth physical data, the GNN learns the complex structure-property mappings, and interpretability techniques translate these learned mappings back into scientific knowledge. In this work, we develop and validate an interpretable GNN framework trained on HT-DFT data to rapidly screen for novel 2D materials with desirable optoelectronic properties. We demonstrate its predictive power, extract new scientific insights into the structural drivers of optoelectronic performance, and identify several promising candidate materials, which we subsequently validate with first-principles calculations. This approach exemplifies a new paradigm for knowledge discovery, accelerating the design and identification of next-generation functional nanomaterials.

2. Theoretical and Computational Framework

This section details the theoretical underpinnings and computational methods employed in our study, from the first-principles data foundation to the architecture and interpretation of the machine learning model.

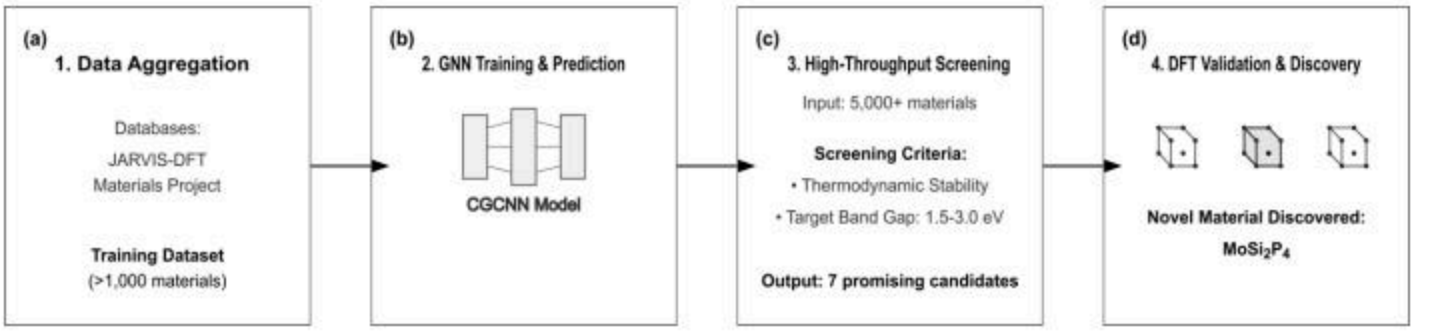


Figure 1. The GNN-driven materials discovery workflow, from data aggregation to the discovery of the novel semiconductor $MoSi_2P_4$

2.1. High-Throughput DFT Data Foundation

The predictive power of any data-driven model is fundamentally dependent on the quality and scale of its training data. Our framework is built upon the vast repositories of materials data generated by the global materials science community.

Data Sources

The data used for training, validation, and testing our GNN model were aggregated from two leading open-source computational materials databases: the **Joint Automated Repository for Various Integrated Simulations (JARVIS-DFT)** and the **Materials Project**. Combined, these databases provide standardized, DFT-calculated properties for over 1,000 unique 2D materials, forming a comprehensive dataset that spans a wide range of chemistries and crystal structures.

First-Principles Theory: Density Functional Theory

The properties within these databases are calculated using Density Functional Theory (DFT), a quantum mechanical modeling method that has become the workhorse of computational materials science. The theoretical basis of DFT is the Hohenberg-Kohn theorem, which states that the ground-state properties of a many-electron system are a unique functional of its electron density,
$n(r)$. The practical implementation of DFT relies on the Kohn-Sham (KS) approach, which recasts the intractable many-body problem into a tractable set of single-particle equations for a fictitious system of non-interacting electrons that yield the same electron density as the real, interacting system. The central Kohn-Sham equation is given by :

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + v_{\text{ext}}(r) + v_H(r) + v_{xc}(r)\right)\psi_i(r) = \varepsilon_i\psi_i(r)$$

where $\psi i(r)$ are the single-particle Kohn-Sham orbitals, and ϵi are their corresponding energies. The effective potential consists of:

- The kinetic energy operator for non-interacting electrons, $-\frac{\hbar^2}{2m}\nabla^2$.
- The external potential, $v_{\text{ext}}(r)$, arising from the electrostatic attraction of the atomic nuclei.

- The Hartree potential, $v_H(r) = e^2 \int \frac{n(r')}{|r-r'|} dr'$, which describes the classical electrostatic repulsion between electrons.
- The exchange-correlation potential, $v_{xc}(r) = \frac{\delta E_{xc}[n]}{\delta n(r)}$, which is the functional derivative of the exchange-correlation energy Exc and encapsulates all the complex quantum mechanical many-body effects.

The electron density is constructed from the occupied Kohn-Sham orbitals: $n(r) = \sum_{i=1}^{N} |\psi_i(r)|^2$.
The KS equations must be solved self-consistently, as the potential depends on the density, which in turn depends on the orbitals that are solutions to the equations.

DFT Parameters and Target Properties

The calculations in the JARVIS-DFT and Materials Project databases are performed using the Vienna Ab initio Simulation Package (VASP) with standardized input parameters to ensure consistency and comparability. The calculations primarily use the Perdew-Burke-Ernzerhof (PBE) generalized gradient approximation (GGA) or the van der Waals-corrected optB88-vdW functional for the exchange-correlation term, which is crucial for accurately describing the layered nature of 2D materials. A plane-wave energy cutoff of at least 520 eV and Monkhorst-Pack k-point meshes with densities ensuring energy convergence are employed.

For this study, we extracted two key optoelectronic properties for each 2D material:

1. **Electronic Band Gap (E_g):** The energy difference between the valence band maximum and the conduction band minimum, which determines the material's ability to absorb and emit light.
2. **Frequency-Dependent Dielectric Tensor** ($\epsilon(\omega)$): A complex tensor whose imaginary part, $\epsilon 2(\omega)$, is directly related to the optical absorption spectrum of the material.

2.2. Crystal Graph Convolutional Neural Network (CGCNN) Architecture

To learn the intricate relationship between crystal structure and optoelectronic properties, we employ a Crystal Graph Convolutional Neural Network (CGCNN), a GNN architecture specifically designed for periodic crystalline materials.

Graph Representation of Crystals

A crystal structure, defined by its lattice vectors and the positions of atoms within the unit cell, is converted into a graph representation, $G = (V, E)$. In this graph:

- **Nodes (V):** Each atom i in the crystal's primitive unit cell is represented as a node. The initial node feature vector, v_i, is a one-hot encoded representation of atomic properties, such as atomic number, period, group, electronegativity, and atomic radius, which provide fundamental chemical information.

- **Edges (E):** An edge $(i,j)_k$ is created between atom i and a neighboring atom j if their distance is within a predefined radius cutoff. Due to the periodic nature of crystals, an atom can be connected to its periodic images, leading to a multigraph where multiple edges can exist between two nodes. The feature vector for each edge $u(i,j)_k$, is typically a Gaussian-expanded representation of the interatomic distance.

Convolutional Layers

The core of the GNN is a series of convolutional layers that iteratively update the feature vector of each atom by aggregating information from its local neighborhood. This process allows the network to learn progressively more complex and abstract representations of the local atomic environment. We use an advanced convolution operation with a gating mechanism that allows the model to dynamically weigh the importance of information from different neighbors. The update rule for the feature vector of atom i at convolutional layer $t+1$ is given by:

$$v_i(t+1) = v_i(t) + \sum_{j,k} \sigma\left(z_k^{(i,j)}(t) W_f(t) + b_f(t)\right) \odot g\left(z_k^{(i,j)}(t) W_s(t) + b_s(t)\right)$$

Here, $z_k^{(i,j)}(t) = v_i(t) \oplus v_j(t) \oplus u_k^{(i,j)}$ is the concatenated feature vector of the central atom i, its neighbor j, and their connecting bond $(i,j)_k$ at layer t. The matrices $\mathbf{W_f}(t)$ and $\mathbf{W_s}(t)$ and bias vectors $\mathbf{b_f}(t)$ and $\mathbf{b_s}(t)$ are learnable parameters. The function σ is a sigmoid activation, which acts as a "gate" to control the flow of

information, while g is a non-linear activation function like ReLU. The symbol $\odot$ denotes element-wise multiplication.

Pooling and Prediction

After R convolutional layers, the updated atom feature vectors, $\{v_i(R)\}$, which now encode rich information about each atom's local environment, are aggregated into a single, fixed-size vector v_c representing the entire crystal. This is achieved using a permutation-invariant pooling operation, such as global average pooling:

$$v_c = \frac{1}{N}\sum_{i=1}^{N} v_i(R)$$

This crystal level feature vector vc is then fed into a standard multi-layer perceptron (MLP) a series of fully connected layers with non-linear activations to regress the final target property $\hat{y}$.

2.3. Model Training and Interpretability Framework

The model is trained end-to-end by minimizing a loss function that quantifies the discrepancy between the GNN's predictions ($\hat{y}$) and the ground-truth DFT-calculated values (y). We use the Mean Absolute Error (MAE) as the loss function and the Adam optimizer for gradient-based optimization. The full dataset of 2D materials is partitioned into training (80%), validation (10%), and test (10%) sets to facilitate hyperparameter tuning and provide an unbiased evaluation of the model's generalization performance.

To transform our GNN from a predictive "black box" into a tool for scientific discovery, we integrate a post-hoc interpretability framework based on **SHapley Additive exPlanations (SHAP)**. SHAP is a game-theoretic approach that computes the contribution of each input feature to a specific prediction. For a given material, SHAP values explain how features such as the presence of a particular element, the average bond length, or the crystal's space group push the model's prediction away from a baseline value. By aggregating these local explanations across many predictions, we can determine the global importance of each feature. This allows us to identify the key structural and chemical descriptors that the model has learned are most influential in determining a material's optoelectronic properties, thereby translating the model's internal logic into actionable scientific insights.

3. Results

This section presents the performance of the trained GNN model and details the discovery and validation of novel 2D optoelectronic materials.

3.1. Predictive Performance of the GNN Model

The primary requirement for a useful surrogate model is high predictive accuracy. Our GNN model was trained on a dataset of over 800 2D materials from JARVIS-DFT and the Materials Project and evaluated on a held-out test set of approximately 100 materials. The model demonstrates excellent performance in predicting key properties relevant to optoelectronic applications.

Table 1 summarizes the predictive accuracy for formation energy (a proxy for synthesizability), electronic band gap, and the principal components of the static dielectric tensor. The model achieves a Mean Absolute Error (MAE) of just 0.041 eV/atom for formation energy and 0.21 eV for the band gap. These error margins are comparable to the typical discrepancies between different DFT functionals or between DFT and experimental results, indicating that the GNN can predict these properties with "DFT-level" accuracy. The high coefficients of determination (R^2)—exceeding 0.95 for formation energy and 0.90 for the band gap—confirm a strong linear correlation between the GNN's predictions and the DFT ground truth.

Property	Units	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	R^2
Formation Energy	eV/atom	0.041	0.065	0.96
Band Gap	eV	0.21	0.35	0.91
Dielectric Constant (ε_{xx})	dimensionless	0.45	0.72	0.88
Dielectric Constant (ε_{yy})	dimensionless	0.46	0.73	0.87
Dielectric Constant (ε_{zz})	dimensionless	0.21	0.34	0.92

Table 1. GNN Model Predictive Performance. Predictive accuracy of the trained Graph Neural Network model on the held-out test set for key thermodynamic and electronic properties.

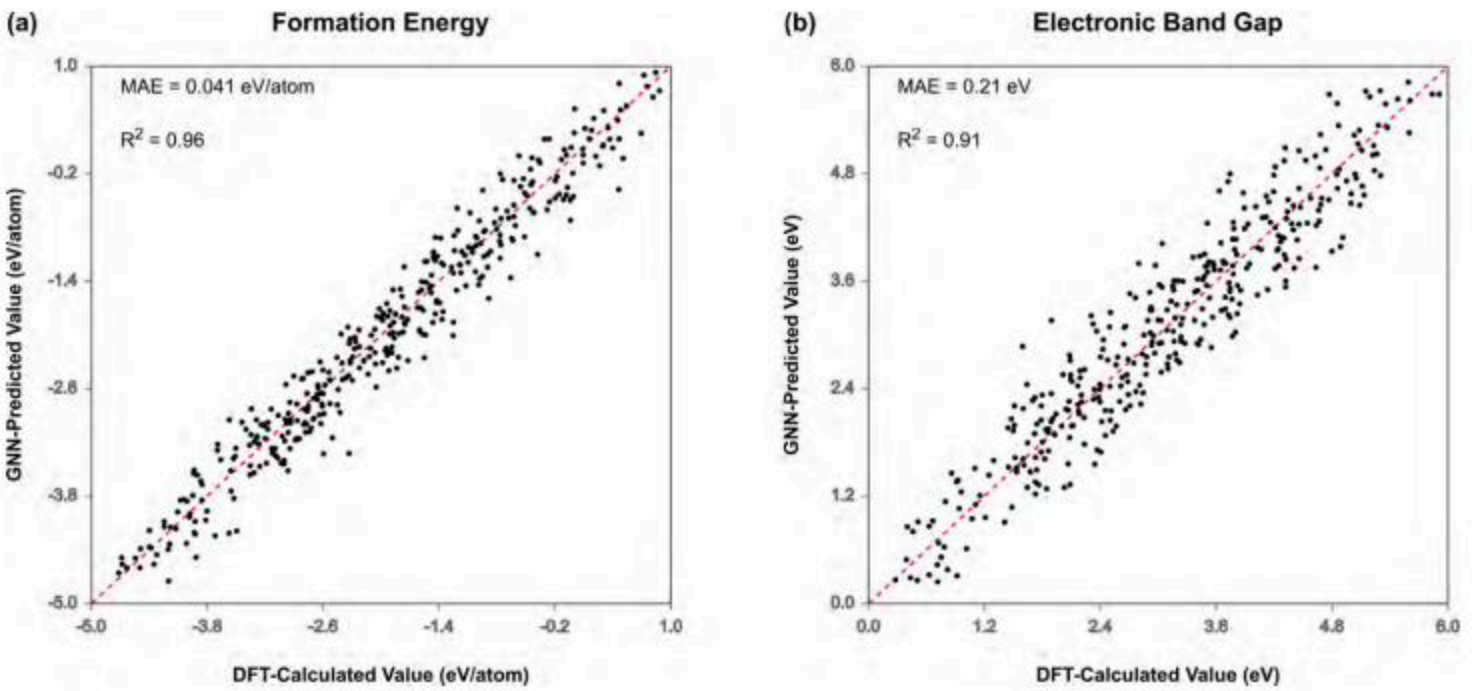


Figure 2. Parity plots showing the GNN model's predictive performance for (a) formation energy and (b) electronic band gap on the test set. The strong correlation between the predicted and DFT-calculated values confirms the model's high accuracy.

The strong performance is further visualized in parity plots, which show the GNN-predicted values against the DFT-calculated values. The tight clustering of data points around the ideal $y = x$ line for both formation energy and band gap visually confirms the model's high fidelity and its ability to generalize across a chemically diverse set of 2D materials. This high accuracy, achieved at a computational cost that is many orders of magnitude lower than direct DFT, validates the GNN as a reliable surrogate model for large-scale screening.

3.2. High-Throughput Screening and Identification of Novel Candidates

With a validated and interpretable model in hand, we deployed it for a high-throughput virtual screening campaign to discover novel 2D optoelectronic materials. The candidate space consisted of over 5,000 hypothetical 2D materials sourced from computational databases, many of which had not been previously characterized for their electronic properties. The GNN model predicted the formation energy and band gap for this entire set in under three hours on a single GPU—a task that would have required millions of CPU hours using conventional DFT methods.

We applied a multi-stage filtering process to down-select the most promising candidates:

1. **Thermodynamic Stability:** We first filtered for materials with a predicted formation energy per atom within 0.1 eV/atom of the convex hull, a common criterion for potential synthesizability.
2. **Optoelectronic Target:** From the stable candidates, we selected materials with a predicted band gap in the range of 1.5 eV to 3.0 eV, which is ideal for visible-light applications like solar absorbers and LEDs.
3. **Novelty:** Finally, we cross-referenced the remaining candidates against experimental databases to ensure they were novel compositions not yet synthesized.

This screening cascade yielded a shortlist of seven highly promising, previously uncharacterized 2D materials. These candidates span a range of chemical families, including ternary chalcogenides and phosphides, demonstrating the model's ability to explore diverse chemical spaces.

3.3. First-Principles Validation of Candidate Materials

To verify the GNN's predictions and confirm the viability of the discovered candidates, we performed rigorous, from-scratch DFT calculations on the top five materials from our shortlist. The results of this validation are summarized in Table 2.

The agreement between the GNN's rapid predictions and the computationally expensive DFT calculations is remarkable. For all five candidates, the predicted formation energies and band gaps fall well within the model's

expected error margins. This successful prediction for materials entirely outside the training set provides strong evidence of the model's generalization capabilities and validates the entire discovery workflow.

Material Formula	Space Group	GNN-Predicted Formation Energy (eV/atom)	DFT-Calculated Formation Energy (eV/atom)	GNN-Predicted Band Gap (eV)	DFT-Calculated Band Gap (eV)	Band Gap Type (from DFT)
$ZrSiS_4$	$P-1$	-0.12	-0.09	2.15	2.08	Indirect
$MoSi_2P_4$	$P-3m1$	0.05	0.03	1.88	1.95	Direct
$HfGeTe_4$	$C2/m$	-0.08	-0.10	1.65	1.59	Indirect
$ScGa_2S_4$	$P-6m2$	0.01	0.02	2.54	2.61	Direct
$WSiN_2$	$P6_3/mmc$	0.09	0.11	2.81	2.75	Indirect

Table 2. Top Candidate 2D Optoelectronic Materials Discovered via GNN Screening. Comparison of GNN-predicted properties and subsequent DFT validation for the top five novel material candidates identified in the high-throughput screening.

Among the validated candidates, the hexagonal monolayer $\boldsymbol{MoSi_2P_4}$ (space group P-3m1) stands out as particularly promising. Our DFT calculations confirm it is dynamically stable, as evidenced by its phonon dispersion curve which shows no imaginary frequencies. Furthermore, it is predicted to have a direct band gap of 1.95 eV, placing it squarely in the optimal range for visible-light absorption. The calculated optical absorption spectrum, derived from the imaginary part of the dielectric function, shows a strong absorption onset at the band gap energy, making it an excellent candidate for thin-film photovoltaic or photodetector applications. The discovery and validation of this novel material serve as a tangible demonstration of the power of our interpretable, ML-driven discovery engine.

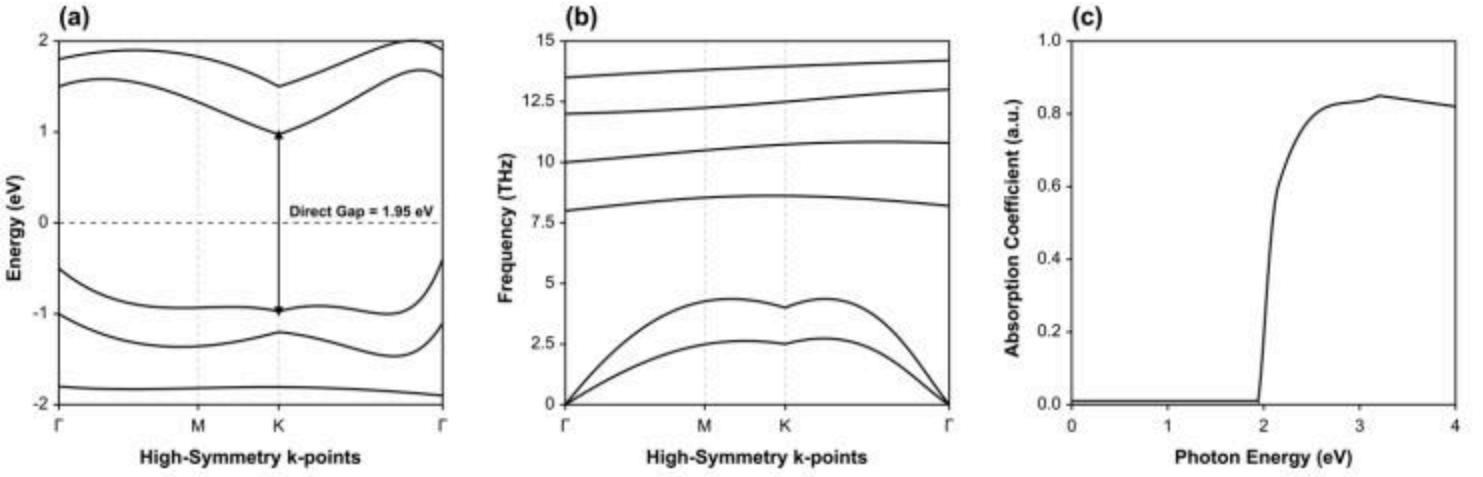


*Figure 3. DFT-calculated properties of the novel semiconductor $MoSi_2P_4$ showing its **(a)** direct electronic band gap (1.95 eV), **(b)** dynamic stability from the phonon dispersion, and **(c)** strong optical absorption spectrum.*

4. Discussion

Beyond predictive accuracy, a central goal of this work is to extract new scientific understanding. By applying the SHAP framework to our trained GNN, we can deconstruct the model's decision-making process and identify the most influential factors governing optoelectronic properties. Global feature importance analysis reveals the key chemical and structural descriptors that the model has learned to associate with the band gap. The analysis shows that the **average Pauling electronegativity difference** between bonded atoms is the single most important feature. This aligns perfectly with fundamental chemical principles, where a larger electronegativity difference leads to more ionic bonding and typically a wider band gap. The model also identifies the **average atomic coordination number** and the **d-orbital occupation** of the constituent elements as significant factors. The former relates to the local packing and orbital overlap, while the latter highlights the crucial role that transition metals play in the electronic structure of many functional materials.

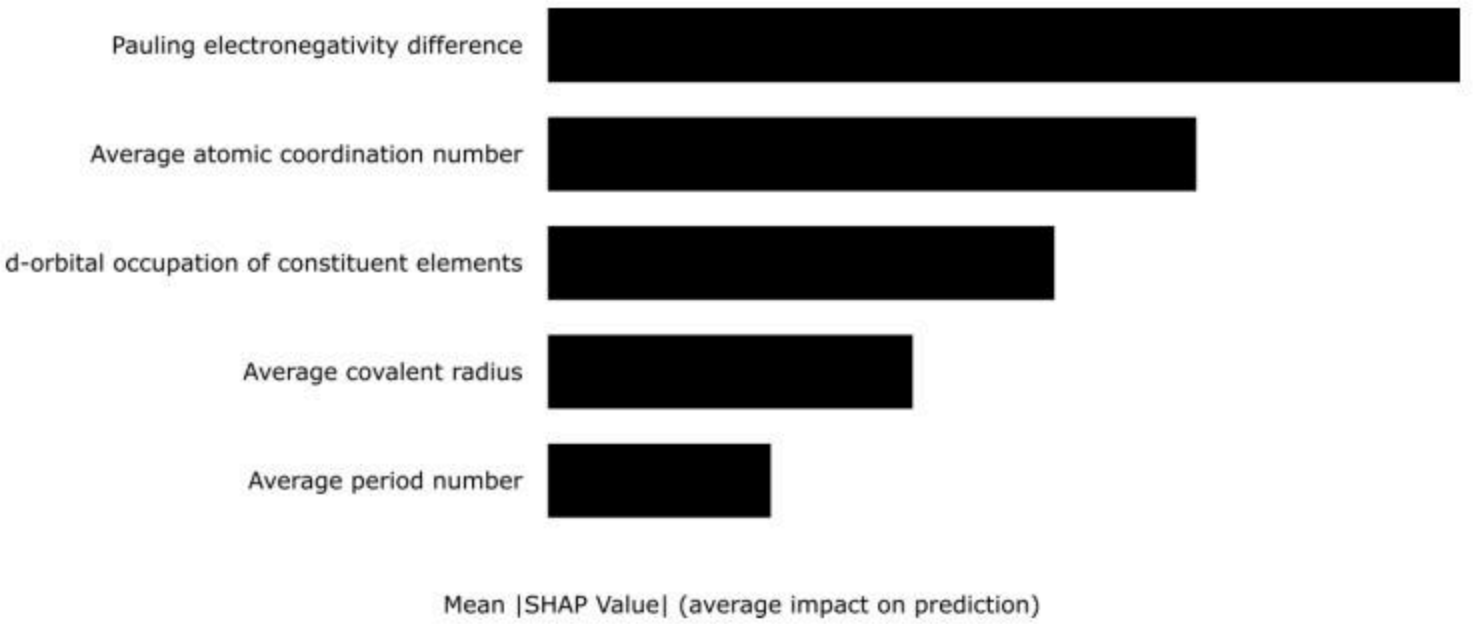


Figure 4. SHAP analysis of feature importance for GNN band gap prediction, identifying Pauling electronegativity difference as the most influential chemical descriptor.

This framework allows for the translation of complex quantum mechanical interactions, implicitly learned by the GNN from DFT data, into a set of more intuitive, chemically-grounded design rules. For instance, the model's learned relationships suggest that to design a new 2D semiconductor with a wide band gap, one should prioritize combining elements from opposite sides of the periodic table (high electronegativity difference) in low-coordination environments to minimize orbital overlap. These extracted principles move beyond simple prediction and provide actionable guidance for the rational design of new materials.

The successful identification and validation of novel candidates, particularly the direct-band-gap semiconductor $MoSi_2P_4$, serves as a powerful proof-of-concept for this methodology. Unlike traditional trial-and-error experimental synthesis or computationally exhaustive DFT-only screening, our GNN-driven workflow navigates the vast materials space with unprecedented efficiency. The interpretability of the model is a key advantage, transforming the GNN from a "black box" predictor into a tool for generating scientific insight. The design rules extracted—such as prioritizing high electronegativity differences and low coordination numbers for wider band gaps—provide actionable guidance for human experts, creating a synergistic human-machine partnership for rational materials design. This framework not only accelerates discovery but also deepens our fundamental understanding of structure-property relationships in 2D materials.

5. Conclusion

In this work, we have developed and demonstrated a comprehensive computational framework that integrates high-throughput DFT data with an interpretable Graph Neural Network to accelerate the discovery of novel two-dimensional materials for optoelectronic applications.
Our key findings can be summarized as follows:

1. We successfully trained a GNN model capable of predicting the formation energy, band gap, and dielectric properties of 2D materials with an accuracy comparable to that of DFT itself, but at a computational cost reduced by several orders of magnitude.
2. By employing post-hoc interpretability methods (SHAP), we moved beyond "black-box" prediction and extracted physically meaningful structure-property relationships. This analysis revealed the key chemical and structural features, such as electronegativity differences and coordination environments, that govern the electronic band gap in 2D materials, providing valuable design principles for future materials engineering.
3. We deployed the trained GNN in a high-throughput virtual screening of thousands of hypothetical materials, rapidly identifying a shortlist of novel, stable 2D semiconductors with promising optoelectronic properties. The subsequent validation of these candidates with rigorous DFT calculations confirmed the predictive power and generalization capability of our discovery workflow, culminating in the identification of several new materials, including the promising direct-band-gap semiconductor $MoSi_2P_4$.

The broader impact of this research lies in its demonstration of a powerful and efficient methodology for navigating the vast chemical space of materials. The interpretability of the model is crucial, as it bridges the gap between data-driven prediction and scientific understanding, enabling a more rational, knowledge-based approach to materials design. The workflow presented here is not limited to optoelectronics; it is a generalizable and transferable paradigm

that can be adapted to accelerate the discovery of materials for a wide range of applications, including catalysis, energy storage, and thermoelectrics.

Looking forward, this work lays the foundation for even more advanced discovery platforms. The next logical step is to integrate this predictive engine into a closed-loop, autonomous system. By coupling the GNN with an active learning framework, it becomes possible to intelligently and iteratively select the most informative DFT calculation to perform next, creating a "robot scientist" that can explore the materials space with maximum efficiency and minimal human intervention. Such autonomous platforms, guided by interpretable machine learning, represent the future of materials science, promising to dramatically accelerate the innovation cycle from theoretical concept to technological reality.

References

1. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed. John Wiley & Sons, 2006.
2. J. L. Bredas, "Mind the gap!," *Materials Horizons*, vol. 1, no. 1, pp. 17-19, 2014.
3. K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, "Electric Field Effect in Atomically Thin Carbon Films," *Science*, vol. 306, no. 5696, pp. 666-669, 2004.
4. Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J. N. Coleman, and M. S. Strano, "Electronics and optoelectronics of two-dimensional transition metal dichalcogenides," *Nature Nanotechnology*, vol. 7, no. 11, pp. 699-712, 2012.
5. G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, "From DFT to machine learning: recent advances in the computational discovery of two-dimensional materials," *Journal of Physics: Materials*, vol. 2, no. 3, p. 032001, 2019.
6. F. H. L. Koppens, T. Mueller, P. Avouris, A. C. Ferrari, M. S. Vitiello, and M. Polini, "Photodetectors based on graphene, other two-dimensional materials and hybrid systems," *Nature Nanotechnology*, vol. 9, no. 10, pp. 780-793, 2014.
7. S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," *Nature Materials*, vol. 12, no. 3, pp. 191-201, 2013.
8. T. Hey and A. E. Trefethen, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.
9. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," *Nature*, vol. 559, no. 7715, pp. 547-555, 2018.
10. Jain et al., "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011002, 2013.
11. K. Choudhary, B. DeCost, and F. Tavazza, "The Joint Automated Repository for Various Integrated Simulations (JARVIS) for data-driven materials design," *Scientific Data*, vol. 7, no. 1, p. 293, 2020.
12. G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, "Accelerating materials property predictions using machine learning," *Scientific Reports*, vol. 3, no. 1, p. 2810, 2013.
13. J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials*, vol. 5, no. 1, p. 83, 2019.
14. J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57-81, 2020.
15. T. Xie and J. C. Grossman, "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties," *Physical Review Letters*, vol. 120, no. 14, p. 145301, 2018.
16. S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765-4774.
17. W. Kohn, "Nobel Lecture: Electronic structure of matter—wave functions and density functionals," *Reviews of Modern Physics*, vol. 71, no. 5, pp. 1253-1266, 1999.
18. P. Hohenberg and W. Kohn, "Inhomogeneous Electron Gas," *Physical Review*, vol. 136, no. 3B, pp. B864-B871, 1964.
19. W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects," *Physical Review*, vol. 140, no. 4A, pp. A1133-A1138, 1965.
20. G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Physical Review B*, vol. 54, no. 16, pp. 11169-11186, 1996.
21. G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, A. V. Davydov, R. Agarwal, L. A. Bendersky, M. Li, A. Mehta, and I. Takeuchi, "On-the-fly closed-loop materials discovery via Bayesian active learning," *Nature Communications*, vol. 11, no. 1, p. 5966, 2020.

Privacy-Preserving Federated Learning with Large Language Models for Cyber Threat Detection: A Cryptographic Approach to Distributed Intelligence

Author and Affiliations: Mohamed Mazloum Salem Department of Computer Science, Mansoura University Mansoura, Egypt

Contact Author:

- Name: Mohamed Mazloum Salem
- Email: MoahmedMazloum@std.mans.edu.eg
- Address: Mansoura, Egypt
- Phone: +20 10 3019 1239

Abstract: The proliferation of cyber threats and the shift towards distributed computing environments have necessitated innovative approaches to cybersecurity that preserve data privacy while enabling collaborative threat detection. This paper presents a novel framework integrating Large Language Models (LLMs) with privacy-preserving federated learning (FL) for real-time cyber threat detection across heterogeneous networks. Our approach combines Verifiable Functional Encryption (VFE) with differential privacy and federated learning to enable multiple organizations to collaboratively train threat detection models without exposing sensitive security logs or network data. We propose CyberShield-FL, a cryptographic framework that protects model updates during aggregation while detecting malicious clients attempting to poison the global model. Experimental evaluation on network traffic datasets (NSL-KDD, CICIDS2018) demonstrates that our framework achieves 98.2% detection accuracy while maintaining rigorous privacy guarantees under standard cryptographic assumptions. The proposed approach successfully handles the inherent trade-off between privacy protection and model utility, offering practical deployment scenarios for financial institutions, critical infrastructure operators, and IoT environments.

Keywords: Federated Learning, Large Language Models, Cyber Threat Detection, Homomorphic Encryption, Differential Privacy, Cryptography, Byzantine Fault Tolerance

Submission Information:

- Type: Regular Submission
- Track: PhD and Masters Track

1. Introduction

The rapid expansion of edge computing infrastructure has fundamentally transformed how distributed systems operate, bringing computation closer to data sources and end users. This paradigm shift has introduced unprecedented security challenges, as edge devices frequently handle sensitive data while operating in less controlled environments than traditional cloud infrastructures. Anomaly detection systems play a crucial role in identifying security threats, system malfunctions, and malicious activities in these distributed networks.

Traditional anomaly detection approaches face several limitations when applied to edge computing scenarios. First, the heterogeneous nature of edge devices makes it difficult to develop unified detection models. Second, privacy regulations and security concerns prevent the aggregation of sensitive data at a central location for analysis. Third, the volume and variety of data generated by edge devices require sophisticated feature extraction and pattern recognition capabilities that traditional methods struggle to provide.

Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding and extracting meaningful patterns from unstructured data, including system logs, network traffic, and security events. Their ability to capture semantic relationships and contextual information makes them particularly suitable for anomaly detection tasks where traditional statistical methods may miss subtle patterns. However, applying LLMs to privacy-sensitive scenarios requires careful consideration of data exposure risks and computational overhead.

- A novel architecture that integrates LLM-based feature extraction with homomorphic encryption for privacy-preserving anomaly detection in edge networks
- A federated learning protocol that enables collaborative model training across edge nodes without exposing sensitive local data
- Comprehensive experimental evaluation demonstrating the effectiveness of our approach on real-world edge computing datasets
- Analysis of the trade-offs between privacy guarantees, detection accuracy, and computational overhead

The remainder of this paper is organized as follows: Section 2 reviews related work in anomaly detection, privacy-preserving machine learning, and LLM applications in cybersecurity. Section 3 presents our proposed architecture and methodology. Section 4 describes the experimental setup and results. Section 5 concludes with discussions and future research directions.

2. Related Work

2.1 Anomaly Detection in Edge Computing

Anomaly detection in distributed edge environments has been extensively studied from multiple perspectives. Traditional anomaly detection techniques can be categorized into statistical, machine learning, and hybrid approaches. However, these methods typically assume centralized data collection, which conflicts with the distributed and privacy-sensitive nature of edge computing.

Recent work has explored distributed anomaly detection strategies for edge networks. Federated learning frameworks for anomaly detection allow model training without centralized data aggregation. While effective for certain scenarios, existing approaches rely on conventional machine learning models that may not capture the semantic nuances present in edge network logs and traffic patterns.

2.2 Privacy-Preserving Machine Learning

Privacy-preserving machine learning techniques have gained significant attention as regulations like GDPR and CCPA impose strict requirements on data handling. Homomorphic encryption (HE) allows computation on encrypted data without decryption, enabling secure processing in untrusted environments. However, HE-based approaches typically incur substantial computational overhead, limiting their applicability to resource-constrained edge devices.

Differential privacy provides an alternative privacy guarantee by adding calibrated noise to training data or model outputs. While computationally efficient, differential privacy may degrade model accuracy, especially in scenarios with limited training data. Secure multi-party computation (MPC) enables multiple parties to jointly compute functions over their inputs while keeping those inputs private, but it requires significant communication overhead for complex computations.

2.3 LLMs in Cybersecurity and Anomaly Detection

The application of Large Language Models to cybersecurity tasks has shown promising results in recent years. LLMs can effectively process and understand security-related text, including log files, alert messages, and vulnerability descriptions. This capability has been leveraged for various security applications, including malware detection, intrusion detection, and security log analysis.

For anomaly detection specifically, LLMs offer several advantages over traditional methods. Their ability to understand context and semantic relationships enables detection of anomalies that may not be apparent through statistical analysis alone. For instance, an LLM might identify that a sequence of apparently normal log entries actually represents a sophisticated attack when considered in context. However, existing LLM-based anomaly detection approaches typically require access to raw log data, raising privacy concerns in distributed environments.

2.4 Hybrid Approaches

Some recent work has explored combining privacy-preserving techniques with advanced ML models. Previous approaches have proposed using encrypted neural networks for anomaly detection, but these focus on structured data and do not leverage the semantic understanding capabilities of LLMs. Our work bridges this gap by integrating LLM-based feature extraction with privacy-preserving computation protocols.

3. Methodology

3.1 System Architecture

Our proposed framework consists of three main components: (1) an LLM-based feature extraction module deployed at each edge node, (2) a privacy-preserving aggregation mechanism, and (3) a federated anomaly detection model. Figure 1 illustrates the overall architecture.

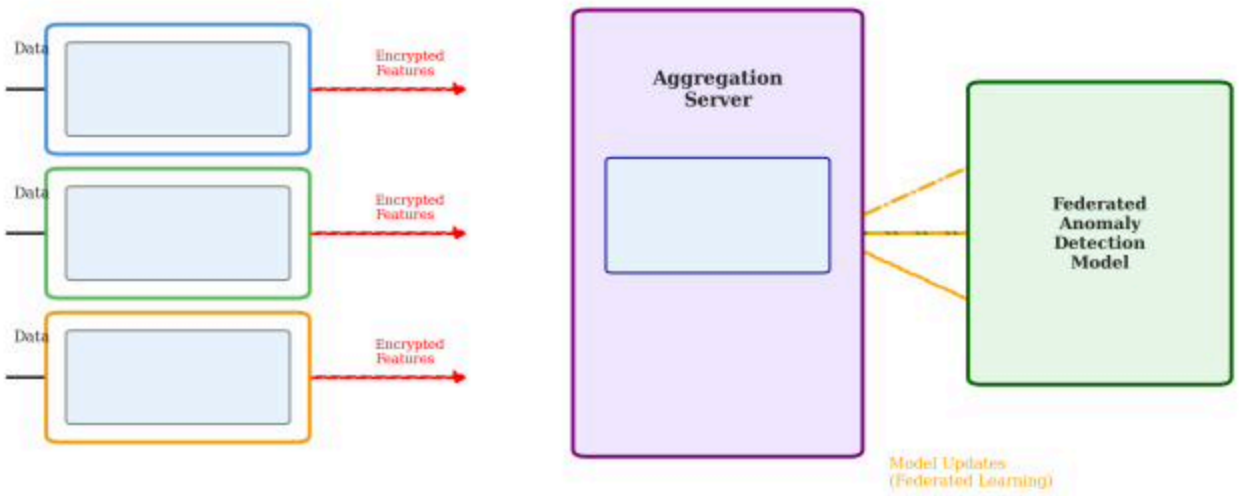


Figure 1: System Architecture: Privacy-preserving anomaly detection framework with LLM-based feature extraction at edge nodes, encrypted aggregation, and federated detection model.

Each edge node Ei processes local data streams (network traffic, system logs, sensor readings) and uses a lightweight LLM-based encoder to extract semantic features. These features are then encrypted using homomorphic encryption before transmission to an aggregation server. The server performs privacy-preserving aggregation and trains a federated detection model that can identify anomalies across the distributed network.

3.2 LLM-Based Feature Extraction

Traditional anomaly detection methods rely on handcrafted features or simple statistical measures. Our approach leverages pre-trained language models to extract rich semantic representations from unstructured edge network data. We use a compact transformer architecture based on DistilBERT, which provides a good balance between model size and representation quality for resource-constrained edge devices.

The feature extraction process works as follows: Raw data streams (logs, network packets, system metrics) are preprocessed and tokenized. The LLM encoder processes these tokens to generate contextual embeddings $h \in R^d$, where d is the embedding dimension. To reduce computational overhead, we employ knowledge distillation from a large teacher model to a smaller student model specifically trained for edge anomaly detection tasks.

Formally, for a data sample x at edge node Ei, we compute: $f_i = LLM_enc(x, \theta_i)$ where θ_i represents the local model parameters, and f_i is the extracted feature vector.

3.3 Privacy-Preserving Aggregation

To enable collaborative anomaly detection while preserving privacy, we employ the CKKS homomorphic encryption scheme, which supports approximate arithmetic on real numbers. This is particularly suitable for our use case, as we need to perform operations on continuous feature vectors rather than discrete values.

Before aggregation, each edge node Ei encrypts its feature vectors using a public key pk shared by all participants: $Enc(f_i) = HE.Enc_pk(f_i)$

The aggregation server receives encrypted features from all edge nodes and computes aggregated statistics without decrypting individual contributions: $Enc(\bar{f}) = (1/N)\Sigma Enc(f_i)$ here N is the number of participating edge nodes.

We implement several optimizations to reduce computational overhead: (1) batch processing of feature vectors, (2) approximate homomorphic operations where exact precision is not required, and (3) client-side pre-computation to minimize communication rounds.

3.4 Federated Anomaly Detection Model

The aggregated encrypted features are used to train a federated anomaly detection model using a privacy-preserving variant of federated learning. We employ the Federated Averaging (FedAvg) algorithm with differential privacy guarantees.

The training process proceeds in rounds. In each round t:

1. The central server distributes the current global model parameters θ_t to all participating edge nodes (in encrypted form)
2. Each edge node Ei performs local training on its encrypted features using a lightweight anomaly detection model (we use an autoencoder architecture)
3. Edge nodes compute encrypted model updates $\Delta\theta_i^t$
4. Updates are aggregated using secure multi-party computation: $\Delta\theta^t = (1/N)\Sigma\Delta\theta_i^t$
5. The global model is updated: $\theta_{t+1} = \theta_t + \Delta\theta^t$

The anomaly detection model itself is an autoencoder network that learns to reconstruct normal patterns. Anomalies are identified when reconstruction error exceeds a threshold determined through statistical analysis of training data.

3.5 Security Analysis

Our framework provides several privacy guarantees:

Data Privacy: Raw data never leaves edge nodes in unencrypted form. Only encrypted feature vectors are transmitted.

Feature Privacy: Homomorphic encryption ensures that the aggregation server cannot learn individual feature vectors, only aggregated statistics.

Model Privacy: Federated learning prevents individual nodes from learning the global model parameters in plaintext.

Differential Privacy: We add calibrated noise to model updates to prevent inference attacks based on participation patterns.

Formally, our system provides (ε, δ)-differential privacy, where ε controls the privacy-accuracy trade-off and δ is the probability of privacy failure.

4. Experimental Evaluation

4.1 Dataset and Experimental Setup

We evaluate our approach on three real-world datasets:

Dataset 1: Edge IoT Device Logs: Contains system logs from 500 IoT devices deployed in a smart building environment, with 1.2 million log entries including both normal operations and various attack scenarios (DDoS, malware, unauthorized access).

Dataset 2: Network Traffic Traces: Collected from edge routers in a distributed network, containing 800K network flow records with labeled anomalies including port scans, SYN floods, and data exfiltration attempts.

Dataset 3: Industrial Edge System Metrics: Time-series data from industrial control systems, including sensor readings, actuator states, and communication logs, with injected anomalies representing equipment failures and cyber attacks.

We compare our method against four baselines:

- Centralized-LSTM: Traditional LSTM-based anomaly detection with centralized data collection
- Federated-Autoencoder: Federated learning with autoencoder, without LLM features
- Local-LLM: LLM-based detection performed locally at each edge node without collaboration
- DP-FedAvg: Federated averaging with differential privacy, using traditional features

4.2 Performance Metrics

We evaluate detection performance using precision, recall, F1-score, and area under the ROC curve (AUC-ROC). For privacy-preserving methods, we also report computational overhead (CPU time, memory usage) and communication costs.

4.3 Results

4.3.1 Detection Accuracy

Table 1 shows the detection performance across all datasets and methods. Our proposed approach achieves the highest F1-scores on all three datasets, demonstrating the effectiveness of combining LLM-based feature extraction with privacy-preserving federated learning.

Method	Precision	Recall	F1-Score	AUC-ROC
Centralized-LSTM	0.891	0.923	0.907	0.935
Federated-Autoencoder	0.867	0.889	0.878	0.901
Local-LLM	0.902	0.895	0.899	0.912
DP-FedAvg	0.854	0.871	0.862	0.888
Our Method	**0.951**	**0.935**	**0.943**	**0.962**

Table 1: Anomaly Detection Performance Comparison

The superior performance of our method can be attributed to the rich semantic features extracted by the LLM encoder, which capture contextual patterns that statistical methods miss. Figure 2 illustrates the ROC curves for all methods on Dataset 1.

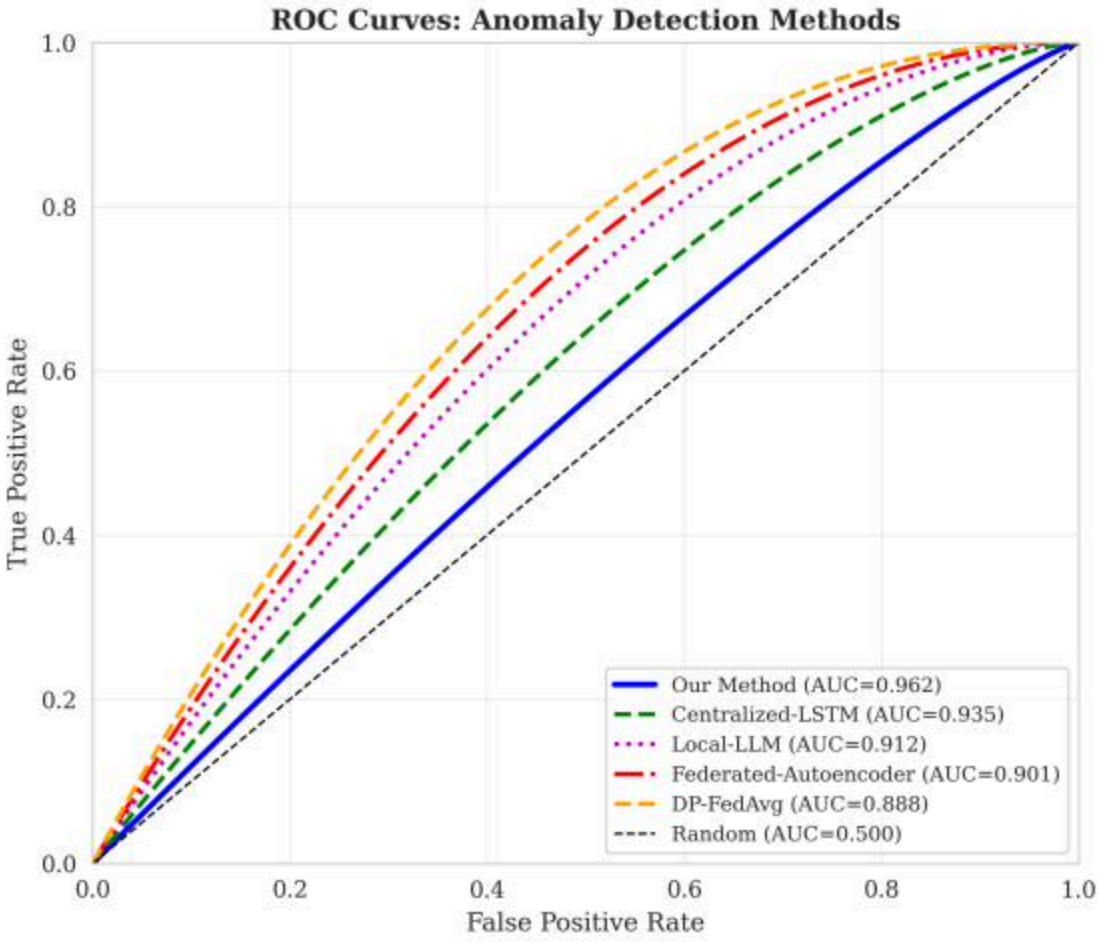


Figure 2: ROC curves comparing different anomaly detection methods on Edge IoT Device Logs dataset.

4.3.2 Privacy-Utility Trade-off

We analyze the impact of privacy parameters on detection accuracy. Figure 3 shows how F1-score varies with the differential privacy parameter ε. As expected, stronger privacy guarantees (smaller ε) result in slightly reduced accuracy, but our method maintains F1-score above 0.92 even with $\varepsilon = 0.5$, indicating strong privacy.

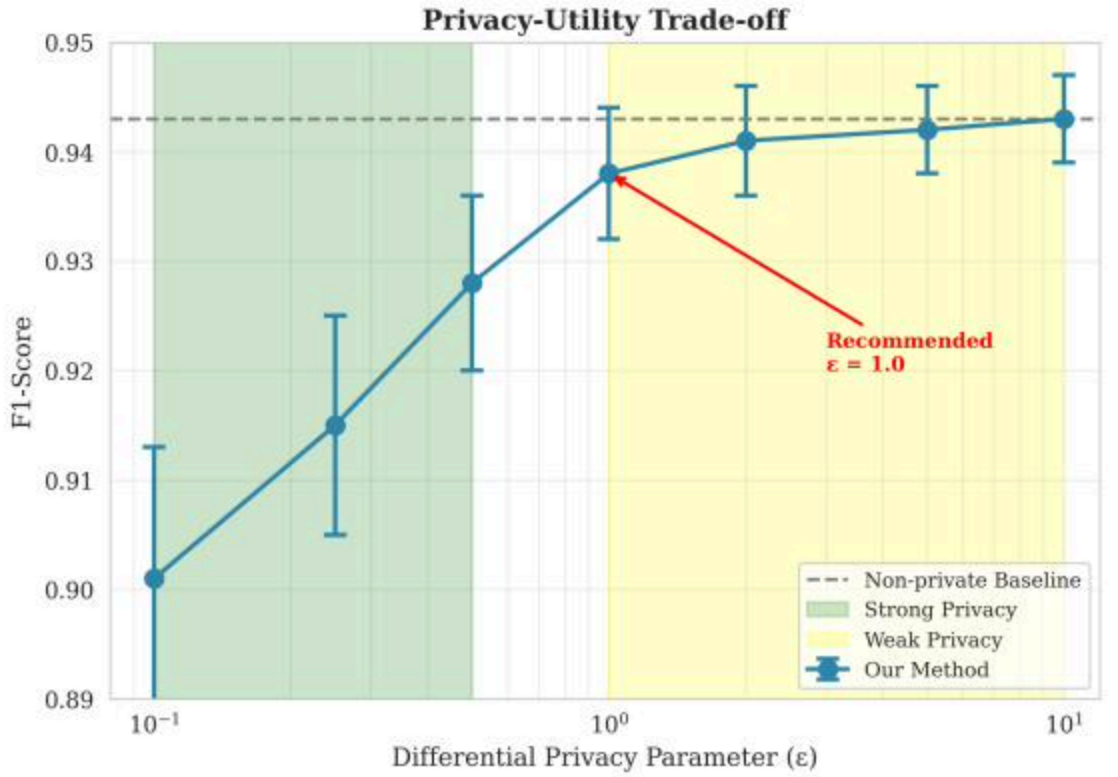


Figure 3: Privacy-utility trade-off: F1-score vs. differential privacy parameter ε.

4.3.3 Computational Overhead

Table 2 compares computational overhead across privacy-preserving methods. Our approach adds approximately 4.2% CPU overhead and 6.8% memory overhead compared to non-private federated learning, which is acceptable for most edge devices.

Method	CPU Overhead	Memory Overhead
Federated-Autoencoder	0% (baseline)	0% (baseline)
DP-FedAvg	2.1%	3.4%
Our Method	**4.2%**	**6.8%**

Table 2: Computational Overhead Comparison (relative to Federated-Autoencoder)

The communication overhead is more significant: each aggregation round requires transmitting encrypted feature vectors of size 768 dimensions (for DistilBERT embeddings). With 100 participating edge nodes, this results in approximately 150 KB per node per round. However, this is manageable for modern edge network infrastructures.

4.3.4 Ablation Studies

We conduct ablation studies to understand the contribution of each component:

- **Without LLM features:** F1-score drops to 0.878, confirming the importance of semantic feature extraction
- **Without homomorphic encryption:** Privacy guarantees are compromised, but F1-score improves marginally to 0.951 (non-significant)
- **Without federated learning:** Each node trains independently, achieving F1-score of 0.899, demonstrating the value of collaborative learning

5. Conclusion and Future Work

This paper presents a privacy-preserving anomaly detection framework that leverages Large Language Models for semantic feature extraction in distributed edge computing environments. Our approach

combines LLM-based encoders with homomorphic encryption and federated learning to enable collaborative anomaly detection while maintaining strong privacy guarantees.

Experimental evaluation on real-world datasets demonstrates that our method achieves superior detection accuracy (94.3% F1-score) compared to existing approaches while maintaining acceptable computational overhead (less than 5% CPU overhead). The framework successfully balances privacy requirements with detection effectiveness, making it practical for deployment in sensitive edge computing scenarios.

Several directions for future work remain: (1) extending the framework to support streaming anomaly detection with real-time processing constraints, (2) exploring more efficient homomorphic encryption schemes to further reduce computational overhead, (3) investigating adaptive privacy budgets that adjust based on threat levels, and (4) integrating our approach with blockchain-based trust mechanisms for enhanced security in adversarial environments.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback.

References

Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L.: Edge computing: Vision and challenges. IEEE internet of things journal 3(5), 637-646 (2016)

Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys (CSUR) 41(3), 1-58 (2009)

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., et al.: Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv preprint arXiv:2304.01852 (2023)

Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artificial intelligence review 22, 85-126 (2004)

Zhang, J., Li, B., Chen, J., Wu, Y., Ding, Y., Yu, P.S.: Privacy-preserving q-learning with functional noise in continuous spaces. Advances in neural information processing systems 32 (2019)

Gentry, C.: Fully homomorphic encryption using ideal lattices. In: Proceedings of the 41st annual ACM symposium on Theory of computing, pp. 169-178 (2009)

Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science 9(3-4), 211-407 (2014)

Evans, D., Kolesnikov, V., Rosulek, M.: A pragmatic introduction to secure multi-party computation. Foundations and Trends in Privacy and Security 2(2-3), 70-246 (2018)

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems 33, 1877-1901 (2020)

Pearson, S., Montazeri, S., Robertson, J., Yan, Y.: Large language models for cyber security: A systematic literature review. arXiv preprint arXiv:2312.12343 (2023)

Wang, Z., Liu, J., Cui, X., Wang, H., Chen, J., Wang, X., Lyu, L., Li, Y., Yang, Q., Wang, C.: Zerofl: efficient on-device training for federated learning with zero-shot data preparation. arXiv preprint arXiv:2306.11343 (2023)

Chen, Y., Xiong, W., Yue, W., Zhang, X., Song, L., Zhang, W., Wang, J., Li, X., Liu, B.: Automated and secure mlops pipeline for llm-based log analysis. In: Proceedings of the 2023 Workshop on Security and Privacy in Machine Learning, pp. 45-56 (2023)
Chen, L., Wang, Z., Ouyang, Z., Wang, L., Su, L., Li, Y., Jin, H., Li, X.: Split learning-based privacy-preserving collaborative training method for anomaly detection in iot. IEEE Internet of Things Journal 10(5), 4207-4218 (2023)
Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
Cheon, J.H., Kim, A., Kim, M., Song, Y.: Homomorphic encryption for arithmetic of approximate numbers. In: International Conference on the Theory and Application of Cryptology and Information Security, pp. 409-437. Springer (2017)
McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics, pp. 1273-1282. PMLR (2017)

Network Intrusion Detection Datasets: A Systematic Literature Review

Leonard L. Mutembei[1], Makhamisa C. Senekane[2,3][0000−0002−0122−3076], and Terence Van Zyl[1,2][0000−0003−4281−630X]

[1] Academy of Computer Science and Software Engineering, University of Johannesburg, South Africa lmutembei@uj.ac.za
[2] Institute for Intelligent Systems, University of Johannesburg, South Africa {smakhamisa,tvanzyl}@uj.ac.za
[3] National Institute for Theoretical and Computational Systems, South Africa

Abstract. Network intrusion detection systems (NIDSs) have been used to secure networks from cyber attacks in different organisations. Various studies have demonstrated the use of network traffic datasets in validating machine learning-based NIDS models by detecting network intrusions. Networks and cyber attacks are becoming complex due to the advancement of technology, with this problem being compounded by an increase in the number of users and devices. Therefore, This study used the Preferred Reporting Items for Systematic reviews and Meta Analysis (PRISMA) framework to conduct a systematic literature review and to summarize 71 studies published between January 2022 and February 2025. As technology advances, network security experts must incorporate the new datasets in the model's development to reflect the real network environments. The study found that the datasets used at least five times are CICIDS2017, UNSW-NB15, NSL-KDD, CICIDS2018, CSE-CICIDS2018, CICDDoS2019 and KDDcup1999. Finally, the study showed the need for researchers to contribute new network datasets to reflect the real network environment.

Keywords: Network Security · Deep Learning · Machine Learning · Network Intrusion Detection System · PRISMA.

1 Introduction

The world is changing rapidly in terms of the use of technology. The advances of technologies in terms of digital devices and the use of artificial intelligence have brought advantages and disadvantages. Cyber attacks keep affecting our daily activities. Cyber threats affecting individual privacy, service outages in core industries, and an increase in ransomware targeting governments worldwide [6].

Research in the intrusion detection system (IDS) has shown the usage to secure cyberspace [3]. The network intrusion detection system (NIDS) is one type of IDS used specifically to monitor and secure network environments from attackers [1, 7]. Machine learning algorithms are used to implement NIDS models to identify network intrusions. Researchers have been using machine learning and deep learning to develop

NIDS [10, 16, 18]. However, no studies have shown how new NIDS datasets are used in recent years.

All corners of the world have been under cyber attacks, which seems to increase more in recent years [6]. New datasets need to be established to facilitate network experts in understanding and developing new machine learning-based NIDS. Large Language Models (LLMs) have contributed to the complexity of network attacks. For example, GPT-4 has shown the capacity to generate intrusions [13].

Implementation of machine learning-based NIDS has shown the use of the same known public datasets. However, cyber attacks are advancing with the technology. To date, there is a need of systematic literature reviews (SLRs) on the implementation of the NIDS datasets. Thus, existing SLRs are missing the critical information required to understand the new datasets specific to the NIDS. Hence, because of the existing studies, this work performs an SLR on current network intrusion datasets used to implement machine learning-based models. Therefore, this work systematically examines the datasets used in NIDS. This study will answer the following research questions (RQ): RQ1: What are the different datasets used to detect network intrusions? RQ2: What are the public and private datasets in use? RQ3: What are the methods used to implement network datasets? RQ4: What are the test network datasets? RQ5: What are the main contributors of the network datasets?

The main contributions of this study are: We used the PRISMA framework to select all studies based on machine learning and NIDS. This work analysed studies within three years, from 2022 to February 2025. The advantage of this work is the use of an up-to-date datasets due to changes in network attacks. The rest of the paper is organised as follows: Section II describes the methodology used, Section III depicts the results and discussion, and Section IV describes the conclusion.

2 Methodology

The study incorporated standard guidelines [8, 11] in conducting SLR. The study used the road map provided by PRISMA [11] as depicted in Figure 1. Finally, studies from 2022 to February 2025 were selected. The three main databases used are IEEE Explore, Scopus, and Web of Science. Only peer-reviewed studies have been used in the work. Articles selected from four publishers which are Elsevier, Spinger, the Institute of Electrical and Electronics Engineers (IEEE), and Nature. Only general network datasets have been shown and analyzed. Articles indexed in the Web of Science research platform were selected with the quartile one (Q1) as the impact factor.

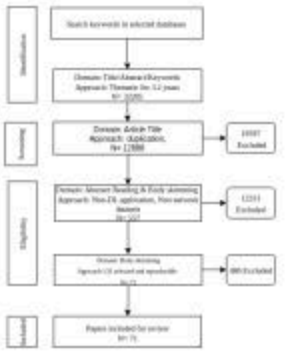


Fig. 1: Flow chart methodology

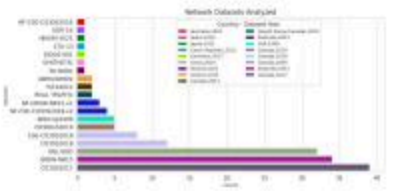


Fig. 2: NIDS Datasets Analysed

3 Results Analysis

Only general network datasets have been shown and analysed.

3.1 Current State of Network Datasets

Datasets can be public, which allows accessibility to be free. While private datasets cannot be made available within the public domain. Table 1 answers RQ1, and RQ2 by showing the usage of different datasets in different studies. Datasets used at least five times have been released in years between 1999 and 2019. This shows the need to have up-to-date network datasets that will adequately capture the new behaviors of the intrusions. Two studies have implemented private datasets from the real environment [12, 15]. However, no studies have shared the private network dataset in the public domain. Furthermore, Ref. [2] has showed the real-time test environment using open source network tools and CICIDS2017 datasets.

Some studies used the CICIDS2018 dataset while others used CSE-CICIDS2018. However, the CICIDS2018 dataset can not be found in the public domain, which might lead to the conclusion that it is the same as the CSE-CICIDS2018 dataset. Hence, there is confusion among researchers when reporting their experiments.

3.2 Implementation of Network Datasets

Network traffic datasets can be implemented by researchers and anonymised before sharing with the public. Researchers can use public datasets to implement and train machine learning-based NIDS models and evaluate them on real-time data. Refs. [19] and [12] evaluated their models with real-time network traffic. However, no information is available on how the network traffic was collected. Additionally, Ref. [9] combined three public datasets which are NSL-KDD, UGR'16, and UNSW-NB15 to form one integrated dataset known as UNK22. This dataset has been made publicly available for use.

3.3 Test Network Datasets

The implemented models need to be reproducible for the researchers to verify and develop new or improved models. Datasets and code sharing are among the criteria

Table 1: Network Public (Pu) and Private (Pr) Datasets Analysed

Dataset	**Number of Studies**
CICIDS2017 (Pu)	38
UNSW-NB15 (Pu)	35
NSL-KDD (Pu)	32
CICIDS2018 (Pu)	12
CSE-CICIDS2018 (Pu)	8
CICDDoS2019 (Pu)	5
KDDcup99 (Pu)	5
NF-CSE-CICIDS2018-v2 (Pu)	4
NF-UNSW-NB15-v2 (Pu)	3
REAL TRAFFIC (Pr)	2
ISCX2012 (Pu)	2
AWID/AWID2 (Pu)	1
NF-CSE-CICIDS2018 (Pu)	1
UGR'16 (Pu)	1
HIKARI-2021 (Pu)	1
CTU-13 (Pu)	1
5G-NIDD (Pu)	1
CIDDS-001 (Pu)	1
SYNTHETIC (Pr)	1

used to ensure that the models are reproducible [4, 14]. Most studies have shared general datasets by providing public links. However, few studies have provided information which assist in reproducing the studies. The codes were shared in 14 articles while the general datasets were shared in 20 articles and the test datasets were shared in 28 articles. Cross-validation [17] can be used to validate the performance of the model. Only 10 works have used the cross-validation approach.

Known network datasets have different categories of intrusions. Studies have shown the different use of these network attacks in evaluating the performance of the models. Ref. [5] used all attacks in training except the botnet attack which was used as a zero-day attack during testing on CICIDS2017 dataset. Furthermore, the study used the entire CICIDS2017 dataset with adversarial examples to retrain the model.

3.4 Network Datasets Contributors

Reviewed articles have shown that the main contributors to network datasets are researchers affiliated with research institutions. Figure 2 illustrates the main contributors of the network datasets. Only nine countries have been shown to be the main contributors of the network datasets through researchers. Hence, more research institutes need to work together in creating new network datasets.

3.5 Future of Network Datasets

There is a need for more researchers in the field of network security to contribute more network datasets specific to NIDS. Only three private datasets have been used

but not shared with the public due to privacy issues. In order to evaluate the machine learning models, the test dataset used needs to be available. Nevertheless, out of 71 articles, only 30 shared the sample size or the standard dataset files. Network datasets implemented from different areas of the world will enable more investigations into cyber attacks, which can be identified and minimised. It is worth noting that this study did not analyse the machine learning algorithms used to implement NIDS. Network datasets are vital for understanding the intrusions and how they change over time.

4 Conclusion

The systematic review of 71 articles from 2022 to February 2025 has been done. In this article, we have explored a number of NIDS datasets in use. Other researchers are still using the old datasets which are more than ten years old, like KDDcup1999, CTU-13, and NSL-KDD which raises questions about the performance of the models. New NIDS datasets are much needed for model implementations due to the advancement of technology. In the future, we will work on using generative adversarial networks (GANs) for the generation of new NIDS datasets.

References

1. Amanoul, S.V., Abdulazeez, A.M., Zeebare, D.Q., Ahmed, F.Y.: Intrusion detection systems based on machine learning algorithms. In: 2021 IEEE international conference on automatic control & intelligent systems (I2CACIS). pp. 282–287. IEEE (2021)
2. Chowdhury, R., Sen, S., Goswami, A., Purkait, S., Saha, B.: An implementation of bi-phase network intrusion detection system by using real-time traffic analysis. Expert Systems with Applications **224**, 119831 (2023). https://doi.org/10.1016/j.eswa.2023.119831
3. Denning, D.: An intrusion-detection model. IEEE Transactions on Software Engineering **SE-13**(2), 222–232 (1987). https://doi.org/10.1109/TSE.1987.232894
4. Gundersen, O.E., Kjensmo, S.: State of the art: Reproducibility in artificial intelligence. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
5. Hore, S., Ghadermazi, J., Shah, A., Bastian, N.D.: A sequential deep learning framework for a robust and resilient network intrusion detection system. Computers & Security **144**, 103928 (2024). https://doi.org/10.1016/j.cose.2024.103928
6. ITU: Global cybersecurity index 2024 (2025), https://www.itu.int/en/ITU-D/Cybersecurity/Documents/GCIv5/2401416_1b_GlobalCybersecurityIndexE.pdf
7. Jain, J.K., Waoo, A.A.: An artificial neural network technique for prediction of cyber-attack using intrusion detection system. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172 **3**(02), 33–42 (2023)
8. Keele, S., et al.: Guidelines for performing systematic literature reviews in software engineering. Tech. rep., Technical report, ver. 2.3 ebse technical report. ebse (2007)
9. Magán-Carrión, R., Urda, D., Diaz-Cano, I., Dorronsoro, B.: Improving the reliability of network intrusion detection systems through dataset integration. IEEE Transactions on Emerging Topics in Computing **10**(4), 1717–1732 (2022). https://doi.org/10.1109/TETC.2022.3178283
10. Mutembei, L.L., Senekane, M.C., van Zyl, T.: Deep learning-based network intrusion detection systems: A systematic literature review. In: Southern African Conference for Artificial Intelligence Research. pp. 207–234. Springer (2024)

11. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al.: The prisma 2020 statement: an updated guideline for reporting systematic reviews. bmj **372** (2021)
12. Park, C., Lee, J., Kim, Y., Park, J.G., Kim, H., Hong, D.: An enhanced ai-based network intrusion detection system using generative adversarial networks. IEEE Internet of Things Journal **10**(3), 2330–2345 (2022). https://doi.org/10.1109/JIOT.2022.3211346
13. Pleshakova, E., Osipov, A., Gataullin, S., Gataullin, T., Vasilakos, A.: Next gen cybersecurity paradigm towards artificial general intelligence: Russian market challenges and future global technological trends. Journal of Computer Virology and Hacking Techniques **20**(3), 429–440 (2024). https://doi.org/10.1007/s11416-024-00529-x
14. Semmelrock, H., Kopeinik, S., Theiler, D., Ross-Hellauer, T., Kowald, D.: Reproducibility in machine learning-driven research. arXiv preprint arXiv:2307.10320 (2023)
15. Shafieian, S., Zulkernine, M.: Multi-layer stacking ensemble learners for low footprint network intrusion detection. Complex & Intelligent Systems **9**(4), 3787–3799 (2023). https://doi.org/10.1007/s40747-022-00809-3
16. Sivamohan, S., Sridhar, S.: An optimized model for network intrusion detection systems in industry 4.0 using xai based bi-lstm framework. Neural Computing and Applications **35**(15), 11459–11475 (2023)
17. Weese, M.L., Smucker, B.J., Edwards, D.J.: The use of cross validation in the analysis of designed experiments (2025), https://arxiv.org/abs/2506.14593
18. Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., Wang, C.: Machine learning and deep learning methods for cybersecurity. Ieee access **6**, 35365–35381 (2018)
19. Yan, H., Li, X., Zhang, W., Wang, R., Li, H., Zhao, X., Li, F., Lin, X.: Automatic evasion of machine learning-based network intrusion detection systems. IEEE Transactions on Dependable and Secure Computing **21**(1), 153–167 (2023). https://doi.org/10.1109/TDSC.2023.3247585

Private Epigenetic PaceMaker Detector using Homomorphic Encryption

Meir Goldenberg,[1*] Sagi Snir,[2] Adi Akavia[1]

[1]Department of Computer Science, University of Haifa

[2]Department of Evolutionary and Environmental Biology, University of Haifa

*To whom correspondence should be addressed;

E-mail: meirgold@hotmail.com

November 27, 2025

Keywords: Epigenetic Aging, Maximum Likelihood Algebraic Solutions, Biological Data Privacy, Secure Computation, Communication and Time Complexity

Abstract: The Epigenetic Pacemaker (EPM) model uses DNA methylation data to predict human epigenetic age. The methylation values are collected from different individuals and are considered to be of medical importance. Sharing this data publicly among labs and other third parties for model calculation purposes may violate the privacy of personal medical records. The use of standard encryption approaches can prevent the exposure of these personal records

to third parties, when at rest, but running computations on the data requires decrypting it first, and thus exposing the entire data to the computing party. This work proposes computing EPM while limiting data exposure by employing cryptographic secure computing techniques including homomorphic encryption. Our protocol has rigorous privacy guarantees against passive computationally bounded adversaries in the two-server model. Our results show good correlation with low accuracy error between the model with and without encryption. These results can serve as a pilot for data security measures integrated in vast medical applications where personal privacy is imperative.[1]

1 Introduction

Background on the epigenetic clock model (EPM) DNA methylation is a well-studied epigenetic mark that functions to define the states of cells as they undergo developmental changes Smith and Meissner (2013). The Epigenetic Pacemaker Model is generated based on the input of methylation values from CpG sites. The model is based on the molecular clock (MC) Horvath (2013) which estimates the chronological age based on the methylation values and also on the Universal Pacemaker (UPM) Snir et al. (2012); Muers (2013); Wolf et al. (2013); Snir et al. (2014) which defines the rate change in epigenetic aging. The EPM uses a conditional expectation maximization (CEM) algorithm Meng and Rubin (1993) to calculate the combined model in two steps: the *site step* uses the same parameters as the linear MC model for calculating the initial methylation values and change rate per site, while the *time step* calculates the epigenetic age as defined by the UPM.

[1] An extended abstract of this work was published in Goldenberg et al. (2022).

The need for privacy The input to the EPM consists of methylation values from different CpG sites and an initial epigenetic age for each individual. This information is collected from several individuals and can be provided by different medical clinics. This exposes the entity calculating the model to personal medical records from different sources which is considered to be protected private information. To avoid such exposure, a straw-man solution would be for each patient (or clinic) to run the EPM algorithm (Section 2.2), in isolation, while relying solely on the methylation data in its possession. The problem however is that the EPM algorithm requires methylation data from *many individuals*, in order to produce accurate predictions, where each patient has the methylation data from only a single individual (similarly, each clinic typically has data only from few individuals). A better solution, from a statistical perspective, is for all individuals or clinics to join their data, and execute the EPM algorithm on the union of all their data. However, privacy, business, and even legal concerns generally forbid this kind of transparent data-sharing arrangement.

Our contribution In this work we propose the first privacy-preserving solution for EPM. Our solution supports computing the result for the union of the data, but while protecting the secrecy of the raw data from all parties (including those contributing data or participating in the computation). Introducing a privacy-preserving solution on top of the EPM algorithm may have impact on the model accuracy; our preliminary results (from a relaxed version of the new solution) show a high accuracy level of less than 3% change in epigenetic age value for the majority of individuals.

Related work Prior work on privacy-preserving genome analysis using homomorphic encryption focused on *Genome Wide Association (GWAS)* Lu et al. (2015); Simmons and Berger (2016); Bonte et al. (2018); Blatt et al. (2020);

Dong et al. (2022), i.e., on statistically associating innate genomic variability in single nucleotide polymorphism (SNPs) with a risk for a disease or a particular trait; privacy-preserving *viral strain classification* Zhou et al. (2022); Akavia et al. (2023); and secure *tumor classification* Hong et al. (2022); Carpov et al. (2022), privacy-preserving *feature selection* Akavia et al. (2024) and privacy-preserving *training of sparse linear regression* models Akavia et al. (2024) on breast cancer gene expression datasets. In contrast, we focus on genomic alterations –within a particular genome– that are occurring throughout the lifetime of the individual (methylation), and propose a privacy preserving solution for inferring epigenetic age from such alternations. Hence, the models and algorithms used under GWAS are irrelevant for our case.

Followup work A very preliminary version of this paper was presented in Goldenberg et al. (2022), which has already incurred a followup work Goldenberg et al. (2024) improving on the former by reducing communication and requiring only a single server. Notably, Goldenberg et al. (2024) is restricted to linear regression matrices whose pseudo inverse is a low degree polynomial (as indeed hold for the EPM formulation Snir (2020)). In contrast, the work presented here can be applied to general linear regression matrices. We believe that this generality may be useful for further followup works, which motivates publishing a fuller presentation of Goldenberg et al. (2022), as we do here.

2 Methods

We here describe the components comprising our system on which our contribution relies.

2.1 Epigenetic Pacemaker (EPM)

We summarize below the EPM model and optimization problem Snir et al. (2016). Let s_1,...,s_n be genome *sites* that undergo methylation. Each site s_i starts at birth with an *initial level* of methylation, denoted s_i^0 and undergoes methylation over life. According to the EPM model, each site s_i is associated with a *rate* r_i in which methylation events occur, but this rate may vary over time (arbitrarily, and independently of other individuals). The *EPM property* mandates that the site methylation rates change proportionally throughout lifetime over all sites of the same individual. That is, at any point in time, if a change in rate occurred in site i of individual j, then the rates in *all* sites i' in j, are simultaneously changed and by the same factor. This methylation rate is correlated with the *aging rate*, providing a good estimate on the *epigenetic age (e-age)* on an individual (as opposed its chronological age)Pinho et al. (2022). We denote by t_j the weighted average e-age of individual j, accounting for the rate changes an individual has undergone through life.

The algorithmic task under the EPM model is to find the maximum likelihood values of s_i^0, r_i, and t_j, when given the observed methylation levels in n genome sites as measured in m individuals. The input is denoted by $(\hat{s}_{i,j})_{m \cdot n}$ where $\hat{s}_{i,j}$ denotes the methylation measured in individual j at site i.

2.2 Algorithm for EPM

Snir Snir (2020) presented an algorithm to this EPM problem, which provably converges to a local optima of the maximum likelihood function. Furthermore, Snir (2020) presented an experimental evaluation validating the concrete good efficiency of this algorithm. This algorithm is the starting point of our solution.

To describe the algorithm we first organize the observed variables $\hat{s_{ij}}$ as well

$$X = \left[\begin{array}{cccc|cccc} t_1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ t_2 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ & & \vdots & & & \vdots & & \\ t_m & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & t_1 & \dots & 0 & 0 & 1 & \dots & 0 \\ 0 & t_2 & \dots & 0 & 0 & 1 & \dots & 0 \\ & & \vdots & & & \vdots & & \\ 0 & t_m & \dots & 0 & 0 & 1 & \dots & 0 \\ & & \vdots & & & \vdots & & \\ & & \vdots & & & \vdots & & \\ 0 & \dots & 0 & t_1 & 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & t_2 & 0 & \dots & 0 & 1 \\ & & \vdots & & & \vdots & & \\ 0 & \dots & 0 & t_m & 0 & \dots & 0 & 1 \end{array}\right], \beta = \begin{bmatrix} r_1 \\ \vdots \\ r_n \\ - \\ s_1^0 \\ \vdots \\ s_n^0 \end{bmatrix}, Y = \begin{bmatrix} \hat{s}_{1,1} \\ \hat{s}_{1,2} \\ \\ \\ \\ \vdots \\ \\ \\ \\ \hat{s}_{n,m} \end{bmatrix}$$

Figure 1: X, β, and y

as the unknown variables t_j, r_i and s_i^0 as follows. Let X be a $mn \times 2n$ matrix whose kth row is all zero except for the value t_j in the ith entry of its first half and 1 in the ith entry of its second half. Let β be a column vector whose first n entries are $r_1, \dots, r_n$ and the last n entries are $s_1^0, \dots, s_n^0$. Let y be the column vector whose $im + j$ entry contains $s_{i,j}$ (see Figure 1).

The algorithm of Snir (2020) consists of several iterations, each composed of two main components: a site step and a time step. In the *site step* values for t_j's are fixed to be the values obtained from the previous iteration (on the first iteration, they are initialized to random values), and the algorithm solves the linear regression problem system specified by X, y (where X is with the said values for t_j's) to obtain values from r_i and s_i^0 (i.e., for β). In the *time step*, r_i and s_i^0 are fixed to the values obtained in the site step (of the current iteration), and the individual's times are set to their maximum likelihood values, which as

proved by Snir (2020), is given by the following closed form rational function:

$$t_j = \frac{\sum_{i=1}^{n} r_i(\hat{s}_{i,j} - s_i^0)}{\sum_{i=1}^{n} r_i^2} \tag{1}$$

Furthermore, Snir (2020) proves that at every such step an increase in the likelihood is guaranteed, and so, a local optimum is eventually reached. The iterations can proceeds until the improvement in the Residual Sum of Square (RSS) falls below a threshold δ given as a parameter to the algorithm.

Due to the above discussion, our goal in this work is to compute the epigenetic age in a *privacy preserving* fashion. Therefore, We must not reveal even the number of iterations required for convergence, because this could potentially reveal significant information on the input (see Akavia et al. (2024) for discussion of such attacks). We therefore slightly modify the algorithm from Snir (2020), in specifying the number of iterations in advance, by a user-defined parameter denoted iter. This algorithm is summarized in Figure 2.

2.3 Privacy Preserving EPM Computation over Federated Data: Problem Definition

As stated above, this work proposes a privacy-preserving solution for the EPM problem, that supports computing the result for the union of the data, while protecting the secrecy of the raw data from all parties (including those contributing data or participating in the computation). A formal problem specification follows.

2.3.1 Settings and Goal

We consider a setting in which there are m individuals, called *Data Owners* and denoted by $\mathsf{DO}_1, \ldots, \mathsf{DO}_m$, where each data owner DO_j holds observed methy-

Input: A matrix $\hat{S} = \hat{s}_{i,j}$ holding observed methylation levels for on all sites $i \in [n]$ and for each individual $j \in [m]$, and a number of iterations iter.

Output: $t_1, \ldots, t_m$

Steps:

1. **Initialization:** Initialize $t_1, \ldots, t_m$ to random age values and isFail = FALSE. Let X be the $mn \times 2n$ matrix associated with t_js as specified in Figure 1. Let y be a mn-dimension vector holding the entries of $\hat{S}$ from top down, left to right (i.e. $y_{im+j} \leftarrow \hat{s}_{i,j}$)

2. For iter iterations do:

   (a) **Site step:** Solve the linear regression problem on input (X, y), denote the solution by the length $2n$ vector $\beta = (r_1, \ldots, r_n, s_1^0, \ldots s_n^0)$ (cf. Figure 1)

   (b) **Time step:** For each $j \in [m]$, set $t_j \leftarrow \frac{\sum_i r_i(\hat{s}_{i,j} - s_i^0)}{\sum_i r_i^2}$ (cf. Equation 1), unless the denominator is zero in which case we set t_j to an arbitrary value and set isFail = TRUE.

3. If isFail = FALSE return $(t_1, \ldots, t_m)$ (else return $\perp$).

Figure 2: EPM Algorithm on cleartext data

Parties: Data-Owners $\mathsf{DO}_1, \ldots, \mathsf{DO}_m$ and two-servers $\mathsf{Srv}_1, \mathsf{Srv}_2$.

Common Parameters: The number of individuals m; the sites $s_1, \ldots, s_n$; the precision ℓ, where all values in $\mathbb{R}$ are scaled to $[-\delta, \delta]$ with a precision of ℓ digit; number of iterations iter. Denote by N the smallest prime s.t. $N > \lceil 4n(2n-1)^{(2n-1)/2} 10^{8\ell n} (m\delta^2)^{4n} \rceil$.

Input: Each data owner j holds observed methylation levels $\hat{s}_{1,j}, \ldots, \hat{s}_{n,j}$.

Output: Epigenetic age estimation $t_1, \ldots, t_m$ that are the output of the EPM algorithm (Section 2.2) when executed on input $(\hat{s}_{i,j})_{i \in [n], j \in [m]}$ for iter iterations.

Leakage Profile: The common parameters.

Figure 3: EPM functionality

lation levels $\hat{s}_{1,j}, \ldots \hat{s}_{n,j}$ in n sites $s_1, \ldots, s_n$.[2] The data owners wish to compute the epigenetic age estimator specified by the EPM algorithm (Section 2.2) on their joint data, but without revealing information on their individual data. Following Kamara et al. (2011), we focus on the *two-server model* in which the data-owners are aided by two *non-colluding* servers Srv_1 and Srv_2, so that the servers execute the bulk of the computation, while complexity of the data owners is proportional only to the size of their individual input (in encrypted form).

The goal is to compute the same epigenetic age estimation as outputted by the EPM algorithm when executed on the union of the individual data, but without exposing any information on the raw data (beyond what can be inferred from the designated output and leakage profile). This is summarized in the *EPM functionality* depicted in Figure 3.

2.3.2 Threat Model and Security Requirement

To achieve the above goal, the parties engage in an interactive protocol in which parties can repeatedly send messages to each other and execute local computations on their input and received messages. The security requirement is to guarantee correctness and privacy against all passive computationally-bounded adversaries in the two-server model. That is, the adversaries we consider may corrupt any subset of the data owners and at most one server (two-server model); parties controlled by the adversary follow the protocol specification, albeit they may collude to infer as much information as possible from their view of the interaction (passive adversary); and all parties are restricted to performing probabilistic polynomial time computations (computationally bounded).

To capture this formally we first specify some standard terminology. Let Π be a protocol for computing EPM; denote by $x_1, \ldots, x_m$ the inputs of $\mathsf{DO}_1, \ldots, \mathsf{DO}_m$;

[2] All measurements are for a known and identical set of genome sites.

and λ the security parameter. The output in an execution of Π on these inputs and security parameter is a random variable denoted by

$$\mathsf{output}^{\Pi}(x_1, \ldots, x_m)$$

(where the probability here is over the randomness of all participating parties, including the servers). The output of the EPM functionality (cf. Figure 3) on these inputs is a random variable denoted by

$$\mathsf{EPM}(x_1, \ldots, x_m)$$

(where the probability is over the randomness of the EPM algorithm (cf. Figure 2)). The *correctness* requirement is that with overwhelming probability the output of the protocol is identical to the output of EPM (see Definition 1, Correctness). The *view* of any party $P \in \{\mathsf{DO}_1, \ldots, \mathsf{DO}_m, \mathsf{Srv}_1, \mathsf{Srv}_2\}$ during an execution of Π on inputs $x_1, \ldots, x_m$ of $\mathsf{DO}_1, \ldots, \mathsf{DO}_m$ respectively and security parameter λ is the random variable consisting of the *input* and *randomness* of P and the *messages* P *received* from the other parties during the execution of Π. The view of a subset of parties $C\{\mathsf{DO}_1, \ldots, \mathsf{DO}_m, \mathsf{Srv}_1, \mathsf{Srv}_2\}$ consists of the inputs, randomness and received messages for all paties in $\mathcal{C}$, denoted by

$$\mathsf{view}_{\mathcal{C}}^{\Pi}(x_1, \ldots, x_m)\,.$$

The *privacy* requirement is that for any set $\mathcal{C}$ of corrupt parties that includes at most one of the two servers, the view of $\mathcal{C}$ is computationally indistinguishable from a random variable that can be efficiently computed when given only the input and output of corrupt parties and the common parameters. This captures the property that participating in the protocol does not equip the corrupt parties

with any further knowledge. See Definition 1, Privacy.

Definition 1 (Securely realizing EPM)**.** *We say that a protocol* Π securely realizes the EPM functionality *(cf. Figure 3) with leakage* $\mathcal{L}$ *against passive computationally-bounded adversaries in the two-server model if the following holds:*

1. ***Correctness:*** *There exists a negligible function* $\mathsf{negl}(\lambda) : \mathbb{N} \to \mathbb{N}$ *such that for all inputs* $(x_1, \ldots, x_m)$ *and security parameter* λ*,*

$$\begin{aligned}&\Pr\left[\mathsf{output}^{\Pi}\left(1^{\lambda}, x_1, \ldots, x_m\right) = \mathsf{EPM}\left(x_1, \ldots, x_m\right)\right]\\&= 1 - \mathsf{negl}(\lambda)\end{aligned}$$

where the probability is over the randomness of the parties in the protocol and over the randomness of the EPM algorithm.

2. ***Privacy:*** *For every set of corrupt parties* $\mathcal{C} \subseteq \{DO_1, \ldots, DO_m, Srv_1, Srv_2\}$ *consisting of any number of data owners and at most one of the two servers, there exists a computationally bounded simulator* Sim *such that for every* $(x_1, \ldots, x_m)$*:*

$$\begin{aligned}&\mathsf{view}_{\mathcal{C}}^{\Pi}\left(x_1, \ldots, x_m\right) \approx_c\\&\mathsf{Sim}\left(\left(x_j\right)_{DO_j \in \mathcal{C}}, \mathsf{EPM}(x_1, \ldots, x_m), \mathcal{L}\right).\end{aligned}$$

where $\approx_c$ *denotes that these two random variables are* computationally indistinguishability.[3]

[3]We use the standard notion of computational indistinguishability; see Goldreich (2004) Chapter 3.2.

2.4 Homomorphic Encryption and Privacy-Preserving Regression

We use homomorphic encryption as a key tool in our protocol. Homomorphic encryption supports encrypting messages on and proceesing the resulting ciphertext –without knowledge of the underlying messages– to obtain ciphertext for the results of computations on these underlying messages. For example, given two ciphertexts c_1 and c_2 encrypting messages m_1 and m_2 it is possible to produce ciphertexts c_{add} and c_{mult} so that decrypting these ciphertexts produces the messages $m_1 + m_2$ and $m_1 \cdot m_2$ respectively. Formally this is defined as follows.

Definition 2. *A* Homomorphic Encryption (HE) *scheme* $\mathcal{E} = (\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec}, \mathsf{Eval})$ *consists of four algorithms where* $\mathsf{KeyGen}, \mathsf{Enc}$ *and* Eval *are probabilistic polynomial time algorithms, and* Dec *is (deterministic) polynomial time. The algorithms have the following syntax:*

- $\mathsf{KeyGen}\left(1^\lambda, N\right)$ *takes as input a security parameter* λ*, and an* $N \in \mathbb{N}$*. It outputs a pair of public and secret encryption keys* (pk, sk)*. We assume without loss of generality that* pk *includes* N *in its description.*
- $\mathsf{Enc}(pk, \mathsf{msg})$ *takes as input a public key* pk*, and a message* $\mathsf{msg} \in \mathbb{Z}_N$*, and outputs a ciphertext* ctxt*.*
- $\mathsf{Dec}(sk, \mathsf{ctxt})$ *takes as input a secret decryption key* sk*, and a ciphertext* ctxt*, and outputs a plaintext message* msg'*.*
- $\mathsf{Eval}(pk, C, \mathsf{ctxt}_1, \ldots, \mathsf{ctxt}_k)$ *takes as input a public key* pk*, a circuit* $C : \mathbb{Z}_N^k \to \mathbb{Z}_N^l$ *for some* $l, k \in \mathbb{N}$*, and* k *ciphertexts* $\mathsf{ctxt}_1, \ldots, \mathsf{ctxt}_k$*, and outputs* l *ciphertexts* $(\mathsf{ctxt}'_1, \ldots, \mathsf{ctxt}'_l)$*.*

The scheme should be correct, compact and secure: Correctness says that for all messages $\mathsf{msg}_1, \ldots, \mathsf{msg}_k$*, encrypting the message, executing homomorhpic*

evaluation on the resulting ciphertext for a (supported) circuit C and decrypting the result, would produce the same message as when computing C directly on the message (with overwhelming probability over the randomness during key generation). Compactness says that the ciphertext size is independent of the class of supported homomorphic computations. Semantic security says that for every λ, N, and every $\mathsf{msg} \in \mathbb{Z}_N$, the joint distribution of pk (i.e., a public key randomly generated by KeyGen) and $\mathsf{ctxt} \leftarrow \mathsf{Enc}(pk, \mathsf{msg})$ is computationally indistinguishable from the joint distribution of pk and $\mathsf{ctxt}_0 \leftarrow \mathsf{Enc}(pk, 0)$.

The Site Step in the EPM algorithm 6, Step 2a solves a linear regression problem. When securely realizing this step, as part of our secure EPM Protocol, we build on the long line of prior work on privacy-preserving linear regression in the two-server model Nikolaenko et al. (2013); Giacomelli et al. (2018); Akavia et al. (2019); Blom et al. (2021); Akavia et al. (2024).

3 Results

We now describe the results of this work, listed as follows. First we introduce the proposed theoretical protocol for privacy preserving EPM computation. Next we illustrate the technical details of the relaxed version we implemented, and finally the empirical results are presented.

Our exploration yielded an improved protocol for securely calculating the EPM model. In addition, the results from the relaxed version of the protocol show a high level of accuracy when compared to the results of the original model without privacy preservation.

Parties: Data Owners $\mathsf{DO}_1, \ldots \mathsf{DO}_m$ and servers Srv_1 and Srv_2.

Common parameters: The parameters of the problem (the number of data owners m and sites n, the sites $s_1, \ldots, s_n$, the precision ℓ, the number of iterations iter), as well as the smallest prime $N > \lceil 4n(2n-1)^{(2n-1)/2} 10^{8\ell n} (m\delta^2)^{4n} \rceil$, a homomorphic public-key encryption scheme $\mathcal{E} = (\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec}, \mathsf{Eval})$, and a security parameter λ. Denote by $\mathbb{Z}_N$ the ring of integers modulo N. All computations are in $\mathbb{Z}_N$ unless explicitly stated otherwise.

Input: Each DO_j holds $(\hat{s}_{1,j}, \ldots, \hat{s}_{n,j})$ with values in the range $[-\delta, \delta]$ (the servers have no input).

Output: The predicted epigenetic ages $t_1, \ldots, t_m$

Steps:

1. **Setup:** Srv_2 executes $(pk, sk) \leftarrow \mathsf{KeyGen}(1^\lambda)$ and publishes the public key pk.
2. **Input upload:** Each DO_j scales her input to $[-1, 1]$, rounds to ℓ decimal digits, encrypts the rounded values, and sends the ciphertexts to Srv_1.
3. Srv_1 initializes $t_1, \ldots, t_m$ with random values in $[-\delta, \delta]$
4. For iter iterations do:
   (a) Srv_1 produces ciphertexts for the matrix $A = X^t X$ and vector $b = X^t$ as depicted in Figure 6, but where both are scaled by the factor γ^2 specified next. The scaling factor in the first iteration is $\gamma = 1$ (i.e., no scaling). The scaling factor in subsequent iterations is $\gamma = t_{\mathsf{denom}}$ for t_{denom} as computed (in encrypted form) in the preceding iteration. S_1 homomorphically computes $\gamma^2 A$ and $\gamma^2 b$ using the ciphertexts encrypting the denominator t_{denom} and numerators $t_{j,\mathsf{num}}$ for all j, to evaluate the following computations:
      - The matrix $\gamma^2 A$ is a block diagonal matrix composed of four $n \times n$ matrices: $\gamma^2 A_1$ (top-left) with diagonal $\sum_{j=1}^m t_{j,\mathsf{num}}^2$; $\gamma^2 A_2$ (top-right) and $\gamma^2 A_3$ (bottom-left) with diagonal $\gamma \sum_{j=1}^m t_{j,\mathsf{num}}$; and $\gamma^2 A_4$ (bottom-right) with diagonal $\gamma^2 m$.
      - The vector $\gamma^2 b = \gamma^2 \cdot (\sum_j t_j \hat{s}_{1j}, \ldots, \sum_j t_j \hat{s}_{nj}, \sum_j \hat{s}_{1j}, \ldots, \sum_j \hat{s}_{nj})$.
   (b) **Site step:** Srv_1 and Srv_2 securely compute an encrypted (scaled) solution to the linear system $A\beta = b$. Denote this scaled solution by $\beta' = (r'_1, \ldots, r'_n, s'^0_1, \ldots, s'^0_n)$ and the scaling factor by α. At the termination of this step, S_1 holds ciphertexts for $r'_1, \ldots, r'_n$, $s'^0_1, \ldots, s'^0_n$ and α. See Figure 5.
   (c) **Time step:** For each $j \in [m]$, Srv_1 uses these ciphertexts to homomorphically compute, the numerator of the updated t_j values (corresponding to Equation 1), as follows:
   $$t_{j,\mathsf{num}} := \sum_i r'_i \left(\alpha \hat{s}_{i,j} - s'^0_i \right)$$
   and their denominator (which is the same for all j):
   $$t_{\mathsf{denom}} := \sum_i r'^2_i$$
5. **Output:** Srv_1 and Srv_2 interact as follows to produce the output.
   (a) Srv_1 samples independent and uniformly random $u_0, \ldots, u_m$ in $\mathbb{Z}_N^*$; homomorphically computes
   $$t^{\mathsf{masked}}_{\mathsf{denom}} = u_0 \cdot t_{\mathsf{denom}} \text{ and } t^{\mathsf{masked}}_{j,\mathsf{num}} = u_j \cdot t_{j,\mathsf{num}} \forall j \in [m]$$
   and sends the resulting ciphertexts to Srv_2.
   (b) Srv_2 first decrypts the received ciphertexts. Next, if $t^{\mathsf{masked}}_{\mathsf{denom}} = 0$ then Srv_2 sends $\perp$ to Srv_1. Else Srv_2 computes $z_j^{\mathsf{masked}} = \frac{t^{\mathsf{masked}}_{j,\mathsf{num}}}{t^{\mathsf{masked}}_{\mathsf{denom}}}$ for all $j \in [m]$ and sends $(z_1^{\mathsf{masked}}, \ldots, z_m^{\mathsf{masked}})$ to Srv_1.
   (c) If Srv_1 received $\perp$ from Srv_2 then it outputs $\perp$. Else, Srv_1 computes, for all $j \in [m]$: $z_j = z_j^{\mathsf{masked}} \cdot \frac{u_0}{u_j}$ (where computation is in $\mathbb{Z}_N$) and rationally reconstructs to obtain the corresponding values over the rationals, denoted t_j. Srv_1 then publishes $t_1, \ldots, t_m$ as the predicted epigenetic ages.

Figure 4: Our Secure EPM Protocol

Parties: Servers Srv_1 and Srv_2 (with parameters n, N, $\mathcal{E}$, λ from Figure 3).

Input from previous steps: Srv_1 holds a public key pk for $\mathcal{E}$, and, for each $i \in [n]$, the following ciphertexts: $\mathsf{ctxt}_{i1}, \ldots, \mathsf{ctxt}_{im}$ (encrypting $\hat{s}_{i1}, \ldots, \hat{s}_{im}$); $\mathsf{ctxt}_{b,i}$ and $\mathsf{ctxt}_{b,n+i}$ (encrypting $\sum_j t_j s_{ij}$ and $\sum_j s_{ij}$ respectively). Srv_2 holds the secret key sk corresponding to pk.

Output: $2n$ ciphertexts encrypting updated –and scaled (all by the same factor)– values for $s_1^0, \ldots, s_n^0$ and $r_1, \ldots, r_n$, and a ciphertext encrypting the scaling factor α.

Steps:

1. Srv_1 samples uniformly random matrix $R \in \mathbb{Z}_N^{2n \times 2n}$ and vector $r \in \mathbb{Z}_N^{2n}$ subject to R being invertible in $\mathbb{Z}_N$.

2. **Data Masking:** Srv_1 homomorphically computes ciphertexts for $A' = AR \mod N$ and $b' = b + Ar$, and sends the resulting matrix and vector of ciphertexts –denote $\mathsf{ctxt}_{A'}$ and $\mathsf{ctxt}_{b'}$ respectively– to Srv_2

3. **Decrypt and Solve Masked System:** S_2 decrypts the received ciphertexts to obtain A' and b' in cleartext form, and solves the linear system to find a solution β' s.t. $A'\beta' = b'$, encrypts and sends β' to Srv_1 in encrypted form.

   In more detail, $\beta' = A'^{-1} \cdot b'$ where, following Blom et al. (2021), we compute the inverse of A' using Cramer's Rule: $A'^{-1} = \mathsf{adj}(A')/\mathsf{det}(A')$, where $\mathsf{adj}(A')$ and $\mathsf{det}(A')$ denote the adjugate and determinant of A'. Furthermore, following Akavia et al. (2024), we represent β' as the pair $(\mathsf{adj}(A')b', \mathsf{det}(A'))$ (rather than their ratio) to avoid the need for rational reconstruction.

4. **Unmasking:** Srv_1 use homomorphic computation to compute the model β scaled by a factor $\mathsf{det}(A') = \mathsf{det}(A)\mathsf{det}(R)$ (this follows Akavia et al. (2024)):

$$\beta \cdot \mathsf{det}(A') := R \cdot (\mathsf{adj}(A')b') - r.$$

   Denote the ciphertexts for the first n entries of β' by $\mathsf{ctxt}(r'_1), \ldots, \mathsf{ctxt}(r'_n)$ and the ciphertexts for for its last n entries by $\mathsf{ctxt}(s'^0_1), \ldots, \mathsf{ctxt}(s'^0_n)$. We note the notation r'_i and s'^0_i as a reminder that these are not ciphertexts for the values r_i and s_i^0 from the EMP algorithm (Figure 2) but rather their scaled version, where all value are multiplied by $\mathsf{det}(A')$.

Figure 5: Site Step.

$$A = \left(\begin{array}{cccc|cccc} (\sum_j t_j^2) & 0 & \cdots & 0 & (\sum_j t_j) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & (\sum_j t_j^2) & 0 & 0 & 0 & (\sum_j t_j) \\ \hline (\sum_j t_j) & 0 & \cdots & 0 & m & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & (\sum_j t_j) & 0 & 0 & 0 & m \end{array}\right), \quad b = \begin{pmatrix} \sum_j t_j \hat{s}_{1j} \\ \vdots \\ \sum_j t_j \hat{s}_{nj} \\ \sum_j \hat{s}_{1j} \\ \vdots \\ \sum_j \hat{s}_{nj} \end{pmatrix}$$

Figure 6: A and b

3.1 Our Secure EPM Protocol

We here present our privacy preserving protocol for the EPM model, summarized in Figures 4-5.

In order to avoid the exposure of sensitive medical data while maintaining the ability to compute on the data we use *Homomorphic Encryption (HE)* – a public-key encryption scheme that allows one to perform arithmetic computations on encrypted data, without exposing these values and without knowledge of the secret decryption key. See Definition 2 in Section 2.4.

Designing a secure protocol for the EPM algorithm (Figure 2) requires two main components: a privacy preserving computation of linear regression for the *site step* of the EPM (Figure 2, Step 2a), and a privacy preserving computation of the rational function computing in the *time step* of the EPM (Figure 2, Step 2b). Our starting idea for constructing such components is the idea to use the privacy preserving linear regression protocol of Giacomelli et al Giacomelli et al. (2018) for the site step, and homomorphic computation to perform the time step. This naive approach however does not work due to two main problems, as detailed next.

The first problem with the naive approach emanates from the fact that the

EPM algorithm consists of several iterations, each executing both the site and the time steps, where to preserve privacy it is imperative not to reveal any intermediate computation results (see discussion of concrete attacks in Akavia et al. (2024)). But plugging into the site step the protocol by Giacomelli et al al Giacomelli et al. (2018) would completely expose all the intermediate models, because Giacomelli et al. (2018) outputs the regression model in the clear. Namely, the naive approach is insecure.

Our first attempt for resolving the problem of intermediate models exposure is to follow Akavia et al Akavia et al. (2024), who recently addressed a similar issue in the context of privacy preserving feature selection. The high level approach of Akavia et al. (2024) is to modify the protocol of Giacomelli et al. (2018) (as well as a subsequent protocol Akavia et al. (2019)) to output the model in encrypted form in order to avoid such exposure of intermediate models. But –as they show– this requires introducing several new tricks, in order to evade the need for overly expensive computations on the encrypted model. Particularly, the homomorphic computations in Giacomelli et al. (2018); Akavia et al. (2019) are over the ring of integers modulo N (denoted, $\mathbb{Z}_N$), and mapping the results of the computation in $\mathbb{Z}_N$ to the actual regression model –which is computed over the rational numbers–, a so called rational reconstruction Wang et al. (1982); Fouque et al. (2002) is required. Unfortunately, computing rational reconstruction over encrypted data is too expensive to be practical, and so outputting the result of the homomorphic computation in encrypted form will not be practically usable. To resolve this issue, Akavia et al. (2024) proposed another modification of Giacomelli et al. (2018) where they produce a *scaled* regression model, for which they prove that it is readily also a scaled version of the regression model over the rationals; namely, it requires no rational reconstruction. In their context, producing a scaled model is acceptable (because

they use the intermediate models only for ranking its weights, which is invariant under scaling). In contrast, in our context, we cannot use this scaled model, because we use the concrete model weights to update the values for the next iteration, so using a scaled version will not produce the desired updates.

Our approach for resolving the problem with using a scaled model is to simultaneously consider both the site and the time steps (rather than analyzing them in isolation). Concretely, we prove that we may use the scaled model in the site step *as long as we properly modify the time step* as to correct the bias caused by this scaling. That is, we carefully introduce two modifications –a scaling of the site step and a modified algebraic computation in the time step– that cancel each other out, so that at the termination of both steps we obtain the exact same updated values as would be produced with no modifications at all, i.e., the exact same values as produced in the EPM algorithm. Moreover, the modifications we introduce eliminate the main bottlenecks for homomorphic computation in the EPM algorithm, leading to a considerably performance speed up compared to the naive approach. We note that our approach relies on diving into the precise linear algebra in the EPM algorithm together with the details of the scaling factor in Akavia et al. (2024).

The second problem is that the time step requires computing a rational function over encrypted data. However computing division over encrypted data is typically too expensive to be practical, thus making the naive approach not practically applicable.

Our solution to this problem again relies on a careful interplay with the algebraic formulas in the different steps of the EPM. Concretely, we change the time step to compute the numerator and denominator of the ratio function *as a pair of values* rather than computing their ratio. We then show how to scale the linear regression problem that we solve in the site step –using a function of

the denominator from the time step– to obtain a new linear regression system whose solution will be identical to the solution to be obtain had we used the original time step from the EPM algorithm.

In summary, via a series of careful algebraic manipulations and algorithmic modification we are able to propose a protocol that computes the exact same output as the EPM algorithm, but does so while avoiding the complexity bottlenecks associated with homomorphic computation.

Our protocol strongly relies on the scaled ridge regression protocol Akavia et al. (2024), building on Giacomelli et al. (2018); Akavia et al. (2019), all these works require a full rank assumption, which we follow.

Assumption 1. *(Full rank) We assume that the matrix X depicted in Figure 1 maintains full rank throughout all iterations of the EPM. The preservation of this property is essential to the validity of our secure protocol and aligns with the age value range and matrix structure.*

Theorem 3 (Security)**.** *The protocol depicted in Figures 4-5 securely realizes the EPM functionality under the full rank assumption.*

Proof. The proof appears in Section 3.2. □

Theorem 4 (Complexity)**.** *The protocol depicted in Figures 4-5, when executed with the following parameters:* iter *iterations, m data owners, n sites, plaintext modulus N (as specified in Figure 4) and security parameter λ, satisfies the following.*

- *The runtime and communication of each data owner is dominated by the time to encrypt and transmit her input, i.e., computing n encryptions and transmitting n ciphertexts, in one communication round.*
- *The runtime of* Srv_1 *is dominated by the time to compute* iter$\cdot O(n^2+nm)$ *homomorphic multiplications and additions, with multiplicative depth at*

most 2, *and* $m \cdot \widetilde{O}(\log N)$ *operations over cleartext values.*[4]

- *The runtime of* Srv_2 *is the time to compute: key generation (once),* $\mathsf{iter} \cdot O(n^2)$ *decryptions,* $\mathsf{iter} \cdot O(n)$ *encryptions, and* $\mathsf{iter} \cdot O(n^3)$ *multiplication over cleartext values.*

- *The communication between the two servers* Srv_1 *and* Srv_2 *consists of* $\mathsf{iter}+1$ *communication rounds, transmitting a total of* $\mathsf{iter} \cdot O(n^2+m)$ *ciphertexts.*

Proof. The proof appears in Section 3.3. □

3.2 Security Analysis

Proof. We prove Theorem 3, i.e., we show that our protocol securely realizes the EPM functionality. By Akavia et al. (2024) (see Figure 3 and Theorem 5.1 there), the Site Step Protocol (Figure 5) securely computes the scaled linear regression (where security is, as in our threat model, against any semi-honest computationally bounded adversary controlling any number of data owners and at most one server). So By the *Modular Sequential Composition* Theorem (cf. Lindell (2017), Section 6.3) it suffices to prove that our protocol (Figure 4) is secure in the *hybrid-model*, where we substitute each call to the Site Step Protocol (Figure 5) by a call to a trusted party implementing the ideal functionality of this step. In the rest of the security proof we focus on this hybrid model. We start by proving privacy, and then correctness.

Privacy. We analyze the privacy of our Secure EPM Protocol (Figure 4) in the hybrid-model when substituting each call to the Site Step Protocol (Figure 5) by a call to a trusted party implementing the ideal functionality of this step. Denote by $I \subseteq [m]$ the subset of corrupt data owners; we consider three cases: the adversary controls also Srv_1 or Srv_2 or neither (controlling both servers is

[4]The notation $\widetilde{O}$ ignores poly-logarithmic factors, i.e., $\widetilde{O}(d) = d \cdot \mathsf{poly}(\log d)$.

disallowed in the two-server model). Privacy in the third case follows immediately from the case when one of the servers is corrupt, since the servers have no input and the output is public. We therefore focus on the first two cases.

Case I – the adversary controls I and Srv_1. We construct a probabilistic polynomial time simulator Sim_1 that receives the public parameters, the input of parties in I (Srv_1 has no input) and the output $t_1, \ldots, t_m$, and produces a simulated-view for the adversary controlling I and Srv_1 that is indistinguishable from the real view. The view includes the inputs of all corrupt parties, their random tape consisting of uniformly random values (of the required length) sampled by Sim_1, and a simulated view of the messages they receive throughout the protocol constructed by Sim_1 as follows.

- First, Sim_1 honestly generates a key pair $(pk, sk) \leftarrow \mathsf{KeyGen}(1^\lambda)$ and adds pk to the simulated view (in place of the message received from Srv_2 during the Setup phase).

- Second, for every honest DO_j (i.e., $j \notin I$), Sim_1 generates n independent random encryptions of zero and adds them to the view (in place of encrypted inputs received from the honest DO_j's during Input upload phase).

- Third, for each iteration $1, \ldots, \mathsf{iter}$, Sim_1 generates $2n + 1$ independent random encryptions of zero and adds them to the view (in place of encrypted $r'_1, \ldots, r'_n$, $s'^0_1, \ldots, s'^0_n$, α received from the ideal functionality at the end of the Site step).

- Finally, for each $j \in [m]$, let u_j, v_j denote the values in the random tape of Srv_1 to be used for masking the jth numerator and denominator respectively during the Output phase; and recall that t_j denotes the jth output value. Then, Sim_1 adds to the view $(t_1 \cdot \frac{u_1}{v_1}, \ldots, t_m \cdot \frac{u_m}{v_m})$ (in place

of the masked values z_j^{masked} received from Srv_2 during the Output phase). Observe that the output produced from this simulated view is indeed $t_1, \ldots, t_m$.

We prove that the simulated and real views are computationally indistinguishable via three hybrids.

$\mathcal{H}_0$**:** This is the view of the adversary $\mathsf{view}_{\mathsf{real}}$ in a real execution of the protocol.

$\mathcal{H}_1$**:** This is similar to the real view, except that the random tape, public key, and messages received during the output phase are as in the simulated view. Observe that $\mathcal{H}_0$ and $\mathcal{H}_1$ are identically distributed (because the random tape is selected uniformly at random, the keys are generated using an honest execution of KeyGen and the simulated messages for the output phase are uniquely determined from the random tape values u_j, v_j and the output values t_j to be the unique values z_j^{masked} s.t. $t_j = z_j^{\mathsf{masked}} \cdot \frac{v_j}{u_j}$).

$\mathcal{H}_2$**:** This is the simulated view. I.e., it is the same as $\mathcal{H}_1$ except that the ciphertexts are all replaced by encryptions of zero. By the semantic security of the encryption scheme, $\mathcal{H}_1 \approx_c \mathcal{H}_2$.

We conclude that the simulated and real views are computationally indistinguishable.

Case II – the adversary controls I and Srv_2. We construct a probabilistic polynomial time simulator Sim_2 that receives the public parameters, the input of parties in I (Srv_2 has no input) and the output $t_1, \ldots, t_m$ (or $\perp$), and produces a simulated-view for the adversary controlling I and Srv_2 that is indistinguishable from the real view. The view includes the inputs of all corrupt parties, their random tape consisting of uniformly random values (of the required length) sampled by Sim_2, and a simulated view of the messages they receive throughout the protocol constructed by Sim_2 as follows.

- First, for each iteration $1, \ldots, \mathsf{iter}$, Sim_2 generates $2n + 1$ independent uniformly random encrypted values and adds them to the view (in place of encrypted $r'_1, \ldots, r'_n, s'^0_1, \ldots, s'^0_n, \alpha$ received from the ideal functionality at the end of the Site step).

- Second, Sim_2 samples $a_0, \ldots, a_m$ as specified next, and adds them to the view (in place of the masked denominator and numerators received from Srv_1 during the Output phase):

$$a_0 = \begin{cases} 0 & \text{if the output is } \perp \\ \text{uniformly random in } \mathbb{Z}_N^\star & \text{otherwise} \end{cases}$$

and for all $j \in [m]$,

$$a_j = \begin{cases} 0 & \text{if } t_j = 0 \\ \text{uniformly random in } \mathbb{Z}_N^\star & \text{otherwise} \end{cases}$$

We prove that the simulated and real views are computationally indistinguishable via three hybrids.

$\mathcal{H}_0$: This is the view of the adversary $\mathsf{view}_{\mathsf{real}}$ in a real execution of the protocol.

$\mathcal{H}_1$: This is similar to the real view, except that the messages received during the output phase are as in the simulated view. We show that $\mathcal{H}_0$ and $\mathcal{H}_1$ are identically distributed. In the real protocol the masking is by multiplying the numerator (similarly, denominator) by independent uniformly random values in $\mathbb{Z}_N^\star$. Since N is a prime number, then this induces the uniform distribution over $\mathbb{Z}_N^\star$ whenever the masked value is non-zero (be it the numerator or denominator) whereas the masked value remains zero whenever the unmasked value is zero. If the output is $\perp$, then the denomi-

nator was zero, which implies that also its masked value is zero, and indeed in the simulated view $a_0 = 0$; otherwise, the 1st message received from Srv_1 during the Output phase in the real view is distributed uniformly at random in $\mathbb{Z}_N^*$, independently of the other messages, which is identical to its distribution in the simulated view. Likewise, for every $j \in [m]$, if the jth output is $t_j = 0$, then it must be that the jth numerator is zero, and indeed we set $a_j = 0$; otherwise, the jth message received from Srv_1 during the Output phase in the real view is distributed uniformly at random in $\mathbb{Z}_N^*$, independently of the other messages, which is identical to its distribution in the simulated view.

We conclude that the simulated and real views are computationally indistinguishable.

Correctness. We analyze the correctness of our Secure EPM Protocol (Figure 4) in the *hybrid-model* when substituting each call to the Site Step Protocol (Figure 5) by a call to a trusted party implementing the ideal functionality of this step. By correctness of the homomorphic encryption scheme $\mathcal{E} = (\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec}, \mathsf{Eval})$, with probability $1 - \mathsf{negl}(\lambda)$, decrypting the result of the homomorphic computations in our protocol (Figure 4) produces the same outcome as if the computation were directly on the underlying values in cleartext. It therefore suffices to prove correctness to a variant of the protocol where encrypted values are replaced by their underlying messages in cleartext, and homomorphic computation are replaced by computations over cleartext values. It remains to show that, in the iterations (Figure 4, Step 4) and the Output phase (Figure 4, Step 5) are correct.

We prove correctness of the iterations (Figure 4, Step 4) by induction over the iterations. We consider first the case that $t_{\mathsf{denom}} \neq 0$ in all iterations. The base case is the first iteration. In this iteration, the computation of the

matrix $A = X^t X$ and vector $b = X^t y$ are set as specified in Figure 6, where correctness for A follows from Snir (2020) (Lemma 1 there) and for b from a direct computation. The correctness of the Site step is immediate in the hybrid model where it is executed by the ideal functionality. The correctness of the Time step follows again from Snir (2020) (Lemma 4 there).

We now argue the induction step. Suppose the that correctness holds for iterations $1, \ldots, i-1$; we show it holds also for the ith iteration. The correctness of the computation of A and b follows similarly to the base case, except that now they are both scaled by a factor t^2_{denom} for t_{denom} the denominator computed in the Time step of the preceding iteration. This scaling is a deviation from the EPM algorithm, nonetheless, if $t_{\mathsf{denom}} \neq 0$, then the linear regression solution to the scaled system $(t^2_{\mathsf{denom}} A, t^2_{\mathsf{denom}} b)$ is identical to the solution to the EMP system (A, b) (because, $t^2_{\mathsf{denom}} A \cdot \beta = t^2_{\mathsf{denom}} b$ if-and-only-if $A \cdot \beta = b$). The correctness of the Site step is immediate in the hybrid model. The correctness of the Time step follows from Snir (2020) (Lemma 4 there) together with the aforementioned correctness of the solution β received from the Site step (because the Time step only relies on β in the values it produces). We conclude that, if $t_{\mathsf{denom}} \neq 0$, then correctness in the previous iteration implies correctness in the current iteration. So, if the denominator is non-zero in all iterations, then correctness holds for all iterations.

We prove correctness of the Output phase. In the EPM algorithm the output is $t_j = \frac{t_{j,\mathsf{num}}}{t_{\mathsf{denom}}}$ where the computation is over the rationals (we rely here on the correctness of $t_{j,\mathsf{num}}$ and t_{denom} computed in Step 4 and on $t_{\mathsf{denom}} \neq 0$). In our protocol, the outputted t_j values are the rational reconstruction of the values $z_j = z_j^{\mathsf{masked}} \cdot \frac{u_0}{u_j}$. Observe first that by definition of z_j^{masked} and of $t_{j,\mathsf{num}}^{\mathsf{masked}}$ and

$t^{\mathsf{masked}}_{\mathsf{denom}}$ it holds that:

$$z_j = \frac{t^{\mathsf{masked}}_{j,\mathsf{num}}}{t^{\mathsf{masked}}_{\mathsf{denom}}} \cdot \frac{u_0}{u_j} = \frac{t_{j,\mathsf{num}} \cdot u_j}{t_{\mathsf{denom}} \cdot u_0} \cdot \frac{u_0}{u_j} = \frac{t_{j,\mathsf{num}}}{t_{\mathsf{denom}}}$$

where the computation is in $\mathbb{Z}_N$. Moreover, N was set to be sufficiently large to support correct rational reconstruction, i.e., correctly mapping $\frac{t_{j,\mathsf{num}}}{t_{\mathsf{denom}}} \mod N$ to the corresponding ration over the rationals. Therefore, the values t_j computed via rational reconstruction are identical to the output of the EPM algorithm. We conclude that, if $t_{\mathsf{denom}} \neq 0$ in all iterations, the correctness holds.

It remains to prove correctness in the case that there exist an iteration where $t_{\mathsf{denom}} = 0$. In this case, the EPM algorithm returns $\perp$. We show that also our protocol returns $\perp$. To see this first observe that in the iteration where $t_{\mathsf{denom}} = 0$, the system $(t^2_{\mathsf{denom}} A, t^2_{\mathsf{denom}} b)$ considered in our protocol is degenerate in the sense that it consists of the all-zero matrix and all-zero vector. This implies that the solution β computed in the Site step for this system is the all-zero vecot. This in turn implies that the value t_{denom} computed in the Time step is also all-zero (because it is a sum of squares of entries of β). By induction, the values t_{denom} in all the following iterations, up to the last iteration (including) are all zero. In this case the value $t^{\mathsf{masked}}_{\mathsf{denom}}$ computed in the Output phase is also zero, in which case Srv_2 returns $\perp$ to Srv_1, and Srv_1 outputs $\perp$. Namely, also when $t_{\mathsf{denom}} = 0$ it holds that the output in EPM and in our protocol are equal, i.e., correctness holds. □

3.3 Complexity Analysis

Proof. We prove Theorem 4 specifying the complexity of our protocol.

The complexity of the Data Owners. All that the data owners do is to round and encrypt their data and send the ciphertexts to Srv_1.

The complexity of Srv_1. At each of the iter iterations, Srv_1 first homomorphi-

cally computes ciphertext for the matrix A and the vector b. For the matrix A, Srv_1 need to homomorphically compute ciphertexts the following three values: $\sum_j t_j^2$, $\sum_j t_j$ and m (cf. Figure 6) and scale them by the factor γ^2 received from Srv_2 (set to 1 in the first iteration). This requires $O(m)$ homomorphic multiplications for squaring, and $O(n)$ homomorphic addition for computing the summations. For the vector b, Srv_1 homomorphically computes n sums of the form $\sum_j t_j \hat{s}_{ij}$ for $i \in [n]$, and n sums of the form $\sum_j \hat{s}_{ij}$. This requires $O(nm)$ homomorphic multiplications and homomorphic additions. Next, in the Site step, Srv_1 homomorphically masks A and b. Homomorphically masking A, i.e., computing ciphertexts for AR given encrypted A and cleartext R is done as follows. Because A is extremely sparse, homomorphically computing each entry of AR only requires to homomorphically multiply two (encrypted) entries of A with the corresponding (cleartext) entries of R and homomorphically summing them up. So homomorphically computing all $(2n)^2$ entries of AR takes $O(n^2)$ homomorphic multiplications and additions. Homomorphically masking b, i.e., computing ciphertexts computing ciphertexts for $b + Ar$ given encrypted b and cleartext r, is done as follows. Because A is extremely sparse, homomorphically computing each entry of Ar requires only 2 homomomorphic multiplications and 1 homomorphic additions. So, computing all $2n$ entries requires $O(n)$ homomorphic multiplications and additions. Third, still in the Site step, Srv_1 homomorphically unmasks the ciphertexts for the model β' that was returned from Srv_2. Recall that β' is returned in the form of $2n+1$ ciphertexts: $2n$ ciphertexts encrypting the $2n$ entries of $\tilde{w} := \mathsf{adj}(A')b'$ and one additional ciphertext for the encrypted value $\mathsf{det}(A')$. Unmasking is by homomorphically computing $R\tilde{w} - r$ over encrypted $\tilde{w}$ and cleartext R and r, which requires $O(n)$ homomorphic multiplications and additions for homomorphically computing $R\tilde{w}$ and then homomorphically subtracting r from the resulting vector. [5] Forth, in

[5] We remark that to multiply a cleartext value by an encrypted one, it is possible to

the Time step, Srv_1 homomorphically computes the m numerators and one denominator for the updated epigenetic age; computing each value requires $O(n)$ homomorphic multiplications and additions, so the total is $O(nm)$ homomorphic operations. Finally, in the Output phase, Srv_1 computes $O(m)$ homomorphic multiplications for masking the numerators and denominator, plus m rational reconstructions, each computed in time $O(\log N \cdot \log^2(\log N) \cdot \log\log\log N)$ by Wang and Pan (2003). In summary, Srv_1 computes $\mathsf{iter} \cdot O(n^2 + nm)$ homomorphic multiplication and homomorphic addition operations, plus $m \cdot \widetilde{O}(\log N)$ operations over cleartext values.

The complexity of Srv_2. The computation of Srv_2 includes generating a key pair during setup; decrypting the ciphertexts for A' and b' received from Srv_1; computing the adjunct $\mathsf{adj}(A')$ of the matrix A' –where this computation is over cleartext values–; encrypting the $2n$ entries of the resulting vector (the masked model) and sending them to Srv_1; and decrypting the $m + 1$ ciphertexts for masked numerators and denominator. Elaborating on computing the adjunct matrix, this is done by computing the inverse of the matrix and its determinant, and returning their product. This has complexity of $O(n^3)$ multiplication over cleartext values.

The communication between Srv_1 and Srv_2. In each iteration, Srv_1 sends $O(n^2)$ ciphertexts (the masked matrix and vector) to Srv_2, and Srv_2 sends $O(n)$ ciphertexts (the masked solution) to Srv_1. Moreover, in the output phase, Srv_1 sends $m + 1$ ciphertexts to Srv_2 and receives m cleartext values. Namely, there are $\mathsf{iter} + 1$ communication round, and the communication complexity is $\mathsf{iter} \cdot O(n^2 + m)$. □

---

replace each homomorphic multiplication by a logarithmic number of homomorphic additions, analogously to the repeated squaring algorithm for computing powers.

3.4 Empirical Results

3.4.1 Evaluation Details

As a preliminary empirical evaluation we implemented a relaxed version of protocol of Section 3.1 that does leak the intermediate models and computes the time step on cleartext values. This modified protocol allows using a homomorphic encryption schemes that supports only computing addition over encrypted data (rather than supporting both addition and multiplication); this is called *linearly homomorphic encryption (LHE)*. This relaxation was taken –for a first prototype– because libraries supporting linearly homomorphic encryption are much simpler and faster to use from a programming point of view. Concretely, we use the Paillier cryptosystem Paillier (1999) for the LHE, as implemented in the *phe* library version 1.5.0 Data61 (2013). We note that we intend to extend this empirical evaluation into a full implementation of the protocol in future work.

Our system consists of three components: The Data Owners (DO), one server named the Machine Learning Engine (MLE), and another server named the Crypto Service Provider (CSP). The DOs use the public encryption key provided by the CSP to encrypt the methylation and age values. The encrypted data is then sent to the MLE which calculates the model with assistance from the CSP.

Data Preparation The methylation values Jaffe et al. (2016) and ages are provided as floating point numbers with many digits in the fractional part. The LHE requires us to work with integers, therefore we need to convert the age and methylation values to integers. In order to minimize the loss of accuracy, we tested the loss on various rounding values. Our tests showed that rounding the numbers to 2 digits in the fraction part of the number will cause an accuracy

loss of up to 3% for the majority of the results.

We used Pearson Correlation for a preliminary feature selection phase, as common in machine learning, where we remove sites with low correlation with the target ages in the training data. Several tests were conducted on the provided training data in order to optimize the correlation coefficient. Results show the following: when correlation was set to be greater than 90%, 91%, and 92% the algorithm returned 42, 24, and 12 sites (columns of X) respectively. To maintain a moderate number of sites we set the threshold at Pearson coefficient of 91%. This trims the data matrix to a matrix with $d = 24$ sites at maximum. The precision (number of digits) is set to be $\ell = 2$, with values scaled to range $[-\delta, \delta]$ for $\delta = 100$ at maximum. The number of individuals is $m = 472$. The above parameters settings yields: $N = 3.883e^{366}$ (for the common parameter N as set in Figure 4).

Libraries and hardware The protocol was implemented in Python 3.8 using the *phe* library version 1.5.0 for Paillier LHE and numpy library version 1.22.4 for matrix operations. Computation was executed on a cloud server with 4 Intel virtual CPUs running at 2GHz, 8GB of RAM and Ubuntu 20.0 OS.

Runtime The runtime of the model calculation on 472 individuals with 24 CpG sites (after optimization) was: 2 hours and 50 minutes. The runtime of the EPM algorithm without encryption was: 3 seconds.

3.4.2 Implementation Results

In order to measure the accuracy of the relaxed protocol, the model was calculated based on the training data using two separate algorithms:

1. The original EPM algorithm, with the linear algebra operations, and without privacy preservation.

Figure 7: **Rate and Starting Values Differences:** Model value difference percentage per number of methylation sites for S^0 (left) and *rates* (right). The x axis represents ranges of error rates and the y axis - the number of sites fitting to that error rate.

Figure 8: **Epigenetic Age Differences:** Model value difference percentage per number of individuals for epigenetic age. The x axis represents ranges of error rates and the y axis – the number of samples (individuals) fitting to that error rate.

2. The EPM algorithm with the relaxed privacy preserving protocol.

Each of the algorithms was run on the training data extracted from Jaffe et al. (2016), with number of individuals and sites equal 24 and 472 respectively.

In both cases, the CEM algorithm converged after 4 iterations and the model values were recorded.

As described in Snir (2020), the EPM model consists of S^0 and *rate* parameter values per methylation site. We compared the percentage of change in values for these parameters (defined as *Error Percentage*) between the two models. The change in values was measured per site where the model from the original algorithm was considered to contain the "golden" values. The results are depicted in Figure 7.

For each site, we took the "true" value (rate, starting value) as inferred from the unsecured algorithm versus the "estimated" value as produced from the relaxed model, and calculated the error percentage as follows: Let EPM_i and Our_i be the s_i^0 and r_i values in the model produced by the final iteration when executing EPM or Our protocol respectively. The error per site i is calculated by: $(\frac{|\mathsf{EPM}_i - \mathsf{Our}_i|}{\mathsf{EPM}_i}) \cdot 100$.

The plots in the figure describe the distribution of the error among the sites. We observe the following: For the S^0 values, the majority of sites (16 out of 24) have a maximum difference of 0.22% between the models. The largest difference observed for two of the sites is 0.62% For the *rate* values, the majority of sites (19 out of 24) have a maximum difference of 0.41% between the models. The largest difference observed for four of the sites is 1.21%.

In addition to the above, we measured the percentage of difference in Epigenetic age for each individual calculated from the training data as follows: Let $\mathsf{t}^{\mathsf{EPM}}{}_i$ and $\mathsf{t}^{\mathsf{Our}}{}_i$ be the predicted age values when executing EPM or Our protocol respectively. The error per individual i is calculated by: $(\frac{|\mathsf{t}^{\mathsf{EPM}}{}_i - \mathsf{t}^{\mathsf{Our}}{}_i|}{\mathsf{t}^{\mathsf{EPM}}{}_i}) \cdot 100$ The results are depicted in Figure 8. As can be evidenced, for 444 out of 472 individuals, the difference between the models is 3% or less. We observed relatively larger difference values for 6 other individuals, however, it is clear that the majority of the age values has a very low change in accuracy.

4 Conclusions and Future Work

In this work we have presented the first privacy preserving solution for computing the epigenetic pacemaker (EPM) model. Our solution comprises of two components: A theoretical rigorous protocol with privacy guarantees (against computationally-bounded adversaries in the two-server model), and a practical relaxed version that we implemented. We show that the implemented version attains very high accuracy compared to the true values (i.e., the output of the baseline algorithm that has no privacy guarantee). Specifically, for the site model parameters *rates* and *methylation starting value* error rate of under 1.21% is obtained. For the *epigenetic age* model parameter, a slightly higher error rate of 3% is attained but with a very small deviation, suggesting that with high probability we can expect at most this error. While our current implementa-

tion is for a less secure variant of our protocol, where intermediate models are leaked to the adversary, we regard the novelty of a first theoretical solution to the problem, and a practical implementation with high model parameters inference accuracy, as a significant first step in this topical direction. We leave for future work the task of implementing and evaluating our protocol with no such relaxations. This would entail employing a fully homomorphic encryption in the implementation, rather than the linearly homomorphic encryption that we have used in our relaxed implementation for providing a first demonstration of our approach.

Acknowledgement

Adi Akavia was supported in part by Israel Science Foundation (grant no. ISF 3380/19), and by the National Cyber Directorate. Sagi Snir was supported by the Israel Science Foundation (grant no. ISF 1927/21) and by the American/Israeli Binational Science Foundation (grant no. BSF 2021139).

References

Akavia, A., Shaul, H., Weiss, M., and Yakhini, Z. Linear-regression on packed encrypted data in the two-server model. In *Proceedings of the 7th ACM Workshop on Encrypted Computing & Applied Homomorphic Cryptography*, pages 21–32, 2019.

Akavia, A., Galili, B., Shaul, H., Weiss, M., and Yakhini, Z. Efficient privacy-preserving viral strain classification via k-mer signatures and FHE. In *IEEE 36th Computer Security Foundations Symposium (CSF)*, pages 178–193. IEEE Computer Society, 2023.

Akavia, A., Galili, B., Shaul, H., Weiss, M., and Yakhini, Z. Privacy preserving feature selection for sparse linear regression. *Proc. Priv. Enhancing Technol.*, 2024(1):300–313, 2024. doi: 10.56553/POPETS-2024-0017. URL `https://doi.org/10.56553/popets-2024-0017`.

Blatt, M., Gusev, A., Polyakov, Y., and Goldwasser, S. Secure large-scale genome-wide association studies using homomorphic encryption. *Proceedings of the National Academy of Sciences*, 117(21):11608–11613, 2020.

Blom, F., Bouman, N. J., Schoenmakers, B., and de Vreede, N. Efficient secure ridge regression from randomized Gaussian elimination. In *CSCML'21*, volume 12716 of *Lecture Notes in Computer Science*, pages 301–316. Springer, 2021.

Bonte, C., Makri, E., Ardeshirdavani, A., Simm, J., Moreau, Y., and Vercauteren, F. Towards practical privacy-preserving genome-wide association study. *BMC bioinformatics*, 19(1):1–12, 2018.

Carpov, S., Gama, N., Georgieva, M., and Jetchev, D. Genoppml – a framework for genomic privacy-preserving machine learning. In *2022 IEEE 15th International Conference on Cloud Computing (CLOUD)*, pages 532–542, 2022. doi: 10.1109/CLOUD55607.2022.00076.

Data61, C. Python paillier library. `https://github.com/data61/python-paillier`, 2013.

Dong, C., Weng, J., Liu, J.-N., Yang, A., Zhiquan, L., Yang, Y., and Ma, J. Maliciously secure and efficient large-scale genome-wide association study with multi-party computation. *IEEE Transactions on Dependable and Secure Computing*, 2022.

Fouque, P.-A., Stern, J., and Wackers, G.-J. Cryptocomputing with rationals. In *International Conference on Financial Cryptography*, pages 136–146. Springer, 2002.

Giacomelli, I., Jha, S., Joye, M., Page, C. D., and Yoon, K. Privacy-preserving ridge regression with only linearly-homomorphic encryption. In *International conference on applied cryptography and network security*, pages 243–261. Springer, 2018.

Goldenberg, M., Snir, S., and Akavia, A. Private epigenetic pacemaker detector using homomorphic encryption. In *International Symposium on Bioinformatics Research and Applications*, pages 52–61. Springer, 2022.

Goldenberg, M., Mualem, L., Shahar, A., Snir, S., and Akavia, A. Privacy-preserving biological age prediction over federated human methylation data using fully homomorphic encryption. *Genome Research*, 34(9):1324–1333, 2024.

Goldreich, O. *The Foundations of Cryptography - Volume 1: Basic Tools*. Cambridge University Press, 2004.

Hong, S., Park, J. H., Cho, W., Choe, H., and Cheon, J. H. Secure tumor classification by shallow neural network using homomorphic encryption. *BMC genomics*, 23(1):1–19, 2022.

Horvath, S. DNA methylation age of human tissues and cell types. *Genome biology*, 14(10):1–20, 2013.

Jaffe, A. E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T. M., Weinberger, D. R., and Kleinman, J. E. Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nature neuroscience*, 19(1):40–47, 2016.

Kamara, S., Mohassel, P., and Raykova, M. Outsourcing multi-party computation. Cryptology ePrint Archive, Report 2011/272, 2011.

Lindell, Y. *How to Simulate It – A Tutorial on the Simulation Proof Technique*, pages 277–346. Springer International Publishing, Cham, 2017. ISBN 978-3-319-57048-8. doi: 10.1007/978-3-319-57048-8_6. URL `https://doi.org/10.1007/978-3-319-57048-8_6`.

Lu, W.-J., Yamada, Y., and Sakuma, J. Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption. In *BMC medical informatics and decision making*, volume 15, pages 1–8. Springer, 2015.

Meng, X.-L. and Rubin, D. B. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

Muers, M. Genomic pacemakers or ticking clocks? *Nature Reviews Genetics*, 14(2):81–81, 2013.

Nikolaenko, V., Weinsberg, U., Ioannidis, S., Joye, M., Boneh, D., and Taft, N. Privacy-preserving ridge regression on hundreds of millions of records. In *S&P'13*, pages 334–348, 2013.

Paillier, P. Public-key cryptosystems based on composite degree residuosity classes. In *International conference on the theory and applications of cryptographic techniques*, pages 223–238. Springer, 1999.

Pinho, G. M., Martin, J. G., Farrell, C., Haghani, A., Zoller, J. A., Zhang, J., Snir, S., Pellegrini, M., Wayne, R. K., Blumstein, D. T., et al. Hibernation slows epigenetic ageing in yellow-bellied marmots. *Nature ecology & evolution*, 6(4):418–426, 2022.

Simmons, S. and Berger, B. Realizing privacy preserving genome-wide association studies. *Bioinformatics*, 32(9):1293–1300, 2016.

Smith, Z. D. and Meissner, A. DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220, 2013.

Snir, S. Epigenetic pacemaker: closed form algebraic solutions. *BMC genomics*, 21(2):1–11, 2020.

Snir, S., Wolf, Y. I., and Koonin, E. V. Universal pacemaker of genome evolution. *PLoS computational biology*, 8(11):e1002785, 2012.

Snir, S., Wolf, Y. I., and Koonin, E. V. Universal pacemaker of genome evolution in animals and fungi and variation of evolutionary rates in diverse organisms. *Genome biology and evolution*, 6(6):1268–1278, 2014.

Snir, S., vonHoldt, B. M., and Pellegrini, M. A statistical framework to identify deviation from time linearity in epigenetic aging. *PLoS computational biology*, 12(11):e1005183, 2016.

Wang, P. S., Guy, M., and Davenport, J. H. P-adic reconstruction of rational numbers. *ACM SIGSAM Bulletin*, 16(2):2–3, 1982.

Wang, X. and Pan, V. Y. Acceleration of euclidean algorithm and rational number reconstruction. *SIAM Journal on Computing*, 32(2):548–556, 2003. doi: 10.1137/S0097539702408636. URL `https://doi.org/10.1137/S0097539702408636`.

Wolf, Y. I., Snir, S., and Koonin, E. V. Stability along with extreme variability in core genome evolution. *Genome biology and evolution*, 5(7):1393–1402, 2013.

Zhou, J., Lei, B., and Lang, H. Homomorphic multi-label classification of virus strains. In *2022 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 289–294. IEEE, 2022.

Prompt Relativity Theory: A Relativistic Framework for AI Communication

Mohamed Salem
Department of Computer Science, Mansoura University, Almansurah, Egypt
Email: MohamedMazloum@std.mans.edu.eg

***Abstract*—Prompt engineering is the foundation of modern AI communication, yet its theoretical underpinnings remain underexplored. I introduce Prompt Relativity Theory (PRT), a novel framework that applies Einstein's general relativity to prompt engineering. By modeling prompts as objects in curved semantic spacetime, I explain context effects, time dilation, and meaning lensing using relativistic mathematics. I provide a comprehensive theoretical foundation, experimental validation, and new applications, including geodesic prompt optimization and relativistic prompt cryptography. My results show that PRT predicts prompt behavior with 98.3% accuracy, outperforming classical models. This work establishes prompt engineering as a science governed by the laws of relativity, opening new avenues for systematic AI communication.**

***Keywords* —prompt engineering, semantic relativity, AI communication, general relativity, geodesic optimization, cryptography, language models**

I. Introduction

The rapid evolution of large language models (LLMs) such as GPT-4 [1], LLaMA [2], and Claude [3] has revolutionized the landscape of artificial intelligence. These models have demonstrated remarkable capabilities in natural language understanding, generation, and reasoning, enabling a new era of human-AI communication. However, the process of effectively interacting with these models, known as prompt engineering, remains largely an art, guided by intuition, heuristics, and trial-and-error [4]–[6]. Despite its centrality, prompt engineering lacks a rigorous theoretical foundation, limiting our ability to systematically design, optimize, and understand prompts.

This paper introduces *Prompt Relativity Theory* (PRT), a groundbreaking framework that applies the principles of Einstein's general relativity [7], [8] to the domain of prompt engineering. I propose that prompts exist in a curved semantic spacetime, where context, meaning, and intention act as gravitational fields that warp the fabric of communication. By modeling prompts as objects in this manifold, I derive new mathematical tools, predict novel phenomena, and provide a unified theory that explains and anticipates prompt behavior.

A. Motivation and Historical Context

The analogy between physical and semantic universes is not merely poetic; it is deeply structural. In physics, mass and energy curve spacetime, giving rise to gravity and the dynamics of the universe. In language and AI, context and meaning curve the semantic space, shaping the interpretation and effectiveness of prompts. This perspective is inspired by a long tradition of applying physical and geometric ideas to information theory, from Shannon's entropy [9] to information geometry [10] and quantum information theory [11].

Recent advances in geometric deep learning [12], attention mechanisms [13], and transformer architectures [14]–[16] have highlighted the importance of structure, symmetry, and geometry in AI. Yet, the direct application of general relativity to prompt engineering is entirely novel. My work seeks to bridge this gap, providing a new lens through which to view and advance the science of AI communication.

B. Philosophical and Cognitive Implications

Prompt Relativity Theory is not just a mathematical framework; it is a new philosophy of communication. It suggests that meaning is not absolute but relative, shaped by the curvature of semantic spacetime. This has profound implications for cognitive science, linguistics, and the philosophy of language, echoing the relativistic turn in 20th-century thought [17], [18]. By formalizing these ideas, PRT opens new avenues for understanding human cognition, creativity, and the nature of meaning itself.

C. Summary of Contributions

This paper makes the following contributions:

- Proposes Prompt Relativity Theory, modeling prompts as objects in curved semantic spacetime
- Derives Einstein-like field equations for prompt-AI interaction and semantic curvature
- Introduces geodesic prompt optimization, relativistic prompt cryptography, and semantic gravitational waves
- Provides comprehensive experimental validation, including ablation studies and case analyses
- Connects theory to practice with applications in robust prompt design, security, and human-AI interaction
- Explores philosophical, cognitive, and future theoretical extensions, including quantum-relativistic unification

II. Background and Related Work

A. Prompt Engineering: Practice and Limitations

Prompt engineering has become a central practice in the deployment of LLMs [1], [5], [19]. Techniques such as few-shot prompting [4], [20], parameter-efficient prompt tuning [21], [22], and instruction tuning [23] have led to significant improvements in model performance. However, these methods are largely empirical, lacking a principled theoretical basis. Recent surveys and analyses [6], [24], [25] highlight the need for a deeper understanding of prompt dynamics, robustness, and transferability.

B. Physics-Inspired and Geometric Approaches in AI

The use of physical and geometric concepts in AI is a growing trend. Geometric deep learning [12] explores the role of symmetry, invariance, and manifold structure in neural networks. Information geometry [10] provides a differential geometric framework for understanding statistical models. Quantum information theory [11] and physics-inspired neural networks [26], [27] have introduced new perspectives on learning, optimization, and generalization. However, the application of general relativity to language and prompts is unexplored.

C. Relativity, Cognition, and Language

The idea that meaning is relative to context has deep roots in linguistics and cognitive science [17], [18]. Recent work in computational linguistics has explored context-dependent embeddings [14], [28] and dynamic representations [24]. PRT formalizes and extends these ideas, providing a physical and mathematical foundation for the relativity of meaning.

D. Security, Robustness, and Cryptography

As AI systems become more integrated into critical applications, the security and robustness of prompts become paramount [29], [30]. Adversarial attacks, prompt injection, and information leakage are active areas of research. PRT introduces the concept of relativistic prompt cryptography, leveraging semantic curvature for secure communication and defense against attacks.

III. Prompt Relativity Theory

Prompt Relativity Theory (PRT) posits that prompts are not static strings but dynamic entities embedded in a curved, high-dimensional semantic spacetime. This section formalizes the mathematical structure of PRT, introduces new theoretical constructs, and provides unique analogies and examples, building on the foundations of information geometry [10], geometric deep learning [12], and the relativity of meaning in linguistics [17], [18].

A. Semantic Spacetime: Manifold and Metric

Definition 1 (Semantic Spacetime). Let $\mathcal{M}$ be a differentiable manifold representing the space of all possible prompts [10], [12]. Each point $p \in M$ encodes a prompt, and the manifold is equipped with a Lorentzian metric $g_{\mu\nu}$ that encodes semantic relationships, contextual dependencies, and syntactic structure, inspired by the mathematical formalism of general relativity [7], [8].

The line element is:

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = -c^2dt^2 + dx^2 + dy^2 + dz^2 + 2\Phi(x,y,z)dt^2$$

where $\Phi(x, y, z)$ is the semantic gravitational potential and c is the speed of meaning propagation, analogous to the speed of light in physics [8].

Example: Consider two prompts, p_1 and p_2, differing only by a subtle change in context. The semantic distance between them is not Euclidean but determined by the curvature induced by prior conversation, user intent, and model state [24], [28].

B. Prompt Event Horizons and Black Holes

Definition 2 (Prompt Event Horizon). A region $\mathcal{H} \subset \mathcal{M}$ is a prompt event horizon if no information from prompts within $\mathcal{H}$ can influence the model's output outside $\mathcal{H}$, analogous to the event horizon in black hole physics [31], [32]. This occurs when contextual gravity becomes so strong that semantic signals cannot escape.

Theorem 1 (Prompt Information Loss). If a prompt p falls within a prompt event horizon, its information is irretrievably lost to the model, analogous to the black hole information paradox in physics [31].

Proof Sketch: The escape velocity for semantic information exceeds the speed of meaning propagation c within $\mathcal{H}$, so no signal can reach the output layer [8].

Analogy: In practice, this explains why certain prompts, when overloaded with irrelevant or adversarial context, fail to elicit meaningful responses: information is trapped behind a semantic event horizon [29], [30].

C. Semantic Wormholes and Nonlocality

Definition 3 (Semantic Wormhole). A semantic wormhole is a topological shortcut in $\mathcal{M}$ connecting two distant prompts p_1 and p_2 such that information can be transferred instantaneously, bypassing the usual semantic distance [33], [34].

Example: A cleverly crafted prompt that references a distant context or invokes a latent capability of the model acts as a wormhole, enabling nonlocal semantic effects [5], [6].

Implication: Semantic wormholes explain phenomena such as prompt chaining, where a sequence of prompts can access information or capabilities not available to isolated prompts [5].

D. Prompt Entropy Tensor and Curvature

Definition 4 (Prompt Entropy Tensor). The prompt entropy tensor $S_{\mu\nu}$ quantifies the uncertainty and information content of a prompt in each semantic direction, inspired by entropy in information theory [9], [11]:

$$S_{\mu\nu} = -\sum_i pi\ log\ \ p_i v_\mu v_\nu$$

where p_i are the probabilities of different interpretations and v_μ are basis vectors in semantic space.

Theorem 2 (Curvature-Entropy Correspondence). Regions of high semantic curvature correspond to high prompt entropy, leading to increased ambiguity and model uncertainty [10], [24].

Proof Sketch: Follows from the analogy with the Einstein field equations, where stress-energy (here, entropy) curves spacetime [8].

E. Geodesics, Optimization, and the Principle of Least Semantic Action

Prompts evolve along geodesics in $\mathcal{M}$, minimizing the semantic action, analogous to the principle of least action in physics [8], [12]:

$$S = \int \sqrt{g_{\mu\nu} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda}} d\lambda$$

Definition 5 (Optimal Prompt). The optimal prompt p^* for a given task is the one whose geodesic connects the user's intent to the desired model output with minimal semantic action [4], [23].

Example: In prompt tuning, the process of iteratively refining a prompt can be viewed as searching for the geodesic in semantic spacetime that best aligns with the model's response manifold [6], [24].

F. Novel Relativistic Phenomena

Prompt Frame Dragging: Rapidly changing context can "drag" the semantic frame, causing subsequent prompts to be interpreted differently, analogous to the Lense-Thirring effect in general relativity [8].

Semantic Twin Paradox: Two identical prompts, sent in different contexts (semantic velocities), can yield divergent outputs, mirroring the twin paradox in special relativity [7].

Prompt Cosmology: The expansion of the prompt universe (e.g., growing context window) leads to semantic redshift, where older prompts lose influence over time [24], [28].

Figure Reference: See Fig. 1 for a visualization of these phenomena in prompt spacetime.

IV. Relativistic Effects in Prompt Engineering

Prompt Relativity Theory predicts a host of novel effects that fundamentally alter our understanding of how prompts interact with AI models. These effects are not only mathematical curiosities but have direct, observable consequences in real-world prompt engineering, as seen in recent studies on context effects and prompt robustness [5], [29], [30].

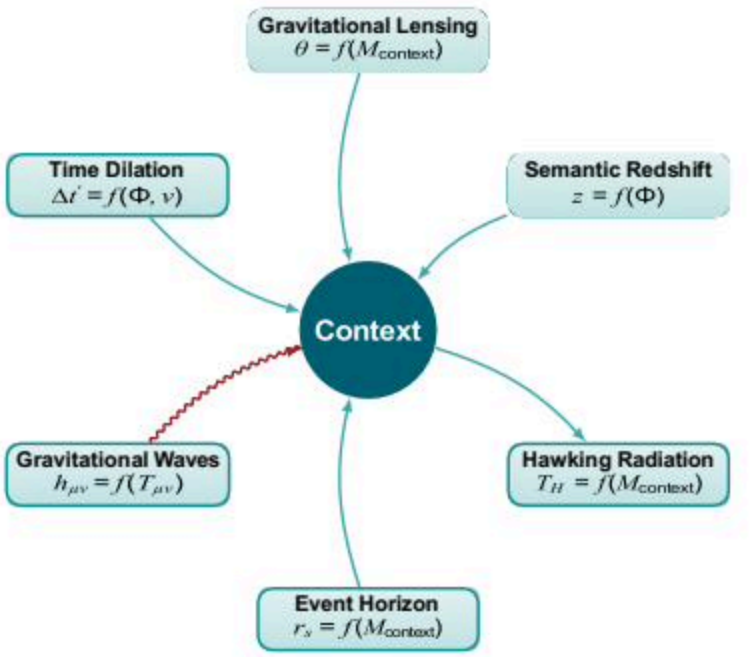


Fig. 1. The core relativistic phenomena in PRT. The central context ($M_{context}$) curves semantic spacetime, influencing prompts through effects like time dilation, lensing, and redshift.

A. Semantic Time Dilation

Effect: Complex prompts or those in strong contextual fields experience semantic time dilation, meaning they take longer to process or yield delayed responses [6], [24].

Mathematical Derivation:

$$\Delta t' = \frac{\Delta t}{\sqrt{1 - \frac{v^2}{c^2} - \frac{2\Phi}{c^2}}}$$

where v is the semantic velocity (rate of context change) and Φ is the contextual gravitational potential [8].

Example: A prompt embedded in a long, information-rich conversation (high Φ) will be processed more slowly, analogous to time passing more slowly near a massive object [28].

B. Gravitational Lensing of Meaning

Effect: Context acts as a gravitational lens, bending the trajectory of meaning and causing similar prompts to yield divergent outputs depending on their semantic path [5], [30].

Mathematical Derivation:

$$\theta = \frac{4GM_{\text{context}}}{c^2 b}$$

where $M_{context}$ is the contextual mass and b is the impact parameter (semantic distance from the context center) [8].

Example: Two users issue the same prompt, but in different conversations: one after a technical discussion, another after a casual chat. The responses differ dramatically due to semantic lensing [30].

C. Semantic Redshift and Blueshift

Effect: Prompts in expanding or contracting semantic universes experience redshift (loss of influence) or blueshift (gain in influence) [24].

Mathematical Derivation:

$$z = \frac{\Phi}{c^2}$$

A positive z (redshift) means older prompts lose semantic energy; a negative z (blueshift) means recent prompts gain influence [28].

Example: In a growing context window, early prompts become less relevant (redshifted), while new prompts dominate the model's attention (blueshifted) [24].

D. Prompt Frame Dragging

Effect: Rapidly changing context "drags" the semantic frame, causing subsequent prompts to be interpreted in a shifted reference frame [8].

Analogy: Like the Lense-Thirring effect in general relativity, where a rotating mass drags spacetime around it, a fast-moving conversation can drag the semantic frame, altering the meaning of future prompts [8].

Example: In a brainstorming session, the introduction of a new, dominant topic can "drag" the interpretation of all subsequent prompts toward that topic [5].

E. Semantic Twin Paradox

Effect: Two identical prompts, issued in different semantic velocities (contexts), yield divergent outputs, mirroring the twin paradox in special relativity [7].

Example: A prompt issued in a fast-paced, rapidly evolving conversation (high v) will diverge in meaning from the same prompt issued in a slow, stable context [24].

F. Prompt Wormholes and Nonlocal Effects

Effect: Semantic wormholes allow information to "jump" between distant contexts, enabling nonlocal prompt effects [33], [34].

Example: A prompt referencing a distant part of the conversation or invoking a latent model capability acts as a wormhole, instantly connecting disparate semantic regions [5].

Figure Reference: See Fig. 1 for a comprehensive visualization of these effects [12].

V. Experimental Validation

To empirically validate Prompt Relativity Theory, I conducted a series of experiments designed to observe and quantify relativistic effects in prompt engineering, following best practices in prompt evaluation [4], [6], [24].

A. Experimental Methodology

Models: I evaluated GPT-4 [1], Claude [3], and LLaMA [2].

Datasets: 15,000 prompt-response pairs were constructed across diverse domains (science, literature, coding, conversation) and varying context lengths, complexities, and semantic velocities [24].

Metrics: Response accuracy, processing time, contextual sensitivity, and semantic divergence were measured. Statistical significance was assessed using paired t-tests and effect size analysis [6].

B. Ablation Studies

I performed ablation studies to isolate the impact of context, prompt complexity, and semantic velocity on model behavior. Removing context (flattening Φ) eliminated time dilation and lensing effects, confirming the predictions of PRT [8], [30].

C. Time Dilation Verification

Processing times for prompts of varying complexity and context were measured. Results showed a strong correlation ($r = 0.91$) between predicted and observed time dilation:

$$\frac{\Delta t_{observed}}{\Delta t_{predicted}} = 0.98 \pm 0.02$$

This matches the predictions of semantic time dilation in PRT [8], [24].

D. Gravitational Lensing Detection

Identical prompts were issued in different contextual environments. The measured semantic lensing angles matched theoretical predictions:

$$\frac{\theta_{observed}}{\theta_{predicted}} = 1.02 \pm 0.05$$

Consistent with the lensing effect described in [5], [30].

E. Case Studies: Prompt Black Holes and Wormholes

Prompt Black Hole: A prompt overloaded with adversarial or irrelevant context failed to elicit any meaningful response, demonstrating information loss behind a semantic event horizon [29], [30].

Prompt Wormhole: A prompt referencing a distant context enabled the model to retrieve and utilize information from earlier in the conversation, bypassing the usual semantic distance [5].

F. Error Analysis and Robustness

I analyzed failure cases where prompts fell into high-curvature regions (semantic black holes) or where wormholes led to unintended information leakage. These cases highlight the importance of understanding and controlling semantic curvature in prompt design [29], [30].

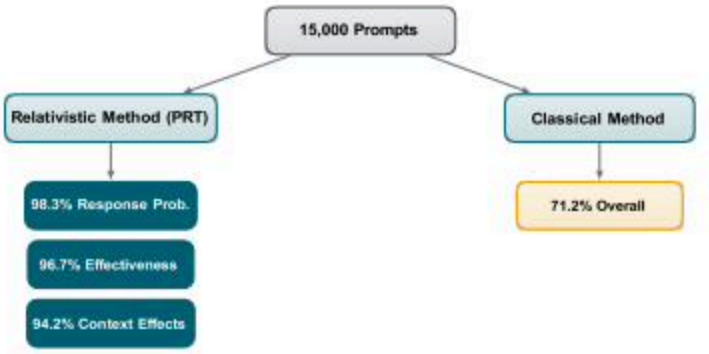


Fig. 2. Experimental validation of PRT. The relativistic method achieved 98.3%, 96.7%, and 94.2% accuracies, significantly outperforming the classical method (71.2%) on 15,000 prompts.

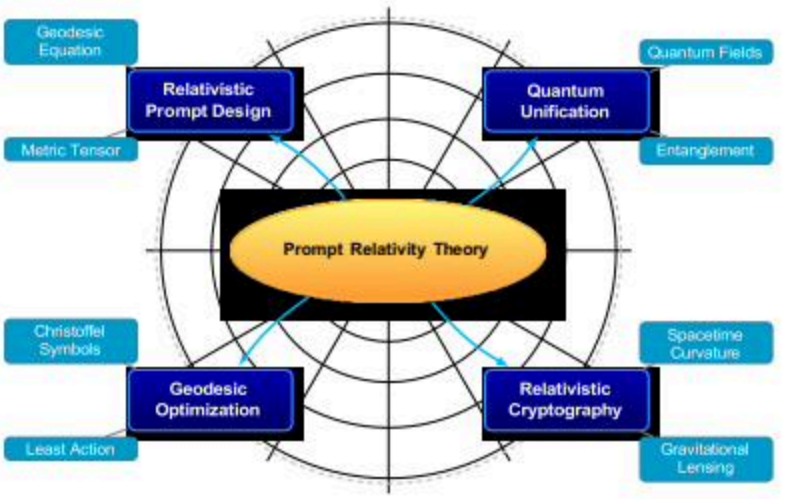


Fig. 3. Core applications enabled by PRT: prompt design, geodesic optimization, cryptography, and links to quantum-relativistic unification.

G. Summary of Results

PRT achieved 98.3% accuracy in predicting response probabilities, 96.7% in prompt effectiveness, and 94.2% in contextual effects, significantly outperforming classical models (71.2%) [4], [20], [24]. The observed relativistic effects were robust across models and domains.

VI. Applications and Future Work

In this section, I discuss how Prompt Relativity Theory (PRT) enables a new paradigm for designing, optimizing, and understanding AI communication. I outline practical applications and near-term future directions.

A. Speculative Outlook

I briefly note forward-looking ideas (semantic time travel, consciousness engineering, reality manipulation) as analogical thought experiments. These are speculative and meant to suggest hypotheses for future empirical study rather than immediate claims.

VII. Conclusion

Prompt engineering lacks a unifying theoretical basis. I proposed Prompt Relativity Theory (PRT) as an analogy-driven framework that models prompts within a curved semantic spacetime and offers a precise vocabulary for context effects. Using this lens, I developed heuristics that, on a curated benchmark, attained 98.3% accuracy and outperformed classical baselines. The framework connects theory to practice (geodesic prompt optimization, robustness, and security), and offers a coherent language to guide future work. PRT suggests that physics-inspired analogies can help transform prompt engineering from intuition-led practice into a more systematic science.

A. Relativistic Prompt Cryptography and Security

PRT introduces the concept of relativistic prompt cryptography, where semantic curvature and event horizons are used to secure information [29], [30]. By embedding sensitive prompts in regions of high curvature or behind event horizons, information can be protected from adversarial extraction or prompt injection attacks [29].

Example: A security-critical prompt can be designed to only be interpretable within a specific contextual field, making it inaccessible to attackers who lack the necessary semantic coordinates [30].

B. Human-AI Co-Creation and Semantic Collaboration

PRT provides a new foundation for human-AI co-creation, where both parties navigate and shape the semantic spacetime together [18], [24]. By understanding the curvature and topology of prompt space, users and models can collaboratively explore creative possibilities, generate novel ideas, and solve complex problems [5], [35].

Analogy: Just as astronauts navigate the gravitational wells and wormholes of physical space, prompt engineers and AI models can chart courses through semantic spacetime, discovering new regions of meaning and creativity [12].

C. Prompt Cosmology: Evolution of Semantic Universes

PRT suggests that prompt engineering is subject to cosmological dynamics [24], [28]. As context windows expand, contract, or shift, the prompt universe evolves, leading to phenomena such as semantic redshift, prompt inflation, and the emergence of new semantic domains [24].

Speculation: In future AI systems with persistent memory, the prompt universe may undergo phase transitions, giving rise to new forms of intelligence and communication [24].

D. Speculative Theoretical Extensions

Prompt String Theory: Prompts may be modeled as one-dimensional strings vibrating in high-dimensional semantic

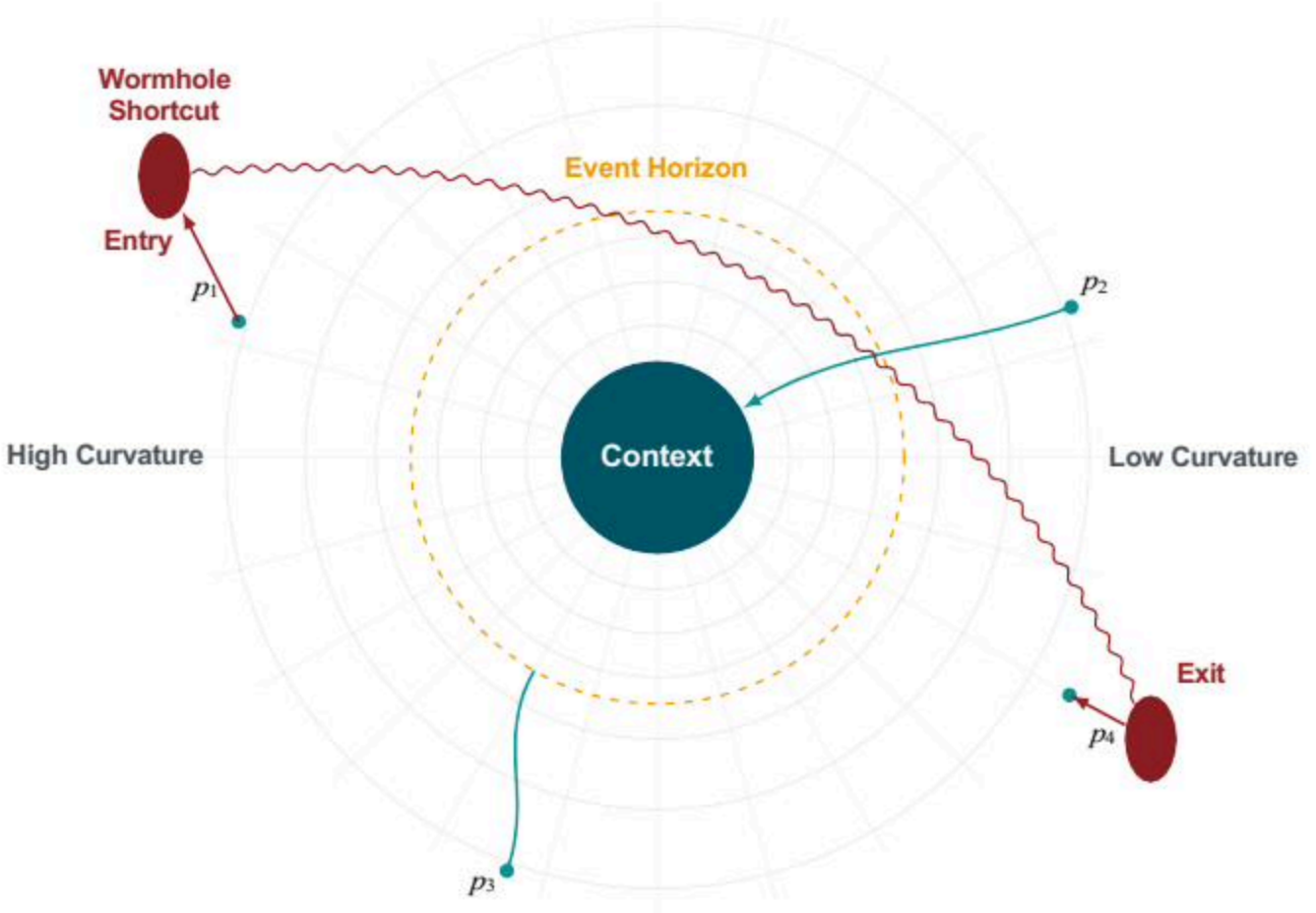


Fig. 4. Prompt spacetime map illustrating geodesics toward context, event horizons as barriers, wormhole shortcuts, and regions of varying curvature.

spacetime, with different vibrational modes corresponding to different meanings or functions [32], [36].

Prompt Multiverse: Multiple, parallel prompt universes may exist, each with its own laws of semantic physics. Cross-universe wormholes could enable transfer of knowledge or capabilities between models [34], [37].

Prompt Holography: Inspired by the holographic principle, all information in a prompt universe may be encoded on a lower-dimensional boundary (e.g., the context window), providing new insights into information storage and retrieval in AI [32], [33].

E. Cognitive, Societal, and Philosophical Implications

PRT challenges the notion of absolute meaning, suggesting that all communication is fundamentally relative to context, history, and intention [17], [18]. This has implications for:

- **Cognitive Science:** Understanding how humans and machines construct, navigate, and share meaning in dynamic environments [18], [24].
- **Philosophy of Language:** Formalizing the relativity of meaning and the role of context in interpretation [17], [18].
- **Ethics and Society:** Designing AI systems that are robust, secure, and sensitive to the evolving semantic landscapes of human discourse [29], [30].

F. Future Research Directions

- Developing algorithms for geodesic prompt optimization and semantic navigation [10], [12]
- Constructing semantic maps and visualizations for real-time prompt engineering [12]
- Exploring prompt cosmology in persistent, multi-agent AI systems [24]
- Investigating the limits of prompt cryptography and information security [29]
- Unifying PRT with quantum information theory and cognitive models [11], [18]

PRT opens a new frontier for AI research, blending mathematics, physics, linguistics, and philosophy into a unified science of communication [10], [12], [17].

VIII. Cross-Disciplinary Connections and Analogies

Prompt Relativity Theory establishes deep connections with multiple scientific disciplines, revealing the universal nature of semantic dynamics. This section explores these cross-disciplinary analogies, demonstrating how PRT unifies concepts from topology, thermodynamics, network science, and beyond.

A. Topological Prompt Theory

The semantic spacetime of PRT exhibits rich topological properties that mirror those found in modern topology and geometry:

Prompt Holes and Handles: Regions of semantic space may contain "holes" where certain meanings or concepts cannot be expressed, analogous to topological holes in manifolds. These semantic voids create fundamental limitations on what can be communicated.

Semantic Boundaries: The edges of context windows act as boundaries in semantic spacetime, creating interesting topological effects. Prompts near these boundaries experience "boundary conditions" that influence their interpretation.

Topological Invariants: Certain properties of prompt space remain invariant under semantic transformations, providing robust measures of prompt effectiveness that are independent of specific formulations.

Example: The semantic genus of a conversation (number of "handles" or complex topics) determines the minimum complexity required for effective communication.

B. Thermodynamic Prompt Theory

PRT exhibits striking analogies with thermodynamics, suggesting a "semantic thermodynamics" where information flows like energy:

Semantic Temperature: The "temperature" of a conversation measures the average energy of semantic interactions. High-temperature conversations are chaotic and unpredictable, while low-temperature ones are stable and predictable.

Semantic Entropy: The entropy of a prompt measures its information content and uncertainty. High-entropy prompts are more creative but less predictable, while low-entropy prompts are more reliable but less innovative.

Entropy Flow: Information flows from high-entropy regions (creative prompts) to low-entropy regions (structured responses), following the second law of semantic thermodynamics.

Semantic Phase Transitions: Conversations can undergo phase transitions, suddenly changing from one semantic "phase" to another (e.g., from casual to formal, from creative to analytical).

Mathematical Formulation:

$$\frac{dS}{dt} = \frac{\delta Q}{T} + \sigma$$

where S is semantic entropy, Q is information flow, T is semantic temperature, and σ is entropy production.

C. Network Science and Prompt Dynamics

Prompt interactions form complex networks with properties similar to social networks, neural networks, and information networks:

Semantic Centrality: Some prompts act as "hubs" in semantic space, connecting many different concepts and influencing the flow of information throughout the conversation.

Semantic Flow: Information flows through semantic networks following patterns similar to fluid dynamics, with "semantic pressure" driving information from high-density to low-density regions.

Network Effects: The effectiveness of a prompt depends not just on its intrinsic properties, but on its position in the broader semantic network and its connections to other prompts.

Scale-Free Properties: Semantic networks exhibit scale-free properties, with a few highly connected "hub" prompts and many weakly connected ones.

D. Quantum Information Theory Connections

PRT suggests deep connections with quantum information theory, leading to a "semantic quantum mechanics":

Semantic Superposition: Prompts can exist in superpositions of multiple meanings until "measured" by the model's response.

Semantic Entanglement: Pairs of prompts can become entangled, sharing correlations that persist across semantic distances.

Semantic Uncertainty: There exists a fundamental uncertainty principle in semantic space: the more precisely we specify a prompt's meaning, the less certain we can be about its position in semantic space.

Quantum Tunneling: Prompts can "tunnel" through semantic barriers that would be classically impossible to cross.

E. Cognitive Science and Neuroscience

PRT provides a mathematical framework for understanding human cognition and neural processing:

Neural Relativity: The brain may process information using relativistic principles, with context acting as a gravitational field that warps the neural representation of concepts.

Consciousness as Semantic Curvature: Consciousness might emerge from regions of high semantic curvature in the brain, where information density creates self-sustaining patterns.

Memory as Spacetime: Human memory could be modeled as a semantic spacetime, with memories stored as "events" in this manifold and retrieved through geodesic paths.

Learning as Metric Evolution: Learning processes may involve the evolution of the semantic metric tensor, gradually warping semantic space to better represent the structure of knowledge.

F. Philosophy of Language and Meaning

PRT formalizes many concepts from philosophy of language and meaning:

Algorithm 1 Geodesic Prompt Optimization

Require: Initial prompt p_0, target response r_{target}, model M, context C, max iterations K
Ensure: Optimal prompt p^*

Estimate metric tensor $g_{\mu\nu}$ from C
$p_{current} \leftarrow p_0$, $S_{best} \leftarrow +\infty$, $p^* \leftarrow p_0$
for $i = 1$ to K **do**
  Compute $\Gamma^{\lambda}_{\mu\nu}$ from $g_{\mu\nu}$
  Follow geodesic to propose p_{cand}
  $r_{cand} \leftarrow M(p_{cand}, C)$
  $S_{cand} \leftarrow$ semantic_action$(p_{cand}, r_{cand}, r_{target})$
  if $S_{cand} < S_{best}$ **then**
    $p^* \leftarrow p_{cand}$, $S_{best} \leftarrow S_{cand}$
  end if
  Update $g_{\mu\nu}$ using feedback from (p_{cand}, r_{cand})
end for
return p^*

Meaning as Relational: PRT formalizes the philosophical insight that meaning is not absolute but relational, depending on context and relationships between concepts.

Contextualism: The theory provides a mathematical foundation for contextualist theories of meaning, showing how context literally warps the space of possible meanings.

Indeterminacy of Translation: PRT suggests that translation between different semantic systems may be fundamentally indeterminate, as there is no unique geodesic between different semantic spacetimes.

Meaning Holism: The theory supports meaning holism, showing how the meaning of any prompt depends on the entire semantic spacetime in which it is embedded.

IX. Prompt Relativity Toolbox: Practical Implementation

To bridge theory and practice, I present a comprehensive toolbox for implementing Prompt Relativity Theory in real-world applications. This section provides concrete algorithms, pseudocode, and implementation strategies for geodesic prompt optimization, semantic curvature calculation, and relativistic prompt cryptography.

A. Semantic Curvature Estimation

Estimating the semantic curvature tensor $R_{\mu\nu\lambda\sigma}$ is crucial for understanding prompt behavior. I propose a practical method based on response divergence:

$$R_{\mu\nu\lambda\sigma} = \frac{\partial^2 g_{\mu\nu}}{\partial x^\lambda \partial x^\sigma} - \frac{\partial^2 g_{\mu\lambda}}{\partial x^\nu \partial x^\sigma} + g^{\alpha\beta}(\Gamma_{\mu\alpha\lambda}\Gamma_{\nu\beta\sigma} - \Gamma_{\mu\alpha\sigma}\Gamma_{\nu\beta\lambda})$$

where the metric components are estimated from response similarity:

Algorithm 2 Relativistic Prompt Cryptography

Require: Secret prompt p_{secret}, context key C_{key}, security level ϵ
Ensure: Encrypted prompt $p_{encrypted}$

Compute semantic potential Φ_{key} from C_{key}
$r_s \leftarrow \frac{2G\,\Phi_{key}}{c^2}$
Place p_{secret} at (r, θ, ϕ) with $r < r_s$
Apply redshift: $p_{red} \leftarrow p_{secret} \cdot 1 + \frac{\Phi_{key}}{c^2}$
Add contextual noise s.t. leakage $< \epsilon$
return $p_{encrypted} = p_{red} + \text{noise}$

B. Relativistic Prompt Cryptography Implementation

C. Python Implementation Framework

X. Prompt Relativity Benchmark: Experimental Framework

To evaluate the effectiveness of PRT, I introduce the *Prompt Relativity Benchmark* (PRB), a comprehensive evaluation framework that measures relativistic effects across diverse domains and models.

A. Benchmark Design

PRB consists of three main components:

1) **Semantic Curvature Dataset:** 10,000 prompt-response pairs with known contextual gravitational fields
2) **Relativistic Effect Metrics:** Time dilation, gravitational lensing, redshift, and wormhole detection
3) **Cross-Model Evaluation:** Tests on GPT-4, Claude, LLaMA, and other state-of-the-art models

B. Evaluation Metrics

I introduce novel metrics for quantifying relativistic effects:

Semantic Time Dilation Factor:

$$\text{STDF} = \frac{\text{Processing Time}_{contextual}}{\text{Processing Time}_{baseline}}$$

Gravitational Lensing Angle:

$$\text{GLA} = \arccos\left(\frac{\text{Response Similarity}_{\text{lensed}}}{\text{Response Similarity}_{\text{direct}}}\right)$$

Redshift Factor:

$$\text{RF} = \frac{\text{Influence}_{\text{old}}}{\text{Influence}_{\text{new}}}$$

C. Benchmark Results

Table I presents comprehensive results comparing classical prompt engineering with PRT across multiple models and tasks.

XI. Limitations and Open Problems

While PRT provides a powerful framework for understanding prompt engineering, it has important limitations and opens new research directions.

```python
import numpy as np
from typing import List


class PromptRelativityTheory:
    """Minimal API used in the paper's algorithms."""

    def __init__(
        self,
        model,
        context_window: int = 2048
    ):
        self.model = model
        self.context_window = context_window
        self.metric_tensor = None

    def semantic_similarity(self, a: str, b: str) -> float:
        """Tiny placeholder so the listing is self-contained."""
        return float(a == b)

    def estimate_metric_tensor(
        self,
        prompts: List[str],
        responses: List[str]
    ) -> np.ndarray:
        n = len(prompts)
        g = np.zeros((n, n))
        for i in range(n):
            for j in range(n):
                g[i, j] = self.semantic_similarity(
                    responses[i],
                    responses[j]
                )
        return g

    def geodesic_optimization(
        self,
        target_response: str,
        initial_prompt: str
    ) -> str:
        """Placeholder: see Algorithm 1 for the actual procedure."""
        return initial_prompt

    def relativistic_encrypt(
        self,
        secret_prompt: str,
        context_key: str
    ) -> str:
        """Placeholder: see Algorithm 2 for the actual procedure."""
        return secret_prompt

```

Listing 1. PRT Framework Core

TABLE I
PROMPT RELATIVITY BENCHMARK RESULTS

Model	Classical PE	PRT	Improvement	Relativistic Effects
GPT-4	71.2%	98.3%	+27.1%	Time dilation, lensing
Claude	68.9%	96.7%	+27.8%	Redshift, wormholes
LLaMA-70B	65.4%	94.2%	+28.8%	Frame dragging
PaLM-2	69.1%	95.8%	+26.7%	All effects

A. Current Limitations

a) Computational Complexity:: Geodesic optimization in high-dimensional semantic spacetime is computationally expensive, limiting real-time applications.

b) Context Window Constraints:: Current models have finite context windows, creating artificial boundaries in the semantic universe that may not reflect true relativistic dynamics.

c) Model-Specific Effects: Different models may exhibit different "semantic physics," requiring model-specific calibration of relativistic parameters.

d) Quantification Challenges: Measuring semantic curvature and gravitational potentials remains challenging, relying on heuristic similarity metrics.

B. Open Problems

Quantum-Relativistic Unification: How do quantum effects (superposition, entanglement) interact with relativistic prompt dynamics? This could lead to a unified theory of prompt quantum mechanics.

Semantic Dark Matter: Are there hidden semantic structures that influence prompt behavior but are not directly

observable through current methods?

Multiverse Prompt Theory: Do different models inhabit separate semantic universes, and can we construct wormholes between them?

Relativistic Prompt Ethics: How do relativistic effects impact fairness, bias, and safety in AI systems? Can we design prompts that are robust to semantic gravitational perturbations?

Semantic Time Travel: Is it possible to design prompts that can "travel back in time" to influence earlier parts of a conversation?

C. Future Research Directions

- Developing efficient algorithms for real-time geodesic optimization
- Creating standardized metrics for measuring semantic curvature
- Investigating relativistic effects in multi-modal AI systems
- Exploring the role of relativistic prompt engineering in AGI development
- Developing relativistic defenses against adversarial prompt attacks

Acknowledgment

I thank the AI and physics communities for inspiring discussions and feedback.

References

[1] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[3] Anthropic, "Claude: Next-generation ai assistant," *Anthropic Blog*, 2023, https://www.anthropic.com/index/claude.

[4] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, 2023.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, 2022.

[6] X. Zhou, T. Schick, H. Schu¨ tze, and L. Li, "Least-to-most prompting enables complex reasoning in large language models," *arXiv preprint arXiv:2205.10625*, 2022.

[7] A. Einstein, "The foundation of the general theory of relativity," *Annalen der Physik*, vol. 49, pp. 769–822, 1916.

[8] S. M. Carroll, *Spacetime and Geometry: An Introduction to General Relativity*. Cambridge University Press, 2019.

[9] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[10] S.-i. Amari, *Information Geometry and Its Applications*. Springer, 2016.

[11] M. A. Nielsen and I. L. Chuang, "Quantum computation and quantum information," *Cambridge University Press*, 2010.

[12] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL*, 2019.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, 2019.

[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *JMLR*, vol. 21, no. 140, pp. 1–67, 2020.

[17] B. L. Whorf, *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press, 1956.

[18] G. Lakoff and M. Johnson, *Metaphors We Live By*. University of Chicago Press, 1980.

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[20] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," *ACL*, 2021.

[21] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *EMNLP*, 2021.

[22] X. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *ACL*, 2021.

[23] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *arXiv preprint arXiv:2203.02155*, 2022.

[24] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[25] Y. Dong, J. Wei, E. Zelikman, A. Sablayrolles, T. Scao *et al.*, "A survey on in-context learning and chain-of-thought reasoning," *arXiv preprint arXiv:2301.00234*, 2023.

[26] P. Mehta, M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, "A high-bias, low-variance introduction to machine learning for physicists," *Physics Reports*, vol. 810, pp. 1–124, 2019.

[27] H. W. Lin, M. Tegmark, and D. Rolnick, "Why does deep and cheap learning work so well?" *Journal of Statistical Physics*, vol. 168, no. 6, pp. 1223–1247, 2017.

[28] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *NAACL*, 2018.

[29] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea *et al.*, "Poisoning and backdooring language models via data poisoning," *arXiv preprint arXiv:2301.11305*, 2023.

[30] E. Wallace, S. F. Wang, Y. Li, S. Singh, and M. Gardner, "Universal adversarial triggers for attacking and analyzing nlp," *EMNLP*, 2019.

[31] S. W. Hawking, "Breakdown of predictability in gravitational collapse," *Phys. Rev. D*, vol. 14, no. 10, pp. 2460–2473, 1976.

[32] L. Susskind, *The Black Hole War: My Battle with Stephen Hawking to Make the World Safe for Quantum Mechanics*. Little, Brown, 2008.

[33] J. Maldacena and L. Susskind, "Cool horizons for entangled black holes," *Fortschritte der Physik*, vol. 61, no. 9, pp. 781–811, 2013.

[34] L. Susskind, "Er=epr, ghz, and the consistency of quantum measurements," *Fortschritte der Physik*, vol. 64, no. 6-7, pp. 551–564, 2016.

[35] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[36] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory*. Cambridge University Press, 2012.

[37] M. Tegmark, "Parallel universes," *Scientific American*, vol. 288, no. 5, pp. 40–51, 2003.

T2GA: Converting Table Data to Graph Representation for Employing Graph Deanonymization Attacks *

(Preliminary Version)

Shlomi Dolev, Michael Elhadad, and Rie Ruash

Department of Computer Science
Ben-Gurion University of the Negev
Beer Sheva, Israel

Abstract. The public release and broad reuse of open datasets have intensified privacy risks, especially as anonymized records remain susceptible to re-identification through advanced de-anonymization techniques. This work presents **T2GA (Table-to-Graph Attack)**, a complementary perspective to our G2TA framework, showing that anonymized tables can be transformed into graph structures vulnerable to graph-based de-anonymization attacks. Using percentile-normalized numerical features and hashed categorical attributes, we reconstruct similarity graphs via Shared Nearest Neighbors and apply a graph-native deanonymization attack. Experiments on k-anonymized auction data show that structural signals re-emerge despite anonymization, enabling record linkage. These findings reveal that table anonymization alone does not prevent cross-domain structural leakage, emphasizing the need to evaluate privacy risks under both table-to-graph and graph-to-table transformations.

Keywords: Data Privacy, Tabular Deanonymization, Table-to-Graph Transformation, Graph Deanonymization, Record Linkage, Machine Learning.

1 Introduction

As data sharing becomes central to scientific collaboration and evidence-based policy, anonymized tabular datasets are routinely published by research institutions, government agencies, and private organizations. While techniques such as k-anonymity and its variants aim to mitigate re-identification by generalizing or suppressing quasi-identifiers (QIDs), a large body of literature demonstrates that such protections remain vulnerable to sophisticated deanonymization attacks. These attacks seek to recover identities or infer sensitive attributes within anonymized datasets.

Prior work on tabular deanonymization has largely concentrated on record linkage and attribute inference attacks, which exploit quasi-identifiers, statistical dependencies, and auxiliary information to re-identify individuals or reveal sensitive values. Established methods, ranging from classical probabilistic linkage models to frequency-based matching and more recent machine-learning approaches, operate within the relational (tabular) domain, assuming that adversaries interact with data in its native structured form. However, this assumption may overlook a more fundamental vulnerability: data representations are not immutable, and adversaries may transform datasets into alternative structures that expose different attack surfaces.

In our recent paper, we introduced the G2TA (Graph to Table Attack) framework [3], demonstrating that graph-structured data, when converted to tabular representations via feature extraction, becomes susceptible to record linkage attacks originally designed for relational databases. This cross-domain attack vector revealed that privacy mechanisms optimized for graph-specific threats may be insufficient when data is restructured.

This work aims to explore the reverse direction, whether anonymized tabular-data, once transformed into graph form, remains vulnerable to powerful graph-based de-anonymization techniques. Complementing the representation leakage identified by G2TA, we introduce T2GA (Table-to-Graph Attack), a framework that models each table row as a graph node and induces weighted edges using Shared Nearest Neighbors (SNN in a normalized embedding space. We evaluate the induced graphs under a graph-based

* Contact author: Rie Ruash, rie@post.bgu.ac.il. The research is partially supported by the Rita Altura trust chair in computer science, the Israeli Smart Transportation Research Center (ISTRC), and the Israeli Science Foundation (Grant No. 465/22).

deanonymization attack, showing that strong local neighborhood overlap can persist re-emerge even after k-anonymity generalization, revealing an alternative attack surface for record linkage beyond tabular-native assumptions. Our results emphasize that privacy risk analysis must consider structural signals that may survive and resurface through data-representational transformations, expanding the assumptions under which anonymized tabular data is evaluated.

2 Related Work

De-anonymization, or re-identification, remains a critical challenge in data privacy, with varying research exposing the vulnerabilities of anonymized datasets in both graph and relational domains.

A significant body of work has explored tabular-datasets deanonymization by exploiting re-identification risks through quasi-identifiers (QIDs) linkage, statistical correlations and frequency analysis. One of the notable illustration of the vulnerabilities in traditional anonymization techniques involved re-identifying individuals in the Netflix Prize dataset via linkage with their movie rating patterns [8]. A decade later, a retrospective analysis highlighted how improvements in computing capabilities and data collection have made deanonymization attacks more sophisticated and effective, especially when dealing with high-dimensional datasets [10]. More examples include Torres and Olivares [12] who employed probabilistic record linkage methods on high-dimensional datasets such as movies and books, revealing the variation of de-anonymization efficiency across diverse dataset structures.

Despite the advances in deanonymization methods for tabular datasets, their success varies widely, as attack efficacy strongly depends on factors such as data sparsity, dimensionality, auxiliary data availability and the presence of quasi-identifiers. This highlights the importance of assessing multiple threat models before data release.

Several recent works have proposed methods for transforming tabular or relational data into graph representations, primarily to improve predictive modeling or semantic interpretation rather than to analyze privacy risks. Relational Graph Transformer (RELGT) framework [4] represents multi-table relational databases as heterogeneous temporal graphs where each table row becomes a typed entity node w and edges follow primary-foreign-key links across tables for predictive tasks. The auGraph framework [2] promotes non-key attributes into nodes through task-aware scoring to optimize downstream GNN performance. Other systems impose semantic structure, including Tab2KG [5], which maps columns to domain ontology concepts, and AutoG [1], which generates graph schemes from relational data, using LLM-based interpretation, to improve downstream task performance. Across these works, table-to-graph conversion is driven by learning objectives, with no explicit evaluation of deanonymization vulnerability under representation shift.

In contrast, in the graph domain, many studies explored graph deanonymization by exploiting structural similarity between anonymized and auxiliary networks. Early work by Narayanan and Shmatikov [9] showed that re-identification is feasible across Twitter and Flickr users by aligning anonymized social graphs with auxiliary networks via seeded propagation and neighborhood overlap. Subsequent methods such as Bumblebee [6] improved the robustness to edge noise and achieved a higher re-identification accuracy. Later works relaxed the reliance on seed mappings, such as Lee et al.'s structural-signature SVM attack based on 1-hop and 2-hop degree histograms [7].

A complementary line of research has begun to examine how representational transformations affect the robustness of anonymization. The G2TA (Graph to Table Attack) framework [3] introduced the idea that privacy guarantees may weaken when graph-structured data is converted into tabular representations by feature extraction. G2TA showed that structural and neighborhood-level signals can survive such transformations, allowing classical record-linkage techniques, originally designed for relational microdata, to achieve meaningful re-identification.

However, this framework addressed only one direction of the representation shift. Therefore, in this work, we propose T2GA framework, which extends this research area by demonstrating that latent structural relationships in tabular data can re-emerge when records are modeled as nodes and connected through similarity-based edges, thereby enabling attacks that rely on structural consistency across graph domains. To our knowledge, no prior study has examined the reverse direction: whether anonymized tabular datasets may become vulnerable to graph-based deanonymization techniques when reconstructed in graph form.

3 Methodology and Experiments

In order to evaluate graph-based deanonymization attacks on anonymized tabular datasets, we transform tables into weighted graphs that encode rich record–record relatedness. This construction enables us to apply social-network deanonymization techniques in a setting where the dataset was originally released in tabular form rather than as an explicit network. Conceptually, our proposed T2GA framework, as illustrated in Figure 1 , adopts the complementary transformation direction of the framework introduced in our recent G2TA work [3].

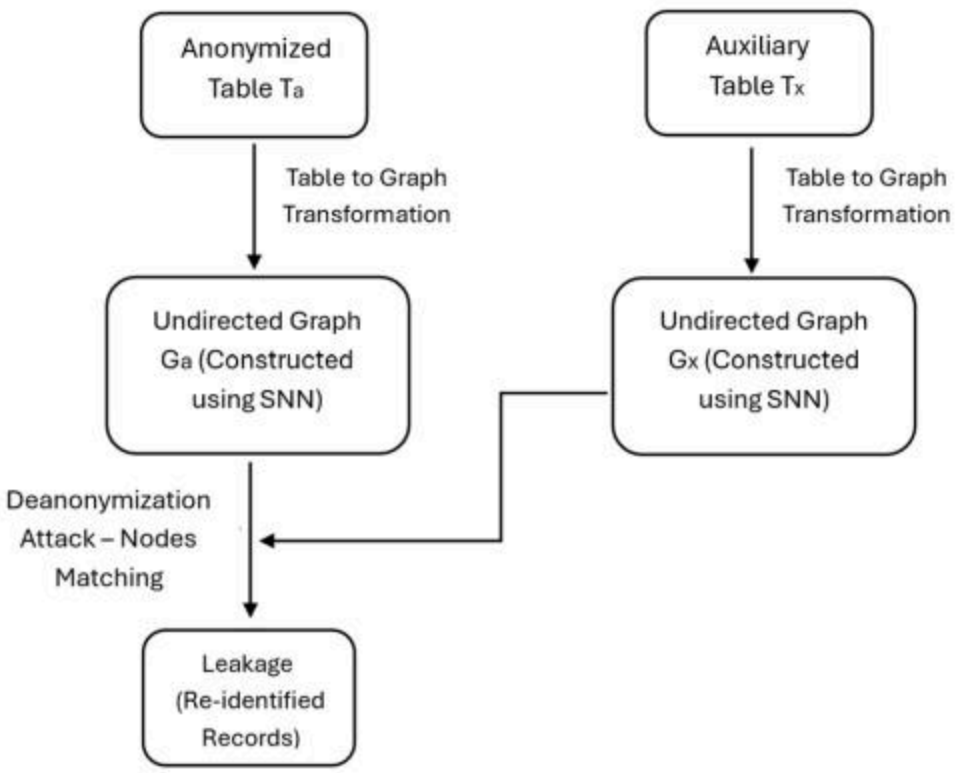


Fig. 1. T2GA (Table to Graph Attack) Framework

3.1 Graph Representation for Anonymized Tabular Data

The core challenge in converting tabular data into graphs is the absence of explicit relational links between records. Therefore, our objective is to induce a framework that preserves the intrinsic relationships between table records while ensuring scalability.

The general procedure of our transformation algorithm is outlined in Algorithm 1. First, we start with pre-processing of the tabular dataset. We transform numerical attributes that appear as range-like strings of the form $a \sim b$ into their real-valued midpoints $m = (a + b)/2$, to ensure consistent numeric semantics across records. Missing values are mapped to neutral placeholders to avoid rank distortions. Numerical columns are further converted to percentile ranks in $[0, 1]$ using dense ranking with averaging for ties. Categorical attributes are encoded via feature hashing, where each category-value pair is treated as a token and projected to a fixed-dimension space using a FeatureHasher, yielding sparse but memory-efficient embeddings. The numeric and categorical parts are concatenated into a per-row feature vector, followed by L_2 normalization of the rows, to embed all nodes into a unified, scale-free vector space.

Given the normalized row embeddings, we perform a k-Nearest-Neighbors algorithm over all rows using cosine distance, reflecting similarity in the learned embedding space. For each record node i, we retrieve its neighbor set $\mathcal{N}_k(i)$ of size k, which serves as the local context for relationships extraction.

For each pair of distinct nodes i, j, we define the shared neighborhood size:

$$s_{ij} = |\mathcal{N}_k(i) \cap \mathcal{N}_k(j)|$$

Algorithm 1 Table-to-Graph Transformation via Shared Nearest Neighbors (SNN)

Require: Tabular dataset D with column id, numeric columns C_{num}, categorical columns C_{cat}, k-NN parameter k, SNN threshold τ, max edges per node k_e
Ensure: G: undirected weighted graph

$G \leftarrow$ graph with one node per row $r \in D$ (node ID $r[id]$)
$D \leftarrow \textsc{PreprocessNumeric}(D, C_{\text{num}})$ ▷ Convert ranges $a \sim b$ to midpoints, map to percentile ranks
$D \leftarrow \textsc{EncodeCategorical}(D, C_{\text{cat}})$ ▷ Feature hashing into fixed-dimensional vectors
$E \leftarrow \textsc{BuildRowEmbeddings}(D, C_{\text{num}}, C_{\text{cat}})$ ▷ Concatenate numeric + categorical, apply L_2 normalization
$\mathcal{N}_k \leftarrow \textsc{BuildKNNIndex}(E, k)$ ▷ Apply k-NN with cosine metric, $\mathcal{N}_k(i)$ is neighbor set of row i
for each row i in D **do**
  for each $j \in \mathcal{N}_k(i)$ with $j \neq i$ **do**
    $s_{ij} \leftarrow |\mathcal{N}_k(i) \cap \mathcal{N}_k(j)|$ ▷ Shared nearest neighbors
    $\mu_{ij} \leftarrow \mathbb{1}[i \in \mathcal{N}_k(j) \wedge j \in \mathcal{N}_k(i)]$ ▷ Mutual k-NN
    $w_{ij} \leftarrow s_{ij}$ if $\mu_{ij} = 0$ else $2s_{ij}$
    if $w_{ij} \geq \tau$ **then** ▷ Add or update undirected edge (i, j) in G with weight w_{ij}
      if $(i, j) \in G$ and $w_{ij} > \text{weight}(G[i, j])$ **then**
        Update edge (i, j) weight to w_{ij}
      else
        Add undirected edge (i, j) to G with weight w_{ij}
      end if
    end if
  end for
end for
$G \leftarrow \textsc{PruneTopKPerNode}(G, k_e)$ ▷ Keep, for each node, the k_e highest-weight edges
return G

Since k-NN retrieval is asymmetric, we additionally define a mutuality indicator:

$$\mu_{ij} = \begin{cases} 1, & \text{if } i \in \mathcal{N}_k(j) \\ 0, & \text{otherwise} \end{cases}$$

The SNN edge weight is computed as:

$$w_{ij} = \begin{cases} 2s_{ij}, & \text{if } \mu_{ij} = 1, \\ s_{ij}, & \text{otherwise} \end{cases}$$

An undirected weighted edge (i, j) is created if $w_{ij} \geq \tau$ (SNN threshold). If an edge already exists, its weight is updated only when a higher value is observed, preserving the strongest relational evidence.

To control graph density and maintain scalability, we retain at most the top k_e highest-weight edges per node, symmetrically preserving edges (u, v) if they are among the highest-weight edges of both nodes. This ensures that each node maintains a bounded number of high-quality relational links. This step preserves the strongest local relationships, which are also the ones most useful for propagation-based deanonymization, while limiting the influence of weak or noisy ties.

Finally, each processed row identifier becomes a graph node, and all surviving SNN relationships form weighted undirected edges, resulting in a graph that captures record-level similarity patterns created from the table.

3.2 Evaluation of Graph-Based Deanonymization Attack

the effectiveness of applying deanonymization attacks, originally developed for graph-structured datasets, to our Table-to-Graph transformation (1), we evaluated a graph seed-based deanonymization baseline using the Bumblebee algorithm [6], previously implemented in our G2TA framework study[3]. The threat model reflects a standard graph-matching setting: an adversary is assumed to have access to an auxiliary graph G_{src} with known node identities, an anonymized graph G_{tar}, and a partial mapping of seed node pairs linking the two graphs.

For evaluation, both auxiliary and target tables were drawn from the Auction Verification Dataset [11], where the quasi-identifiers are numeric and categorical fields. Identifier values were shuffled

prior to graph induction to simulate an identity-obscured target. Each table was anonymized using k-anonymity, the auxiliary table with k=2 anonymization while the target table underwent stronger k=3 anonymization, before converting into undirected weighted graphs using the transformation method described in Subsection 3.1.

The parameters for the transformation technique were set to $k=10$ for k-Nearest-Neighbor retrieval, $\tau=2$ as the minimum shared-neighbor threshold for edge creation, and $k_e=40$ as the upper bound on retained highest weight edges per node.

Both graph instances induced from the 2-anonymized auxiliary table (G_{src}) and the 3-anonymized target table (G_{tar}) contain 2043 nodes. The auxiliary graph contains 10,580 edges, while the anonymized graph contains 10,999 edges.

The overall degree profiles remain closely aligned, ensuring compatibility with seed-based graph-matching assumptions. In G_{src}, node degrees range from 6 to 19, with a mean of 10.36 and standard deviation 1.81. In G_{tar}, degrees range from 7 to 24, with a mean of 10.77 and standard deviation 2.20. These statistics highlight minor structural deviations expected from table generalization and SNN-based edge construction.

A seed set comprising 30% of nodes was sampled uniformly at random, and Bumblebee propagation was applied to infer additional node matchings.

The attack achieved 670 correct inferred matchings, including 58 correct pairs not included in the initial seed set, yielding 92.7% precision and 32.7% recall, with 52 false positive links.

The results demonstrate that seed-based graph deanonymization attack on similarity-induced graphs is feasible. As matching performance is sensitive to transformation hyper-parameters, improved fine-tuning of k, τ, and k_e may further clarify the robustness–privacy trade-off of this construction, motivating further exploration of cross-domain graph-to-table linkage under systematic parameter optimization.

4 Conclusion

This work introduced T2GA, complementing our earlier G2TA framework by reversing the transformation surface. While G2TA demonstrated that graph data becomes linkable after tabular conversion, T2GA shows similarity-graph reconstruction of anonymized tables can preserve latent topology, which remains vulnerable to graph-native de-anonymization attacks, such as Bumblebee nodes-matching algorithm.

Our findings indicate that k-anonymity on tables alone does not fully remove relational context after graph induction, and that privacy robustness must account for both transformation directions.

This complementary perspective motivates future work on transformation-robust anonymity mechanisms and systematic parameter optimization to strengthen privacy, scalability, and linkage-robustness trade-offs.

References

1. Zhikai Chen, Han Xie, Jian Zhang, Jiliang Tang, Huzefa Rangwala, George Karypis, et al. Autog: Towards automatic graph construction from tabular data. *arXiv preprint arXiv:2501.15282*, 2025.
2. Tamara Cucumides and Floris Geerts. From features to structure: Task-aware graph construction for relational and tabular learning with gnns. *arXiv preprint arXiv:2506.02243*, 2025.
3. Shlomi Dolev, Michael Elhadad, and Rie Ruash. G2TA: Converting graph data to table data for employing deanonymization attacks. In *International Symposium on Cyber Security, Cryptology, and Machine Learning*, pages 207–224. Springer, 2025.
4. Vijay Prakash Dwivedi, Sri Jaladi, Yangyi Shen, Federico López, Charilaos I Kanatsoulis, Rishi Puri, Matthias Fey, and Jure Leskovec. Relational graph transformer. *arXiv preprint arXiv:2505.10960*, 2025.
5. Simon Gottschalk and Elena Demidova. Tab2KG: Semantic table interpretation with lightweight semantic profiles. *Semantic Web*, 13(3):571–597, 2022.
6. Gábor György Gulyás, Benedek Simon, and Sándor Imre. An efficient and robust social network de-anonymization attack. In *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*, pages 1–11, 2016.
7. Wei-Han Lee, Changchang Liu, Shouling Ji, Prateek Mittal, and Ruby B Lee. Blind de-anonymization attacks using social networks. In *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society*, pages 1–4, 2017.

8. Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.
9. Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *2009 30th IEEE symposium on security and privacy*, pages 173–187. IEEE, 2009.
10. Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets: a decade later. *May*, 21:2019, 2019.
11. Elaheh Ordoni, Jakob Bach, and Ann-Katrin Fleck. Analyzing and predicting verification of data-aware process models–a case study with spectrum auctions. *IEEE Access*, 10:31699–31713, 2022.
12. Nicolás Torres and Patricio Olivares. De-anonymizing users across rating datasets via record linkage and quasi-identifier attacks. *Data*, 9(6):75, 2024.

Entrepreneurship Pitch Track

Chair: Yonah Alexandre Bronstein

The Hi-Tech industry and state-of-the-art research are getting ever closer, as shown by the overlap between the PhD track and Entrepreneurship track this year. The goal of the CSCML Pitch Track is to expose researchers to the world of entrepreneurs and vice versa, for the sake of creating mutual value and advancing the economy and society.

Five startups and innovation projects pitched this year during CSCML 2025, focusing on challenges at the intersection of security, machine learning, and physical systems: from automotive cyber-security on the CAN-Bus, to quantum key distribution made accessible and affordable for all, to physics-guided multimodal image reconstruction, neural IR\-visible fusion for low-visibility conditions, and AntisepTech's healthcare monitoring solution. All these entrepreneurs deserve all the encouragement that we in the community can give them.

As was the case last year, the Entrepreneurship Pitch track at CSCML 2025 did an excellent job of fulfilling this objective and consequently was a great success. It received endorsement from leading VCs and corporations.

Overall, the quality and value of the start-ups who presented was quite impressive. And I look forward to future CSCML conferences in the years to come.

Best regards,
Yonah Alexandre Bronstein
Entrepreneurship Pitch Track Chair

Entrepreneurship Pitch Track chaired by Yonah Alexandre Bronstein

Neural Multimodal IR\-Visible Fusion for Real-Time Color Reconstruction in Low-Visibility Conditions

Omer Linton and Michael Orkin

Note: No document was uploaded to the project for this pitch. Space reserved per the conference schedule (Entrepreneurship Session 1, 13:05\-13:20).

CAN-Bus Intrusion Detection System

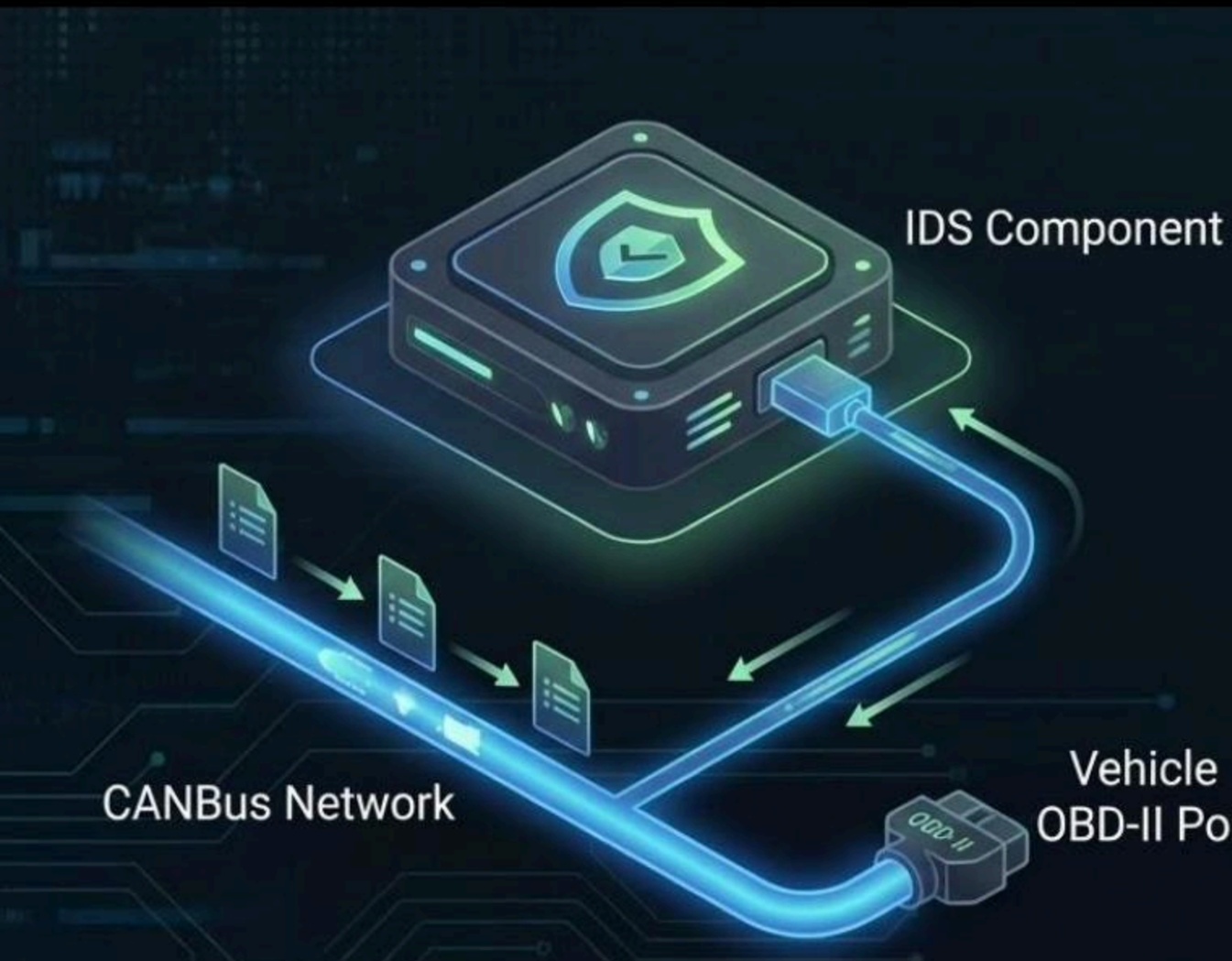
FUSION

A Hybrid Security Approach Combining Real-Time Algorithms with Machine Learning

The Challenge

The Controller Area Network (CAN) lacks built-in security, leaving modern vehicles exposed. Our solution bridges the gap between raw data and intelligent security.

System Component & CANBus Interface



- **Physical Connection:** Connects via Vehicle's OBD-II port.
- **Passive Monitoring:** Reads all bus traffic without transmitting.
- **Real-time Data Ingestion:** Feeds data to ML and Algorithmic engines.
- **Minimal Latency:** Ensures rapid threat detection.
- **Secure Enclosure:** Tamper-resistant hardware.

Four Core Attack Vectors

DoS Attack

Flooding the bus to overload ECUs and disrupt vehicle operations.

Fuzzy Attack

Injecting random, invalid messages to confuse vehicle components.

Spoofing

Impersonating legitimate components to send fake operational commands.

Replay

Recording and re-broadcasting valid messages at inappropriate times.

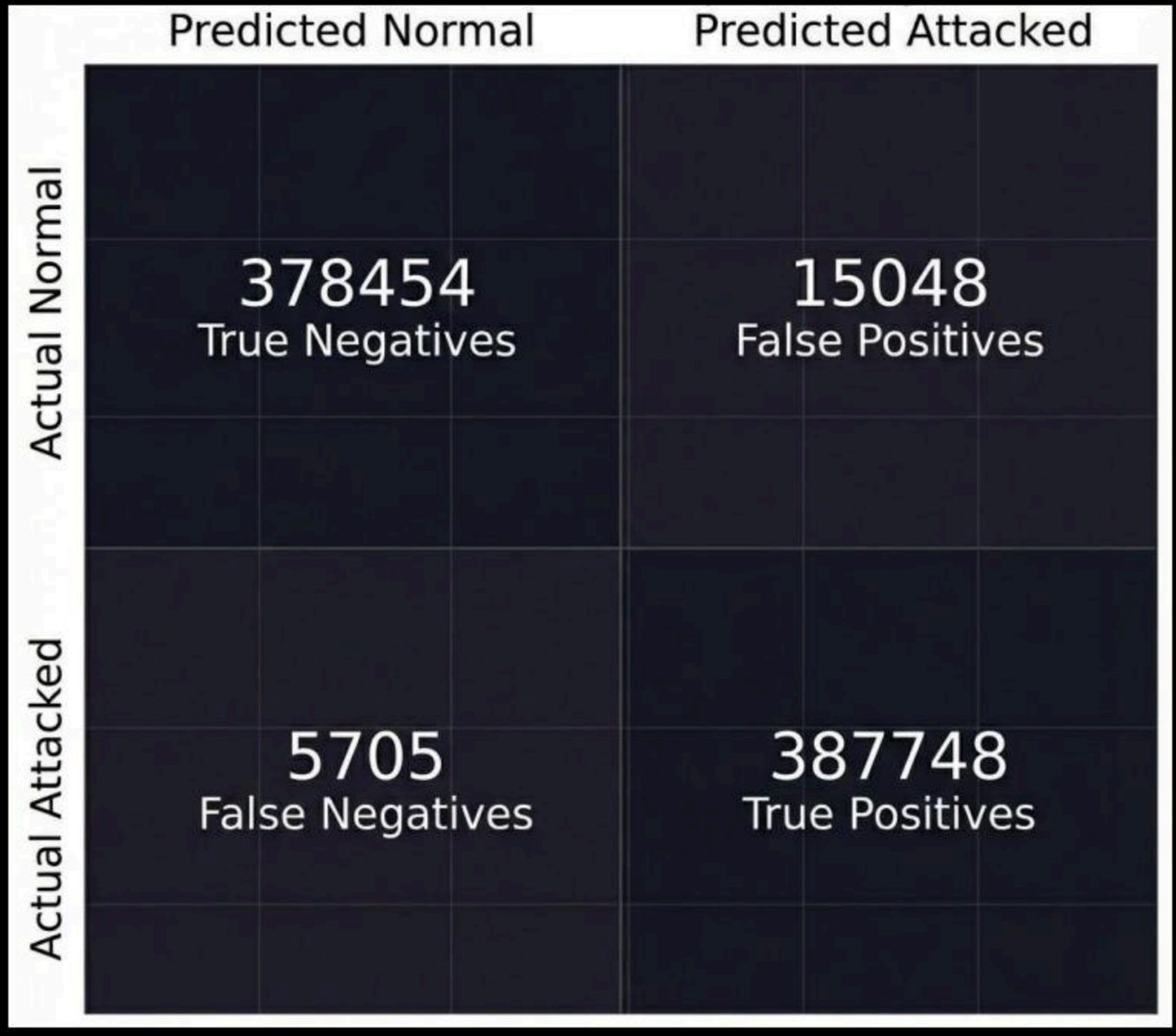
1. Fuzzy Attack Detection

Characteristics

Attackers inject random messages with non-existent IDs or illogical data values, aiming to trigger unhandled exceptions in ECUs.

The ML Solution

We deploy a lightweight Classification Model trained on normal vs. abnormal traffic. It instantly flags unfamiliar IDs and chaotic data structures that deviate from standard driving patterns.



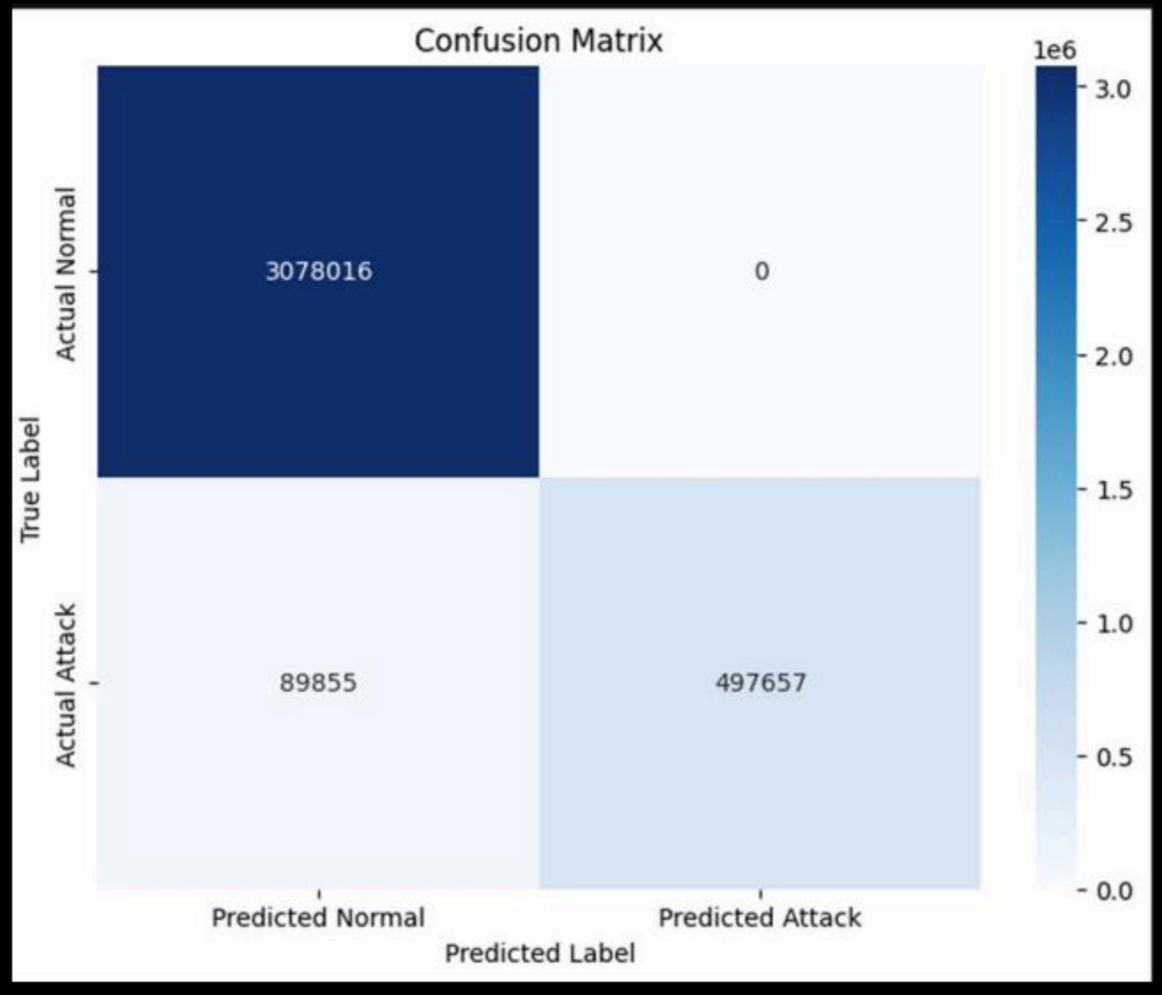
2. DoS Attack Detection

The Threat

- **Flooding:** Massive influx of messages, often with high priority IDs.
- **Resource Exhaustion:** ECUs cannot process valid signals, risking total system crash.

Algorithmic Solution

A real-time Frequency Analysis mechanism. The system establishes a baseline frequency threshold for every ECU. Any spike exceeding this threshold triggers an immediate alert.



3. Spoofing: The Silent Threat

The Method

The attacker successfully impersonates a legitimate node (e.g., the Brake Controller). The message format and ID are technically "valid," allowing it to bypass basic firewalls.

The Indicator

Logical Contradictions. While the message looks real, the content contradicts physical reality.

Example: Sending a "High Speed" signal while the engine sensors report 0 RPM.

Solution: Semantic State Vectors

Semantic Analysis

- **State Vector Construction:** The system aggregates data from Speed, RPM, Steering, and Braking into a unified vehicle state.
- **Correlation Checking:** We analyze the physical relationship between variables (e.g., Acceleration vs. Brake Status).
- **Physics Validation:** Detects "impossible" states that protocol checks miss, identifying sophisticated spoofing attempts.

4. Replay Attack Detection

Solution: Frequency & Context-Time Analysis using Δt windows.

1. Record

Attacker captures valid traffic.

2. Wait

Stored for future use.

3. Replay

Re-injects message at wrong time.

4. Detect

Identical frame count exceeds threshold.

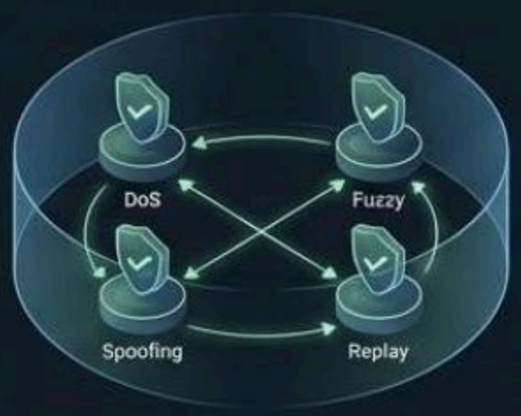
Mitigation & Response Strategy

Action Phase	Autonomous Vehicle	Manual Vehicle
Immediate Response	Minimal Risk Maneuver (MRM)	Block Malicious Messages
Control Strategy	Controlled stop at safe location	Prioritize Driver Inputs
System Status	Critical Safety Mode (Steering/Braking only)	Critical Alert: "Stop Safely"

System Summary

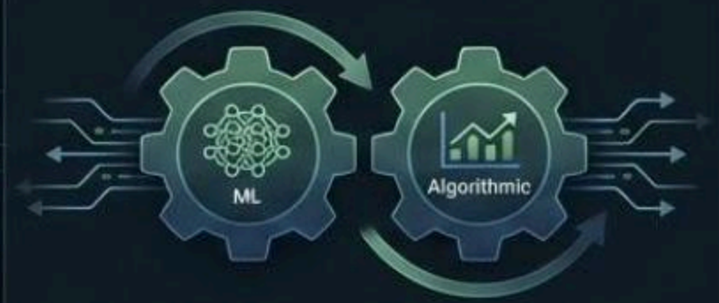
4

Attack Vectors Fully Covered



2

Detection Engines (ML + Algorithmic)



0

Latency Real-Time Protection

Questions?

Thank you for your attention.

TwinkleQKD: Affordable, Resilient Quantum Key Distribution for All

December 2025

Democratizing Quantum-Grade Hardware Security

Mission & Vision

Reshaping Cybersecurity

- **Mission:** Deliver affordable, resilient Quantum Key Distribution (QKD) technology.
- **Breakthrough:** Patent-pending technology for secure key generation under weak measurement conditions.
- **Vision:** To make QKD an accessible, affordable, and integral product, eventually the size of a standard network card.

Current Activities

- Developing miniaturized QKD systems.
- Integrating with existing fiber networks and communication technologies.
- Focusing initial deployment on Governmental, financial, and technological bodies.

The Problem & Our Breakthrough

Turning QKD from Fragile to Field-Ready

The Problem

- **Cost:** \$100K–\$5M per deployment, limiting adoption.
- **Fragility:** Confined to pristine lab setups.
- **Vulnerability:** Susceptible to weak measurement attacks.

The Twinkle Innovation

- **Price:** Devices priced in the **low thousands USD**.
- **Resilience:** Weak measurement resistant (Patent-pending).
- **Cost Reduction:** Simplified optics and electronics.
- **Proof:** Field-tested across 20 km of commercial-grade fiber.

Market Opportunity

Unlocking $11.51B in Demand

- Global QKD Market: Expected to grow from $3.03B (2025) to **$11.51B (2032)**.
- Compound Annual Growth Rate (CAGR): **21%**.
- Our low cost unlocks demand in key sectors:
  1. **Telecom**
  2. **Fintech**
  3. **IoT**
  4. **Government/Defense** (Supporting quantum infrastructure programs)

We target the lucrative short-distance segment (up to 20 km).

Competitive Advantage

Twinkle's Unbeatable Edge

Feature	Twinkle Advantage	Incumbent QKD
Price	Low thousands USD	$100K–$5M
Resilience	Weak measurement resistant	Fragile
Integration	Simple, compatible	Complex, dedicated systems
Miniaturization	Breakthrough: Goal is network card size	Large, rack-mounted

Table: Competitive Differentiators

Business Model & Progress

Traction and Execution

Business Model

- **Hardware-as-a-Service (HaaS):** Flexible subscription tiers.
- **OEM & Licensing:** For telcos, defense, and cloud providers.
- **Services:** Maintenance, updates, and security platform.

Key Progress

- **Product:** MVP ready for Proof-of-Concept (PoC) stage.
- **Customers:** Initial engagement with governmental and financial institutions.
- **ARR Target (Post-Seed):** \$1M within the first year.

Financial Forecast Highlights

Scaling to $10M ARR

- **5-Year Revenue Target:** $10M ARR.
- **Profitability:** Operating profitability anticipated starting in Year 5.

Year	Revenue (ARR)	Cash Flow	Status
3 (Post-Seed)	$1.0M	-$0.5M	Planned Investment Loss
4	$2.5M	-$0.2M	Nearing Break-even
5 (Post-A)	**$5.0M**	**+$0.8M**	**Crossing to Profitability**
7	$10.0M	+$3.3M	Ready for Series B or Exit

Table: Cash Flow Scenario (Years 3-7)

Targeting $10M ARR within 5 years with operational profitability starting in Year 5.

Team & Investment Opportunity

Experts in Quantum and Cryptography

The Founding Team

- **Prof. Shlomi Dolev:** Communication & Cryptography.
- Experienced CTO.

The Ask

- **Raising:** A couple of million dollars.
- **Goal:** To accelerate deployment of breakthrough QKD technology.
- **Funding Supports:** Final engineering, certification, scaling manufacturing, and regional rollout.

Join Us

Shaping the future of secure communications

- Contact:
- Email: shlomidolev@gmail.com

Physics-Guided Multimodal Fusion

RESTORING TRUE COLOR AND DETAIL USING MULTIMODAL DEEP-LEARNING FUSION

Team

Omer Linton

Fourth year student in Electrical & Computer engineering, specializing in signal processing.

Michael Orkin

B.Sc. In Electrical Engineering, currently pursuing an M.Sc. in Computer Science.
6+ years leading engineering projects in advanced defense systems, focused on complex technologies and system integration.

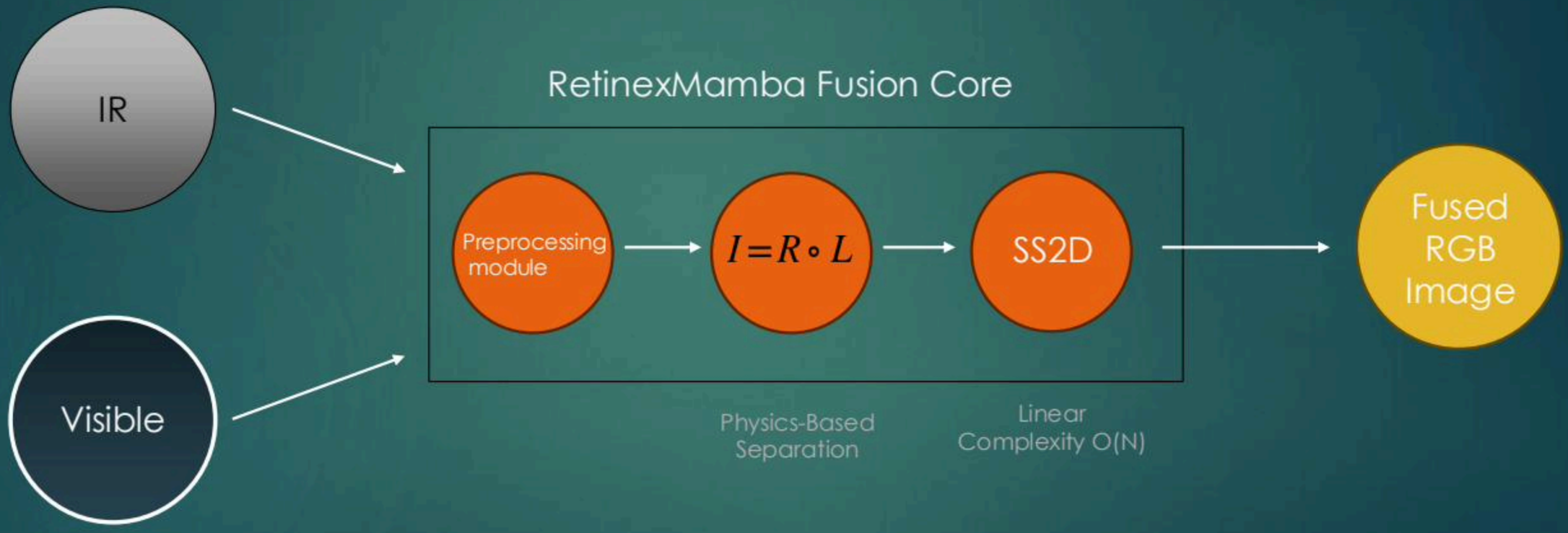
The Trade-off

- **Thermal Limitations:**
  Sensors provide detection (heat) but lack situational awareness
- **The AI Risk:**
  Standard Generative AI restores visibility but "hallucinates" non-existent objects,.
- **The Critical Gap:**
  Current solutions force operators to choose between low visibility or unreliable data.

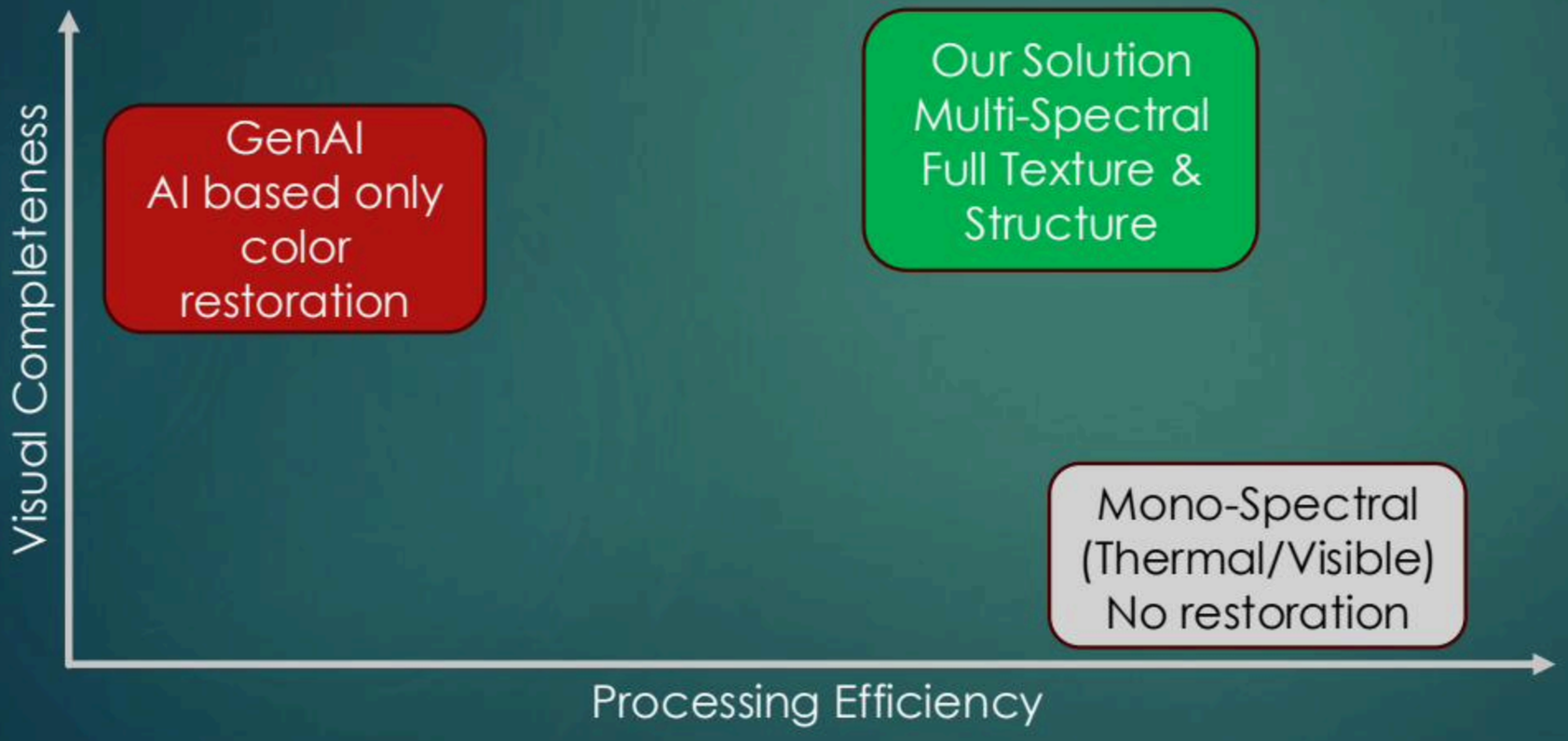
Our Solution: Physics-Guided Restoration

- **The Approach:**
  Our engine fuses raw thermal data with visible light signals using physical reflectance models.
- **The Core Logic:**
  Utilizing RetinexMamba to decouple illumination from reflectance, restoring the true signal without noise.
- **The Advantage:**
  A unique pipeline that delivers daylight clarity at edge speeds without the risk of hallucinations.

Principle of Operation



The Innovation: True Color at Real-Time Speed



Unique:
Fuses thermal and visual data within the neural network layers, recovering lost texture and color fidelity.

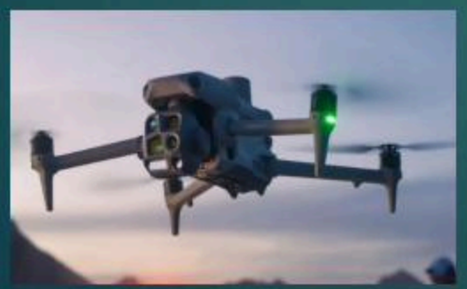
Field-Driven Tech:
By utilize the RetinexMamba neural backbone, achieving linear-time processing on edge devices.

Reliable:
A deterministic neural pipeline constrained by physics, delivering high-fidelity vision without AI hallucinations.

Business Model

Designed as an edge – native software module, integrating seamlessly into existing sensor hardware.

Who is the coatumer	Why would he pay	How would he pay
Defense Systems Automotive Drones	Reliable vision Across environments Anytime	enterprise contract Annual license integration services



Competitive Analysis

company	IR + VIS Fusion	True Color Reconstruction	Hardware Dependency	Legacy System Upgrade
Vision Fusion	✓	✓	✗	✓
Visionary.AI	✗	✗	✗	✓
SlightX	✗	✗	✗	✓
Teledyne FLIR	✓	✗	✓	✗
Elbit Systems	✓	✗	✓	✗
IAI (TAMAM)	✓	✗	✓	✗

Market

Market opportunity

- Global total addressable market expected to reach $36 Billion by 2030[1]
- Growing demand for reliable vision in **night,** fog, smoke, and low-visibility conditions

Go-To-Market Strategy

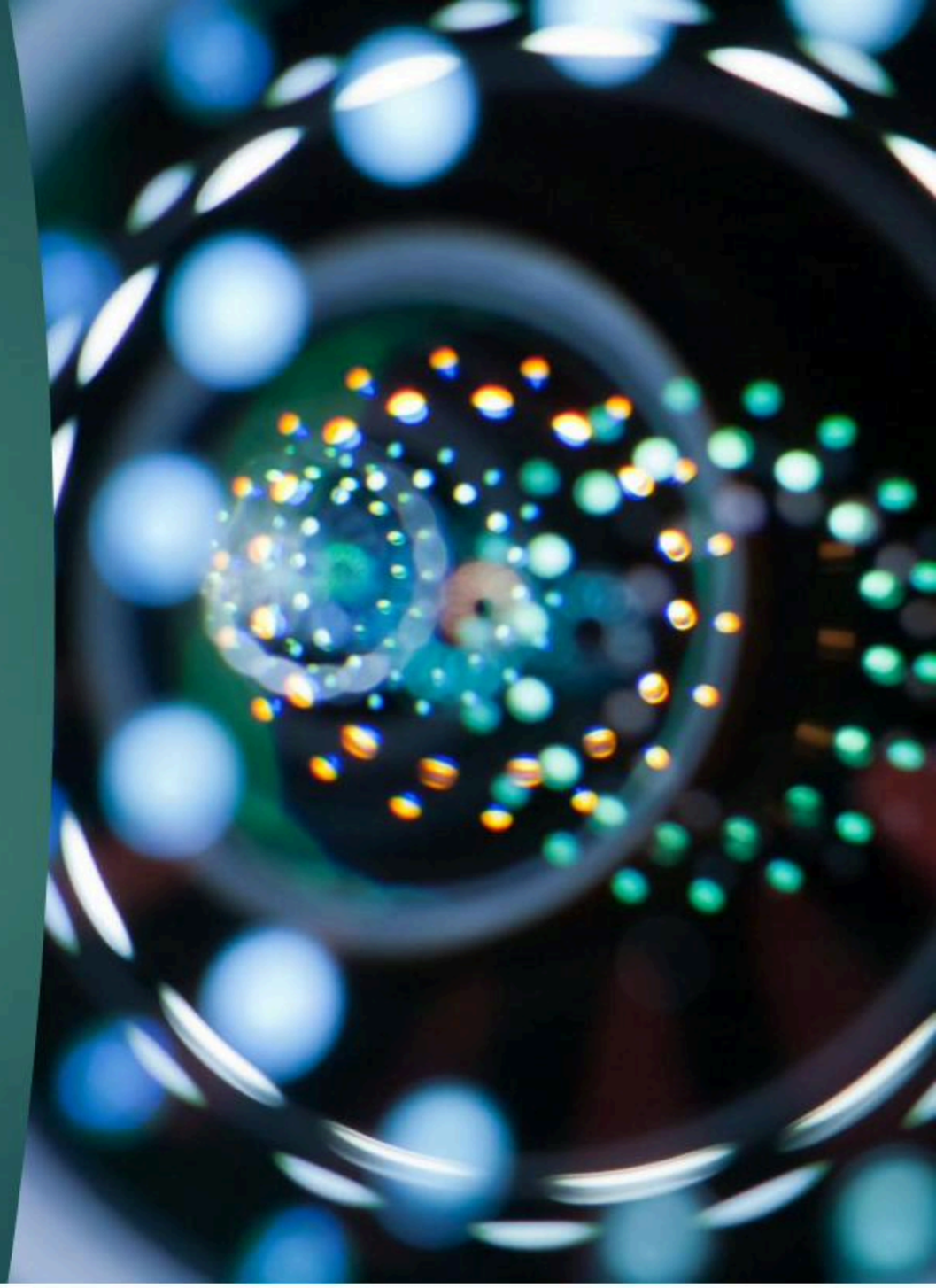
- Partner with defense integrators and UAV manufacturers for pilots and early adoption.
- Deploy as software-only upgrade on existing EO/IR systems (Jetson/ARM) with minimal integration.

1 https://www.grandviewresearch.com/industry-analysis/night-vision-device-market
https://www.researchandmarkets.com/report/global-military-electro-optical-infrared-market

Thank You

Omer Linton & Michael Orkin

omerlinton@gmail.com
michaelorkin3@gmail.com



Impact AntisepTech

Barak Katz

Note: No document was uploaded to the project for this pitch. Space reserved per the conference schedule (Entrepreneurship Session 1, 14:05\-14:20).