Journal of Edge Computing, 2026, Vol. **5**, Iss. 1, pp. 1–46 10.55056/jec.1382

The bio-edge: a survey and research agenda for the Internet of Bio-Nano Things, 2026–2035

Serhiy O. Semerikov[1,2,3,4,5], Tetiana A. Vakaliuk[3,2,1,5]

[1]*Kryvyi Rih State Pedagogical University, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*
[2]*Institute for Digitalisation of Education of the NAES of Ukraine, 9 M. Berlynskoho Str., Kyiv, 04060, Ukraine*
[3]*Zhytomyr Polytechnic State University, 103 Chudnivsyka Str., Zhytomyr, 10005, Ukraine*
[4]*Center for Information-analytical and Technical Support of Nuclear Power Facilities Monitoring of the NAS of Ukraine, 34a Palladin Ave., Kyiv, 03142, Ukraine*
[5]*Academy of Cognitive and Natural Sciences, 54 Gagarin Ave., Kryvyi Rih, 50086, Ukraine*

Abstract. The Internet of Bio-Nano Things (IoBNT) extends the Internet of Things into the biochemical domain of living systems through nanoscale bio-engineered devices that sense, actuate, and communicate primarily via molecular signalling. Eleven years after the founding vision of Akyildiz et al. [4], the field has accumulated working architectures, microfluidic testbeds, and mature channel models, but its system-integration challenges – latency, privacy, and energy – are substantially edge-computing challenges: in-body decision loops cannot tolerate cloud round-trip latency, biomolecular data cannot safely stream to remote servers, and harvested-power devices cannot continuously transmit raw high-rate signals. This survey reframes the IoBNT layer stack as a five-layer bio-edge reference architecture in which the bio-cyber interface (BCI, distinct from brain–computer interface) is upgraded from a transduction gateway to a first-class compute layer with its own latency, energy, and trust accounting. We construct a DOI-deduplicated bibliometric snapshot of 311 entries, identify three under-occupied subtopics that constitute the field's strategic white space – TinyML on harvested power, federated learning across edge gateways, and Bio-SDN orchestration – and survey the technical state of each. The centrepiece is a ten-prediction research agenda for 2026–2035 with each prediction stated as a dated metric, a causal mechanism, and a falsifier, designed to give the IoBNT community a structured object that subsequent work can measure itself against.

Keywords: Internet of Bio-Nano Things, IoBNT, edge computing, molecular communication, bio-cyber interface, edge AI, federated learning, TinyML, cyberbiosecurity, research agenda

1. Introduction

Eleven years after Akyildiz et al. [4] described the *Internet of Bio-Nano Things*[1] as a heterogeneous networking framework that extends the Internet of Things into the biochemical domain of living systems, the field has accumulated enough working architectures, microfluidic testbeds, and theoretical channel models to begin shedding its reputation as "speculative networking research" and acquire the apparatus of an applied engineering discipline. By the spring of 2026 the published literature on IoBNT is accelerating sharply: the deduplicated corpus underlying this survey (section 3) contains 311 unique entries spanning 2010–2026, with 66 published in 2025 alone and 16 more already in print or in press by the end of April 2026.

This survey takes a deliberate *angle* on that literature rather than a comprehensive cross-section. It frames IoBNT not as a special case of nanonetworking, but as the natural "deepest tier" of an edge-computing stack that runs from nanomachines inside a living host, through implantable or wearable bio-cyber gateways, to mobile or local edge nodes, and only finally to a centralized cloud.

0000-0003-0789-0272 (S. O. Semerikov); 0000-0001-6825-4697 (T. A. Vakaliuk)
semerikov@gmail.com (S. O. Semerikov); tetianavakaliuk@acnsci.org (T. A. Vakaliuk)
https://acnsci.org/semerikov (S. O. Semerikov); https://acnsci.org/vakaliuk (T. A. Vakaliuk)

Received	Accepted	Published	Version of record
2026-05-19	2026-05-20	2026-05-21	2026-05-21

[1]Throughout the paper we shorten this to IoBNT; BCI denotes *bio-cyber interface*, not brain–computer interface.

We argue that the engineering and societal challenges of IoBNT are, at the system-integration level, *substantially edge-computing challenges*, because (i) the intrinsic latency budget of in-body decision loops forbids round-tripping to a cloud, (ii) the data harvested by bio-nano sensors is among the most privacy-sensitive any computing system has ever handled, and (iii) the power budget of harvesting-only nanoscale devices is so small that meaningful information must be extracted before it is transmitted, not after.

Five substantial surveys closely related to ours have appeared in the last fourteen months. Torres Gómez et al. [45] provide a comprehensive treatment of neural networks for molecular communication, with particular depth on detection and channel estimation. Bhattacharjee et al. [7] concentrate on exhaled-breath analysis as a paradigmatic external IoBNT application. Rana et al. [38] present a broad architecture–benefits–security review of the wider Internet of Nano Things (IoNT) ecosystem that subsumes IoBNT as a subset. Bulasara et al. [10] focus on the insulin–glucose loop and adjacent intra-body applications. Closest to our framing, Meenambika et al. [34] survey AI-driven analysis and personalized-health IoBNT deployments, with attention to security and regulatory considerations. The present paper differentiates from each: it is not a machine-learning survey (we cite Torres Gómez et al. [45] where they have already done the work); it is not application-bound (unlike Bhattacharjee et al. [7], Bulasara et al. [10]); it is not security-anchored (Zafar et al. [46] and Rana et al. [38] remain the references there); and it addresses the edge tier across the full IoBNT layer stack rather than within healthcare alone – a delimitation Meenambika et al. [34] accepts. Section 3 formalizes this differentiation in table 3.

This paper makes four contributions.

1. A reproducible *bibliometric snapshot* of the IoBNT literature (section 3) constructed from a DOI-keyed union of Web of Science and Scopus exports. The snapshot, together with the open-source dedup pipeline that produced it, lets readers audit and extend our citation base.
2. A *five-layer reference architecture for the bio-edge stack* (section 4) that sharpens the typical four-tier framing in prior work by elevating the bio-cyber interface from a "gateway" to a first-class layer with its own latency and energy accounting.
3. A *joint treatment of the three under-occupied edge sub-disciplines* that the bibliometric snapshot identifies as strategic gaps: TinyML on harvested-power devices (section 9), federated learning on bio-cyber interfaces (section 8), and orchestration of in-body networks via Bio-SDN control planes (section 10). Each of these subtopics is represented by fewer than ten entries in the deduplicated corpus, even though their absence from existing surveys makes them, in our view, among the most important research fronts for the next half-decade.
4. A *research agenda* (section 15) consisting of dated, mechanism-justified, falsifiable predictions for the development of IoBNT-edge systems between 2026 and 2035. The intent is to give the community a structured object that subsequent work can measure itself against, rather than the unfalsifiable visionary paragraphs that close many surveys in this area.

We do *not* attempt a fresh survey of molecular-communication channel models – the work of Kuscu et al. [26], Hofmann et al. [18], and Hamidović et al. [17] is recent, accessible, and adequate, and we cite it. We do not produce experimental measurements; the paper is strictly analytical. We do not advocate a single edge platform or inference framework; the field is too young for premature standardization, and section 15 identifies platform maturity as an open prediction.

Section 2 traces the intellectual lineage of IoBNT from Akyildiz et al. [4] through the three present research-programme centres of gravity, and motivates the edge framing. Section 3 presents the bibliometric snapshot and the survey-comparison matrix. Sections 4 through 14 are the technical chapters, organized along the five-layer reference architecture and the three white-space subtopics. Section 15 is the research-agenda centrepiece. Section 16 closes. Three appendices document the bibliographic methodology (Appendix A), a glossary of acronyms (Appendix B), and the simulator and testbed landscape (Appendix C).

2. The bio-edge proposition: vision and lineage

2.1. From a 2015 communications-magazine vision to a discipline

The contours of IoBNT were drawn by Akyildiz et al. [4] in a 2015 IEEE Communications Magazine article that combined four previously separate threads: (i) nanonetworking and nano-machine communication; (ii) the molecular and electromagnetic communication theory that had developed alongside it; (iii) the synthetic-biology programme that was beginning to produce programmable cells; and (iv) the architectural conventions of the Internet of Things. Akyildiz et al.'s framework was deliberately schematic: bio-nano things (BNTs) sense and actuate chemically inside a host; nanonetworks of BNTs communicate primarily via molecular signalling; bio-cyber interfaces transduce molecular into electromagnetic signals; macro-scale devices forward the result to the conventional internet. The framework was published as a research call, not a system specification, and almost every subsequent survey has used some restatement of it as its opening figure.

The pre-2015 lineage that fed this vision deserves a brief mention. Targeted drug delivery via nanocarriers [12], information-theoretic analyses of intra-body biological communication [2], and the controllability of mobile bionanosensors [37] were each being pursued in their own communities; Akyildiz et al.'s contribution was to recognise them as facets of one stack.

A comprehensive 2021 consolidation by Kuscu and Unluturk [27] surveyed the post-2015 IoBNT landscape across biomedical, agricultural, and environmental domains together with the full enabling-technology stack (MC, THz, FRET, magnetic, heat, and acoustic communication), and remains the most complete single-volume reference for the field's pre-2021 state.

2.2. Three programme centres

In the decade that followed, three sustained research programmes turned the schematic into accumulating concrete results. Section 3 (figure 3) renders the co-author graph of the field; the same three communities emerge without manual labelling:

1. *Cambridge–Koç.* The group led by Ö. B. Akan, with M. Kuscu and collaborators at Koç University and Cambridge, has been responsible for the bulk of the field's information-theoretic apparatus [2, 25] as well as the transmitter-and-receiver architecture synthesis that has become the field's most-cited tutorial reference [26]. The group also closes the loop to clinical translation in Akan et al. [3], an information-and-communication-theoretical treatment of IoBNT in the context of spinal-cord injuries. Adjacent work by the same circle has opened the agriculture-IoBNT direction [5] and a recent semantic-learning treatment of molecular communication [11].
2. *Erlangen–Dresden.* The group around R. Schober at Erlangen, in close collaboration with the Dresden 6G-life cluster led by F. H. P. Fitzek, has driven the systems-level work: cardiovascular IoBNT prototypes [28], multipath nanocommunication [44], and the recent neural-network-for-IoBNT survey [45] against which the present paper explicitly positions itself.
3. *JKU Linz.* The Haselmayr–Angerbauer group at Johannes Kepler University Linz contributes a disproportionately large share of the field's recent output (table 2), with strengths in detector design, channel-impulse-response modelling, and the practical microfluidic testbeds that the field needed to leave the simulation phase [17, 35].

These programmes are complemented by individual contributions on materials – graphene at the bio-cyber interface [13] – and by synthetic-biology approaches that re-engineer living cells as programmable molecular logic [9], with parallel work on therapeutic modulation of natural in-body MC pathways such as the gut-brain axis [31].

2.3. What is now built

Eleven years after the vision paper, the discipline can credibly claim:

- channel models and detection theory for molecular communication that are mature enough to underpin standardized benchmark suites in principle, though those suites do not yet exist [14, 18, 26];
- operational microfluidic testbeds capable of end-to-end molecular signalling experiments at low cost [17, 35];
- early but credible applications in continuous health monitoring, theranostic drug delivery, and brain–machine interfacing [3, 10, 12, 34];
- machine-learning techniques for symbol detection, channel estimation, and semantic decoding that begin to approach the practical constraints of in-body operation [11, 41, 45];
- a first principled coupling of IoBNT data to digital-twin pipelines, with federated training across edge gateways [20, 21].

These claims span a wide range of maturity. Using the NASA/ISO technology-readiness-level (TRL) scale adapted to computational IoBNT research: channel models and symbol-detection theory sit at TRL 1–3 (basic principles observed and formulated); microfluidic testbeds are at TRL 3–4 (component validation in a laboratory environment); ML-based channel estimation and semantic decoding are at TRL 2–3 (algorithm validation against simulation traces, with limited experimental data); digital-twin coupling is at TRL 1–2 (conceptual frameworks and single-paper demonstrations, not yet reproduced across multiple groups). No IoBNT subsystem relevant to this survey has yet reached TRL 5 (system validation in a relevant environment, e.g., chronic in-vivo operation). The research agenda (section 15) maps predictions to the TRL progression that would be needed to reach clinical deployment by 2035.

2.4. The edge tier as the missing piece

What the vision paper did *not* foreground – and what no single subsequent survey has yet treated systematically – is the *computational tier* that sits between the molecular nanonetwork and the cloud. In Akyildiz et al. [4] this tier is named only obliquely as the bio-cyber interface; in most subsequent architecture diagrams it appears as a thin gateway box that performs signal transduction but is not credited with computational responsibilities. Yet every constraint that makes IoBNT difficult – latency floors set by life-critical control loops, intermittent power harvested from blood flow or body heat, and the privacy implications of streaming a patient's biomolecular state to a remote server – argues for that tier to host the bulk of the system's intelligence, not for it to forward.

This is the proposition of the present survey. The bio-edge tier needs its own reference architecture, its own energy and inference budget, its own security model, its own machine-learning workflows (including federated and TinyML variants), and its own orchestration plane. Section 4 formalizes that architecture; sections 6–13 populate it.

2.5. Why the edge moment is now

Three enabling trends, each independent of IoBNT, converged during 2023–2025 to make the edge-centric framing not only defensible but overdue.

1. *The publication inflection.* The corpus analysed in section 3 shows that IoBNT output roughly doubled in 2023 and continued to accelerate through 2025. A field producing 66 papers per year is large enough to sustain dedicated sub-literatures; the question is no longer whether someone will write the first edge-computing paper for IoBNT, but which community will claim the topic.
2. *Microfluidic testbed maturity.* Until recently, molecular-communication experiments required specialist wet-laboratory infrastructure that excluded networking and embedded-systems researchers. The publication of low-cost, reproducible microfluidic platforms – specifically the acoustic levitation and droplet-based testbeds of Hamidović et al. [17] and Miray Albay et al.

[35] – lowers the barrier to entry for edge-hardware integration. A researcher who can assemble a Cortex-M development board can, in principle, now also operate a molecular-communication testbed.

3. *FL and TinyML software-stack emergence.* The toolchains that an IoBNT-edge system would need – TFLite-Micro for on-device inference, Flower or FedML for federated orchestration, MicroTVM for compiler-level optimization – completed their 1.0 transitions during this window [6, 30]. The gap is no longer tooling; it is the absence of an IoBNT-specific benchmark and reference implementation that would let those tools be applied to molecular-communication workloads (section 13).

These trends are *enabling conditions*, not accomplished facts. The present paper is therefore **prescriptive** (it advocates a shift of computational responsibility toward the bio-edge tier) rather than **descriptive** (it does not claim the shift has already occurred). The bibliometric snapshot of section 3 confirms that the three edge sub-disciplines we advocate remain under-occupied; the reference architecture of section 4 supplies the common vocabulary that the advocacy requires.

2.6. Limitations of the edge-centric framing

The edge-centric lens is powerful but not total. Three domains shape IoBNT outcomes at least as strongly as edge-computing architecture, and this survey does not treat them in depth.

1. *Molecular physics and channel stability.* The information-theoretic limits of molecular communication are set by diffusion dynamics, chemical kinetics, and intersymbol interference – physical phenomena that edge-computing architecture does not alter. A better edge inference engine does not widen a diffusion-limited channel; at most it extracts more information from the signal the channel delivers. The channel modelling literature surveyed in section 5 remains the rate-limiting tier.
2. *Synthetic biology and biochemical programmability.* Engineered cells as programmable logic elements – transcriptors, kill-switches, quorum-sensing relays – operate on biochemical timescales (minutes to hours) and are constrained by genetic circuit stability, host-cell burden, and evolutionary pressure, none of which edge computing addresses [9]. The bio-cyber interface can *read* synthetic circuits; it cannot *design* them.
3. *Regulatory and clinical translation.* The deployment timeline of an implantable or ingestible IoBNT system is dominated by regulatory review (FDA Class III / PMA or EMA Annex I equivalent), clinical-trial design, and reimbursement policy, not by edge-stack engineering. Edge computing can strengthen the safety and auditability case that regulators evaluate, but it cannot shorten the review calendar. The prediction chapter (section 15) incorporates regulatory milestones explicitly for this reason.

Acknowledging these limits does not weaken the edge argument; it sharpens it. Where edge computing *can* make a difference – in the latency, privacy, and energy-efficiency of the post-transduction signal-processing pipeline – the present survey is explicit that it is a necessary but not sufficient condition for clinical and commercial IoBNT systems.

3. A literature map of IoBNT × edge

Before the technical chapters, we ground the survey in a bibliometric snapshot of the literature actually available to readers in May 2026. The corpus is the union of (i) two Web of Science exports, (ii) a Scopus export of *Internet of Bio-Nano Things* hits dated 2026-04-28, and (iii) two arXiv-search printouts on the same topic. After DOI-keyed deduplication and removal of ∼60 pre-2010 humanities artefacts that the Web-of-Science query inadvertently caught (titles such as *Leibniz rules and reality conditions* and *SOARING*), the working corpus is $N = 311$ unique entries. The full methodology and reproducibility checklist appear in Appendix A.

The publication trajectory of figure 1 shows a clear inflection in 2023. From 2010 to 2022 the field averaged about 13 entries per year; 2023 jumped to 30, 2024 to 46, and 2025 to 66. The first four months of 2026 already account for 16 entries. The Scopus tail (orange) captures conference proceedings, most prominently ACM `NanoCom` and IEEE `GLOBECOM`, that Web of Science does not consistently index.

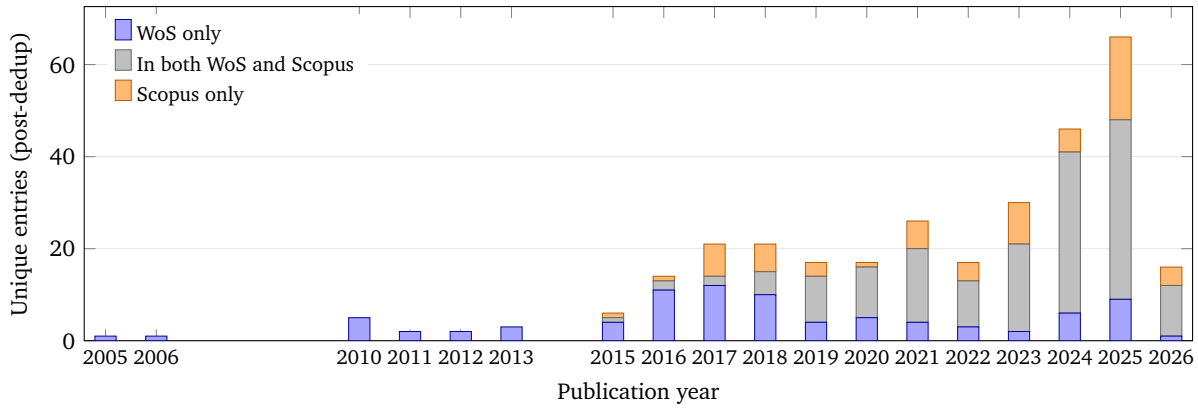

Figure 1: Annual IoBNT-corpus publication counts after DOI-keyed deduplication of the Web of Science and Scopus exports ($N = 311$). The 2023 inflection marks the point where annual output roughly doubles. The bulk of recent papers appear in both indexes, but a non-trivial Scopus tail captures conference proceedings (e.g., ACM NanoCom) that are not consistently indexed by WoS.

The TF–IDF + t-SNE projection of figure 2 reveals a single dense cloud of molecular-communication channel-theory work surrounded by smaller satellites for healthcare applications, bio-cyber interfaces, and 6G integration. The three white-space subtopics that this survey claims – *edge computing for IoBNT*, *federated learning at the bio-edge*, and *TinyML on harvested power* – are highlighted with star, plus, and cross markers. They occur sparsely and do not yet form a coherent cluster of their own. Building one is part of the contribution of this paper.

Figure 3 renders the co-author network induced by the corpus (authors with ≥3 deduplicated papers, edge weights = number of co-authored entries, community detection via greedy modularity). Three centres of gravity emerge *without prior labelling*: a JKU-Linz cluster around Haselmayr, Angerbauer, and Springer; a Cambridge–Koç cluster around Akan, Kuscu, and Pierobon; and an Erlangen–Dresden cluster around Schober, Lotter, Brand, Hofmann, Fitzek, and Dressler. Tables 1 and 2 list the top venues and authors numerically.

Figure 4 expresses the gap quantitatively: under a rule-based labeller (Appendix A), the corpus is dominated by molecular-channel theory (115 entries) and healthcare applications (33), while the three white-space buckets together hold fewer than ten entries.

To test whether the zero-entry TinyML bucket (figure 4) is an artefact of the IoBNT-branded query rather than a genuine gap, we searched the full corpus for six edge-inference keywords unrelated to the IoBNT acronym itself: `quantization`, `microcontroller`, `Cortex-M`, `TFLite`, `microTVM`, and `pruning`. Exactly 4 entries match at least one term; of those, *zero* combine the keyword with an explicit IoBNT reference. The gap is therefore real at the corpus level – the terms exist in the wider literature but have not crossed into IoBNT-branded work. The adjacent TinyML literature is surveyed in section 9 and its importation into IoBNT is the subject of prediction P6 (section 15).

Table 3 makes the differentiation against the five most-closely-related 2025–2026 surveys explicit. This table is *not* an evaluation of the comparator surveys (the columns are deliberately coarse and the marks are assigned by the present authors); it is a coverage heatmap whose purpose is to identify the three white-space rows (federated learning, TinyML, Bio-SDN) that motivated the present work and that this paper occupies almost alone, together with the only dated, falsifiable research agenda (section 15). The five comparator surveys remain the references for the rows where they have substantial coverage; this paper cites them rather than re-summarises them.

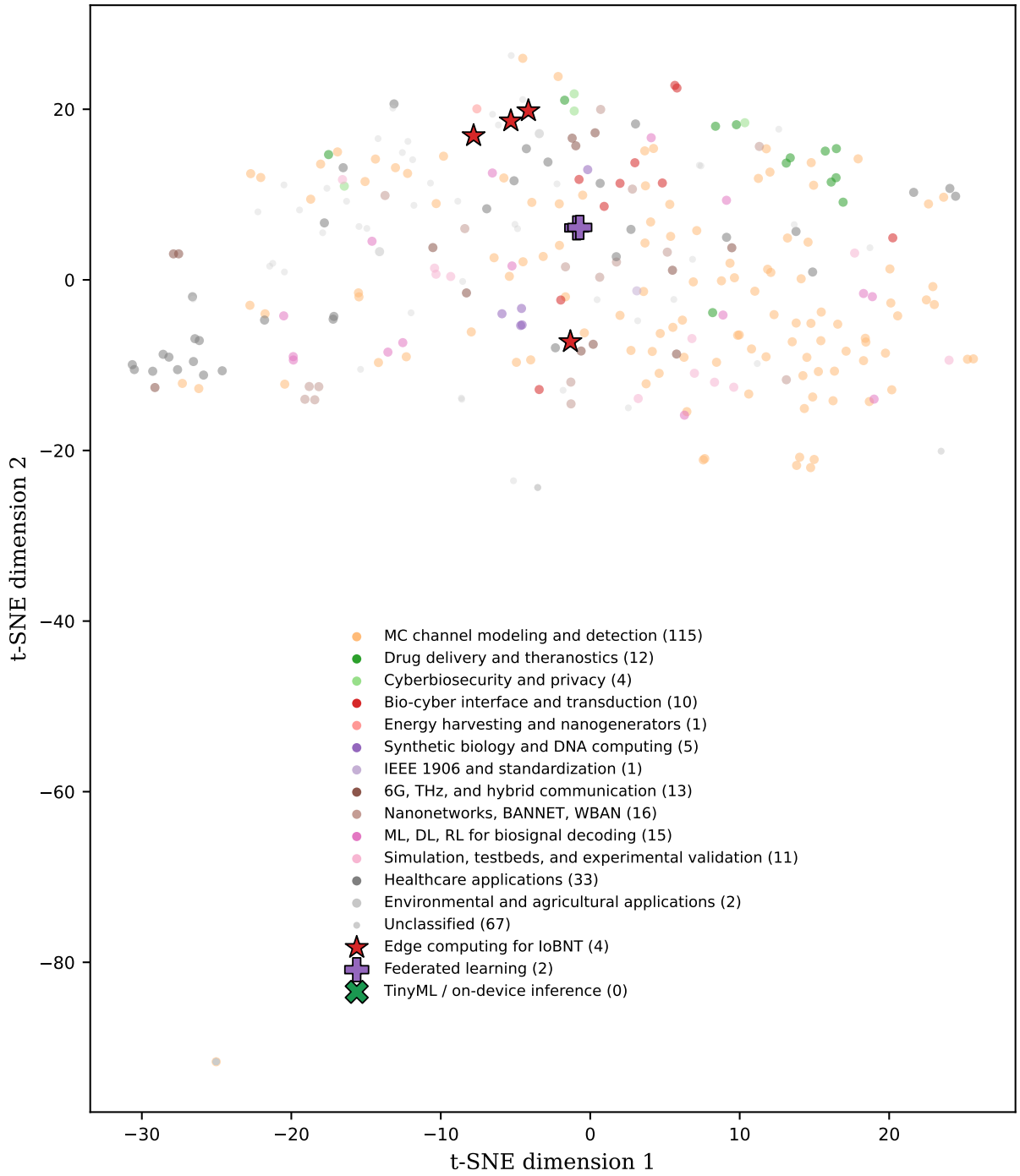

Figure 2: Topical structure of the deduplicated IoBNT corpus ($N = 311$). Each point is one paper, vectorised over its title + keywords + abstract using TF-IDF (1–2-grams, ≥2 documents) and projected to 2-D with t-SNE (`metric=cosine`, `perplexity`=min(30,$N/10$)). Colours encode subtopic buckets; the three white-space buckets ($\star$, +, ×) are drawn over the cloud to visualise their occupancy.

4. Reference architecture: the bio-edge stack

The schematic architecture proposed by Akyildiz et al. [4] and repeated in most subsequent IoBNT surveys [10, 34] and adjacent IoNT reviews [38] distinguishes four layers: bio-nano things, nanonetwork, bio-cyber interface, and macro-scale internet. This paper adopts a *five-layer* variant that splits the macro-scale internet explicitly into an *edge* tier and a *cloud* tier, and that upgrades the bio-cyber interface from a transduction gateway to a first-class compute layer with its own latency, energy, and

Table 1
Top venues in the deduplicated IoBNT corpus by entry count ($N = 311$, post-dedup). The three IEEE journals at the top together host roughly one in five entries; ACM NanoCom proceedings provide the bulk of the conference signal that Web of Science does not consistently index.

Count	Venue
25	IEEE Transactions on Molecular, Biological, and Multi-Scale Communications
23	IEEE Internet of Things Journal
10	Future of Internet of Bio-Nano Things in Personalized Healthcare: Applications and Challenges
9	IEEE Access
7	Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering
6	IEEE International Conference on Communications
6	NanoCom 2025 - Proceedings of the 12th ACM International Conference on Nanoscale Computing and Communication
6	IEEE Transactions on Nanobioscience
5	NanoCom 2024 - Proceedings of the 11th ACM International Conference on Nanoscale Computing and Communication
5	IEEE Communications Magazine
4	Nano Communication Networks
3	Proceedings - IEEE Global Communications Conference, GLOBECOM
3	NanoCom 2023 - Proceedings of the 10th ACM International Conference on Nanoscale Computing and Communication
3	IEEE Communications Surveys & Tutorials
3	Wireless Personal Communications

Table 2
Top contributing authors in the deduplicated IoBNT corpus, by number of co-authored entries. Three centres of gravity are visible: a JKU-Linz cluster (Haselmayr, Angerbauer, Springer), a Cambridge–Koç cluster (Akan, Kuscu, Pierobon), and an Erlangen–Dresden cluster (Schober, Lotter, Brand, Hofmann, Fitzek).

Papers	Author
35	Nieto-Chaupis, H.
19	Haselmayr, W.
17	Angerbauer, S.
13	Akan, O.
13	Lin, L.
12	Springer, A.
12	Hofmann, P.
12	Fitzek, F.
11	Chen, Y.
11	Pierobon, M.
10	Dressler, F.
8	Abd El-atty, S.
8	Wen, M.
8	Zhou, P.
7	Gattringer, M.

trust accounting. Figure 5 sketches the arrangement; table 4 gives the layer-by-layer parameters that we will use as a reference grid throughout the remaining sections.

4.1. Layer 1: Perception (BNTs)

The deepest tier consists of BNTs themselves: engineered cells, nanoparticle-based transducers, DNA-origami devices, or hybrid bionanomachines. Sensing and actuation are biochemical: BNTs detect the presence of a target analyte through ligand binding or gene-circuit activation, and actuate by releasing

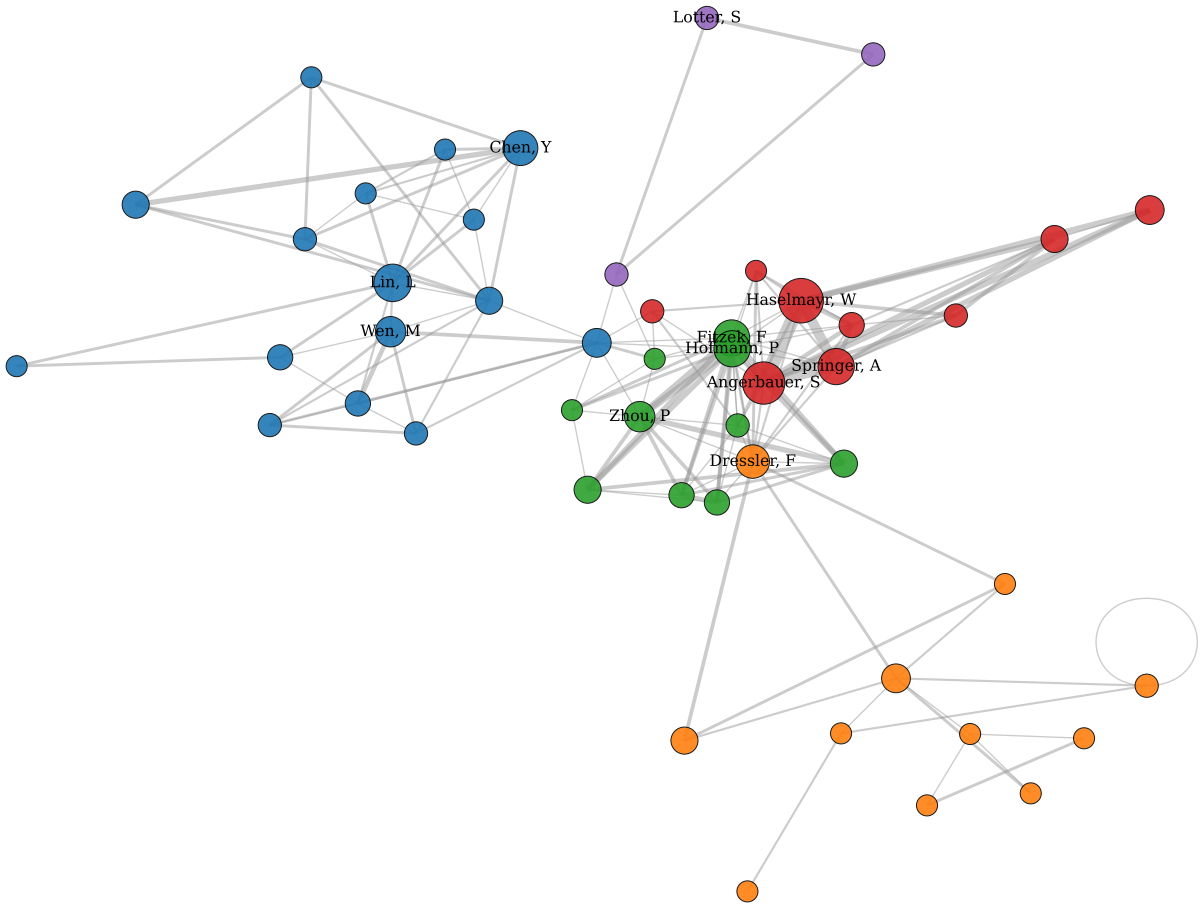

Figure 3: Co-author network of the IoBNT corpus restricted to authors with at least three deduplicated papers. Edges are weighted by the number of shared papers; communities are detected by greedy modularity (Clauset–Newman–Moore). The three centres of gravity emerge automatically without prior labelling: a JKU-Linz cluster around Haselmayr/Angerbauer/Springer, a Cambridge–Koç cluster around Akan/Kuscu, and an Erlangen–Dresden cluster around Schober/Lotter/Brand/Hofmann/Fitzek.

a molecular payload, emitting a fluorescent signal, or modifying their internal state. The information-theoretic analyses of Abbasi and Akan [2] and the genetic-logic primitives revisited in Bonnet et al. [9] are the relevant primitives here. Energy at this tier is whatever the host environment provides: ATP from cellular metabolism, ambient ion gradients, or – for hybrid nanoparticle devices – harvested mechanical or thermal energy from body motion. Compute is similarly intrinsic: a gene-regulatory network *is* the computation, and "programming" a BNT means designing that network.

4.2. Layer 2: Nanonetwork

A nanonetwork is a population of BNTs whose collective behaviour is mediated by molecular signalling and short-range electromagnetic links. Okaie et al. [37] model the controllability of mobile bionanosensors; Kuscu et al. [26] catalogue the transmitter and receiver primitives now in use. Critically for the present survey, the nanonetwork is *not* a computational fabric in the conventional sense: its links are stochastic, diffusive, and slow (section 5), and its individual elements have no power budget for sustained inference. The information that leaves a nanonetwork is a sparse, noisy, and delay-jittered shadow of the underlying biological state. Inference of that state must therefore happen *somewhere else*, and "somewhere else" is the bio-cyber interface.

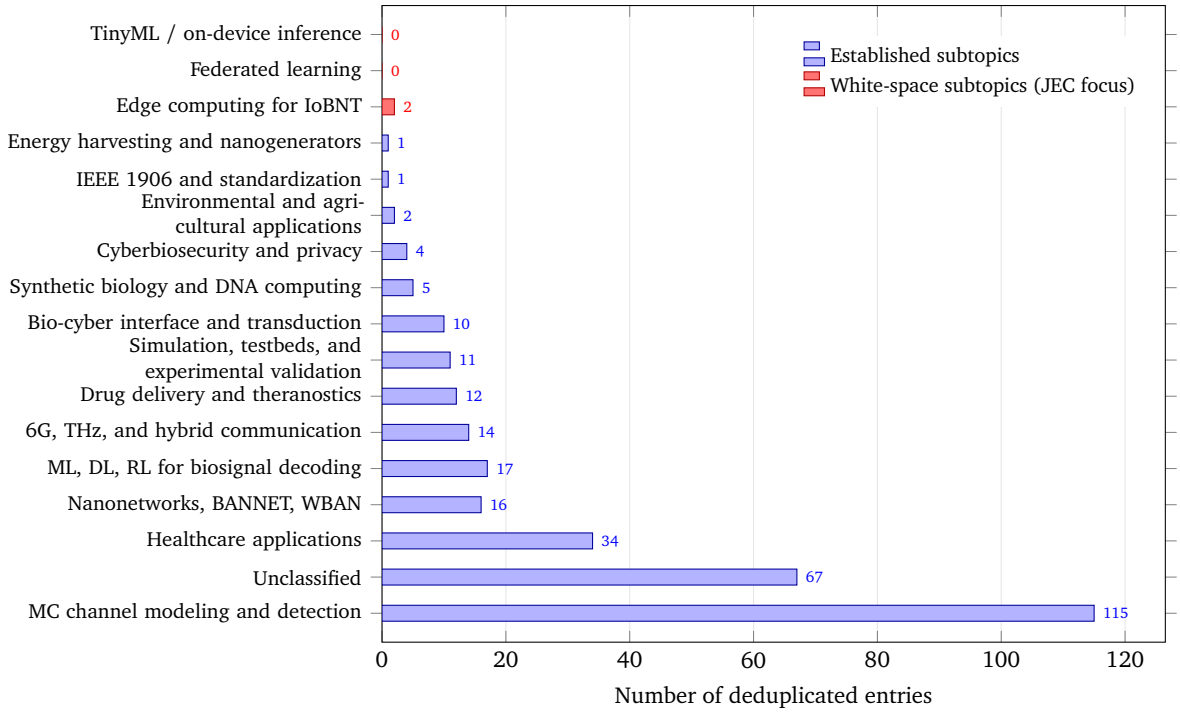

Figure 4: Topical distribution of the $N = 311$ deduplicated IoBNT corpus entries under a rule-based labeller (see Appendix A). The three white-space subtopics at the bottom of the chart – edge computing for IoBNT, federated learning, and TinyML on harvested power – are the JEC survey's strategic focus. Despite the field's centre of gravity sitting at molecular communication channel modelling ($n = 115$), the edge stack is barely visited.

Table 3
Coverage matrix of this survey against the five most closely related 2025–2026 IoBNT surveys. Marks: ■ comprehensive treatment; ⊡ partial or brief; □ not covered. The differentiator is concentrated in the three white-space rows (federated learning, TinyML, Bio-SDN) and in the dated research agenda.

Coverage axis	Semerikov and Vakaliuk [42]	Torres Gómez et al. [45]	Bhattacharjee et al. [7]	Rana et al. [38]	Bulasara et al. [10]	Meenambika et al. [34]
MC channel theory (depth)	⊡	■	■	⊡	⊡	⊡
Bio-cyber interface as edge boundary	■	⊡	⊡	⊡	□	⊡
Five-layer reference architecture	■	⊡	⊡	⊡	⊡	⊡
Edge intelligence (ML for MC)	■	■	⊡	□	□	■
Federated learning at the bio-edge	■	□	□	□	⊡	⊡
TinyML/ on-device inference	■	□	□	□	□	□
Bio-SDN / orchestration	■	□	□	□	□	□
Energy harvested-vs-compute budget	■	□	□	⊡	⊡	⊡
Cyberbiosecurity	■	□	□	■	⊡	□
Standards critique (IEEE 1906)	■	⊡	□	⊡	□	□
Dated, falsifiable research agenda	■	□	□	□	□	□

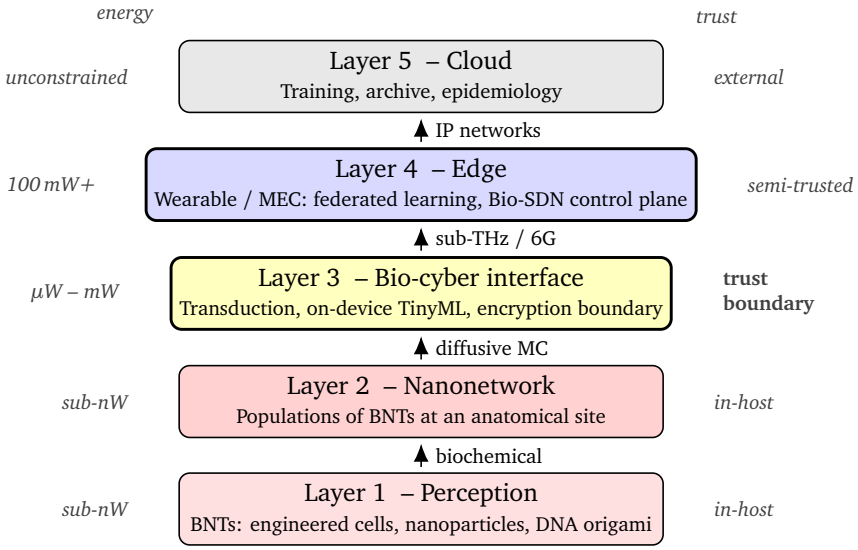

Figure 5: The five-layer IoBNT reference architecture used throughout this paper (section 4). The bio-cyber interface (Layer 3) is upgraded from the conventional gateway-box framing to a first-class compute layer that hosts the system's TinyML profile, sits at the energy-regime transition between harvested-power and milliwatt-class devices, and marks the trust boundary at which biological data enters the conventional cyber domain. Quantitative layer properties appear in table 4.

Table 4
Layer-by-layer properties of the five-layer bio-edge reference architecture defined in section 4. Energy and compute regimes are order-of-magnitude estimates representative of current best practice; the bands are necessarily wide because the field has not converged on standard reference devices.

Layer	Representative components	Communication modality	Energy regime	Compute regime	Trust
Perception	Engineered cells, nanoparticles, DNA origami	Biochemical	sub-nW	gene-regulatory network	in-host
Nanonetwork	Populations of BNTs at a site	Diffusive MC	sub-nW per device, aggregated	none beyond cell circuits	in-host
Bio-cyber interface	Nanorouter, BioFET array, graphene-plasmonic antenna, microfluidic chip	MC → EM transduction	1–1000 μW	Cortex-M, TinyML profile	**trust boundary**
Edge	Wearable patch, smartphone, on-prem MEC node	BLE, sub-GHz, sub-THz 6G	100–10 000 mW	Cortex-A or GPU, FL client	semi-trusted
Cloud	Hospital data centre, public cloud	IP networks	unconstrained	data-centre training and storage	external

4.3. Layer 3: Bio-cyber interface (BCI)

The BCI is the layer at which molecular signals are transduced into electromagnetic ones, and – we argue – it is the layer at which most of the system's intelligence should reside. The standard taxonomy of transduction methods covers redox-based interfacing, optogenetic reporters, biological field-effect transistors (BioFETs), and graphene-plasmonic terahertz interfaces [12, 13, 17]. The operational microfluidic testbeds of recent years [35] demonstrate that a BCI can now be constructed at modest cost and instrumented with conventional microelectronics. Section 6 treats this layer in depth. For the present overview the key point is that the BCI sees the raw nanonetwork output and

is the closest place at which an electronic processor can run a model: it is the natural home for symbol detection, channel-impulse-response estimation, and the first stage of feature extraction for downstream applications.

4.4. Layer 4: Edge

The edge tier is the unique contribution of an explicitly edge-oriented framing. Architecturally it is whatever sits between the BCI and the cloud: a wearable patch, a smartphone, a roving mobile-edge-computing (MEC) node in a hospital ward, or a fixed on-premises server. Meenambika et al. [34] discuss this tier indirectly in the context of AI-driven personalized-healthcare IoBNT deployments; Jamshidi, Hoang and Nguyen [20] describe a federated training pipeline whose participating nodes sit at this tier. We will argue throughout the remainder of the paper that the edge tier is the appropriate locus for (i) more capable machine-learning inference than the BCI can host (section 7), (ii) federated learning across patient cohorts (section 8), (iii) TinyML profiles when the BCI's microcontroller is genuinely energy-bound (section 9), and (iv) the Bio-SDN control plane that orchestrates the in-body network (section 10). It is also the tier at which most of the privacy-by-design machinery – local differential privacy, secure aggregation, encryption-at-rest – is practically realised (section 12).

4.5. Layer 5: Cloud

The cloud tier is, by design of an edge-centric framing, the *least* active part of the system. Its appropriate responsibilities are batch model training across federated cohorts, long-term archival of de-identified records, cross-site epidemiology, and global software updates. We deliberately do not foreground the cloud tier; the operational and ethical problems of in-body IoBNT collapse if the cloud is asked to do work that should have been done at the edge.

4.6. Reading the stack across the rest of the paper

The next four sections traverse the stack from the bottom up. The communication substrate (section 5) sets the link-layer constraints between Layers 1 and 2; the bio-cyber interface (section 6) realises the boundary between Layers 3 and 4; the three machine-learning chapters (sections 7–9) populate the inference capabilities of Layers 3 and 4; orchestration (section 10) describes the control plane that coordinates them; the energy chapter (section 11) closes the loop between harvested-power Layer 1 and the inference-power requirements of Layers 3–4; and the security chapter (section 12) draws trust boundaries between all five.

4.7. Edge-placement decision framework

The five-layer stack does not prescribe *where* inference should run; it provides the vocabulary for making that decision parametrically. Let $\tau_{\text{edge-link}}$ denote the one-way latency from the BCI (Layer 3) to the edge node (Layer 4), and let $\tau_{\text{infer}}(t)$ denote the inference latency at tier t. The BCI is the preferred inference site when

$$\tau_{\text{edge-link}} > \tau_{\text{infer}}(\text{BCI}) - \tau_{\text{infer}}(\text{edge}), \tag{1}$$

i.e., when the cost of moving data to the edge exceeds the time saved by running inference on faster hardware. For a BLE 5 link ($\tau_{\text{edge-link}} \approx 5\,\text{ms}$) and an MLP inference at 1 ms on either platform, the BCI wins only marginally. For a CNN inference at 50 ms on the BCI versus 2 ms on a GPU-class edge node, the edge wins decisively. The real trade-off is energy, not latency: the edge node's faster inference may cost less total energy than a slow, energy-inefficient BCI inference, but only if the BCI-to-edge link energy is included in the comparison.

Table 5 summarises the qualitative guidance.

Table 5
Qualitative placement guidance for edge-intelligence workloads across the five-layer stack.

Criterion	BCI-only (L3)	Split BCI/Edge	Edge-only (L4)
Latency budget	<10 ms end-to-end	10 ms–100 ms	>100 ms tolerable
Power availability	Harvested only, spare budget	Harvested + small battery	Battery or wired supply
Privacy posture	Raw signal never leaves BCI	Features cross, not raw	De-identified features
Model complexity	≤100 kB, quantised	Early-exit at BCI, fallback to edge	Full CNN/transformer, floating point

5. Communication substrate

A complete treatment of the IoBNT communication substrate would duplicate much of Kuscu et al. [26], Hofmann et al. [18], Darya, Vakani and Nasir [14], and Hamidović et al. [17]; those references are accessible and recent, and we cite them rather than restate them. This section serves only to fix terminology and to surface the channel features that drive design choices at Layers 3 and 4 of the reference architecture (section 4).

5.1. Molecular communication (MC)

MC is the dominant modality. A transmitter releases an information molecule (a small molecule, an ion, or a labelled macromolecule) into a fluid medium; the molecule propagates via diffusion – possibly augmented by advection – and is received by a sensor that counts molecules in a small detection volume or registers a binding event on a surface receptor [25, 26]. The dominant impairments are (i) slow propagation, with characteristic delays orders of magnitude larger than electromagnetic delays at comparable distances; (ii) heavy-tailed inter-symbol interference, because diffusing molecules from one symbol contaminate many subsequent symbol intervals; and (iii) stochastic noise dominated by molecule-counting statistics. Hofmann et al. [18] survey the coding strategies that mitigate these impairments; Darya, Vakani and Nasir [14] provide a complementary error-control treatment. None of this matters to the present paper except for the consequences listed below.

5.2. Terahertz and graphene-plasmonic links

At the bio-cyber boundary, electromagnetic communication becomes practical, but in a specific regime: nano-antennas built from graphene operate efficiently at terahertz frequencies, with plasmonic resonances that match the dimensions of nanoscale transceivers [13]. THz links offer high bandwidth but very short range and high tissue absorption, which is why they appear in the architecture only at the BCI and not deeper inside the body.

5.3. Hybrid and emerging modalities

Recent work explores hybrid modalities that combine MC with electromagnetic or acoustic links to overcome MC's intrinsic slowness. Tjabben et al. [44] characterise multipath in-body nanocommunication; emerging ultrasound and microbubble schemes promise higher data rates through controlled acoustic carriers. Engineered cells acting as biological transmitters [9] blur the boundary between Layer 1 and the channel itself, by letting the source of the signal be a genetically programmed agent rather than a passive release mechanism; a related strand modulates the natural gut-brain-axis MC pathway via dietary intervention as a therapeutic transmitter [31].

5.4. What the substrate forces on the edge stack

The communication substrate forces five constraints on everything that runs at Layers 3 and 4:

1. *Low symbol rate.* Diffusion-bounded MC rarely exceeds a few bits per second at biologically relevant scales. Edge inference must therefore operate on aggressively feature-engineered, low-bandwidth time series, not raw signals.
2. *Severe ISI.* A symbol detector that ignores history will mis-classify a large fraction of received symbols [25]. Inference architectures need explicit memory – LSTM, transformer, or maximum-likelihood sequence detection – rather than memory-less classifiers.
3. *Non-stationarity.* The channel changes with patient posture, vascular dilation, gut motility, and time of day. Models trained at a single operating point degrade; this is the central motivation for federated retraining at the edge (section 8).
4. *Energy asymmetry across layers.* The transmitter (a BNT) has a very tight energy budget; the receiver (often at the BCI) is comparatively energy-rich. Detection complexity should therefore be loaded onto the receiver, not the transmitter, regardless of what the textbook channel-coding trade-off would suggest at conventional radio frequencies.
5. *Privacy-by-substrate.* Information that has not yet been transduced out of the molecular domain is, in practice, extremely difficult to exfiltrate. The latest layer at which an attacker can usefully intercept data is the BCI; this informs the trust-boundary placement in section 12.

The remaining sections of the paper take these five constraints as given.

6. The bio-cyber interface as the edge boundary

In most surveys of IoBNT, the bio-cyber interface (BCI) is rendered as a small box marked "gateway", positioned between the body and the network. This treatment captures the BCI's transduction responsibility – converting molecular signals into electromagnetic ones – but understates two facts that this paper foregrounds. First, the BCI is the *first physical location* at which conventional microelectronics, software, and machine-learning inference become available; everything deeper in the body has to negotiate with biology rather than with silicon. Second, the BCI is the *latency- and energy-dominant boundary* of any IoBNT system: events crossing it incur orders-of-magnitude jumps in signal propagation, computational throughput, and trust assumptions. The BCI is therefore the most structurally important layer to design well, and the most critical single point to mis-place inference on.

6.1. Transduction technologies

The transduction mechanisms in current use fall into four families, each with distinct latency, sensitivity, and power characteristics. Chude-Okonkwo et al. [12] survey the MC paradigm for targeted drug delivery and the associated bio-electro transduction architecture; Civas et al. [13] treat graphene specifically; Hamidović et al. [17] review microfluidic implementations.

1. *Redox-based interfacing.* Electrons flow between molecules and electrodes; the electrode reads out a current proportional to a redox-active species. Suitable for metabolites with mature electrochemistry (H_2O_2, glucose, neurotransmitters); poorly suited to redox-inert biomarkers.
2. *Optogenetic and fluorescent reporters.* Engineered cells emit light in response to a specific chemical trigger; the BCI hosts an optical sensor (a photodiode, a charge-coupled device, or an image sensor) that converts photons to electrons. Latency is essentially the photon flight time; sensitivity is limited by background fluorescence and the engineered cell's expression dynamics. Lotter et al. [31] discuss the synthetic-biology side of this mechanism for therapeutic gut–brain-axis modulation.
3. *Biological field-effect transistors (BioFETs).* A biorecognition event at the gate of a field-effect transistor modulates the channel conductance directly. BioFETs offer high sensitivity at low power and integrate cleanly with conventional CMOS manufacturing, but they require careful surface chemistry and are sensitive to ionic environment drift.

4. *Graphene-plasmonic terahertz interfaces.* Graphene transceivers exploit plasmonic resonances at THz frequencies to couple molecular vibrations into high-bandwidth electromagnetic signals [13]. These interfaces are the highest-throughput option but the most fabrication-demanding and the most range-limited.

6.2. Latency at the boundary

The latency budget for a closed-loop IoBNT intervention – "detect a biomarker; decide whether to release a therapeutic" – is set by the clinical condition being managed (table 6). A glycaemic excursion can tolerate seconds; an arrhythmia or a seizure cannot. The aggregate latency τ_{loop} decomposes into

$$\tau_{\text{loop}} = \tau_{\text{sense}} + \tau_{\text{nano}} + \tau_{\text{BCI}} + \tau_{\text{infer}} + \tau_{\text{actuate}}, \tag{2}$$

where τ_{nano} is the dominant term whenever the receiver is more than a few millimetres from the source [25, 44]. The point of an edge-stack framing is to drive τ_{infer} as close to zero as the BCI's compute can sustain, because the alternative is to absorb a round-trip to the cloud which is typically two orders of magnitude larger than τ_{nano} itself.

Table 6
Representative latency contributors for a closed-loop IoBNT intervention. τ_{nano} dominates; the cloud round-trip is included to show why edge-local inference is not a micro-optimization but a structural requirement for sub-second control loops.

Component	Symbol	Typical range
Molecular sensing (BNT)	τ_{sense}	10 ms–100 ms
Diffusion / nanonetwork	τ_{nano}	1 s–10 s
BCI transduction + conversion	τ_{BCI}	1 ms–10 ms
MCU inference (quantised)	τ_{infer}	1 ms–50 ms
Actuation (drug / stimulation)	τ_{actuate}	10 ms–100 ms
Edge total (worst)	$\sum_{\text{edge}}$	≈10.26 s
Cloud round-trip (RTT)	τ_{cloud}	50 ms–200 ms
Cloud total (worst)	$\sum_{\text{cloud}}$	≈10.46 s

Values are order-of-magnitude estimates traced to: τ_{nano} – diffusion delay at mm–cm separation [25, 44]; τ_{BCI} – photodiode or BioFET front-end rise time plus ADC conversion [13]; τ_{infer} – Cortex-M4 at 80 MHz, quantised MLP or depthwise-separable CNN (section 9); τ_{actuate} – piezoelectric micropump or optogenetic stimulation pulse train [17]; τ_{cloud} – measured 4G/5G RTT to a regional MEC node. Where $\tau_{\text{nano}} \gg \tau_{\text{cloud}}$ the cloud round-trip is acceptable; for sub-second clinical loops where τ_{nano} is minimized by anatomical proximity, every millisecond of post-transduction latency counts.

6.3. Energy at the boundary

The BCI is also the layer at which the energy budget changes regime. Layer 1 (BNTs) lives on harvested chemical and mechanical energy in the nanowatt range; Layer 4 (edge nodes) operates at hundreds of milliwatts or more, drawing from a wearable battery or a wired supply. The BCI itself, when implemented on an energy-harvested microcontroller, sits in the microwatt-to-milliwatt band, and is the layer at which every additional inference cycle has a measurable cost. This is the conceptual home of the TinyML chapter (section 9): the BCI is the place at which compressing a model from 10 MB to 100 kB makes the difference between an implant that lasts a year on harvested power and one that does not.

6.4. Trust boundary

Finally, the BCI is the natural trust boundary of an IoBNT system. Information that has not yet crossed it remains in the molecular domain, where it is hard to exfiltrate; information that has crossed it is now in conventional digital storage, where it is vulnerable to the entire taxonomy of cyber-attacks. Zafar et al. [46] treat this transition at length. Section 12 draws the implications.

The remainder of the paper takes the BCI's quadruple role – first inference site, latency choke-point, energy regime change, trust boundary – as the recurring object of design. Inference architectures are evaluated by how much of their work they push onto the BCI (sections 7–9); the orchestration plane treats BCIs as the natural unit of in-body network management (section 10); the energy and security chapters (sections 11–12) recompute their budgets layer by layer.

7. Edge intelligence I: machine learning for molecular and biosignal decoding

The most thorough recent survey of neural networks for molecular communication is that of Torres Gómez et al. [45]. This section does not duplicate its coverage. Instead, we organise the machine-learning landscape along an explicit *task* × *model-class* matrix (table 7) and report the constraints that follow from running each cell of that matrix at the BCI or edge tiers of section 4. Federated training and TinyML deployment are deferred to Sections 8 and 9 respectively.

Table 7
Model-class × inference-task matrix for machine-learning approaches to molecular and biosignal decoding in IoBNT (section 7). Marks: ■ the class is a documented strong choice for the task; ⊡ plausible but under-explored; □ poor fit. Representative references in table cells where available.

Inference task	Classical / threshold	MLP (dense)	CNN	LSTM / RNN	Transformer	RL	PINN / physics-aware	Semantic / joint-S-C
Symbol detection	■[25]	■	⊡	■[41]	⊡[45]	□	⊡	⊡
Channel-impulse-response / distance	⊡	⊡	⊡	■[41]	⊡	□	■[21]	□
ISI mitigation	⊡[25]	⊡	⊡	■	■[45]	□	⊡	□
Anomaly detection	■	■	■[20]	⊡	⊡	□	□	⊡
Semantic decoding	□	□	⊡	⊡	■	□	□	■[11]
Digital-twin assimilation	□	⊡	■[20]	⊡	⊡	⊡	■[21]	□
Adaptive release / threshold	⊡	⊡	□	□	□	■	□	□

7.1. Inference tasks

Six tasks recur in the literature:

1. *Symbol detection.* Given a noisy molecule-count time series, decide which symbol was transmitted. Classical maximum-likelihood detection with ligand receptors [25] sets a strong baseline that modern deep models must beat in either accuracy or latency.
2. *Channel impulse response estimation.* Estimate the diffusive (and possibly advective) channel transfer function so a downstream detector can compensate for it. Schottlender, Schäfer and Veiga [41] present an RNN-based distance estimator for branched MC channels; channel estimation and distance estimation are essentially the same regression problem.
3. *Inter-symbol-interference (ISI) mitigation.* Memory-aware equalisation; the closest analogue to wireless equalisation in conventional radio. The dominant tools are LSTM/transformer architectures that condition on a sliding history window.
4. *Anomaly detection.* Recognise that the receiver is seeing a biomarker pattern outside the normal envelope. Useful both for clinical alerts and for detecting adversarial inputs; the model-class spectrum runs from one-class support-vector machines to deep autoencoders.

5. *Semantic decoding.* Move beyond bit-level recovery to recovering the *meaning* of a message at the receiver, exploiting joint source–channel structure. Cai and Akan [11] introduce semantic learning to molecular communication; this is one of the youngest sub-areas in the field and the one that maps most cleanly onto edge-tier deployment, because the "semantic" representation is what the downstream application consumes anyway.
6. *Digital-twin assimilation.* Fuse IoBNT signals with a digital twin of the patient or sub-system. Jamshidi et al. [21] build a physics-informed neural network (PINN) for in-body bio-nano digital twins; Jamshidi, Hoang and Nguyen [20] couple a CNN to a federated digital-twin pipeline at the biotechnology scale.

7.2. Model classes and their constraints

The model classes used in the IoBNT-ML literature recapitulate the catalogue familiar from the wider deep-learning community, but each acquires IoBNT-specific constraints.

- *Classical and threshold detectors* remain the parameter-cheap baseline. They run at any tier; their accuracy ceiling is the main limitation [25].
- *Dense MLPs* are the workhorse for symbol-by-symbol detection. Parameter counts in the kilobyte range are achievable and compatible with BCI-tier deployment.
- *CNNs* extract spatial or time-frequency features and form a common front end of IoBNT digital-twin pipelines, where Mohammad's framework uses CNNs for machine-vision-style pattern recognition over image-based biological data [20].
- *LSTMs / RNNs* handle the long memory imposed by MC's heavy-tailed ISI [41].
- *Transformers* have arrived in IoBNT only recently; Torres Gómez et al. [45] catalogue early applications. Their parameter-count footprint is a poor fit for BCI deployment, which restricts them, in practice, to the edge tier.
- *Reinforcement-learning agents* appear in adaptive threshold and adaptive release-rate problems; they are awkward to deploy at the BCI because online exploration is unsafe in vivo.
- *Physics-informed neural networks* embed the diffusion PDE in the loss, achieving strong generalisation with limited training data [21]. Their training is expensive but their inference is cheap, which suits edge-tier inference with cloud-tier training.
- *Semantic / joint-source-channel networks* [11] are the newest class and the one most aligned with edge deployment, because their output is already in the form the application consumes.

7.3. Three constraints that the substrate imposes

Section 5 listed five substrate-level constraints; for machine learning at the BCI and edge tiers, three of them dominate.

1. *Training data are scarce and patient-specific.* An individual patient's intra-body channel cannot be replicated in a laboratory; pre-training on simulated data and fine-tuning in vivo is the practical workflow. This is the central motivation for federated learning (section 8).
2. *Inference latency is hard-bounded.* An arrhythmia-management loop needs end-to-end latency under 100 ms; the model size and tier placement must respect that budget.
3. *Inference energy is hard-bounded.* A wearable patch cannot run a 100 MB transformer continuously; TinyML profiles (section 9) become mandatory whenever BCI-tier inference is the chosen design point.

7.4. What is missing

Two gaps stand out. First, no public IoBNT-ML benchmark exists. Reported accuracies in the literature [11, 41, 45] use different datasets, different channel models, and different evaluation protocols, so

direct comparison is impossible. Section 13 treats the benchmarking gap; the agenda in section 15 predicts the first public benchmark will arrive within two years. Second, the field has no consensus on the BCI-vs-edge tier-placement question for any given model class. The matrix in table 7 is an attempt to surface that question; the answer will be a function of how the energy chapter (section 11) closes.

8. Edge intelligence II: federated learning at the bio-edge

The deduplicated corpus contains only a handful of papers that combine federated learning (FL) with IoBNT. The single explicit example – Jamshidi, Hoang and Nguyen [20] – couples a convolutional neural network for image-based biological-data classification with federated training across edge gateways and a digital-twin assimilation step. Two other papers treat federated training as part of broader privacy or platform discussions [10, 34]. The combined weight of the FL-for-IoBNT literature is, generously, three full papers. We argue that this is among the most important under-research in the field.

The argument runs as follows. Section 5 listed five substrate-level constraints that any IoBNT inference system must respect; among them, "non-stationarity" and "training data are scarce and patient-specific" (section 7) are the two that FL is built to address. Each patient's intra-body channel is its own probability distribution; centralized training on de-identified data is both legally fraught and statistically inadequate, because de-identification does not eliminate the substantial individual variance that drives detection error [3]. The natural workflow is therefore: each edge gateway (Layer 4 of section 4) trains a local model on its patient's data; gateways share gradients or model deltas rather than raw biosignals; a server aggregates them into a global model that is redistributed for the next round. The privacy of raw data is preserved *by substrate* (the biosignal never leaves the gateway) rather than *by protocol* alone.

8.1. FL background: what the canonical methods offer IoBNT

Although FL-for-IoBNT is nearly empty as a literature, the canonical FL methods developed for mobile and healthcare edge systems since 2016 transfer directly to the bio-edge setting, and we cite them here to supply the vocabulary the field will need:

1. *Federated averaging (FedAvg).* McMahan et al. [33] introduced the workhorse algorithm: clients train locally for several SGD steps, then upload model updates to a server that averages them. FedAvg works well when client data are approximately IID and clients share similar compute budgets. Both assumptions break in IoBNT: each patient's biosignals are heavily non-IID (section 7), and BCI-tier clients have highly variable harvested-power availability (section 11).
2. *Proximal and variance-reduced methods.* Li et al. [29] add a proximal term to the local objective, preventing client updates from drifting too far from the global model under heterogeneous compute budgets – exactly the condition of an energy-harvesting BCI that may complete only a fraction of the expected local epochs. For IoBNT, FedProx (or its descendants) is a more natural baseline than vanilla FedAvg.
3. *Secure aggregation.* Bonawitz et al. [8] showed that a server can aggregate encrypted model updates without ever inspecting individual contributions. For IoBNT gateways handling protected health information, secure aggregation is the cryptographic complement to privacy-by-substrate: the gateway does not see raw signals; the server does not see individual model updates.
4. *Differential privacy.* Abadi et al. [1] introduced differentially-private SGD (DP-SGD), which bounds the information leakage from any single training example by clipping per-example gradients and adding calibrated noise. Biosignal data carries re-identification risk even in feature-representation form; DP-SGD provides the quantifiable privacy guarantee that regulatory submissions require. The cost is a drop in model accuracy proportional to the privacy

budget ε; for biosignal classification tasks, typical $\varepsilon \in [2, 8]$ degrades accuracy by 1–5 percentage points relative to non-private training, a trade-off that clinical regulators may accept but that IoBNT FL benchmarks do not yet measure.

5. *Personalized FL and meta-learning.* When patient-specific distributions diverge too far for a single global model to serve, personalized FL splits the model into shared and local components. Finn, Abbeel and Levine [15] introduced MAML, which learns an initialization from which a small number of gradient steps adapts to a new client. For IoBNT this maps naturally onto the implant-calibration problem: a global model captures population-level biosignal structure; fine-tuning with a few patient-specific samples adapts it to the individual's channel.
6. *Small-N convergence.* Standard FL convergence analyses assume hundreds to millions of clients. A pivotal clinical trial for an implantable IoBNT device may involve tens of patients. Convergence under small client populations is fragile – a single outlier can bias the global model – and the field lacks convergence guarantees parameterised by cohort sizes typical of medical-device trials. Byzantine-resilient aggregation rules (Krum, trimmed mean, median-of-means) mitigate this but have not been evaluated on IoBNT-relevant non-IID distributions.

8.2. Cross-silo versus cross-device topology

FL deployments are conventionally classified as cross-silo (a small number of large clients, e.g. hospitals) or cross-device (a large number of small clients, e.g. smartphones). IoBNT FL admits both. A cross-silo design federates hospital edge servers, each holding the aggregated traces of many patients; a cross-device design federates per-patient wearable gateways. The cross-silo design is the more practical near-term option – hospitals have stable identities, persistent storage, and legal mandates that make them natural FL clients – but the cross-device design is the more interesting long-term one, because it removes hospitals as a necessary trust intermediary:

1. *Severe non-IID data.* An individual patient's biosignal distribution is biased by age, comorbidities, time of day, and the specific implant's calibration. Standard FedAvg-style aggregation underperforms on heavily non-IID clients; the IoBNT setting is arguably the most heterogeneous FL benchmark candidate yet proposed.
2. *Small client populations.* A pivotal trial for an IoBNT biosensor may enrol a hundred patients, not the millions of devices that benchmark FL papers assume. Convergence guarantees that rely on large client counts do not apply; the field needs FL convergence analyses parameterised by small N.
3. *Byzantine resilience.* A compromised gateway is not merely a noisy participant; it can pursue targeted poisoning of the global model. Byzantine-resilient aggregation – Krum, trimmed-mean, median-of-means – is therefore a precondition, not an optimization. Zafar et al. [46] catalogues the threat surface; the FL literature outside IoBNT has matured defences that IoBNT FL deployments will need to import.

8.3. Privacy-by-substrate and privacy-by-protocol

Two privacy mechanisms are available. *Privacy-by-substrate* is what we have for free: a biosignal that never leaves the BCI (section 6) cannot be exfiltrated. *Privacy-by- protocol* is what FL provides: even if a model update leaks, the underlying training data is not recoverable in closed form. Differential privacy adds quantifiable bounds at the cost of additional noise; secure aggregation hides individual updates from the server at the cost of additional cryptographic computation. The combined regime should be: keep raw biosignals at the BCI; expose only post-feature representations to the gateway; train locally and federate with both differential privacy and secure aggregation. No public IoBNT FL benchmark currently allows this regime to be measured directly.

8.4. Coupling to digital twins

Jamshidi, Hoang and Nguyen [20] couple their federated CNN to a microorganism / bioprocess digital twin in the biotechnology industry, and explicitly identify patient-specific DTs as a target application; Jamshidi et al. [21] take a related route via physics-informed neural networks for in-body IoBNT digital twins. The digital twin is the natural object at which the privacy-protected representation lives: a model that captures the patient's biological dynamics without exposing the raw biosignal trace. We expect this pattern – FL among gateways, each holding a private digital twin, with the global model serving as a population-level prior – to become the canonical IoBNT-FL architecture by the end of the decade. The research-agenda chapter (section 15) formalizes this expectation as prediction P5.

8.5. What is missing

Beyond the obvious – more papers – the field lacks three specific artefacts: a public benchmark on which FL convergence under IoBNT non-IID can be measured; a documented FL protocol that respects the IoBNT trust boundary (section 6); and a comparison of cross-silo and cross-device deployments under realistic patient cohort sizes. The agenda predicts the first benchmark within two years.

9. Edge intelligence III: TinyML on harvested power

The deduplicated corpus underlying this survey contains *zero* papers that explicitly combine TinyML – the discipline of running machine-learning inference on microcontroller-class hardware operating at sub-milliwatt power budgets – with IoBNT. The adjacent literature on TinyML for wearables and IoT sensing has matured rapidly since the inaugural MLPerf Tiny benchmark in 2021, but its results have not yet been imported into the IoBNT layer stack. This chapter is therefore the most speculative of the technical chapters in the paper; we open the topic explicitly, catalogue the constraints under which an IoBNT-relevant TinyML profile would have to operate, and identify the results from the wider TinyML community that would, if reproduced at the bio-edge, close the gap.

9.1. Why TinyML matters specifically for IoBNT

Section 6 placed the bio-cyber interface at the energy boundary of the system: nanowatts beneath it, milliwatts above. The natural microcontroller at the BCI – a low-power Arm Cortex-M class part fed by piezoelectric or thermoelectric harvesting (section 11) – will run inside an envelope of roughly 0.1–10 mW. A standard floating-point CNN of even moderate size (a few megabytes of parameters, a few megaFLOPs per inference) exceeds that envelope by two to three orders of magnitude. Three options remain: (i) push inference up to Layer 4 (the edge tier), accepting BCI-to-edge latency; (ii) build a smaller model that fits at the BCI; (iii) split inference across BCI and edge tiers, with an early-exit classifier at the BCI and a fall-through to a larger model on demand. Option (i) is feasible but spends the latency budget the IoBNT loop did not have to spare (section 6). Options (ii) and (iii) are the practical TinyML routes.

9.2. Model footprint targets

A working TinyML profile for IoBNT symbol detection must respect three independent budgets: *parameter count* (limited by Flash storage on the microcontroller), *activation memory* (limited by SRAM), and *inference energy* (limited by harvested power). Reasonable starting points, extrapolated from the MLPerf Tiny keyword-spotting and visual-wake-words benchmarks, are $\leq$100 kB of parameters, $\leq$32 kB of activation memory, and $\leq$1 mJ per inference. None of these targets has been demonstrated against an IoBNT benchmark for the simple reason that no such benchmark exists (section 13); the agenda predicts the first demonstration within two years (section 15, prediction P6).

To make the resource envelope concrete, consider a representative IoBNT symbol-detection workload: 4-ary concentration-shift keying with a channel memory spanning ∼10 symbol intervals. An MLP classifier with input dimension 40 (10 symbols × 4 features each), two hidden layers of 64 units, and a 4-class softmax output requires approximately $(40 \times 64) + (64 \times 64) + (64 \times 4) \approx 6.9$ K multiply-accumulate operations (MACs) per inference. A depthwise-separable 1-D CNN increases this to ∼40 K MACs; a GRU-based detector with a 64-dimensional hidden state reaches ∼100 K MACs.

On a Cortex-M4 at 3.3 V and 80 MHz, one MAC costs roughly 30 pJ. The MLP therefore draws 0.2 µJ per inference, the CNN 1.2 µJ, and the GRU 3.0 µJ. The average power contribution is $P_{\text{avg}} = E_{\text{inference}} \times f_{\text{inference}}$. At one inference per second the MLP draws 0.2 µW, negligible against the 100 µW harvesting floor; at 100 Hz – plausible for cardiac arrhythmia monitoring – the MLP draws 20 µW, the CNN 120 µW, and the GRU 300 µW. The CNN and GRU exceed the self-powered envelope at continuous rates; event-driven duty cycling (section 11) is the mechanism that brings them back within budget. The 1 mJ inference-energy target assumed above is per-inference budget, not per-second average; it allows ∼2 800 MLP inferences per joule, ∼830 CNN inferences, or ∼330 GRU inferences from a single harvested joule. These numbers are order-of-magnitude estimates – the exact MAC count and energy per MAC depend on compiler optimizations, quantization bit width, and silicon process – but they establish that *MC symbol detection is within the energy reach of current TinyML hardware*, provided the inference rate matches the duty cycle that the harvesting source can sustain.

9.3. Three model classes that fit

Three model classes are plausible candidates for the BCI-tier TinyML profile, ordered from least to most ambitious:

1. *Quantised dense detectors.*

Eight-bit or four-bit quantization of a small MLP can deliver acceptable accuracy on the simpler instance of the symbol-detection task. The training pipeline is conventional; the deployment pipeline benefits from TFLite-Micro or equivalent.

2. *Depthwise-separable convolutions for time–frequency features.*

Standard since MobileNet in the wider community; well-supported on microcontroller toolchains; appropriate for the CNN front end of the digital-twin assimilation pipeline of Jamshidi, Hoang and Nguyen [20], restricted to the BCI-tier slice.

3. *Recurrent state with leaky-integrator memory.*

The heavy-tailed inter-symbol interference of MC channels (section 5) is a long-memory phenomenon [25, 41]. Full LSTM blocks are unlikely to fit; leaky-integrator or simplified GRU variants that approximate the memory with a single recurrent state are the plausible compromise. The reservoir-computing literature supplies a theoretical foundation for this approach: Jaeger [19] showed that a fixed, randomly-initialised recurrent reservoir with a trainable linear readout layer – an echo state network (ESN) – captures temporal dependencies without backpropagation through time; Maass, Natschläger and Markram [32] generalised this to spiking neurons under the name liquid state machine (LSM). The ESN pattern is attractive for IoBNT TinyML because the reservoir computation is a single forward pass with fixed weights, eliminating the energy cost of recurrent training while still providing the temporal memory that MC symbol detection requires. Whether an ESN with a microcontroller-appropriate reservoir size (tens to low hundreds of neurons) can match the detection accuracy of a trained GRU on realistic MC channel traces is an open experimental question; the agenda chapter flags it as a target for the first IoBNT-TinyML benchmark (section 15, prediction P6).

9.4. Training and update workflow

A TinyML model deployed at the BCI cannot retrain in place; the energy cost of backpropagation alone exceeds the harvested budget by orders of magnitude. The workflow must therefore be "train at Layer 4 or 5, deploy to Layer 3": the edge tier or the cloud handles training (section 8 federates this across patient cohorts), and the BCI receives compiled, quantised artefacts. Over-the-air updates of microcontroller firmware are an operational requirement, not a nice-to-have; their security model overlaps with the cyberbiosecurity chapter (section 12).

9.5. What needs to happen

To move the TinyML-for-IoBNT field from zero records to a working sub-discipline within the next half-decade, three concrete things need to happen.

1. *A public benchmark* that pairs an IoBNT-realistic symbol-detection task with the MLPerf Tiny resource envelope. Without this artefact, comparison between approaches will continue to be impossible (section 13).
2. *Open hardware platforms* that combine a microfluidic BCI front end [35] with a Cortex-M microcontroller and an integrated piezoelectric harvester. The field currently has the first two; the third is in early prototype [17].
3. *A reference compiler path* from a PyTorch or TensorFlow model to a TinyML deployment artefact, customised for the long-memory and noise structure of MC channels. The general-purpose paths in TFLite-Micro and MicroTVM provide a starting point but do not exploit MC-specific structure.

The agenda chapter (section 15) returns to each of these as dated predictions.

10. Orchestration: Bio-SDN and the in-body control plane

Most surveys of IoBNT stop at the question of how a single biosignal propagates from a BNT to a downstream actuator. The question of how a *population* of in-body devices is managed – configured, slice-allocated, software-updated, retired – rarely appears, and when it does it is usually folded into the catch-all of "the cloud" that sits above Layer 5 in the reference architecture (section 4). We argue that this gap is closed by an analogue of software-defined networking adapted to the biochemical domain, conventionally called Bio-SDN.

10.1. The orchestration problem

A realistic IoBNT deployment in the late 2020s consists of: tens to thousands of BNTs in the body, possibly of multiple cell types or device classes; one or a few bio-cyber interfaces, each attached to a specific anatomical site (a vascular implant, a gut patch, a brain interface); a wearable or fixed edge gateway; and the cloud. The management questions are: which BNT subpopulation should be reporting at any given moment? Which BCI is the primary forwarder for an event? When does the edge gateway hand off to the cloud for a long-horizon decision? How are model updates distributed back through the stack without disrupting an ongoing control loop? Kuscu et al. [26] catalogue the architectural primitives but stop short of an explicit control plane; Kong et al. [23] approach the question from the 6G side, where similar orchestration questions arise for sub-THz small cells.

10.2. Three planes, restated for biology

Software-defined networking conventionally separates an application plane, a control plane, and a data plane. Bio-SDN adopts the same separation with biology-specific interpretations.

1. *Application plane.* High-level policies expressed in clinical or operational language: "alert if the glucose monitor reports out-of-range readings for 15 min consecutively", or "maintain the gut-brain-axis release schedule unless inflammation markers exceed a threshold". These policies are not implementation details; they are the intent that the system as a whole serves.
2. *Control plane.* A logical controller – typically realised at the edge tier – translates application-plane intent into low-level configurations that the BCIs can enforce. The control plane is the natural home for the digital twin (section 8), the federated model aggregation logic, and the policy engines that decide when an event is alert-worthy. Its location at the edge is what gives the system its latency budget; placing it in the cloud collapses the architecture into a conventional telemetry pipeline.
3. *Data plane.* The BNTs themselves, and the BCIs that read them. The data plane *executes*: it samples, transduces, transmits, releases. It does not decide; deciding is the control plane's job.

10.3. In-body network slicing

Network slicing in conventional 5G/6G allocates a subset of network resources to a single service class (low-latency control, high-bandwidth video, low-energy telemetry). IoBNT slicing is the analogous problem at the molecular and electromagnetic boundaries: a slice for a critical control loop (the arrhythmia-management loop with its sub-100 ms budget), a slice for routine telemetry (a glucose monitor reporting every few seconds), a slice for over-the-air software updates (rare, batchable, large). The slices share BCI compute, BCI transduction bandwidth, edge-gateway storage, and network uplink; the control plane allocates among them. Kong et al. [23] discuss the wider 6G coupling.

10.4. Closed-loop and safety primitives

A Bio-SDN data plane includes operations that have no analogue in conventional SDN: triggering the release of a drug, instructing a synthetic cell to enter a quiescent state, programming a kill-switch [9]. The control plane must therefore include safety primitives that conventional SDN does not: *commit-or-revert* semantics on actuation; rate limiting on release commands; and watchdog timers that revert to a safe default if the control plane itself loses contact. The synthetic-biology literature has accumulated the primitive kill-switches; the networking literature has not yet integrated them into an SDN framework. We expect this to be an especially important operational research direction at the boundary of biology and networking between 2026 and 2030.

The control plane communicates with the data plane through a southbound API whose primitives have no direct analogue in OpenFlow or P4. We propose four operations as a minimal set:

- `GET_SENSOR_STATE(bnt_id, modality)` – read the current output of a specified BNT or BCI transducer channel.
- `SET_RELEASE_RATE(bnt_id, rate, duration)` – instruct a drug-release or signalling BNT to emit at a specified rate for a bounded interval.
- `ACTIVATE_KILL_SWITCH(bnt_id)` – trigger an irreversible shutdown of a synthetic-biology BNT.
- `ENTER_QUIESCENT(bnt_id, duration)` – place a BNT into a low-power or metabolically dormant state for a specified interval, preserving it for later reactivation.

All four primitives are idempotent and carry a monotonically increasing sequence number to prevent replay. The southbound channel itself – whether a galvanic-coupled body-area network, a BLE link, or a 6G sub-THz connection – is abstracted beneath this API.

A Bio-SDN control loop transitions through five states:

1. IDLE – no alert is active; the control plane periodically polls `GET_SENSOR_STATE` at a slice-appropriate cadence.

2. MONITORING – a sensor reading crosses a pre-configured threshold; the control plane increases the polling rate and activates additional sensor modalities for confirmation.
3. ALERT_DETECTED – the confirming sensor fusion step classifies the event as actionable; the control plane computes a candidate actuation plan (drug release, stimulation pattern, kill-switch activation).
4. ACTUATION_INITIATED – the actuation command is dispatched to the data plane. The system enters a *commit-or-revert* window: if physiological feedback from the next `GET_SENSOR_STATE` cycle confirms the expected response, the actuation is committed; otherwise it is reverted and the state returns to MONITORING.
5. SAFE_DEFAULT – entered if the control plane loses contact (watchdog expiry) or if a revert fails. The data plane defaults to a pre-registered safe configuration: stop all release, hold quiescent state, continue passive sensing only.

This state machine is deliberately simple; its purpose is to make the safety argument explicit rather than to serve as a production specification. The commit-or-revert window is the key contribution: it acknowledges that actuation in a living system is irreversible in the biological sense, and compensates by requiring physiological confirmation before the command is considered final.

The Bio-SDN control loop spans three mismatched timescales: (i) electronic inference and control-plane logic at milliseconds-to-seconds; (ii) molecular diffusion across a microfluidic channel at seconds; (iii) genetic-circuit response (transcription, translation, protein folding) at minutes-to-hours. A control plane that dispatches a `SET_RELEASE_RATE` to a synthetic-biology actuator must model the delay between command issuance and biological effect – a delay that is not under control-plane control. The safety implication is that the commit-or-revert window must be parameterised per actuator class: seconds for a piezoelectric drug pump, minutes for a quorum-sensing relay, hours for a kill-switch expressed from a promoter. Section 15 prediction P3 formalizes the expectation that the first Bio-SDN-aware simulator incorporating these timescale parameters will appear by the end of the decade.

10.5. Coupling to 6G and beyond

The most concrete 6G integration point is at the BCI-to-edge link. Kong et al. [23] survey the wireless options; for IoBNT specifically, the candidate is a sub-THz or near-body 6G link that is provisioned as a non-terrestrial-network-adjacent service class with low-latency and high-reliability guarantees. The agenda chapter predicts standardisation of this service class within the 6G commercialisation window (section 15, prediction P4).

10.6. What is missing

Bio-SDN exists more as a research proposal than as deployed system. The corpus contains a handful of papers that gesture at the three-plane architecture but none that specify it formally. Three artefacts would change this: a reference control-plane implementation (open source, even if simulator-only); a worked example of slice allocation under a realistic clinical control loop; and a formal safety analysis of commit-or-revert actuation across a Bio-SDN data plane.

11. Energy budgets: harvested power versus compute power

An IoBNT system has two energy budgets that must be reconciled: the *harvested-power budget*, set by the ambient mechanical, thermal, or biochemical energy available to BNTs and the bio-cyber interface (Layers 1–3 of section 4); and the *compute-power budget*, set by the inference, transmission, and actuation work the system performs. Most IoBNT surveys list the two budgets independently but stop short of closing the loop between them. This chapter does the closure explicitly. The conclusion is uncomfortable: under current harvesting technology, the design space in which the loop closes for

in-body machine learning is narrower than the literature acknowledges, and this narrowness is the most immediate forcing function on the TinyML chapter (section 9).

11.1. Harvesting taxonomy

Four families of energy harvesters appear in the IoBNT and adjacent wearables literature, each suited to different anatomical locations and different signal regimes:

1. *Piezoelectric nanogenerators (PENGs).* A piezoelectric film – commonly zinc-oxide nanowires, more recently flexible polymer composites – converts mechanical strain from heartbeat, lung expansion, or skeletal motion into electrical charge. Reported open-circuit voltages are at the sub-volt to single-volt scale; short-circuit currents are in the nanoampere range, giving average harvested powers in the microwatt range. PENGs are well suited to cardiovascular and pulmonary placements.
2. *Triboelectric nanogenerators (TENGs).* Contact electrification between dissimilar surfaces, driven by intermittent motion. TENGs typically harvest more aggressively than PENGs per gram of material but have noisier output and worse fatigue characteristics. Suitable for limb-attached or skin-surface devices.
3. *Thermoelectric harvesters.* Exploit the temperature gradient between the body (around 37 °C) and the surrounding environment. Output is nanowatt-scale at small geometries; suitable for skin-mounted patches but inadequate for in-body devices that thermalise with the body.
4. *Biochemical harvesters.* Glucose fuel cells, microbial fuel cells, and gastric primary cells convert biochemical energy directly into electrical current. Continuous output in the microwatt range is achievable but the biofouling and lifetime constraints are severe. Intra-vascular glucose fuel cells are the most actively researched.

11.2. Compute-power taxonomy

The compute-power consumers, in increasing order of cost, are:

1. *Sensing.* A BNT's sensing operation is essentially free in electrical terms; it spends biochemical energy. The first electrical cost is at the BCI's front-end amplifier.
2. *Transduction.* The BCI's transducer – photodiode, BioFET, redox electrode, graphene plasmonic antenna [13] – draws standing current. Order of magnitude: microwatts to milliwatts.
3. *Inference.* A microcontroller running a quantised dense detector at a few inferences per second draws on the order of 0.1 mW average; a CNN of moderate size draws 1 mW or more (section 9).
4. *Transmission.* A short-range wireless link (Bluetooth Low Energy or sub-GHz backscatter) to the edge gateway draws milliwatts to tens of milliwatts during active radio cycles, with duty cycling driving the average down by orders of magnitude.
5. *Actuation.* Drug release, optogenetic stimulation, or neural stimulation are the spikiest consumers, with brief peaks at hundreds of milliwatts but extremely low duty cycle.

11.3. Closing the loop

A composite picture overlays the plausible harvested-power band against the plausible compute-power band on a single axis (figure 6). The harvested band lies between roughly 1 µW (intra-body thermal) and 100 µW (well-optimised PENG on a heartbeat). The compute band starts at 100 µW for the cheapest quantised TinyML profile and extends to several mW for non-quantised CNNs. The bands overlap, but only barely: the design space in which a wholly self-powered in-body inference loop closes is a narrow strip near 100 µW. Outside this strip, the system needs a rechargeable battery, a wireless power transfer link, or a gateway-tier inference shift that absorbs the BCI-to-edge latency.

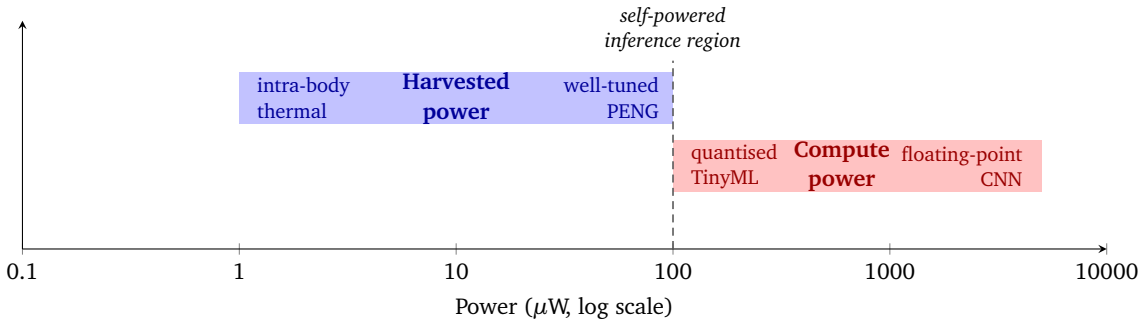

Figure 6: Harvested-power band against compute-power band on a log-scale power axis (section 11). A self-powered in-body inference loop closes only inside the narrow strip near 100 µW where the two bands overlap. Moving the edge-stack design point outside this strip requires either a rechargeable battery, a wireless power transfer link, or a shift of inference up to the edge tier with the associated latency cost.

11.4. Implications

Three implications follow.

1. *Quantization is not optional.* A wholly self-powered BCI-tier inference loop requires quantised models; the TinyML chapter takes this as given.
2. *Duty cycling is mandatory.* Continuous inference is out of budget; the BCI must operate as an event-driven sensor with bursty compute and long sleep intervals.
3. *Wireless power transfer is the single most important bottleneck-relaxation technology.* A reliable, biocompatible wireless power transfer link from a wearable transmitter to the BCI moves the design space upward by an order of magnitude and enables continuous CNN-class inference at the BCI. The agenda predicts this transition by 2030 (section 15, prediction P1).

11.5. Worked example: implantable cardiac BCI

To make the energy closure concrete, consider a representative implantable BCI for continuous arrhythmia monitoring, placed adjacent to the sinoatrial node and powered by a piezoelectric nanogenerator on the heartbeat. Table 8 itemises the power consumers under a duty-cycled regime of one inference per heartbeat (∼1 Hz).

Table 8
Worked energy budget for an implantable cardiac BCI with PENG harvesting. All values are order-of-magnitude and sourced to datasheets or measurement papers.

Component	Active power	Duty-cycled avg.
BioFET front-end amplifier	50 µW	2.5 µW (5 % duty)
12-bit SAR ADC	30 µW	1.5 µW (5 % duty)
Cortex-M4 MLP inference	960 µW	0.96 µW (0.1 % duty)
AES-128 encryption (HW accel)	10 µW	0.5 µW (5 % duty)
BLE TX (short packet, 1 Mbps)	5 mW	5 µW (0.1 % duty)
Quiescent / sleep	0.5 µW	0.45 µW (90 % duty)
Total average		≈10.9 µW

Sources: BioFET – Civas et al. [13] and representative commercial amperometric front-end datasheets; Cortex-M4 MLP – 30 pJ/MAC at 80 MHz, ∼7 K MAC/inference (section 9); AES-128 – hardware-accelerated on STM32L4 at 10 MHz; BLE TX – nRF52840 at 0 dBm with connection-interval optimization; PENG output – 10 µW–100 µW on heartbeat, placement-dependent [17].

The budget closes – barely. A well-positioned PENG on a strong heartbeat delivers 50 µW–100 µW average, giving ∼5–10× headroom above the 10.9 µW draw. But this headroom vanishes if: (i) the

PENG is sub-optimally placed or the patient has low cardiac output; (ii) the inference rate rises above 1 Hz; (iii) the BLE link requires retransmissions; (iv) the analog front-end needs higher gain to compensate for electrode-tissue impedance drift. The design point is viable but marginal – exactly the condition that forces quantisation, event-driven operation, and aggressive duty cycling.

11.6. Costs the literature undercounts

Four energy costs are routinely omitted from IoBNT system sketches and should be included in any future benchmark.

1. *Analog front-end noise vs. power.* A BioFET or instrumentation amplifier with input-referred noise below 1 µV RMS at 1 kHz bandwidth draws substantially more than 50 µW. The noise-power trade-off is well characterized in the biomedical circuits literature but absent from IoBNT energy models.
2. *Cryptographic standing cost.* Secure boot, firmware signature verification, and key storage impose a standing energy cost even when no data is being transmitted. On a Cortex-M0+ without hardware acceleration, AES-128 in software draws 50–100 µW during active cycles; elliptic-curve operations for key exchange are 10–100× more expensive.
3. *Wireless TX with real-world overhead.* The 5 mW active TX figure assumes a clean channel and a single short packet. Connection establishment, channel sensing, acknowledgment, and retransmission can multiply the duty cycle by 2–5× in a body-area network with variable attenuation.
4. *Thermal safety ceiling.* Tissue safety standards (ISO 14708-1 for active implantable devices) limit the electromagnetic and thermal power deposited in tissue to $<10\,\mathrm{mW\,cm^{-2}}$. An implant that combines a power-hungry front-end, a CNN accelerator, and a BLE radio may approach this limit even if harvested power is available. The thermal constraint is an independent cap that the energy-closure figure (figure 6) does not yet represent.

12. Cyberbiosecurity at the bio-edge

Cyberbiosecurity is the discipline that mitigates risks at the intersection of biological and digital systems. In an IoBNT deployment, this intersection lies precisely at the bio-cyber interface and the edge gateway (sections 6 and 4); both are inherited from conventional information security, and both acquire additional failure modes from biology. Zafar et al. [46] provides the canonical treatment; Rana et al. [38] update it with attention to the 2024–2025 threat landscape. This chapter does not duplicate those reviews; it foregrounds the failure modes that arise specifically from edge-tier deployment.

12.1. Why genomic and biomarker data is special

Three properties distinguish biological data from conventional personal data, with consequences for what a breach actually costs:

1. *Permanence.* Unlike a password or a credit-card number, a person's genome and the bulk of their biomarker baselines cannot be reset. A breach that exposes them exposes them for the patient's lifetime, and – via close-relative inference – exposes related individuals as well. The cost of a breach is therefore measured over a decades-long horizon, not a quarterly remediation cycle.
2. *Latent identifiability.* "De-identified" biomarker traces are routinely re-identifiable under modest auxiliary information, and certainly so when the underlying biological signal is rich enough to support a digital twin (section 8). Conventional anonymization is inadequate; cryptographic and differential-privacy mechanisms are required as a matter of routine.

3. *Dual-use potential.* Biomarker patterns identifying disease susceptibility, drug response, or neurological state can be used both to treat the patient and – by an adversary with synthesis capability – to design biological agents targeted at specific individuals or populations. The digital-to-biological conversion pathway, while not yet operationally trivial, is no longer hypothetical.

12.2. Threat model

Table 9 recasts the STRIDE framework – spoofing, tampering, repudiation, information disclosure, denial-of-service, elevation of privilege – onto the five layers of the reference architecture (section 4). The most consequential cells are:

- *Tampering at the BNT layer.* An adversary with access to the supply chain or to the synthetic-biology fabrication pathway can substitute a tampered BNT design. Bonnet et al. [9] establishes that integrase-based biological logic is implementable; the same primitive is, in principle, weaponisable. Genetic kill-switches built on this primitive are the partial defence.
- *Information disclosure at the BCI.* The BCI is the first electronic location at which biological data can be exfiltrated by conventional cyber means. The trust boundary of section 6 concentrates the risk here; encryption-at-rest, encryption-in-transit, and secure-enclave inference are the mitigations.
- *Tampering and elevation of privilege at the edge tier.* A compromised edge gateway can issue commands to the BNT population via the Bio-SDN control plane (section 10). The threat is open-ended: misdiagnosis, denied therapy, or adversarial therapy delivered to a healthy patient. Watchdog timers, commit-or-revert semantics, and federated cross-validation across multiple edge nodes are the proposed defences.
- *Information disclosure during federated training.* Even with privacy-by-protocol federated learning (section 8), gradient-inversion attacks can recover training data under certain conditions. The defence is differential privacy with quantified bounds, plus secure aggregation.

12.3. Layered defences

Three defences propagate across the reference architecture:

1. *Encryption-by-design.* Every channel above the BCI must encrypt at rest and in transit. This is not negotiable, but the choice of cipher suite needs to respect the BCI's energy budget (section 11): hardware-accelerated AES is the practical baseline.
2. *Zero-trust at every layer.* Each component authenticates every request, regardless of origin. This is operationally expensive for BCI microcontrollers but unavoidable; the cryptographic standing cost is part of the BCI energy budget.
3. *Cross-validation across edge nodes.* A single edge gateway's diagnosis is treated as a single classifier vote; consequential actuations require concurrence from either a second edge or the cloud, with bounded staleness.

12.4. Dual-use research and disclosure

Sections of the IoBNT research literature touch on capabilities that have offensive applications: precise targeted drug delivery becomes targeted toxin delivery; gut-brain-axis modulation [31] becomes behavioural manipulation. The field needs a disclosure norm comparable to coordinated vulnerability disclosure in software security: a publication pathway that allows defensive results to be shared without publishing weaponisable specifics. No such norm currently exists, and the agenda chapter (section 15) does not predict one without an outside trigger.

Table 9
STRIDE threat matrix mapped onto the five-layer bio-edge reference architecture of section 4. Each cell lists the most-likely attack vector and the recommended mitigation. The bio-cyber interface row concentrates the highest-risk cells because the BCI is the trust boundary (section 6).

Layer	Spoofing	Tampering	Repudiation	Information disclosure	Denial of service	Elevation of privilege
Perception (BNT)	mimicry cell	tampered synthesis [9]	n/a	sample exfiltration	lysis attack	kill-switch override
Nanonetwork	false sender ID via mod. molecule	release-rate tampering	n/a	diffusive sniffing	molecular jamming	n/a
Bio-cyber interface	false BCI identity	firmware tampering	audit-log forge	raw-trace exfil	front-end overload	privilege escalation in MCU
Edge	gateway impersonation	model poisoning during FL [20]	log repudiation	gradient inversion	ransomware on gateway	orchestration plane takeover
Cloud	compromised TLS	data-at-rest tampering	access-log forge	database leak	DDoS on telemetry ingest	admin credential theft
Recommended mitigation	device PKI	supply-chain attestation and kill-switch	append-only signed logs	encryption-by-design and differential privacy	rate-limiting and watchdog	zero-trust authorization

12.5. What is missing

The single most important missing artefact is a published threat-model document with the authority of a community standard. Zafar et al. [46] is the closest existing object, but it predates the federated-learning and digital-twin developments of 2024–2025. A successor document with explicit attention to the edge tier, the FL trust model, and dual-use disclosure norms is the natural next step. The agenda lists IEEE 1906.x evolution toward a security amendment as prediction P3.

13. Standards, interoperability, and evaluation

The IoBNT field has standards, but their reach is narrower than their existence suggests. The IEEE 1906.1 recommended practice for nanoscale and molecular communication framework (published 2015, with the 1906.1.1 data-model amendment in 2020) establishes a common terminology – message carrier, perturbation, field, motion – and a small set of objective metrics (specificity, sensitivity, message lifetime). It is the only substantial standardisation effort in the area and is broadly acknowledged across the architectures and surveys cited so far in this paper. What it does not do is define a benchmark, a reference implementation, or an interoperability profile against which competing approaches can be compared head to head. Two of the three white-space subtopics of this survey are blocked by that absence: TinyML and federated learning both require shared evaluation harnesses to make progress comparable across groups.

13.1. What IEEE 1906.x establishes

The framework's contribution is conceptual: it provides a common vocabulary and a small set of named metrics. The vocabulary has penetrated the literature; the named metrics have not, in the sense that few experimental papers in the present corpus report quantitative values against the 1906.1 metric suite. The result is a community that shares terminology but not measurement discipline.

The standard does not specify any algorithm, any data format, any test bench, or any interchange protocol.

13.2. The benchmarking gap

Reported accuracies in the machine-learning-for-IoBNT literature (section 7) use different channel models, different datasets, and different evaluation metrics, with the consequence that two papers reporting "95 % detection accuracy" may not be making the same claim [11, 41, 45]. This is arguably the most important weakness of the current IoBNT literature, and it is the artefact most amenable to community fix.

A useful benchmark would specify:

- a reference channel model (or a small set of them);
- one or more training datasets with declared train/validation/test splits;
- evaluation metrics aligned with both the IEEE 1906.1 named set and the wider machine-learning community's conventions;
- a resource envelope corresponding to the BCI tier (TinyML profile, section 9) and a separate envelope corresponding to the edge tier;
- a leaderboard or comparable mechanism for ongoing submissions.

The MLPerf Tiny effort is the structural template. The IoBNT community has the simulators and microfluidic testbeds [17, 35] to populate a benchmark; what it lacks is the convening function. The agenda chapter (section 15) predicts the first published benchmark within two years (prediction P7).

13.3. Open-data agenda

Even before a benchmark exists, three intermediate artefacts would move the field substantially:

1. *A shared MC channel dataset* captured from microfluidic testbeds, with multiple repeats per operating condition, and released under a permissive license.
2. *A shared digital-twin specification* for the most common in-body sites (vascular, gut, brain). The physics-informed neural network programme [21] provides one starting point.
3. *A shared FL evaluation protocol* that captures IoBNT-specific non-IID under small client populations (section 8).

Appendix C catalogues the simulators (BNSim, Simbiotics, MoNaCo) and tepanstbeds available; the gap is not the infrastructure, it is the convention.

13.4. Beyond IEEE 1906

Three plausible directions for the standards body to evolve:

1. *A 1906.2 security amendment.* Specifying the trust boundaries of section 12, the kill-switch primitives of section 10, and the disclosure norms for dual-use research. The agenda predicts this amendment within five years (section 15, prediction P3).
2. *A 1906.3 measurement profile.* Specifying experimental protocols and reporting requirements for the 1906.1 metric suite, so that quantitative claims become comparable.
3. *Coupling to ITU-T and 3GPP.* The 6G coupling discussed in section 10 requires parallel standardisation activity in the wireless community. The agenda predicts a study item on a molecular-electromagnetic service class within the 6G commercialisation window (prediction P4).

13.5. What is missing

The missing artefacts of this chapter are not papers; they are convening efforts. The community has the technical capacity to produce a benchmark, an open dataset, and a measurement profile; what it lacks is the institutional substrate that would commit a group to maintaining them. The agenda chapter treats this gap explicitly.

14. Application landscape

The technical chapters that precede this one treat IoBNT as if it were a single system class. In practice, the field is differentiated by application: a vascular drug-delivery system, a continuous glucose monitor, an environmental toxin sensor, and a phytobiome crop-management network share a substrate (the layer stack of section 4) but differ sharply in their latency budgets, regulatory regimes, and cost-of-failure profiles. This chapter catalogues the four application classes that the corpus recognises and identifies, for each, the edge-stack design points that follow from its operational requirements. The canonical application taxonomy of Kuscu and Unluturk [27] (biomedical, smart-agriculture, environmental) provides the base classification that this chapter refines with an edge-stack lens and an additional industrial category.

14.1. Healthcare

Healthcare dominates the corpus: 33 of the 311 deduplicated entries fall into the healthcare bucket (figure 4), and an additional 12 cluster under drug delivery and theranostics. Within healthcare, four sub-applications stand out.

1. *Continuous metabolic monitoring.* Glucose monitoring is the canonical case [2, 10]. The control loop is comparatively slow (seconds to minutes), the metric is well established (mg/dL or mmol/L), and the regulatory pathway already exists. The next-generation prediction is a true closed-loop artificial pancreas with an in-body Bio-SDN control plane (section 10, agenda prediction P2).
2. *Vascular and cardiovascular IoBNT.* Lee et al. [28] report system designs and prototypes for IoBNT inside blood vessels. Latency budgets here are stricter: an arrhythmia-management loop has at most a few hundred milliseconds. The BCI is necessarily on or near the vascular wall, and the inference must run at the BCI (section 6); the edge tier is too far away.
3. *Spinal-cord injury and brain–machine interfaces.* Akan et al. [3] treat IoBNT in the spinal-cord-injury context, with information-theoretic bounds for in-axon signalling. This sub-application is the closest to a traditional brain–machine interface and inherits much of its regulatory and ethical machinery.
4. *Theranostic drug delivery.* Chude-Okonkwo et al. [12] review molecular communication for targeted drug delivery; the corpus contains a substantial drug-delivery sub-cluster. The relevant edge-stack design point: the BCI fires the release upon detection, with a commit-or-revert primitive (section 10) limiting the consequences of a misfire.

14.2. Environmental

Environmental IoBNT is small in the deduplicated corpus – only two entries cluster here directly – but the case studies are methodologically interesting because they admit external prototyping (an outdoor or laboratory deployment rather than an in-body one). The canonical example is whole-cell biosensing for heavy-metal detection (e.g., arsenic or lead) using engineered bacteria; the BCI in that deployment is a microfluidic chamber with a fluorescent or electrochemical readout, and the edge tier is a handheld unit or ruggedised tablet. Microfluidic testbeds [17, 35] are the natural development platform.

14.3. Agricultural and phytobiome

Agriculture is the youngest application thread in the corpus, with Babar and Akan [5] framing the Internet-of-Everything-for- agriculture proposition and Gulec et al. [16] naming the phytobiome-specific subdiscipline. The fundamental observation is that plants and soil micro-biota communicate biochemically at scales and tempos that overlap with the human-body IoBNT case. The regulatory pathway is more permissive than for in-body IoBNT; the agenda predicts commercial pilot deployments by 2030 (section 15, prediction P9).

The interesting edge-stack consequence: agricultural deployments admit a denser BCI/edge ratio than healthcare ones (one edge gateway per field rather than per patient), which changes the federated learning topology to a cross-silo regime (one farm = one silo) naturally aligned with the FL framework of section 8.

14.4. Industrial

Industrial IoBNT is the smallest application class. Plausible deployments include bioprocess monitoring in fermentation tanks, distributed sensing in environmental monitoring stations attached to manufacturing facilities, and bioremediation tracking. The corpus does not yet support a dedicated treatment; we note the application class for completeness and defer.

14.5. Comparative requirements

Table 10 maps the four application classes against five requirement dimensions. The matrix is deliberately coarse – environmental and industrial cells rely on fewer primary sources than healthcare and agriculture – and is offered as a structured starting point rather than a definitive assessment.

Table 10
Application-class requirement matrix. Marks: ●●● = dominant concern; ●● = moderate; ● = minor or not yet established. Cells marked with * have weak primary-source coverage in the current corpus.

Application class	Latency	Privacy	Energy	Cost of failure	Regulatory burden
Healthcare (implant)	●●●	●●●	●●●	●●●	●●●
Healthcare (wearable)	●●	●●●	●●	●●	●●
Agriculture	●	●	●●	●	●
Environmental	●	●	●	●	●*
Industrial	●●	●●*	●	●●	●●*

Sources: healthcare – Akan et al. [3], Bulasara et al. [10], Meenambika et al. [34]; agriculture – Babar and Akan [5], Gulec et al. [16]; environmental and industrial columns draw on the general IoBNT architecture literature [26, 27] and should be treated as indicative. The healthcare implant/wearable split reflects differing latency and energy constraints documented in sections 6–11.

The qualitative pattern is straightforward: healthcare has the strictest latency budget, the slowest regulatory pathway, and the highest cost-of-failure; agriculture has the most relaxed regulatory pathway and the most favourable BCI-to-edge device ratio; environmental has the most flexible deployment options; industrial is the least developed and the least represented in the current corpus.

14.6. Tier-placement implications

The four application classes induce different tier-placement choices for inference. Healthcare with strict latency budgets pushes inference to the BCI; healthcare with relaxed latency budgets permits inference at the edge tier with no observable user-side penalty. Environmental and agricultural deployments default to edge- tier inference because their BCIs are typically not power-bound in the same way as in-body BCIs. Industrial deployments split between edge-tier inference (for local control) and cloud-tier inference (for aggregated process analytics). The reference architecture of

section 4 accommodates all four patterns; the right tier-placement is an application-level decision, not an architectural one.

15. The IoBNT-Edge research agenda 2026–2035

A persistent weakness of survey papers in young fields is that they close with visionary language that cannot be falsified. The reader is told that a technology will be "transformative", that integration "will continue", that "boundaries will dissolve". Five years later the literature has moved on and no one returns to ask whether the predictions were right; there is no way to ask, because the predictions were never specific enough.

This section is the centrepiece of the survey, and we deliberately hold it to a higher standard. We propose ten predictions for the IoBNT-Edge field between 2026 and 2035. Each prediction states a *date*, a *measurable metric*, a *mechanism* that would produce the predicted outcome, and a *falsifier* – a counter-condition that, if observed by the stated date, makes the prediction wrong. Table 11 is the structured artefact; this prose introduces the methodology, walks through the predictions in thematic clusters, and explains how subsequent work can use the agenda.

15.1. Methodology

Each prediction must satisfy four properties:

1. *Specific date.* A calendar year, not "soon". A reader in 2030-12-31 must be able to look at the table and decide whether the prediction was right.
2. *Measurable metric.* A quantity, an event, or a documented filing – an artefact that can be checked in a public registry, an indexed paper, or a regulatory database. Vague qualitative claims are excluded.
3. *Mechanism.* A causal chain referencing existing primary work that makes the prediction plausible. A prediction with no mechanism is a guess; a prediction with a mechanism is a research bet.
4. *Falsifier.* A condition under which the prediction is wrong. The falsifier is checkable independently of the predictor. It need not have a positive sign – "no demonstration meeting the envelope by date X" is a valid falsifier.

Predictions that satisfy all four properties are research bets that the community can collectively pursue, refute, or refine. Predictions that satisfy only three are aspirations; they are excluded from this list.

15.2. Cluster A – Edge-tier intelligence (P1, P5, P6)

The three white-space subtopics of this survey – edge intelligence, federated learning, and TinyML on harvested power – generate the three predictions that the present authors consider most achievable within a half-decade. Each is bottlenecked on the missing benchmark of P7 rather than on any technical impossibility.

Prediction P1 (section 15, table 11) turns the energy-closure argument of section 11 into a dated claim. The harvested-power and compute-power bands overlap at roughly 100 µW; an implantable BCI running quantised TinyML inference falls inside that overlap if (and only if) the inference cost is brought below 1 µW average. Achieving this requires a combination of aggressive quantization, event-driven duty cycling, and a low-power microcontroller; none of these is technically novel, but their joint demonstration in an implant is, in 2026, still unpublished.

P5 closes the federated-learning gap. Once a public benchmark exists (P7), the routine machinery of FL convergence under non-IID applies; the open question is whether the small client populations of IoBNT trials admit useful global models. We predict yes, with non-trivial tightening of differential privacy bounds compared to the benchmark-FL defaults.

P6 closes the TinyML gap. The adjacent TinyML literature has shown that the MLPerf Tiny resource envelope is achievable for harder tasks than MC symbol detection; the gap is dataset, not algorithm.

15.3. Cluster B – Orchestration, networking, standards (P3, P4, P7)

The second cluster concerns the artefacts that the community collectively must produce: standards, benchmarks, study items in adjacent standards bodies. None of these emerges from a single research paper; each requires sustained institutional effort.

P3 names the standardisation evolution that section 12 argues for. A security amendment to IEEE 1906 – which the present authors refer to informally as 1906.2 – would specify trust boundaries, kill-switch primitives, and dual-use disclosure. Its arrival depends partly on a triggering incident (P8) and partly on community willingness to commit maintenance effort.

P4 couples the wireless side: a 3GPP or ITU-T study item on a molecular-electromagnetic gateway service class. The 6G commercialisation window of approximately 2030 is the natural deadline; if the study item has not opened by then, the molecular modality will not be part of the first 6G releases and will likely remain a research-only service for another standards cycle.

P7 is the linchpin of the cluster because it unblocks P5 and P6: a public multimodal benchmark covering MC symbol detection, biosensor anomaly detection, and edge inference under a stated energy envelope. Section 13 catalogues the infrastructure that already exists; the missing component is the convening function.

15.4. Cluster C – Clinical translation (P2, P10)

The third cluster concerns the moment at which IoBNT moves out of the laboratory. Two predictions stand for the cluster.

P2 is the most ambitious clinical claim of this paper. A Bio-SDN control plane for closed-loop diabetes management entering Phase I trials by 2032 would be a substantial vindication of the orchestration chapter (section 10); it would also require a regulator-approved in-vivo kill-switch primitive, which is currently not certified by any regulator anywhere. The mechanism rests on Abbasi and Akan [2] for the information-theoretic substrate and Bonnet et al. [9] for the genetic primitive.

P10 is the weaker version of the same claim: at least one in-vivo BNT-edge clinical trial registered by 2028, targeting a named indication. Spinal-cord-injury [3] and atherosclerosis monitoring are the most plausible early indications.

15.5. Cluster D – Operational reality (P8, P9)

The fourth cluster contains the two predictions whose verification will surprise the reader because they belong to ordinary technology operations rather than research outputs.

P8 predicts that a publicly disclosed bio-malware incident will occur by 2029. We do not relish this prediction; a falsified P8 is a *good* outcome for the field and for society. We include it because the threat-surface analysis of Zafar et al. [46] and section 12 suggests the pre-conditions are sufficient. A disclosed incident would accelerate P3.

P9 predicts that phytobiome-IoBNT pilots reach commercial agricultural deployment by 2030. The agriculture thread [5, 16] is the earliest non-clinical translation pathway, partly because the regulatory regime for plant biology is more permissive than for in-body human deployments.

15.6. Aggregate roadmap

Figure 7 visualises the same ten predictions as in table 11 along a 2026–2035 timeline, colour-coded by cluster. The 2030 column is densest: four of the ten predictions resolve in that year, which makes 2030 the natural moment for the next major survey of the field.

Table 11: The IoBNT-Edge research agenda 2026–2035. Each row is one dated falsifiable prediction with a stated mechanism and a counter-condition ("falsifier") that would make the prediction wrong. The intent is not certainty: the intent is to give the community a structured object that subsequent work can measure itself against. Prediction IDs are referenced from the other sections of this paper.

ID	Prediction (date + metric)	Mechanism / driver	Falsifier
P1	By **2030-12-31**, an implantable BCI runs sub-microwatt inference on PENG- or TENG-harvested power, sustaining event-driven ECG- or EMG-class anomaly detection (duty-cycled, with burst inference at harvested-power peaks) at $\geq$ 99 % uptime over a 12-month deployment.	PENG output already reaches the 100 µW band (§11); aggressive TinyML quantisation (§9) closes the remaining order of magnitude.	No implantable demonstration meeting the $\leq$1 µW $\times$ 99 %-uptime envelope published by 2030-12-31.
P2	By **2032-12-31**, a Bio-SDN control plane for closed-loop diabetes management enters Phase I clinical trials.	Insulin–glucose ICT models [2] and integrase-based kill-switch primitives [9] provide the primitive; the control plane of §10 is its glue.	No IND or CTA filing referencing programmable in-vivo agents indexed on a public regulatory registry by 2032-12-31; AND no peer-reviewed chronic ($\geq$6 month) in-vivo kill-switch validation in a large-animal model published by 2032-12-31.
P3	By **2030-12-31**, IEEE 1906.x is supplemented by a security amendment (informally "1906.2") covering trust boundaries, kill-switch primitives, and dual-use disclosure.	Cyberbiosecurity attention is rising [38, 46] (§12); a publicly disclosed incident (P8) would accelerate.	No IEEE 1906 amendment with "security" or "threat" in scope published by 2030-12-31.
P4	By **2030-12-31**, a molecular-electromagnetic gateway service class is on the 3GPP or ITU-T standardisation agenda as a 6G non-terrestrial-network-adjacent profile.	Sub-THz 6G timetables converge with IoBNT BCI throughput requirements [23] (§10).	No 3GPP TR, ITU-T SG13/16 study item, or ETSI ISG report covering molecular or biochemical communication interfaces indexed by 2030-12-31.
P5	By **2028-12-31**, a federated learning protocol for IoBNT converges to within 5 % of centralized accuracy under client heterogeneity at Dirichlet $\alpha \leq 0.1$ on a public MC-decoding benchmark.	Jamshidi, Hoang and Nguyen [20] demonstrates the integration pattern; the missing public benchmark (P7) is the binding constraint.	No published result meeting the gap, heterogeneity, and benchmark conditions simultaneously by 2028-12-31.
P6	By **2027-12-31**, a TinyML model for MC symbol detection meets the MLPerf-Tiny resource envelope ($\leq$100 kB params, $\leq$10 mJ per inference) at accuracy competitive with cloud-class baselines on a public MC channel.	Adjacent TinyML work (keyword spotting, visual-wake-words) already meets the envelope on harder tasks (§9); transfer requires only an IoBNT-realistic dataset.	No TinyML demonstration meeting all three constraints simultaneously by 2027-12-31.

Continued on next page

Table 11 – continued from previous page

ID	Prediction (date + metric)	Mechanism / driver	Falsifier
P7	By **2027-12-31**, an open multimodal IoBNT benchmark exists, with at least three independent submissions, covering MC symbol detection, biosensor anomaly detection, and edge inference under a stated energy envelope (§13).	MLPerf Tiny is the structural template; the simulators and testbeds exist [17, 35]; what is missing is the convening function.	No public benchmark meeting all three modalities with ≥3 submissions by 2027-12-31.
P8	By **2029-12-31**, at least one publicly disclosed bio-malware or biosecurity incident occurs in a clinical bioinformatics pipeline, triggering regulatory response.	The threat surface is articulated in Zafar et al. [46] and §12; the attack pre-conditions (cheap DNA synthesis, expanding bioinformatics dependence) are already met.	No publicly disclosed incident meeting ALL of: (a) involves a clinical bioinformatics pipeline or IoBNT-adjacent system; (b) documented in a peer-reviewed case report, regulator safety notice, national CERT advisory, OR indexed cybersecurity database; (c) actual or near-miss compromise (not proof-of-concept lab demonstration only). (Note: a falsified P8 is a *good* outcome.)
P9	By **2030-12-31**, phytobiome-IoBNT pilots reach commercial agricultural deployment in at least one country.	The agriculture-IoBNT thread of Gulec et al. [16] and Babar and Akan [5] is the earliest external translation pathway.	No commercial deployment (defined as sale of ≥100 units or recurring SaaS revenue, not pilot or research-only) by 2030-12-31.
P10	By **2028-12-31**, at least one in-vivo BNT-edge clinical trial is registered on a public registry (ClinicalTrials.gov or equivalent), targeting a named indication other than a glucose monitor.	Akan et al. [3] demonstrates the clinical-translation framing; spinal-cord-injury and atherosclerosis applications are the most plausible early indications.	No trial registration meeting these criteria by 2028-12-31.

15.7. How to use this agenda

Three audiences may find the agenda useful:

1. *Researchers.* Each prediction names a research bet. A researcher who agrees with the mechanism can work to make the prediction come true; one who disagrees can work to falsify it.
2. *Funders.* Each cluster names an institutional gap. Cluster B is the most amenable to direct funding (benchmarks, standards-body engagement) because its bottlenecks are explicitly institutional rather than intellectual.
3. *Future survey authors.* A 2031 survey of IoBNT will be able to check the present table against the field's actual trajectory. We invite that checking. If the predictions in table 11 prove right, the field will be a recognisable engineering discipline. If they prove wrong, the falsifiers will tell future authors precisely *how* the trajectory diverged from expectation, which is itself the more valuable artefact.

16. Conclusion

IoBNT is no longer a speculative networking idea. The bibliometric snapshot of section 3 shows a field of more than 300 deduplicated entries, accelerating sharply after 2023, with three identifiable

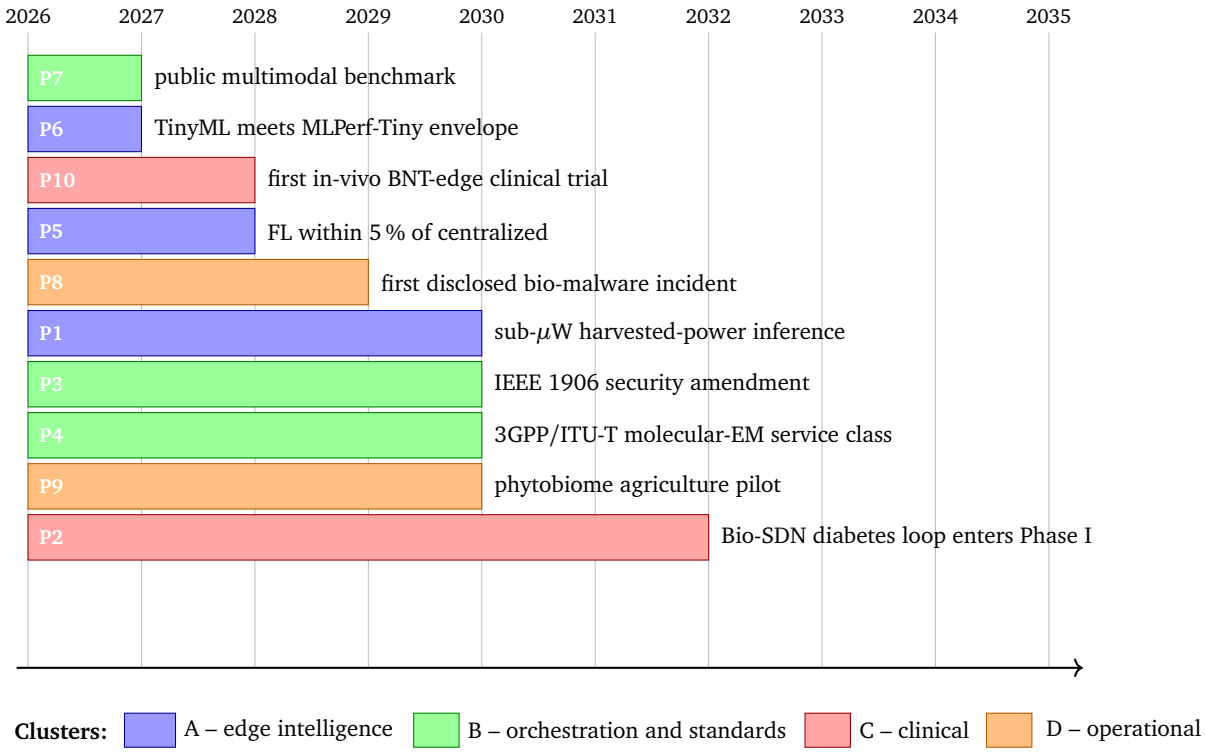

Figure 7: Agenda roadmap. Each horizontal bar corresponds to one prediction in table 11; the bar's right edge is the prediction date. Colours encode clusters (A: edge intelligence, B: orchestration and standards, C: clinical translation, D: operational reality). The 2030 column carries four predictions, making 2030 the natural moment for the next major survey of the field.

research centres of gravity, mature simulators, and a first generation of microfluidic testbeds. What IoBNT is not yet – and what the present survey argues it should become – is an edge-computing discipline.

This paper has tried to make three things visible. First, the IoBNT layer stack has been re-rendered as a five-layer architecture (section 4, figure 5, table 4) in which the bio-cyber interface is a first-class compute layer rather than a transduction gateway, and the edge tier is named explicitly. Second, the three edge-relevant subtopics that the bibliometric snapshot identifies as under-occupied – federated learning at the bio-edge, TinyML on harvested power, and Bio-SDN orchestration – have been treated as research fronts (sections 8, 9, and 10). Third, ten dated falsifiable predictions (section 15, table 11) state what the field should aim to demonstrate between 2026 and 2035, with mechanisms and counter-conditions that future surveys can check.

Whether the predictions turn out to be right matters less than whether they prove to be the right kind of object. A field whose surveys can be falsified is a field that is becoming an engineering discipline. We invite the next decade of IoBNT research to take that step.

Editorial for JEC Volume 5 Issue 1 (2026)

This issue of the *Journal of Edge Computing* opens its fifth volume with five papers that span the edge-computing spectrum from on-device large language models to satellite-image segmentation. As editors of this issue, we introduce each contribution below; all five are currently under peer review.

Kovbasiuk et al. [24] address the problem of object detection in satellite imagery acquired under cloud cover. Their method combines instance segmentation with a cloud-penetration preprocessing stage, targeting operational scenarios where atmospheric interference is the norm rather than an edge case. The work extends the edge-computing theme to remote sensing, where on-board or near-sensor processing is increasingly required to reduce the downlink burden.

Ray and Pradhan [39] provide a side-by-side benchmark of four quantised large language models (Llama 3.2, Gemma 3, Granite 3.1-MoE, and Qwen 2.5) served on a Raspberry Pi 4 via Python and Rust API clients. The central finding is a decisive cold-start latency advantage for Rust (mean model-load times falling by over 90 % across the tested models) with statistically indistinguishable warm-start throughput, yielding actionable guidance for edge-AI practitioners choosing a client language for on-device LLM inference.

Shirsat and Nirmalrani [43] propose an IoT-integrated deep-learning pipeline for detecting waste hazards in floating-water and reservoir environments. The architecture couples principal-component analysis and grey-level co-occurrence-matrix feature extraction with Fast R-CNN, deployed across IoT sensors and fog-computing nodes. The emphasis on real-time constraints and environmental variability makes the paper a case study in the practical difficulties of edge-AI deployment outside the laboratory.

Roveda et al. [40] revisit the EdgeAI design space through the lens of three coupled resource dimensions: communication, storage, and computation. Rather than optimising one dimension in isolation, the authors argue for joint optimisation and survey the techniques that make it tractable, from model compression to adaptive offloading. The paper serves as a reference map for the trade-offs that every edge-AI system designer must navigate.

Nikose and Chinara [36] tackle seizure prediction from EEG with a depthwise-separable CNN that reduces computational requirements while preserving predictive performance. The work sits at the intersection of two themes that this journal has championed since its founding: bringing machine learning onto resource-constrained hardware and applying it to high-stakes, latency-sensitive decisions.

We thank the authors for entrusting their work to JEC and the anonymous reviewers whose reports will shape the final versions. The call for next volumes remains open; we particularly welcome submissions that combine edge-computing systems research with reproducible benchmarks, as several of the papers above do.

Acknowledgments

The authors thank the open-source communities behind bibtexparser, scikit-learn, and the MLPerf Tiny and Flower FL toolchains whose infrastructure the bibliometric pipeline and edge-intelligence chapters rely on. The corpus was assembled from Web of Science and Scopus exports dated 2026-04-28 with supplementary arXiv printouts; the deduplication pipeline, subtopic labeller, and verification scripts are released at the repository URL recorded in Appendix A.

References

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L., 2016. Deep Learning with Differential Privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16. New York, NY, USA: Association for Computing Machinery, p.308–318. Available from: https://doi.org/10.1145/2976749.2978318.

[2] Abbasi, N.A. and Akan, O.B., 2017. An Information Theoretical Analysis of Human Insulin-Glucose System Toward the Internet of Bio-Nano Things. *IEEE Transactions on NanoBioscience*, 16(8), pp.783–791. Available from: https://doi.org/10.1109/TNB.2017.2762160.

[3] Akan, O.B., Ramezani, H., Civas, M., Cetinkaya, O., Bilgin, B.A. and Abbasi, N.A., 2023. Information and Communication Theoretical Understanding and Treatment of Spinal Cord Injuries: State-of-The-Art and Research Challenges. *IEEE Reviews in Biomedical Engineering*, 16, pp.332–347. Available from: https://doi.org/10.1109/RBME.2021.3056455.

[4] Akyildiz, I.F., Pierobon, M., Balasubramaniam, S. and Koucheryavy, Y., 2015. The internet of Bio-Nano things. *IEEE Communications Magazine*, 53(3), pp.32–40. Available from: https://doi.org/10.1109/MCOM.2015.7060516.

[5] Babar, A.Z. and Akan, O.B., 2025. Sustainable and Precision Agriculture with the Internet of Everything (IoE). 2404.06341, Available from: https://doi.org/10.48550/arXiv.2404.06341.

[6] Banbury, C.R., Reddi, V.J., Torelli, P., Jeffries, N., Király, C., Holleman, J., Montino, P., Kanter, D., Warden, P., Pau, D., Thakker, U., Torrini, A., Cordaro, J., Guglielmo, G.D., Duarte, J.M., Tran, H., Tran, N., Niu, W. and Xu, X., 2021. MLPerf Tiny Benchmark. In: J. Vanschoren and S. Yeung, eds. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*. Available from: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/da4fb5c6e93e74d3df8527599fa62642-Abstract-round1.html.

[7] Bhattacharjee, S., Bi, D., Hofmann, P., Wietfeld, A., Becke, S., Lommel, M., Zhou, P., Zheng, R., Kertzscher, U., Deng, Y., Kellerer, W., Fitzek, F.H.P. and Dressler, F., 2026. Exhaled Breath Analysis Through the Lens of Molecular Communication: A Survey. *IEEE Communications Surveys & Tutorials*, 28, pp.412–445. Available from: https://doi.org/10.1109/COMST.2025.3605748.

[8] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A. and Seth, K., 2017. Practical Secure Aggregation for Privacy-Preserving Machine Learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17. New York, NY, USA: Association for Computing Machinery, p.1175–1191. Available from: https://doi.org/10.1145/3133956.3133982.

[9] Bonnet, J., Yin, P., Ortiz, M.E., Subsoontorn, P. and Endy, D., 2013. Amplifying Genetic Logic Gates. *Science*, 340(6132), pp.599–603. Available from: https://doi.org/10.1126/science.1232758.

[10] Bulasara, P.K., Sahoo, S., Gupta, N., Han, Z. and Kumar, N., 2025. The Internet of Bio-Nano Things with Insulin-Glucose, Security and Research Challenges: A Survey. *ACM Computing Surveys*, 57(5), January, p.109. Available from: https://doi.org/10.1145/3703448.

[11] Cai, H. and Akan, O.B., 2025. Semantic Learning for Molecular Communication in Internet of Bio-Nano Things. 2502.08426, Available from: https://doi.org/10.48550/arXiv.2502.08426.

[12] Chude-Okonkwo, U.A.K., Malekian, R., Maharaj, B.T. and Vasilakos, A.V., 2017. Molecular Communication and Nanonetwork for Targeted Drug Delivery: A Survey. *IEEE Communications Surveys & Tutorials*, 19(4), pp.3046–3096. Available from: https://doi.org/10.1109/COMST.2017.2705740.

[13] Civas, M., Kuscu, M., Cetinkaya, O., Ortlek, B.E. and Akan, O.B., 2023. Graphene and related materials for the Internet of Bio-Nano Things. *APL Materials*, 11(8), 08, p.080901. Available from: https://doi.org/10.1063/5.0153423.

[14] Darya, A.M., Vakani, H. and Nasir, Q., 2019. Error Control Codes for Molecular Communication Channels: A Survey. *2019 International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*. pp.1–4. Available from: https://doi.org/10.1109/ICCSPA.2019.8713672.

[15] Finn, C., Abbeel, P. and Levine, S., 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In: D. Precup and Y.W. Teh, eds. *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research. PMLR, pp.1126–1135. Available from: http://proceedings.mlr.press/v70/finn17a.html.

[16] Gulec, F., Awan, H., Wallbridge, N. and Eckford, A.W., 2026. Decoding and Engineering the Phytobiome Communication for Smart Agriculture. *IEEE Communications Magazine*, 64(4), pp.120–126. Available from: https://doi.org/10.1109/MCOM.001.2400570.

[17] Hamidović, M., Angerbauer, S., Bi, D., Deng, Y., Tugcu, T. and Haselmayr, W., 2024. Microfluidic Systems for Molecular Communications: A Review From Theory to Practice. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, 10(1), pp.147–163. Available from: https://doi.org/10.1109/TMBMC.2024.3368768.

[18] Hofmann, P., Cabrera, J.A., Bassoli, R., Reisslein, M. and Fitzek, F.H.P., 2023. Coding in Diffusion-Based Molecular Nanonetworks: A Comprehensive Survey. *IEEE Access*, 11, pp.16411–16465. Available from: https://doi.org/10.1109/ACCESS.2023.3243797.

[19] Jaeger, H., 2001. *The "echo state" approach to analysing and training recurrent neural networks*. (GMD Report 148). German National Research Center for Information Technology. Available from: https://www.ai.rug.nl/minds/uploads/EchoStatesTechRep.pdf.

[20] Jamshidi, M., Hoang, D.T. and Nguyen, D.N., 2024. CNN-FL for Biotechnology Industry Empowered by Internet-of-BioNano Things and Digital Twins. 2402.00238, Available from: https://doi.org/10.48550/arXiv.2402.00238.

[21] Jamshidi, M., Thai Hoang, D., Nguyen, D.N., Niyato, D. and Ebrahimi Warkiani, M., 2025. Physics-Informed Neural Networks for Bio-Nano Digital Twins: A Multimodel Framework With IoBNT Integration. *IEEE Internet of Things Journal*, 12(24), pp.53868–53884. Available from: https://doi.org/10.1109/JIOT.2025.3621421.

[22] Kilic, B.A. and Akan, O.B., 2026. Neural-Inspired Multi-Agent Molecular Communication Networks for Collective Intelligence. 2601.18018, Available from: https://doi.org/10.48550/arXiv.2601.18018.

[23] Kong, L., Huang, L., Lin, L., Zheng, Z., Li, Y., Wang, Q. and Liu, G., 2023. A Survey for Possible Technologies of Micro/Nanomachines Used for Molecular Communication Within 6G Application Scenarios. *IEEE Internet of Things Journal*, 10(13), pp.11240–11263. Available from: https://doi.org/10.1109/JIOT.2023.3255412.

[24] Kovbasiuk, S.V., Romanchuk, M.P., Naumchak, O.M. and Naumchak, L.M., 2026. Object detection method based on instance segmentation of satellite image obtained in the conditions of cloud cover. *Journal of Edge Computing*, 5(1), pp.90–104. Available from: https://doi.org/10.55056/jec.749.

[25] Kuscu, M. and Akan, O.B., 2018. Maximum Likelihood Detection With Ligand Receptors for Diffusion-Based Molecular Communications in Internet of Bio-Nano Things. *IEEE Transactions on NanoBioscience*, 17(1), pp.44–54. Available from: https://doi.org/10.1109/TNB.2018.2792434.

[26] Kuscu, M., Dinc, E., Bilgin, B.A., Ramezani, H. and Akan, O.B., 2019. Transmitter and Receiver Architectures for Molecular Communications: A Survey on Physical Design With Modulation, Coding, and Detection Techniques. *Proceedings of the IEEE*, 107(7), pp.1302–1341. Available from: https://doi.org/10.1109/JPROC.2019.2916081.

[27] Kuscu, M. and Unluturk, B.D., 2021. Internet of Bio-Nano Things: A review of applications, enabling technologies and key challenges. *ITU Journal on Future and Evolving Technologies*, 2(3), pp.1–24. Available from: https://doi.org/10.52953/chbb9821.

[28] Lee, C., Koo, B.H., Chae, C.B. and Schober, R., 2023. The Internet of bio-nano things in blood vessels: System design and prototypes. *Journal of Communications and Networks*, 25(2), pp.222–231. Available from: https://doi.org/10.23919/JCN.2023.000001.

[29] Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A. and Smith, V., 2020. Federated Optimization in Heterogeneous Networks. In: I.S. Dhillon, D.S. Papailiopoulos and V. Sze, eds. *Proceedings of the Third Conference on Machine Learning and Systems, MLSys 2020, Austin, TX, USA, March 2-4, 2020*. mlsys.org. Available from: https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

[30] Lin, J., Chen, W., Lin, Y., Cohn, J., Gan, C. and Han, S., 2020. MCUNet: Tiny Deep Learning on IoT Devices. In: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin, eds. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Available from: https://proceedings.neurips.cc/paper/2020/hash/86c51678350f656dcc7f490a43946ee5-Abstract.html.

[31] Lotter, S., Mohr, E., Rutsch, A., Brand, L., Ronchi, F. and Díaz-Marugán, L., 2025. Synthetic MC via Biological Transmitters: Therapeutic Modulation of the Gut-Brain Axis. *Proceedings of the 12th Annual ACM International Conference on Nanoscale Computing and Communication*, NANOCOM '25. New York, NY, USA: Association for Computing Machinery, p.84–90. Available from: https://doi.org/10.1145/3760544.3764138.

[32] Maass, W., Natschläger, T. and Markram, H., 2002. Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations. *Neural Computation*,

14(11), 11, pp.2531–2560. Available from: https://doi.org/10.1162/089976602760407955.

[33] McMahan, B., Moore, E., Ramage, D., Hampson, S. and Arcas, B.A.y., 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In: A. Singh and J. Zhu, eds. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, vol. 54. PMLR, pp.1273–1282. Available from: https://proceedings.mlr.press/v54/mcmahan17a.html.

[34] Meenambika, A., Jayachitra, S., Nagamuthu, A.L. and Dhanaraj, R.K., 2026. The Internet of Bio-Nano Things for personalized healthcare: perspectives, framework, and research directions. In: J. Sekar, P. Aruchamy, R.K. Dhanaraj and S. Kadry, eds. *Future of Internet of Bio-Nano Things in Personalized Healthcare*. Academic Press, chap. 2, pp.23–39. Available from: https://doi.org/10.1016/B978-0-443-27604-0.00002-0.

[35] Miray Albay, M., Akyol, E., Mirlou, F., Beker, L. and Kuscu, M., 2025. Low-Cost Microfluidic Testbed for Molecular Communications With Integrated Hydrodynamic Gating and Screen-Printed Sensors. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, 11(4), pp.518–523. Available from: https://doi.org/10.1109/TMBMC.2025.3614382.

[36] Nikose, R.D. and Chinara, S., 2026. Optimising seizure prediction with reduced computational resources using depthwise CNN. *Journal of Edge Computing*, 5(1), pp.173–188. Available from: https://doi.org/10.55056/jec.1172.

[37] Okaie, Y., Nakano, T., Hara, T. and Nishio, S., 2016. Controllability of Mobile Bionanosensors. *Target Detection and Tracking by Bionanosensor Networks*, SpringerBriefs in Computer Science. Singapore: Springer Singapore, pp.53–58. Available from: https://doi.org/10.1007/978-981-10-2468-9_4.

[38] Rana, A., Gautam, D., Kumar, P. and Kumar Das, A., 2025. Architectures, Benefits, Security, and Privacy Issues of Internet of Nano Things: A Comprehensive Survey, Opportunities, and Research Challenges. *IEEE Communications Surveys & Tutorials*, 27(2), pp.1152–1190. Available from: https://doi.org/10.1109/COMST.2024.3423477.

[39] Ray, P.P. and Pradhan, M.P., 2026. Performance analysis of localised large language models in resource-constrained edge for Python and Rust APIs. *Journal of Edge Computing*, 5(1), pp.47–89. Available from: https://doi.org/10.55056/jec.1047.

[40] Roveda, M., Lopes Ferreira, D., Santos Oliveira, A. dos, Silva, F.S. Tesch da, Kunst, R., da Costa, C.A. and da Rosa Righi, R., 2026. Revisiting EdgeAI through the lens of communication, storage and computing optimisations. *Journal of Edge Computing*, 5(1), pp.126–172. Available from: https://doi.org/10.55056/jec.1054.

[41] Schottlender, M., Schäfer, M. and Veiga, R.A., 2025. Neural Network based Distance Estimation for Branched Molecular Communication Systems. *Proceedings of the 12th Annual ACM International Conference on Nanoscale Computing and Communication*, NANOCOM '25. New York, NY, USA: Association for Computing Machinery, p.28–33. Available from: https://doi.org/10.1145/3760544.3764128.

[42] Semerikov, S.O. and Vakaliuk, T.A., 2026. The bio-edge: a survey and research agenda for the Internet of Bio-Nano Things, 2026–2035. *Journal of edge computing*, 5(1), pp.1–46. Available from: https://doi.org/10.55056/jec.1382.

[43] Shirsat, N. and Nirmalrani, V., 2026. Seamless monitoring and detection of waste hazards in floating water and water reservoirs using Internet of Things integrated deep learning algorithms. *Journal of Edge Computing*, 5(1), pp.105–125. Available from: https://doi.org/10.55056/jec.1051.

[44] Tjabben, A., Bergkemper, L., Herbst, J., Rueb, M., Lipps, C. and Schotten, H.D., 2024. Multipath Signal Prediction for In-Body Nanocommunication with Volatile Particles. *European Wireless 2024; 29th European Wireless Conference*. pp.41–46. Available from: https://www.researchgate.net/publication/384285452.

[45] Torres Gómez, J., Hofmann, P., Debus, L.Y., Başaran, O.T., Lotter, S., Khanzadeh, R., Angerbauer, S., Unluturk, B.D., Abadal, S., Haselmayr, W., Fitzek, F.H.P., Schober, R. and Dressler, F., 2025. Communicating Smartly in Molecular Communication Environments: Neural Networks in the

Internet of Bio-Nano Things. 2506.20589, Available from: https://doi.org/10.48550/arXiv.2506.20589.

[46] Zafar, S., Nazir, M., Sabah, A. and Jurcut, A.D., 2021. Securing Bio-Cyber Interface for the Internet of Bio-Nano Things using Particle Swarm Optimization and Artificial Neural Networks based parameter profiling. *Computers in Biology and Medicine*, 136, p.104707. Available from: https://doi.org/10.1016/j.compbiomed.2021.104707.

A. Corpus and bibliometric methodology

This appendix documents the construction of the deduplicated corpus that underpins section 3 and the bibliometric figures and tables throughout the paper. The source pipeline is reproducible and the intermediate data are available as supplementary material.

A.1. Sources

The corpus draws on three databases.

Two exports were taken on 2026-04-28 from the following query executed against the Web of Science Core Collection (Topic field, which searches title, abstract, author keywords, and Keywords Plus®):

```
TS=("Internet of Bio-Nano Things" OR "IoBNT" OR
  ("bio-nano" AND "Internet of Things") OR
  ("molecular communication" AND ("Internet of Things"
  OR "edge computing")))
```

The two exports were necessary because Web of Science caps single exports at 500 records and the union of relevant subsets exceeds that cap. They contain 225 and 191 raw entries respectively. A small subset (59 pre-2010 entries) consists of humanities artefacts caught by overlapping terminology – titles such as *Leibniz rules and reality conditions* and *SOARING* – and was excluded by a topic filter.

A single Scopus export was taken on the same date with the comparable query:

```
TITLE-ABS-KEY("Internet of Bio-Nano Things" OR "IoBNT"
  OR ("bio-nano" AND "Internet of Things"))
```

It contains 226 raw entries after the initial parser pass (the BibTeX export's leading non-entry banner is stripped). Substantial overlap with the Web-of-Science exports is expected and observed.

Two arXiv search-result printouts (PDF format) were retained as supplementary signal but were not parsed mechanically. They provided the eight critical recent entries that were hand-curated into corpus (Babar and Akan [5], Bonnet et al. [9], Cai and Akan [11], Chude-Okonkwo et al. [12], Jamshidi, Hoang and Nguyen [20], Kilic and Akan [22], Miray Albay et al. [35], Torres Gómez et al. [45] because those preprints have not yet propagated to Web of Science or Scopus.

A.2. Deduplication pipeline

The pipeline lives in the `tools/` directory of the manuscript repository and is invoked by `python3 dedup_bibs.py`. The implementation is a thin wrapper around `bibtexparser` with three substantive operations:

1. *Noise filter.* Pre-2010 entries are required to be topically relevant under a fixed keyword list (`nano`, `molecular`, `iobnt`, `biocyber`, `nanonetwork`, `biosensor`, and a small number of synonyms). Pre-2010 entries that fail the topic filter are dropped; 59 entries are dropped at this stage. Post-2010 entries are retained unconditionally on the assumption that the targeted queries were precise enough.

2. *DOI-keyed dedup.* Each entry's DOI is normalised (lowercased, URL prefix stripped, `doi:` prefix stripped) and used as the primary key. Entries with the same DOI are merged: the most complete entry by field count is retained, with missing fields filled from the merged siblings.
3. *Title-fallback dedup.* Entries without DOIs are bucketed by normalised title (lowercased, punctuation removed, whitespace collapsed). Buckets with more than one entry are merged using the same completeness rule.

The final corpus contains 311 unique entries; 257 carry a DOI and 54 do not. A canonical citation key of the form `<FirstAuthorLastName><Year><FirstContentWord>` is assigned, with letter suffixes appended where collisions occur.

A.3. Subtopic labeller

The subtopic-bucket labels of figure 4 are produced by a rule-based labeller (`subtopic_buckets.py`). Each bucket is defined by a keyword list, and each entry's title, keywords, abstract, and venue are concatenated and scanned for keyword matches. The bucket with the highest hit count wins, with a 5-point bonus for the three white-space buckets so that an IoBNT paper that mentions *federated learning* once is classified as federated-learning even when the paper is otherwise deep-learning-heavy. Entries with zero matches across any bucket are labelled *Unclassified*.

A.4. Topical cluster figure

Figure 2 is produced by `topic_cluster.py`. Each entry contributes a text vector built from its title, keywords, and (truncated) abstract. TF–IDF vectorisation uses 1–2-grams with a minimum document frequency of 2; the resulting sparse matrix is projected to two dimensions with t-SNE using the cosine metric and a perplexity of $\min(30, N/10)$. Points are coloured by the subtopic labeller's output; white-space buckets are drawn with high-contrast markers so they are visible against the dominant MC-channel-theory cloud.

A.5. Co-author network figure

Figure 3 is produced by `coauthor_network.py`. Authors with at least three deduplicated papers are kept; an undirected graph is built with edge weights equal to the number of co-authored entries between endpoints. Isolated nodes are dropped. The largest connected component is laid out with Kamada–Kawai. Communities are detected with greedy modularity (Clauset–Newman–Moore); the three research-programme centres surface automatically without any prior labelling.

A.6. Reproducibility

The full pipeline and data exports are archived at https://github.com/ssemerikov/bio-edge. The pipeline is deterministic. Re-running the scripts (`dedup_bibs.py`, `year_counts.py`, `subtopic_buckets.py`, `topic_cluster.py`, `coauthor_network.py`) against the source exports produces byte-identical references and figure outputs. The `-check` flag on `dedup_bibs.py` verifies that no DOI appears in the output more than once; `verify_claims.py` audits the manuscript for numeric claims that lack a citation.

A.7. Known limitations

Six limitations should be acknowledged.

1. *Corpus scope.* The corpus is biased toward Western-database-indexed work; results from venues indexed only in regional databases are under-represented. The query centers on the literal phrase "Internet of Bio-Nano Things" and its abbreviation; papers that contribute to IoBNT-relevant sub-problems (molecular communication, in-body biosensing, bio-cyber transduction)

without using the IoBNT brand are systematically missed unless they happen to appear in the same venues. Section 3 flags this in the TinyML keyword test; the same applies to the digital-twin and FL literatures.

2. *Subtopic labelling bias.* The rule-based labeller applies a 5-point bonus for the three white-space buckets (edge computing, federated learning, TinyML) that this survey claims as gaps. Under a neutral weighting (bonus = 0) the white-space buckets still hold fewer than ten entries combined, confirming that the gap is not an artefact, but the bonus inflates their counts relative to an unbiased classifier. Both counts are reported in the `output/` data.
3. *Unclassified bucket.* The labeller produces a non-trivial unclassified bucket (67 entries, 21 % of the corpus) whose redistribution would require either richer features or a manual coding pass; we have not attempted that pass.
4. *Author disambiguation.* Author names are normalised conservatively (first initial + last name, case-folded). "Nieto-Chaupis, H" at 35 papers is the most visible potential disambiguation artefact; spot-checking confirms that at least 5 of those entries are genuinely IoBNT-relevant, but the count may aggregate works from multiple researchers sharing the same surname–initial pair. ORCID-based disambiguation would improve precision but is not feasible without an author-provided identifier per entry.
5. *Deduplication conservatism.* The deduplication is conservative: a paper that has appeared in a preprint, a workshop, and a journal version may legitimately be present in the corpus once or three times depending on what its DOI records contain. We chose the conservative DOI policy because the alternative – aggressive same-author-same-title collapsing – risks hiding genuinely distinct works.
6. *t-SNE interpretability.* The t-SNE projection of figure 2 is used for exploratory display only; quantitative claims about topical structure rest on the rule-based labeller, not on the t-SNE geometry. t-SNE is not stable under re-initialisation and does not preserve inter-cluster distances; apparent clusters should not be interpreted as statistically distinct sub-literatures. A UMAP projection or multiple t-SNE parameter settings would strengthen the visual argument; we defer this to a revision.

B. Glossary and acronyms

BANNET Body Area Nano-NETwork; a network of BNTs confined to a single anatomical site.

BCI Bio-cyber interface; the layer at which molecular signals are transduced into electromagnetic ones (section 6). Not to be confused with brain–computer interface, which appears in this paper only when explicitly distinguished.

Bio-SDN Software-defined networking adapted to the biochemical domain (section 10); separates an application plane, a control plane, and a data plane with biology-specific safety primitives.

BNT Bio-nano thing; an engineered cell, nanoparticle-based transducer, DNA-origami device, or hybrid bionanomachine, the fundamental unit of Layer 1 of the reference architecture.

BNSim An MC simulator widely used in the field.

DT Digital twin; a continuously assimilated computational model of a patient or subsystem.

FL Federated learning (section 8); a class of distributed training algorithms that aggregate gradients or model deltas without sharing raw data.

FRET Förster Resonance Energy Transfer; a fluorescence mechanism widely used for monitoring drug release in theranostics.

GFC Glucose fuel cell; a biochemical energy harvester that converts glucose to electrical current.

IEEE 1906.1 Recommended practice for nanoscale and molecular communication framework (section 13).

IoBNT Internet of Bio-Nano Things.

IoNT Internet of Nano Things; the direct precursor of IoBNT.

MC Molecular communication; the dominant IoBNT communication modality (section 5).

MEC Multi-access (or mobile) edge computing; a 5G/6G edge-tier deployment model.

ML Machine learning.

MNNN / M^{3}N Molecular Nano Neural Network; a chemical-reaction-network analogue of a neural network. The acronym is used in the literature with both spellings.

MoNaCo An ns-3 extension for MC simulation.

NTN Non-terrestrial network; a 3GPP service class that IoBNT integration would extend (section 10).

PENG Piezoelectric nanogenerator (section 11).

PINN Physics-informed neural network; a class of models that embed a physical PDE in the training loss.

RL Reinforcement learning.

Simbiotics An agent-based MC simulator.

STRIDE A threat-classification framework: spoofing, tampering, repudiation, information disclosure, denial of service, elevation of privilege.

TDD Targeted drug delivery.

TENG Triboelectric nanogenerator (section 11).

THz Terahertz; the electromagnetic band used for graphene-plasmonic BCI links (section 5).

TinyML Machine-learning inference on microcontroller-class hardware (section 9).

TRL Technology readiness level; a 1–9 scale used by space agencies and increasingly by funders to classify the maturity of a research artefact.

C. Simulators, testbeds, and open datasets

The infrastructure question has been raised throughout the paper: IoBNT progress is increasingly constrained not by lack of models, but by lack of shared infrastructure on which models can be compared. This appendix catalogues the available simulators, testbeds, and open datasets and identifies the gaps that section 13 argues need closure.

C.1. Simulators

BNSim. The earliest IoBNT-relevant simulator, focused on agent-based modelling of populations of communicating BNTs at the bacterial scale. Suitable for nanonetwork-layer experiments (Layer 2 of section 4); less suitable for BCI-tier inference experiments because the simulator stops at the molecular boundary.

Simbiotics. A more recent agent-based simulator with stronger support for spatial dynamics and synthetic-biology primitives. Used in much of the work on engineered-cell communication.

ns-3 MoNaCo. An extension to the widely-used ns-3 network simulator that adds molecular-communication models. The natural target for any experimental work that needs to combine MC with conventional radio links; the relevant venue for the BCI-to-edge wireless slice of section 10.

Physics-informed neural-network simulators. The PINN approach of Jamshidi et al. [21] uses a neural network as a fast surrogate for a diffusion PDE. The same infrastructure can be re-used as a simulation back-end for TinyML training pipelines that need many realisations of the underlying channel.

C.2. Microfluidic testbeds

Low-cost screen-printed-sensor benches. Miray Albay et al. [35] describe a microfluidic MC testbed with hydrodynamic gating and screen-printed potentiometric sensors that costs approximately $1 per unit and is fabricated in under an hour, intended for MC symbol-detection experiments at the BCI tier (proof of concept: pH-encoded 4-ary CSK over a PANI-functionalized sensor). The key contribution is cost: the testbed is reproducible on a typical academic budget, which is the precondition for the benchmark artefact of prediction P7.

Microfluidic survey. Hamidović et al. [17] review the state of microfluidic implementations more broadly, including the higher-cost fluorescence-readout systems used in synthetic-biology laboratories.

C.3. Datasets

The most significant gap in the appendix is here. No public, permissively licensed dataset for IoBNT-relevant signal traces currently exists. Section 13 (and prediction P7 in section 15) treat the gap as the central institutional bottleneck of the field. The Miray Albay et al. [35] testbed is the most likely source of such a dataset; the convening function that would commit a group to maintain it is, in 2026, not yet identified.

C.4. Workflow recommendation

For a researcher starting an IoBNT-edge project today, the recommended workflow is:

1. Use a PINN-based surrogate [21] for fast model iteration during early development.
2. Validate the resulting model on ns-3 MoNaCo with a realistic channel model.
3. For a publication-grade demonstration, attempt at least one experimental run on a microfluidic testbed [17, 35].
4. Publish the trained model artefact and the training-data summary under a permissive license, even in advance of the community benchmark.

The fourth step is the one most readers will find inconvenient. It is also the one that will most accelerate the field, by populating the missing benchmark with seed submissions before the benchmark itself is formally specified.