未来
6G-Mirai

6G-MIRAI-HARMONY - Initial results on AI-based baseband design for 6G communication

Project information

Project name	Machine Intelligence based Radio Access Infrastructure
Project acronym	6G-MIRAI-HARMONY
Grant agreement	101192369
Call	HORIZON-JU-SNS-2024
Topic	HORIZON-JU-SNS-2024-STREAM-B-01-05
Type of action	HORIZON JU Research and Innovation Actions
Start date	1 April 2025
Duration	36 months

Document information

Associated WP	WP1
Associated Task	T1.2
Associated Deliverable	D1.2
Main authors	Rouaa Diab (Editor) (Fraunhofer), Renato L.G. Cavalcante (Fraunhofer), Noor Ul Ain (Fraunhofer), Ismael Adalid (ISRD), Md Munjure Mowla (ISRD), Robbert Beerten (KUL), Tanguy Kerdoncuff (Ericsson), Adrian Garcia-Rodriguez (Ericsson), Mahdi Shahabi (SEQUANS)
Contributors	
Reviewers	Henning Sanneck (Apple), Jian Song (Apple), Apostolos Destounis (Ericsson), Jochen Fink (Fraunhofer)
Type	R
Dissemination level	PU
Comment	

Document revision history

Version	Date	Changes	Author
V1	05.05.26	First draft	(See main author list)
V2	18.05.26	Second draft	-
V3	09.06.26	Reviewed and revised version	-
V4	26.06.26	Final/Submitted version	-

Co-funded by the European Union

Disclaimer

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the SNS JU (granting authority). Neither the European Union nor the granting authority can be held responsible for them. 6G-MIRAI project has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation program under Grant Agreement No 101192369.

Contents

Acronyms and definitions

3GPP	3rd Generation Partnership Project
4G	Fourth Generation (of mobile communication networks)
5G	Fifth Generaiton (of wireless communication networks)
6G	Sixth Generation (of mobile communication networks)
ADC	Analog-to-Digital Converter
AI	Artificial Intelligence
AoA	Angle-of-Arrival
AoD	Angle-of-Departure
AP	Access Point
BBU	Baseband Processing Unit
BER	Bit Error Rate
BM	Beam Management
BO	Bayesian Optimization
BS	Base Station
CAPEX	Capital Expenditure
CF	Cell-Free
CF-mMIMO	Cell-Free Massive MIMO
CPU	Central Processing Unit
CQI	Channel Quality Indicator
CSI	Channel State Information
DDPG	Deep Deterministic Policy Gradient
DFT	Discrete Fourier Transform
DL	Downlink
DNN	Deep Neural Network

Co-funded by the European Union

EBS	Exhaustive Beam Sweeping
EC	Entropy Coder
EMD	Earth Mover’s Distance
EV	Eigenvector
GNN	Graph Neural Network
GPR	Gaussian Process Regression
GPU	Graphics Processing Unit
InH	Indoor Hotspot
IoT	Internet of Things
KVI	Key Value Indicators
LMMSE	Linear Minimum Mean Square Error
LS	Least Squares
LTE	Long Term Evolution
MIMO	Multiple Input Multiple Output
ML	Machine Learning
mmWave	Millimeter-Wave
MSE	Mean Squared Error
MU	Multiuser
Near-RT RIC	Near-Real-Time RAN Intelligent Controller
NMSE	Normalized Mean Squared Error
NR	New Radio
NRX	Neural Receiver
NTC	Non-linear Transform Coding
NW	Network
NZP-CSI-RS	Non-Zero Power Channel State Information Reference Signals
OFDM	Orthogonal Frequency Division Multiplexing
OPEX	Operational Expenditure

PHY	Physical Layer
PMI	Precoding Matrix Indicator
QoS	Quality of Service
RAN	Radio Access Network
RB	Resource Block
RE	Resource Element
Rec. Ag.	Receiver-Agnostic
Rec. Aw.	Receiver-Aware
ResNet	Residual Network
RI	Rank Indicator
RRH	Radio Remote Head
RS	Reference Signal
RSRP	Reference Signal Received Power
RT	Ray Tracing
SE	Spectral Efficiency
SNR	Signal-to-Noise Ratio
SQ	Scalar Quantizer
SSB	Synchronization Sequence Blocks
Tbps	Terabits Per Second
TDD	Time Division Duplex
UatF	Use-and-then-Forget
UE	User Equipment
UL	Uplink
UMa	Urban Macro
UMi	Urban Micro
URA	Uniform Rectangular Array
UT	User Terminal

VQ	Vector Quantization
WP	Work Package
XGMF	XG Mobile Promotion Forum

Executive Summary

This deliverable presents initial results on AI-based baseband processing developed within the scope of the 6G-MIRAI-HARMONY project. The work reported herein is carried out under WP1, Task 1.2.

Given the stringent requirements of future 6G systems—such as extreme adaptability, scalability, and performance under highly dynamic network conditions—WP1 aims to advance the development of next-generation wireless systems in which learning-based methods are embedded directly into the physical layer design.

The main objective of WP1 is to define, design, and realize an air interface that is inherently AI-native, treating artificial intelligence and machine learning not as auxiliary components, but as fundamental building blocks of the communication system. This paradigm enables seamless integration of AI/ML techniques into the air interface, supporting a more adaptive, efficient, and future-proof wireless communication framework.

In Task 1.2, the focus is on AI-based baseband design, covering key physical layer and supporting radio resource management functionalities through learning-based methods. In this deliverable, we present initial results on AI/ML-based techniques for prediction, overhead reduction, and system-level optimization.

Co-funded by the European Union

1 Introduction

The development of wireless communications has largely relied on enhancing network performance by increasing the number of available resources in the time, frequency, and spatial domains, coupled with powerful model-driven resource allocation techniques. Over the last two decades, Multiple Input Multiple Output (MIMO) has emerged as a key technology in both research and industry, with practical implementation in 4G LTE using relatively small-scale antenna arrays, followed by a massive MIMO expansion in 5G. Therefore, massive MIMO is now a well-established concept in both theory and practice, although it continues to evolve in more advanced research directions.

However, one limitation it does not fully address is the cell-edge user problem. This has led to significant interest in cell-free massive MIMO in recent years, where one of the main goals is to achieve a more homogeneous Quality of Service (QoS) distribution across the coverage area. While coordinated multipoint can be viewed as an earlier manifestation of cell-free MIMO, several practical limitations at the time made large-scale realization difficult. However, due to advancements in wireless networks over the years, particularly in distributed processing and fronthaul capabilities, these approaches have become more feasible, making distributed implementations an active research topic in both academia and industry.

The distributed nature and larger antenna arrays of future 6G networks introduce both challenges and opportunities. For example, network resource allocation problems become more complex not only due to increased dimensionality, but also due to the need to exchange messages and information between different network elements in distributed implementations. Moreover, optimization problems must account for different information and message-exchange mechanisms, as well as varying update rates of exchanged information, such as channel state information, when designing allocation strategies.

Due to its success in fields such as image processing and other data-rich applications, Artificial Intelligence (AI)/Machine Learning (ML) has also been widely adopted in wireless communications. The ability of data-driven models to learn directly from data makes them attractive for tackling complex, high-dimensional problems arising from the distributed and large-scale nature of 6G networks, and has opened a wide range of opportunities to revisit and enhance traditional approaches, as well as to develop novel AI/ML-based methodologies.

While this area has been studied for some time, the increased computational capabilities of modern Central Processing Units (CPUs) and the parallel processing power of Graphics Processing Units (GPUs), together with the availability of large-scale data, have contributed to the recent AI/ML boom. In 5G systems, AI/ML has primarily been used as an add-on to enhance traditional model-driven techniques. However, in 6G, the vision is to move towards more AI-native network designs, where learning-based methods are integrated into the fundamental architecture rather than serving as mere enhancements.

However, purely data-driven approaches may be insufficient in many cases, and from an industry perspective, a complete replacement of traditional model-based methods with AI/ML is currently considered impractical, as these approaches were actively explored and validated over many years, frequently reaching close-to-optimal performance. Thus, hybrid frameworks that combine AI/ML methods with well-established model-based techniques are increasingly adopted.

Contextualizing the discussion above within the 6G-MIRAI (Machine Intelligence-based Radio Access Infrastructure)-HARMONY project framework, we provide initial insight into the work being carried out in **Task 1.2** in **Work Package 1 (WP1)** on AI-based baseband design, with contributions from **Fraunhofer HHI**, **Ericsson France**, **SEQUANS**, **ISRD**, and **KU Leuven**.

We begin by highlighting the role of AI/ML in reducing network overhead in Section 2. Specifically, we present an end-to-end fronthaul compression technique and address Channel State Information (CSI) overhead reduction through antenna selection combined with AI/ML-based compensation, including contributions to standardization activities and 3rd Generation Partnership Project (3GPP) discussions. Section 3 focuses on the use of AI/ML for prediction in cell-free MIMO networks. In particular, we investigate channel prediction as a means of reducing CSI estimation overhead and subsequently address predictive beam selection. Section 4 presents a hybrid approach that combines conventional methods with AI/ML techniques for indoor access point placement. At the end of each subsection, we discuss how the presented work relates to the target Key Value Indicators

(KVIs) defined in Deliverable D3.1 of the project [1], covering environmental, economic and societal aspects. Finally, Section 5 summarizes the initial results obtained in this work, while Section 6 provides references.

2 AI/ML-based Overhead Reduction

As discussed in the introduction, the ever-increasing number of serving antennas in future AI-native 6G networks leads to significant overhead due to the large overhead created by measuring the channel on each of these antennas and processing the received signal from each of these antennas.

This section presents two techniques aimed at addressing these challenges. First, we consider the reduction of fronthaul signaling overhead required for distributed MIMO operation. To this end, an AI/ML-based compression algorithm for fronthaul rate minimization is presented, where signal reconstruction is optimized to maximize downstream task performance rather than signal fidelity alone. The proposed end-to-end receiver-aware compression approach is evaluated in terms of goodput versus fronthaul rate and is shown to outperform conventional methods, such as non-receiver-aware compression schemes.

Second, we address the overhead associated with CSI acquisition by reducing the number of antenna ports used for CSI estimation. In massive MIMO systems, utilizing all antenna ports for CSI acquisition may become inefficient. Since CSI-Reference Signal (RS) signals are transmitted over only a subset of the available spatial resources, the missing information must be recovered through AI/ML-based reconstruction techniques. The proposed approach is evaluated in terms of throughput as a function of Signal-to-Noise Ratio (SNR) and is shown to outperform baseline methods, including linear interpolation and CSI estimation using all antenna ports. Furthermore, the effects of different antenna muting patterns is shown.

2.1 Receiver Aware Neural Fronthaul Compression (KU Leuven)

The evolution of wireless communications toward Massive MIMO has dramatically increased the size of base-station antenna arrays [2], which is a primary enabler for the high throughput requirements of 5G and beyond. Modern network architectures are often disaggregated according to predefined 3GPP splits and decouple the Baseband Processing Unit (BBU) from the Remote Radio Head (RRH) to improve scalability and flexibility and to

Co-funded by the European Union 6GSNS

reduce operational costs [3][4]. The most prominent of these splits is the so-called split 7.2 where the physical layer itself is split between the BBU and the RRH. However, transmitting all receiver signals from the RRH to the BBU imposes immense fronthaul capacity requirements; thus, these signals must be sufficiently compressed. Designing such a compression algorithm is highly challenging, as it invariably forces downstream tasks to work with limited-precision reconstructions of those signals.

Signal processing under limited precision has been a significant topic of interest in wireless systems over the past few years. Existing works on uplink (UL) decoding can be broadly categorized into two categories: 1) hardware-constraint-aware processing where the signal processing algorithms must operate under low-precision Analog-to-Digital Converters (ADCs) at the receive antennas [5] and 2) capacity-constraint-aware processing that maximizes decoding capability under limited fronthaul capacity [6]. In this work, we consider the latter where the signal is preprocessed before being quantized by a Quantizer with a fixed configuration. By using an additional entropy coder, the fronthaul rate minimization can be recast as an entropy minimization problem, a well-known problem in machine learning. Traditionally, many compression algorithms are optimized only for signal fidelity; however, our main objective is to design fronthaul compression algorithms to account for downstream tasks. Additionally, traditional signal processing struggles to efficiently exploit the complex spatio-temporal and frequency-domain correlations inherent in Orthogonal Frequency Division Multiplexing (OFDM) signals and require knowledge of each user's channel or its statistics.

This shortcoming results in inefficient compression and degraded downstream performance. Furthermore, these methods often rely on complex, iterative optimization algorithms [6]. To address these limitations, we propose a novel end-to-end learned fronthaul compression framework based on Non-linear Transform Coding (NTC). This paradigm, which has recently achieved strong performance in image processing through the work of Ballé et al. [7][8][9], formulates compression as an end-to-end optimization problem in which a neural network encodes the signal into a latent space, which is subsequently quantized and compressed by an Entropy Coder (EC).

Simultaneously, Neural Receivers (NRXs), where the UL decoder is replaced by a neural network, have recently gained traction in the wireless community [10]. By designing both the fronthaul compression algorithm and the UL decoder as neural networks, they become fully differentiable and can be optimized as part of a fully end-to-end system.

Ultimately, current fronthaul compression schemes:

- Require perfectly known channel statistics or known channels
- Cannot optimize for end-to-end metrics such as BER

Our work addresses both shortcomings directly. We propose a receiver-aware fronthaul framework that directly optimizes fronthaul compression for symbol decoding without altering the receiver. The most related work is presented by Bian and Günduz [11] who have explored fronthaul compression via distortion minimization through Vector Quantization (VQ) on the one hand and NTC on the other hand. Their approach, however, does not account for downstream tasks.

To address these issues, we present our main contribution in the next. Which consists of the design of a neural fronthaul compression that is fully aware and optimized for a further downstream UL receiver.

2.1.1 Neural Compression

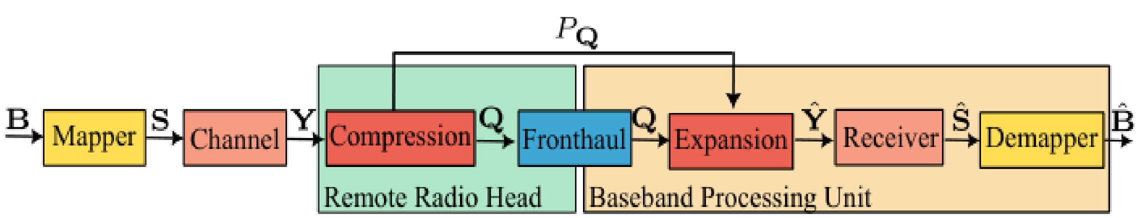

Figure 1: Our system model, its components and the associated signals.

Our central idea comprises a learned fronthaul compression framework based on NTC, where a quantized signal's entropy is optimized and the signal is subsequently entropy coded. Our central innovation is a "**receiver-aware**" strategy. Instead of only trying to reconstruct the waveform as accurately as possible, the compression algorithm is optimized end-to-end to maximize downstream decoding performance—effectively discarding irrelevant noise while preserving the crucial features for the baseband receiver.

Pipeline: The proposed system treats the entire compression and decoding process as a single, differentiable neural pipeline. The flow operates in four main steps:

- **Preprocessing**: The raw fronthaul signal is first converted into the angular domain using a deterministic discrete Fourier transform (DFT), making the signal sparser and easier to compress.
- **Compression (RRH)**: The Analysis Transform neural network maps the preprocessed signal into a lower-dimensional latent space. The signal is then quantized and passed through an Entropy Coder to minimize the data rate required for transmission.
- **Expansion (BBU)**: Upon reaching the BBU, the entropy decoder and Synthesis Transform network expands the compressed signal back to its original dimensions.
- **Receiver (BBU)**: A graph neural network (GNN) acts as the uplink decoder, estimating the final transmitted bitstream directly from the reconstructed signal.

2.1.2 Results

Our simulation environment uses the 3GPP Urban Micro (UMi) channel model with one single-antenna user, an RRH with 64 antenna elements and a BBU. Our simulation considers a window of 14 consecutive time symbols and 48 adjacent subcarriers at a carrier frequency of 2.14 GHz.

To verify our results, we first validate the trainable fronthaul compression algorithm by training purely for signal fidelity. We train a separate model for each point on the rate-distortion pareto curve. This result is shown in Figure 2Figure 2. We find that even for a relatively high $\frac{E_b}{N_0}$, we need a rather large fronthaul rate for a modest normalized mean squared error. For example, a fronthaul rate of 100 bits per channel use (time-frequency resource element) can only guarantee an normalized mean squared error (NMSE) of -5 dB at an $\frac{E_b}{N_0}$ of -2 dB.

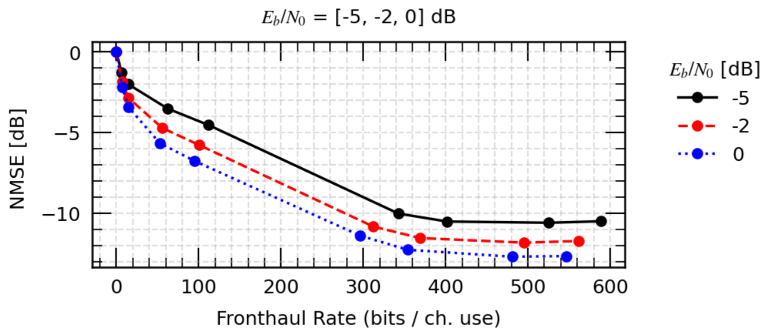

Figure 2: NMSE as a function of fronthaul rate

We then highlight the superiority of our receiver-aware (Rec. Aw) method by comparing it to the receiver-agnostic (Rec. Ag.) method which only minimizes the NMSE of the fronthaul signal. We also compare our results with those of the NRX without fronthaul compression, a Least Squares receiver with a simple scalar quantizer (LS SQ) and a NRX with scalar quantizer (NRX SQ). Our result shows the achieved goodput per resource element as a function of the required fronthaul rate to send this resource element from the RRH to the BBU. Both values are computed as the averages over the full window.

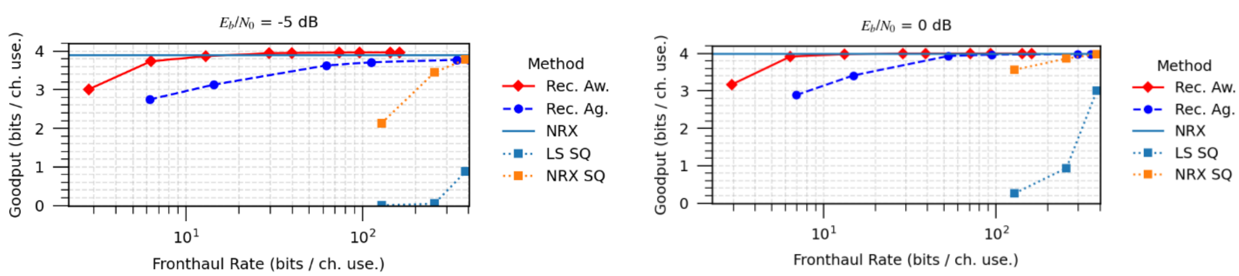

Figure 3: Achieved goodput as a function of fronthaul rate for different Eb/N0

The fundamental difference lies in the optimization goal: receiver-agnostic compression uses Mean Squared Error (MSE) to reconstruct the original antenna samples as accurately as possible, which wastes capacity by preserving irrelevant noise. In contrast, receiver-aware compression treats the system as a unified pipeline, optimizing directly for decoding accuracy.

The performance gap becomes most evident in high-noise environments (i.e. E_b/N_0 = -5 dB) where the receiver-aware seems able to effectively filter the signal from the noisy received vector. The receiver-aware model achieves a goodput of 3.87 bits using a fronthaul rate of only 12.95 bits. The receiver-agnostic model requires a higher average fronthaul rate (13.41 bits) yet yields a lower average goodput of only 3.19 bits per resource element.

We conclude that our Neural Fronthaul compression scheme is can significantly reduce the fronthaul requirement and we can improve performance significantly by considering the full end-to-end system and making the fronthaul aware of the downstream receiver.

2.1.3 KVI Discussion

Environmental:

Fronthaul rate minimization leads to more efficient utilization of the fronthaul link and will reduce the energy consumption associated with data transmission between the BBU and RRH by increasing the efficiency (in terms of fronthaul rate a.f.o. the total throughput) of the fronthaul. These gains are enabled by the AI/ML model, which learns compact latent representations that reduce redundancy in the transmitted information. However, the energy consumption associated with training and inference should also be considered.

Economic:

Lower fronthaul rate requirements may reduce Operational Expenditure (OPEX) through more efficient utilization of fronthaul resources and lower transport load. Reduced capacity requirements may also contribute to lower Capital Expenditure (CAPEX). These benefits are enabled by the AI/ML-based compression approach, although its computational and deployment costs should also be taken into account.

Societal:

Unlike traditional approaches that rely on explicit user-specific channel information, the proposed scheme reduces dependence on such information, thereby limiting the amount of user-specific channel information that must be estimated, stored, or exchanged. Furthermore, efficiently enabling the offloading of the signal decoding from an edge-deployed base station to a central processing cloud. This may contribute to privacy-related objectives and reduce concerns associated with data handling.

Co-funded by the European Union

2.2 AI-Based CSI-RS Overhead Reduction (Ericsson)

The evolution toward 6G radio access networks is characterized by a rapid increase in the number of antenna ports enabled by large-scale and highly dense antenna arrays [12,13]. While such architectures provide substantial gains in spatial resolution, beamforming flexibility, and spatial multiplexing, they also exacerbate the overhead associated with downlink CSI-RS. In conventional New Radio (NR) systems, CSI acquisition relies on transmitting CSI-RS on all active antenna ports, allowing the user equipment (UE) to estimate the full port-domain channel and report key feedback quantities such as rank indicator (RI), channel quality indicator (CQI), and precoding matrix indicator (PMI) [14].

As the number of antenna ports grows to several hundred in envisioned 6G deployments, full-port CSI-RS transmission becomes increasingly inefficient, consuming a non-negligible fraction of the available time–frequency resources as shown on Figure 4.This issue has been recognized in 6G studies within 3GPP where Ericsson has several contributions. Additionally, Ericsson has submitted a white paper on this topic in XG Mobile Promotion Forum (XGMF) 6G Radio Technology Project in Japan. For a comprehensive list of contributions, the reader is referred to Section 2.2.5.

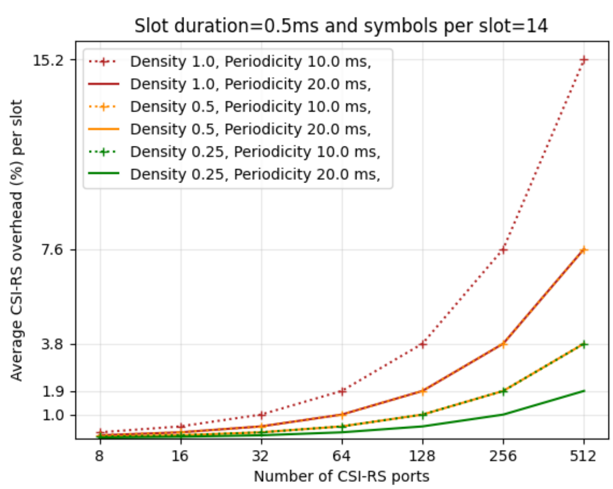

Figure 4: Average CSI-RS overhead for different number of CSI-RS ports, periodicity and frequency density.

To solve this problem, one option is to use spatial CSI-RS overhead reduction, where only on a subset of CSI-RS ports are transmitted, while the remaining ports are muted. When CSI-RS is transmitted only on a subset of antenna ports, the UE observes an incomplete version of the wireless channel. Recovering the missing channel coefficients is a non-trivial inference problem, especially for aggressive port reduction ratios (e.g., sounding only 25% of the ports). This motivates the investigation of AI/ML-based approaches, which are capable of learning structured relationships directly from data.

As shown on Figure 5, which illustrates the workflow of the method, the reconstruction can be performed at the UE side or at the network (NW) side. In the NW situation, the UE compresses the CSI and feedback the CSI based only on the subset of port transmitted, without any reconstruction; note that there exists different flavors to compress the CSI in the case. But in the following we focus on the development of UE-side spatial-domain CSI overhead reduction and compare it against various non-AI benchmarks. This choice is motivated by the fact that UE-side reconstruction represents a more natural evolution of existing 5G NR mechanisms, as it can largely reuse the current CSI acquisition, compression, and feedback framework without requiring fundamental changes to the network architecture. As such, UE-side AI-based CSI reconstruction constitutes a sensible first step toward AI-enhanced CSI acquisition in 6G and has therefore been the primary focus in the early phases of 6G studies.

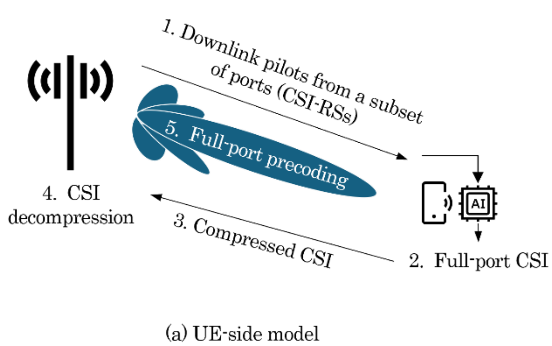

Figure 5: Illustration of (a) UE-side and (b) network-side model variants for spatial CSI-RS overhead reduction

The choice of CSI-RS ports to be sounded defines the muting pattern, and the fraction of sounded ports determines the port reduction ratio (e.g., 50% or 25% of ports). Figure 6 shows potential muting patterns each exhibiting different spatial sampling characteristics, which intuitively lead to distinct performance.

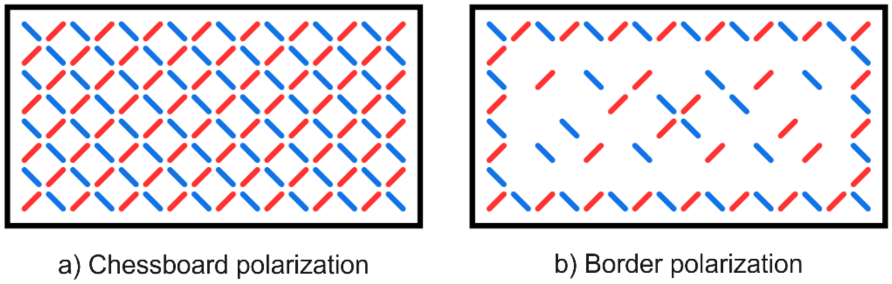

Figure 6: Examples of good CSI-RS partial port sounding patterns with 50% of ports (left) and 25% (right)

2.2.1 UE-side AI/ML-based channel state information (CSI) reconstruction

The AI/ML-based CSI reconstruction model is trained in a supervised manner using statistical channel realizations and evaluation assumptions following 6G 3GPP TSG RAN WG1 #124 (evaluation methodology for link-level simulation of DL based CSI acquisition) associated with the 3GPP Study of 6G Radio Study Item RP-251881. Additionally, the Table 1 list the specific configuration used between the different possibilities agreed. The input to the UE model is obtained from CSI-RSs transmitted on a subset of antenna ports and subcarriers, with muted ports set to zero. The output is the reconstructed full channel matrix corresponding to all CSI-RS ports. One RE (Resource Element) per RB (Resource Block) is used for both the input and the output, potential future work would consider also some frequency reconstruction with for instance one RE per 2 RB.

A convolutional neural network with residual (ResNet-like) connections is used to learn spatial and frequency-domain correlations and infer the missing channel coefficients. Training is performed offline, using data characterizing a wide variety of wireless channel conditions, i.e., different UE locations in an Urban Macro deployment. In the current setting, the model is supposed to work on any Base Station (BS) in urban area. During evaluation, the trained model is used for inference at the UE, without online learning. In the current study, the model is UE-sided and trained for a single BS antenna array configuration and a fixed CSI-RS muting pattern.

2.2.2 UE-side non-AI/ML-based channel state information (CSI) reconstruction

Three representative non-AI/ML based CSI reconstruction methods approaches are considered as benchmarks:

- **All-port**: Legacy 5G NR approach with all CSI-RS ports transmitted.
- **Linear interpolation**: The channel coefficients of muted ports are estimated by interpolating between neighboring sounded ports.
- **Combined partial and full CSI-RS transmission**: In this approach illustrated in Figure 7, sparse CSI-RS transmissions are complemented by occasional full-port CSI-RS transmissions. The full-port measurements are used to estimate wideband spatial covariance, which is then exploited (e.g., via LMMSE-based interpolation) to reconstruct muted ports during sparse transmissions. This method improves reconstruction accuracy but partially offsets the overhead reduction gains due to the presence of periodic full-port CSI-RS transmissions.

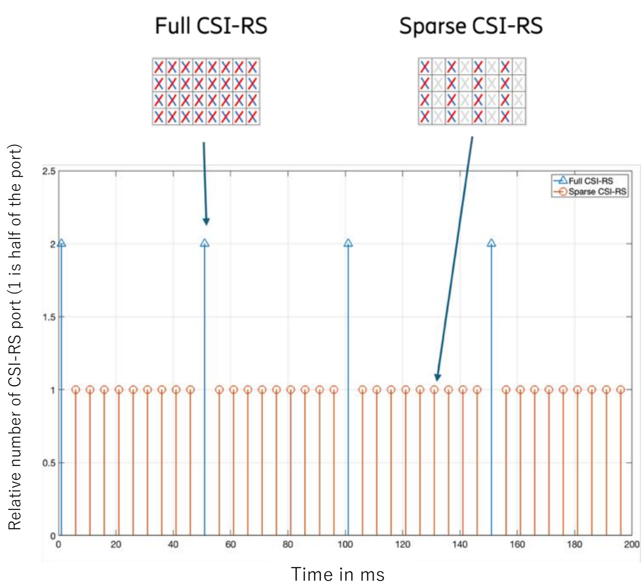

Figure 7: Illustration of the combined partial and full CSI RS transmission benchmark. Time in ms.

2.2.3 Throughput Performance

Figure 8 shows the link-level downlink throughput as a function of SNR for the considered CSI-RS overhead reduction schemes with CSI feedback type I and type II and the 25%

border polarization muting pattern display on Figure 6. Type I and Type II feedback are two possible ways for the UE to report the estimated channel to the NW. They are based on agreed list of matrices (codebook) which are reported by the UE. Type II is more complex but contains more information about the channel which can usually then be used in multiuser MIMO (MU-MIMO) condition to reduce the inference between UE. The result presented are obtained using the configuration provided in Table 1.

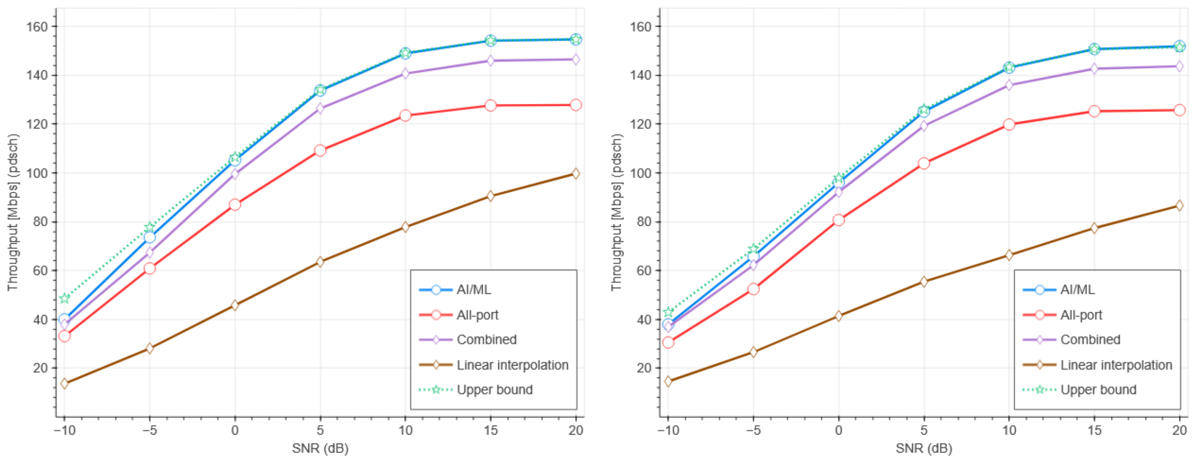

Figure 8: Downlink throughput vs. SNR with Type II CSI feedback (left) and Type I CSI feedback (right). Urban Macro deployment, 256 base station antennas, 4 UE antennas

Across the entire SNR range, the AI-based approach consistently outperforms non-AI baselines and approaches the upper bound corresponding to ideal CSI reconstruction. The gain is particularly pronounced at higher SNRs compared to All-port methods that suffer from high CSI-RS overhead. In contrast, linear interpolation exhibits significant throughput degradation.

Table 1: Common evaluation assumptions used for training and evaluation of AI/ML model and non-AI benchmarks for CSI-RS overhead reduction

Parameters	Value
Carrier frequency	7GHz
Duplex	TDD
TDD DL/UL ratio	60% DL slots and 40% UL slots
Allocation Bandwidth	20 MHz
Subcarrier spacing	30 kHz
BS antenna configuration	256 ports: (M, N, P, Mg, Ng, Mp, Np) = (32,16,2,1,1,8,16), (dH,dV) = (0.5, 0.5) λ
UE antenna configuration	4 ports: (M, N, P, Mg, Ng, Mp, Np) = (1,2,2,1,1,1,1)
MIMO rank	1 or 2 depending on rank adaptation
UE number for MU-MIMO	1
CSI feedback delay	4 ms
Muting pattern	CSI-RS transmission with half or a quarter of the ports (Figure 66)
Feedback type	Rel-19 Type I Scheme B or Rel-19 eType II
CSI-RS periodicity	10 ms
Channel estimation	Practical channel estimation for DMRS and CSI-RS
BS Tx power constraint	Per TRP power constraint
EVM	No radio impairments
UE speed	3 km/h for both indoor and outdoor
Channel model	Urban Macro
PDCCH	Reserved the first 2 symbols of each DL slot for PDCCH

2.2.4 Impact of port reduction ratio

The performance of AI-based CSI reconstruction is strongly influenced by both the port reduction ratio and the associated muting pattern. Consequently, a significant part of this work was devoted to the evaluation multiple candidate muting patterns in order to identify configurations offering a favorable trade-off between CSI-RS overhead reduction and reconstruction performance. Two representative and highly performant muting patterns corresponding to half- and quarter-port sounding are illustrated in Figure 6. The associated throughput performance is shown in Figure 9, where the AI/ML-based approach remains very close to the upper bound for both port reduction ratios. Furthermore, quarter-port sounding provides better performance than half-port sounding for both the AI/ML method and the advanced non-AI baseline (Combined), while simple linear interpolation exhibits clear performance limitations, particularly at aggressive port reduction ratios.

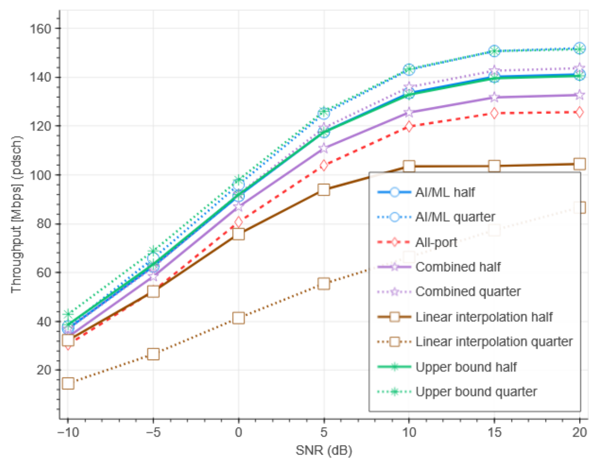

Figure 9: Throughput vs. SNR for two muting patterns presented in Figure 6

2.2.5 KVI Discussion

Environmental:

The reduction in CSI-RS antenna ports decreases the amount of reference signaling transmitted over the air interface, leading to more efficient utilization of radio resources, and thus reducing the energy required for CSI acquisition. These benefits are enabled by the AI/ML-based signal reconstruction approach; therefore, the overall environmental impact should also account for the computational resources and energy consumption associated with model training and inference to ensure a net efficiency gain. The model is trained offline, meaning that the computationally intensive training phase is not part of normal network operation and only inference is required during deployment.

Economic:

Signal reconstruction is performed at the UE side and remains compatible with the existing CSI acquisition and feedback framework. This allows the current network architecture and procedures to be reused without limited modifications, potentially reducing deployment complexity, implementation effort, and associated CAPEX and OPEX costs. However, the computational and hardware requirements associated with AI/ML deployment should also be considered when assessing the overall impact and additionally, the large gain obtained by this method are useful mainly for high number of antenna (e.g. 1024) which limit its interest in current deployment.

Societal:

The proposed models are trained using statistical channel realizations and assumptions aligned with 3GPP guidelines rather than datasets containing user-specific information. Consequently, the model development process does not require the collection or processing of personal user data, which may support privacy-related objectives and reduce concerns associated with data handling. Further studies might consider real data to have more efficient AI/ML model, but the approach to gather those data are out of scope of the study.

2.2.6 External Output

[1] Ericsson, "TDoc R1-2505180: AI/ML Use Cases for 6GR Air Interface," 3GPP TSG-RAN WG1 Meeting #122, Bengaluru, India, August 25-29, 2025.

[2] Ericsson, "TDoc R1-2506762: AI/ML Use Cases for 6GR Air Interface," 3GPP TSG-RAN WG1 Meeting #122bis, Prague, Czech, October 13–17, 2025.

[3] Ericsson, "TDoc R1-2508365: AI/ML Use Cases for 6GR Air Interface," 3GPP TSG-RAN WG1 Meeting #123, Dallas, USA, Nov 17-21, 2025.

[4] Ericsson, "TDoc R1-2601038: Views on aspects of downlink-based CSI Acquisition," 3GPP TSG-RAN WG1 Meeting #124, Gothenburg, Sweden, February 9-13, 2026.

[5] Ericsson, "TDoc R1-2603259: Views on aspects of downlink-based CSI Acquisition," 3GPP TSG-RAN WG1 Meeting #124bis, Malta, MT, April 13-17, 2026.

[6] A. Garcia-Rodriguez, "Keynote: The role of artificial intelligence in the 6G air interface," 6th IEICE-CS International Conference on Emerging Technologies for Communications, Osaka, Japan, November 26-28, 2025. [Online]. Available: https://zenodo.org/records/17858948.

3 AI/ML-based Prediction in Cell-free MIMO

This section focuses on AI/ML-based prediction techniques for wireless communications, specifically CSI prediction and beam selection.

First, we consider an indoor hotspot scenario characterized by low user mobility, where frequent CSI acquisition may become redundant and place unnecessary burden on system resources. To address this, AI/ML is employed to predict the channel state while reducing CSI measurement effort. The effectiveness of the proposed approach is evaluated through a similarity analysis, demonstrating satisfactory prediction accuracy.

Next, we address beam management, which traditionally relies on exhaustive beam search procedures. In this work, AI/ML is used to predict the optimal narrow beam thereby reducing the need for exhaustive search while maintaining effective beam selection performance, demonstrated through high classification accuracy.

3.1 AI-based channel prediction for reducing channel estimation overhead in CF mMIMO (SEQ)

CSI acquisition is a fundamental enabler for coherent transmission, beamforming, user-centric cooperation, and resource allocation in cell-free massive MIMO (CF-mMIMO) systems. However, obtaining accurate CSI requires periodic CSI-RS measurements, channel estimation procedures, and CSI reporting mechanisms, all of which consume radio resources and increase system complexity. As the number of distributed access points (APs) and connected devices increases, the overhead associated with CSI acquisition becomes a significant scalability challenge.

This issue is particularly relevant in Indoor Hotspot (InH) deployments, which are representative of many envisioned industrial Internet of Things (IoT) and smart-manufacturing scenarios. In such environments, user terminals (UTs) often exhibit low mobility and the wireless channel evolves relatively slowly over time. Consequently, strong temporal, spatial, and frequency-domain correlations can be observed across consecutive channel realizations. These correlations create opportunities to reduce CSI acquisition overhead by exploiting previously observed channel information and partial channel measurements rather than relying exclusively on frequent pilot-based estimation procedures.

Co-funded by the European Union

Within this context, the main objective of this work is to reduce CSI measurement effort and CSI reporting overhead while maintaining sufficiently accurate channel knowledge for communication and resource allocation functions. To achieve this objective, two complementary approaches are investigated. The first approach focuses on AI-based CSI prediction, where future channel realizations are inferred from previously observed CSI. The second approach focuses on AI-based channel deduction, where partial channel observations are combined with historical CSI information to reconstruct the complete channel state. Both approaches follow a hybrid methodology that combines conventional channel estimation techniques with AI-assisted inference mechanisms.

3.3.1 Problem Formulation

Consider a user-centric CF-mMIMO deployment operating in a low-mobility InH environment. Under conventional operation, CSI acquisition requires frequent CSI-RS transmissions, channel measurements, and CSI reporting procedures to maintain accurate channel knowledge at the network side. While this approach provides reliable CSI estimates, it introduces substantial signaling overhead, particularly when channel updates are performed frequently.

The key idea explored in this work is that, due to the relatively slow channel evolution in low-mobility InH scenarios, a significant portion of the CSI information can be inferred from previously available channel observations. Consequently, the frequency of CSI acquisition procedures can potentially be reduced without significantly degrading CSI accuracy. In particular, we assume there are some moving objects in the environment when generating the dataset from the available database to provide some sorts of dynamic scenarios. However, limited UT mobility can be considered as well through changing the UT location inside the predefined user grids in the database.

The investigated approaches therefore aim to answer the following research question:

How can temporal and spatial channel correlations be exploited to reduce CSI acquisition overhead while maintaining sufficiently accurate CSI estimates for CF-mMIMO operation?

To address this question, two complementary solutions are investigated. The first solution exploits temporal channel correlation through channel prediction, whereas the second solution exploits both temporal and spatial channel correlations through channel deduction from partial CSI observations.

3.3.2 AI-based CSI Prediction

The first investigated approach focuses on reducing CSI acquisition overhead through channel prediction. The underlying assumption is that, in low-mobility InH scenarios, consecutive channel realizations remain highly correlated over multiple transmission intervals. As a result, future channel states can be inferred from previously estimated CSI without requiring channel estimation at every update instant.

The considered channel model follows the conventional decomposition of the wireless channel into large-scale and small-scale fading components. The large-scale component captures path-loss and shadowing effects, whereas the small-scale component reflects multipath propagation and Doppler-induced channel variations. Due to the limited mobility of the considered users, channel evolution can be approximated through a channel-aging process in which future channel realizations remain strongly correlated with previously observed channel states. In other words, the channel data available from previous time instances and the mobility of the user terminals and their new locations are predicted, and the current channel data is updated according to the channel aging model and predicted locations. Based on this observation, CSI prediction is formulated as a sequential learning problem where the objective is to estimate future channel realizations using historical CSI observations. While conventional channel-aging models can provide useful baseline predictions, their performance may degrade when channel conditions vary over time or when environmental dynamics become difficult to model analytically.

To improve adaptability, a reinforcement-learning framework based on Deep Deterministic Policy Gradient (DDPG) is currently being investigated. The proposed framework utilizes historical CSI observations together with channel evolution statistics to learn adaptive prediction policies. The AI component continuously adjusts its prediction strategy according to the observed channel behavior, while conventional channel models provide a physically meaningful representation of channel evolution. This combination results in a hybrid AI-assisted and model-based prediction framework.

The primary objective of this approach is to reduce the frequency of CSI acquisition procedures while preserving sufficient CSI accuracy for subsequent communication tasks. By exploiting temporal channel correlation, CSI-RS measurements and reporting operations can potentially be performed less frequently, thereby reducing signaling overhead.

Figure 10 illustrates the considered channel prediction framework according to predicting the mobility of user terminals and their new locations, and highlights the interaction between historical CSI observations, channel evolution modeling, and AI-assisted prediction.

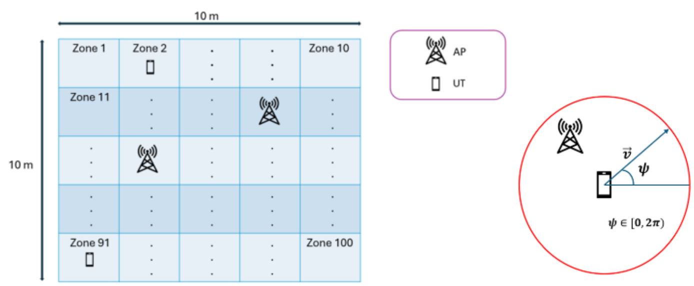

Figure 10. Predicting UT mobility and up-to-date location based on the speed and direction and updating the channel data according to the channel aging model

3.3.3 AI-Based Channel Deduction from Partial CSI Observation

Although channel prediction can reduce CSI acquisition overhead, prediction accuracy generally decreases as the prediction horizon increases. Furthermore, unexpected environmental changes may introduce prediction errors that accumulate over time. To address these limitations, a second complementary approach referred to as channel deduction is being investigated.

The fundamental idea is to combine a limited number of current CSI observations with historical CSI information to reconstruct the complete channel state. Instead of relying entirely on channel prediction or performing full CSI estimation, only a subset of channel observations is acquired through conventional CSI measurement procedures. These partial observations are then combined with previously available CSI to infer the remaining channel information.

The proposed approach exploits the strong spatial, temporal, and frequency-domain correlations typically observed in low-mobility InH environments. Historical CSI observations provide information regarding channel evolution trends, while partial CSI measurements provide real-time information about the current channel conditions. By jointly processing both sources of information, the reconstruction framework can infer complete CSI estimates while requiring significantly fewer channel measurements.

Unlike conventional interpolation techniques, the proposed framework leverages AI-based learning mechanisms to capture complex relationships between historical and current CSI observations. At the same time, unlike purely data-driven approaches, conventional CSI measurements remain an integral part of the framework and continuously provide reliable channel references. Consequently, the proposed solution can be viewed as a hybrid AI-assisted and measurement-assisted CSI acquisition approach.

Figure 11 illustrates the proposed channel deduction framework and the interaction between historical CSI information and partial channel observations based on a hybrid approach for estimation/prediction [15]. The figure also highlights how the reconstructed CSI is obtained through the fusion of measured and inferred channel information.

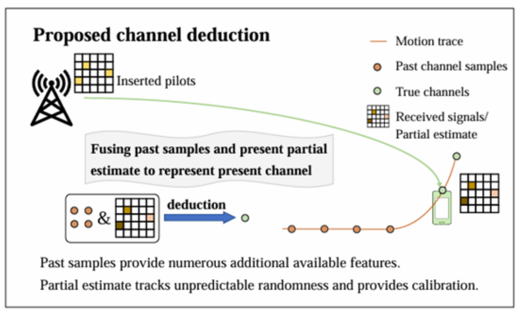

Figure 11: channel deduction method to address limitations of estimation and Prediction

3.3.4 Preliminary Results and Discussion

The current stage of the work focuses on establishing the theoretical foundations and AI-based architectures required for low-overhead CSI acquisition in low-mobility InH scenarios.

For the channel prediction approach, analytical channel-evolution models have been developed to characterize temporal channel correlation under low mobility conditions. For instance, in a low-mobility scenario with a speed limit of less than 1 m/s. We have investigated the CSI accuracy in terms of similarity obtained from NMSE metric. It is assumed that there is some infrequent pilot-based CSI acquisition at the estimation points, and in the meantime when there is no up-to-date data, the channel coefficients are predicted according to prediction of user mobility and the channel aging model. Preliminary investigations confirm that the considered InH environment exhibits strong temporal correlation, creating favorable conditions for CSI prediction which results in similarity up to 90 percent. The CSI overhead may be further reduced through predicting eigenvectors (EVs), instead of channel coefficients, and recover the full channel from the EVs at the cost of reduced similarity as shown in Figure 12. Initial simulation results indicate that prediction accuracy remains satisfactory within the coherence interval of the considered scenarios, suggesting that CSI acquisition procedures may potentially be performed less frequently than in conventional systems.

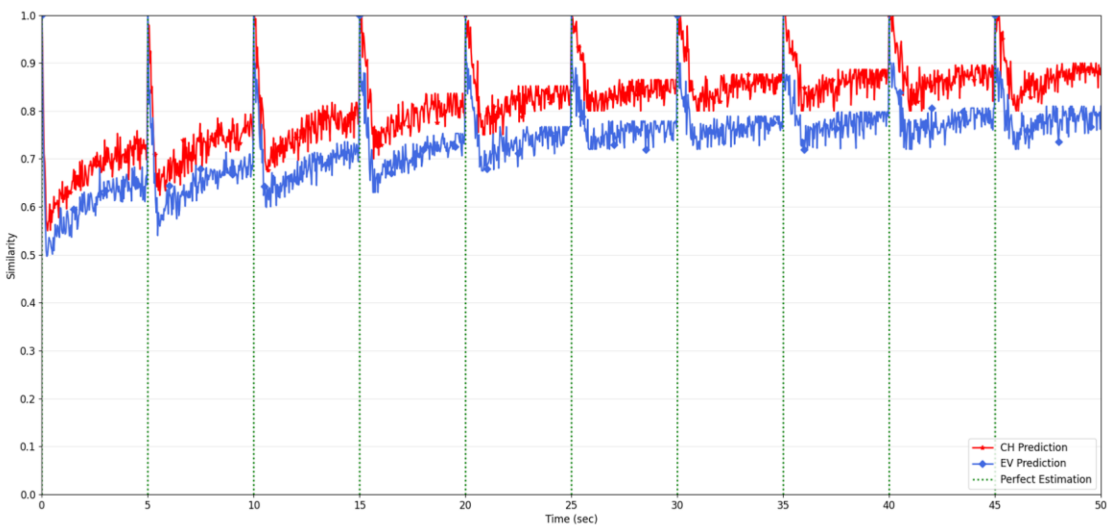

Figure 12: Similarity analysis for low-mobility scenario with infrequent channel estimation

For the channel deduction approach, the framework architecture has been established, and initial evaluations are currently underway. Preliminary observations indicate that partial CSI measurements can effectively compensate for prediction of inaccuracies while requiring substantially fewer pilot resources than full CSI estimation procedures.

3.3.5 KVI Discussion

Environmental:

Reduced feedback frequency of CSI estimation lowers signaling overhead and radio resource usage, essentially reducing processing power and energy consumption for channel acquisition (e.g., less average active antenna units and CSI feedback decoding workload at a massive MIMO cell site). However, this must be balanced against the computational cost and energy required for AI/ML-based prediction and deduction/inference using historical channel data.

Economic:

Reducing CSI estimation and feedback frequency may primarily lower OPEX through reduced processing load and signaling/backhaul costs, respectively. The AI/ML-based prediction reduces the need for frequent measurements but introduces additional computational requirements that must be considered in operational cost. Reduced CSI feedback frequency may drive long-term CAPEX reductions by increasing lifespan of existing infrastructure due to lowered processing stress on physical hardware.

Societal:

The approach relies on historical channel observations for prediction and deduction. Storage, handling, and processing of such historical channel data should be considered in relation to privacy and data governance aspects.

3.2 ML-based beamforming algorithm from real-time traffic prediction for CF mMIMO (ISRD)

The evolution of mobile networks toward 6G is driven by the necessity to support extreme data rates, projected on the order of terabits per second (Tbps), and deterministic latency for next-generation applications. To achieve these PHY layer capacities, 6G architectures must extensively exploit the massive bandwidths available at higher frequencies. However, electromagnetic wave propagation at these extreme frequencies is severely vulnerable to physical blockages, atmospheric absorption, and profound free-space path loss. To close the link budget, networks must deploy MIMO systems that form highly directional beams to concentrate power toward the UE.

Simultaneously, to mitigate the severe inter-cell interference inherent in traditional topologies, 6G introduces the CF-mMIMO paradigm. In a Cell-Free (CF) architecture, a vast number of geographically distributed APs which conceptually map to disaggregated O-RUs coherently and jointly serve UEs utilizing the same time-frequency resource blocks. By eliminating artificial cellular boundaries, CF-mMIMO significantly enhances spectral efficiency at the MAC. However, this distributed topology exponentially complicates the Beam Management (BM) framework. Coordinating highly directional spatial beams across a massive, disaggregated network of multi-panel APs transforms the traditional spatial search space into a high-dimensional optimization problem. Executing legacy heuristic algorithms in such a dense, interference-limited environment is computationally intractable and prohibitive for strict latency budgets.

To master this dimensionality challenge and optimize Radio Access Network (RAN) traffic throughput, ML and AI emerge as indispensable technological enablers. Native to modern, disaggregated network architectures such as the O-RAN framework, which supports intelligence hosting via the Near-Real-Time RAN Intelligent Controller (Near-RT RIC), ML algorithms can efficiently orchestrate complex beamforming policies across distributed nodes.

3.2.1 Motivation

As previously discussed, as telecommunications networks evolve toward 6G, the reliance on MIMO becomes critical to meet stricter capacity and latency requirements. Traditionally, according to [16], during initial access, a first coarse beam is established and selected via the Synchronization Signal Block (SSB) (P1 procedure). This is followed by a fine beam refinement process (P2 and P3 procedures) using Non-Zero Power Channel State Information Reference Signals (NZP-CSI-RS).

The exhaustive beam sweeping approach presents a severe bottleneck. In the baseline methodology, the network must transmit multiple CSI-RS resources (e.g., 12 distinct beams, as modeled in our baseline simulation) across the time-frequency grid. The UE calculates the Reference Signal Received Power (RSRP) for each beam, and the optimal beam (for the P2 procedure) is then selected based on the maximum RSRP measured by the UE. This traditional methodology generates a massive CSI-RS beam measurement overhead that scales multiplicatively with respect to the number of beams at the base station (M) and the user equipment (N), resulting in an MxN complexity.

The core limitation of this traditional approach is the spectral efficiency penalty imposed by the pilot overhead required for beam selection. Network capacity, C_{eff}, is reduced proportionally according to the number of resource elements dedicated to CSI-RS sweeping rather than user data payload.

$$C_{eff} = \left(1 + \frac{N_{CSI}}{N_{Total}}\right) B \cdot Log_2(1 + SNR)$$

3.2.2 Proposed Deep Learning solution

To resolve this, we propose a ML predictive beam selection algorithm. Rather than relying on exhaustive RSRP (from CSI-RS) measurements for the P2 procedure, solution leverages environmental and spatial features such as spatial loss, angles of arrival/departure, SSB transmission angle, transmitter position, receiver position, path loss, and scatterer positioning to directly infer the optimal fine beam via a Deep Neural Network (DNN).

To extract this measurement dataset, it has been developed a simulator that models a scattering MIMO channel, free-space path loss, and phased Uniform Rectangular Arrays (URAs). The extracted dataset inherently captures the complex, nonlinear relationships between scatterer locations and optimal beam trajectories. However, as a critical observation for future 6G innovation, this model assumes deterministic knowledge of environmental scatterer positions.

By leveraging contextual telemetry and environmental learning rather than relying exclusively on exhaustive L1 probing, AI models can proactively infer optimal (or suboptimal) transmission strategies.

Figure 13 conceptually illustrates this AI-augmented workflow, showing how intelligent beam decision blocks trigger optimal narrow beam refinement based on contextual channel feedback, establishing the foundation for the predictive methodology.

Co-funded by the European Union

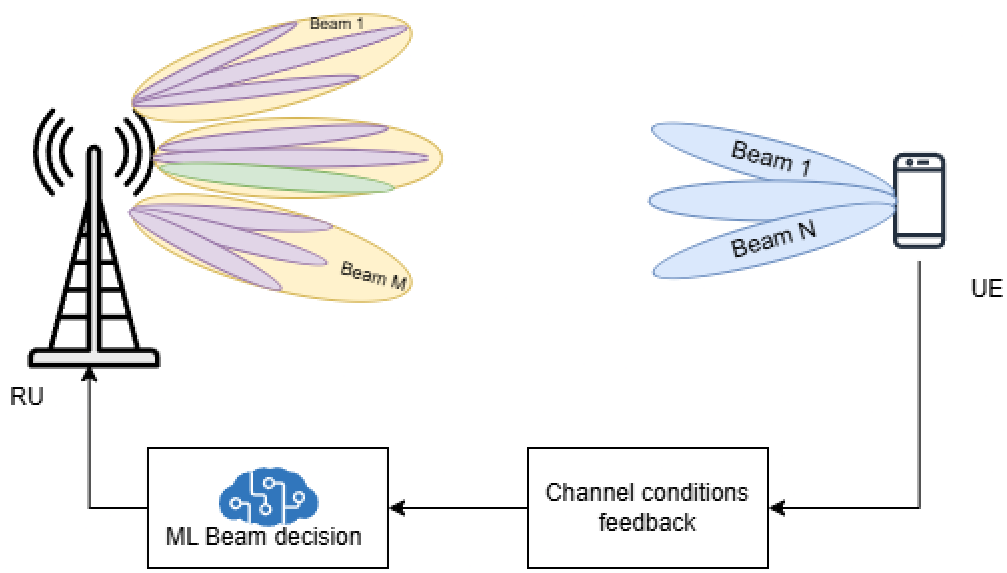

Figure 13: ML based approach for Beam Management selection

The architecture replaces the brute-force RSRP measurement cycle with a predictive classification model. By standardizing features using Z-score normalization and employing an optimized multi-layer perceptron (with dropout and batch normalization for regularization), the model maps spatial inputs directly to the optimal narrow beam.

3.2.3 Results

A downlink channel scenario was configured, incorporating the presence of a signal scatterer. As illustrated in Figure 14, the simulation models the environment following the initial SSB coarse beam selection (P1 procedure), during the subsequent process of determining the narrow beam.

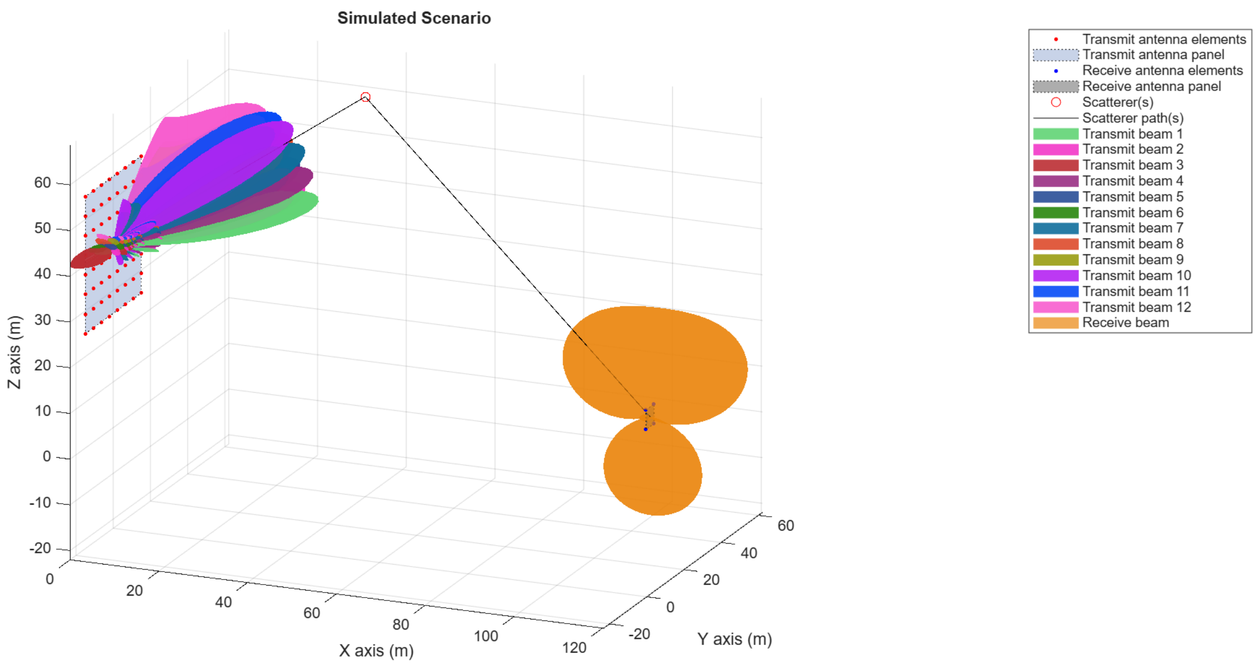

Figure 14: Spatial scene simulated

The simulation is executed across various scatterer and UE positions. The metrics extracted from these dynamic configurations are utilized to characterize the system and identify the optimal narrow beam (CSI-RS).

Furthermore, an initial DNN was developed using the generated dataset to classify the optimal narrow beam. Features obtained from the simulator serve as the input vector, while the target label corresponds to the optimal narrow beam index (corresponding to the highest RSRP beam), which is mapped into 12 distinct categorical classes (Beam 1 through Beam 12). To evaluate the generalization capability of the model, the dataset is partitioned using an 80/20 split for training and validation, respectively.

Figure 15 illustrates the confusion matrix for the CSI-RS narrow beam prediction on the validation dataset. The proposed DNN architecture achieves high classification accuracy across all categories, with most beams exceeding an 85% true positive rate. Misclassifications are predominantly confined to adjacent beam indices, an expected outcome given the spatial overlap of neighboring beam patterns in practical wireless channels.

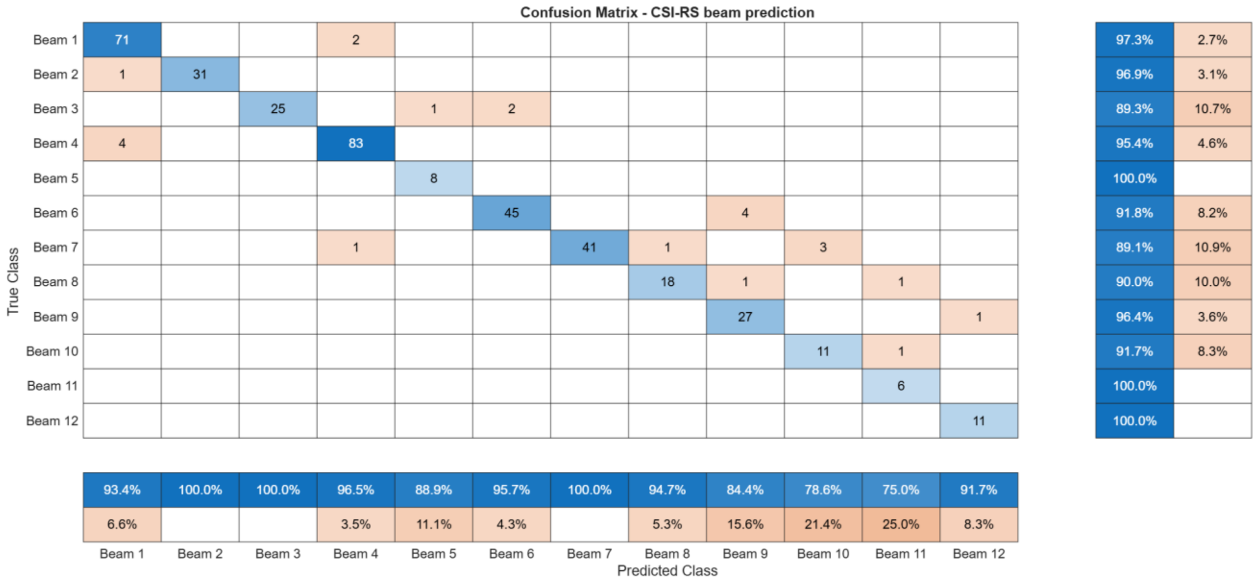

Figure 15: Confusion matrix of narrow beam prediction

3.2.4 KVI Discussion

Environmental:

Replacing exhaustive CSI-RS beam sweeping with predictive DNN drastically reduces the O-RU active RF transmission time and baseband processing load. This lowers energy consumption per bit and enhances material efficiency by maximizing the capacity of existing CF-mMIMO deployments without requiring denser infrastructure over-provisioning.

Economic:

Eliminating the *MxN* measurement complexity reduces the computational overhead and power consumption at the O-DU, directly lowering OPEX. Furthermore, by minimizing L1 pilot overhead and maximizing the effective capacity, C_{eff}, operators can support Tbps throughputs on existing hardware, deferring massive infrastructure CAPEX.

Societal:

The predictive algorithm infers the optimal beam exclusively using L1/L2 spatial and physical telemetry (e.g., Angle of Arrival, path loss, scatterer topology). By avoiding Layer 3 (L3) or application-layer payload inspection, the network achieves strict user data protection and privacy by design.

Co-funded by the European Union

4 AI-based Access Point Placement for Indoor Cell-Free Networks (Fraunhofer HHI)

The evolution toward 6G wireless communication is driving a transition from conventional cellular architectures to user-centric CF Massive MIMO systems. In CF networks, multiple distributed APs jointly serve users simultaneously over the same time-frequency resources, effectively eliminating traditional cell boundaries. This architecture significantly enhances macro-diversity, reduces path loss, and mitigates inter-cell interference. Indoor environments, however, present unique challenges due to their highly site-dependent propagation behavior, including strong multipath propagation, reflections, diffractions, and attenuation caused by walls and construction materials. As a result, inaccurate AP placement can significantly reduce coverage quality, spectral efficiency (SE), and energy efficiency.

Traditional indoor deployment strategies commonly rely on grid-based layouts, heuristic rules, or simplified propagation models. Although computationally efficient, these approaches do not accurately capture the complex characteristics of real indoor environments. More sophisticated optimization methods often consider AP placement independently, without jointly optimizing physical layer (PHY) operations such as beamforming and power allocation. For indoor radio environments, statistical channel models can be insufficient and require a more precise characterization of site-specific propagation. This is achieved using a ray-tracing simulator that accounts for reflections, diffractions, and material-dependent propagation effects. Another major challenge lies in the strong coupling between AP deployment and PHY optimization. While beamforming and uplink power control can be optimized for a fixed network layout, determining the optimal AP locations is challenging due to the combinatorial nature of the search space and the high computational cost associated with evaluating a single network configuration. Evaluation of a single configuration requires computationally expensive ray-tracing, channel generation followed by PHY optimization, rendering exhaustive search approaches impractical.

Co-funded by the European Union

6GSNS

4.1 Methodology

To address these challenges, this work proposes an AI-driven hierarchical optimization framework. AP placement is treated as an outer-loop optimization problem guided by a surrogate learning model, while the inner loop performs PHY optimization through joint beamforming and power allocation to maximize the minimum user SE (max-min fairness criterion). In addition, site-specific propagation is modeled using ray tracing (RT), enabling accurate characterization of indoor wireless channels while motivating the need for computationally efficient optimization techniques.

The proposed framework is formulated as a bilevel optimization problem consisting of two interconnected stages:

1. The outer optimization stage determines the optimal AP locations within the feasible 3D deployment space. This problem is combinatorial, high-dimensional, and inherently non-convex.
2. For a fixed AP deployment, PHY parameters including beamforming vectors and user uplink transmit powers are jointly optimized through centralized resource-allocation algorithms under per-user power constraints to improve user performance and fairness.

4.1.1 Physical layer (PHY) optimization:

The PHY optimization framework incorporates three main components:

- **Team-MMSE Combining:** A decentralized receive combining strategy that enables cooperation among APs with limited backhaul signaling [17].
- **Fixed-Point Iterative Optimization:** Iterative algorithms are used to solve the weighted max-min SINR optimization problem with guaranteed convergence properties [18,19].
- **Ergodic Rate Evaluation:** Network performance is evaluated using tractable lower bounds on achievable ergodic rates, specifically the Use-and-then-Forget (UatF) bound [20].

4.1.2 Access Point placement:

Since evaluating each AP deployment requires both ray-tracing simulations and PHY optimization, direct exhaustive search and gradient-based optimization become computationally infeasible. To overcome this limitation, a surrogate AI-model is trained to learn the mapping from given AP locations to the resulting minimum SE achieved among the users. Further, the framework employs a Bayesian Optimization (BO) pipeline for efficient exploration of the AP placement space. The BO algorithm navigates the search space guided by an acquisition function. The acquisition function balances between the exploration and exploitation of the search space.

More specifically, a Gaussian Process Regression (GPR) [21] model is used as the surrogate function to make inferences about the expensive-to-evaluate objective function. GPR models are proven to be beneficial in making inferences about functions. GPR models employ a kernel function to learn any prior knowledge about the structure of the function of interest. Conventional kernel functions are not well suited for AP placement problems because AP configurations are unordered sets. To address this, a custom kernel based on the Earth Mover's Distance (EMD) [22] is utilized. This kernel captures geometric similarities between different AP layouts by measuring the minimum transformation cost between configurations, enabling more effective modeling of spatial relationships.

The overall pipeline is illustrated below:

Co-funded by the European Union 6GSNS

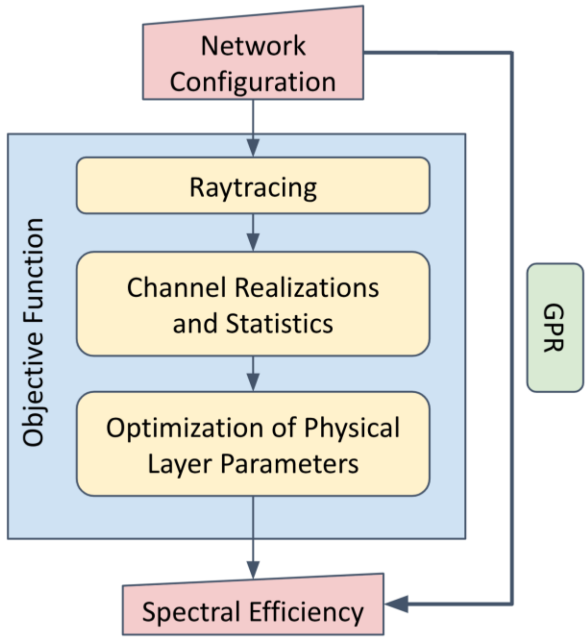

Fig 16. Evaluation pipeline of the objective function. Each evaluation is computationally expensive due to the use of a detailed physical-layer simulation. The GPR model is trained to make inferences about the objective function.

4.2 Results

For each AP deployment, site-specific channel coefficients are generated using a ray-tracing simulator. This produces deterministic multi-path channel responses between APs and users.To capture small-scale fading and evaluate ergodic performance, multiple channel realizations are generated by assigning independent random phases to the ray-tracing paths while preserving the large-scale propagation characteristics of the environment.

The optimization algorithm is initialized with 15 manually selected network configurations. Each network configuration denotes the spatial information (three-dimensional position and antenna directivity) of all the access points in the network. For each configuration, the objective function is evaluated to obtain the minimum SE achieved in the network. The resulting dataset, consisting of pairs of network configurations and their corresponding minimum SE values is used to train the GPR surrogate model.

The Bayesian optimization algorithm is then executed for 30 iterations using the trained surrogate model. The optimization process identifies the network configuration, yielding the highest minimum SE among all evaluated candidates. The configuration proposed by our

GPR model consistently achieves a higher minimum SE compared to manually selected configurations. A comparison of the achieved SE for manually selected configuration versus GPR proposed is provided in Fig 17. In this set of experiments, for each configuration, 3 APs are placed which jointly serve 6 users.

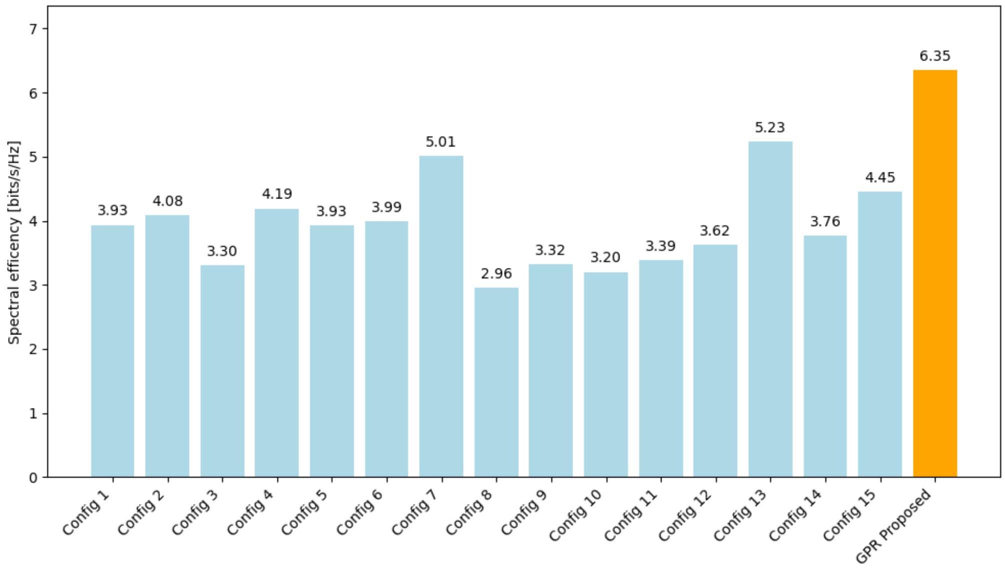

Fig 17. Comparison of achieved SE between manually selected network configurations (in blue) versus GPR-proposed configuration (in orange).

Overall, the proposed GPR-based Bayesian optimization framework enables efficient AP placement optimization for indoor cell-free networks by significantly reducing the number of expensive objective-function evaluations required. By combining site-specific ray-tracing simulations with surrogate-model-guided search, the framework identifies high-quality AP deployments that achieve near-optimal max-min spectral efficiency while maintaining practical computational complexity.

4.3 KVI Discussion

Environmental:

The proposed AI-based access point placement framework reduces network planning effort and deployment costs by minimizing the number of expensive simulation and optimization evaluations required to identify high-performing indoor network layouts. By improving spectral efficiency and resource allocation, the solution can also lower infrastructure requirements and operational expenditures while supporting more reliable wireless connectivity in enterprise and industrial environments.

Economic:

By optimizing access point locations and jointly considering physical-layer resource allocation, the framework improves network energy efficiency and reduces unnecessary infrastructure deployment. More efficient use of radio resources can decrease overall power consumption of indoor wireless networks, contributing to lower energy demand and supporting the sustainability objectives of future 6G communication systems.

Societal:

The project contributes to the development of reliable, high-capacity indoor wireless connectivity that can support digital transformation across workplaces, public buildings, and industrial facilities. Enhanced network coverage, fairness, and user experience enable more inclusive access to advanced digital services, supporting productivity, innovation, and future smart-building applications.

Co-funded by the European Union

5. Conclusion

In this deliverable, we have demonstrated through initial results the potential of AI/ML to transform communication systems towards an AI-native 6G air interface, improving adaptability, scalability, and resource efficiency in distributed network environments, and supporting the requirements of future 6G systems.

The baseband processing methodologies presented in this deliverable address a broad range of network optimization problems inherent to AI-native and cell-free network architectures. We investigate AI/ML-based compression techniques for fronthaul data, a key aspect of distributed implementations. We further demonstrate the superiority of AI/ML-based approaches for CSI-RS overhead reduction and reconstruction compared to baseline techniques, highlighting their ability to learn complex channel structures directly from data.

Prediction is used as an inherent AI/ML capability to reduce the frequency of CSI acquisition and to replace computationally heavy conventional procedures. This is demonstrated in channel estimation for cell-free systems, where prediction reduces CSI acquisition overhead, and in beam management, where it replaces exhaustive beam sweeping procedures.

Furthermore, for access point placement in cell-free networks, the combinatorial and high-dimensional nature of the problem renders traditional optimization approaches intractable. We show that AI/ML methods, in combination with classical optimization tools, enable efficient high-dimensional optimization and scalable decision-making in distributed network environments.

Overall, our initial results demonstrate the potential of AI/ML for baseband processing, whether as a replacement or in combination with traditional techniques, to enable scalable, AI-native, and distributed 6G network architectures.

Co-funded by the European Union

6 References

[1] 6G-MIRAI-HARMONY, D3.1: “D3.1 Definition of initial common scenarios, data management and benchmarking methodology”, January 2026, https://6g-mirai-harmony.eu/wp-content/uploads/2026/02/6G-MIRAI_Deliverable_D3.1_Scenarios_Data_Bench_v1.0.pdf

[2] T. L. Marzetta, “Massive MIMO: An Introduction,” Bell Labs Technical Journal, vol. 20, pp. 11–22, 2015.

[3] V. Q. Rodriguez, F. Guillemin, A. Ferrieux, and L. Thomas, “Cloud-RAN functional split for an efficient fronthaul network,” in 2020 International Wireless Communications and Mobile Computing (IWCMC), (Limassol, Cyprus), pp. 245–250, IEEE, June 2020.

[4] L. Gavrilovska, V. Rakovic, and D. Denkovski, “From Cloud RAN to Open RAN,” Wireless Personal Communications, vol. 113, pp. 1523–1539, Aug. 2020.

[5] Atzeni and A. Tölli, “Channel Estimation and Data Detection Analysis of Massive MIMO with 1-Bit ADCs,” IEEE Transactions on Wireless Communications, vol. 21, pp. 3850–3867, June 2022.

[6] Y. Khorsandmanesh, E. Bjornson, and J. Jalden, “Joint Uplink-Downlink Fronthaul Bit Allocation in Fronthaul-Limited Massive MU-MIMO Systems,” Jan. 2026. arXiv:2601.17423.

[7] J. Ballé, V. Laparra, and E. P. Simoncelli, “End-to-end Optimized Image Compression,” in International Conference on Learning Representations, Feb. 2017.

[8] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” in International Conference on Learning Representations, Feb. 2018.

[9] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, “Nonlinear Transform Coding,” IEEEJournal of Selected Topics in Signal Processing, vol. 15, pp. 339–353, Feb. 2021.

[10] S. Cammerer, F. A. Aoudia, S. Dörner, M. Stark, J. Hoydis, and S. ten Brink, “Trainable Communication Systems: Concepts and Prototype,” IEEE Transactions on Communications, vol. 68, pp. 5489–5503, Sept. 2020.

[11] C. Bian, Y. Shao, and D. Gunduz, “Towards AI-Native Fronthaul: Neural Compression for NextG Cloud RAN,” June 2025. arXiv:2506.06925.

[12] 3GPP RP-251881, "New SID: Study on 6G Radio," Czech Republic, June 9-13, 2025.

[13] E. Björnson et. al, "Towards 6G MIMO: Massive Spatial Multiplexing, Dense Arrays, and Interplay Between Electromagnetics and Processing," Jan. 2024. [Online]. Available: https://arxiv.org/abs/2401.02844.

[14] E. Dahlman, S. Parkvall, and J. Skold, 5G NR: The next generation wireless access technology. Academic Press, 2020.

[15] Chen, Zirui, et al. "Channel deduction: A new learning framework to acquire channel from outdated samples and coarse estimate." *IEEE Journal on Selected Areas in Communications* 43.3 (2025): 944-958.

[16] 3GPP TR38.802, "Study on New Radio Access Technology Physical Layer Aspects", September, 2017.

[17] L. Miretti, E. Björnson, and D. Gesbert, "Team MMSE precoding with applications to cell-free massive MIMO," IEEE Trans. Wireless Commun., vol. 21, no. 8, pp. 6242–6255, 2022.

[18] R. L. G. Cavalcante, L. Miretti, and S. Stanczak, "Characterization of the weak Pareto boundary of resource allocation problems in wireless networks – implications to cell-less systems," ICC 2023, pp. 5010–5016.

[19] C. J. Nuzman, "Contraction approach to power control, with non-monotonic applications", in IEEE GLOBECOM 2007 - IEEE Global Telecommunications Conference, 2007, pp. 5283-5287.

[20] E. Bjornson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," Foundations and Trends in Signal Process., vol. 11, no. 3-4, pp. 154–655, 2017.

[21] R. Garnett, M. Osborne, and S. Roberts, "Bayesian optimization for sensor set selection," in Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks, 04 2010.

[22] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," International Journal of Computer Vision, vol. 40, pp. 99–121, 11 2000.

Co-funded by the European Union