

6G-MIRAI: Initial Datasets for Benchmarking and Validation

Project information

Project name	Machine Intelligence based Radio Access Infrastructure
Project acronym	6G-MIRAI
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	3
Associated Task	T3.1, T3.2
Associated Deliverable	D3.2
Main authors	Robbert Beerten (Editor) (KUL)
Contributors	Cel Thys (KUL), Utku Uçak (HHI), Fariba Armandoust (HHI), Mahdi Shahabi (SEQ), , Ismael Adalid (ISR D), Paul Almasan (Telefonica), Tanguy Kerdoncuff (Ericsson)
Reviewers	Rodney Martinez Alonso (KUL), Efstathios Katranaras (SEQ), Mahdi Shahabi (SEQ)
Type	DATA
Dissemination level	Public
Comment	

Document revision history

Version	Date	Changes	Author
V0.1	13/04/2026	Table of contents	Robbert Beerten
V0.5	01/07/2026	Initial contribution	All
V1	29/09/2026	Final additions	All

Disclaimer

The content of this document reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains. While the information contained in the documents is believed to be accurate, the authors(s) or any other participant in the 6G-MIRAI consortium make no warranty of any kind with regard to this material including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Neither the 6G-MIRAI Consortium nor any of its members, their officers, employees, or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein. Without derogating from the generality of the foregoing neither the 6G-MIRAI Consortium nor any of its members, their officers, employees, or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

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.

© 6G-MIRAI Consortium. This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation, or both. Reproduction is authorised provided the source is acknowledged.

Contents

Contents	3
Acronyms and Definitions	5
Table Of Partners	7
Executive Summary	8
1 Introduction.....	9
2 FR3 PA Response (KU Leuven)	10
2.1 Dataset Table	10
2.2 Scenario.....	10
2.3 Dataset Content.....	12
3 FR2 Beam Management (IS-Wireless)	13
3.1 Dataset Table	13
3.2 Scenario.....	13
4 UL Power Control Dataset (IS-Wireless).....	15
4.1 Dataset Table	15
4.2 Scenario.....	16
4.3 Dataset Content.....	16
5 CSI-RS Measurement (Fraunhofer)	18
5.1 Dataset Table	18
5.2 Scenario.....	18
5.3 Dataset Content.....	20
5.3.1 Example File List.....	21
5.3.2 Metadata Example.....	21
6 3GPP datasets for AI/ML-based CSI-RS overhead reduction (Ericsson)	22
6.1 Dataset Table	22
6.2 Scenario.....	23
7 Site-Specific MIMO Channel Generation with GenAI (Telefónica)	26
7.1 Dataset Table	26
7.2 Scenario.....	27
8 Dynamic Channel Deduction (Sequans)	29

8.1	Dataset Table	29
8.2	Scenario.....	29
8.3	Dataset Content and Preprocessing Pipeline	30
9	Outlook	32
	References	33

Acronyms and Definitions

ASD	Angular Standard Deviation
ASD	Angular Standard Deviation
COTS	Commercial Off The Shelf
CSI	Channel State Information
CSI-RS	Channel State Information Reference Signal
DMRS	Demodulation Reference Signal
DU	Distributed Unit
DUT	Device-Under-Test
FR3	Frequency Range 3
MIMO	Multiple-Input-Multiple-Output
NZP-CSI-RS	Non-Zero Power Channel State Information Reference Signal
NZP-SI-RS	Non-Zero Power Channel State Information Reference Signal
O-DU	O-RAN Distributed Unit
O-RU	O-RAN Radio Unit
PA	Power Amplifier
RSRP	Reference Signal Received Power
RU	Radio Unit
SE	Spectral Efficiency
SINR	Signal to Interference and Noise Ratio
TDD	Time Division Duplexing
TR	Technical Report
TRP	Transmission and Reception Point
TX	Transmitter
UE	User Equipment

Uma	Urban Macro
Umi	Urban Micro
USRP	Universal Software Radio Peripheral

Table Of Partners

Short Name	Partner
Ericsson France	ERICSSON FRANCE
Fraunhofer	FRAUNHOFER GESELLSCHAFT ZUR FORDERUNG DER ANGEWANDTEN FORSCHUNG EV
Telefonica	TELEFONICA INNOVACION DIGITAL SL
ISRD	ISRD SP Z O.O. (IS-WIRELESS)
SEQ	SEQUANS COMMUNICATIONS SA
KUL	KATHOLIEKE UNIVERSITEIT LEUVEN

Executive Summary

This deliverable presents the initial datasets and simulation environments developed and collected in the 6G-MIRAI-HARMONY project to support the benchmarking and validation of AI/ML-based solutions for future 6G radio access networks. The work contributes to the project's objective of establishing a common framework for data, benchmarking, and validation, enabling reproducible evaluation of the solutions developed across the project.

The deliverable describes the collected datasets and simulation environments together with the corresponding generation or collection procedures, data characteristics, formats, relevant scenario and system assumptions, and supporting metadata. Particular attention is given to the documentation and organisation of the datasets to facilitate their interpretation, reuse, and reproducibility.

The datasets and associated methodologies presented in this deliverable constitute an initial release and will continue to evolve as the project progresses. Further datasets, refinements, and validation results will feed into the subsequent benchmarking and demonstration activities and will ultimately contribute to the consolidated results on common scenarios, datasets, validation, and benchmarking reported later in the project.

1 Introduction

This deliverable presents the initial collection of datasets and simulation environments developed, generated, measured, or prepared within the 6G-MIRAI-HARMONY project to support the development, training, benchmarking, and validation of AI/ML-based solutions for future wireless communication systems.

Due to the heterogeneous nature of the considered use cases, the datasets differ in their origin, structure, level of processing, and intended application. Some datasets originate from physical measurement campaigns, while others are generated through system-level or link-level simulation environments. Where applicable, the corresponding scenario assumptions, measurement or simulation configurations, metadata, preprocessing procedures, and dataset structures are documented to support their correct interpretation and reuse. Particular attention is given to the documentation and organisation of the datasets to support their interpretation, reuse, and reproducibility, building on the methodological framework established in the initial WP3 activities, documented in deliverable D3.1 [1].

The remainder of this deliverable is organized according to the individual datasets and their associated use cases. Section 2 describes the KU Leuven FR3 PA response dataset, including its measurement setup and dataset content. Section 3 presents the FR2 beam-management dataset and the associated beam-refinement scenario by IS-Wireless. Section 4 describes the uplink power-control dataset for Cell-Free network deployments by IS-Wireless. Section 5 presents the CSI-RS measurement dataset by Fraunhofer, including its file structure and metadata. Section 6 introduces the Ericsson 6G 3GPP datasets intended for AI/ML-based CSI-RS overhead reduction and describes the corresponding simulation environment. Section 7 presents the dataset for site-specific MIMO channel generation using generative AI by Telefónica. Section 8 describes the dataset and preprocessing pipeline for dynamic channel deduction and prediction by Sequans.

2 FR3 PA Response (KU Leuven)

This dataset provides aligned input and output IQ (In-phase and Quadrature) measurements of three amplifiers intended for FR3 frequency range (7.125 GHz to 24.25 GHz). The devices-under-test (DUTs) are: Analog Devices' ADL9006, Talent Microwave TLPA2G22G-43-43-HS and Qorvo TGA3042. The data captures the complex nonlinearities and memory effects inherent in high-frequency wideband amplification. It supports research on Digital Predistortion (DPD) and PA behavioural modelling using methods such as Volterra series or Neural Networks, as well as power efficiency analysis in 6G and beyond wireless systems. Captures are provided in formats compatible with MATLAB and Python/OpenDPD.

2.1 Dataset Table

Dataset Name	FR3 PA response dataset
Contributor	KU Leuven
Intended use case	Power Amplifier (PA) behavioural modelling, digital pre-distortion, HW impairment simulation
Measured / Simulated / Synthetic / Hybrid	Measurement data from instrument testbench
Status	Initial Dataset
Estimated Volume	Approx. 7MB / capture, with more captures being added in time.
Scenario	Standalone PA in a VNA-based testbench. Carrier frequency inside the FR3 band, single carrier M-QAM waveform. Multiple captures of input & output waveforms for each PA device-under-test, to support training/testing/validation split.
Dataset Repository	https://doi.org/10.48804/306IIM
License	https://creativecommons.org/licenses/by/4.0/

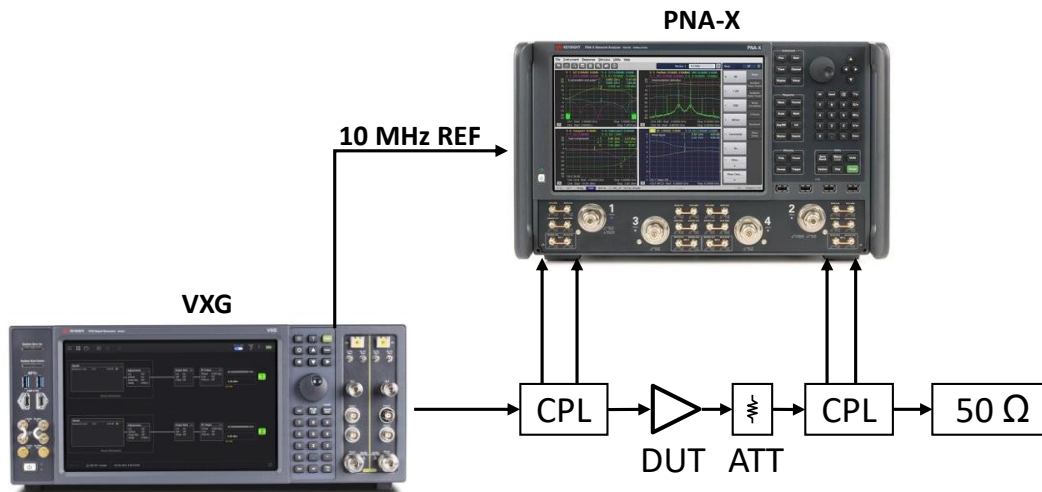
Table 1: Dataset Table for the PA Response Measurement

2.2 Scenario

Below is a picture of the measurement setup used for dataset generation. The device under test (DUT) is connected to the vector network analyzer (VNA, mentioned as 'PNA-X' on the diagram) using two bidirectional couplers. The setup further includes a high power 50 Ohm load and an output attenuator to limit power into the instrument. The input waveform is created offline at baseband and is generated at RF by a vector signal generator (VXG) using a sampling rate of 3 Gigasamples-per-second (Gsps). Both instruments share a 10MHz

reference clock. The input and output waveforms are captured on the PNA-X using vector averaging to improve the noise floor.

Figure 1: Measurement Setup for the PA measurement



The table below lists key scenario parameters, including mean input and output power for the amplifier under test. There are three amplifiers measured in the current work: ADL9006 from Analog Devices, Talent Microwave's TLPA2G22G-43-43-HS and Qorvo's TGA3042. In each case they are driven with a 100 MHz input signal using 64-QAM and root-raised-cosine (RRC) pulse shaping. The waveforms are sampled by the VNA at a rate of 1.28 Gsps. The RMS input power level is set such that the amplifier is backed off from the saturation region.

Parameter	DUT 1: ADL9006	DUT 2: TLPA2G22G	DUT 3: TGA3042
Carrier frequency	15.0 GHz	15.0 GHz	7.225 GHz
Signal Bandwidth	100 MHz	100 MHz	100MHz
Modulation	64-QAM	64-QAM	64-QAM
Pulse shaping	RRC, $\alpha = 0.35$	RRC, $\alpha = 0.35$	RRC, $\alpha = 0.35$
RMS Input Power	-5 dBm	-20 dBm	-22dBm
RMS Output power	10 dBm	33 dBm	25 dBm
Output Attenuator	10 dB	30 dB	30 dB
DAC rate (VXG)	3 Gsps	3 Gsps	3 Gsps
Sampling rate (PNA)	1.28 Gsps	1.28 Gsps	1.28 Gsps

Table 2: Parameters for the the different PAs that were measured

2.3 Dataset Content

Filename	Content
"ADL9006_64QAM_100MHz_n2dBm/*.mat" "ADL9006_64QAM_100MHz_n2dBm/*.csv" "TLPA2G22G-43-43-HS_64QAM_100MHz_n20dBm/*.mat" "TLPA2G22G-43-43-HS_64QAM_100MHz_n20dBm/*.csv" "TGA3042_64QAM_100MHz_n22dBm/*.mat" "TGA3042_64QAM_100MHz_n22dBm/*.csv"	Captures of PA input and output complex waveforms. Each waveform consists of 262144 IQ samples. For each DUT there are 3 independent captures, each taken at a different time ("Train", "Test", "Extra"). Captures are provided in CSV and MAT format
main_script.m	Matlab script showing how to load and explore the measured data

Table 3: File Contents of the different PA Measurements

3 FR2 Beam Management (IS-Wireless)

The scenario is designed to represent a realistic mmWave propagation environment in which directional transmission is required to compensate for the high path loss associated with higher carrier frequencies. It focuses specifically on the downlink P-2 beam-management procedure, where the network refines the transmit direction after the initial beam-acquisition stage while the UE maintains a fixed receive beam. The following subsections describe the propagation assumptions, antenna configuration, beam-generation method, angular sampling strategy, and beam-selection criterion used in the simulation.

3.1 Dataset Table

Dataset Name	FR2 Beam Management
Contributor	MATLAB
Intended use case	Beam selection
Measured/ Simulated /Synthetic /Hybrid	From MATLAB simulation measurements are extracted from the beam selection considering the scenario channel model. Data extracted is used as part of ML based approach
Status	N/A
Estimated Volume	20 MB
Scenario	UMa, UMi
Dataset Repository	N/A
License	N/A

Table 4: Dataset Table for the FR2 Beam Management simulations

3.2 Scenario

This scenario focuses on the downlink transmit-end beam refinement procedure, to simulate a realistic scattering environment. The setup is designed to address the challenges of Frequency Range 2 (FR2), where millimeter-wave (mmWave) frequencies between 24.25 GHz and 52.6 GHz are subject to significant path loss and penetration loss

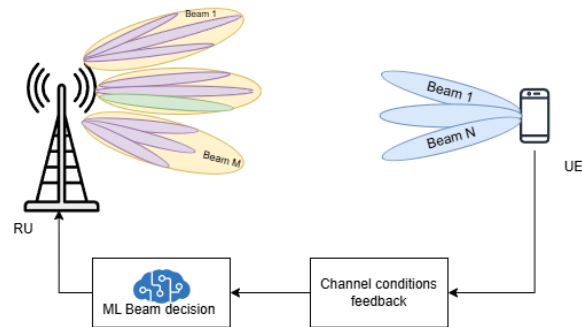


Figure 2: Considered Scenario

To mitigate these effects, the configuration implements Beam Management Procedure 2 (P-2) as defined in TR 38.802, which specifies the network-side sweep of refined transmit beams via CSI-RS while the UE receive beam remains fixed, but leaves the beam-generation method and codebook design entirely to implementation. Accordingly, the transmit beams in this scenario are not selected from a stored, quantized codebook table; instead, each Non-Zero-Power Channel State Information Reference Signal (NZP-CSI-RS) resource is mapped to an array steering vector computed for a specific azimuth/elevation direction. While this expression is mathematically DFT-like, it is evaluated at continuous angles rather than restricted to a finite set of pre-quantized phase states, so it constitutes a custom, non-standardized beam set rather than a formal DFT codebook.

The angular spacing between the resulting beams is derived from the transmit array's own half-power beamwidth approximately $12\text{--}13^\circ$ for the 8×8 array and carrier frequency used here with beam directions distributed at roughly this spacing across the angular range identified during the initial acquisition phase (P-1), so that adjacent beams overlap close to their -3 dB points and minimize coverage gaps. Throughout this process, the User Equipment (UE) maintains a fixed receive beam to isolate the performance of the transmitter's refined beams. The optimal transmit beam is ultimately selected by evaluating the Reference Signal Received Power (RSRP) measurements for each CSI-RS resource, ensuring the highest possible directivity and link budget efficiency for the communication session.

4 UL Power Control Dataset (IS-Wireless)

This section describes the synthetic dataset used to support the evaluation of uplink power-control strategies in RU-rich cell-free massive MIMO and O-RAN-style deployments. The dataset is designed to provide a controlled and reproducible simulation input in which the number of O-RUs is comparatively high with respect to the number of active UEs, allowing different power-control approaches to be assessed under dense distributed-radio conditions. Rather than providing precomputed performance labels, the dataset captures the underlying network geometry, mobility, associations, and channel statistics required by the project simulation pipeline. The following subsections describe the scenario configuration, dataset contents, and its use as a common input for conventional and learning-based uplink power-control methods.

4.1 Dataset Table

Dataset Name	Simulated dataset
Contributor	ISRD
Intended use case	Dynamic uplink power allocation
Measured / Simulated / Synthetic / Hybrid	Simulated
Status	N/A
Estimated Volume	280 .npz files, 4.13GB total, ~130.400 timestamps per scenario
Scenario	500x500m, 4DUs, 3-5 RUs per DU (randomly placed), 6-10 UEs, 0- km/h
Dataset Repository	N / A
License	N / A
Description	This is a synthetic RU-rich uplink power-control dataset, not a measured over-the-air dataset; the raw .npz files contain geometry, mobility, O-RU/O-DU mapping, and spatial covariance matrices, while SE/SINR/TX-power/reliability values are produced later during simulator runs.

Table 5: Dataset Table for the UL Power Control simulations

4.2 Scenario

We use a synthetic RU-rich uplink power-control dataset generated by the project scenario-generation pipeline. The dataset models a cell-free massive MIMO / O-RAN-style deployment over a $500 \text{ m} \times 500 \text{ m}$ area with 4 O-DUs and 12–20 O-RUs, corresponding to 3–5 O-RUs per O-DU. Each scenario contains 6–10 UEs with either static or low-mobility trajectories at 0, 2, 4, or 6 km/h. The RU-rich configuration is designed to evaluate uplink power-control behavior in an environment where the number of radio units is high relative to the number of active users.

The raw scenario files store O-DU, O-RU, and UE geometry, UE mobility, O-RU/O-DU associations, and per-time-block spatial covariance matrices for 4-antenna O-RUs. The dataset is used as a common simulator input for open-loop, fixed closed-loop, adaptive closed-loop, and RL-based closed-loop uplink power-control policies. Final performance quantities such as SINR, spectral efficiency, transmit power, serving-cluster behavior, and reliability-proxy metrics are generated during simulator/evaluator runs rather than stored as static labels in the dataset.

4.3 Dataset Content

Parameter	Value / Range
Dataset type	Synthetic simulated RU-rich uplink scenarios
Area	$500 \text{ m} \times 500 \text{ m}$
O-Dus	4
O-RUs per O-DU	3, 4, or 5
Total O-Rus	12, 16, or 20
UEs	6–10
O-RU antennas	4
UE speeds	0, 2, 4, 6 km/h
Scenario duration	10, 20, 40 s

Control/coherence interval	50 ms
Time blocks	200, 400, or 800
Channel covariance	Local-scattering covariance, ASD = 20°

Content	Description
O-DU positions	Fixed O-DU coordinates arranged on a 2×2 deployment grid.
O-RU positions	Random O-RU positions inside each O-DU sub-region, with O-RU and O-DU IDs.
UE trajectories	Per-time-block UE x/y positions and UE IDs.
UE movement directions	Complex-valued UE direction vectors for each time block
O-RU/O-DU mapping	Mapping between each O-RU and its serving O-DU.
Spatial covariance	Per-time, per-UE, per-O-RU 4×4 covariance matrices.
Scenario metadata	UE count, O-DU count, O-RUs per O-DU, speed, duration, coherence interval, seed, and generator settings.
SE / SINR / TX-power outputs	Generated during simulator or evaluator runs, not stored as static raw dataset labels.
Reliability metrics	Postprocessed from simulated SE/SINR traces; current packet-level reliability is represented through proxy metrics unless traffic-specific experiments are enabled.

5 CSI-RS Measurement (Fraunhofer)

The Fraunhofer HHI CSI-RS MIMO Dataset is being developed in collaboration with SEQUANS to support the development and validation of machine learning algorithms related to CSI-RS. It will contain measured multi-antenna channel estimates and associated metadata in SigMF format. Measurements will cover indoor over-the-air line-of-sight and non-line-of-sight scenarios, as well as emulated indoor ray-tracing and outdoor 3GPP channel scenarios. The setup combines USRP software-defined radios, antenna arrays, and a PROPSIM F64 channel emulator, using OpenAirInterface and MATLAB 5G Toolbox as separate measurement platforms. Measurements will operate at FR1 carrier frequencies and support configurations with up to 16 transmit antenna ports.

5.1 Dataset Table

Dataset Name	Fraunhofer HHI CSI-RS MIMO Dataset
Contributor	Fraunhofer HHI
Intended use case	Collaboration with SEQUANS to provide measurement data to validate their CSI-RS-related machine learning algorithm development
Measured / Simulated / Synthetic / Hybrid	Measurements using USRPs, O-RUs, and COTS UEs with OpenAirInterface, GNURadio, and MATLAB, over-the-air with antennas or with a channel emulator
Estimated Volume	2 GB
Scenario	<ul style="list-style-type: none"> Indoor, over-the-air, LOS/NLOS Emulated indoor with ray tracing Emulated outdoor
Dataset Repository	Not published yet
License	{Copyright License}

Table 6: Dataset Table for the CSI-RS MIMO measurements

5.2 Scenario

The Fraunhofer HHI measurement setup contains several USRP SDRs, commercial O-RUs, COTS UEs, and a PROPSIM F64 Channel Emulator. It is possible to configure the hardware for several over-the-air or emulated MIMO scenarios. The plan is to run measurement software in several hardware configurations to capture MIMO channel state information (CSI).

Concretely defined measurement scenarios are:

- Indoor over-the-air LOS and NLOS measurements with different placement of the user equipment (UE)
- Outdoor 3GPP channel model measurements with the channel emulator
- Indoor ray-tracing channel models with the channel emulator

OpenAirInterface (OAI) and MATLAB 5G Toolbox will be used as two separate measurement software stacks. OAI will be used as a real-time end-to-end system. In this measurement scenario, an OAI-gNB with CSI-RS will be deployed and an OAI-UE will connect to it. The UE will then perform downlink channel estimation using CSI-RS pilots. The data will be exported using OAI's T-tracer framework and converted to SigMF using an internally developed tool.

The limitation of the OAI-based measurement setup is the maximum CSI-RS ports currently supported by OAI. Because of this, only up to 12 Tx ports are possible. A MATLAB 5G Toolbox-based setup will be used to perform measurements with 16 antennas

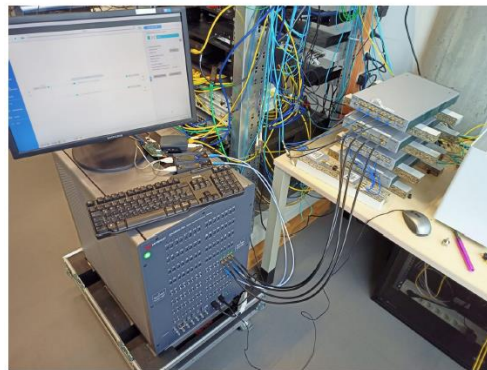


Figure 3: USRPs connected to the channel emulator

For measurements with 16 transmit antennas, MATLAB 5G Toolbox will be used to create a downlink signal with SSBs and 16-port CSI-RS. This signal will be transmitted in a loop from the antenna array with USRPs and received with a receiver USRP and recorded to an IQ file. Channel estimation and SigMF archiving will be implemented as a post-processing step.

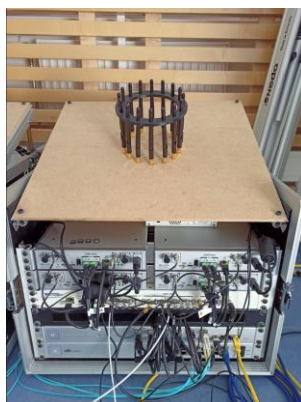


Figure 4: A 16-antenna measurement setup with 4 x USRP N310s connected to a circular array

The measurements will be conducted at FR1 carrier frequencies with a 15.36 Msps sampling rate, up to 16 antenna ports.

5.3 Dataset Content

The dataset will contain multi-antenna CSI data in SigMF format. SigMF is an open-source signal measurement format that has two main file extensions: .sigmf-data and .sigmf-meta. The first extension, .sigmf-data, contains the binary signal data and the .sigmf-meta file with the same name contains metadata about the measurement in JSON format. Basic metadata fields contain information about measurement parameters like the carrier frequency, sampling rate, Tx or Rx antenna array index.

There will be two measurement configurations in terms of the antenna array and the wireless channel. The first configuration consists of the over-the-air measurements with a circular array in an indoor lab environment. The second configuration will use the channel emulator and connect USRPs to the channel emulator with RF cables. The channel emulator can be configured to define artificial arrays and channel models.

5.3.1 Example File List

- csirs-indoor-los-2x1_UE_PHY_DL_CSIRS_CHANNEL_ESTIMATE_FREQ_000001.sigmf-data
- csirs-indoor-los-2x1_UE_PHY_DL_CSIRS_CHANNEL_ESTIMATE_FREQ_000001.sigmf-meta
- csirs-indoor-los-2x1_UE_PHY_DL_CSIRS_CHANNEL_ESTIMATE_FREQ_000002.sigmf-data
- csirs-indoor-los-2x1_UE_PHY_DL_CSIRS_CHANNEL_ESTIMATE_FREQ_000002.sigmf-meta
- csirs-indoor-los-2x1_UE_PHY_DL_CSIRS_CHANNEL_ESTIMATE_FREQ_000003.sigmf-data
- csirs-indoor-los-2x1_UE_PHY_DL_CSIRS_CHANNEL_ESTIMATE_FREQ_000003.sigmf-meta
- csirs-indoor-los-2x1_UE_PHY_DL_CSIRS_CHANNEL_ESTIMATE_FREQ_000004.sigmf-data
- csirs-indoor-los-2x1_UE_PHY_DL_CSIRS_CHANNEL_ESTIMATE_FREQ_000004.sigmf-meta

5.3.2 Metadata Example

```
{ "global": {
  "core:datatype": "ci16_le",
  "core:version": "1.2.5",
  "core:sample_rate": 15360000.0,
  "core:author": "Fraunhofer HHI",
  "core:description": "CSI RS Channel Estimates from indoor LOS environment."
},
"captures": [
  {
    "core:sample_start": 0,
    "core:frequency": 3319680000.0,
    "core:datetime": "2026-05-27T08:50:26.228091+00:00"
  }
],
"annotations": [
  {
    "core:sample_start": 0,
    "core:sample_count": 512,
    "core:label": "UE_PHY_DL_CSIRS_CHANNEL_ESTIMATE_FREQ",
    "oai:event_id": 343,
    "oai:record_message_index": 3,
    "oai:occurrence_index": 1,
    "oai:buffer_field": "chest_f",
    "oai:gNB_ID": 0,
    "oai:UE_ID": 0,
    "oai:frame": 172,
    "oai:slot": 0,
    "oai:antenna": 0,
    "oai:port": 0
  }
]
}
```

6 3GPP datasets for AI/ML-based CSI-RS overhead reduction (Ericsson)

These datasets are used by Ericsson in their contributions to 3GPP standardization as per the 3GPP way of working, where contributors are typically required i) follow a common set of agreed evaluation assumptions, and ii) implement proprietary confidential algorithms to evaluate and compare the performance impact of the feature under study. Due to the confidential nature of the algorithms implemented to generate the data, which in many cases are close to those available in commercial products, the actual datasets cannot be publicly shared. However, the key components of the datasets can be reproduced in other radio simulation tools without confidential components via implementation of the following 3GPP-agreed settings. To be more concrete, it would be possible for anyone to use public 3GPP channel model with configuration bellow, together with additional noise added to simulate pilot transmission followed by any denoising algorithm, which would be a good approximation of the data for the task. But we can't directly share those data as the denoising process is confidential.

6.1 Dataset Table

Dataset Name	6G 3GPP dataset
Contributor	Ericsson
Intended use case	Train and evaluate AI/ML on CSI-RS reconstruction based on partial CSI-RS transmission
Measured / Simulated / Synthetic / Hybrid	Urban Macro TR38.901 (3GPP) synthetic
Estimated Volume	~ 40G for AI/ML training and ~10G for validation
Scenario	Described in detail bellow
Dataset Repository	N/A
License	N/A

Table 7: Dataset Table for the AI/ML-based CSI-RS overhead reduction measurements

6.2 Scenario

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 2)
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

Table 8: Dataset setting used for AI/ML model training

The dataset setting shown in Table 8 is used to train an AI/ML model at the UE side to reconstruct the full port channel based on partial pilot (CSI-RS) transmitted by the BS toward the UE. The input of the AI/ML model is the partial pilot received after UE denoising. As shown in Figure 5 below, only a subset of the CSI-RS port is used to transmit downlink pilot toward the UE. The AI/ML is trained with the ground truth being the entire channel with all ports. To allow future training based on real data, the ground truth part of the dataset is based

on practical denoising not the actual genie channel that would not be available in practical conditions.

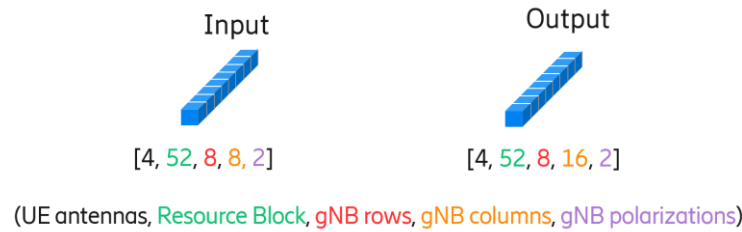


Figure 5: Example of input and output tensor dimension of the dataset (complex value).

This dataset setting is designed to train an AI/ML model at the UE side to reconstruct the full multi-port channel from partially transmitted downlink pilots (CSI-RS) sent by the BS. The model input consists of the received partial CSI-RS after UE-side denoising, where only a subset of CSI-RS ports is used for pilot transmission, as illustrated in Figure 6. The model is trained using the full-port channel as ground truth. Figure 6 provides an example of the total shape of the input and output with a basic partial sounding pattern with only half of the gNB columns. To ensure applicability to future training with real measurements, the ground truth is obtained through practical denoising rather than an idealized genie channel, which would not be available under realistic operating conditions. All SNR conditions are kept during training, even very low SNR, such that the AI/ML model is able to work under any condition.

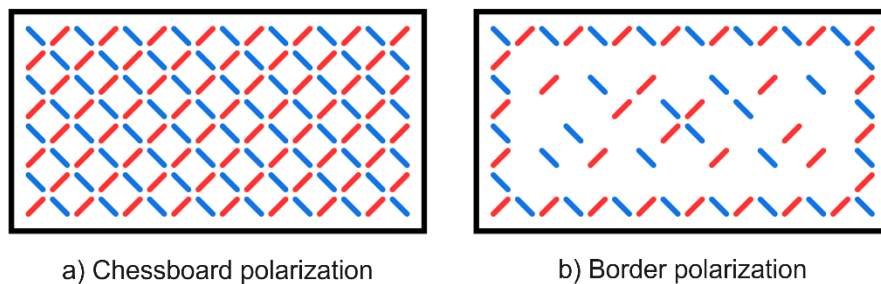


Figure 6: Examples of good CSI-RS partial port sounding patterns with 50% of ports (left) and 25% (right).

Below is a list of 3GPP contributions where those configurations were used to generate simulation results:

- Ericsson, "TDoc R1-2505180: AI/ML Use Cases for 6GR Air Interface," 3GPP TSG-RAN WG1 Meeting #122, Bengaluru, India, August 25-29, 2025.
- Ericsson, "TDoc R1-2506762: AI/ML Use Cases for 6GR Air Interface," 3GPP TSG-RAN WG1 Meeting #122bis, Prague, Czech, October 13-17, 2025.

- Ericsson, “TDoc R1-2508365: AI/ML Use Cases for 6GR Air Interface,” 3GPP TSG-RAN WG1 Meeting #123, Dallas, USA, Nov 17-21, 2025.
- Ericsson, “TDoc R1-2601038: Views on aspects of downlink-based CSI Acquisition,” 3GPP TSG-RAN WG1 Meeting #124, Gothenburg, Sweden, February 9-13, 2026.
- Ericsson, “TDoc R1-2603259: Views on aspects of downlink-based CSI Acquisition,” 3GPP TSG-RAN WG1 Meeting #124bis, Malta, MT, April 13-17, 2026.

7 Site-Specific MIMO Channel Generation with GenAI (Telefónica)

This dataset is used to train generative artificial intelligence (GenAI) models for site-specific MIMO channel modeling. Specifically, it was created using ray-tracing in a realistic urban environment, providing site-specific MIMO radio channel samples for a realistic deployment scenario in downtown London. The data is used to train and evaluate the models developed in respective WP1 work.

The dataset comprises three scenario variants at different frequencies and propagation conditions: 28 GHz (millimeter-wave) line-of-sight (LoS) channels and 3.5 GHz (sub-6 GHz) channels under both LoS and non-line-of-sight (NLoS) conditions. We leveraged Sionna RT [2] to simulate real site-specific channel values, allowing reproducible experiments and eliminating the need for costly measurement campaigns. Key parameters and scenario details are summarized below, followed by a description of how the data is generated.

7.1 Dataset Table

Dataset Name	Simulated site-specific wireless MIMO channel data
Contributor	Telefónica
Intended use case	AI-native RAN (physical-layer) modelling; e.g. generative channel augmentation for CSI compression & beam alignment
Measured / Simulated / Synthetic / Hybrid	Simulated: high-fidelity raytracing of real city site; fully synthetic channel data
Estimated Volume	11 GBs
Scenario	Urban microcell in downtown London ($200 \times 200 \text{ m}^2$ area with real building geometry from OpenStreetMap; single BS with 32-antenna UPA, UEs randomly distributed outdoors within $\sim 100 \text{ m}$ radius from the BS 3 variants: 28 GHz (LoS), 3.5 GHz (LoS), 3.5 GHz (NLoS)
Dataset Repository	https://huggingface.co/datasets/PaulAlm/GenAI_Channel_Modeling_Datasets
License	Not specified.

Table 9: Dataset Table for the site-specific wireless MIMO channel data simulations

7.2 Scenario

The scenario consists of a single base station (BS) deployed in the center of a dense urban environment spanning an area of 200×200 meters [3]. The city layout was obtained from OpenStreetMap [4] and the BS transmits on either millimeter-wave (28 GHz) or sub-6 GHz (3.5 GHz) frequencies, with corresponding BS antenna heights of 8.5 m and 25 m, respectively. Each BS uses a 32-element (4×8) uniform planar array (UPA), and each user equipment (UE) has a 4-element (2×2) UPA. UEs are uniformly distributed inside a 100 m radius circle around the BS to represent typical user distributions within a cell, excluding indoor positions.

In this dataset, we leveraged Sionna RT [2] to launch the ray tracing simulations for each UE position to generate the channel state matrix. Using the computed paths with ray tracing, a full complex channel matrix H is constructed following the standard MIMO channel equation. Specifically, the channel matrix has dimensions: $N_u \times N_b$ where $N_u=4$ (UE antennas) and $N_b=32$ (BS antennas). Each sample in the dataset is thus a MIMO channel realization tied to a specific set of 2D coordinates.

There are three dataset variants created to evaluate GenAI-based models under progressively more challenging conditions. All three variants share the same base scenario (i.e., London urban area with identical BS/UE antenna setups and ray-tracing parameters), but differ in frequency and propagation conditions:

- 28 GHz LoS: 28 GHz carriers with LoS users only. All UE positions in this subset have at least one unobstructed path to the BS, modeling a pure line-of-sight millimeter-wave scenario.
- 3.5 GHz LoS: 3.5 GHz carriers with LoS users only. All UE positions in this subset have at least one unobstructed path to the BS, modeling a pure line-of-sight millimeter-wave scenario.
- 3.5 GHz LoS+NLoS: 3.5 GHz carriers, with LoS and NLoS users. This dataset includes the LoS users from the previous dataset but has a new set of NLoS UEs.

The increasing complexity, from a straightforward LoS scenario to a mixed LoS+NLoS scenario at 3.5 GHz, provides increasing multipath richness, enabling a systematic testing of generative model performance under varying propagation complexity.

For each dataset variant, the UE locations are divided into non-overlapping training, validation, and test subsets. Only channels from training positions are used to train the GenAI models, while the fidelity and task performance of generated channels are evaluated on unseen test positions, ensuring robust generalization. Each sample consists of a structured MIMO channel tensor (real and imaginary components) annotated by the corresponding UE coordinate and scenario information (e.g., LoS vs NLoS label).

8 Dynamic Channel Deduction (Sequans)

This dataset is used by Sequans to train, validate, and benchmark deep-learning-based spatio-temporal channel deduction and prediction algorithms against conventional linear spatio-temporal predictors. The dataset is generated using the standardized DeepMIMO framework, leveraging high-fidelity 3D ray-tracing simulations to capture realistic physical environment dynamics.

8.1 Dataset Table

Dataset Name	Preprocessed DeepMIMO O2_Dynamic Channel Dataset
Contributor	DeepMIMO (raw dataset) / Sequans (preprocessed dataset)
Intended use case	Spatio-temporal channel tracking, dynamic pilot interpolation, and multi-step CSI prediction.
Measured / Simulated / Synthetic / Hybrid	Simulated (High-fidelity 3D ray-tracing generated via Remcom Wireless InSite).
Estimated Volume	~8.5 GB
Scenario	Urban V2X vehicular corridor with dynamic user trajectories (O2_Dynamic scenario).
Dataset Repository	https://deepmimo.net/ (Raw dataset generator)
License	Open-source for academic and collaborative research use.

Table 10: Dataset Table for the dynamic channel deduction simulations

8.2 Scenario

The selected scenario is the O2_Dynamic outdoor V2X environment, representing a dense urban street grid. This scenario is specifically chosen to evaluate algorithmic performance under user-mobility conditions, where Doppler spreads and rapid phase transitions make conventional linear estimation methods highly sub-optimal.

The physical propagation paths are calculated using deterministic 3D ray tracing, which natively captures the geometric layout of the buildings, foliage, and dynamic vehicle coordinates. The key configuration parameters used by Sequans to generate the channel matrices are outlined in Table 11 below:

Parameter	Value / Configuration
Active Ray-Tracing Scenario	O2_Dynamic (Sub-6 GHz V2X)
Carrier Frequency (f_c)	3.5 GHz
Multipath Components (L)	25 strongest paths
BS Transmit Array (N_t)	64 antennas (Uniform Linear Array - ULA)
Antenna Spacing	0.5λ
UE Receive Configuration (N_r)	Single antenna (SISO/SIMO evaluation)

Table 11: Key configuration setup for channel matrices generation

8.3 Dataset Content and Preprocessing Pipeline

The raw dataset contains sequence-based complex channel tensors $H \in \mathbb{C}^{N_r \times N_t \times N_c \times T}$, where T represents the time dimension. To make this data suitable for deep-learning-based channel deduction, Sequans applies a specific preprocessing pipeline:

- **Pilot Masking (Sparse Sampling):** A downsampling operator is applied to the full channel grid to simulate a highly sparse pilot pattern (e.g., keeping only 20% of the subcarriers), yielding the input pilot matrix H_t^0 .
- **Temporal Windowing:** The time-series data is split into overlapping sequences of length n (representing the number of past available channel states used for deduction).
- **Noise Injection:** Complex AWGN is added to the sparse pilot elements to simulate realistic SNR conditions ranging from -5 dB to 25dB.
- **Data Export Structure:** The processed samples are stored as serialized .npz or .mat tensors structured as:
 - **Input Tensor:** Sparse, noisy pilot history over n past frames: $[Batch\ Size, N_t, N_{pilot}, n, 2]$ (where the last dimension represents the real and imaginary components).
 - **Ground Truth Tensor:** Full, noise-free channel matrix at the target deduction timestamp: $[Batch\ Size, N_t, N_c, 2]$.

9 Outlook

The datasets presented in this deliverable constitute an initial basis for the development, evaluation, and validation of AI/ML-enabled methods within the 6G-MIRAI-HARMONY project. They cover a diverse set of technical challenges, including power-amplifier characterization, beam management, uplink power control, CSI-RS processing, channel modelling, and dynamic channel prediction.

A key objective of the subsequent work is to increase the maturity and usability of the datasets as well as the alignment of datasets usage across the project. Depending on the individual use case, this may include the generation or collection of additional samples. For measurement-based datasets, further measurement campaigns may provide broader coverage of relevant scenarios. For simulation-based datasets, additional parameter combinations and scenario realizations may be generated to improve statistical coverage and support more comprehensive evaluations.

Furthermore, the datasets described in this deliverable originate from different partners and are produced using different measurement platforms, simulation environments, and processing pipelines. Consequently, continued effort will be required to maintain harmonised, clear descriptions of data formats, parameter definitions, scenario assumptions, and evaluation procedures.

Finally, this deliverable contributes to the consolidation of the project's scenarios, datasets, validation procedures, and benchmarking methodology.

References

- [1] 6G-MIRAI deliverable D3.1, "Definition of initial common scenarios, data management and benchmarking methodology", Feb. 2026
- [2] J. Hoydis, *et al.* "Sionna RT: Differentiable ray tracing for radio propagation modeling". *2023 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2023. p. 317-321.
- [3] S. Beyraghi, M. Sadeghian, F.B. Ismail, A. Lozano, P. Almasan, G. and Geraci, 2026. Site-Specific MIMO Channel Generation via Diffusion and Flow Matching: Fidelity, Efficiency, and Downstream Utility. arXiv preprint arXiv:2606.20098.
- [4] M. Haklay and P. Weber, "OpenStreetMap: User-Generated Street Maps," in IEEE Pervasive Computing, vol. 7, no. 4, pp. 12-18, Oct.-Dec. 2008