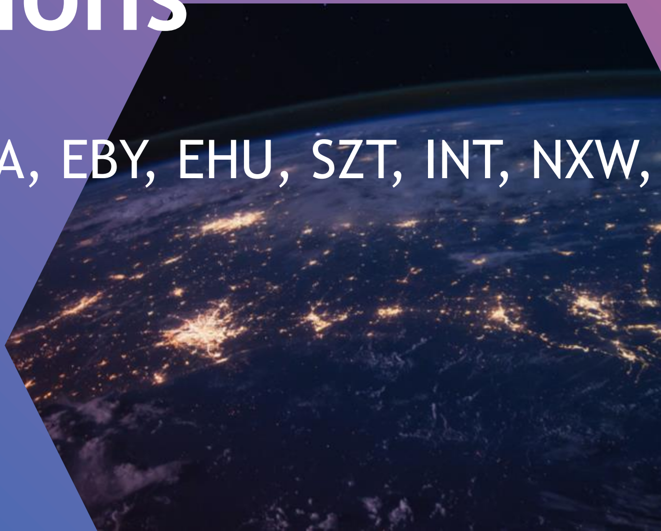
Hexa-X: WP4 – Deliverable D4.3

AI-driven communication & computation co-design: initial solutions

NOF, EAB, ATO, BCO, CEA, EBY, EHU, SZT, INT, NXW, NOG, ORA, OUL, UPI, WIN

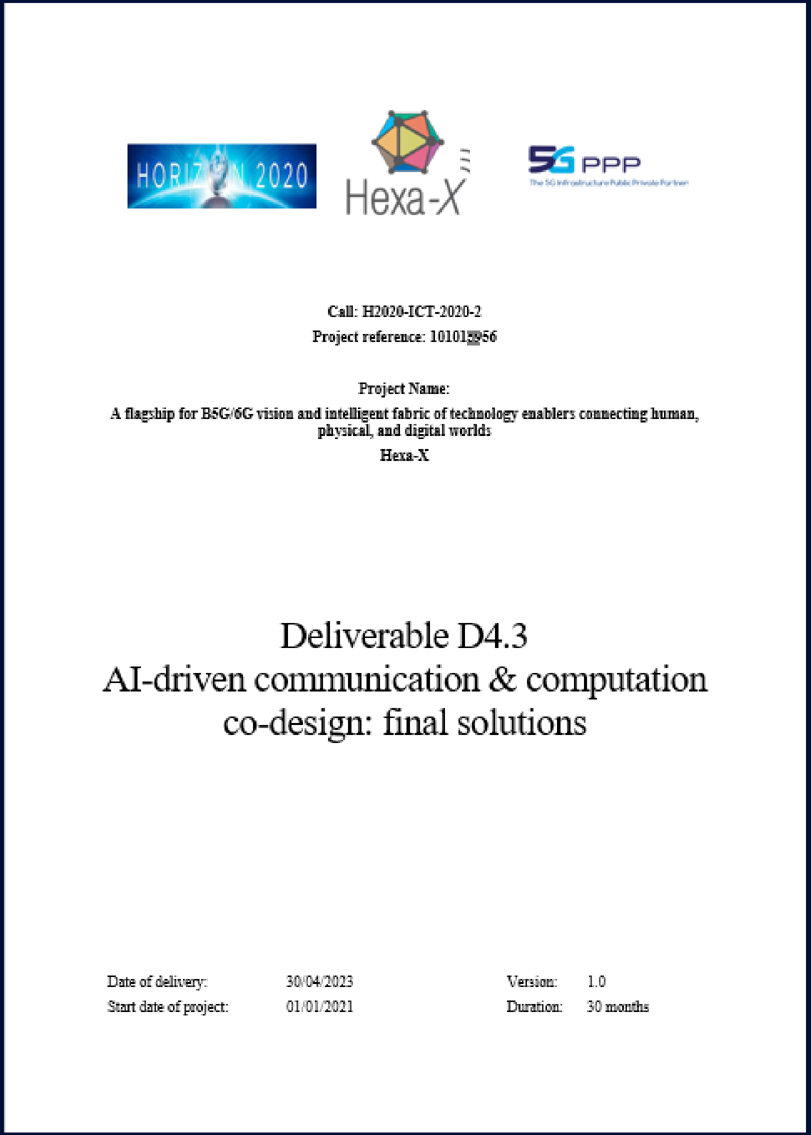
01.05.2023

hexa-x.eu



Mission and Scope

- Hexa-X WP4 (***AI-driven communication and computation co-design***) develops concepts for **AI-based air-interface design** and aims to deliver a **secure and sustainable 6G distributed learning platform** able to optimally support and address distributed edge workloads and learning/ inferencing mechanisms
- This report is the third and final deliverable of project Hexa-X WP4, building on D4.1 and D4.2, and detailing the final set of solutions provided by the technical tasks in the work package, i.e., T4.2 and T4.3 giving summarizes the demonstration activity.
- Technical areas of focus are:
  - **Network performance enhancement using AI/ML in 6G**
  - **6G network as an efficient AI platform**
  - **AI/ML as an enabler for 6G network sustainability**
  - **Privacy, security & trust in AI-enabled 6G**
  - **Demonstration activities - Federated eXplainable AI (FED-XAI) demo**



Call: H2020-ICT-2020-2

Project reference: 101015956

Project Name:

A flagship for B5G/6G vision and intelligent fabric of technology enablers connecting human, physical, and digital worlds

Hexa-X

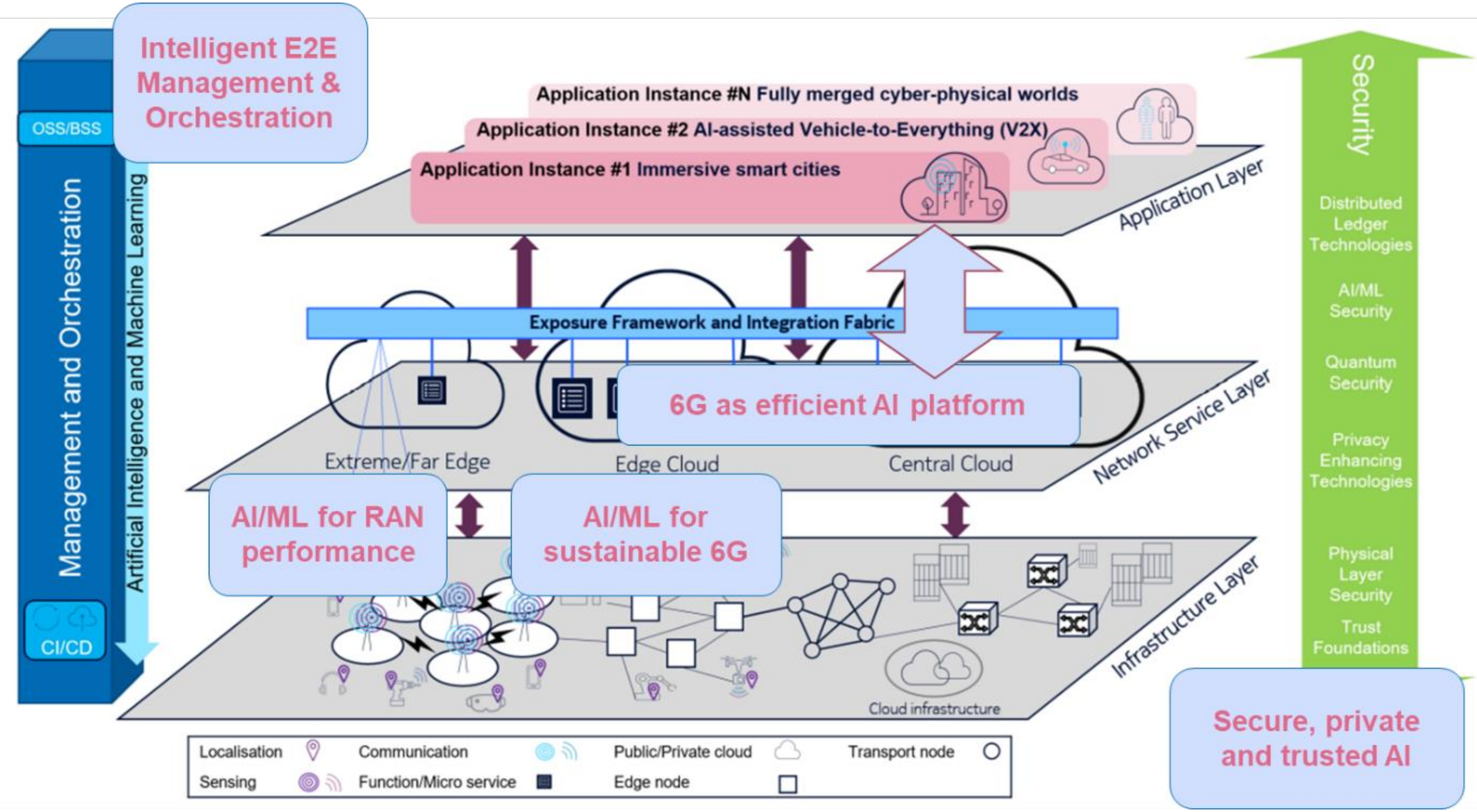
Deliverable D4.3

AI-driven communication & computation co-design: final solutions

Date of delivery:	30/04/2023	Version:	1.0
Start date of project:	01/01/2021	Duration:	30 months

AI-driven communication and compute solutions

- Future 6G network functions and use cases will be intertwined with various forms of learning and intelligence in many aspects including air interface design, data management, optimality of compute and processing functions, network automation & service availability.
- The right-hand illustration shows how Hexa-X WP4 addresses all the above domains with technical enablers, algorithms, joint solution proposals for communication and computation to help fulfilling Hexa-X Connecting Intelligence research challenges.



Intelligent E2E Management & Orchestration
OSS/BSS
Management and Orchestration
Artificial Intelligence and Machine Learning
CI/CD
Application Instance #N Fully merged cyber-physical worlds
Application Instance #2 AI-assisted Vehicle-to-Everything (V2X)
Application Instance #1 Immersive smart cities
Application Layer
Exposure Framework and Integration Fabric
6G as efficient AI platform
Extreme/Far Edge
Edge Cloud
Central Cloud
Network Service Layer
AI/ML for RAN performance
AI/ML for sustainable 6G
Cloud infrastructure
Infrastructure Layer
Security
Distributed Ledger Technologies
AI/ML Security
Quantum Security
Privacy Enhancing Technologies
Physical Layer Security
Trust Foundations
Secure, private and trusted AI
Localisation
Sensing
Communication
Function/Micro service
Public/Private cloud
Edge node
Transport node

Contribution of WP4 technical enablers in network architectural blocks

Hexa-X WP4 quantifiable targets

Quantifiable target #	Title	WP4 task of relevance (*)
T1	Increased AI algorithm robustness to system parameter volatility, lower complexity and significant Bit Error Rate (BER)/ BLock-Error Rate (BLER) gain, as compared to classical approaches	T4.2
T2	Increased AI algorithm robustness to system parameter volatility, lower complexity and efficient resource utilisation and rate gain as compared to classical approaches	T4.2
T3	Resilient communication and compute network services for distributed AI applications in large scales	T4.3
T4	The accuracy of an XAI model within (<10%) of “black box” solutions	T4.3
T5	Energy reduction of a factor of (>10) at the infrastructure level and a factor of (>100) at the user devices’ side, as a result of (network & application) workload offloading and learning/ inferencing task delegations	T4.3
T6	Increased trustworthiness of AI through privacy and security enhancing technologies and AI network intrusion detection capability	T4.3

(*) Task 4.2: AI-driven air interface design
Task 4.3: Methods and algorithms for sustainable and secure distributed AI

Network performance enhancement using AI/ML in 6G

Network performance enhancement using AI/ML in 6G

- **Main emphasis:** how can AI/ML-based solutions enhance the network performance in a quantifiable way?

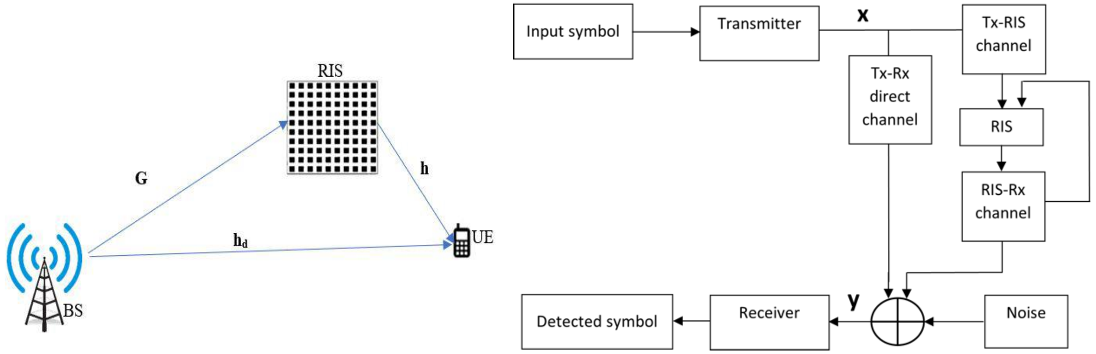
- The first part of the chapter focuses on **radio access network performance improvements over classical design methods**
  - Communication reliability improvements
  - Bit-rate and spectral efficiency improvements
  - Designs accounting for nonlinear distortion

- In the latter part of the chapter, the focus shifts to **improvements in E2E network operation & management**
  - AI/ML-based predictive orchestration
  - Distributed AI for automated UPF scaling in low-latency network slices

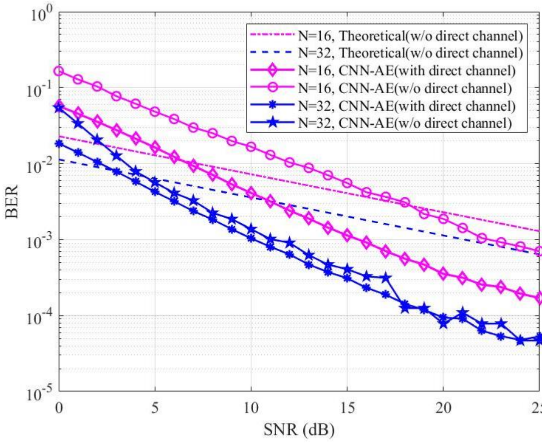
- **KPIs:**
  - Bit Error Rate (BER)/ Block Error Rate (BLER)
  - Channel estimation error
  - Complexity
  - Bit rate or spectral efficiency
  - Flexibility
  - Mobility support
  - Latency
  - Network energy efficiency
  - Inferencing accuracy

ML-based end-to-end learning of RIS-assisted communication systems

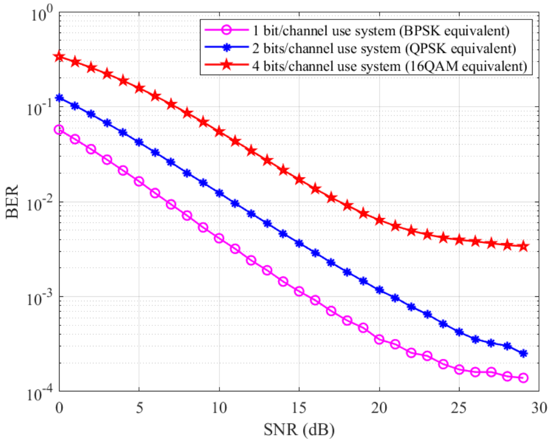
- **Problem/ challenge to be addressed**
  - Reducing the computational complexity and improving the performance of signal processing tasks in RIS-assisted communication systems.
  - Overcome suboptimal solution approaches in modular level optimisation tasks by end-to-end optimisation of the communication system.
- **Final proposed solution**
  - A CNN-based autoencoder to jointly optimise the transmitter, receiver, and the RIS to learn the transmit signals at the BS and reflection coefficients of the RIS, minimising the end-to-end symbol detection error.
  - The autoencoder jointly optimises the sub-tasks of the transmitter, the receiver, and the RIS such as encoding/decoding, channel estimation, phase optimisation, and modulation/demodulation.
- **Evaluation towards 6G KPIs/ KVIs**
  - Improved BER/BLER performance
- **Quantifiable “Connecting Intelligence” targets**
  - T1: improved end-to-end BER/BLER and low complexity processing



The RIS-assisted communication system model and system architecture.



The BER performance (BPSK) of the proposed CNN autoencoder for RIS-assisted communication vs baseline (theoretical) for different RIS sizes.



The BER performance of the CNN autoencoder for RIS-assisted system for higher communication rates.

AI-based enhanced beam selection

The problem

- Beam scanning has increasing overhead for
  - higher frequencies
  - large antenna arrays
  - D-MIMO

Final proposed solution

- Beam identification inspired by compressed sensing
- Exploiting channel sparsity in angular domain
- For N antennas, only $M \ll N$ measurements needed
- Scenario-specific dictionary optimization
  - Sparse AI/ML-based decoder
  - Based on learned iterative soft thresholding algorithm (LISTA)

Highlights

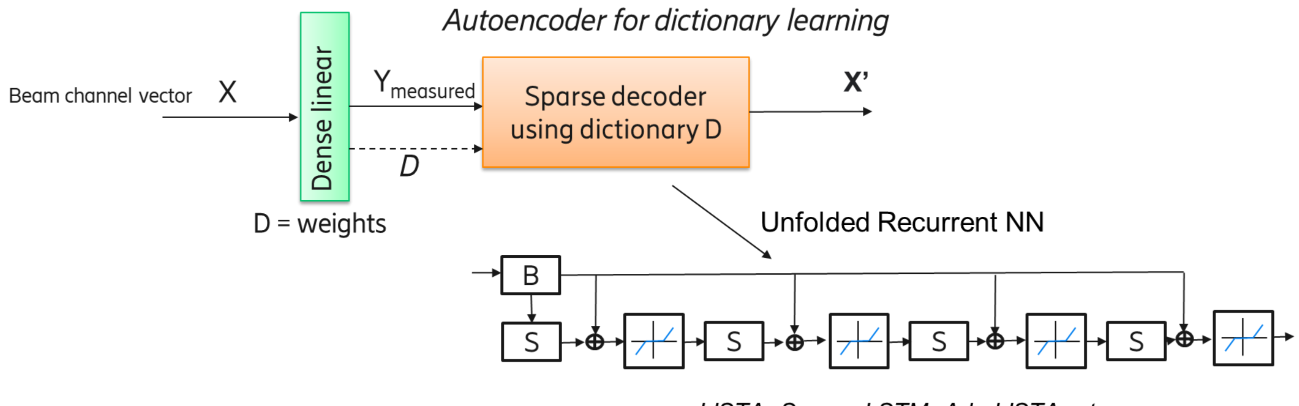
- Significant gain from both components:
  - Optimized dictionary
  - Trained neural sparse decoder
- Shows better generalization properties than MLP decoders

Evaluation towards 6G KPIs/KVIs

- Reduced beam scanning time (5-20% of baseline), lower connection drop, faster recovery

Quantifiable targets

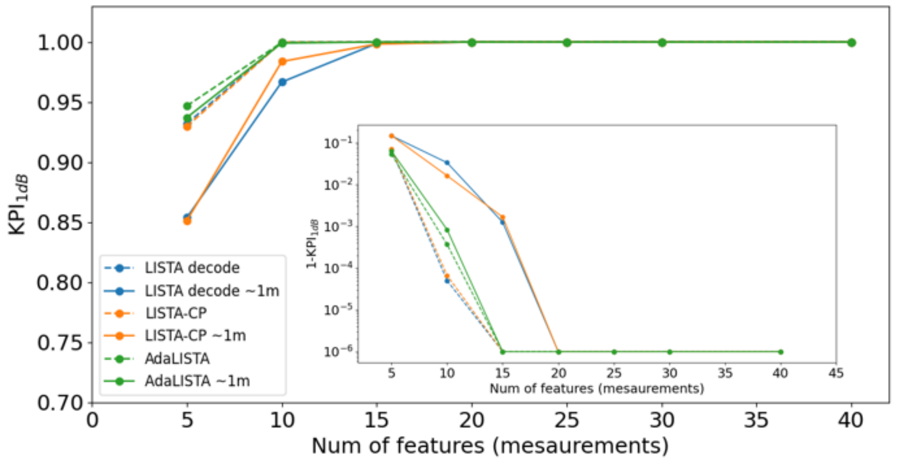
- Contributing to (T2) efficient resource utilisation by enabling shorter beam scanning bursts



LISTA, Sparse LSTM, Ada-LISTA, etc.

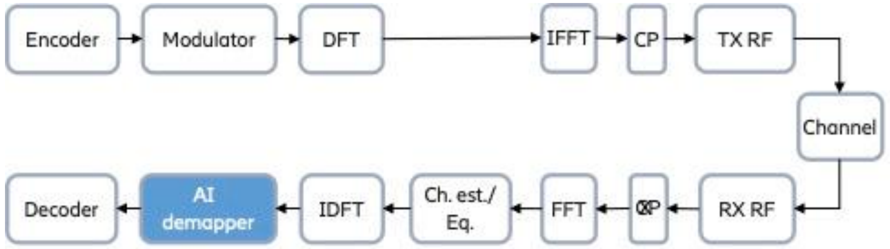
~25cm vs ~1m sampling distance: NLoS

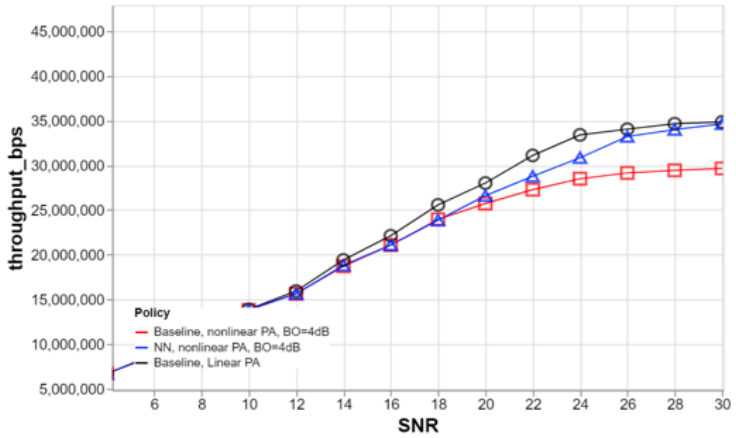
(dense, 15 layers, optimized D, with norm)

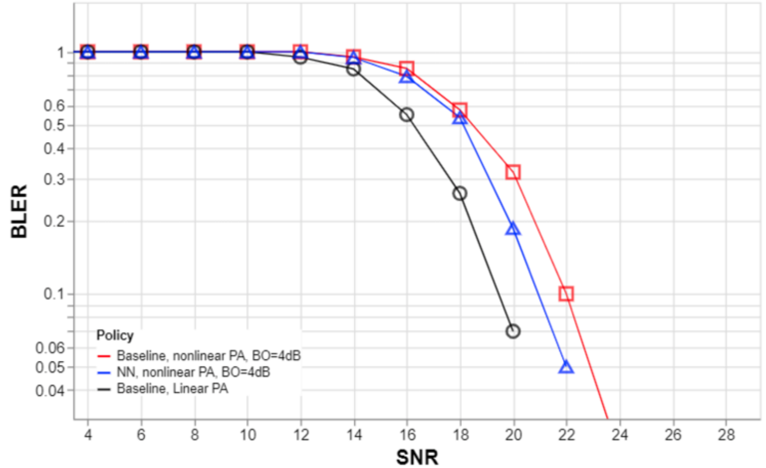


AI-empowered receiver for PA non-linearity compensation

- Problem/ challenge to be addressed
  - Power Amplifier (PA) non-linearity degrades throughput
  - Classical methods compensate PA non-linearity at transmitter side
    - PA power back-off → low energy efficiency
    - Digital-pre-distortion (DPD) → high complexity at transmitter side
- Final proposed solution
  - Neural network (NN)-based demapper to compensate non-linearities at receiver
  - Integrated with legacy methods for equalization and channel estimation
  - Inputs: equalized symbols (split in real and imaginary part) and SNR estimate
  - Outputs: soft bits used as input to the LDPC decoder
- Evaluation towards 6G KPIs/ KVIs
  - KPIs: Coverage, Spectral efficiency, Energy efficiency, Throughput, BER, BLER
  - KVIs: sustainability
- Quantifiable “Connecting Intelligence” targets
  - Spectral efficiency - Up to 20% increase
  - Energy efficiency - Up to 70% increase in Power Added Efficiency (PAE)
  - Throughput - Up to 20% increase
  - BLER: 1 dB gain @ 10% BLER for 64QAM

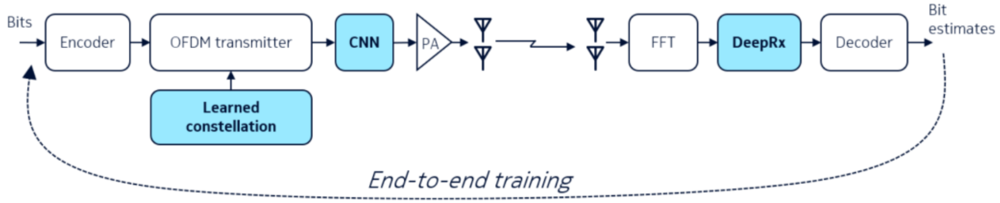


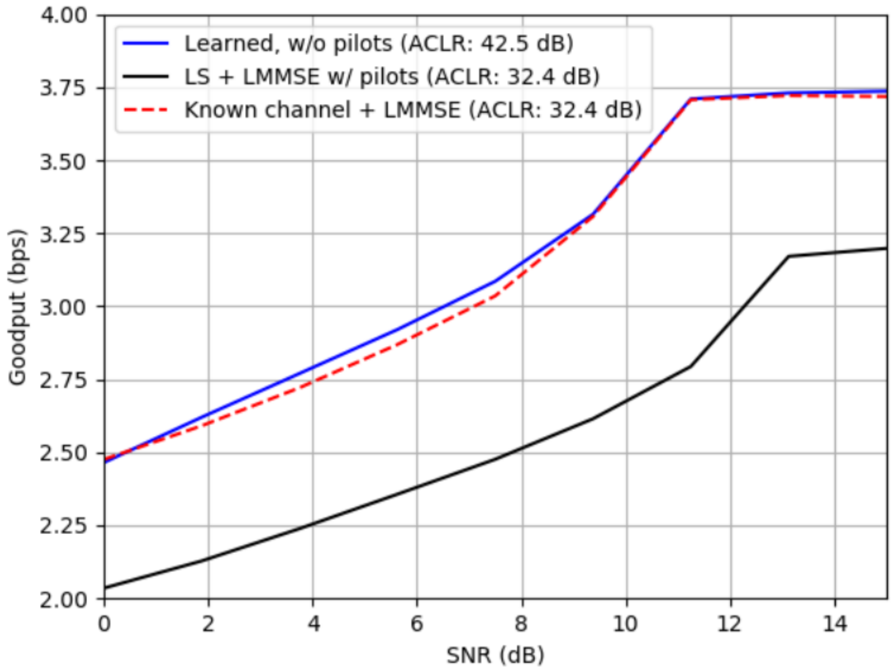




AI-Based Enhancements for Sub-THz

- **Problem/ challenge to be addressed**
  - How to facilitate pilotless transmissions and learn a waveform that is more resilient against hardware impairments?
  - Evaluations carried out at sub-THz
- **Final proposed solution**
  - Learn a waveform and a receiver jointly
  - The learned waveform is more resistant against nonlinear distortion and can be detected without any pilots by the jointly learned convolutional receiver (DeepRx)
- **Evaluation towards 6G KPIs/ KVIs**
  - Higher bit rate, increased spectral efficiency
- **Quantifiable “Connecting Intelligence” targets**
  - T1 by providing a BLER improvement, and T2 by improving the throughput
  - The BLER gain is approximately 2 dB, while the throughput improvement is in the order of 20-30%





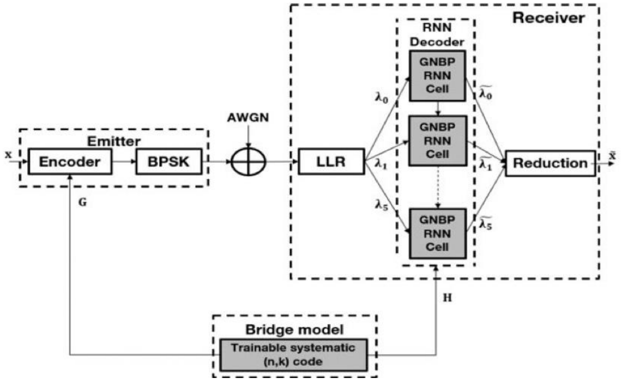
Neural network and machine learning aided channel (de)coding for constrained device

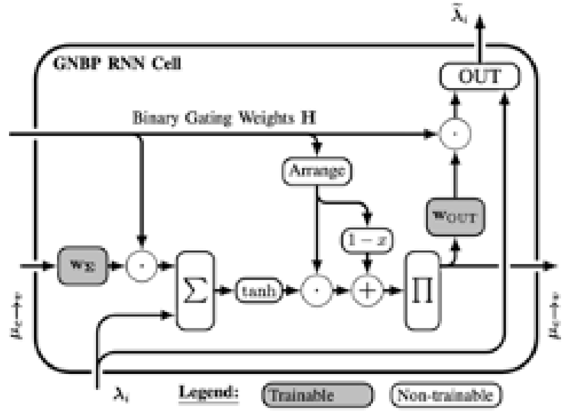
- **Problem/ challenge to be addressed**
  - Improve the efficiency of Forward Error Correction (FEC) mechanisms for short packets in IoT use-cases

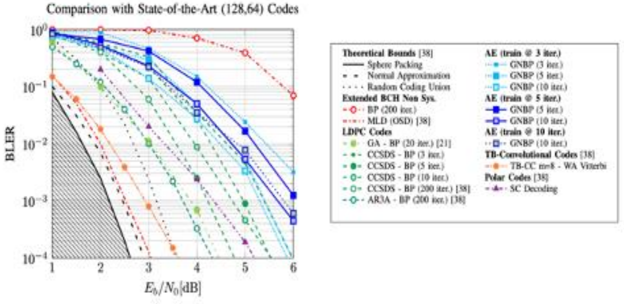
- **Final proposed solution**
  - Design and optimisation of linear block codes and decoders modelled jointly in an auto-encoder model
  - Decoder inspired by Belief Propagation structures currently in use for the decoding of 5G LDPC codes

- **Evaluation towards 6G KPIs/ KVIs**
  - Bit Error Rate (BER)/ BLock-Error Rate (BLER) gain.
  - Complexity gain

- **Quantifiable “Connecting Intelligence” targets**
  - *T1, T2,T4*

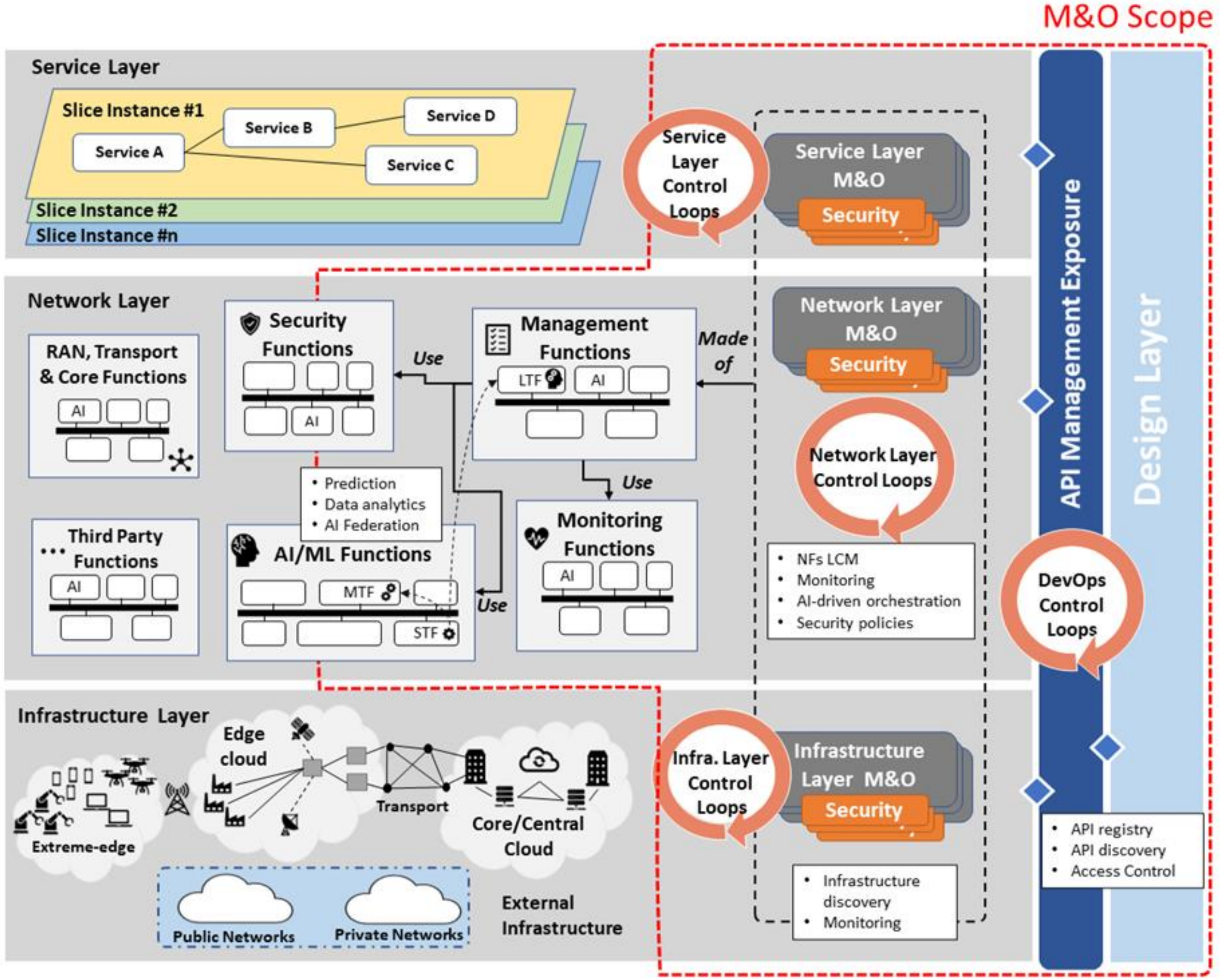






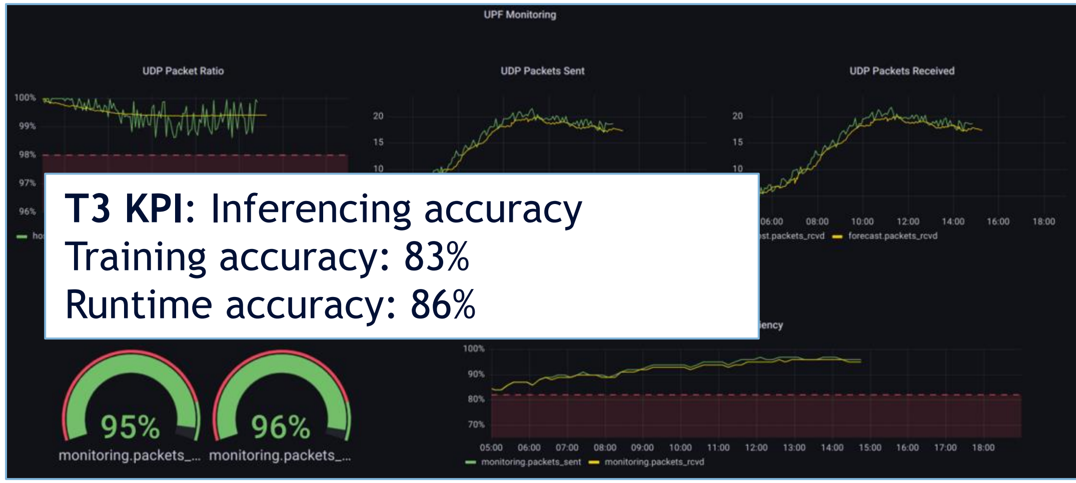
AI/ML-based predictive orchestration (ATO)

- **Problem/ challenge to be addressed**
  - Integration of AI/ML techniques into management and orchestration operations in 6G mobile networks
  - Legacy management and orchestration approaches lack the capabilities to face with 6G mobile networks requirements (i.e., integration of the extreme-edge domain, manage the compute-continuum as a whole, flexible and dynamic operations across domains, etc.)
- **Final proposed solution**
  - **Classify three types of M&O forecasting algorithms:**
    - Long-term forecasting
    - Mid-term forecasting
    - Short-term forecasting
  - Integrate Predictive orchestration as a particular approach of the more generic AI-based orchestration concept
  - Design a mapping between Hexa-X WP6 M&O architecture and the proposed algorithms and functions.
  - **Outputs:** Design and conceptual mapping to WP6 M&O architecture Building blocks.
- **Evaluation towards 6G KPIs/ KVIs**
  - **KPIs:** Energy efficiency, latency, programmability, elasticity, scalability, automation, resiliency.
  - **KVIs:** sustainability and trustworthiness
- **Quantifiable “Connecting Intelligence” targets**
  - T2
  - T3
  - T5

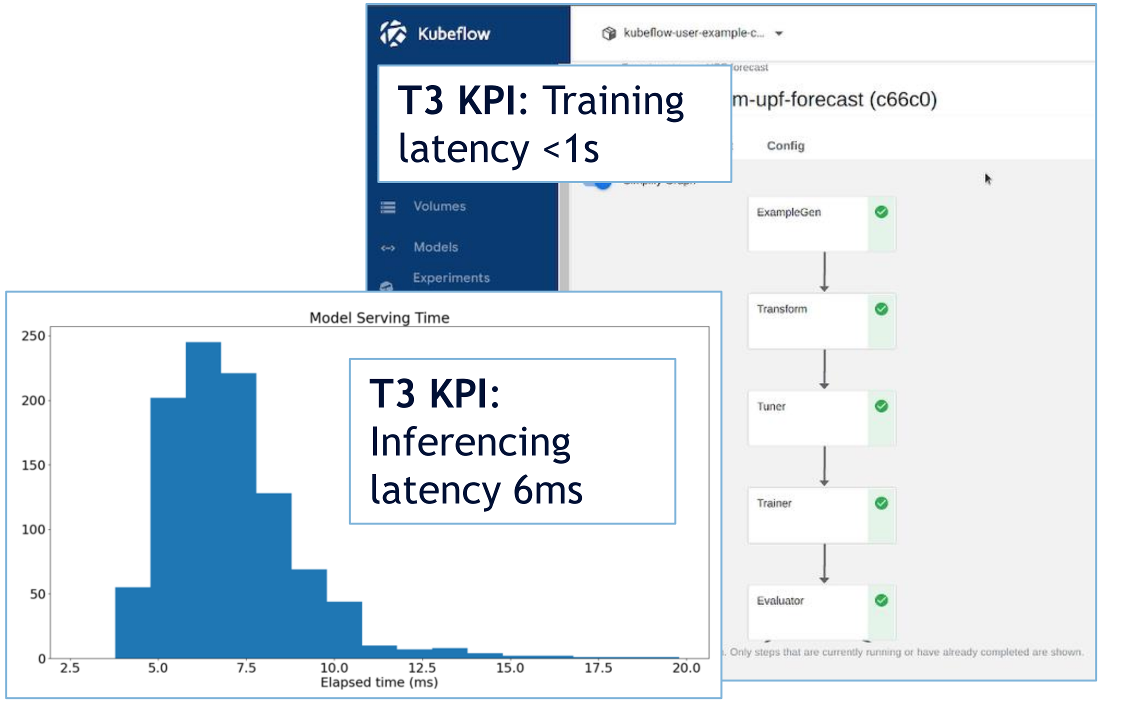


Distributed AI for automated UPF scaling in low-latency network slices (NXW)

- **Problem/challenge to be addressed**
  - Current orchestration solutions based on centralized AI may not be optimized for B5G/6G low-latency use cases with distributed UPFs
  - The traffic required for monitoring and training data distribution for centralized AI can lead to network congestion
- **Final proposed solution**
  - Architectural enhancement for orchestrating distributed AI functions, some closer to UPFs at the edge and others at the core
  - Use of application, network and infrastructure data collected at the edge for local decisions, e.g., inferencing and proactive edge resource management
- **Evaluation towards 6G KPIs/KVIs**
  - **Results obtained in the lab validation with a synthetic dataset covering 6 weeks of UPF simulated traffic for urban mobility patterns**
    - ✓ Inferencing latency (T3): 6.8ms (Target: 30s)
    - ✓ Inferencing accuracy (T3) (Avg over 5hr):
      - Training accuracy: 83% (Target: 89%)
      - Runtime accuracy: 86% (Target: 83%)
    - ✓ Training latency (T3): <1s (Target: < 1min)
- **Quantifiable “Connecting Intelligence” targets**
  - T3: The inferencing at the edge compute can help reducing the inferencing latency. Inferencing accuracy can be affected due to limited resource at the edge



T3 KPI: Inferencing accuracy
Training accuracy: 83%
Runtime accuracy: 86%

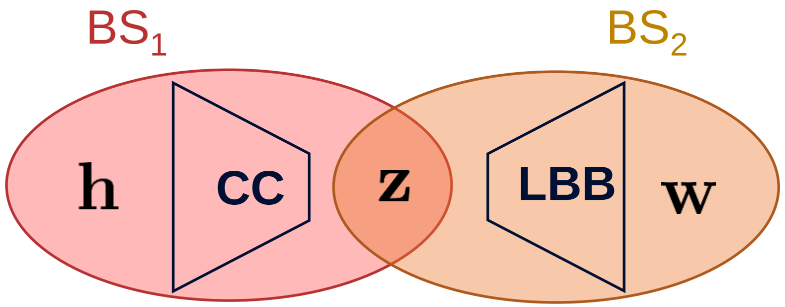


T3 KPI: Training latency <1s

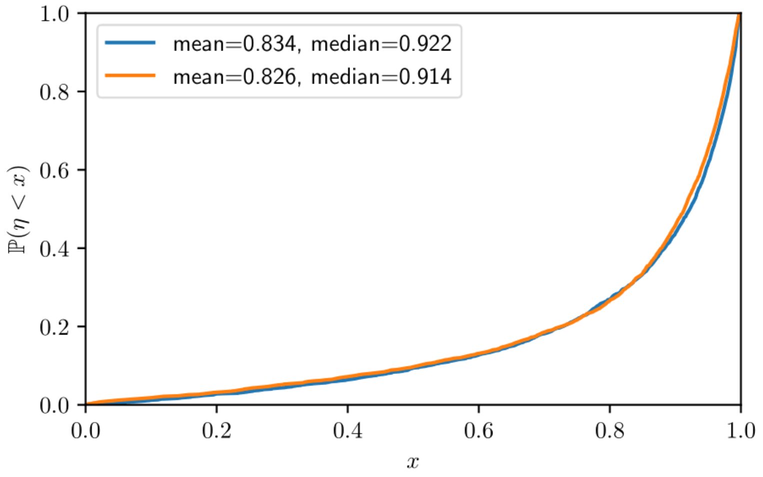
T3 KPI: Inferencing latency 6ms

Channel charting based beamforming

- **Problem/ challenge to be addressed**
  - Beam search can be very time consuming when using a lot of antennas at the base station. Location-based beamforming method help reduce the computational burden but rely on the precise knowledge of users' locations which are not always available.
- **Final proposed solution**
  - Take an already learned channel chart as input for a location-based beamforming NN. This allows a base station to choose appropriate precoders based on chart locations instead of spatial locations. This alleviate the need for a precise estimation of user locations and opens the way to several applications such as channel mapping in space and frequency.
- **Evaluation towards 6G KPIs/KVIs**
  - Precoder correlation to channel
- **Quantifiable "Connecting Intelligence" targets**
  - T2: reuse of channel chart for low-complexity precoding



Schematic view of the proposed method. First, channel measurments **h** are used to learn a chart **z**. A NN then takes the learned chart as input to produce corresponding precoders **w**. The two steps could be done at two different base stations.



CDF of the correlations. Blue curve corresponds to the original LBB at BS1. Orange curve corresponds to CC at BS1 and LBB at BS2 at different frequencies.

6G network as an efficient AI platform

6G network as an efficient AI platform

Enable and enhance the global operation of AI services, with **computing as a native part of future networks**

- Proposed solutions address the following problem:
  1. **Network services and data structures for AI applications**
     - AIaaS - seamless exploitation of network knowledge
     - Flexible compute workload assignment, CaaS
     - AI workload placement for energy, knowledge sharing and trust optimisation
  2. **Efficient inference for distributed AI**
     - Scalable and resilient deployment of distributed AI
     - Joint communication and computation orchestration for edge inference
     - Goal-oriented communication approach for edge inference
     - Network impairment resilience of autonomous agents
  3. **Efficient training for distributed AI**
     - Centralized training and decentralized execution (CTDE) approach to multi-cell multi-user MIMO
     - Federated ML model load balancing at the edge
     - Frugal Federated Learning
- Relevant KPIs/KVIs
  - AI agent availability, reliability, latency
  - Network and UE energy reduction, i.e., energy efficiency
  - Inferencing accuracy
  - Resource efficiency and complexity

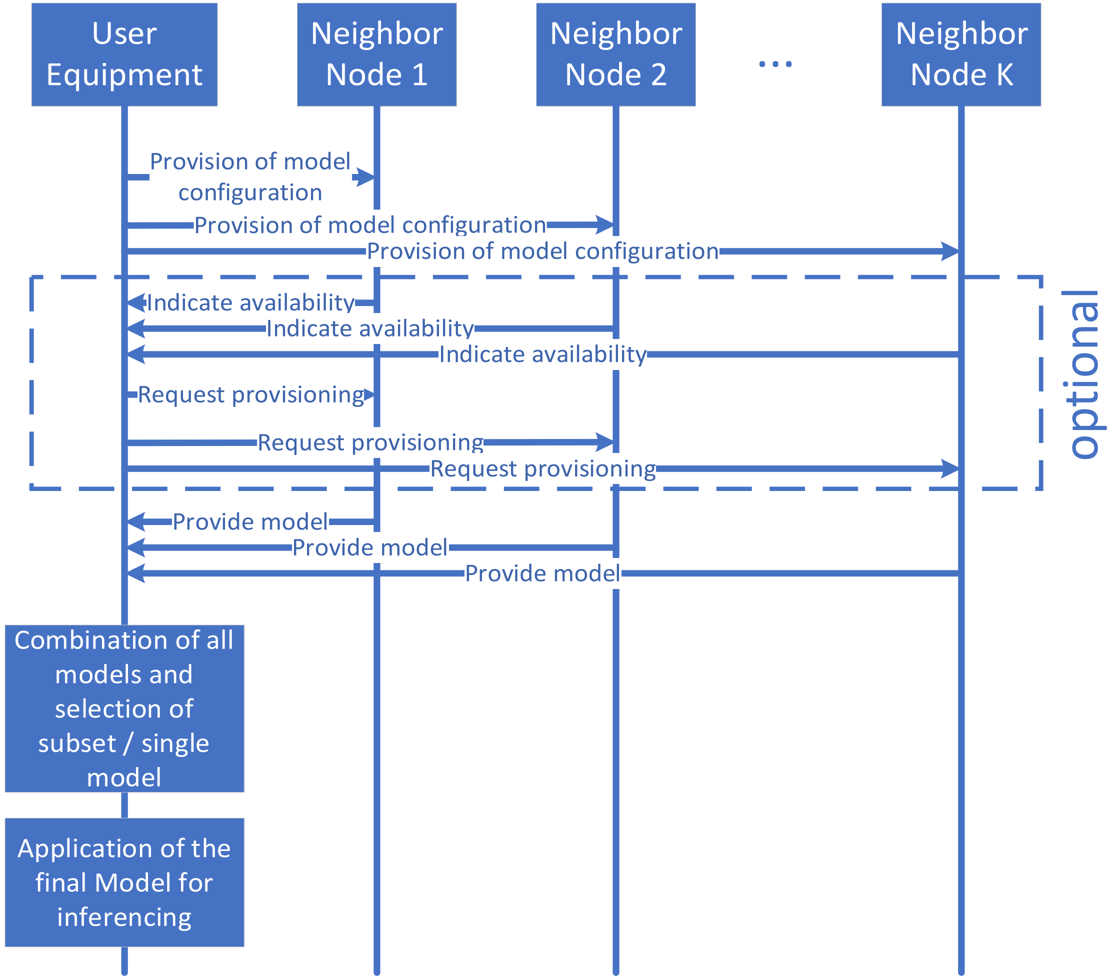
AIaaS - seamless exploitation of network knowledge

- **Problem/ challenge to be addressed**
  - How to enable a UE carrying an ML model *keep it up-to-date in mobility/ connection interruption regimes*.

- **Final proposed set of solutions**
  - Relevant data structures for AI Service (AIS)-assisted inferencing
  - Data structures relevant to AIS discovery to be used for service interoperabilit

- **Targeted 6G KPIs/ KVIs**
  - (On device) AI agent availability, AI agent reliability

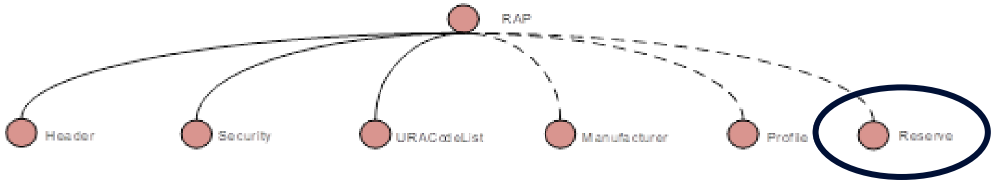
- **Quantifiable "Connecting Intelligence" targets**
  - T3: aim is to enable "seamless learning" and learning scalability by introducing some *filtering criteria* during *AI agent discovery*
  - T5: the design goal is to *route* an inferencing task to the most relevant and available AI agent with a maximum tolerable E2E *latency and energy consumption level*.



Signalling flow for requesting/delivering of new relevant ML models from a multitude of nodes per some filtering criteria

Flexible computing workload assignment (Compute-as-a-Service, CaaS)

- **Problem/ challenge to be addressed**
  - How to delegate/ distribute (generic) processing tasks across the network,
  - How to choose the appropriate representation of the code depending on the platform characteristics (heterogeneity) and container design for code (indicating code suitability for specific Hexa-X Use Cases and for "High Risk" AI applications as determined by the Draft European AI Act).
- **Final proposed set of solutions**
  - Extension of definiton of Radio Application Package (RAP) format to accomodate for indication of suitability for specific Hexa-X User Cases,
  - Extension of definiton of Radio Application Package (RAP) format to accomodate for indication of suitability for specific „High Risk" AI applications defined by the European AI Act.
- **Targeted 6G KPIs/ KVIs**
  - AI agent availability
  - Network energy efficiency
  - Flexibility
- **Quantifiable "Connecting Intelligence" targets**
  - T5, as aiming to facilitate flexible workload offloading



Bit 1 set to "1" means suitability for 1. Biometric identification and categorisation of natural persons:

(a) AI systems intended to be used for the 'real-time' and 'post' remote biometric identification of natural persons;

Bit 2 set to "1" means suitability for 2. Management and operation of critical infrastructure:

(a) AI systems intended to be used as safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity.

Bit 3 set to "1" means suitability for 3. Education and vocational training:

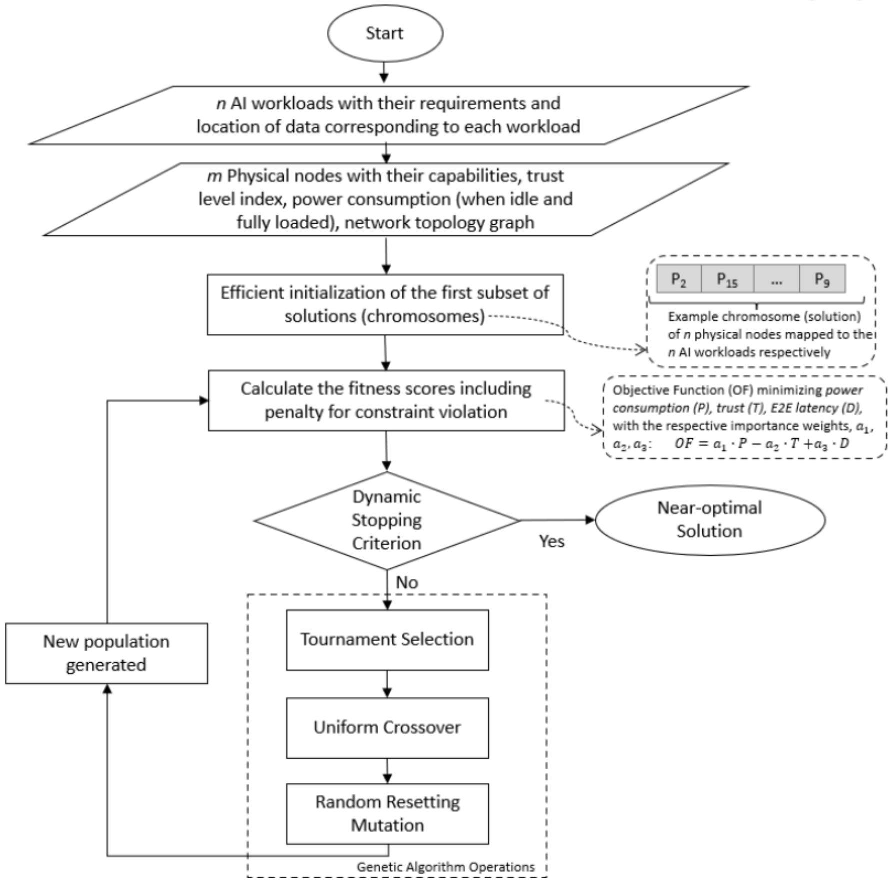
Bit 3 set to "1", sub-Bit 1 set to "1" means suitability for (a) AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions;

Bit 3 set to "1", sub-Bit 2 set to "1" means suitability for (b) AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions and for assessing participants in tests commonly required for admission to educational institutions.

Radio Application Package containing a specific bit which indicates that the manufacturers guarantees compliance for specific AI "High Risk" categories

AI workload placement for energy, knowledge sharing and trust optimisation

- **Problem/ challenge to be addressed**
  - Dealing with the trust, traffic, and energy consumption problems, that physical nodes who undertake the execution of AI algorithms/workloads, face
  - The main challenge is to optimise the placement of the AI algorithms/workloads to the various physical nodes with respect to the energy consumption of the overall network towards sustainability, the traffic, and the trust of these physical nodes
- **Final proposed solution**
  - Algorithm solving the optimisation problem following a meta-heuristic technique, for allocating the AI algorithms/workloads to physical nodes.
  - Input:
    - AI algorithms'/workloads' computational requirements, level of criticality,
    - physical nodes' computational and communication capabilities, energy consumption, trust level and data vicinity.
  - Output:
    - efficient (near optimal) AI placement
- **Evaluation towards 6G KPIs/ KVIs**
  - Up to 23% reduction of power consumption compared to baseline (random feasible placement)
  - Up to 98% reduction of E2E latency depending on number of AI workloads, compared to baseline (random feasible placement)
- **Quantifiable "Connecting Intelligence" targets**
  - T3: increase availability with efficient workload placement
  - T5: AI workload placement accounting network energy consumption



Start

n AI workloads with their requirements and location of data corresponding to each workload

m Physical nodes with their capabilities, trust level index, power consumption (when idle and fully loaded), network topology graph

Efficient initialization of the first subset of solutions (chromosomes)

P_2	P_{15}	...	P_9

Example chromosome (solution) of *n* physical nodes mapped to the *n* AI workloads respectively

Calculate the fitness scores including penalty for constraint violation

Objective Function (OF) minimizing *power consumption (P), trust (T), E2E latency (D),* with the respective importance weights, a_1, a_2, a_3: $OF = a_1 \cdot P - a_2 \cdot T + a_3 \cdot D$

Dynamic Stopping Criterion

Yes

Near-optimal Solution

No

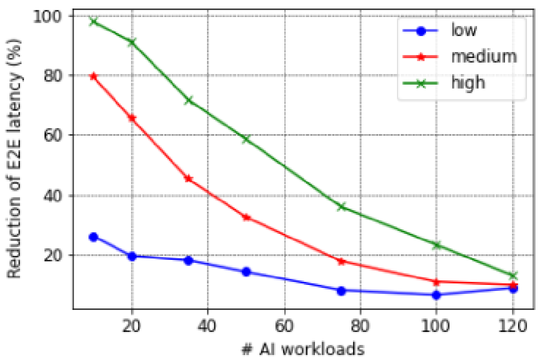
Tournament Selection

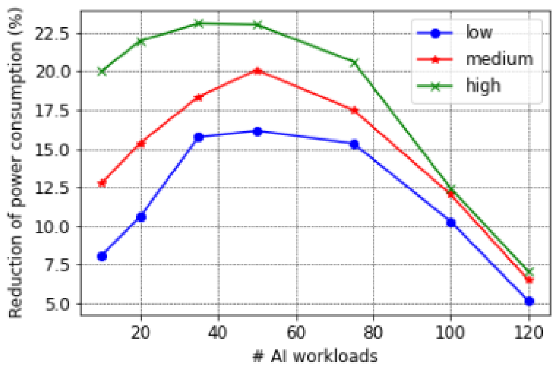
Uniform Crossover

Random Resetting Mutation

New population generated

Genetic Algorithm Operations





Resilient deployment of distributed AI - Technical Summary

The problem

- Sensor sharing in Hyperconnected resilient applications
- Sensor data and AI logic distributed over a multitude of wireless devices
  - Overlapping inputs – unnecessarily high load
  - Variable input quality – application-level feedback
  - Delay sensitive application – early inference results needed

Final proposed solution

- Proposed AI application and communication system architecture
- Incremental evaluation framework
- Evaluated on ANN-to-SNN converted neural networks
- Application – network interface
  - Application to provides continuous ms-level feedback on input quality
  - Network to prioritize device-edge communications accordingly

Highlights

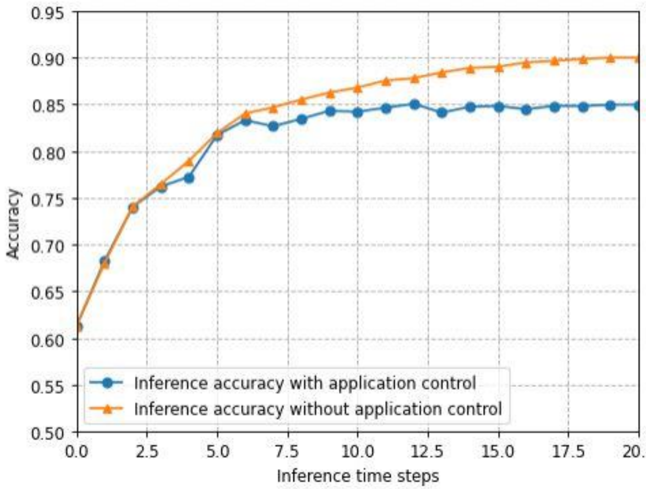
- Accuracy-latency trade-off can be controlled
- Significant load reduction due to application-network joint control

Targeted 6G KPIs/ KVIs

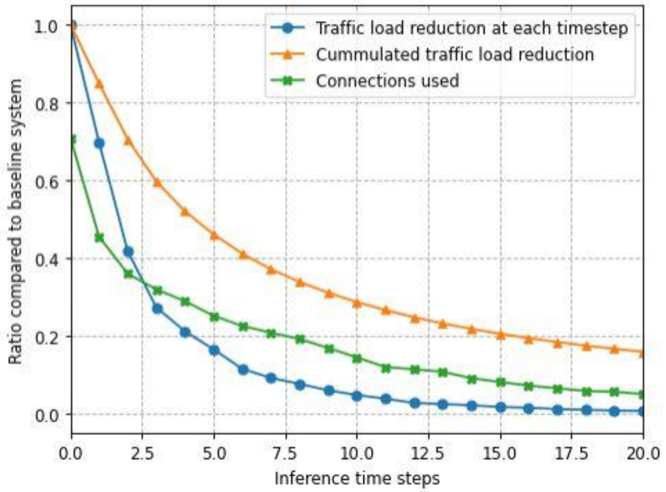
- Enables flexible trade-off between *inferencing latency* and *accuracy*
- Increased *device density* can be reached due to reduced aggregated edge load

Quantifiable “Connecting Intelligence” targets

- T3: the proposed joint communication-control solution for distributed AI inference enables reduced edge load and increased device density, while maintaining resilient operation



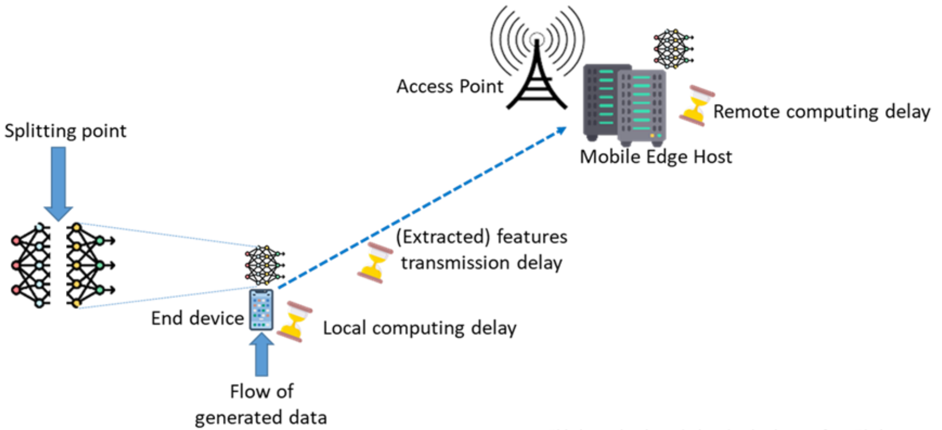
Accuracy-latency trade-off can be controlled with incremental inference



Load reduction due to the application-network joint control is significant, down to 20%

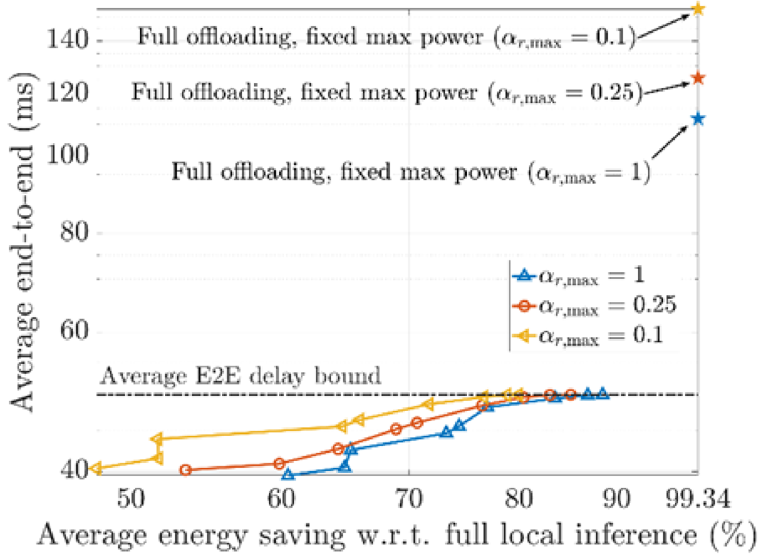
Joint communication and computation resource orchestration for edge inference

- **Problem/ challenge to be addressed**
  - Dynamically classify data collected by end users
  - Challenge: deal with time-varying connect-compute network conditions including
    - Availability of edge computing resources
    - Data arrivals
    - Wireless channels
- **Final proposed solution**
  - DNN splitting for partial computation offloading with
    - Adaptive selection of DNN splitting point
    - Device transmit power optimisation
- **Evaluation towards 6G KPIs/ KVIs**
  - Device energy consumption (communication and computing)
  - End-to-end delay (communication and computing)
- **Quantifiable "Connecting Intelligence" targets**
  - T5 on energy consumption reduction - simulations based evaluations show high gain w.r.t. benchmark solutions (i.e., full offloading and full local inference)



This image has been designed using images from Flaticon.com

Reference scenario

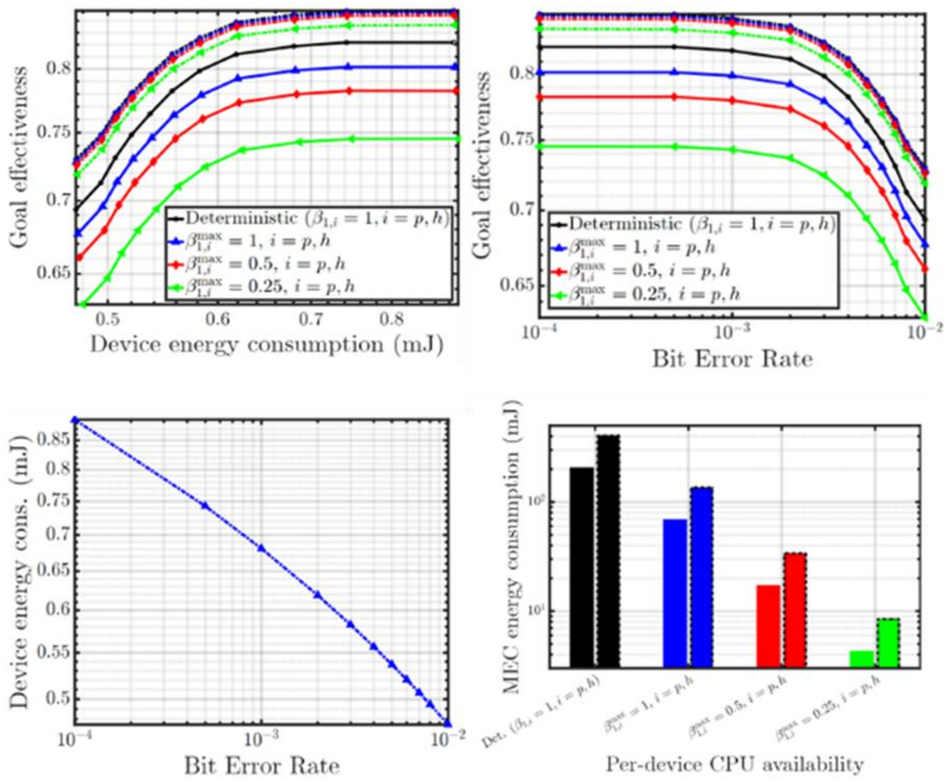


trade-off between average E2E delay and energy saving for different MEH's computing resource availabilities

Goal-oriented communication approach for edge inference

- **Problem/ challenge to be addressed**
  - Strike the best trade-off between energy, latency, and accuracy of an edge inference task (e.g., image classification)
- **Final proposed solution**
  - Base the success of communication on application performance (e.g., confident inference in time), rather than bit-quality metrics
- **Evaluation towards 6G KPIs/ KVIs**
  - Goal-effectiveness, defined as correct (or confident) inference on time
  - Device energy consumption
  - Edge server energy consumption

  Evaluated w.r.t. classical communication metric (e.g., BER)
- **Quantifiable “Connecting Intelligence” targets**
  - T5 on energy consumption reduction, also at the network side thanks to cooperative inference across multiple edge servers



Goal-effectiveness, device energy consumption, and edge server energy consumption, as a function of bit-level quality metrics (i.e., BER)

Network impairment resilience of autonomous agents

- **Problem/ challenge to be addressed**
  - Use of ML to predict mobility/ connection interruption regimes and provide resilience for the UE / AI agent.
- **Final proposed solution**
  - Deploy data analytics methods to predict quality issues ahead of time to prepare the UE/AI agent for connectionless operation.
  - Can be implemented both on agent and network side
- **Evaluation towards 6G KPIs/ KVIs**
  - AI and computation: safety, maintainability/recovery. Signalling of incidents ahead in time.
  - Use cases: Massive twinning, Robots to Cobots, AI partners, V2X
  - Real-time intelligent decisions based on distributed data - Agents interpret intents and surroundings, perform challenging and risky tasks
- **Quantifiable “Connecting Intelligence” targets**
  - T3 (AI agent availability and reliability),

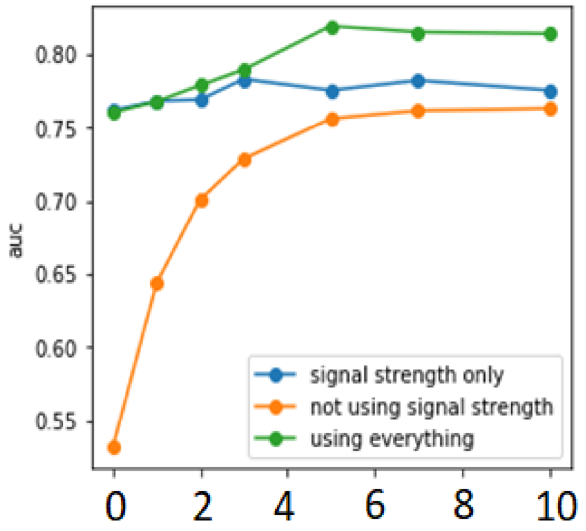
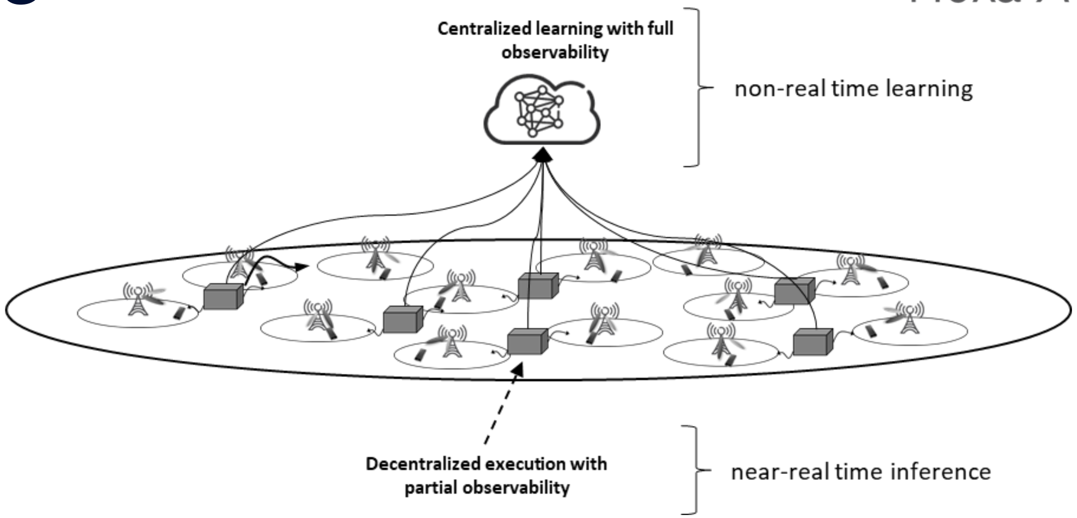


Figure: sample measurement using smartphone own data collection

Centralized Training and decentralized execution (CTDE) approach to multi-cell multi-user MIMO

- **Problem/challenge to be addressed**
  - Problem: multi-cell and multi-user precoding problem to minimize interference
  - Challenge
    - Learning to coordinate multiple agents (or base stations) with partial observability
    - Multi-dimensional continuous action space
- **Final proposed solution**
  - CTDE framework based on multi-agent deep deterministic policy gradient (MA-DDPG) algorithm
    - Learn multi-dimensional continuous action policy in a centralized manner with a shared critic (taking CSI, precoding vectors of every cell as input)
    - Decentralized inference (taking local CSI, no inter-cell data sharing as input)
  - Pre-processing step to handle "phase ambiguity" and reach faster convergence and better performance
- **Evaluation towards 6G KPIs/ KVIs**
  - Interference in multi-cell environment is mitigated to achieve maximum weighted sum rate (i.e., pareto-boundary of rate region) in two cell two UE scenario.
  - This result shows that the proposed solution can learn a pareto-optimal beamforming strategy.
- **Quantifiable “Connecting Intelligence” targets**
  - Increased agent density (decentralized inference) (T3)
  - Improved inferencing accuracy or improved latency (T3)

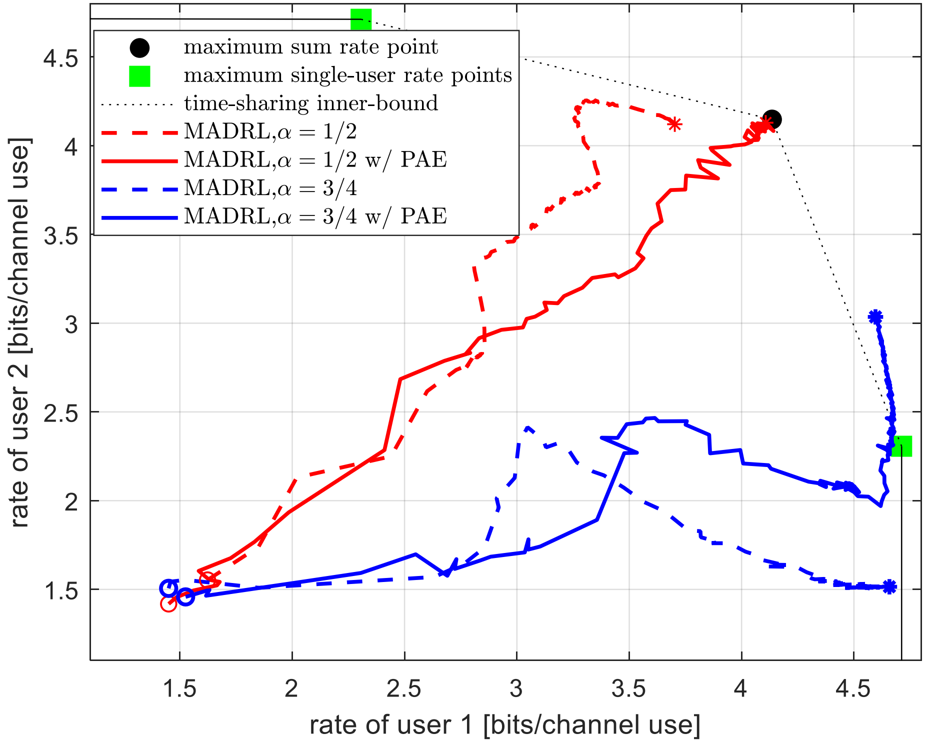


Centralized learning with full observability

non-real time learning

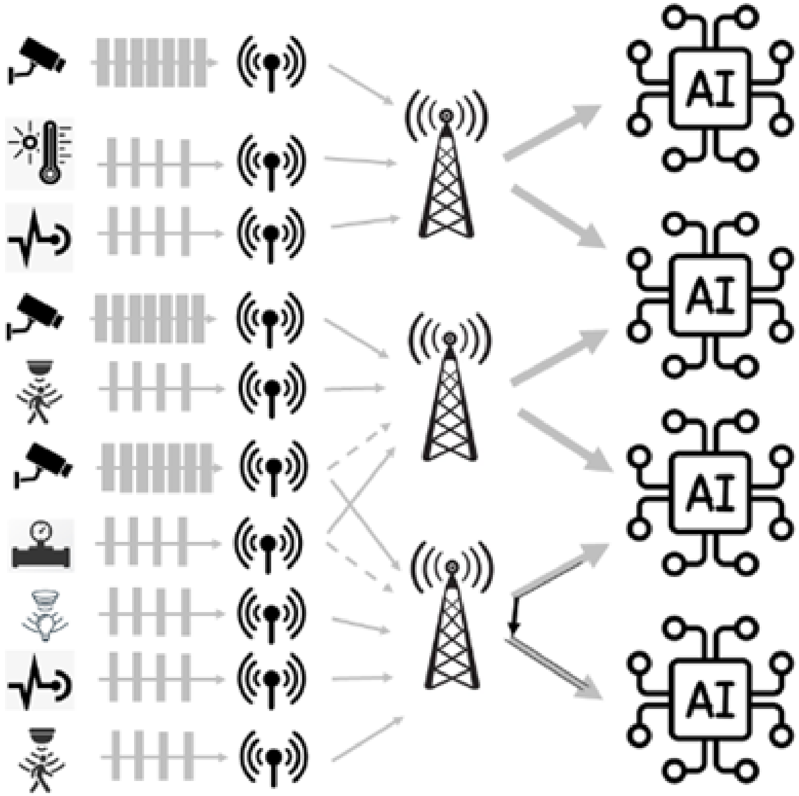
Decentralized execution with partial observability

near-real time inference



Federated ML model load balancing at the edge

- **Problem/ challenge to be addressed**
  - In mobile sensor streams with highly skewed and nonstationary data distributions, remedy imbalance w.r.t. amount and type of sensor
  - Negative impact 1: slow tasks that delay the completion of the whole stage – latency, energy
  - Negative impact 2: uneven knowledge of the environment resulting on suboptimal accuracy
- **Final proposed solution**
  - A dynamic reconnection solution to provide load balancing to remedy potential hot spots and data type diversity to ensure quality balance for the federated learners.
  - Experiments with Simulation of Urban Mobility (SUMO) generator [http://sumo.dlr.de/index.html]
- **Evaluation towards 6G KPIs/ KVIs**
  - AI agent availability, Latency, Energy reduction
  - Optimal assignment of resources
  - Increasing AI agent density.
  - Relevant use cases: Interacting and cooperative mobile robots, Smart city – sensors connected to AI services
- **Quantifiable “Connecting Intelligence” targets**
  - T2 (complexity gain (reducing the processing time)
  - T5 (inference & E2E latency)



Federated Learning application

- Middle node: hot spot with too much data
- Bottom node: insufficient data with one type (camera) missing
- Reconnection decision with state migration between the bottom AI agents.

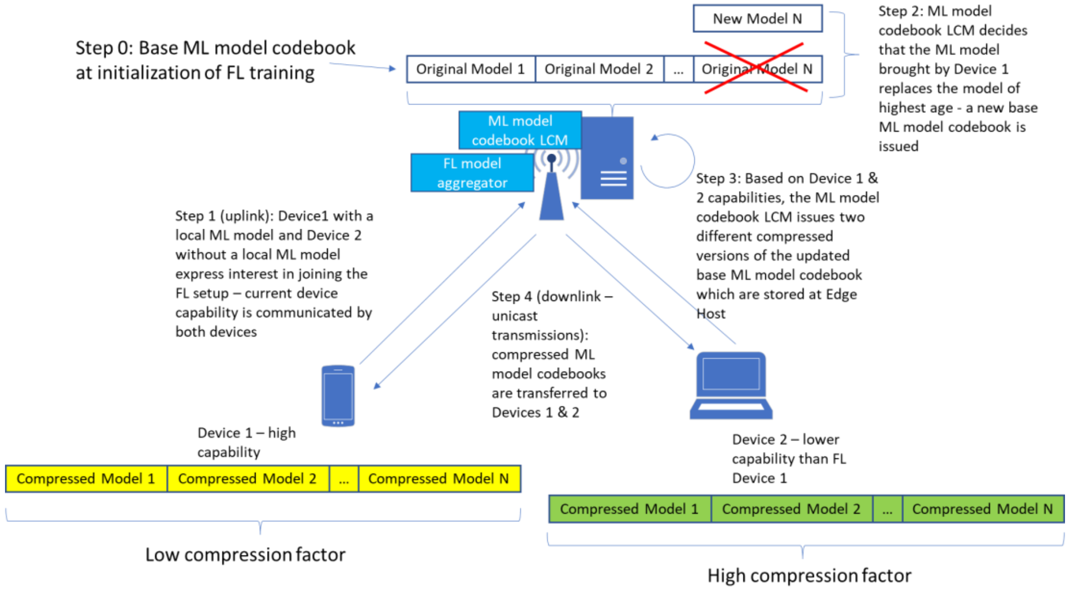
Frugal Federated Learning (FL)

- **Problem/ challenge to be addressed**
  - Dealing with over-the-air learning scenarios where available radio and storage resources are limited at device and edge RAN.

- **Final proposed set of solutions**
  - Device capability-adaptive and & overhead over-the-air FL
  - FL contributing device reporting to FL aggregator of its capabilities and learning status during FL training
  - ML model codebook - construction & maintenance operations
  - ML model codebook Lifecycle Manager (LCM) and model similarity score
  - FL training methods applying the proposed device capability-adaptive ML model codebook-based FL training

- **Targeted 6G KPIs/ KVIs**
  - Inferencing accuracy
  - Latency
  - End-to-end energy efficiency

- **Quantifiable “Connecting Intelligence” targets**
  - T3, T5, as aiming to enable large scale deployment of AI agents by economising needed radio, computing and storage resources.



Step 0: Base ML model codebook at initialization of FL training

Step 2: ML model codebook LCM decides that the ML model brought by Device 1 replaces the model of highest age - a new base ML model codebook is issued

Step 1 (uplink): Device1 with a local ML model and Device 2 without a local ML model express interest in joining the FL setup – current device capability is communicated by both devices

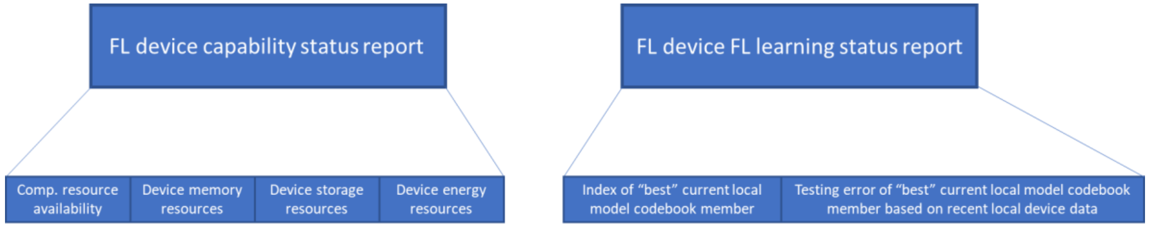
Step 3: Based on Device 1 & 2 capabilities, the ML model codebook LCM issues two different compressed versions of the updated base ML model codebook which are stored at Edge Host

Step 4 (downlink – unicast transmissions): compressed ML model codebooks are transferred to Devices 1 & 2

Device 1 – high capability

Device 2 – lower capability than FL Device 1

Example of producing different device-specific ML model codebooks, derived by a base ML model codebook and tailored to device capabilities -- FL initialisation/ establishment stage.


FL device capability status report
Comp. resource availability
Device memory resources
Device storage resources
Device energy resources
FL device FL learning status report
Index of “best” current local model codebook member
Testing error of “best” current local model codebook member based on recent local device data

Proposed status report by each FL device -- for upload to the FL aggregator in each FL training round.

AI/ML as an enabler for 6G network sustainability

AI/ML as an enabler for 6G network sustainability

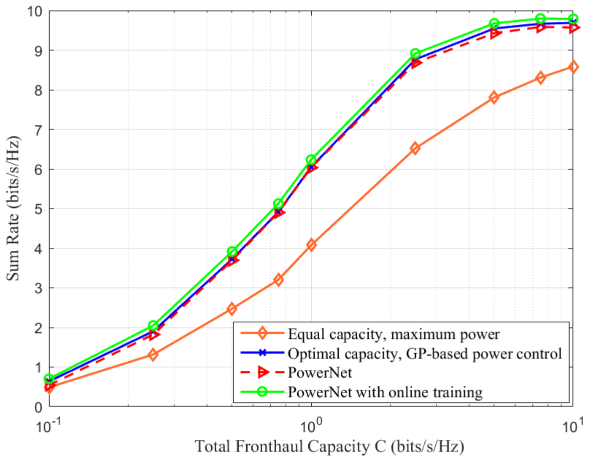
Improving 6G energy efficiency with low-complexity AI solutions

Proposed final solutions in 3 main categories

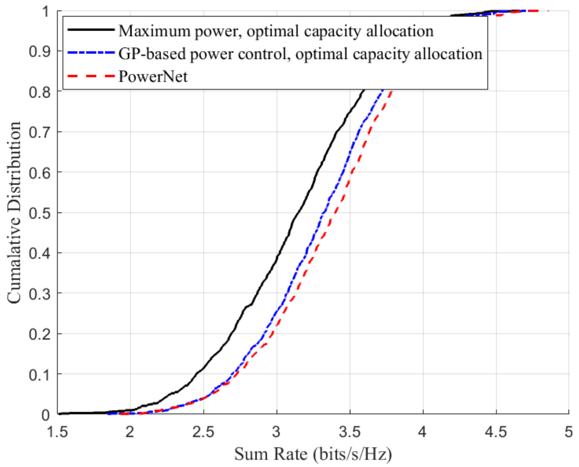
1. Universal functional approximation property of AI/ML
2. Moving complexity from inference into offline training
3. AI/ML to acquire more additional data from network for energy saving

Low complexity radio resource allocation in cell-free massive MIMO

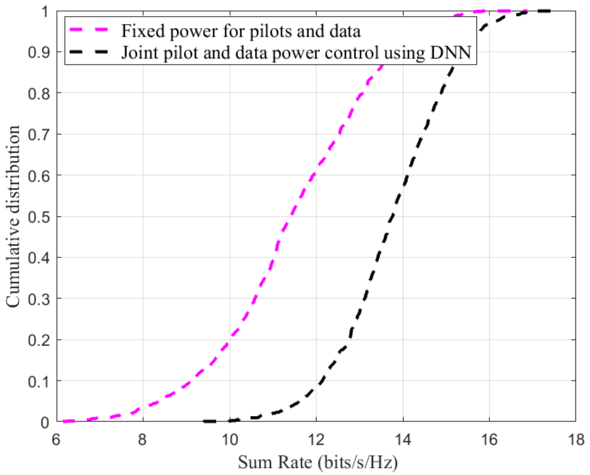
- **Problem/ challenge to be addressed**
  - Reducing the computational complexity in radio resource allocation tasks in cell-free massive MIMO networks.
  - Improve the flexibility/adaptability of the resource allocation algorithms to varying system configurations.
- **Final proposed solution**
  - An unsupervised learning-based DNN to learn the optimal resource allocations in a data-driven manner to achieve sum rate maximisation objective.
  - Problem 1: The DNN PowerNet proposed in Hexa-x D4.2 for joint power control and fronthaul capacity allocation to maximise the system sum rate shown to be able to adapt to different system parameters such as number of users, total fronthaul capacity etc.
  - Problem 2: Joint pilot power and data power control in cell-free massive MIMO uplink transmission considered to improve the system sum rate.
- **Evaluation towards 6G KPIs/ KVIs**
  - Complexity gain
  - Flexibility
- **Quantifiable "Connecting Intelligence" targets**
  - T2: rate gain via low complexity and efficient resource utilisation



Average sum rate performance with different power control and capacity allocation algorithms for 50 access points and 10 users in the cell-free massive MIMO network. The DNN results are obtained using the model trained with total capacity C = 1 bits/s/Hz.



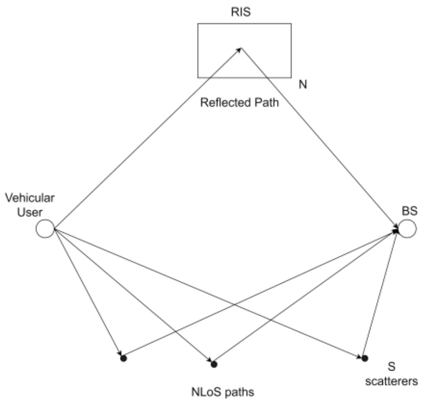
Cumulative distribution of the sum rate for 50 access points and 5 users in the cell-free massive MIMO network, obtained using the model trained with 10 users.



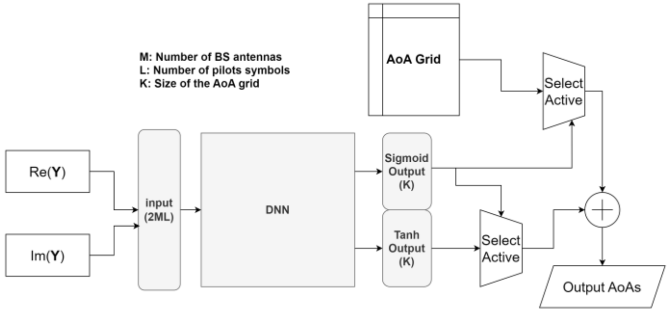
Sum rate performance comparison for fixed power transmission and pilot and data power control using the proposed unsupervised learning approach.

ML-based channel estimation for RIS-assisted systems with mobility

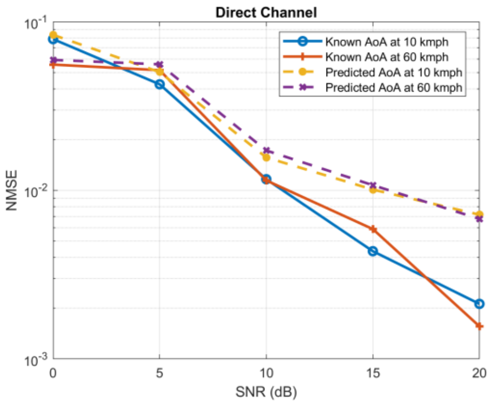
- **Problem/ challenge to be addressed**
  - Uplink channel estimation for an RIS-assisted mmWave vehicular network
  - Estimating sparse angular parameters and Doppler shifts using ML
- **Final proposed solution**
  - Use a sparse mmWave angular domain channel model (angle of arrivals, complex path gains, Doppler shifts)
  - Neural network model to predict AoAs (discrete point + error)
  - Optimization-based algorithm to estimate Doppler shifts, path gains
  - Use DeepMIMO for channel simulation
- **Evaluation towards 6G KPIs/ KVIs**
  - Channel Estimation Error
  - Spectral Efficiency
- **Quantifiable “Connecting Intelligence” targets**
  - T2: Improved channel estimation accuracy under mobility, while reducing the pilot overhead



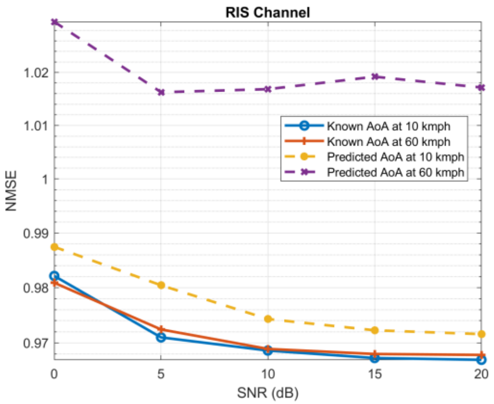
System model.



DNN architecture for AoA prediction.



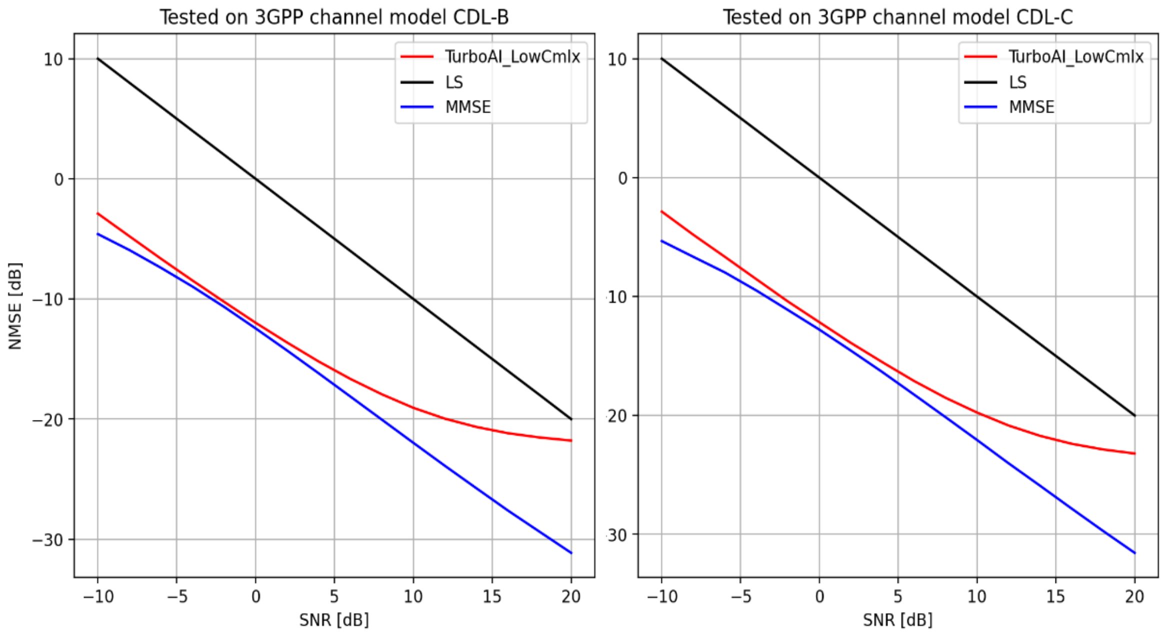
Variation of NMSE with SNR (dB) for direct channel.



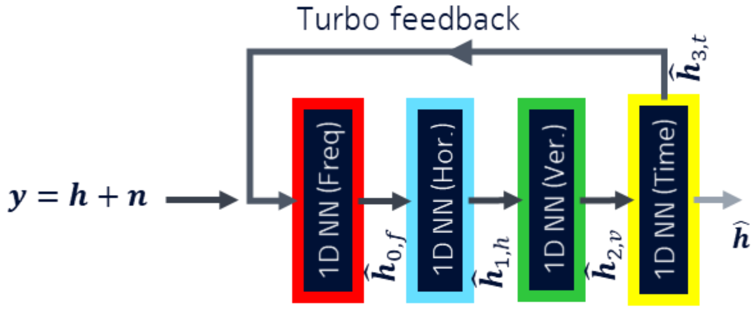
Variation of NMSE with SNR (dB) for RIS channel.

Generalizable low complexity channel estimation using neural networks

- **Problem/ challenge to be addressed**
  - Generalize the channel estimation performance beyond learning of the training data distribution, i.e. prepare the NN for channel model mismatch
  - Reduce the sample/computational complexity
- **Final proposed solution framework**
  - Use Turbo-AI structure, which is inspired by MMSE formulation iteratively.
  - To save computation complexity use Turbo AI architecture for different domains of the channel, i.e., Spatial, Frequency, time, separately.
  - Extensive training on different channel models can help to alleviate dependency on specific model parameters.
- **Targeted 6G KPIs/ KVIs**
  - Channel Estimation Error
  - Spectral Efficiency
- **Quantifiable "Connecting Intelligence" targets**
  - T1

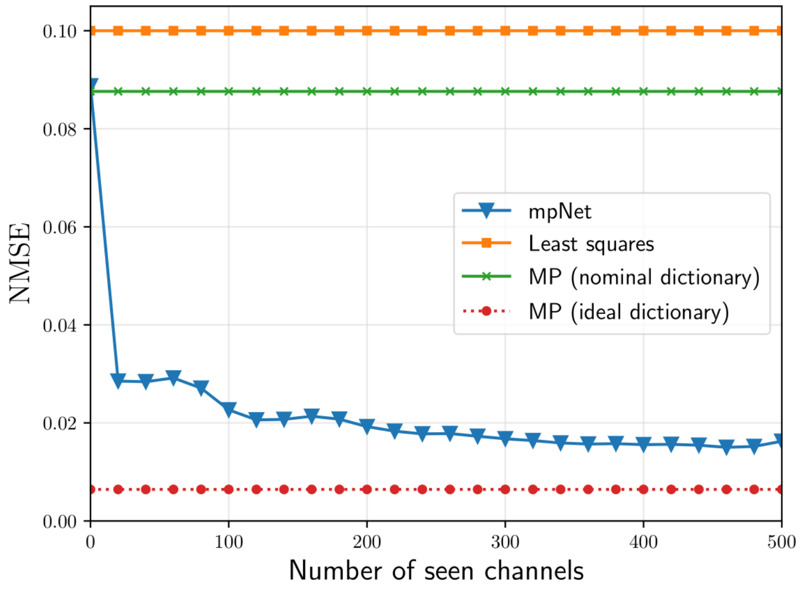


The NN is trained on CDL-A,-D & -E, while tested on dataset generated according to CDL-B and CDL-C, to show TurboAI's generalizability beyond the definition. The performance degradation is due to the channel model mismatch.



Deep unfolding for efficient channel estimation

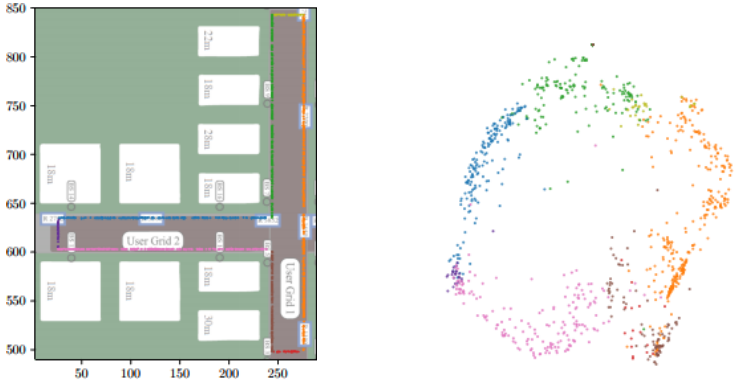
- **Problem/ challenge to be addressed**
  - Channel estimation algorithms based on physical models are theoretically very accurate but also very sensitive to hardware impairments and modifications.
- **Final proposed solution**
  - View the channel estimation algorithm as a NN (deep unfolding technique) that can be optimized and thus adapt in real time to incoming data. The new idea is to use structured dictionary and a hierarchical search within it to drastically reduce both sample complexity and time complexity.
- **Evaluation towards 6G KPIs/KVIs**
  - Channel Estimation Error
- **Quantifiable "Connecting Intelligence" targets**
  - T2: improved channel estimation and reduced complexity



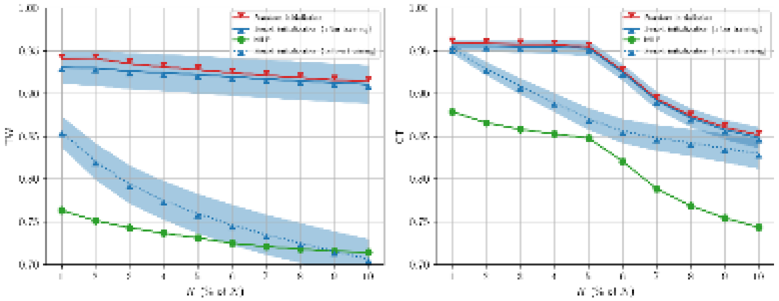
The model's channel estimation performance compared to baselines.

Hybrid model for channel charting

- **Problem/ challenge to be addressed**
  - Channel charting aims at localizing users relatively to one another in an unsupervised manner (without requiring access to GNSS , using only channels). It can be used for several applications, ranging from resource or pilot allocations to beam prediction. Most existing channel charting methods rely on the second order moment of channels and are thus computationally expensive.
- **Final proposed solution**
  - A hybrid model for the task of channel charting:
    - Structure of a model-based neural network with few parameters
    - Smart initialization based on an specifically designed channel distance measure and dimensionality reduction method (Isomap)
    - Training using a triplet loss exploiting temporal information obtained from the channel collection process
- **Evaluation towards 6G KPIs/KVIs**
  - Charts of better quality (TW, CT)
- **Quantifiable “Connecting Intelligence” targets**
  - T2: low complexity on-the-fly channel charting



The real path the user follows (left) and the final channel chart produced by the hybrid model (right).

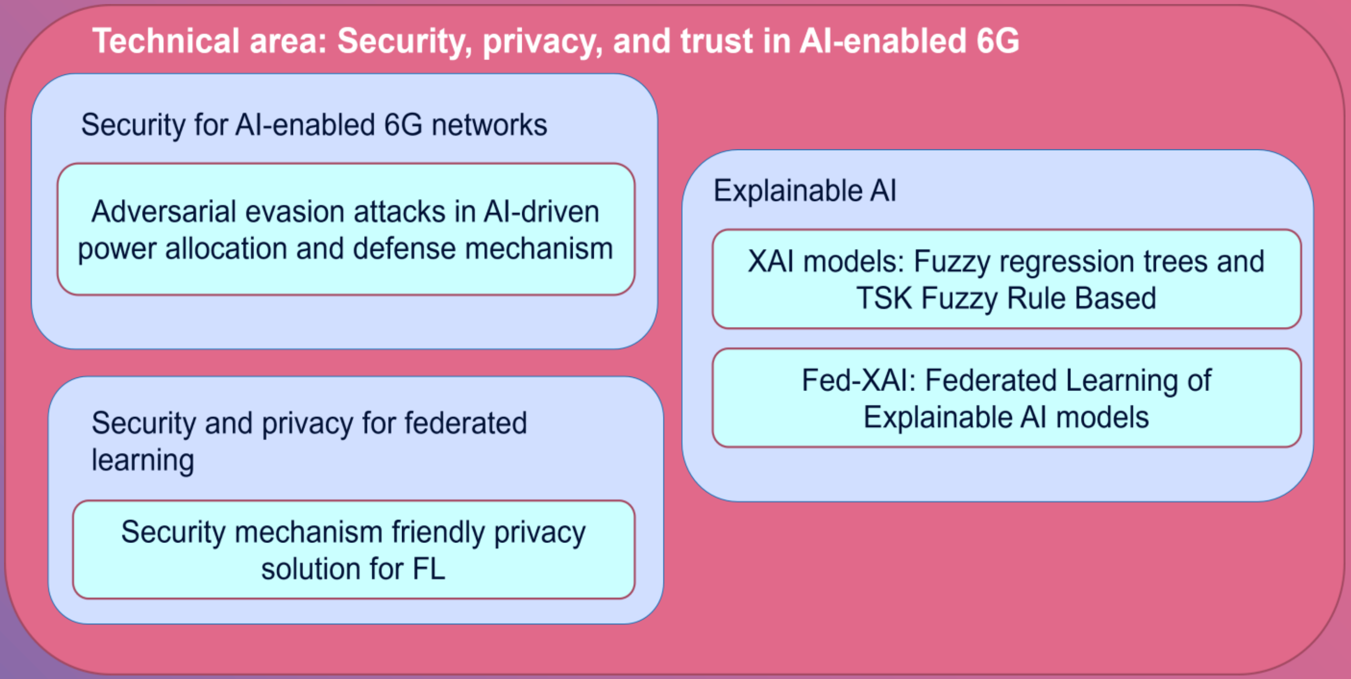


Comparison of the proposed approach to several variants on channels from the DeepMIMO dataset. Trustworthiness (TW) and continuity (CT) are given (the higher the better) as a function of the size of the considered neighbourhood K.

Privacy, security & trust in AI-enabled 6G

Privacy, security & trust in AI-enabled 6G

- The use of ML on massive amount of data is steadily increasing in time
- Cyber-attacks can be detected thanks to in-network AI/ML functionality
- Trustworthiness in AI/ML becomes critical for AI-pervasive 6G because AI/ML-based decisions are done for autonomy of communication and detection of cyber-attacks
- D4.3 focuses on possible adversarial attacks to AI/ML and mitigation technique to increase the robustness of AI model, also focus on how privacy of AI/ML can increase, and how to better interpret the AI/ML via explainable AI.



Technical area: Security, privacy, and trust in AI-enabled 6G

Security for AI-enabled 6G networks

Adversarial evasion attacks in AI-driven power allocation and defense mechanism

Security and privacy for federated learning

Security mechanism friendly privacy solution for FL

Explainable AI

XAI models: Fuzzy regression trees and TSK Fuzzy Rule Based

Fed-XAI: Federated Learning of Explainable AI models

Security of AI-driven power allocation for D-MIMO

- **Tech contribution**
  - Highly complex calculations for optimal power control problem in D-MIMO systems is needed.
  - Instead, usage of a well-trained AI model to approximate exact solution.
  - Motivation: Vulnerability of AI models against adversarial attacks. Malicious UE's or malicious RU's in the network manipulating the pilot signals or channel relate data being transmitted to DU (control unit) in an attempt to degrade the performance of AI-driven power allocation functionality.
  - Aim: Evaluating the success of adversarial attacks against the target AI model and define a defense mechanism.
- **Final solution approach**
  - Carefully crafted perturbations are applied to the input of the AI model which degrade the performance of the network in terms of both spectral and energy efficiency
  - Comparison of the risk associated with adversarial attacks with conventional attack threats.
  - A defense mechanism to mitigate the effects of such attacks.
- **Targeted 6G KPIs/ KVIs**
  - AI security & privacy
  - Adversarial attack success rate
  - Adversarial defense success rate
- **WP4 quantifiable targets**
  - **T6:** The security of the AI can be evaluated by estimating the success rate of the adversarial attacks and how much the defense mechanism decrease the success rate of adversarial attacks.
- **Evaluation results**
  - Effects of the adversarial attacks on per-user Spectrum Efficiency and average Energy Efficiency with different amount of perturbations and different levels of input information.
  - Simulation environment:
    - In MATLAB to generate our experimental training and test data sets.
    - 16 RUs distributed on a uniform grid in a 500m × 500m area.
    - 4 UEs and their locations are randomly chosen in the same region at each trial.
  - Adversarial attacks with optimized perturbations can significantly degrade the performance of the network in terms of both spectral and energy efficiency. In our experiments, we observed 60.51% success rate of adversarial attacks. In addition, the mitigation method decreases the success rate of attacker 37.29%.

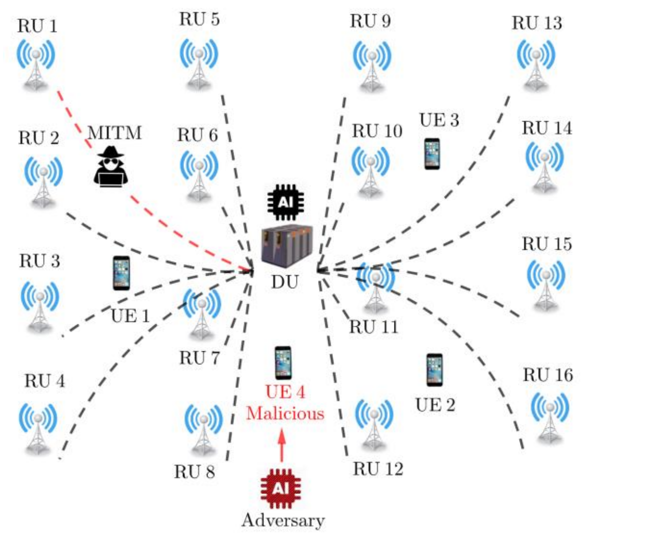


Fig1. D-MIMO network with potential attacks

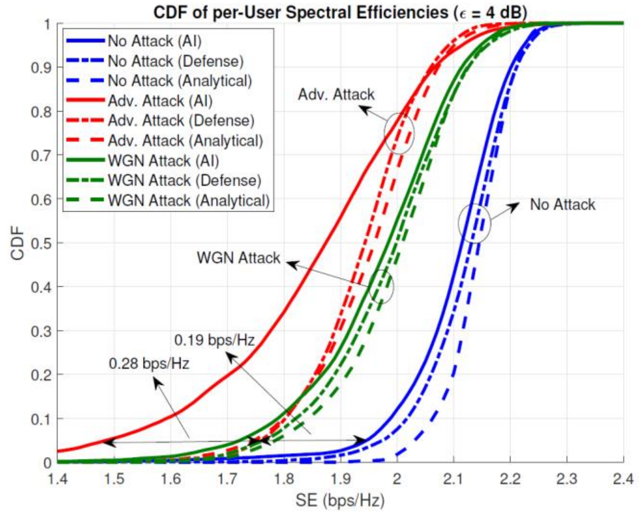


Fig2. CDF of per-user SEs for different cases

Security mechanism friendly privacy solutions for federated learning

- **Problem/challenge to be addressed**
  - FL is a privacy aware method however model updates that are sent to the FL server may leak some information about the clients.
  - Challenge: providing security and privacy at the same time
  - Aim: Enhance the FL privacy and prevent malicious behavior of the clients
- **Final solution approach**
  - Introduce a multi-hop communication along with blind signature to hide the identity of the clients and prevent malicious behavior of clients (i.e. model modification, involving multiple model update in one round of FL).
  - Extend the work: introduce and utilize a detector entity to the network to ensure that local model updates arrive to the server i.e. the clients can not drop the packets of other clients.
- **Targeted 6G KPIs/KVIs as defined by WP4**
  - AI Privacy:
  - Model accuracy
  - Total Overhead
- **WP4 quantifiable targets**
  - T6:It is hard to provide concrete numbers about the evaluation of privacy enhancement, but it has been proven with security arguments that the solution hides the identities of the local model updates.
  - We expect the accuracy of the model generated at server after applying defense mechanism to be (>90%).
  - In case of total overhead, the proposed method is expected to have (<30%) overhead after applying the defense mechanism.
- **Evaluation results**
  - Our proposed method brings 10.76 % overhead to the typical FL when communication speed is 504 Mb/sec
  - The accuracy of the model is remained unchanged.

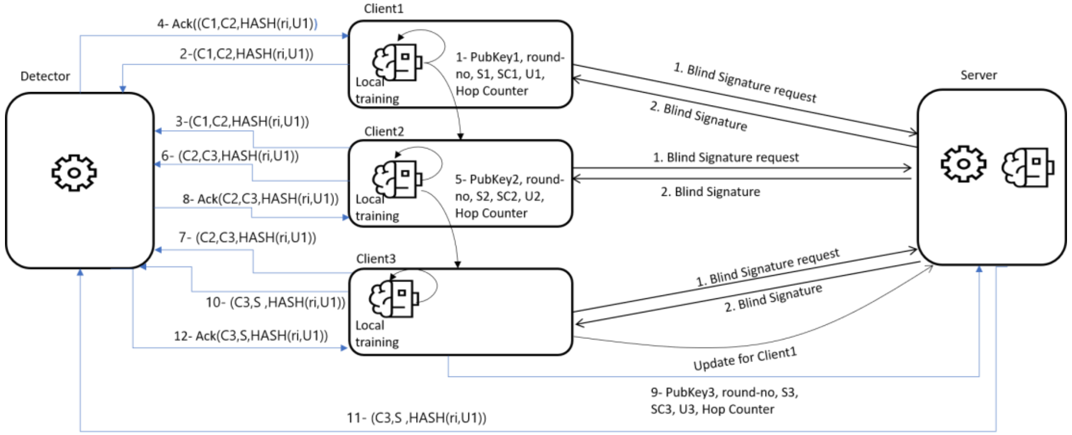

Client1
4- Ack((C1,C2,HASH(ri,U1))
2-(C1,C2,HASH(ri,U1))
Detector
Local training
1- PubKey1, round-no, S1, SC1, U1, Hop Counter
1. Blind Signature request
2. Blind Signature
Server
3-(C1,C2,HASH(ri,U1))
Client2
6- (C2,C3,HASH(ri,U1))
8- Ack(C2,C3,HASH(ri,U1))
Local training
5- PubKey2, round-no, S2, SC2, U2, Hop Counter
1. Blind Signature request
2. Blind Signature
7- (C2,C3,HASH(ri,U1))
Client3
10- (C3,S ,HASH(ri,U1))
12- Ack(C3,S,HASH(ri,U1))
Local training
1. Blind Signature request
2. Blind Signature
Update for Client1
9- PubKey3, round-no, S3, SC3, U3, Hop Counter
11- (C3,S ,HASH(ri,U1))

Fig 1. Proposed interactions of server, clients, and detector.

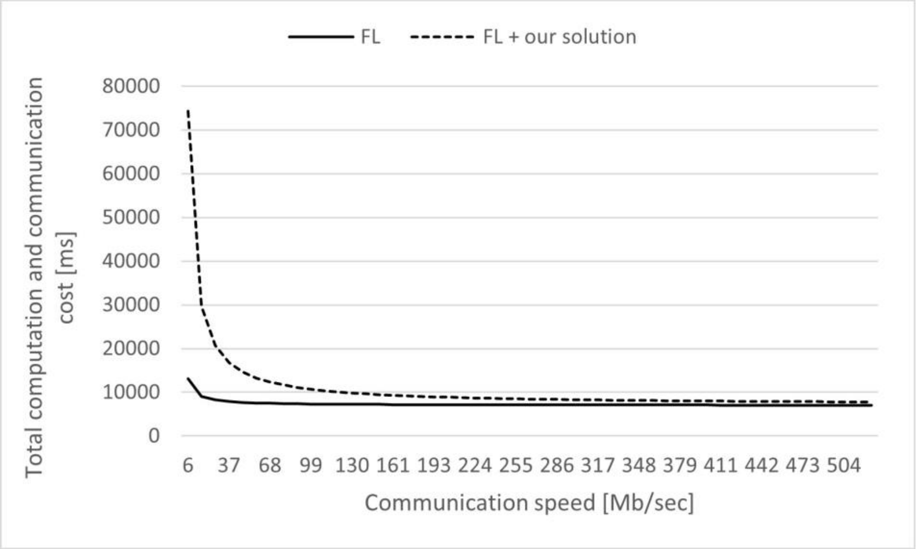
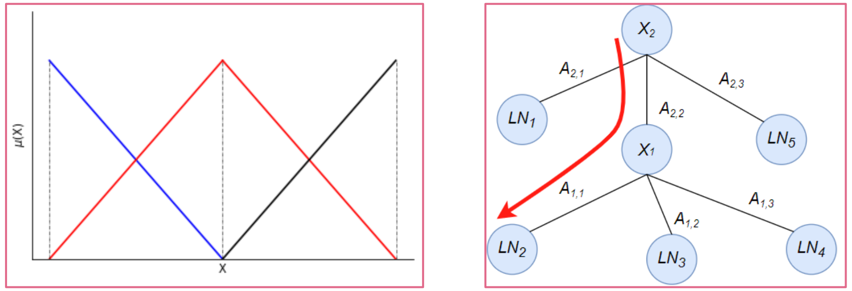


Fig 2. Overhead of our solution depending on the communication speed

XAI models: Fuzzy regression trees and TSK Fuzzy Rule Based Systems

- **Problem/ challenge to be addressed**
  - Explainability as a requirement towards Trustworthy AI
    - Rule based systems (RBSs) and Decision Trees (DTs) considered highly inherently interpretable models
    - Concepts from fuzzy set theory can further boost modelling capability and interpretability
- **Final proposed solution**
  - Four variants of Fuzzy RT (FRT) as combination of different local regression models and different inference strategies
  - Firtst-order Takagi-Sugeno-Kang Fuzzy Rule-based Systems (TSK-FRBSs) with enforced interpretability
- **Evaluation towards 6G KPIs/ KVIs**
  - Inference accuracy (measured as MSE on regression task): gain of XAI model (TSK) vs less interpretable state-of-art solution in the range [-4%,+37%]
  - Explainability: high level of interpretability ensured by the adoption of inherently interpretable models
- **Quantifiable "Connecting Intelligence" targets**
  - **T4:** accuracy of an XAI model within (<10%) of "black box" solutions



(Left) Strong triangular uniform fuzzy partition over a generic input attribute. Three fuzzy sets represent («low», «medium», «high») values, respectively. (Right) Toy multi-way FRT. A path from the root to a leaf represents a rule.

IF

longitude (x_1) is Low *AND* latitude (x_2) is Medium *AND* housingMedianAge (x_3) is Medium *AND* totalRooms (x_4) is Low *AND* totalBedrooms (x_5) Low *AND* population (x_6) is Low *AND* households (x_7) is Low *AND* medianIncome (x_8) is Medium

THEN

$$\mathbf{medianHouseValue} = 0.83 - 1.08x_1 - 0.95x_2 + 0.08x_3 + 0.41x_4 + 2.18x_5 - 5.29x_6 + 0.27x_7 + 1.28x_8$$

Example of a rule from TSK model generated on the *california* benchmark dataset (regression problem).

	TSK		TSK [FSN+20]	
	Train	Test	Train	Test
WI	1.28	1.38	1.48	1.52
TR	24.06	40.30	32.07	62.93
MO	3.99	6.36	4.49	8.22
CA	4.78	4.81	4.62	4.64

Results: our proposed XAI model (TSK with maximum matching) vs less interpretable state-of-art solution (FSN+20)

Fed-XAI - Federated Learning of Explainable AI models

- **Problem/ challenge to be addressed**
  - Enabling *collaborative training of explainable AI models* without violating privacy of users
    - Main challenge: traditional protocols (e.g., Federated Averaging) are not immediately amenable to FL of highly interpretable models such as Decision Trees and Rule-Based Systems
- **Final proposed solution**
  - Fed-XAI: Federated Learning of Rule-Based Systems
    - Definition of a novel approach for Federated TSK FRBS
- **Evaluation towards 6G KPIs/ KVIs**
  - Inference accuracy of XAI models learned in federated fashion higher than those locally learned: evaluation on benchmark datasets in terms of MSE with gain in the interval [1x,2x]
  - Explainability: high level of interpretability ensured by the adoption of inherently interpretable models
- **Quantifiable "Connecting Intelligence" targets**
  - **T4:** accuracy of an XAI model within (<10%) of "black box" solutions

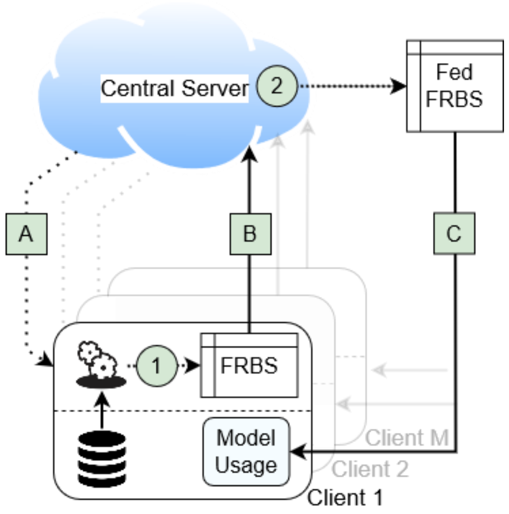
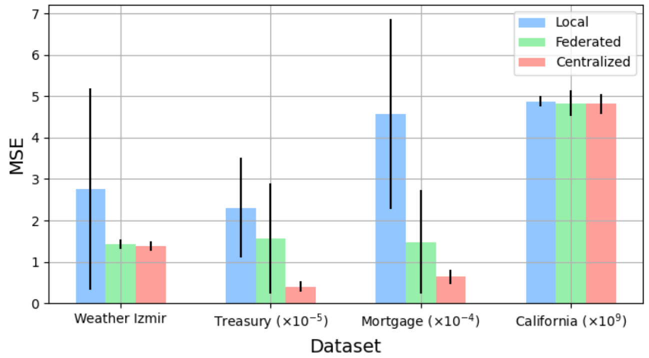


Illustration of federated learning of XAI models:

- [A], [B], [C] communication steps
- (1) local learning (2 aggregation

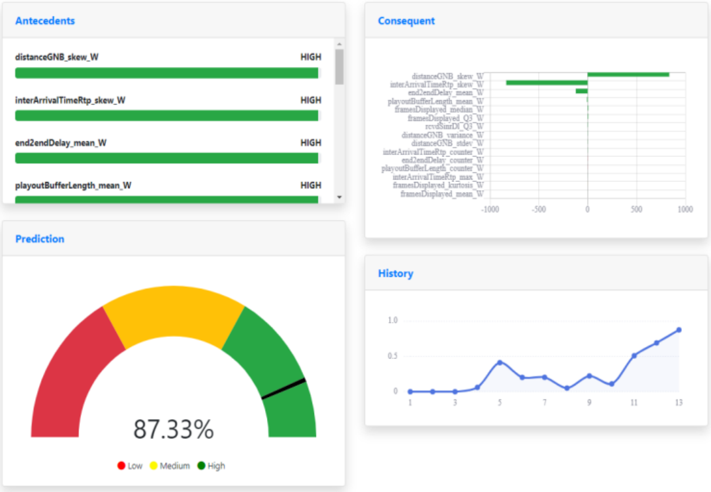


Experimental results: average MSE on four regression datasets. Comparison between local, federated and centralized learning schemes. Error bars represent standard deviation.

Demonstration activities - Federated eXplainable AI (FED-XAI) demo

Demo: Federated eXplainable AI (FED-XAI)

- **Problem/ challenge to be addressed**
  - Fed-XAI models target at forecasting QoE (regression problem)
    - Several instances of vehicular User Equipment (UE), connected to a B5G/6G network, receive (or send) a video stream, whose perceived quality is crucial for the availability of advanced driving assistance systems, such as see-through (or tele-operated driving)
- **Final proposed solution**
  - Development of a framework for Fed-XAI:
    - Intel OpenFL library for FL process, extended for Fed-XAI support
    - Container as de-facto standard for lightweight virtualization
    - Messages are exchanged via RESTful APIs over HTTPS for security
- **Evaluation towards 6G KPIs/ KVIs**
  - Inference accuracy:
    - Federated Learning (FL) compared with Local Learning (LL) setting
    - Fed-TSK compared with standard FL of NNs
  - Explainability: high level of interpretability ensured by the adoption of inherently interpretable models
- **Quantifiable “Connecting Intelligence” targets**
  - **T4:** accuracy of an XAI model within (<10%) of “black box” solutions



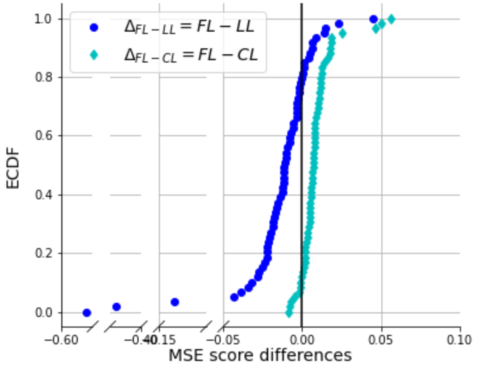
Antecedents

distanceGNB_skew_W	HIGH
interArrivalTimeRtp_skew_W	HIGH
end2endDelay_mean_W	HIGH
playoutBufferLength_mean_W	HIGH

Prediction

87.33%

Snapshot of Fed-XAI dashboard: predictions and related explanations



ECDF of the differences of test MSE scores between FL and LL (dark blue) and between FL and Centralized learning (light blue).

FL - TSK		FL - NN	
MSE	R^2	MSE	R^2
0.066	0.559	0.059	0.606

Average values of test MSE scores and R^2 scores obtained with the FL setting by the TSK and NN models

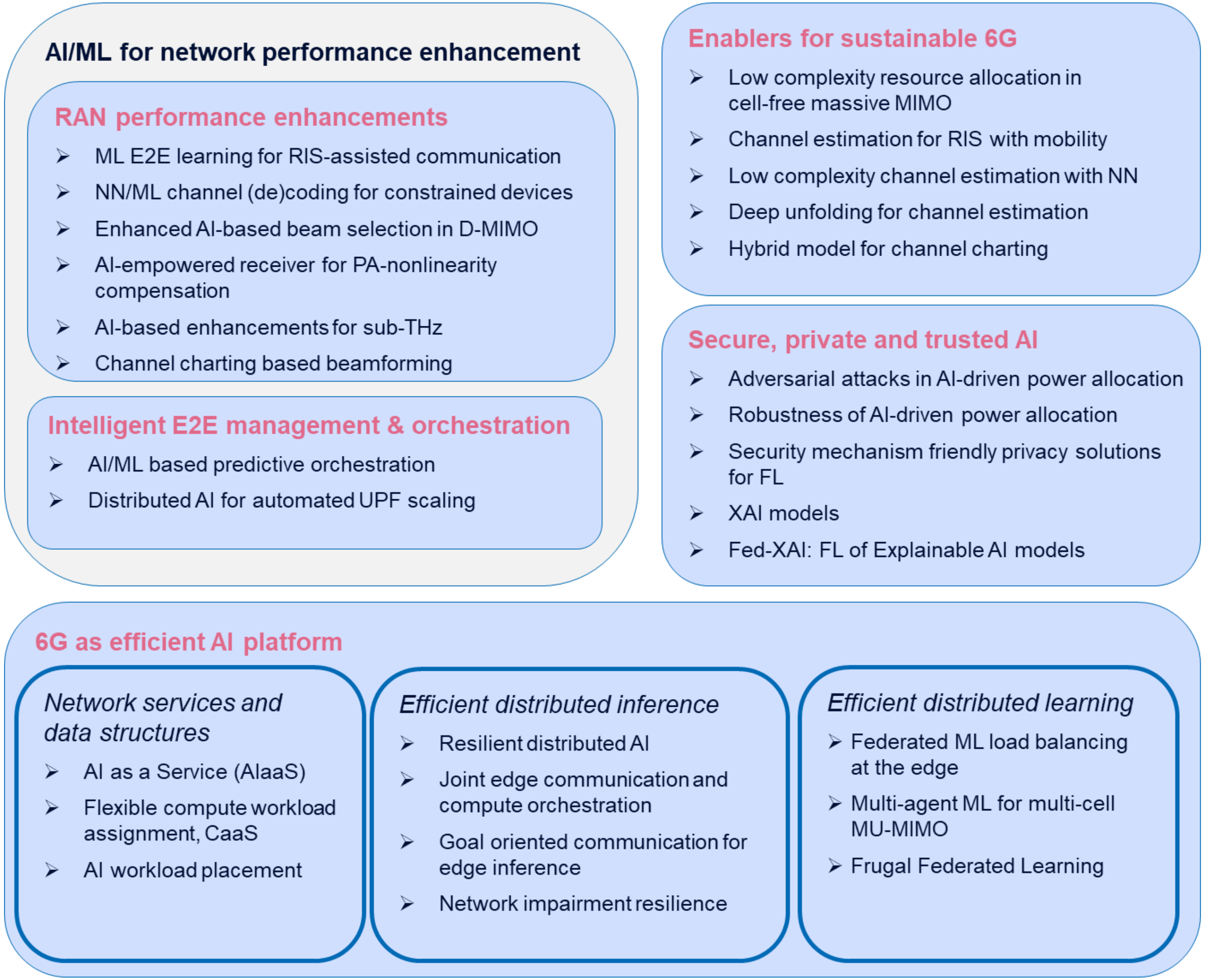
Conclusions on tackling the Connecting Intelligence research challenge

Summary

- AI-driven communication & computation co-design will constitute a major leap forward by 6G systems over previous generations. Instead of addressing AI and computation tasks on higher layers only, corresponding enablers will be deeply anchored in future 6G system and will rely on AI- and compute-native design approaches.
- D4.3 provides new innovative proposals and detailed studies of corresponding key enablers – further adding details, substantive insight and evaluation results over the previous Hexa-X Deliverables D4.1 and D4.2.
- Technical areas of focus are the following:
  - Network performance enhancement using AI/ML in 6G
  - AI/ML as an enabler for 6G network sustainability
  - 6G network as an efficient AI platform
  - Privacy, security & trust in AI-enabled 6G
  - Demonstration activities - Federated eXplainable AI (FED-XAI) demo
- Finally, all new approaches have been evaluated again Hexa-X Key Performance Indicators (KPIs) and Key Value Indicators (KVIs) as they have been defined on a project wide level. Chapter 8 of D4.3 summarizes the *Hexa-X quantified targets in Connecting intelligence towards 6G*.

Summary of the AI-driven solutions of the document by their domain of application

- The technical solutions described in this report contribute to various network domains.
- A corresponding breakdown to technical solutions in the illustration on the right-hand side.



AI/ML for network performance enhancement

RAN performance enhancements

- ML E2E learning for RIS-assisted communication
- NN/ML channel (de)coding for constrained devices
- Enhanced AI-based beam selection in D-MIMO
- AI-empowered receiver for PA-nonlinearity compensation
- AI-based enhancements for sub-THz
- Channel charting based beamforming

Intelligent E2E management & orchestration

- AI/ML based predictive orchestration
- Distributed AI for automated UPF scaling

Enablers for sustainable 6G

- Low complexity resource allocation in cell-free massive MIMO
- Channel estimation for RIS with mobility
- Low complexity channel estimation with NN
- Deep unfolding for channel estimation
- Hybrid model for channel charting

Secure, private and trusted AI

- Adversarial attacks in AI-driven power allocation
- Robustness of AI-driven power allocation
- Security mechanism friendly privacy solutions for FL
- XAI models
- Fed-XAI: FL of Explainable AI models

6G as efficient AI platform

Network services and data structures

- AI as a Service (AIaaS)
- Flexible compute workload assignment, CaaS
- AI workload placement

Efficient distributed inference

- Resilient distributed AI
- Joint edge communication and compute orchestration
- Goal oriented communication for edge inference
- Network impairment resilience

Efficient distributed learning

- Federated ML load balancing at the edge
- Multi-agent ML for multi-cell MU-MIMO
- Frugal Federated Learning

Thank you!

HEXA-X.EU

Hexa-X

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101015956.