Alliance for AI, IoT and Edge Continuum Innovation

AI Factories for the Energy sector

Release 1.0

AIOTI WG Energy

2026

© AIOTI. All rights reserved.

Executive Summary

Europe's electricity system and its AI infrastructure are becoming one planning problem. Artificial intelligence (AI) is moving into grid operation and industry, while the compute behind it is becoming a major new load on the grid. Data centres used about 415 TWh in 2024, 1.5% of global electricity. Their demand grew 17% in 2025, and 50% for AI-focused centres.

Europe's answer on the compute side is the EuroHPC AI Factories: nineteen selected, with thirteen Antennas, and a call open for up to seven Gigafactories. None is dedicated to energy. Energy stakeholders can use them, but today they do so through isolated projects.

Our proposal: use the AI Factories Europe is already building for the Energy sector and make sure they support the grid and do not strain it. We call this the ***Energy AI Factory***. It is a usage pattern and reference architecture, not a new facility.

What we found:

- **Power, not compute, is the binding constraint.** Grid stability and access to electricity now limit AI growth more than computing capacity.
- **AI facilities behave unlike conventional data centres.** Racks of 100–200 kW need liquid cooling. Load can swing by 30–60% within milliseconds, which can destabilise low-inertia grids unless facilities are built with storage, ramp control, and fault ride-through.
- **The gap to operations is data and regulation as much as technology.** Energy data are fragmented, smart-meter and grid-measurement coverage is uneven across Member States, and AI that acts on critical infrastructure is high-risk under the EU AI Act, alongside NIS2 and the CER Directive.

The pattern: Three tiers work as one loop. At the grid edge, compact models run in milliseconds to seconds. AI Gates, sector-aware intermediaries such as Competence Centres and Digital Innovation Hubs, check data, screen regulation, and validate use cases. AI Factories then provide large-scale training and simulation, and the retrained models return to the edge. Pilots at BI-REX and the IT4LIA AI Factory support individual stages. The complete loop for an energy model has not yet been run end to end.

Recommendations:

- **European Commission and EuroHPC:** recognise energy within the AI Factory network, through a dedicated Energy AI Factory or an energy vertical across existing ones, as called for by the transmission system operators' association ENTSO-E. Add standing allocations for continuous retraining.
- **Commission and Member States:** fund and formalise the AI Gate ecosystem.
- **Commission, ENTSO-E and the DSO Entity:** deliver a mandated Common European Energy Data Space built on interoperable, sovereignty-preserving data sharing.
- **AI Factories and regulators:** build compliance into AI Factories from the first training run and clarify when continuous retraining requires a new conformity assessment.
- **Grid operators and AI Factory operators:** plan grid and compute together, and design AI Factories as grid-supportive flexibility assets.

Annex I gives AI Factories a checklist of what is needed to serve the Energy sector.

© AIOTI. All rights reserved.

Table of Contents

Executive Summary 2

Table of Contents 3

List of Figures 5

List of Tables 5

List of Acronyms 6

1 Introduction 8

1.1 Context 8

1.2 Purpose and Scope 8

1.3 Intended Audience 9

1.4 Organisation of This Paper 9

2 The Convergence of AI and Energy 9

2.1 The Digital Transformation of Energy Systems 9

2.2 CEI Infrastructure and Challenges 11

2.3 The Emergence of AI Factories 12

2.4 Gaps between AI Engineering and Operational Energy Environments 13

3 AI Factories Today, HEROES, HPC 15

3.1 Definition and Core Principles 15

3.2 AI Factories and Status 16

3.3 Differences between Traditional Data Centres and AI Factories 18

3.4 EUROHPC HEROES 19

4 Energy AI Factories Core Components and Capabilities 21

4.1 Data Management and Governance Frameworks 21

4.2 Network-data-compute Orchestration across the CEI-HPC Continuum 22

4.3 AI Lifecycle Management Aspects 22

4.4 Security, Privacy, Data Sovereignty 22

4.5 Interoperability and Standards 23

4.6 Regulatory and Compliance Considerations 23

5 Challenges Facing the Energy Sector 24

© AIOTI. All rights reserved.

5.1 Rising Energy Demand and Electrification, Increasing Power Consumption....24
5.2 Renewable Integration and Grid Volatility....25
5.3 Cooling Technologies and Efficiency Strategies....26
5.4 Infrastructure Aging and Operational Complexity....26
5.5 Data Fragmentation across Energy Systems....27
6 Use cases and Applications....28
6.1 Continuous Training and Deployment of Energy Models....28
6.2 Energy System Digital Twins and Simulations....29
6.3 Autonomous Grid and Asset Management....30
6.4 Predictive Maintenance for Energy Assets....31
6.5 Energy Trading and Market Optimization....32
6.6 Industrial Energy Efficiency....33
6.7 Cross-sector Alignment....34
7 Unified Vision, Architecture of an Energy AI Factory....36
7.1 Data Ingestion from the Grid and Industrial Systems....37
7.2 High-performance AI Infrastructure....38
7.3 Model Development and Lifecycle Management....39
7.4 Real-time Inference at the Grid Edge....40
7.5 Alignment with Energy Regulations....40
7.6 Sustainability....41
8 Future Outlook and Recommendations....42
8.1 Emerging technologies and trends....42
8.2 Evolution of AI-enabled energy systems....42
8.3 Recommendations....43
References....45
Contributors....48
Annex I: Checklist, Requirements to become an Energy AI Factory....48
About AIOTI....52

© AIOTI. All rights reserved.

List of Figures

Figure 1: The Energy AI Factory is the energy-oriented use of existing AI Factories. It combines their compute with energy domain knowledge, software engineering and data communications, and applies it to the main challenges of the Energy sector. 8
Figure 2: grid transformation and data centre development converge into a single planning problem, connected by the CEI-HPC continuum. 10
Figure 3: EuroHPC AI Factories (19) and AI Factory Antennas (13) across Europe. Markers indicate host countries, not host cities; countries with two AI Factories show two markers. Source: EuroHPC JU. 16
Figure 4: Energy AI Factory components and cross-cutting requirements. 21
Figure 5: Energy AI factory, challenges for the Energy sector. 24
Figure 6: The AI Gate pattern: validation at the gate, scale-up in the AI Factory, deployment and monitoring at the edge, with drift-triggered retraining. Source: BI-REX. 29
Figure 7: (a) Large-format 3D-printed wind-turbine components prepared for experimental testing against design calculations; (b) Large Eddy Simulation of a motorcycle and rider in a wind tunnel run on HPC. Source: BI-REX PoC catalogue 30
Figure 8: (a) Agile quadruped robot with LiDAR surveying the underground canals of Bologna; (b) PLC and remote I/O module connected over a private 5G standalone network. Source: BI-REX PoC catalogue. 31
Figure 9: – (a) Acoustic anomaly-detection pipeline on a CNC machine: time-series storage, neural network, real-time dashboard and remote operator alert; (b) audio signal and anomaly score. Source: BI-REX PoC catalogue. 32
Figure 10: Reserve marketplaces in Finland overseen by Fingrid (FFR, FCR-D, FCR-N, aFRR, mFRR). Source: Fingrid / Predictricity project. 33
Figure 11: (a) Edge interface of the non-invasive energy-tracking kit (phase currents and waveforms) on a legacy machine; (b) measured vs. predicted power absorption used for unsupervised anomaly detection. Source: BI-REX PoC catalogue. 34
Figure 12: Roles in the AI Gate model: IT4LIA as upstream AI Factory, BI-REX as AI Gate, downstream adopters, and the joint KPIs foreseen by BI-REX and CINECA. Source: BI-REX. 35
Figure 13: proposed software architecture pattern for an Energy AI Factory. 36
Figure 14: what the grid sees from an Energy AI Factory (30–60% swings, forced oscillations, UPS-driven flapping), how the facility is built to handle it, and what the grid gains (frequency reserves, cooling load shifting, waste-heat reuse). 37
Figure 15: Energy AI Factory layered infrastructure: compute, and network together. 38

List of Tables

Table 1: Overview of EuroHPC AI Factories: hosting countries, systems, and sectoral focus, and whether the AI Factory considers services relevant to the Energy sector. 17
Table 2: Latency tiers of an Energy AI Factory. Market timescales are those of the Finnish reserve markets (Section 6.5) 40

© AIOTI. All rights reserved.

List of Acronyms

Acronym	Meaning
5G SA	Fifth-generation mobile network, Standalone
ACER	EU Agency for the Cooperation of Energy Regulators
AI	Artificial Intelligence
AIOTI	Alliance for the Internet of Things Innovation
AMR	Autonomous Mobile Robot
aFRR	automatic Frequency Restoration Reserve
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
BI-REX	Big Data Innovation and Research Excellence (Italian national Competence Centre)
BSC	Barcelona Supercomputing Center
CEI	Cloud-Edge-IoT
CER	Critical Entities Resilience (Directive)
CFD	Computational Fluid Dynamics
CI/CD	Continuous Integration / Continuous Delivery
CNC	Computer Numerical Control
CODECO	Cognitive, Decentralised Edge-to-Cloud Orchestration
CPU	Central Processing Unit
DER	Distributed Energy Resource
DERA	(ENTSO-E) European (energy) Data Exchange Reference Architecture
DESAP	Digitalisation of Energy System EU Action Plan
DSO	Distribution System Operator
DSOE	EU DSO Entity
EAR	Energy Aware Runtime
ECMWF	European Centre for Medium-Range Weather Forecasts
EDC	Eclipse Dataspace Components
EDIH	European Digital Innovation Hub
ENISA	European Union Agency for Cybersecurity
ENTSO-E	European Network of Transmission System Operators for Electricity
ERP	Enterprise Resource Planning
ESG	Environmental, Social and Governance
EU	European Union
EuroCC	European network of National Competence Centres for HPC
EuroHPC JU	European High-Performance Computing Joint Undertaking
FAIR	Findable, Accessible, Interoperable, Reusable
FBK	Fondazione Bruno Kessler
FCR-D / FCR-N	Frequency Containment Reserve for Disturbances / for Normal operation
FEM	Finite Element Method
FFR	Fast Frequency Reserve
GDPR	General Data Protection Regulation
GENCI	Grand Équipement National de Calcul Intensif (French national HPC agency)
GPU	Graphics Processing Unit

© AIOTI. All rights reserved.

Acronym	Meaning
GRETAP	Generic Routing Encapsulation (TAP) tunnelling
HEROES	Hybrid Eco Responsible Optimized European Solution
HLRS	High-Performance Computing Center Stuttgart
HPC	High-Performance Computing
HPCQS	HPC Quantum Simulator (EuroHPC quantum-HPC hybrid incubator)
IDS	International Data Spaces
IEA	International Energy Agency
IED	Intelligent Electronic Device
IFAB	International Foundation Big Data and Artificial Intelligence for Human Development
IoT	Internet of Things
IP	Intellectual Property
KPI	Key Performance Indicator
LiDAR	Light Detection and Ranging
mFRR	manual Frequency Restoration Reserve
ML	Machine Learning
MLOps	Machine Learning Operations
NIS2	Network and Information Security Directive (2)
NTUA	National Technical University of Athens
OS	Operating System
OT	Operational Technology
PCSS	Poznań Supercomputing and Networking Center
PLC	Programmable Logic Controller
PMU	Phasor Measurement Unit
PNRR	Piano Nazionale di Ripresa e Resilienza (Italian Recovery and Resilience Plan)
PoC	Proof of Concept
PUE	Power Usage Effectiveness
QKD	Quantum Key Distribution
RTU	Remote Terminal Unit
SCADA	Supervisory Control and Data Acquisition
SME	Small and Medium-sized Enterprise
SMR	Small Modular Reactor
TEF	Testing and Experimentation Facility
TSO	Transmission System Operator
UPS	Uninterruptible Power Supply
W4E	Weather4Energy
WG	Working Group

© AIOTI. All rights reserved.

1 Introduction

Europe is building shared AI compute but has not yet worked out how the Energy sector should use it or how it affects the grid.

Why look at AI Factories from an energy perspective? This section sets out the context of the paper, its purpose and scope, the readers it is written for, and how the rest of the paper is organised.

1.1 Context

Europe's energy system is being reshaped by electrification, distributed generation, and growing dependence on digital technologies, while artificial intelligence is moving from experimentation to operational use across the sector. At the same time, the compute that powers AI is becoming a major new load on the grid. Europe has responded to the compute side of this shift with the EuroHPC AI Factories [1], shared ecosystems that combine supercomputing capacity, data, and talent. None of them is dedicated to the Energy sector, and how energy stakeholders can use them, and how they in turn affect the grid, is still largely unaddressed.

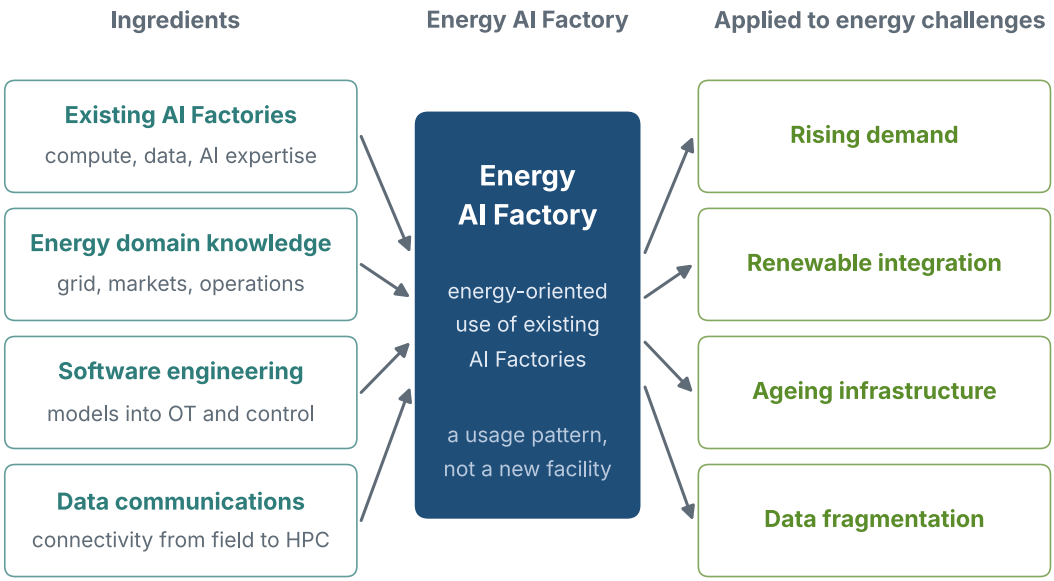


Figure 1: The Energy AI Factory is the energy-oriented use of existing AI Factories. It combines their compute with energy domain knowledge, software engineering and data communications, and applies it to the main challenges of the Energy sector.

1.2 Purpose and Scope

This white paper was prepared by the Working Group on Energy of the Alliance for the Internet of Things Innovation (AIOTI) and builds on the group's earlier work on AI in energy. It examines how AI Factories, High-Performance Computing (HPC) and the Cloud-Edge-IoT (CEI) continuum can address four challenges of the Energy sector: rising demand, renewable integration, ageing infrastructure, and data fragmentation. It also looks at the opposite direction, namely what the growth of AI-optimised data centres means for the energy system.

The paper proposes **a reference architectural pattern** for the energy-oriented use of existing AI Factories, **which it calls the Energy AI Factory**, and closes with recommendations and a checklist for AI Factories to be enabled to support relevant services of the Energy sector (Annex I). It

© AIOTI. All rights reserved.

focuses on Europe, on the electricity system and on industrial energy use. It does not propose new hardware or evaluate individual vendors or products.

In this paper, an **Energy AI Factory** means the energy-oriented use of existing AI Factories. It applies their shared compute, data, and expertise to Energy sector challenges through energy-specific data pipelines, intermediaries called AI Gates, and the operational requirements of the sector. It is a usage pattern and a reference architecture, not a new type of facility.

1.3 Intended Audience

The paper is written for transmission and distribution system operators, utilities and aggregators, data centre and HPC operators, industrial energy users, technology providers, and policymakers and funding bodies responsible for European AI and energy strategy. It assumes familiarity with energy systems or with digital infrastructure, but not with both.

1.4 Organisation of This Paper

Section 2 describes the convergence of AI and energy: the digital transformation of energy systems and the growth of data centres as one planning problem, the IoT-Edge-Cloud infrastructure that connects them, the emergence of AI Factories, and the gaps between AI engineering and operational energy environments. Section 3 presents AI Factories as they exist today, including their definition, the status of the EuroHPC network, how they differ from traditional data centres, and EuroHPC HEROES as an energy-aware HPC precedent. Section 4 describes the core components and capabilities of an AI Factory: data management and governance, orchestration across the continuum, AI lifecycle management, security and sovereignty, interoperability and standards, and regulatory compliance. Section 5 sets out the challenges facing the Energy sector, from rising demand and renewable integration to cooling, ageing infrastructure, and data fragmentation. Section 6 maps these challenges to use cases and applications, supported by proof-of-concept evidence. Section 7 brings the elements together in a reference architecture for the Energy AI Factory. Section 8 closes with future trends and five recommendations for policy, ecosystem, and technical action.

2 The Convergence of AI and Energy

Grid digitalisation needs more compute, and compute is becoming a grid load, so the two must be planned as one system.

Where do the energy system and AI infrastructure meet? This section sets out four parts of that convergence: the digital transformation of energy systems, the CEI-HPC infrastructure that carries grid data, the emergence of AI Factories, and the gaps between AI engineering and operational energy environments.

2.1 The Digital Transformation of Energy Systems

Energy systems are evolving in two directions: (i) grid transformation and (ii) data centre development, as represented in Figure 2.

Grids have been shifting from centralized, predictable generation toward distributed, bidirectional, and highly dynamic architectures involving *Distributed Energy Resources (DERs)*, including flexible loads such as electric vehicles and heat pumps, as well as active prosumer participation. This shift depends on pervasive sensing, real-time analytics, and increasingly autonomous decision-making.

The second direction is the transformation of data centres themselves, which is now unavoidable. Data centres are no longer passive IT facilities sitting on the grid. They are

© AIOTI. All rights reserved.

becoming energy and infrastructure systems in their own right. A modern AI-optimised data centre is, functionally, a large industrial load with its own supply requirements, cooling and water demands, and siting constraints, and its growth is now a first-order variable in energy systems planning.

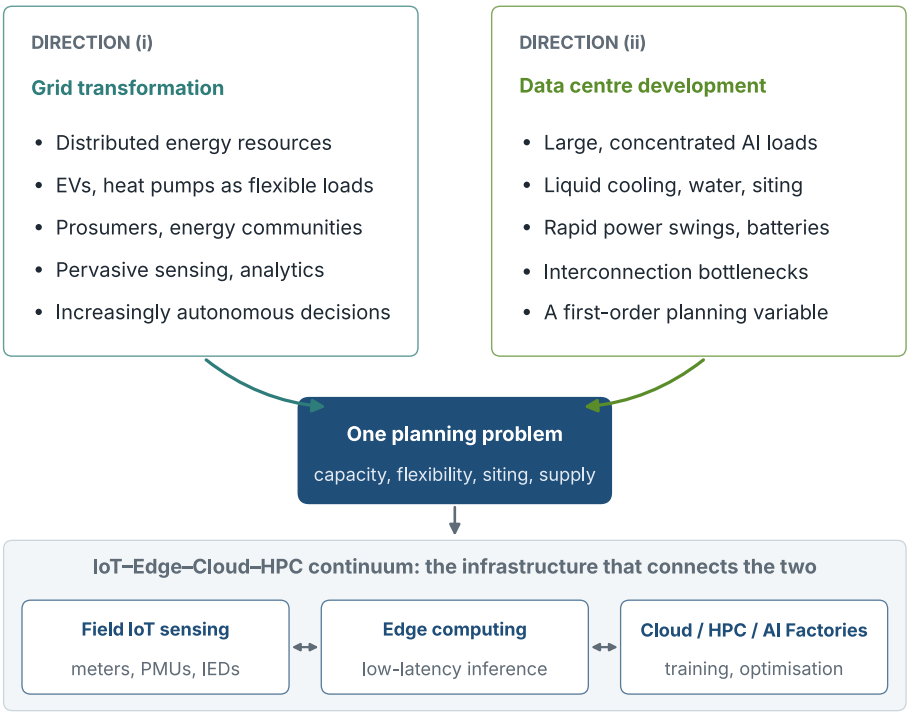


Figure 2: grid transformation and data centre development converge into a single planning problem, connected by the CEI-HPC continuum.

AIOTI's WG Energy has been addressing this landscape, with core topics that include the impact of advanced digitalisation on the Energy sector, AI and digital twins, distributed generation and energy communities, sector integration, grid flexibility and resilience, and cross-domain interoperability across the computing continuum. Among these, AI and digital twins are central, because manageing the converged system depends on accurate models of its behaviour. Energy systems are inherently complex: they consist of heterogeneous, interconnected assets whose interactions produce emergent behaviour that individual assets do not show. Manageing and operating them optimally requires stakeholders to capture and emulate this behaviour accurately under varying conditions. First-principles models may be considered, but their accuracy can degrade as systems age or change over time. AI can approximate real system behaviour, either by adjusting an existing model to match the real-world system or by building a model where none exists. In both cases the result is an energy system digital twin. Training and running such models at scale, and keeping them aligned with the physical system, is where HPC and AI Factory resources become relevant (Section 6.2).

Meeting this load is reshaping generation planning. According to IEA, renewables remain the fastest-growing supply source for data centres, growing at roughly 22% annually and meeting close to half of incremental demand through 2030, but natural gas and coal together still meet over 40% of additional near-term demand, and nuclear, including a fast-growing pipeline of small modular reactor (SMR) offtake agreements with data-centre operators, which grew from 25 GW at end-2024 to 45 GW today, is becoming a strategic supply option specifically because of this load. In some regions, data centres already represent a material share of national or state-level demand (e.g. over 20% in Ireland), forcing grid operators to treat hyperscale and AI Factory-class facilities as anchor loads requiring dedicated planning, not incremental additions absorbed by existing headroom.

© AIOTI. All rights reserved.

Power is not the only constraint. Rising accelerator power densities are pushing thermal management beyond air-cooling limits toward direct-to-chip and immersion liquid cooling. This substantially reduces direct water consumption relative to evaporative cooling but does not eliminate the underlying electricity demand, which remains the dominant environmental cost. Beyond cooling, water availability, grid interconnection queue times, transformer and gas-turbine supply-chain bottlenecks, and land and siting constraints are now as material to build-out timelines as capital availability.

AI workloads also introduce a qualitatively different load profile from legacy IT: rapid, large swings in power draw tied to training and inference cycles. These can exceed what onsite gas generation can reliably follow, which is driving adoption of onsite battery storage, both to protect facility reliability and, potentially, to offer flexibility services to the grid.

For the Energy sector, this means AI data centres cannot be designed purely as consumers of grid capacity to be accommodated. The digital transformation of energy systems and the infrastructural transformation of data centres **are converging into a single planning problem**. The remainder of this paper treats that convergence as its organizing frame, examining how existing AI Factories can be used as ***energy-oriented AI Factories***: shared compute, data, and expertise resources applied to Energy sector challenges, rather than treating grid digitalisation and data centre growth as separate tracks.

2.2 CEI Infrastructure and Challenges

Applying AI to the grid, including through the AI Factory resources discussed in this paper, depends on a flexible and resilient infrastructure that carries and processes grid data. This infrastructure spans three tiers with fundamentally different constraints: field-level IoT sensing, edge computing near the point of measurement, and cloud and HPC-class compute for training and system-wide optimisation. Together they form the **CEI continuum**. This section examines where that continuum currently stands in Europe, and why it remains a persistent bottleneck.

Field-level IoT deployment is now substantial, but uneven across Europe. By the end of 2024, about 63% of Europe's electricity consumers had smart meters, with more than 195 million units deployed across the EU 27+3 region, and the installed base is projected to reach approximately 285 million meters, or around 80% penetration, by 2029 [2]. That growth is far from evenly distributed: ACER reports 80% smart-meter deployment in half of EU member states, but penetration below 20% in seven member states, with Germany specifically flagged among the lowest-ranked countries according to ACER [3]. This lack of balance impacts directly AI Factory-scale energy applications, since forecasting and grid-optimisation models are only as good as the field-level data feeding them, and a fragmented rollout means training-data quality and availability vary sharply by member state. The regulatory foundation for this rollout traces back to EU electricity market directives, which have progressively required member states to pursue smart-metering targets since 2009, most recently through the Directive on Common Rules for the Internal Market for Electricity (Directive (EU) 2019/944).

Edge computing is becoming essential for time-sensitive grid functions. A growing body of peer-reviewed work makes the case that cloud-only architectures cannot meet the latency requirements of modern grid operations. Time-sensitive applications, e.g., coordinated electric-vehicle control, online state estimation, and online demand response, depend on bounded latency characteristics that only edge computing can provide, and protection schemes increasingly use intelligent electronic devices as edge-computing nodes for fast, low-latency short-circuit diagnosis[i] [4]. A recent systematic review of edge computing in power systems similarly finds that traditional cloud-based solutions face high latency, limited bandwidth, and cybersecurity vulnerabilities that make them less effective for real-time smart-grid applications, since edge computing minimizes communication latency by reducing the distance data must travel before being analysed [5].

© AIOTI. All rights reserved.

Edge AI is following the same trajectory: a review of Edge AI for the Internet of Energy finds that connecting sensors and IoT devices to edge platforms enables real-time energy monitoring and rapid local decision-making, with model-optimisation and pruning tools allowing AI inference to run efficiently on resource-constrained field devices, therefore forming a foundational layer for reliable, efficient, and timely processing at the grid edge [6].

Furthermore, the continuum has no single standard architecture yet, which is precisely the gap AIOTI's own work in this area is trying to close. AIOTI's WG Standardisation, working with the EUCloudEdgeIoT.Eu and Open Continuum projects, has published *Towards a Computing Continuum Reference Architecture* [7], which provides a taxonomy, analyses what it takes to bridge the gap from IoT to edge and cloud, and offers an architecture analysis together with a perspective on validation. This effort sits alongside a broader European push towards the development of a "smart-energy OS" as a framework for enabling flexibility in energy consumption by controlling power load in IoT-enabled systems, reflecting a broader need for architectural concepts capable of abstracting away the underlying heterogeneity of sensor and system environments.

Practical constraints remain significant even where edge and cloud tiers are individually mature. Connecting them reliably across the Energy sector's heterogeneous legacy base considering SCADA systems, RTUs, and proprietary substation protocols is not an easy task. Reviews of the field note recurring difficulty in reducing communication overhead across the CEI continuum and in achieving effective federated learning across the constrained, RESTful environments typical of field devices meaning that even well-architected reference models still face real deployment friction at the device layer [8].

Hence, an uneven but growing field-level IoT penetration, an accumulating scientific and industrial consensus that bounded latency edge computing is required rather than optional for real-time grid functions, and an as-yet-unfinished European reference architecture for the computing continuum define the infrastructure gap that an Energy AI Factory (Section 6) must bridge. The HPC-class compute discussed in Section 3 only delivers value for the Energy sector if it can reliably ingest from, and push validated models back down to, this still-maturing edge layer.

2.3 The Emergence of AI Factories

AI Factories provide a candidate architectural response to three problems that recur throughout Sections 2.1 and 2.2: fragmented energy data ecosystems, a persistent gap between AI engineering and *Operational Technology (OT)* environments, and unresolved challenges of scalability, trust, and compliance.

AI factories can be defined as hubs that use the supercomputing capacity of the EuroHPC Joint Undertaking to develop trustworthy, cutting-edge generative AI models, helping improve areas including health, manufacturing, climate, and finance. They can be seen as dynamic, interconnected, systems of ecosystems that bring together computing power, data, and talent, linking supercomputing centres, universities, SMEs, industry, and financial actors across Europe. The Commission identified the establishment of AI Factories as a strategic priority in the 2024 AI Innovation Package, reinforced by the AI Continent Action Plan, and the underlying EuroHPC regulation was formally amended in 2024 to include their development and operation [9]. The AI Factory European network, which prioritises access for AI startups and SMEs and plans at least nine new AI-optimised supercomputers to more than triple EuroHPC AI computing capacity, is catalogued in detail in Section 3.2.

Fragmentation in current energy data ecosystems. This paper's Section 2.2 already documented uneven smart-meter penetration and the absence of a single standard CEI reference architecture. AIOTI's own prior work in this area, the *AI in Energy* white paper [10], points to the same fragmentation from the data-governance side, noting that interoperability through data

© AIOTI. All rights reserved.

spaces and standardization for data exchange through semantic ontology are treated as core pillars precisely because AI systems are central components across most of the energy use cases examined, and cannot function reliably without them. An AI Factory does not by itself resolve this fragmentation, but it concentrates the compute and governance capacity needed to enforce common data contracts across DSOs, TSOs, aggregators, and industrial actors, as further addressed in section 3.1.

Gaps between AI engineering and operational environments. The difficulty of embedding AI models into live operational infrastructure is well documented outside the Energy sector and applies directly within it. Industry analysis of AI adoption in utilities specifically finds that embedding models into outage-management systems, asset-management platforms, or field-service workflows is rarely straightforward, and that small integration gaps can block adoption entirely, while outputs in regulated environments must remain explainable and defensible for trust and compliance reasons. This mirrors a broader pattern identified in enterprise AI research: large-scale surveys consistently report that organizations widely experiment with AI but that most deployments remain limited to pilot projects rather than integrated, production-grade systems, with legacy-system integration and governance gaps cited as major obstacles to scaling. AI Factories, by providing standardized, validated compute and model-lifecycle infrastructure are positioned as one mechanism for narrowing this pilot-to-production gap. However, to achieve this, AI Factories must be designed with OT-grade reliability and safety requirements in mind from the outset.

Challenges in scalability, trust, and compliance. Trust remains the central constraint on scaling AI beyond pilots. Recent industry research on responsible AI adoption finds that nearly 60% of organizations cite knowledge and training gaps as the primary barrier to implementing responsible-AI practices, and that organizations which treat AI trust as a core operational capability are better positioned to scale AI adoption to its full potential [11]. For energy specifically, AIOTI's own *AI in Energy* work reports that a dedicated practice on AI trustworthiness was under development to support pilot projects in developing and manageing their AI systems throughout the AI lifecycle, explicitly taking into account regulation and standardization requirements, directly anticipating the AI-lifecycle and compliance considerations which are further elaborated in Sections 4.3 and 4.6 of this paper.

Why this matters for energy specifically. Unlike general-purpose AI Factories, an Energy sector deployment inherits safety-critical infrastructure obligations that most AI Factory use cases to date have not had to confront at the same intensity: control-loop latency requirements, critical-infrastructure security norms, and sector-specific regulation spanning network codes and unbundling rules. The remainder of this paper argues that these requirements justify the energy-oriented use of existing AI Factories, referred to here as the *Energy AI Factory* (Section 1.2), as a distinct architectural pattern, rather than assuming that general-purpose AI Factory infrastructure will transfer to energy without modification.

2.4 Gaps between AI Engineering and Operational Energy Environments

Three main gaps between the engineering of AI across the continuum, and the existing operational energy environments are debated in this section: scalability, compliance, and data access.

Scalability. For an Energy AI Factory, scalability is not only an organisational question (moving AI from pilots to production) but a concrete compute-access question that EuroHPC's own AI Factories programme is already trying to solve. Access to AI Factory infrastructure is tiered specifically to address this gradient [12]: Playground access provides limited resources for entry-level users and initial experimentation, Fast Lane access offers a lighter-weight process for users already familiar with HPC with up to 50,000 GPU hours, and Large-Scale access caters to development of large AI models and applications requiring more than 50,000 GPU hours. For energy specifically, the remaining scalability gap is less about compute access, which the AI

© AIOTI. All rights reserved.

Factory tiering model addresses reasonably well, and more about whether a model validated at Fast Lane or Large Scale can then be embedded into control-loop and asset-management systems that were never designed with model-serving interfaces in mind (Section 6.3).

Compliance. Energy sector AI Factory deployments inherit a denser compliance stack than most other AI Factory use cases, because energy is explicitly named as critical infrastructure under EU law. Under the EU AI Act, AI systems intended for use as safety components in the management and operation of critical infrastructure including the supply of electricity, gas, and heating are classified as high-risk under Annex III, point 2, with that classification drawing on the sectoral scope of the Critical Entities Resilience Directive (EU) 2022/2557 [13]. This has direct, practical consequences for AI Factory deployments: high-risk classification triggers obligations around risk management, data governance, technical documentation, and conformity assessment before a system can be deployed in live grid operations. The classification is not automatic for every energy AI use case. The European Commission's draft guidelines confirm a narrower gateway requiring cumulative conditions, including that the system function as a genuine "safety component," so that, for example, a predictive-maintenance tool that only alerts a human operator may fall outside the high-risk category while a system that automatically adjusts grid operations would clearly qualify. Volkova et al. provide a review of the resulting compliance landscape concludes that the Energy sector has effectively been categorized as a high-risk field of application for AI wherever safety components are involved in manageing critical infrastructure, spanning both physical and digital critical infrastructure as defined in the underlying directive [14]. Hence, for an Energy AI Factory, this means compliance work, i.e., conformity assessment, documentation, human-oversight design needs to begin at the model-development stage inside the AI Factory.

Data discovery and access. Even where energy data exists and is not withheld for competitive reasons, it is frequently not *findable* or *reusable* in a form an AI Factory's training pipelines can consume directly. This is a documented, sector-specific problem: a systematic study testing FAIR (Findable, Accessible, Interoperable, Reusable) data practices across 80 databases relevant to the low-carbon energy transition found that operationalising the FAIR principles remains an unresolved challenge for the domain, documenting concrete bottlenecks from both human and machine perspectives on discoverability and reuse [15]. Recent project-level work shows a practical path forward: the **GREEN.DAT.AI** initiative [16] demonstrates how a shared, FAIR-compliant data catalog layered over federated, sovereignty-preserving connectors provides the ingestion governance an Energy AI Factory requires at its front end.

© AIOTI. All rights reserved.

3 AI Factories Today, HEROES, HPC

Nineteen AI Factories are selected, and they differ from conventional data centres in ways the grid will feel.

What are AI Factories, and how do they differ from the data centres the grid already knows? This section sets out four topics: the definition and core principles of AI Factories, their status across Europe, their differences from conventional data centres, and EuroHPC HEROES as an example of energy-aware computing.

3.1 Definition and Core Principles

An AI Factory is officially defined as a "one-stop shop" and "dynamic ecosystem" built around the EU's supercomputers, managed by the European High-Performance Computing Joint Undertaking, with a primary mission to develop trustworthy, cutting-edge AI models[1]. Building on the EuroHPC ecosystem introduced in Section 2.3, highly relevant in the context of this white paper is the fact that the Energy sector's own transmission system operators have now explicitly called for this model to be applied to energy: in its response to the European Commission's consultation on the Strategic Roadmap for Digitalisation and AI in the Energy Sector, ENTSO-E lists among its eight key priorities the **"establishing [of] EU 'AI factories' and secure computing clusters for energy-relevant AI model training"** [16]**.**

The European power system is undergoing structural transformation on two fronts simultaneously. First, TSO–DSO coordination itself is digitalising: the Joint Task Force of ENTSO-E and the DSO Entity, established under the European Commission's *Digitalisation of Energy System EU Action Plan (DESAP)*, has published a joint report underscoring the critical need for digitalisation to enhance grid operation, planning, and customer integration in order to meet European carbon-reduction targets [18]. Second, data centres themselves are becoming systemically relevant participants in that same power system: ENTSO-E's 2026 report [19][20] on data centres and the power system finds that as their scale and electricity demand expand, data centres are evolving into systemically relevant electricity users whose behaviour increasingly influences power system operation, while also noting potential for data centres to support grid stability through flexible consumption and market participation .

These two trends converge directly on the Energy AI Factory concept: it sits at the exact intersection of "grid digitalisation needs more compute" and "compute infrastructure is becoming a grid-relevant asset in its own right", as debated in section 2.1.

ENTSO-E's consultation response frames these three principles concretely for the energy domain. On interoperability, ENTSO-E calls for a federated architecture connecting TSO and DSO digital twins based on open standards. On resilience and governance, it calls for a clear governance model, for example an EU-level Digital & Data Office supported by an expert council, as essential to guarantee trust, interoperability, and security across the digitalised power system. On sovereignty, this echoes the broader rationale already established for EuroHPC AI Factories generally, which are framed as critical infrastructure akin to energy grids or transport networks, intended to ensure the EU can innovate without depending on the strategic decisions of other global powers. This framing basically reinforces the fact that the AI Factory infrastructure is being built to specifically serve the needs of the energy grid.

ENTSO-E's response explicitly calls on Europe to move beyond pilots and create a mandated, sustainably funded Common European Energy Data Space dedicated to electricity, designed

© AIOTI. All rights reserved.

for near-real-time operational data and built to enable coordinated TSO–DSO processes, flexibility platforms, and future federated digital twins. Furthermore, its roadmap for digitalisation explicitly proposes two priority federated digital twins: a TSO–DSO operational twin integrating structural and real-time grid data for congestion management, security analyses, and coordinated control, developed under a broader "Federated Approach to Digital Twins" research stream.

This positions digital twins not as a standalone use case but as a core architectural component. Because operational digital twins require near-real-time model inference rather than periodic batch updates, the compute and data layers of an Energy AI Factory must be designed to serve them continuously.

An Energy AI Factory must therefore support the full chain: ingesting near-real-time operational data via the Common European Energy Data Space, training and validating models on EuroHPC-tiered compute, and deploying federated digital twin and forecasting models back into TSO–DSO control environments. Governance has to oversee this entire pipeline, not only the point of deployment.

3.2 AI Factories and Status

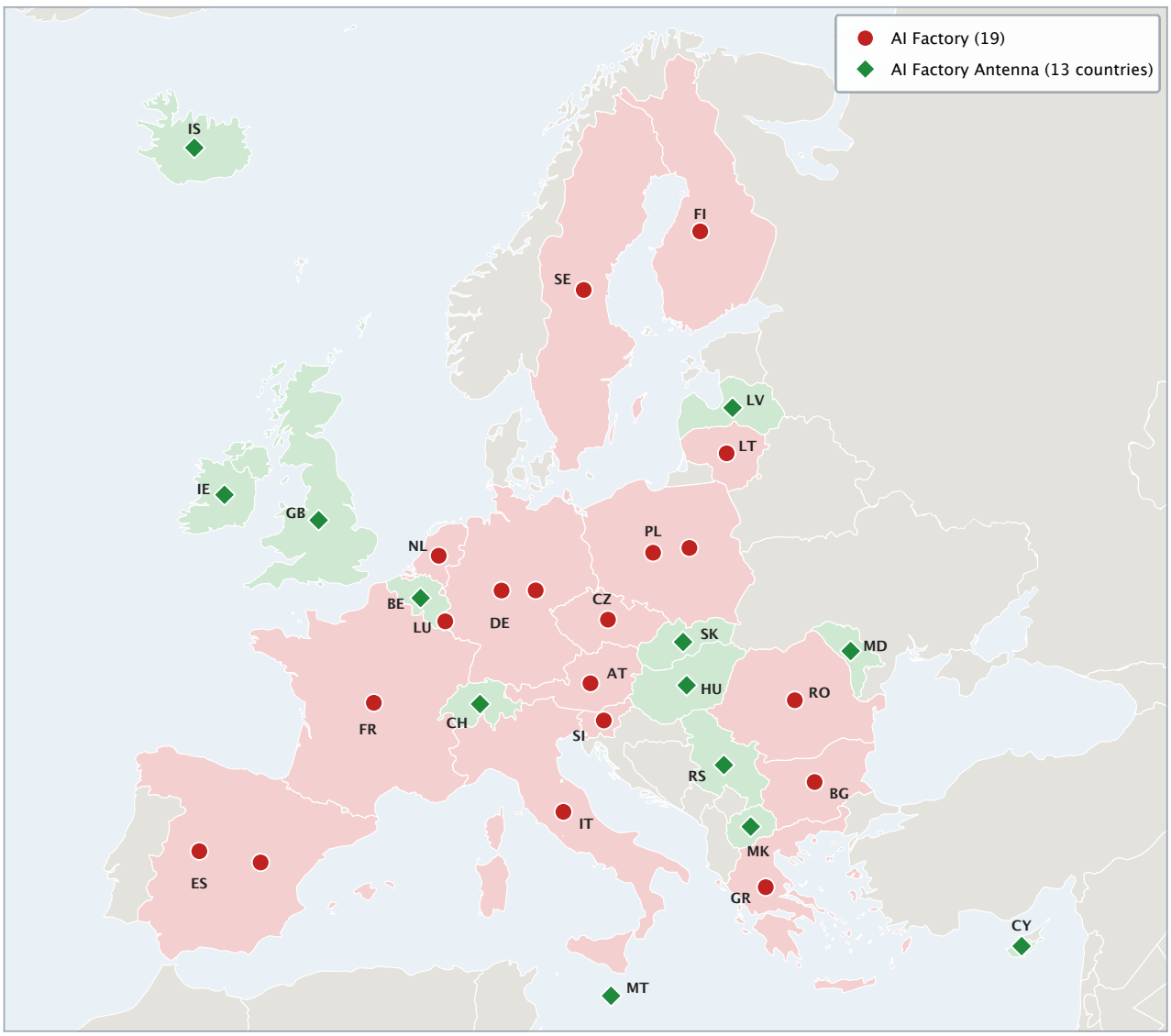


Figure 3: EuroHPC AI Factories (19) and AI Factory Antennas (13) across Europe. Markers indicate host countries, not host cities; countries with two AI Factories show two markers. Source: EuroHPC JU.

© AIOTI. All rights reserved.

As of November 2026, EuroHPC JU is overseeing the implementation of 19 AI Factories across Europe, alongside 13 AI Factory Antennas (Figure 3) that extend national reach without requiring a full supercomputer deployment[1]. The most recent selection round took place in October 2025, after three rounds in total: seven AI Factories were selected in December 2024 (Finland, Germany, Greece, Italy, Luxembourg, Spain, Sweden); six more followed in March 2025 (Austria, Bulgaria, France, Germany, Poland, Slovenia); and a third round, following the cut-off of 30 June 2025 under call EUROHPC-2024-CEI-AI-02, added six more (Czechia, Lithuania, the Netherlands, Poland, Romania, Spain). Not all of this infrastructure is newly built: the IT4LIA AI Factory, for example, has offered services since April 2025 on existing infrastructure, mainly the Leonardo supercomputer, while a new AI-optimised system for it was contracted in April 2026. Hence, the AI Factories network is being layered onto an already-substantial European HPC base rather than built from a green field. In parallel, EuroHPC launched a call for up to seven AI Gigafactories on 30 July 2026, with submissions due on 12 November 2026 and selection expected in early 2027.

Purpose and types of sectoral data. Individual AI Factories are explicitly framed around different sectoral priorities rather than a single generic mandate. Examples include IT4LIA [23] in Italy, positioned as a strategic initiative transforming the national and European AI landscape through advanced supercomputing infrastructure and a comprehensive service portfolio; the Spanish 1HealthAI [24] factory, aiming to position European One-Health products, services, research, and the broader economy among global leaders, spanning climate–health interactions, genomics, personalised medicine, sustainable agri-food systems, blue biotechnology, and environmental health; and national factories such as CZAI [25] (Czechia), focused on practical applications, technological infrastructure, and skills development, NLAIF (Netherlands), intended to bridge AI research breakthroughs and their rapid, large-scale application.

The Commission frames the sectoral coverage of AI Factories as spanning health, manufacturing, climate, finance, and space, and no AI Factory in the current 19-site network is dedicated specifically to energy. In the Commission's general messageing, energy appears mainly as a design concern (**the energy efficiency of AI Factory infrastructure**) and not as a target application domain; among individual sites, only Lithuania's LitAI Factory lists green energy among its priorities, alongside cybersecurity, manufacturing, and health technology.

Table 1 summarises the 19 AI Factories from the perspective of the Energy sector. Two sites name energy among their priorities, always alongside other sectors: Poland's PIAST-AI Factory (under sustainability) and Lithuania's LitAI Factory (green energy). Five further sites show an overlap with energy and climate, namely LUMI, HammerHAI (HLRS), PHAROS, IT4LIA, and the BSC AI Factory. The remaining twelve relate to energy only indirectly, as 1HealthAI does through climate and environmental health, or are general-purpose, with no identified energy focus. No AI Factory in the 19-site network has services specifically dedicated to the Energy sector.

Energy aspects. At present, energy-relevant HPC work therefore sits mostly in adjacent EuroHPC research and innovation projects, and in general-purpose AI Factories with partial sectoral overlap, not in a dedicated AI Factory. The HEROES project [26], for example, focuses on workflows of strategic importance in the renewable energy field, integrating energy-aware HPC/cloud scheduling via tools such as EAR, and the TwinREV project runs on PHAROS.

Table 1: Overview of EuroHPC AI Factories: hosting countries, systems, and sectoral focus, and whether the AI Factory considers services relevant to the Energy sector.

Country	AI Factory	Supercomputer / system	Sectoral focus	Relevance to energy
Poland	PIAST-AI	New system (PCSS, Poznań)	Cybersecurity, health and life sciences, public sector, space, robotics, sustainability	**Energy named as a priority** (under sustainability: energy, agriculture, climate)
Lithuania	LitAI	—	Cybersecurity, intelligent manufacturing, health technology, green energy	**Energy named as a priority** (green energy)

© AIOTI. All rights reserved.

Finland	LUMI AI	LUMI	Open, multi-sector ecosystem	**Energy and climate overlap**
Germany	HammerHAI	New standalone system (HLRS Stuttgart)	Manufacturing, engineering, automotive, mobility	**Energy and climate overlap**
Greece	PHAROS	DAEDALUS	Mainly health, also general	**Energy and climate overlap.** Hosts the TwinREV energy project
Italy	IT4LIA	Leonardo (+ Bologna Tecnopolo upgrade); IPAZIA HPC Center of BI-REX as AI Gate	National/industrial AI	**Energy and climate overlap**; energy-related proofs of concept via the BI-REX AI Gate
Spain	BSC AI	MareNostrum 5 (AI-accelerator upgrade)	Large-scale generative model training	**Energy and climate overlap**
Spain	1HealthAI	—	One Health: climate–health, genomics, agri-food	Indirect (climate and environment)
Germany	JAIF	JUPITER (exascale)	General/academic, federated learning	No energy focus identified
France	GENCI	Alice Recoque (per JAIF partnership reference)	Exascale capacity, federated learning with JAIF	No energy focus identified
Poland	GAIA	—	Healthcare, space, large language models	No energy focus identified
Czechia	CZAI	KarolAIna (based on Karolina)	Practical applications, infrastructure, skills	No energy focus identified
Netherlands	NLAIF	—	Healthcare, public safety, agriculture	No energy focus identified
Luxembourg	(unnamed)	MeluXina	General	No energy focus identified
Sweden	MIMER	—	General	No energy focus identified
Austria	AI:AT	—	General	No energy focus identified
Bulgaria	BRAIN++	Discoverer	General	No energy focus identified
Slovenia	SALIF	Vega	General	No energy focus identified
Romania	RO AI	—	General (SME-focused)	No energy focus identified

3.3 Differences between Traditional Data Centres and AI Factories

Traditional data centres are optimized for general-purpose compute, storage, and enterprise workloads with comparatively stable, predictable power draws. AI Factories differ from this baseline in ways that are not incremental but structural in power density, cooling architecture, demand behaviour, and underlying purpose. Each of these differences has direct implications for how an AI Factory interacts with the electricity grid it depends on.

Power density. The single clearest technical differentiator is power draw per rack. Industry survey data puts the global data-centre mean at around 7.6 kW per rack, with 5–9 kW the most common band [27]. AI-optimised facilities depart sharply from this: standard racks typically use 7–10 kW, while AI-capable racks can demand 30 kW to over 100 kW, averaging more than 60 kW in dedicated AI facilities, with the newest NVIDIA GB200 NVL36 and NVL72 servers reaching roughly 66 kW and 120 kW per rack, respectively [28]. However, industry roadmaps already point toward 235 kW racks with next-generation "Vera Rubin" systems and megawatt-scale racks in subsequent generations, a shift serious enough that vendors describe it as requiring a fundamental redesign of power distribution and cooling rather than an incremental upgrade [29].

Cooling architecture. Traditional data centre cooling was designed around air-cooling ceilings that cannot support AI-class density. Cooling accounts for roughly 7% of electricity consumption in efficient hyperscale facilities but can exceed 30% in less-efficient, non-AI-optimised ones [30]. AI Factories are increasingly moving to liquid and immersion cooling not as an efficiency optimisation but as a physical necessity, since air cooling supports up to roughly 20 kW per rack,

© AIOTI. All rights reserved.

liquid cooling up to around 100 kW, and immersion cooling 200+ kW while pushing power usage effectiveness (PUE) toward near-1.0 [31].

Demand volatility and grid interaction. Traditional data centres present relatively flat, predictable loads to the grid. AI workloads do not: GPU clusters can swing from idle to full power within milliseconds, producing peak-to-average power ratios that conventional *Uninterruptible Power Supply (UPS)* systems were not designed to absorb, which is why lithium-ion battery systems in AI facilities are now routinely oversized several-fold to buffer sudden surges, with battery energy storage increasingly treated as critical infrastructure sitting between UPS and backup generation. This is the same volatility pattern already discussed in Section 2.1 as a grid-stability concern: AI Factories are not just larger loads than traditional data centres, they are structurally spikier ones, with direct implications for how a grid operator must plan reserves and flexibility around them.

Beyond the physical differences, AI Factories are defined by mission rather than just hardware: they are purpose-built to develop trustworthy, cutting-edge AI models within a public–private research and industrial ecosystem (Section 3.1), rather than to provide generic compute or storage capacity to arbitrary tenants. This mission orientation is what justifies public co-funding and sovereignty framing that a conventional commercial data centre would not attract.

The aforementioned described differences mean an AI Factory cannot be planned, sited, or grid-connected using the same assumptions as a traditional data centre, a constraint that applies with particular force to any Energy AI Factory, since a facility whose stated purpose is to serve the grid should not simultaneously be a source of the demand volatility and connection complexity documented above without deliberate co-design between its compute infrastructure and the grid capacity it draws on, as shall be discussed in section 6.

3.4 EUROHPC HEROES

Within the EuroHPC JU's research and innovation portfolio, the project most directly relevant to an Energy AI Factory is HEROES (Hybrid Eco Responsible Optimized European Solution) funded under the Horizon 2020 call topic EuroHPC-02-2019 on "HPC and data centric environments and application platforms," alongside three sibling projects (ACROSS, NextSim, OPTIMA) launched in the same March 2021 cohort. HEROES ran for a two-year period with €328,346 in Horizon 2020 funding against a total project budget exceeding €890,000, led by the French HPC solution provider UCIT in a consortium of four European enterprises delivering HPC services to industrial clients. Unlike the general-purpose AI Factories catalogued in Section 3.2, HEROES was explicitly energy-oriented from inception. Although the project was designed to create value across multiple industrial sectors over time, its initial focus specifically addressed workflows of strategic importance in the renewable energy field, alongside manufacturing applications where HPC supports the design of more energy-efficient products such as energy-efficient vehicles. The consortium included direct energy-sector advisory partners such as Meteosim, UL Renewables, EDF, and Dallara, providing guidance on workflows relevant to these use cases, complemented by a major supercomputing research centre contributing specific expertise in energy management and resource optimisation.

HEROES' core innovation is a platform-selection module [32]: a decision-support layer that lets end users submit complex simulation and AI/ML workflows and choose the best execution environment, HPC data centres or cloud infrastructure, to achieve their goals within time, budget, and energy-efficiency constraints. For the energy-aware compute framing debated in this white paper, HEROES went beyond a purely theoretical platform-selection concept: EAR, the system software for data-centre energy management, was integrated directly into the HEROES platform, and the project's decision-making module was integrated into OKA, a pre-existing data-science platform for HPC environments, improving OKA's effectiveness in the process. In other words, HEROES demonstrates the exact mechanism an Energy AI Factory would need at system level: compute scheduling that treats energy cost and efficiency as a first-class

© AIOTI. All rights reserved.

optimisation input rather than an afterthought, in a project explicitly seeded with renewable-energy use cases and energy-sector industrial partners.

HEROES sits within a wider set of EuroHPC-funded application and skills projects including LIGATE (drug design), MICROCARD (disease modelling), NextSim (aerospace engineering), and the HPCQS quantum-HPC hybrid incubator that together illustrate the EuroHPC JU's second core objective: developing the applications, algorithms, and software that run on its supercomputers for public and private users across sectors including energy and climate research. This positions HEROES not as a standalone curiosity but as one instance of a broader EuroHPC pattern: sector-specific application development riding on top of shared, general-purpose HPC infrastructure, precisely the pattern this paper argues should be formalised into a dedicated Energy AI Factory rather than left as a project-level initiative with a fixed two-year funding horizon.

HEROES's outputs did not end with the project's Horizon 2020 funding period. Several of the platform's components are already available for integration into current computing infrastructures through Do it Now, a commercial venture of three consortium partners (UCit, France; HPCNow!, Spain; Do IT Systems, Italy), based in Lattes, France [33]. This gives the Energy AI Factory concept proposed in Section 7 a concrete, existing starting point for its energy-aware compute-orchestration layer, instead of requiring that capability to be built from scratch.

What HEROES does not provide yet and what distinguishes a project-level initiative from the AI Factory model proper is the long-term, sovereign, EuroHPC-governed infrastructure and access-tier model that has been presented in section 2.4, and that a formally designated AI Factory carries. HEROES demonstrates that energy-aware HPC scheduling for renewable-energy workflows is technically achievable and has already been built once; it does not by itself constitute the standing, funded, sector-dedicated infrastructure that ENTSO-E's consultation response is explicitly asking EuroHPC to establish.

© AIOTI. All rights reserved.

4 Energy AI Factories Core Components and Capabilities

Three capabilities (data, orchestration, lifecycle) and three cross-cutting requirements (security, interoperability, compliance) define what an AI Factory must provide for energy.

What must an AI Factory offer to serve the Energy sector? This section sets out six components: data management, orchestration across the computing continuum, AI lifecycle management, security and sovereignty, interoperability and standards, and regulatory compliance.

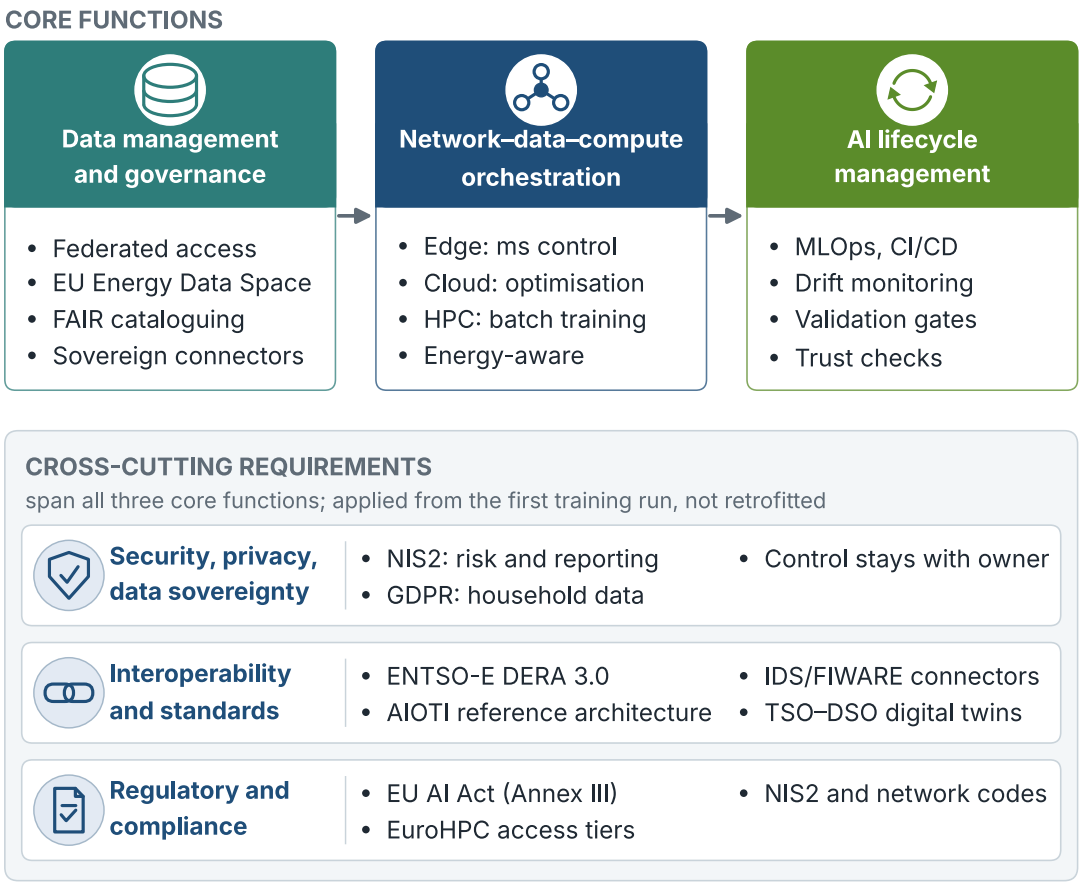


Figure 4: Energy AI Factory components and cross-cutting requirements.

4.1 Data Management and Governance Frameworks

An Energy AI Factory's data layer must federate across DSOs, TSOs, aggregators, and industrial actors without requiring physical data centralisation, operationalising the mandated Common European Energy Data Space called for by ENTSO-E and addressing the human and machine discoverability bottlenecks documented across low-carbon energy databases. Recent project-level work shows what the governance layer underneath such a dataspace looks like in practice: the GREEN.DAT.AI initiative builds an AI-ready data space whose Data Space Common Services enforce FAIR data management through a shared data catalog, while its architecture built on data-sovereignty principles and services such as an Eclipse *Dataspace Components (EDC)* connector ensures interoperability and GDPR compliance simultaneously.

An Energy AI Factory's data-governance framework should therefore be assessed against three requirements that the sector has already articulated: (i) a mandated Common European Energy Data Space, as called for by ENTSO-E; (ii) FAIR-compliant data cataloguing, as demonstrated by GREEN.DAT.AI and supported by the FAIR study of low-carbon energy databases; and (iii)

© AIOTI. All rights reserved.

sovereignty-preserving connectors that leave control of the data, including compliance with the GDPR where personal data is involved, with the originating party and not with the AI Factory.

4.2 Network-data-compute Orchestration across the CEI-HPC Continuum

Orchestration must span tiers with fundamentally different latency and reliability requirements: IoT and edge for millisecond-scale protection and control, cloud for near-real-time optimisation, and HPC for batch training and long-horizon simulation. Building on the AIOTI WG Standardisation Computing Continuum Reference Architecture described in section 3.2 and which frames CEI convergence as a system-of-systems problem, an Energy AI Factory must extend energy-aware orchestration from HPC and cloud to the grid edge. Two EU-funded projects cover complementary parts of this path. EuroHPC HEROES (Section 3.4) provides a platform-selection module that ranks HPC and cloud offers by time, budget, and energy consumption. CODECO [35] and its follow-up Eclipse technology KuDECO [36] provides cognitive, cross-layer orchestration of computing, data, and network services across the CEI continuum, aiming to reduce latency and power consumption from the far edge to the cloud, with energy among its use cases.

Within the latency and reliability limits of each tier, workload placement decisions (edge inference, cloud optimisation, or HPC training) should take energy into account, and not be driven by performance alone.

4.3 AI Lifecycle Management Aspects

Energy-domain models require lifecycle management adapted to safety-critical operational technology, not generic MLOps. The standard MLOps paradigm covering the end-to-end conceptualization, implementation, monitoring, deployment, and scalability of ML products, including CI/CD automation, versioning, continuous training and evaluation, and continuous monitoring for model performance and data drift provides the baseline vocabulary for this. That baseline has already been adapted for safety-critical scientific control-system contexts closely analogous to grid operations: a recent AI-ready control-system architecture for a particle accelerator complex defines lifecycle management around researching, developing, testing, deploying, monitoring, and improving ML models specifically within operational systems, built on a prerequisite of high-quality, well-understood control-system data. Industry-specific guidance reinforces this: research on MLOps in energy and materials companies finds that MLOps establishes key practices across the full application life cycle — data management, model development and deployment, and live model operations — that increase productivity, speed, and reliability while reducing risk when properly supported end-to-end. AIOTI's own prior work anticipates the trust dimension of this lifecycle directly: its *AI in Energy* trustworthiness practice embeds regulatory and standardisation checks across each stage of this lifecycle, mitigating the risks of silent "update opacity" when models are retrained.

For an Energy AI Factory, lifecycle management must therefore combine standard MLOps rigor (versioning, drift monitoring, CI/CD) with control-system-grade validation gates before any model is permitted to influence a live control loop.

4.4 Security, Privacy, Data Sovereignty

Energy infrastructure is critical infrastructure, so an Energy AI Factory inherits both general AI-security practice and sector-specific cybersecurity law. The relevant EU legal instrument is the NIS2 Directive (Directive (EU) 2022/2555), which establishes a unified legal framework for cybersecurity across 18 critical sectors, explicitly including energy (electricity, district heating and cooling, gas, oil, hydrogen), and requires essential entities to adopt risk-management measures and meet strict incident-reporting timelines. ENISA reported more than 200 cyber incidents in the Energy sector in 2023 alone, and NIS2 raises the compliance bar specifically because the sector remains a high-value target [37]. Recent academic work shows AI itself can

© AIOTI. All rights reserved.

be part of the compliance answer: a 2026 study examining NIS2 compliance in the Energy sector using AI-driven cyber threat intelligence found that AI-based anomaly detection and enrichment tools can support several key NIS2-related operational tasks in energy OT environments, identifying unusual network and host telemetry, enriching findings with asset and event relationships, and producing structured, reportable alert objects, provided human oversight and transparency of analysis remain part of the AI-supported workflow [38]. On sovereignty specifically, this reinforces the broader AI Factories rationale already established in Section 3.1: **AI Factories are framed as critical infrastructure akin to energy grids or transport networks**, intended to ensure the EU can innovate without depending on the strategic decisions of other global powers, a framing that applies with particular directness when the AI Factory in question serves the energy grid itself. On privacy, the GDPR-compliance mechanisms already discussed in Section 4.1 (EDC Connectors, data-sovereignty-preserving dataspace architecture) provide the practical toolset for keeping prosumer and household-level energy data compliant even as it flows through shared AI Factory infrastructure.

4.5 Interoperability and Standards

Standards are the connective tissue between the dataspace, orchestration, and compliance layers. ENTSO-E's own technology roadmap points to a dedicated data-exchange standard for the sector: the "European (energy) data exchange reference architecture 3.0" (DERA 3.0), which provides recommendations for implementing a common data exchange architecture aligned with other sector standardisation processes. This sits alongside the IDS/FIWARE-based dataspace connector approach already used by GREEN.DAT.AI and recommended more broadly for energy dataspaces, and alongside AIOTI's general-purpose computing continuum reference architecture. ENTSO-E's consultation response goes further and specifies the design principle these standards should serve: a federated architecture connecting TSO and DSO digital twins based on open standards, rather than proprietary or siloed platforms.

For an Energy AI Factory, the interoperability requirement is therefore not to invent new standards but to converge two already-existing, converging standards tracks, the general-purpose computing-continuum/dataspace standards (AIOTI, IDS/FIWARE) and energy-sector-specific data exchange standards (ENTSO-E's DERA 3.0) into a single reference architecture.

4.6 Regulatory and Compliance Considerations

Compliance for an Energy AI Factory spans three distinct regimes that must be satisfied simultaneously: EU AI Act risk classification, EuroHPC's own evolving regulatory framework, and sector-specific network and cybersecurity regulation. First, as analysed in Section 2.4, AI systems functioning as genuine safety components in critical electricity, gas, and heating infrastructure qualify as high-risk under Annex III, point 2 of the EU AI Act (aligned with the CER Directive (EU) 2022/2557), triggering mandatory risk management, data governance, technical documentation, and human oversight, whereas purely advisory operator-alert tools may fall outside the high-risk gateway. Second, in parallel, the EuroHPC regulatory framework underpinning AI Factories themselves has been evolving to formally accommodate this kind of sector-specific deployment: the EuroHPC regulation was amended in 2024 to include the development and operation of AI Factories, with a further 2025/2026 amendment extending this to AI gigafactories and a dedicated quantum-technologies pillar. Third, layered on top of both is NIS2's incident-reporting and risk-management regime for the Energy sector specifically alongside electricity market unbundling and network code rules.

For an Energy AI Factory, the practical implication is that compliance work — AI Act conformity assessment, NIS2 risk management, EuroHPC access-tier eligibility — needs to begin at the model-development stage inside the AI Factory, not be retrofitted once a model reaches deployment, since retrofitting compliance onto an already-trained model is far costlier than designing for it from the first training run.

© AIOTI. All rights reserved.

5 Challenges Facing the Energy Sector

Energy availability, not computing capacity, now limits AI deployment.

What does an AI-driven load mean for the Energy sector? This section sets out five challenges: rising demand and electrification, renewable integration and grid volatility, cooling and efficiency, ageing infrastructure and operational complexity, and data fragmentation across energy systems.

The rapid expansion of AI and the deployment of hyperscaler AI factories present unprecedented challenges for energy systems worldwide [39]. Energy availability, and not just computational capacity, are jointly emerging as the primary binding constraint on AI deployment today [40].

AI data centres function as a distinct load category, characterised by high power density, rapid and large-scale power transients, and specific power quality profiles [39]. These technical attributes create difficulties for long-term resource adequacy and transmission planning due to mismatched development timelines, as the speed of data centre development far outpaces traditional grid planning [40]. They also strain real-time grid balancing and introduce risks to system stability, such as voltage and frequency deviations and converter-driven instabilities [39]. The economic and environmental footprints associated with this new type of consumer require a coordinated strategy, encompassing data centre technologies, grid-enhancing solutions, and new policy frameworks [47].

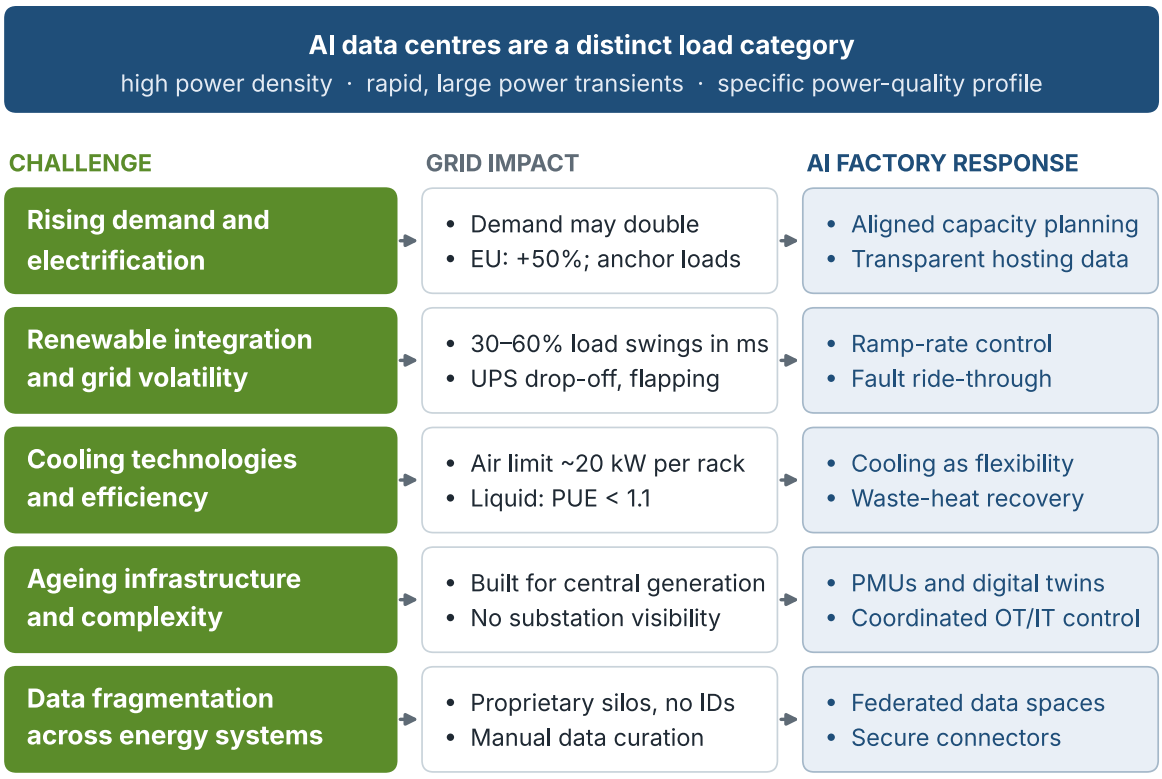


Figure 5: Energy AI factory, challenges for the Energy sector.

5.1 Rising Energy Demand and Electrification, Increasing Power Consumption

The integration of AI Factories into the European energy landscape occurs against the backdrop of a broader electrification of the economy, encompassing electric vehicles, heat pumps, and industrial processes. This convergence is driving an unprecedented rise in overall

© AIOTI. All rights reserved.

electricity demand, within which AI-optimised data centres act as systemically relevant industrial loads.

Recent global assessments indicate that electricity consumption by data centres could double by 2030, driven heavily by the rapid deployment of high-density accelerated servers required for AI [41][42]· In Europe specifically, total data centre electricity demand is expected to increase by more than 50 percent between 2025 and 2030, with growth heavily concentrated in hyperscale and colocation facilities operating at the 100–200+ kW rack densities detailed in Section 3.3.

This exponential growth in power consumption creates a dual challenge for transmission system operators. Firstly, the massive scale of individual AI Factories can exacerbate local transmission capacity constraints, forcing operators to treat these facilities as anchor loads that require dedicated planning rather than incremental additions. Secondly, the rapid development cycles of AI facilities often outpace the timeline required to upgrade grid infrastructure, which typically spans five to ten years. Consequently, securing access to power has become the most significant bottleneck for data centre operators, underscoring the urgent need for aligned capacity planning and transparent grid hosting data.

The geographic clustering of these energy intensive facilities further strains transmission corridors particularly around established metropolitan technology hubs where grid headroom is already heavily constrained. In some European regions, data centres already account for a disproportionate share of total regional electricity demand, forcing system operators to implement strict hosting capacity limits. Consequently, future planning models must integrate digital infrastructure growth as a primary variable rather than a secondary load assumption, ensuring that regional generation adequacy and transmission expansion keep pace with accelerated computational deployments.

5.2 Renewable Integration and Grid Volatility

The transition towards variable renewable energy sources, such as wind and solar, already introduces significant supply-side uncertainty for grid operators [43] . At the same time, the operational profile of AI data centres introduces unprecedented demand-side volatility. Unlike legacy facilities with steady demand curves, AI training workloads can cause aggregate facility power consumption to swing by 30 to 60 percent within milliseconds as computational resources are allocated. This rapid, software-driven load variation can excite electromechanical modes in the power system, causing load-induced forced oscillations that propagate across wide geographical areas. Furthermore, these steep and unpredictable load ramps challenge real-time balancing operations, increasing the risk of voltage and frequency deviations.

During grid disturbances, the protection mechanisms of AI data centres can inadvertently exacerbate instability. Uninterruptible power supply systems are designed to prioritise the continuity of critical IT loads. At the first sign of a voltage dip, these systems can instantaneously transition to battery mode, entirely removing hundreds of megawatts of demand from the electrical grid. This sudden load loss can transform a localised voltage incident into a system-wide frequency disturbance. When the fault clears, the uncoordinated reconnection of these massive loads can depress grid frequency back to critical thresholds, risking repeated cycles of disconnection and reconnection, a phenomenon known as flapping. Mitigating these integration risks requires grid-interactive data centre architectures that coordinate with the electrical grid through mechanisms such as ramp rate controls, oscillation damping, and advanced fault ride-through capabilities [44].

These stability challenges are exacerbated during periods of low system inertia, which occur when high shares of variable renewables displace synchronous generation. Because power electronic converters in modern data centres and their backup systems do not inherently provide rotational inertia, sudden load steps can trigger rapid frequency deviations. Establishing

© AIOTI. All rights reserved.

rigorous technical standards for fault ride-through and active power recovery is therefore essential to prevent automated load shedding from destabilising the wider pan-European transmission grid.

5.3 Cooling Technologies and Efficiency Strategies

The escalating thermal output of high-density artificial intelligence accelerators is pushing traditional air-cooling architectures to their physical and economic limits. As noted in Section 3.3, where cooling overhead ranges from roughly 7% in optimised hyperscale sites to over 30% in legacy facilities, rack power densities exceeding the ~20 kW air-cooling ceiling and rising power densities are forcing the entire sector to fundamentally rethink thermal management.

To address these challenges, operators are rapidly transitioning towards advanced liquid cooling solutions, including direct-to-chip and immersion cooling. These technologies offer significantly higher heat transfer efficiency compared to conventional air conditioning, allowing facilities to maintain optimal operating temperatures for tightly packed server arrays without thermal throttling [45]. By replacing energy-intensive air handlers with liquid loops, advanced designs can achieve Power Usage Effectiveness values below 1.1, with some leading hyperscalers reporting values below 1.05. Furthermore, moving to liquid cooling can cut direct water use by 70 to 90 percent relative to traditional evaporative cooling, which is a critical sustainability metric.

Despite these efficiency gains, liquid cooling does not eliminate the underlying electricity demand, which remains a dominant environmental cost. However, the thermal inertia inherent in cooling systems - where buildings and coolant circuits retain cold for minutes to hours - presents a unique opportunity for grid operators. Cooling loads can be temporarily reduced or shifted without immediately affecting IT operations, providing short-duration flexibility to the grid. Additionally, the elevated return temperatures of liquid coolants are well suited for waste heat recovery applications, such as district heating networks or industrial processes[2]. Integrating these advanced cooling strategies addresses the immediate thermal challenges of AI infrastructure while aligning data centre operations with broader European energy efficiency and sustainability targets.

To maximize the benefits of these thermal innovations, modern energy AI factories are increasingly adopting closed-loop liquid architectures equipped with automated flow control. These systems dynamically adjust coolant distribution based on real-time computational workloads and ambient weather conditions. By pairing advanced liquid loops with on-site thermal energy storage, operators can decouple cooling energy consumption from immediate peak grid pricing periods, transforming thermal management systems into active participants in demand-side flexibility programs.

5.4 Infrastructure Aging and Operational Complexity

The rapid integration of high-density artificial intelligence infrastructure into the Energy sector intersects with a compounding operational challenge, which is the ageing state of existing electrical grids. Much of Europe's transmission and distribution infrastructure was designed decades ago for centralised, predictable generation models rather than bidirectional, highly dynamic loads. The physical and operational complexity of manageing these legacy assets is amplified by the scale of modern AI factories, which draw power at levels historically associated with heavy industrial manufacturing. Furthermore, transmission system operators face significant

© AIOTI. All rights reserved.

difficulties because local grid capacity is frequently strained by the high geographic concentration of these facilities [46].

This operational complexity is further compounded by the convergence of operational technology and digital information technology environments within modern energy systems. Data centres rely on complex internal power distribution chains, including main switchgear, uninterruptible power supply systems, and automated transfer switches, all of which must coordinate seamlessly to maintain high reliability. When ageing grid components experience stress from rapid load fluctuations or external grid faults, the risk of cascading operational failures increases significantly. Addressing this growing complexity requires a coordinated approach that involves upgrading physical substation assets, deploying advanced monitoring tools such as high-resolution measurement units, and implementing sophisticated control systems to safely manage the interaction between legacy grid infrastructure and modern, software-driven energy loads.

Moreover, the lack of real-time visibility into local substation loading compounds the difficulty of integrating massive industrial loads onto ageing distribution networks. Transmission system operators frequently lack the high-resolution telemetry needed to monitor fast-acting power transients generated by accelerated computing clusters. Modernising substation infrastructure through the widespread deployment of synchronised phasor measurement units and advanced digital twin simulations is vital to bridge the gap between legacy grid assets and the rigorous demands of next-generation AI infrastructure.

5.5 Data Fragmentation across Energy Systems

The effective deployment of artificial intelligence and energy AI factories is heavily constrained by widespread data fragmentation across modern power systems. Although the Energy sector generates vast quantities of operational telemetry from smart meters, sensors, supervisory control and data acquisition systems, and real-time market signals, this information frequently remains locked inside proprietary silos and incompatible formats. Operational technology environments were historically designed for local control and internal security rather than cross-domain data sharing, resulting in severe barriers to data discoverability, accessibility, and reuse.

This structural fragmentation prevents transmission system operators, distribution system operators, and industrial actors from creating the unified data pipelines required to feed advanced machine learning models and training algorithms. Compounding the FAIR discoverability bottlenecks identified in Section 2.4, many existing energy databases lack persistent identifiers and standardised metadata structures, meaning that cross-domain integration still requires extensive manual curation [48].

Overcoming these digital silos requires moving beyond isolated repositories towards standardised semantic ontologies, federated data spaces, and secure connectors that enable seamless data exchange as exemplified by the GREEN.DAT.AI dataspace governance and connector architecture mentioned in Section 4.1. Integrating these frameworks ensures that data can flow securely from edge devices through to centralised high-performance computing clusters without compromising proprietary boundaries or regulatory security.

Addressing these systemic data silos is especially urgent for the implementation of federated digital twins and cross-border energy balancing platforms. Without harmonised semantic models and automated data validation pipelines, AI trained on fragmented historical data risk producing inaccurate forecasts for grid congestion and renewable generation. Aligning energy data spaces with established European interoperability frameworks ensures the high data quality and low-latency exchange necessary to support autonomous grid operations at scale.

© AIOTI. All rights reserved.

6 Use cases and Applications

The use cases show where AI Factories add value for energy and where the full pathway from gate to factory to edge still has to be proven.

Where does an AI Factory add value for energy? This section sets out seven use cases: continuous training and deployment of energy models, digital twins and simulation, autonomous grid and asset management, predictive maintenance, energy trading and market optimisation, industrial energy efficiency, and cross-sector alignment. Each states the problem, how an AI Factory helps, an example and the open challenges, and provides open issues.

6.1 Continuous Training and Deployment of Energy Models

The following texts are contributed by BI-REX, Italian national Competence Centre for Industry 4.0 and coordinator of the EDIH BI-REX++. They build on BI-REX's pilot-line Proof-of-Concept portfolio and on IPAZIA, BI-REX's industrial HPC cluster, and propose a recurring pattern, the **AI Gate**, through which a downstream infrastructure validates use cases before handing them over to an AI Factory and brings the results back to the edge. Where the full gate-to-factory pathway has not yet been executed, this is stated explicitly.

Problem. Energy models are not trained once. Load and consumption profiles drift with seasons, production mix and tariffs; assets are added, replaced or re-parametrised; sensors fail or are recalibrated. A model that performed well at commissioning silently degrades unless it is monitored and retrained, and in an operational context every retraining cycle must pass the same validation gates as the original release. For industrial operators, and for the SMEs that both consume and supply energy equipment, the bottleneck is rarely the algorithm: it is the lack of a repeatable pipeline linking field data, compute for retraining and controlled redeployment at the edge, which is a main reason why deployments stall at pilot stage.

How an AI Factory helps. An AI Factory provides the scale component of this loop: GPU capacity for periodic retraining on large multi-site datasets, MLOps tooling adapted to HPC schedulers (experiment tracking and model and artefact versioning, as offered in the IT4LIA service portfolio), federated training where data cannot leave its owner, and sandboxed trust evaluation before release. What an AI Factory cannot do on its own is decide which models are mature enough to deserve that capacity, nor operate the last mile into OT environments. We therefore propose an **AI Gate pattern** in three stages: (i) **Gate**: a downstream infrastructure (Competence Centre or EDIH) checks data readiness, performs a first AI Act screening and validates the use case in a test-before-invest setting on mid-scale, on-premise compute; (ii) **Factory**: only validated pipelines are promoted to AI Factory resources for large-scale (re)training; (iii) **Edge**: the retrained model returns to the operational environment, where drift monitoring generates the trigger for the next cycle.

Example. BI-REX operates IPAZIA [50], an industrial HPC cluster at its Bologna pilot line with CPU nodes and NVIDIA L40, H100 and H200 GPU nodes, Slurm scheduling, an open-standard software stack, data residency in Italy and NIS2-aligned access practices. Because IPAZIA uses the same scheduler and open toolchains as EuroHPC systems, a pipeline validated there can be moved to an AI Factory as a change of scale rather than of platform. The individual stages have already been exercised on the pilot line [51]: an unsupervised model trained on the power-absorption time series of a metal powder-bed-fusion machine detects energy anomalies in near real time, with inference granularity adapted to the sensors available; a non-invasive kit tracks the energy absorption of legacy machines at the edge; and a 5G edge-cloud architecture separates model training in the cloud from low-latency inference on a private 5G edge node. BI-REX has been formally recognised as AI Gate of the Italian AI Factory IT4LIA; the collaboration with

© AIOTI. All rights reserved.

CINECA, the IT4LIA coordinator, foresees that validated use cases and industrial datasets move upstream for compute-intensive processing, while IT4LIA redirects to BI-REX the companies needing data readiness, regulatory guidance or test-before-invest. The complete gate-factory-edge retraining loop for energy models has not yet been executed end-to-end; it is proposed here as a reference use case.

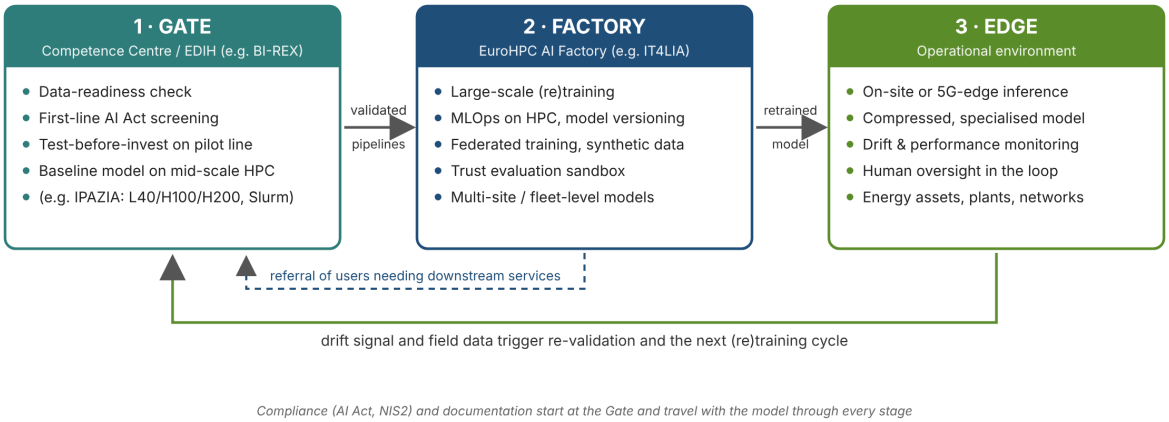


Figure 6: The AI Gate pattern: validation at the gate, scale-up in the AI Factory, deployment and monitoring at the edge, with drift-triggered retraining. Source: BI-REX.

Open challenges. (a) Cadence: AI Factory access is organised in calls and fixed allocations, whereas drift-driven retraining is continuous; a standing, lightweight allocation for validated pipelines would be needed. (b) Portability: containers, data movement and identity management across gate, factory and edge are not yet harmonised. (c) Regulation: it is unclear when retraining amounts to a substantial modification under the AI Act, triggering renewed conformity assessment. (d) Sustainability: AI Gates have no dedicated funding or engagement model, which limits their ability to run this loop beyond pilots.

6.2 Energy System Digital Twins and Simulations

Problem. Digital twins of energy assets and networks depend on two ingredients that pull in opposite directions: high-fidelity physics (CFD, finite-element, power-flow or hydraulic models) that is too slow for operational use, and real-time data streams that require fast inference. Both the design of energy components (turbine blades, heat exchangers, power-electronics cooling) and the operation of networked infrastructures depend on bridging this gap, typically through surrogate or physics-informed AI models trained on large volumes of simulation data.

How an AI Factory helps. Generating the training data for a surrogate means running hundreds to thousands of parametric simulations, which is an HPC workload; training physics-informed or graph-based surrogates on those data is an AI workload. An AI Factory can host both, together with simulation data lakes and digital-twin engineering services (IT4LIA lists surrogate modelling, physics-informed world models and digital twins among its manufacturing-domain services). The resulting compact surrogate is light enough to run next to the asset or in the operator's control room.

Example. Using its HPC resources, BI-REX delivers high-fidelity simulation as a test-before-invest service: full FEM homogenisation of a woven carbon-fibre composite in Code_Aster (six load cases, extracting the nine independent elastic constants from the weave geometry alone) and Large Eddy Simulation of a motorcycle and rider in a wind tunnel, resolving wake and vortex structures. Another PoC combines CFD optimisation and generative design to embed heat sinks directly in printed circuit boards through additive manufacturing of electronics, lowering the junction temperature of hotspot components: a thermal-management problem shared by

© AIOTI. All rights reserved.

power electronics and AI hardware. Energy hardware is already on the pilot line: large-format 3D-printed wind-turbine components were tested experimentally against design calculations, and a turbine-blade prototype was produced by Directed Energy Deposition in high-entropy alloys. On the operational side, BI-REX has prototyped twins whose interface is identical for simulated and real assets (a fleet twin of autonomous mobile robots, software-in-the-loop virtual commissioning) and, with partners, a decision-support system for water utilities based on network digital twins, anomaly detection and optimisation, an approach directly transferable to distribution grids. In the AI Gate pattern, single high-fidelity runs and pilot surrogates are validated on IPAZIA; parametric campaigns and large-scale surrogate training are the natural candidates for promotion to the AI Factory.

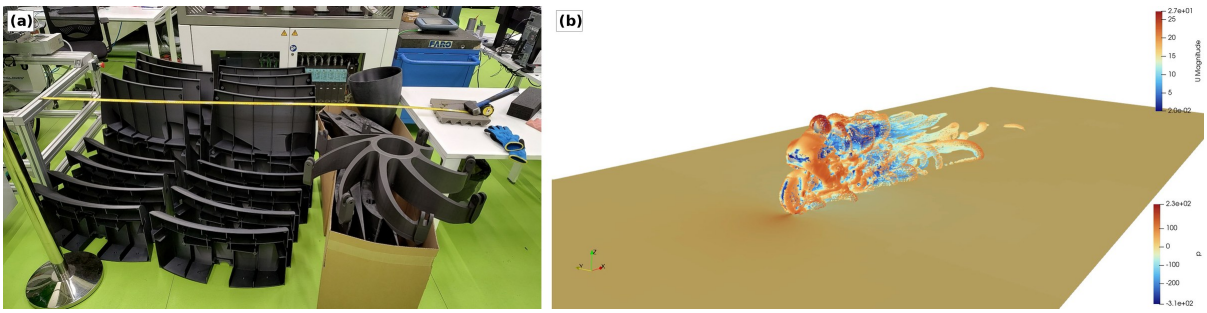


Figure 7: (a) Large-format 3D-printed wind-turbine components prepared for experimental testing against design calculations; (b) Large Eddy Simulation of a motorcycle and rider in a wind tunnel run on HPC. Source: BI-REX PoC catalogue.

Open challenges. Surrogate accuracy outside the training envelope must be quantified before twin outputs inform operations; simulation datasets reach terabyte scale and need FAIR curation; commercial solver licences on shared HPC are often restrictive; and closing the sim-to-real gap requires continuous calibration against field data, which links this use case to 5.1. Interoperability between industrial twins and the federated TSO-DSO twins called for by ENTSO-E (Section 2.1) is still an open design question.

TWIN-REV Application

TwinREV, the digital platform developed within the CRETE VALLEY project by the National Technical University of Athens, for planning and operating Renewable Energy Valleys with high renewable energy penetration. TwinREV integrates AI-based forecasting of photovoltaic and wind generation, alongside broader renewable generation and electricity demand forecasting, with power-flow simulations and flexibility management tools. Its Transformer-based models provide day-ahead and intraday forecasts to support energy communities, renewable energy managers and grid operators in anticipating variations in supply and demand and coordinating flexible resources. TwinREV has also been included as an artifact in PHAROS, the Greek AI Factory, providing a concrete link between a European energy demonstration project and the AI Factory ecosystem. This creates opportunities to further develop, evaluate and scale renewable forecasting services through PHAROS's AI expertise and computing infrastructure, supporting their adaptation to additional sites and energy communities.

6.3 Autonomous Grid and Asset Management

Problem. Grid and generation assets are dispersed, often in hazardous or hard-to-reach places (substations, tunnels, culverts, hydraulic works) and are still largely inspected by hand at fixed intervals. Autonomy has two faces here: robots that inspect and monitor assets, and control systems that act on them. Both depend on robust perception and reliable wireless connectivity in industrial environments; anything that acts on the grid also inherits the safety and human-oversight obligations of high-risk AI.

© AIOTI. All rights reserved.

How an AI Factory helps. Perception models for inspection (3D point clouds, thermal and visual imagery, acoustic signatures) and control policies for robots and agents are increasingly trained in simulation and refined on multi-site data, which requires GPU-intensive training and large simulation campaigns. The AI Factory supplies this capacity, together with sandboxes to stress-test robustness before field release. The downstream gate supplies what the AI Factory lacks: real assets, real radio conditions and the operational context in which autonomy is validated.

Example. BI-REX has deployed an agile quadruped robot with LiDAR to survey the historic underground canals of Bologna, producing point clouds of confined infrastructure; the same platform, with an extensible mission system, has been used with a telecom operator for autonomous intrusion detection, a pattern directly relevant to the perimeter security of substations and generation sites. On the control side, a PoC with an automation vendor connects a PLC to remote I/O modules over a private 5G standalone network using layer-2 tunnelling (GRETAP, WireGuard), so that controller and actuators behave as if wired; this removes cabling across large sites for soft-real-time applications and lets one central controller drive distributed actuators. A digital twin of an AMR fleet predicts congestion and triggers rerouting, virtual traffic lights and mission handover: a small-scale analogue of autonomous coordination among distributed resources. These capabilities were validated on BI-REX's pilot line and partner sites, not on grid assets. In the AI Gate pattern, the transfer would run as follows: field data and simulation scenarios collected downstream, perception and policy training at scale in the AI Factory, validation back on the gate's test infrastructure before any operational use.

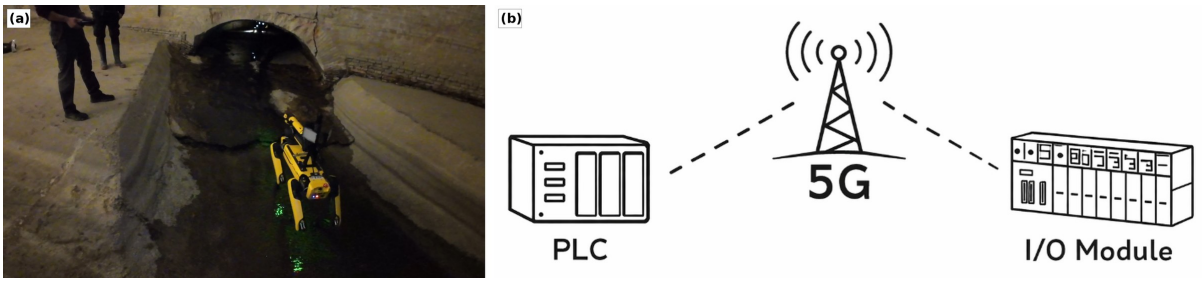


Figure 8: (a) Agile quadruped robot with LiDAR surveying the underground canals of Bologna; (b) PLC and remote I/O module connected over a private 5G standalone network. Source: BI-REX PoC catalogue.

Open challenges. Soft-real-time 5G control is not protection-grade: millisecond-level grid functions still require deterministic links and local intelligence. Systems that autonomously adjust grid operations fall squarely within the AI Act high-risk category, so human oversight and documentation must be designed from the first prototype. Transfer from factories to substations (electromagnetic environment, safety rules, site access) is non-trivial, and inspection data from critical infrastructure are subject to NIS2 and security constraints that limit their use on shared compute.

6.4 Predictive Maintenance for Energy Assets

Problem. Unplanned failures of rotating machinery, power electronics and thermal equipment are a major cost for generators, network operators and energy-intensive industry, and ageing infrastructure makes the problem worse. Two obstacles recur: failures are rare, so labelled data are scarce; and each operator sees only its own fleet, so models are trained on too little variety to generalise.

How an AI Factory helps. An AI Factory can pre-train time-series and acoustic foundation models on pooled, multi-asset data, generate synthetic fault data to compensate for rare events, and run federated training so that operators benefit from each other's failure history without exchanging raw data. The resulting models are then specialised and compressed for edge

© AIOTI. All rights reserved.

deployment. As noted in Section 2.4, a predictive-maintenance tool that only alerts a human operator may fall outside the AI Act high-risk category, which makes this a pragmatic first use case for an Energy AI Factory.

Example. BI-REX's pilot-line portfolio contains several building blocks validated on real machines: an acoustic sensor and ML model detecting anomalies during CNC machining and alerting operators remotely; a neural network locating defects and estimating their severity during metal powder-bed fusion, so that compromised builds are stopped early; the unsupervised energy-signature anomaly detection described in 6.1; and multi-protocol data collection from heterogeneous machines that builds the historical record maintenance models need. Audio classification models compact enough to run on low-power microphones have also been developed for adaptive street lighting, showing that inference can run on very constrained edge devices. The pilot line also covers the step after the prediction: a turbine-blade component produced by Directed Energy Deposition demonstrates a process that can also repair damaged parts or add features to existing components. In the AI Gate pattern these single-machine models are validated on IPAZIA, which lists predictive maintenance among its target workloads; moving to fleet-scale, cross-operator models is the natural candidate for an AI Factory.

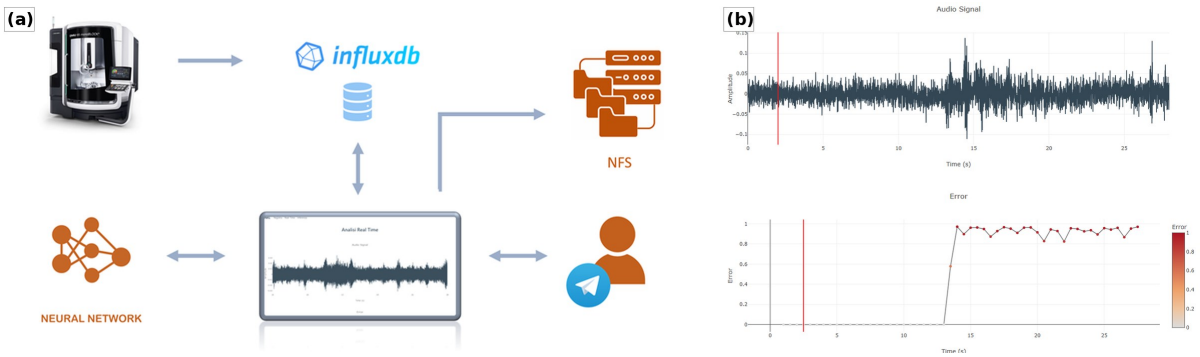


Figure 9: – (a) Acoustic anomaly-detection pipeline on a CNC machine: time-series storage, neural network, real-time dashboard and remote operator alert; (b) audio signal and anomaly score. Source: BI-REX PoC catalogue.

Open challenges. Transferring from manufacturing machines to energy assets (turbines, pumps, transformers) requires domain data that must come from energy partners; operators are reluctant to share failure data with competitors, so federated and synthetic-data approaches must be shown to be trustworthy; explanations must be meaningful to maintenance crews; and edge devices impose strict memory and power budgets on compressed models.

6.5 Energy Trading and Market Optimization

The energy market has a broad spectrum, spanning from baseload, spot and the emerging demand response – frequency balancing market. For example, in Finland, Fingrid oversees the Finnish frequency balancing market, which at the time of the writing of this paper, consists of 5 different markets, each with their own requirements, price levels, and also market behaviours: 1) *FFR (Fast Frequency Reserve)* where energy resources are activated during major frequency deviation particularly when the grid has low inertia within one second or less, 2) *Frequency Containment Reserve for disturbances (FCR-D)* where energy resources are activated within seconds and 3) Frequency Containment Reserve for normal operation where energy resources are activated within a few minutes, 4) Automatic Frequency Restoration Reserve where activations of resources are within five minutes, and 5) Manual Frequency Restoration Reserve where activations of resources need to be fulfilled within 15 minutes.

Energy resource owners are presented with opportunities to exploit these markets and obtain incentives, however predicting ahead of the time the price level and optimally schedule

© AIOTI. All rights reserved.

participations of energy resources in the appropriate energy markets, taking into account the primary usage of the energy resources is challenging task. AI could be used to address this challenge where, 1) AI can be used to predict the price level of energy market based on various parameters such as historical price, weather, grid frequency, 2) AI can be used to optimize the provisions of energy resources to appropriate markets.

The Predictricity [52] project was led by Aalto University, carried out between 2019 to 2022 which investigates and develops AI-based solutions that attempt to predict the frequency balancing markets and enable optimal participations of energy resources in order to leverage optimal incentives. One example is that reinforcement learning based AI agent is used to optimally allocate and schedule energy resources based on the predicted price level. The project involved the participation of several Finnish companies, such as Siemens Energy, St1, Tietoevry, and others, with a notable pilot of one of the major shopping centres in Finland located in the Helsinki region. The project validated and demonstrated the feasibility and opportunities of using AI for trading energy resources in emerging electricity markets and optimize participations.



Figure 10: Reserve marketplaces in Finland overseen by Fingrid (FFR, FCR-D, FCR-N, aFRR, mFRR). Source: Fingrid / Predictricity project.

6.6 Industrial Energy Efficiency

Problem. Manufacturing SMEs rarely know where their energy goes: meters sit at plant level, legacy machines have no digital interface, and thermal and electrical consumption are recorded in separate systems, if at all. Meanwhile customers and lenders increasingly ask for product-level carbon footprints and ESG reporting. Efficiency opportunities exist, but they stay invisible without the sub-metering, data integration and analytics that most SMEs cannot build alone.

How an AI Factory helps. Taken one plant at a time, most efficiency models are small and do not need supercomputing. The value of an AI Factory appears at aggregation level: models for load disaggregation, consumption benchmarking and anomaly detection trained across hundreds of plants and machine types; language models that turn heterogeneous energy and production data into ESG reports; and, through a gate, a channel that brings many SMEs to those models without each company negotiating compute access. Here the AI Gate is not only a filter but the realistic route from AI Factory capacity to small industrial users.

© AIOTI. All rights reserved.

Example. As Competence Centre and EDIH coordinator, BI-REX has supported the digital and AI transition of more than 500 SMEs under the Italian Recovery and Resilience Plan, around 40% of which asked for help on data and lifecycle management, including AI model development [53]. Energy efficiency recurs across its Proof-of-Concept work: (i) non-invasive tracking of energy absorption on legacy production lines with external sensors and an edge interface, enabling near-real-time costing as a low-cost alternative to revamping; (ii) for a manufacturer, an audit mapping thermal and electrical energy vectors into an ESG-oriented data model, automated meter acquisition, and algorithms for early detection of waste and consumption anomalies, with real-time KPI dashboards; (iii) for a machine builder, measurement of the use-phase energy consumption and carbon footprint of automatic machines, benchmarked against Best Available Techniques to guide eco-design; (iv) dashboards computing CO_2-equivalent per part sold from ERP and consumption data; (v) for business hotels, a system correlating energy use with each overnight stay and generating personalised ESG reports with generative AI. These PoCs were delivered on BI-REX infrastructure. The AI Factory step, pooling anonymised or synthetic consumption data across companies to train shared models, is proposed here and builds on a BI-REX PoC that generates statistically faithful synthetic time series while measuring re-identification risk.

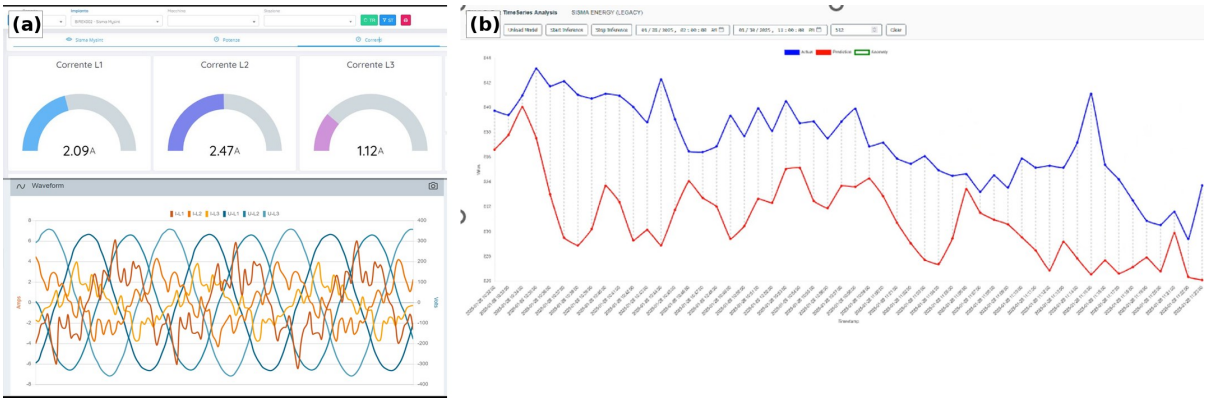


Figure 11: (a) Edge interface of the non-invasive energy-tracking kit (phase currents and waveforms) on a legacy machine; (b) measured vs. predicted power absorption used for unsupervised anomaly detection. Source: BI-REX PoC catalogue.

Open challenges. Sub-metering hardware remains the first cost barrier; data from different plants are not comparable without common semantics (Section 4.5); SMEs fear that energy data reveal production volumes and margins, so data-sharing incentives and privacy guarantees are needed; and the energy footprint of the AI pipeline itself must stay well below the savings it enables.

6.7 Cross-sector Alignment

Problem. Energy, digital infrastructure and industry are converging into a single planning problem (Section 1.1), yet no AI Factory in the current network is dedicated to energy (Section 2.2). Energy-relevant AI therefore has to be built across sectors: manufacturers are both large energy consumers and suppliers of energy equipment, and their data, models and test facilities sit in industrial ecosystems that AI Factories do not reach directly. The European Commission positions AI Factories as upstream and EDIHs as downstream infrastructures for adoption [54]; how the two are wired together in practice is still open.

How an AI Factory helps. A general-purpose AI Factory can serve energy use cases if it is fed by sector-aware intermediaries that pre-qualify demand, prepare data and carry results back into operations. Cross-sector alignment then becomes an architectural property: the same factory

© AIOTI. All rights reserved.

serves energy, manufacturing and climate through gates that share methods (data readiness, AI Act screening, test-before-invest) while bringing different domain knowledge.

Example. BI-REX combines the roles such a gate needs: national Competence Centre for Industry 4.0, coordinator of the EDIH BI-REX++, participant in EuroCC, partner of Bologna's House of Emerging Technologies, and operator of the IPAZIA HPC cluster and a manufacturing pilot line. BI-REX has been formally recognised as AI Gate of the IT4LIA AI Factory, coordinated by CINECA. In this role it performs data-readiness checks, first-line AI Act screening and test-before-invest validation, transfers only mature projects and industrial datasets upstream, and receives from IT4LIA the companies that need downstream services; BI-REX and CINECA foresee joint KPIs such as SMEs supported, referral volume, onboarding time and pilot-to-deployment rate. IPAZIA is the technical bridge: a project validated there already runs on a Slurm-based, open-standard stack, so promotion to EuroHPC systems is a change of scale rather than of platform. Bologna's Tecnopolo, which hosts CINECA's Leonardo supercomputer and the ECMWF data centre, shows the value of co-location: the IFAB-led W4E 2.0 project with CINECA, FBK and Eni aims to turn AI-based renewable generation forecasting into an operational service across Italian and European market zones [55]. Cross-cutting enablers already tested by BI-REX (synthetic data for privacy-preserving sharing, quantum key distribution for secure data transfer, private 5G for field data collection) are the components needed to move industrial and energy data safely between gate, factory and edge.

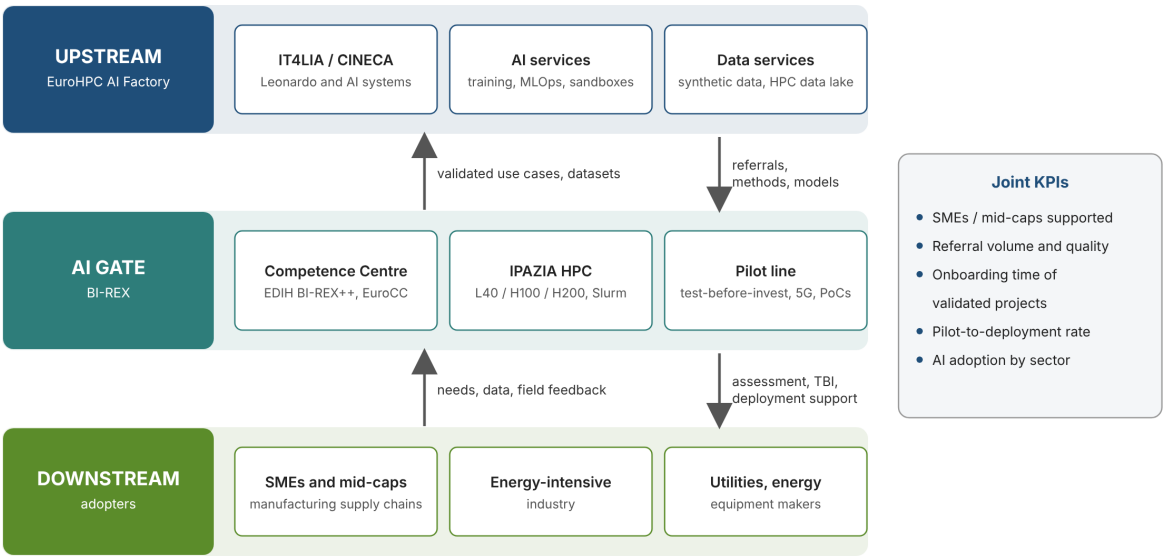


Figure 12: Roles in the AI Gate model: IT4LIA as upstream AI Factory, BI-REX as AI Gate, downstream adopters, and the joint KPIs foreseen by BI-REX and CINECA. Source: BI-REX.

Open challenges. AI Gates lack a structural funding and engagement model; AI Factory access tiers are defined by user category, which complicates mixed public-private consortia; IP and data ownership across gate and factory must be settled before industrial datasets flow upstream; and, without an energy-dedicated AI Factory, energy specifics (network codes, critical-infrastructure security) must be carried by the gates themselves, which calls for coordination among EDIHs, TEFs and AI Factories to avoid duplication.

© AIOTI. All rights reserved.

7 Unified Vision, Architecture of an Energy AI Factory

The Energy AI Factory is a three-tier pattern of grid edge, AI Gate and AI Factory, joined in one loop and co-designed with the grid.

How do the pieces fit together? This section describes the reference architecture (Figure 13), a three-tier pattern of grid edge, AI Gate and AI Factory connected in one loop. It then covers six aspects of how it works: data ingestion, high-performance AI infrastructure, model development and lifecycle management, real-time inference at the grid edge, alignment with energy regulations, and sustainability.

The preceding sections show that the digital transformation of the power system and the growth of AI-optimised compute are converging into a single planning problem (Section 2.1). Nineteen AI Factories and thirteen Antennas have been selected across Europe, yet none is dedicated to energy (Section 3.2), even though ENTSO-E has called for AI factories and secure computing clusters for energy-relevant model training (Section 3.1). An Energy AI Factory therefore cannot be conceived as a larger cloud data centre. It must reconcile the large, volatile load of accelerated computing (Sections 3.3 and 5.1–5.3) with the reliability, latency, and regulatory demands of operational technology (OT) environments.

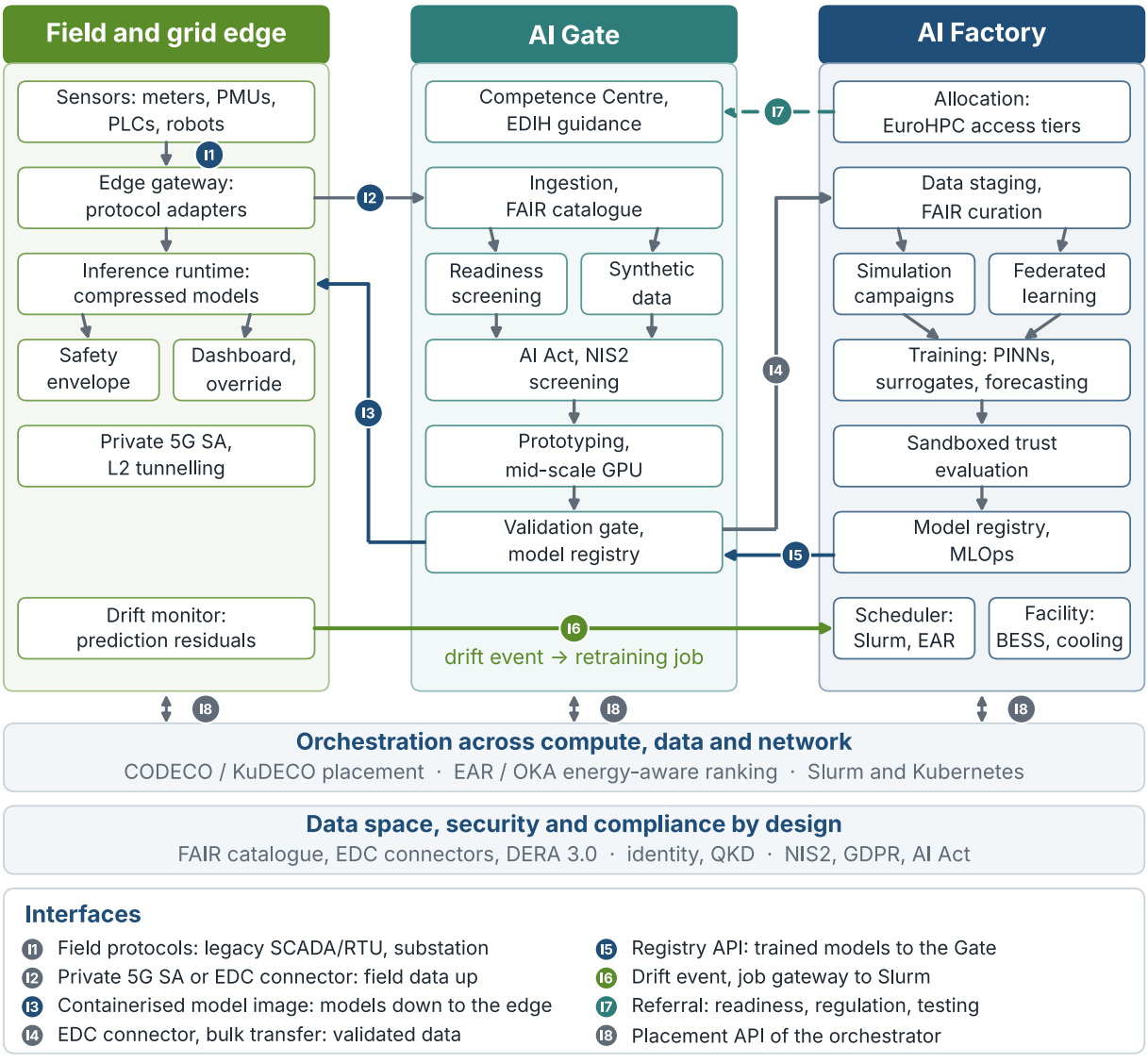


Figure 13: proposed software architecture pattern for an Energy AI Factory.

© AIOTI. All rights reserved.

This section combines the core capabilities of Section 4 (Figure 4), the grid constraints of Section 5 and the use cases of Section 6 into a reference architecture pattern provided in Figure 13, with an example on what the grid gets in Figure 14. As defined in Section 1, it describes a usage pattern for existing AI Factories, not a new facility. The architecture is a federated system of three tiers across the CEI–HPC continuum:

1. **Field and grid edge tier.** Smart meters, phasor measurement units (PMUs), intelligent electronic devices (IEDs), industrial PLCs, and autonomous inspection platforms. This tier performs local inference and protection, from milliseconds to seconds.
2. **AI Gate tier.** Downstream, sector-aware infrastructures such as Competence Centres, European Digital Innovation Hubs (EDIHs) and Testing and Experimentation Facilities (TEFs), exemplified by BI-REX and its IPAZIA cluster. The Gate checks data readiness, generates privacy-preserving synthetic data, performs first-line AI Act and NIS2 screening, and prototypes on EuroHPC-compatible stacks in a test-before-invest setting (Sections 6.1 and 6.7).
3. **AI Factory tier.** EuroHPC AI Factories such as IT4LIA/CINECA or PHAROS. They provide multi-node GPU capacity for pre-training, parametric simulation campaigns, federated cross-operator learning and sandboxed conformity evaluation.

Data move upwards only after the Gate has validated them. Trained and compressed models move downwards. Drift signals from the edge trigger the next retraining cycle, and AI Factories refer users who need data readiness, regulatory guidance or test-before-invest back to the Gate. Individual elements have been exercised at pilot-line scale (Section 6). The complete loop for an energy model has not yet been executed end to end, and it is proposed here as a reference use case. The six subsections next describe how the architecture handles ingestion, compute, model lifecycle, edge inference, regulation and sustainability.

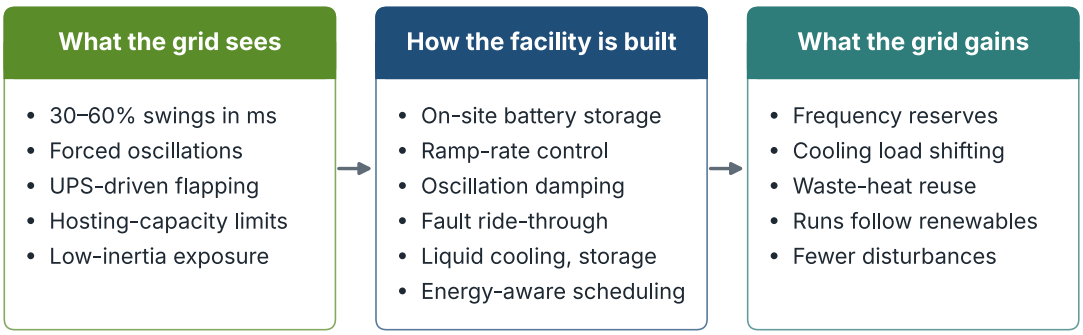


Figure 14: what the grid sees from an Energy AI Factory (30–60% swings, forced oscillations, UPS-driven flapping), how the facility is built to handle it, and what the grid gains (frequency reserves, cooling load shifting, waste-heat reuse).

7.1 Data Ingestion from the Grid and Industrial Systems

The ingestion layer must bridge heterogeneous and geographically uneven sources without physically centralising sensitive operational or commercial data. On the grid side, data originate from smart meters (more than 195 million in the EU27+3 by the end of 2024, with penetration that varies widely across Member States; Section 2.2), substation SCADA telemetry, and PMUs that capture fast transients on ageing distribution networks (Section 5.4). On the industrial side, SMEs rarely have machine-level metering, so ingestion relies on non-invasive sensor kits and multi-protocol edge interfaces on legacy machines. These capture phase currents, power waveforms, acoustic emissions, and thermal data, complemented by LiDAR and private 5G telemetry from autonomous assets.

© AIOTI. All rights reserved.

Because energy data are fragmented and often neither findable nor reusable, raw telemetry cannot flow directly into HPC training pipelines. Ingestion should combine the two standards tracks of Section 4.5:

- energy-specific exchange architectures: ENTSO-E's DERA 3.0 and the Common European Energy Data Space.
- general-purpose dataspace and continuum frameworks: AIOTI's Computing Continuum Reference Architecture and IDS/FIWARE connectors.

As in GREEN.DAT.AI, common services enforce FAIR cataloguing at the ingestion boundary. Sovereignty-preserving connectors keep ownership with the originating TSO, DSO or enterprise and support GDPR compliance. Where confidentiality or security rules prevent sharing raw data, such as cross-operator failure histories or SME production profiles, the AI Gate applies readiness screening, quantum key distribution (QKD) for transport, or synthetic time-series generation with quantified re-identification risk before data are promoted upstream. BI-REX has tested QKD and synthetic data at pilot level.

7.2 High-performance AI Infrastructure

An Energy AI Factory depends on two kinds of infrastructure that must be designed together: the compute that trains and validates energy models, and the network that moves data, models, and control signals between the tiers of the architecture.

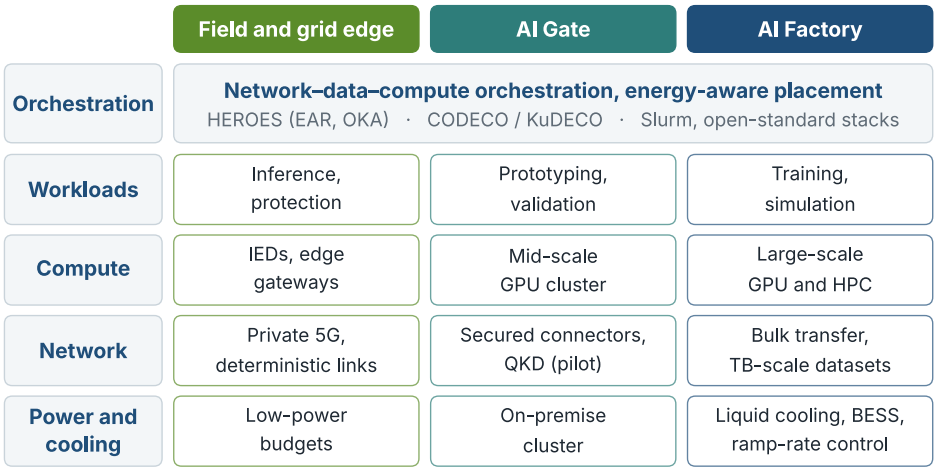


Figure 15: Energy AI Factory layered infrastructure: compute, and network together.

Compute. The compute layer supports two interdependent workload classes. The first is physics-based simulation: hundreds to thousands of parametric CFD, finite-element, hydraulic and power-flow runs that generate training data. The second is large-scale AI training: physics-informed networks, graph-based surrogates, Transformer-based forecasting such as TwinREV on PHAROS, and multi-site models for load disaggregation and anomaly. Rack power ranges from tens of kilowatts to well over 100 kW, against an air-cooling ceiling of roughly 20 kW. Direct-to-chip or immersion liquid cooling is therefore a physical necessity and not an optional efficiency measure.

Electrical interface. An Energy AI Factory that is meant to help stabilise the grid should not itself destabilise it. GPU clusters can swing by 30–60% within milliseconds, which can excite forced oscillations and trigger UPS-driven disconnection and reconnection cycles ("flapping") during voltage dips in low-inertia grids (Section 5.2). The facility must therefore be co-designed with its grid connection, using the mitigations identified in Section 5.2:

© AIOTI. All rights reserved.

- on-site battery storage between the UPS and backup generation.
- ramp-rate control.
- oscillation damping, including synthetic inertia where appropriate.
- fault ride-through capability.

Network. The network links the three tiers (Figure 13) and carries field data upwards, validated data to the AI Factory, and trained models back down. Its requirements differ by link:

- **Field to edge.** Connectivity must cope with the heterogeneous legacy base of SCADA systems, RTUs, and proprietary substation protocols. Private 5G standalone networks with layer-2 tunnelling can provide soft-real-time coordination for robots and distributed actuators, but they are not protection-grade. Millisecond-level protection functions remain on deterministic local links.
- **Edge to Gate and Factory.** Wide-area links add latency and bandwidth limits and expose the system to cybersecurity risk. Data in transit should be secured with sovereignty-preserving connectors and, where required, quantum key distribution, which BI-REX has tested at pilot level. Federated learning over constrained field links remains difficult, so training schemes should limit communication overhead.
- **Gate to AI Factory.** Parametric simulation campaigns produce datasets that reach terabyte scale. Bulk transfer, FAIR curation and identity management between the gate and the factory must be planned for. Containers, data movement, and identity are not yet harmonised across gate, factory and edge, and closing this gap is a prerequisite for the gate-to-factory pathway.

Orchestration across compute and network. Workload placement should take network and compute into account together, and energy as well, within the latency and reliability limits of each tier. Two EU-funded efforts cover complementary parts of this path. HEROES integrated Energy Aware Runtime (EAR) telemetry into the OKA decision-support platform and Slurm-based environments to rank HPC and cloud options by time, budget and energy. CODECO and its successor KuDECO provide cross-layer orchestration of computing, data and network services from the far edge to the cloud, aiming to reduce latency and power consumption. Because downstream Gates such as IPAZIA use the same Slurm-based, open-standard stack as EuroHPC systems, a validated workload moves upstream as a change of scale and not of platform. EuroHPC access tiers (Playground, Fast Lane up to 50,000 GPU hours, and Large Scale above that) should be complemented by standing, lightweight allocations for continuous operational retraining.

7.3 Model Development and Lifecycle Management

Energy models are not trained once. Load profiles drift with season, production mix and tariffs, grid topologies change as distributed resources are added, and assets degrade or are recalibrated. Retraining can also silently invalidate downstream safety assumptions in ways that black-box testing cannot detect (the "update opacity" problem described in Section 4.3). The lifecycle layer must therefore combine industrial MLOps (experiment tracking, dataset lineage, and model and artefact versioning adapted to HPC schedulers, as in the IT4LIA service portfolio) with control-system-grade validation gates before any updated model can influence an operational loop.

The loop is operationalised through the three-stage AI Gate pattern of Section 6.1 (Figure 6):

1. **Gate validation.** A downstream Competence Centre or EDIH validates the problem framing, curates data, and tests prototypes on mid-scale compute.

2. **Factory scale-up.** Mature pipelines are promoted to the AI Factory for compute-intensive training. This includes federated learning across operator fleets (Section 6.4) and physics-

© AIOTI. All rights reserved.

informed surrogate training bounded by quantified uncertainty (Section 6.2), followed by sandboxed trust evaluation.

3. **Edge deployment and drift-triggered retraining.** Models are compressed and containerised for the field. Runtime monitoring of prediction residuals and drift triggers the next cycle through the AI Factory.

7.4 Real-time Inference at the Grid Edge

Cloud and HPC infrastructure cannot execute time-critical grid functions, because of wide-area latency, bandwidth limits and cybersecurity exposure. An Energy AI Factory delivers operational value only if models trained on its supercomputers can be compressed and pushed to resource-constrained nodes at substations, generation sites, prosumer premises and factory floors. Execution is therefore stratified by latency tier, as shown in Table 2.

Table 2: Latency tiers of an Energy AI Factory. Market timescales are those of the Finnish reserve markets (Section 6.5).

Timescale	Example functions	Where it runs
Milliseconds to about 1 s	Protection and short-circuit diagnosis; fast frequency response (FFR in Finland, ≤1 s; Section 6.5)	Substation IEDs and local edge controllers, on deterministic links
Seconds to minutes	Frequency containment (FCR-D, FCR-N); machine anomaly detection; robot and actuator control over private 5G (Sections 6.1, 6.3 and 6.4)	Plant and microgrid edge gateways
5–15 minutes, intraday and day-ahead	Frequency restoration (aFRR, mFRR); reserve bidding; renewable forecasting such as TwinREV (Sections 6.2 and 6.5)	Regional edge or cloud, running models trained in the AI Factory
Hours to days (offline)	Training, parametric simulation, federated learning, retraining (Sections 6.1 and 6.2)	AI Factory (HPC)

Edge nodes must operate within OT environments that were never designed with model-serving interfaces. Compressed networks and physics surrogates should therefore run inside deterministic safety envelopes and standard industrial protocols. Private 5G can provide soft-real-time coordination for robots and distributed actuators, but millisecond-grade protection stays on deterministic local links. Edge outputs, whether an anomaly score or a congestion forecast, should come with explainable operator dashboards and human-in-the-loop override, so that local autonomy never bypasses physical protection.

7.5 Alignment with Energy Regulations

Energy sector AI faces one of the densest regulatory stacks in European law, combining horizontal digital regulation with sector-specific electricity rules. Under the EU AI Act (Regulation (EU) 2024/1689), AI systems that act as safety components in the management and operation of critical infrastructure, including electricity, gas and heating supply, are high-risk under Annex III, point 2, aligned with the Critical Entities Resilience Directive (EU) 2022/2557. The classification follows a cumulative gateway:

- decision-support and predictive-maintenance models that only alert human operators may fall outside it.
- autonomous systems that adjust grid operations trigger risk management, data governance, technical documentation, human oversight, and conformity assessment.

The NIS2 Directive (EU) 2022/2555 adds cybersecurity risk management and incident reporting across sectors that explicitly include energy (Section 4.4). Beyond these directives, an Energy AI Factory must align with:

© AIOTI. All rights reserved.

- Directive (EU) 2019/944 on smart metering, flexibility and the TSO/DSO unbundling rules that govern how data can be shared between regulated and competitive activities.
- emerging technical requirements for data-centre fault ride-through and active power recovery.
- the amended EuroHPC Regulation.
- ENTSO-E's proposal for an EU-level Digital & Data Office.

The architecture embeds compliance by design from the first training run (Section 4.6). Gates perform first-line AI Act screening and NIS2-aligned access control, and AI Factories provide sandboxed conformity evaluation. It remains unclear when drift-driven retraining amounts to a substantial modification that requires renewed conformity assessment (Section 6.1), and regulators need to clarify this.

7.6 Sustainability

Sustainability has two sides: **reducing the footprint of the AI Factory itself**, and **maximising the efficiency and decarbonisation gains its models enable elsewhere**. On the infrastructure side, data-centre electricity demand is projected to double by 2030, with European demand rising by more than 50% between 2025 and 2030 (Section 5.1). An Energy AI Factory should lead by example through four commitments:

- **Thermal management.** Closed-loop direct-to-chip or immersion cooling, targeting a PUE below about 1.1 and a 70–90% cut in direct water use compared with evaporative cooling (Sections 3.3 and 5.3).
- **Waste-heat reuse.** Using the higher return temperatures of liquid coolants for district heating and adjacent industrial processes (Section 5.3).
- **Energy-aware orchestration.** Scheduling batch training and simulation, with tools such as EAR, in line with renewable availability and carbon intensity (Sections 3.4 and 4.2).
- **Grid flexibility.** Thermal storage and on-site batteries that let the facility shift cooling loads and offer reserve services, following ASHRAE's grid-interactive design work and ENTSO-E's analysis of the dual role of data centres (Sections 5.2 and 5.3).

On the application side, the energy and carbon cost of an AI pipeline must stay well below the savings it enables. Pooling anonymised or synthetic consumption data at AI Factory level can support sub-metering of legacy machines, anomaly and waste detection, eco-design benchmarking, per-part CO_2 accounting and ESG reporting for hundreds of SMEs that could not afford dedicated supercomputing. Technical sustainability also needs institutional sustainability: permanent governance and funding for AI Gates, so that validated energy models stay maintained beyond initial grant horizons.

© AIOTI. All rights reserved.

8 Future Outlook and Recommendations

Five actions by the Commission, EuroHPC, grid operators and AI Gates can turn the pattern into working infrastructure.

Where is this heading, and what should be done? This section sets out emerging technologies and trends, the evolution of AI-enabled energy systems, five recommendations to the institutions that can act on them.

This paper has described the Energy AI Factory as a usage pattern: existing AI Factories applied to energy workloads, connected to the grid edge through AI Gates. Parts of the pattern have been exercised on pilot lines. The complete loop for an energy model, from field data through validation and large-scale training back to edge deployment and drift-triggered retraining, has not yet been run end to end. This section looks at where the technology and the energy system are heading, sets out five recommendations, and proposes next steps for the working group. Annex I turns the requirements collected throughout the paper into a checklist for AI Factories.

8.1 Emerging technologies and trends

Smaller, domain-specific models. The models that reach the grid edge are compact. The examples in this paper include Transformer-based renewable forecasting, physics-informed and graph-based surrogates, unsupervised power-signature anomaly detection, and audio classifiers small enough to run on low-power microphones. The emerging pattern is to pre-train on pooled multi-asset data in AI Factories, then specialise and compress the result for devices with tight memory and power budgets. Whether small and medium foundation models for energy can be built and distilled this way at fleet scale has still to be shown.

Agentic and autonomous AI. Reinforcement-learning agents have been demonstrated on narrow tasks, such as bidding in frequency reserve markets and coordinating robot fleets on a manufacturing pilot line. They have not yet been shown on grid assets. As autonomy increases, agents will need to operate inside deterministic safety envelopes with human override, because systems that adjust grid operations fall under the high-risk category of the EU AI Act.

Enabling technologies at pilot level. Synthetic time-series data with measured re-identification risk, quantum key distribution for data transfer, and private 5G standalone networks have been evaluated at BI-REX. They are candidate building blocks for moving data safely between edge, gate, and factory. Private 5G suits soft-real-time coordination but not protection functions. The EuroHPC quantum pillar is a development to watch, but this paper has not assessed its relevance to energy.

Execution by latency tier. Protection functions run in milliseconds at the edge, reserve and forecasting functions in seconds to minutes, and training and simulation offline in the AI Factory. Placing each workload in the right tier, with energy as one of the criteria, is a central orchestration task for the coming years.

8.2 Evolution of AI-enabled energy systems

From pilots to an operating continuum. European energy AI is still largely project-level. HEROES, GREEN.DAT.AI, TwinREV and Predictricity each demonstrate one part of the picture. ENTSO-E points to the next stage: a mandated Common European Energy Data Space, federated TSO–DSO digital twins built on open standards, and a shared reference architecture for the computing continuum.

A two-way relationship between compute and grid. Large AI facilities can no longer be treated as passive loads. Designed well, with on-site storage, ramp-rate control, fault ride-through, thermal storage and energy-aware scheduling, an AI Factory can shift non-urgent training to

© AIOTI. All rights reserved.

times of renewable surplus and offer reserve services in return. This requires connection standards that have not yet been harmonised.

AI Gigafactories. EuroHPC launched a call for up to seven AI Gigafactories on 30 July 2026, with submissions due on 12 November 2026 and selection expected in early 2027. The decision that matters for energy is whether siting, grid connection and workload design follow the requirements of Annex I from the outset, because retrofitting them later is costlier.

Renewable valleys and energy communities. TwinREV, developed in the CRETE VALLEY project, combines forecasting, power-flow simulation and flexibility tools for renewable energy valleys and energy communities. Its inclusion as an artefact in PHAROS shows how a demonstrator can become a service within an AI Factory.

8.3 Recommendations

AIOTI WG Energy formulates five recommendations, addressed to the institutions that can act on them.

1. Recognise energy as a use of the AI Factory network and add access for continuous retraining. *To the European Commission and EuroHPC.* ENTSO-E has called for AI factories and secure computing clusters for energy-relevant model training. This can be met either by a dedicated EuroHPC Energy AI Factory pattern, or by an energy vertical across existing AI Factories with energy and climate expertise, as proposed in this paper. In both cases, access should be extended. Call-based allocations suit one-off projects, whereas operational models drift continuously and need standing, lightweight allocations for retraining. User categories should also be adapted so that mixed consortia of grid operators and industry can participate.

2. Formalise and fund the AI Gate ecosystem. *To the European Commission, Member States, Competence Centres, EDIHs and TEFs.* Gates provide data readiness, first-line regulatory screening and test-before-invest validation, and they bring small and medium industrial users to AI Factory capacity. They currently lack structural funding. Co-funding, standard templates for intellectual property and data ownership, and shared indicators such as organisations supported, referral volume, onboarding time and pilot-to-deployment rate would make the model durable. Gates should also coordinate with AI Factories to avoid duplication, and until an energy-dedicated factory exists, they should carry energy-specific requirements such as network codes and critical-infrastructure security.

3. Operationalise the Common European Energy Data Space and the continuum standards. *To the European Commission, ENTSO-E, the DSO Entity, and standardisation bodies.* The data space should move beyond isolated pilots to a mandated, sustainably funded service for near-real-time electricity and industrial energy data, supported by EU-level data governance. FAIR cataloguing and sovereignty-preserving connectors, following the GREEN.DAT.AI blueprint, should become standard practice. The computing-continuum reference architecture and ENTSO-E's DERA 3.0 should be converged, and the uneven smart-meter and phasor-measurement coverage across Member States should be closed.

4. Embed compliance by design and regulatory sandboxes. *To AI Factories, regulators, and the AI Office.* Documentation, traceability, and sandboxed conformity evaluation should be part of AI Factory MLOps, so that AI Act, NIS2 and CER obligations are managed from the first training run. Regulators should clarify when drift-driven retraining counts as a substantial modification that requires a new conformity assessment, since the answer determines whether continuous retraining is practicable.

5. Plan grid and compute together, and design AI Factories to be grid interactive. *To TSOs, DSOs, regulators, and AI Factory operators.* AI Factory and data-centre growth should enter adequacy and transmission planning as a primary variable, supported by transparent hosting-capacity

© AIOTI. All rights reserved.

data. Harmonised connection requirements are needed for fault ride-through, ramp-rate limiting and oscillation damping. Operators should be encouraged to adopt liquid cooling with heat recovery, energy-aware scheduling, and participation in reserve markets.

© AIOTI. All rights reserved.

References

[1] EuroHPC Joint Undertaking, "AI Factories," 2025. [Online]. Available: https://www.eurohpc-ju.europa.eu/ai-factories_en

[2] International Energy Agency (IEA), "Key Questions on Energy and AI," 2025. [Online]. Available: https://www.iea.org/reports/key-questions-on-energy-and-ai

[3] ACER / CEER, "2025 Retail Monitoring Report – Country Sheets: Electricity," Jul. 2025. [Online]. Available: https://energiaoltre.it/wp-content/uploads/2025/07/2025-Retail-Monitoring-Report-Country-Sheets-Electricity.pdf

[4] Cheng Feng, Yi Wang, Qixin Chen, Yi Ding, Goran Strbac, Chongqing Kang, Smart grid encounters edge computing: opportunities and applications, Advances in Applied Energy, Volume 1, 2021, 100006, ISSN 2666-7924, https://doi.org/10.1016/j.adapen.2020.100006

[5] J. Li, C. Gu, Y. Xiang and F. Li, "Edge-cloud Computing Systems for Smart Grid: State-of-the-art, Architecture, and Applications," in *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 4, pp. 805-817, July 2022, doi: 10.35833/MPCE.2021.000161

[6] Yassine Himeur, Aya Nabil Sayed, Abdullah Alsalemi, Faycal Bensaali, Abbes Amira, Edge AI for Internet of Energy: Challenges and perspectives, Internet of Things, Volume 25, 2024, ISSN 2542-6605, https://doi.org/10.1016/j.iot.2023.101035

[7] AIOTI WG Standardisation, "Towards a Computing Continuum Reference Architecture," Release 20250530, May 2025. [Online]. Available: https://aioti.eu/wp-content/uploads/Towards-a-computing-continuum-reference-architecture-Final.pdf

[8] Violettas, G., & Mamatas, L. (2026). TriCloudEdge: A multi-layer Cloud Continuum. *arXiv preprint arXiv:2602.02121*.

[9] Council of Europe, Artificial Intelligence: Council paves the way for the creation of AI gigafactories. Press release, Jan 2026. Consulted in: September 2026.

[10] AIOTI WG Energy, "AIOTI Paper: AI in Energy," 2025. [Online]. Available: https://aioti.eu/aioti-paper-ai-in-energy/

[11] McKinsey & Company, "State of AI Trust in 2026: Shifting to the Agentic Era," Tech Forward Insights, 2026. [Online]: https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era

[12] EuroHPC Joint Undertaking, "AI Factories Access Modes," 2025. [Online]. Available: https://www.eurohpc-ju.europa.eu/ai-factories/ai-factories-access-modes_en

[13] European Commission, "AI Act Service Desk – Critical Infrastructure," 2025. [Online]. Available: https://ai-act-service-desk.ec.europa.eu/en/critical-infrastructure

[14] A. Volkova, M. Hatamian, A. Anapyanova and H. de Meer, "Being Accountable is Smart: Navigating the Technical and Regulatory Landscape of AI-based Services for Power Grid" in *Proceedings of the 2024 ACM Conference on Information Technology for Social Good (GoodIT'24), 4-6 September 2024, Bremen, Germany* , ACM, Sep.2024. pp. 118–126.

[15] Schwanitz, V.J., Wierling, A., Biresselioglu, M.E. *et al.* Current state and call for action to accomplish findability, accessibility, interoperability, and reusability of low carbon energy data. *Sci Rep* **12**, 5208 (2022). https://doi.org/10.1038/s41598-022-08774-0

[16] GREEN.DAT.AI Consortium, "GREEN.DAT.AI – Energy-efficient AI-ready Data Spaces," Horizon Europe Project. [Online]. Available: https://greendatai.eu/

[17] Segler Consulting, "Europe's AI Gambit: An In-Depth Analysis of the EuroHPC AI Factories and the Quest for Digital Sovereignty," 2025. [Online]. Available: https://www.seglerconsulting.com/europes-ai-gambit-an-in-depth-analysis-of-the-eurohpc-ai-factories-and-the-quest-for-digital-sovereignty

[18] DSOE, TSO-DSO Challenges & Opportunities for the Digital EU Electricity System, Feb 2025. [Online]: https://eepublicdownloads.blob.core.windows.net/public-cdn-container/clean-documents/Publications/RDC%20publications/DESAP_joint_report_on_common_TSO-DSO_challenges___opportunities_v1.1.pdf

[19] ENTSO-E, "New ENTSO-E reports highlight significant progress in European electricity market integration," Jun. 29, 2026. [Online]. Available: https://www.entsoe.eu/news/2026/06/29/new-entso-e-reports-highlight-significant-progress-in-european-electricity-market-integration/

[20] ENTSO-E, "Data centres and the power system: expected trends, challenges, and opportunities," Brussels, Belgium, May 2026. [Online]: https://eepublicdownloads.blob.core.windows.net/public-cdn-container/clean-documents/Reports/2026/ENTSO-E_Data_Centres_260520.pdf

[21] EuroHPC Joint Undertaking, "EuroHPC JU selects six additional AI Factories to expand Europe's AI capabilities," Oct. 10, 2025. [Online]. Available: https://www.eurohpc-ju.europa.eu/eurohpc-ju-selects-six-additional-ai-factories-expand-europes-ai-capabilities-2025-10-10_en

[22] EuroHPC Joint Undertaking, "EuroHPC JU selects AI Factory Antennas to broaden the AI Factories initiative," Oct. 13, 2025. [Online]. Available: https://www.eurohpc-ju.europa.eu/eurohpc-ju-selects-ai-factory-antennas-broaden-ai-factories-initiative-2025-10-13_en

© AIOTI. All rights reserved.

[23] IT4LIA AI Factory, "IT4LIA AI Factory Portal and Service Portfolio (MLOps on HPC, Data pre-processing on HPC, AI Technical Trust Evaluation & Sandbox, Manufacturing & Earth Domain Prototyping)," 2026. [Online]. Available: https://it4lia-aifactory.eu/
[24] European Commission, "HealthAI – Artificial Intelligence Factory (CESGA Galicia)," Reforms and Investments Portal. [Online]. Available: https://reforms-investments.ec.europa.eu/projects/healthai-artificial-intelligence-factory-cesga-galicia_en
[25] Czech Semiconductor Centre, "The Czech Republic has its own AI Factory, including a new AI supercomputer," 2025. [Online]. Available: https://czechsemiconductorcentre.cz/the-czech-republic-has-its-own-ai-factory-including-a-new-ai-supercomputer/
[26] EuroHPC Joint Undertaking, "HEROES – Hybrid Eco-Responsible Offerings for European Supercomputing," Research and Innovation Projects. [Online]. Available: https://www.eurohpc-ju.europa.eu/research-innovation/our-projects/heroes_en
[27] ModulEdge, "AI Data Center Architecture and Operations," Blog. [Online]. Available: https://www.moduledge.com/blog/ai-data-centre
[28] Xin Chen, Xiaoyang Wang, Ana Colacelli, Matt Lee, Le Xie, Electricity demand and grid impacts of AI data centres: Challenges and prospects, Nexus, Volume 3, Issue 3, 2026, 100162, ISSN 2950-1601, https://doi.org/10.1016/j.ynexs.2026.100162
[29] [27] IEEE ComSoc Technology Blog, "Telecom Data Centers Must Be Redesigned for the AI Era with Rack-Scale Architectures & Enhanced Power/Cooling Requirements," May 25, 2026. [Online]. Available: https://techblog.comsoc.org/2026/05/25/telecom-data-centres-must-be-redesigned-for-the-ai-era-with-rack-scale-architectures-enhanced-power-cooling-requirements/
[30] TechPlusTrends, "Power Requirements in AI Data Centers," 2025. [Online]. Available: https://techplustrends.com/power-requirements-ai-data-centres/
[31] EuroHPC Joint Undertaking, "ACROSS, HEROES, NEXTSim and OPTIMA: European High-Performance Computing Joint Undertaking launched four new research and innovation projects," Mar. 29, 2021. [Online]. Available: https://eurohpc-ju.europa.eu/across-heroes-nextsim-and-optima-european-high-performance-computing-joint-undertaking-launched-four-2021-03-29_en
[32] European Commission CORDIS, "Democratising access to high-performance computing (HEROES project results)," Article ID 449981. [Online]. Available: https://cordis.europa.eu/article/id/449981-democratising-access-to-high-performance-computing
[33] HPCWire, "Do IT Systems, HPCNow! And UCit create a Joint HPC Venture", Press Release, Nov 2020. [Online]: https://www.hpcwire.com/aiwire/2020/11/17/do-it-systems-hpcnow-and-ucit-create-a-joint-hpc-venture/
[34] Ferrario, A., & Hatherley, J. (2026). Update Opacity: Epistemic Accessibility and Governance Under AI System Change. *arXiv preprint arXiv:2606.00037*.
[35] Horizon Europe Cognitive, Decentralised Edge-Cloud Orchestration (CODECO), grant agreement 101092696. https://he-codeco.eu.
[36] Eclipse KuDECO, [online]: https://projects.eclipse.org/projects/technology.kudeco. Consulted: November 2026.
[37] ENISA (European Union Agency for Cybersecurity), "Cybersecurity of Critical Sectors," 2025. [Online]: https://www.enisa.europa.eu/topics/cybersecurity-of-critical-sectors
[38] J. Siivola, R. Paronen, U. Tariq, Q. Pham, W. Villegas, I. Tikanmaki, J. Rajamaki, Exploring NIS2 Compliance in the Energy Sector Using AI-Driven Cyber Threat Intelligence, Vol. 21 No. 1 (2026): Proceedings of the 21st International Conference on Cyber Warfare and Security (ICCWS 2026). [Online]: https://papers.academic-conferences.org/index.php/iccws/article/view/4482
[39] E. Ginzburg-Ganz, P. Lifshits, R. Machlev, J. Belikov, Z. Krieger, and Y. Levron, "Technical Challenges of AI Data Center Integration into Power Grids—A Survey," Energies, vol. 19, no. 1, p. 137, Jan. 2026, doi: 10.3390/en19010137 (Footnote 35)
[40] L. Lo Schiavo, "AI Data Centers and Energy Regulation: When Traditional Frameworks Meet Unprecedented Demand Growth," Global Energy Law and Sustainability, vol. 6, no. 2, pp. 157–190, Aug. 2025, doi: 10.3366/gels.2025.0141 (Footnote 36)
[41] International Energy Agency (IEA), "Electricity 2024 – Analysis," Paris, 2024. [Online]. Available: https://www.iea.org/reports/electricity-2024 (Footnote 37)
[42] Electric Power Research Institute (EPRI), "Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption," Product ID 3002028905, May 2024. [Online]. Available: https://www.epri.com/research/products/000000003002028905
[43] P. Tasai, "AI data centres and grid resilience: Insights from Europe and the Middle East," Reactive Technologies, 2026. [Online]. Available: https://reactive-technologies.com/news/ai-data-centres-and-grid-resilience-insights-from-europe-and-the-middle-east/
[44] ASHRAE, "Grid-Interactive Design (Demand Flexibility) | AI Data Center Energy Performance Framework," 2025. [Online]. Available: https://www.ashrae.org/technical-resources/ai-data-centre-framework/grid-interactive-design-demand-flexibility
[45] C. Leonard, "The rise of direct-to-chip cooling as a top AI cooling system," Schneider Electric Blog, Jan. 29, 2026. [Online]. Available: https://blog.se.com/datacentre/2026/01/29/the-rise-direct-to-chip-cooling-top-ai-cooling-system/

© AIOTI. All rights reserved.

[46] Stratview Research, "Data Center Liquid Cooling Market Size & Trend | 2026–2031," 2026. [Online]. Available: https://www.stratviewresearch.com/3624/data-centre-liquid-cooling-market.html
[47] ENTSO-E, "From Energy Demand to Grid Support: ENTSO-E Explores Data Centres' Dual Role," May 8, 2026. [Online]. Available: https://www.entsoe.eu/news/2026/05/08/from-energy-demand-to-grid-support-entso-e-explores-data-centres-dual-role/ (Footnote 43, Endnote iv).
[48] V. J. Schwanitz et al., "Current state and call for action to accomplish findability, accessibility, interoperability, and reusability of low carbon energy data," Scientific Reports, vol. 12, no. 1, p. 5208, Mar. 2022, doi: 10.1038/s41598-022-08774-0
[49] European Commission CORDIS, "Energy-efficient AI-ready Data Spaces (GREEN.DAT.AI) – Horizon Europe Fact Sheet (Grant ID 101070416)," 2023. [Online]. Available: https://cordis.europa.eu/project/id/101070416
[50] BI-REX Competence Centre, "IPAZIA HPC: accessible supercomputing to accelerate innovation and industrial AI," and Technical Brochure, Mar. 2026. [Online]. Available: https://bi-rex.it/en/?p=63749 and https://bi-rex.it/wp-content/uploads/2026/03/ENG-IPAZIA-HPC.pdf
[51] BI-REX Competence Centre, "Proof-of-Concept Catalogue," 2026. [Online]. Available: https://bi-rex.it/proof-of-concept/
[52] Aalto University, "Predictricity – Enabling Distributed Resources of a Smart Grid to Participate in the Electricity Reserve Market using Machine Learning," Research Project (2019–2022). [Online]. Available: https://research.aalto.fi/en/projects/predictricity/
[53] BI-REX Competence Centre, "Operational data reported for BI-REX activities as national Competence Centre and European Digital Innovation Hub (EDIH) under PNRR Mission 1, Component 2," 2026
[54] [European Commission, "Guidelines on the European AI Innovation Ecosystem: Positioning AI Factories as Upstream and EDIHs as Downstream Infrastructures," 2025
[55] International Foundation Big Data and Artificial Intelligence for Human Development (IFAB), "ICSC Innovation Grants 2026: W4E 2.0 (Weather4Energy Extension)," in collaboration with CINECA, FBK, and Eni, 2026.
[56] National Technical University of Athens (NTUA) Decision Support Systems Laboratory (EPU-NTUA), "TwinREV – Digital Twin for Renewable Energy Valley (Crete Arvi Pilot)," 2026. [Online]. Available: https://twin-rev.epu.ntua.gr/.

© AIOTI. All rights reserved.

Contributors

Editors: Rute Sofia (fortiss GmbH/AIOTI WG Energy co-chair), Natalie Samovitch (Enercoutim, AIOTI WG Chair).

Reviewers: Valentina Peniche (AIOTI Secretary General), Vittorio Pettonati (AIOTI)

Contributors:

- Rute Sofia (fortiss GmbH)
- Natalie Samovitch (Enercoutim)
- Fabio Silva (IERC/Tyndall)
- Udayanto Dwi Atmojo (Aalto University)
- Federico Naidi (Bi-Rex)
- Elissaios Sarmas (NTUA / EPU-NTU_)
- Vittorio Pettonati (AIOTI)

Annex I: Checklist, Requirements to become an Energy AI Factory

One list shows what an AI Factory needs to serve the Energy sector well.

What would an AI Factory have to do to serve the Energy sector well? This annex collects the requirements stated throughout the paper into a checklist of nine areas: grid interaction of the facility, cooling and facility sustainability, compute, orchestration and network, data and interoperability, model lifecycle, edge deployment and the operational technology interface, security and sovereignty, regulation and compliance, and access and governance. It is a proposal of the working group, not a certification scheme.

A. Grid interaction of the facility

☐	Requirement	Section
☐	Treat the facility as an anchor load: plan siting and grid connection with the TSO/DSO, accounting for water, interconnection queues, and land	2.1, 5.1, 5.4
☐	Provide transparent load-growth and hosting-capacity data to grid operators for adequacy planning	5.1, 8
☐	Limit facility load swings (30–60% within milliseconds) through ramp-rate control and oscillation damping	3.3, 5.2, 7.2
☐	Install on-site battery storage between UPS and backup generation, sized to absorb surges	3.3, 5.2, 7.2
☐	Provide fault ride-through and active power recovery, and avoid UPS-driven disconnection and reconnection cycles (flapping) in low-inertia conditions	5.2, 7.2, 7.5
☐	Provide synthetic inertia or grid-forming battery capability where appropriate	5.2, 7.2, 8
☐	Support high-resolution measurement (e.g. PMU) of facility load transients so operators can observe them	5.4, 8
☐	Offer flexibility: shift cooling load and provide reserve services or market participation	3.1, 5.3, 7.6, 8

B. Cooling and facility sustainability

☐	Requirement	Section

© AIOTI. All rights reserved.

☐	Use direct-to-chip or immersion liquid cooling for AI-class rack densities (air ceiling about 20 kW per rack)	3.3, 5.3, 7.2
☐	Target a PUE below about 1.1	5.3, 7.6
☐	Cut direct water use by 70–90% relative to evaporative cooling	5.3, 7.6
☐	Use closed-loop cooling with automated flow control and thermal storage to decouple cooling energy from peak periods	5.3, 7.6
☐	Recover waste heat for district heating or industrial processes	5.3, 7.6
☐	Keep the energy and carbon cost of an AI pipeline well below the savings it enables, and report it	6.6, 7.6
☐	Co-locate with, or contract, renewable and firm low-carbon supply (outlook)	8

C. Compute, orchestration, and network

☐	Requirement	Section
☐	Host both parametric simulation campaigns and AI training, with simulation data lakes and digital-twin engineering services	6.2, 7.2
☐	Provide GPU capacity for periodic or continuous retraining on large multi-site datasets	6.1
☐	Serve operational digital twins continuously	3.1
☐	Treat energy as a first-class scheduling input (EAR telemetry, OKA/HEROES platform selection), aligned with renewable availability and carbon intensity	3.4, 4.2, 7.2, 7.6
☐	Orchestrate compute, data, and network together from far edge to cloud with energy in placement decisions within each tier's latency limits	4.2, 7.2
☐	Run the same scheduler and open-standard toolchain as downstream Gates so promotion is a change of scale, not of platform	6.1, 6.7, 7.2
☐	Harmonise containers, data movement and identity across edge, Gate and Factory	6.1, 7.2
☐	Support terabyte-scale bulk transfer between Gate and Factory with FAIR curation	6.2, 7.2
☐	Support deterministic networking	6.3, 7.2
☐	Secure wide-area transport with sovereignty-preserving connectors and, where required, QKD	6.7, 7.2
☐	Limit communication overhead in federated training over constrained field links	2.2, 7.2
☐	Provide licensing for commercial solvers that works on shared HPC	6.2

D. Data, interoperability, and sovereignty

☐	Requirement	Section
☐	Federate across DSOs, TSOs, aggregators and industry without physical centralisation; control stays with the originating party	4.1, 7.1
☐	Connect to the Common European Energy Data Space for near-real-time operational data	3.1, 4.1, 8
☐	Enforce FAIR cataloguing with persistent identifiers and standard metadata at the ingestion boundary	2.4, 4.1, 5.5, 7.1

© AIOTI. All rights reserved.

☐	Use sovereignty-preserving connectors (EDC, IDS/FIWARE) and ensure GDPR compliance	4.1, 4.4, 4.5, 7.1
☐	Converge DERA 3.0 with the computing-continuum and dataspace standards on open standards, avoiding siloed platforms	3.1, 4.5, 8
☐	Apply common semantic models and ontologies across plants and operators	5.5, 6.6
☐	Provide automated data-validation and readiness checks before data move upstream	5.5, 6.1, 7.1
☐	Provide synthetic-data generation with quantified re-identification risk, including synthetic fault data for rare events	6.4, 6.6, 7.1
☐	Ingest from legacy SCADA, RTUs, and proprietary substation protocols, and from non-invasive sensor kits on legacy machines	2.2, 6.6, 7.1
☐	Settle IP and data-ownership terms across Gate and Factory before industrial datasets flow upstream	6.7, 8
☐	Offer data-sharing incentives and privacy guarantees, since SMEs fear revealing volumes and margins	6.6

E. Model lifecycle

☐	Requirement	Section
☐	Provide MLOps adapted to HPC schedulers: experiment tracking, dataset lineage, model and artefact versioning, CI/CD	4.3, 6.1, 7.3
☐	Monitor drift and trigger the next retraining cycle from it	4.3, 6.1, 7.3
☐	Apply control-system-grade validation gates before any model influences a live loop; every retraining passes the same gates as the original release	4.3, 6.1, 7.3
☐	Provide sandboxed trust and robustness evaluation before field release	6.1, 6.3, 7.5
☐	Quantify surrogate accuracy outside the training envelope before twin outputs inform operations	6.2, 7.3
☐	Calibrate continuously against field data to close the sim-to-real gap	6.2
☐	Support federated learning across operator fleets where data cannot leave its owner	6.1, 6.4, 7.3
☐	Keep documentation and epistemic traceability that mitigate update opacity	4.3, 8
☐	Produce explanations that are meaningful to operators and maintenance crews	2.3, 6.4, 7.4

F. Edge deployment and OT interface

☐	Requirement	Section
☐	Compress and containerise models for edge nodes within strict memory and power budgets	2.2, 6.4, 7.3, 7.4
☐	Provide serving interfaces that fit OT systems never designed for model serving	2.4, 7.4

© AIOTI. All rights reserved.

☐	Run each function in its latency tier; keep protection local and keep the Factory out of the real-time loop	4.2, 6.3, 6.5, 7.4
☐	Run edge models inside deterministic safety envelopes using standard industrial protocols	7.4, 8
☐	Provide human-in-the-loop override and explainable dashboards, so local autonomy never bypasses physical protection	6.3, 7.4, 8
☐	Validate on a gate's test infrastructure before any operational use	6.1, 6.3
☐	Design for OT-grade reliability and safety from the outset	2.3
☐	Resolve interoperability between industrial twins and federated TSO–DSO twins	3.1, 6.2

G. Security and sovereignty

☐	Requirement	Section
☐	Meet NIS2 risk-management and incident-reporting obligations	4.4, 4.6, 7.5
☐	Apply NIS2-aligned access practices and data residency	6.1
☐	Respect restrictions on using critical-infrastructure inspection data on shared compute	6.3
☐	Keep prosumer and household data GDPR-compliant on shared infrastructure	4.4
☐	Keep human oversight and transparency in AI-supported security analysis	4.4
☐	Operate as sovereign critical infrastructure, free of dependence on non-EU decisions	3.1, 4.4

H. Regulation and compliance

☐	Requirement	Section
☐	Classify each use case under the AI Act (Annex III, point 2): safety component or advisory only	2.4, 4.6, 7.5
☐	Start compliance at the first training run: risk management, data governance, technical documentation, human oversight, conformity assessment	2.4, 4.6, 7.5, 8
☐	Provide a sandbox for conformity evaluation, built into MLOps	7.5, 8
☐	Track retraining events against the "substantial modification" question, pending regulatory clarification	6.1, 7.5, 8
☐	Align with the CER Directive	2.4, 4.6
☐	Align with the electricity directive: smart metering, flexibility, TSO/DSO unbundling data rules	2.2, 4.6, 7.5
☐	Meet EuroHPC access-tier eligibility and the amended EuroHPC Regulation	4.6, 7.5

I. Access, governance and the AI Gate

☐	Requirement	Section

© AIOTI. All rights reserved.

☐	Offer standing, lightweight allocations for continuous drift-triggered retraining, alongside call-based tiers	2.4, 6.1, 8
☐	Adapt access tiers to mixed public–private TSO/DSO/industry consortia	6.7, 8
☐	Establish a link to an AI Gate for readiness, AI Act screening and test-before-invest, with referral in both directions	6.1, 6.7
☐	Agree joint KPIs with the Gate: SMEs supported, referral volume, onboarding time, pilot-to-deployment rate	6.7, 8
☐	Secure permanent governance and funding for the Gate relationship	6.1, 6.7, 7.6, 8
☐	Coordinate with EDIHs and TEFs to avoid duplication, and carry energy specifics (network codes, critical-infrastructure security) through the gates	6.7
☐	Govern the whole pipeline from ingestion to deployment, with EU-level data governance	3.1, 8
☐	Involve energy partners for domain data, since manufacturing data do not transfer to turbines, pumps, or transformers	6.4
☐	Route SMEs to aggregation-level models through the Gate, without each negotiating compute access	6.6
☐	Start with advisory-only use cases (e.g. predictive maintenance), which may fall outside the high-risk category	2.4, 6.4

Acknowledgements

All rights reserved, Alliance for AI, IoT and Edge Continuum Innovation (AIOTI). The content of this document is provided 'as-is' and for general information purposes only; it does not constitute strategic or any other professional advice. The content or parts thereof may not be complete, accurate or up to date. Notwithstanding anything contained in this document, AIOTI disclaims responsibility (including where AIOTI or any of its officers, members or contractors have been negligent) for any direct or indirect loss, damage, claim, or liability any person, company, organisation or other entity or body may incur as a result, this to the maximum extent permitted by law.

About AIOTI

AIOTI is the multi-stakeholder platform for stimulating AI, IoT and Edge Continuum Innovation in Europe, bringing together small and large companies, academia, researchers, policy makers, end-users and representatives of society in an end-to-end approach. We strive to leverage, share and promote best practices in the AI, IoT and Edge Continuum ecosystems, be a one-stop point of information to our members while proactively addressing key issues and roadblocks for economic growth, acceptance and adoption of the AI, IoT and Edge Continuum Innovation in society. AIOTI contributions goes beyond technology and addresses horizontal elements across application domains, such as matchmaking and stimulating cooperation by creating joint research roadmaps, defining policies and driving convergence of standards and interoperability.