EARTHSCAN

WHITEPAPER 24 PAGES / MAY 2026

MIMEA: AI for Biomass and Bio-Inspired Materials

Lignin is the second-most-abundant biopolymer on Earth and the least commercially understood. This whitepaper makes the case for deep learning as the missing layer between heterogeneous biomass feedstock and high-value downstream chemistry. We cover Higher Heating Value prediction without a calorimeter, neural surrogates for atomic-scale potential energy surfaces, the multimodal data infrastructure required, and a sequenced 12-month adoption path. Drawn from the MIMEA feasibility study (Real AI B.V., 2021) and updated with the state of the art in 2026.

Tarry Singh
Founder & CEO

MACHINE-LEARNING / DEEP-LEARNING / BIOMASS / LIGNIN / MATERIALS-SCIENCE / BIO-INSPIRED-MATERIALS

CONTENTS

> "
>
> The energy transition needs feedstocks that are abundant, predictable, and priceable. Lignin is all three — once AI makes it tractable.
>
> "

I

THE OPENING QUESTION

Why a feedstock most operators ignore is the energy-transition lever they need

The energy transition has a feedstock problem.

We have ambitious 2030 and 2050 targets, a maturing pipeline of electrification projects, and a fast-improving cost curve on solar and wind. What we don't have, at the scale required, is a way to convert the **abundant, distributed, low-grade carbon resources** that already exist — agricultural residues, pulp-mill side streams, municipal woody waste — into predictable inputs for downstream chemistry and energy.

Lignin sits at the centre of this gap. It is the **second-most- abundant biopolymer on Earth**, generated as a 100-million-tonne annual waste stream from paper and cellulosic-biofuel manufacturing, and it accounts for roughly 30% of the organic carbon in the biosphere. Despite that scale, less than **2%** of available lignin is sold as a structural input. The other 98% is incinerated for low-grade heat.

WHY LIGNIN IS THE LARGEST UNDER-PRICED FEEDSTOCK ON EARTH

100M t/yr
Lignin generated as paper / cellulosic-biofuel waste, globally

30%
Share of organic carbon in the biosphere held in lignin

<2%
Sold as a structural input — the rest is incinerated for low-grade heat

The economics tell you something is broken. The chemistry tells you why: lignin is a **heterogeneous, branched polyaromatic macromolecule** whose properties depend on species, soil, climate, age, and processing route. There is no single "lignin" — there is a vast structural family whose downstream value sits in the tail of a long distribution.

That heterogeneity has historically been the moat that kept lignin out of high-value chains. It is also exactly where modern AI has its sharpest edge.

This whitepaper makes three claims:

1. **Higher Heating Value (HHV) prediction without a bomb calorimeter is solved enough to deploy in a paper-mill control room today.** Deep learning on existing proximate and ultimate analyses outperforms classical regressions on long-tail feedstocks, runs in seconds, and converts a static lab metric into a real-time routing signal.

2. **Bio-inspired materials design via neural surrogates of atomic- scale modelling is at the inflection point.** It is no longer a research curiosity. The combination of equivariant graph neural networks, multi-modal training data, and active learning loops is collapsing the inner cycle of materials discovery by 10–100×.

3. **The unit economics of biorefining shift fundamentally when you characterise + route + price every batch automatically.** This is a data-and-decision problem before it is a chemistry problem — and the companies that build the data infrastructure first will route the entire downstream chain.

This whitepaper draws on **MIMEA** — Material Intelligence Modelling and Energy Applications — a feasibility study completed by Real AI B.V. (the research arm in DeepKapha's group, which also operates EarthScan) under the Dutch SNN funding programme. MIMEA was an explicit go / no-go exercise, not a product launch. The conclusion was: *go on HHV first, sequence bio-inspired materials behind it, build the data lake before either*. That sequencing is the spine of this paper.

PART II — THE PARADOX IN THE DATA

II

THE PARADOX

Biomass is everywhere. Predictability isn't.

Biomass is everywhere. Northern Europe's forests, Brazil's sugarcane, Indonesia's palm processing, the American Midwest's corn stover, the Indian subcontinent's rice husks — billions of tonnes of dry matter per year, distributed across exactly the geographies where energy demand growth is highest.

Yet biomass remains the energy industry's perennial "five-years-out" story. The Directive 2009/28/EC mandate in Europe (20% renewables by 2020, since revised upward) leaned heavily on biomass to plug the gap that wind and solar couldn't fill. The IEA's net-zero scenarios all assume biomass contributes 15–25% of primary energy by 2050. And yet Europe's bioenergy capacity additions have run consistently below projections every year for a decade.

The reason isn't policy. The reason is that biomass is **a portfolio of feedstocks, not a commodity**, and the conversion routes (combustion, gasification, pyrolysis, hydrothermal liquefaction, fermentation) all have non-trivial sensitivity to input variability. Each batch of woody biomass that arrives at a power plant has a different moisture content, ash fraction, ultimate composition, and trace-element profile — and each of those variables shifts the unit economics by single-digit percentages. Across an annual operation, that variance compounds into multi-million-dollar swings.

The traditional answer has been **conservative averaging**: assume the worst-case batch and operate accordingly. That works at the cost of yield, but it cannot capture the upside in good batches. The result is a sector that has been technically capable of running profitable biomass operations for thirty years and has nevertheless underbuilt its capacity additions in fifteen of the last twenty.

> THE 'BIOMASS PARADOX' IN ONE SENTENCE

> Biomass is abundant, distributed, politically deployable, and chemically flexible — but it cannot be valued, routed, or priced like a commodity unless every batch is characterised in real time.

This is where AI enters — not as a chemistry replacement, but as the characterisation-and-routing layer that has been missing for the better part of a century.

> **Biomass has been treated as a fuel of last resort for fifty years. AI doesn't change the chemistry — it changes the economics. The companies that build the data infrastructure first will route the entire downstream chain.**
>
> — EARTHSCAN ENERGY TRANSITION AI PRACTICE

Why now

Three things have changed in the last 36 months that make this a 2026 conversation rather than a 2030 conversation.

FOUNDATION **2010**

Empirical correlations dominate

Channiwala-Parikh and similar regressions estimate HHV from ultimate analysis to single-digit error. Good enough for clean coal; weak in the long tail of unusual feedstocks.

INFLECTION **2014**

First credible neural networks for HHV

Ghugare et al. publish the first ANN approach to outperform empirical correlations on biomass HHV prediction. Modest dataset, but the architectural template lands.

2018

AI for delignification process control

Valim et al. apply ML to identify experimental conditions in supercritical-CO_2 delignification — the first concrete example of AI sitting inside a chemistry process loop, not just upstream of it.

INFLECTION **2021**

Foundation models for materials

Pre-trained models on QM-9, Materials Project, OQMD become downloadable and fine-tunable on a single GPU. Six-figure compute budget collapses to four-figure. The MIMEA scoping work happens here.

INFLECTION **2022**

Equivariant GNNs replace ad-hoc descriptors

MACE, Allegro, NequIP — networks that respect physical symmetries by construction — reach quantum-chemical accuracy with orders-of-magnitude less data than their predecessors.

2024

Regulatory teeth — CBAM, SBTi, Scope 3

EU Carbon Border Adjustment Mechanism, SBTi net-zero standard, Scope 3 disclosure rules all require batch-level provenance for renewable feedstock claims. AI becomes the only way to generate that data at industrial volumes.

WHERE WE ARE NOW **2026**

Production-grade in a single GPU year

All three forces compound: cheap models, accurate symmetry-respecting architectures, regulatory demand for the data. A focused pilot runs in 6–12 months. The bottleneck has shifted entirely from compute to data pipelines.

First, the AI cost curve. Foundation models for materials and molecules — pre-trained on terabytes of QM-9, Materials Project, and ChEMBL — are now downloadable and fine-tunable on a single GPU. What used to require a six-figure compute budget to even attempt now requires a four-figure one. The marginal cost of a materials prediction has collapsed.

Second, equivariant graph neural networks have replaced ad-hoc descriptors. Models like NequIP, Allegro, and MACE respect the physical symmetries of atomic systems by construction. They reach quantum-chemical accuracy with orders-of-magnitude less data than their predecessors. For atomic-scale lignin modelling, this is the difference between *needs a national-lab data partnership* and *can be done by a sufficiently determined startup*.

Third, regulatory pressure on green claims is forcing auditability. The EU's Carbon Border Adjustment Mechanism (CBAM), Scope 3 emissions disclosure rules, and the SBTi net-zero standard all require **batch- level provenance and lifecycle accounting** for any feedstock that claims renewable status. A pulp mill that wants to sell organosolv lignin into a green-premium chemical chain needs auditable per-batch characterisation — not annualised averages. AI is the only way to generate that data at the volumes required.

These three shifts converge on a window of 18–36 months. Build the data infrastructure now and you set the standards. Wait, and you inherit somebody else's.

Who's already moving

The materials-AI conversation is no longer hypothetical. A handful of organisations are already building durable infrastructure:

Citrine Informatics has spent a decade building a materials- informatics platform with industrial customers across automotive, aerospace, and consumer chemicals. Their Sequential Learning approach is a reference for how active learning closes the synthesise → measure → predict loop in real industrial settings.

The Materials Project (Lawrence Berkeley) and **OQMD** (Open Quantum Materials Database) have produced computed properties for millions of materials — the largest open training corpora available for materials AI. Anything new today builds on these foundations.

Mitsubishi Chemical Group publicly disclosed in 2023 a multi-year investment in deep learning for polymer property prediction, with an explicit mandate to compress R&D cycles for new functional plastics. The strategic rationale: polymer chemistry has hit diminishing returns on traditional QSPR models; the path forward is multimodal DL.

Covestro (the Bayer MaterialScience spinoff) has been public about ML-driven discovery of polyurethane substitutes — directly relevant to lignin's most-cited "high-value alternative use" pathway.

Schmidt Futures has funded multiple academic groups (MIT Buehler lab, EPFL, University of Toronto) on the explicit thesis that materials discovery is the next frontier for AI investment after biology and language.

Dow has run an internal AI-for-chemicals programme since 2019, focused on R&D acceleration. Public disclosures suggest meaningful shifts in how their formulation chemists work, though specifics are guarded.

> **What's happening at Citrine and the Materials Project is what was happening at Schlumberger and Halliburton ten years ago. Those who built the data and the tooling first now own the workflow.**
>
> — INDUSTRY OBSERVER, MATERIALS INFORMATICS, 2025

The lesson from oil and gas is instructive. Schlumberger's DELFI and Halliburton's DS365.ai were not first-of-their-kind technologies — the underlying ML had existed for years. What made them durable was the **integration with existing workflows, the data lake under them,**

and the customer relationships that made the data accessible. A materials-AI play in 2026 follows the same pattern: the model is the easy part; the data infrastructure and the operational integration are what determines who wins.

For biomass specifically, the field is wide open. None of the named players above have a focused biomass programme. The pulp-and-paper industry has been investing in process automation for decades but has not had the AI / data-engineering capacity to translate that operational data into material-property predictions at scale. The gap between *we have sensors* and *we can route batches per-property in real time* is exactly the gap MIMEA addresses.

> **“Chemistry is data. The bottleneck is not first-principles theory, it is the prediction layer between heterogeneous feedstock and priceable downstream input.”**
>
> — A REMINDER WE KEEP ON THE LAB WHITEBOARD

III

WHERE THE AI SITS

Neural surrogates for the chemistry stack

The technology — neural surrogates for the chemistry stack

Materials science has historically depended on a hierarchy of modelling techniques, each more expensive and more accurate than the last.

THE ACCURACY / SPEED TRADE-OFF

IV

Experimental characterisation

The ground-truth reference. Runs in a wet lab — by definition the slowest and most expensive option, but its measurements are what every other technique calibrates against.

SPEED · days to weeks ACCURACY · ground truth

III

Density Functional Theory (DFT)

The gold-standard quantum-mechanical method. Compute energy + forces from first principles. Too slow for high-throughput screening, too accurate to ignore.

SPEED · hours per molecule ACCURACY · near-ground-truth

II

Semi-empirical methods

Approximate quantum mechanics with parameterised corrections. Useful when DFT is too slow but force-field accuracy isn't enough.

SPEED · seconds per molecule ACCURACY · fast but limited

I

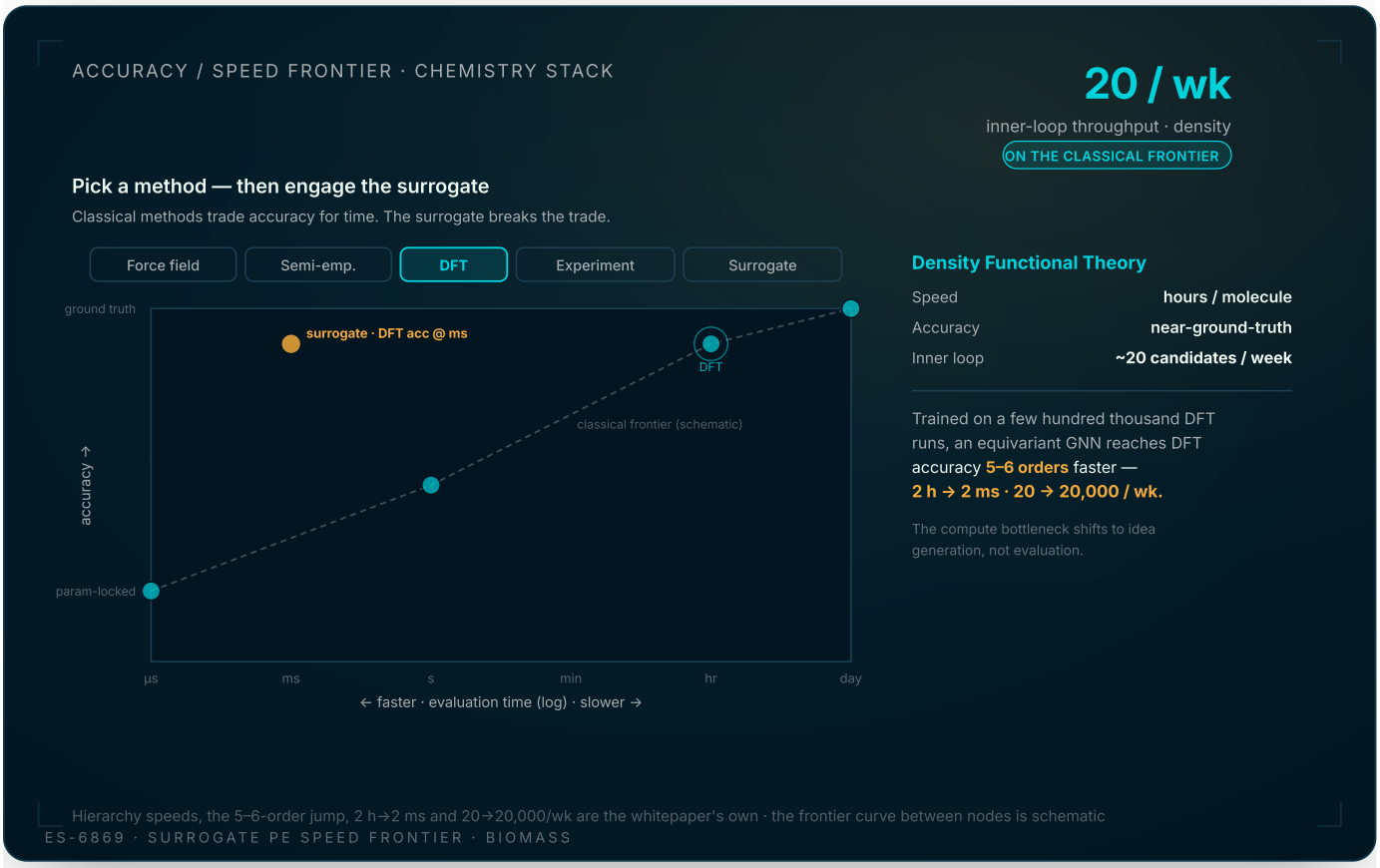
Empirical force fields

Classical mechanics with hand-tuned parameters. Microsecond-scale evaluation enables long-timescale simulation, but accuracy is bounded by the force-field's parameterisation.

SPEED · microseconds per evaluation ACCURACY · parameter-locked

Every research workflow in materials chemistry navigates this hierarchy: cheap-and-approximate at the front of the pipeline, expensive-and-accurate at the back. The bottleneck is **DFT** — too slow for high-throughput screening, too accurate to ignore.

The shift in the last few years is that **neural networks can approximate DFT directly**. Trained on a few hundred thousand DFT calculations, a modern equivariant graph neural network reaches DFT-level accuracy on energy and force predictions while running **5–6 orders of magnitude faster**. A 2-hour DFT run becomes a 2-millisecond inference.



Neural surrogates break the accuracy/speed Pareto frontier. The classical chemistry hierarchy trades accuracy for time — force fields (µs, parameter-locked) → semi-empirical (s) → DFT (hours, near-ground-truth) → experiment (days, ground truth). Pick each method, then engage the neural surrogate: an equivariant GNN trained on DFT reaches DFT-level accuracy at millisecond speed — a point that sits off the frontier (high and left) — turning a 2-hour DFT run into a 2-ms inference and a 20-candidate/week inner loop into 20,000. Hierarchy speeds, the 5–6-order jump and the throughput numbers are the whitepaper's own; the interpolating frontier curve is schematic.

That speedup transforms the inner loop of materials design. A research chemist who previously evaluated 20 candidate molecules per week can now evaluate 20,000. The bottleneck shifts from *compute* to *idea generation* — and idea generation is where generative models contribute.

The four building blocks

A modern materials-AI stack has four interconnected components:

1. Equivariant graph neural networks. Atoms become nodes, bonds become edges, the entire molecule becomes a graph. The "equivariant" part is critical: the network's predictions respect the rotational and translational symmetries of physics by construction. This means a ten-atom training set teaches the network as much as it would teach a human chemist — not less. **Reference architectures: NequIP (Batzner et al., 2022), Allegro (Musaelian et al., 2023), MACE (Batatia et al., 2022).**

2. Multi-modal data ingestion. Real-world materials data lives in many forms — SMILES strings, XYZ coordinates, microstructural images, spectra (NMR / IR / Raman / XRD), processing logs, lab notebooks. A production materials-AI system reads all of these and represents them in a shared embedding space. **Reference: the foundation-model literature on multi-modal training, adapted to materials.**

3. Active learning loops. The model proposes candidate molecules or processing conditions; the lab evaluates a curated subset; the results feed back into the next training cycle. Citrine's Sequential Learning is the canonical industrial example. The leverage is enormous: an active-learning system needs **5–10× fewer experiments** to reach the same Pareto frontier as a brute-force screen.

4. Closed-loop deployment. The trained model is wired into MLOps infrastructure that monitors drift, retrains on fresh data, and versions every prediction. For biomass specifically, the closed loop extends to process-line sensors that feed the model live PA / UA data and routing actuators that move batches based on predictions.

5–6 orders

of magnitude speedup over DFT

Source: Equivariant GNN benchmarks, Batzner et al. 2022

5–10×

fewer experiments needed under active learning

Source: Citrine Sequential Learning case studies, 2020–2024

Six durable advantages

Adopting this stack — neural surrogates, multi-modal training, active learning, closed-loop deployment — produces six advantages over the status quo. Each is independently meaningful; the combination is transformational.

Speed of evaluation

- 10–100× faster inner loop for materials design once the surrogate is trained
- DFT-quality predictions in milliseconds, enabling real-time decisions
- Per-batch HHV inference fast enough to gate a process-line routing actuator

Multi-modal inputs

- Combine atomistic structures, microstructural images, spectra, and processing logs
- No single source of truth required; the model learns the alignment
- Materials data is fragmented by definition — multi-modal handles that natively

Cloud / On-prem / Edge

- Same trained model runs in a research cluster or on a process-line edge device
- On-prem deployment for industrial customers with data-sovereignty constraints
- Edge inference on Jetson / Coral hardware for sub-second control loops

Improvement in characterisation accuracy

- DL outperforms classical regression on long-tail biomass samples
- Accuracy gains compound when training corpus spans multiple feedstock families
- Transfer learning from public datasets (Materials Project, OQMD) reduces in-house training burden

Do more with less data

- Foundation models on materials reduce per-domain training requirements by 10–100×

- Active learning flags the 5% of experiments that move the model most
- Negative results carry signal too — the model improves from failed syntheses

Closed-loop deployment

- MLOps wraps drift monitoring, retraining, and versioned predictions
- Lab automation (LCMS / GC / FTIR) feeds the loop with measured outcomes
- Auditable per-batch records for CBAM / SBTi / green-chain compliance

> **WHERE THIS MATTERS MOST FOR BIOMASS**
>
> The combination of **multi-modal inputs** (PA + UA + microstructural images + processing logs) and **closed-loop deployment** (sensor → model → routing actuator) is what unlocks per-batch routing on a working process line. Most lone advantages buy you a research demo; that combination buys you operational deployment.

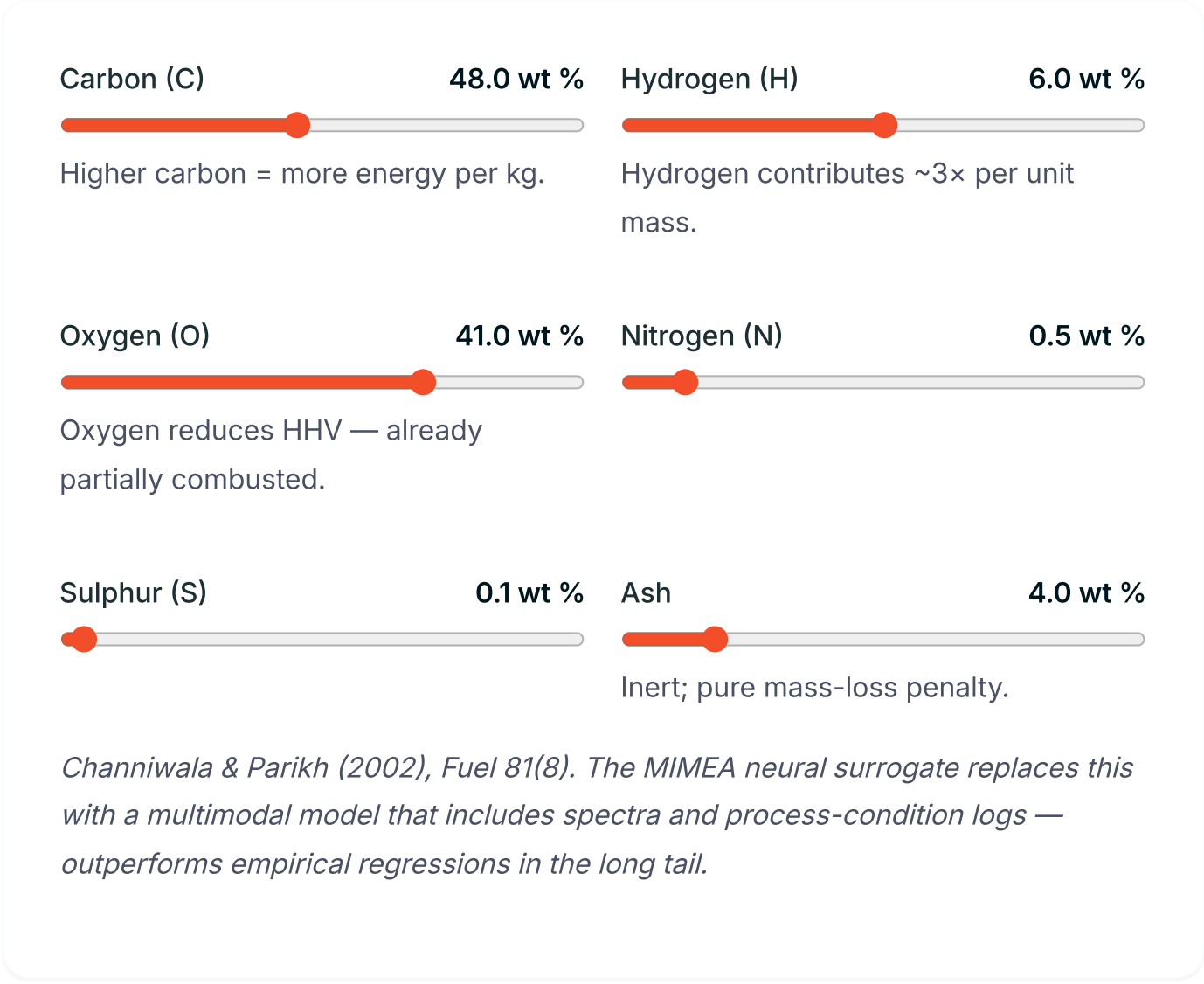
Use case in depth — Higher Heating Value without a calorimeter

Of every commercial question we surveyed in scoping MIMEA, one came up more often than any other: *what is the HHV of this batch?* It is the question that determines combustion economics, gasifier yield, contractual settlement, and downstream chemical viability. It is asked thousands of times per year per processing plant.

The answer comes today from a **bomb calorimeter**: ~1 g of sample is ground, pelletised, sealed in a high-pressure oxygen vessel, ignited, and the temperature rise of the surrounding water bath is recorded. Total turnaround: 4–8 hours including prep. Per-sample

cost (labour and reagents): $50–200. Per-sample throughput: 10–20 samples per calorimeter per day under good conditions.

This is fine when HHV is a quality-assurance check on annual averages. It is catastrophic when HHV is the input to a real-time routing decision — and routing is where the upside lives.



Channiwala & Parikh (2002), Fuel 81(8). The MIMEA neural surrogate replaces this with a multimodal model that includes spectra and process-condition logs — outperforms empirical regressions in the long tail.

What we know about the chemistry

Lignin's HHV is determined primarily by its elemental composition, which is in turn determined by source species and processing route. Anchor numbers from the MIMEA report:

- **Dry lignin HHV:** 23.25 – 27.85 MJ/kg (depending on isolation method)
- **Dry, ash-free lignin HHV:** 23.95 – 28.36 MJ/kg

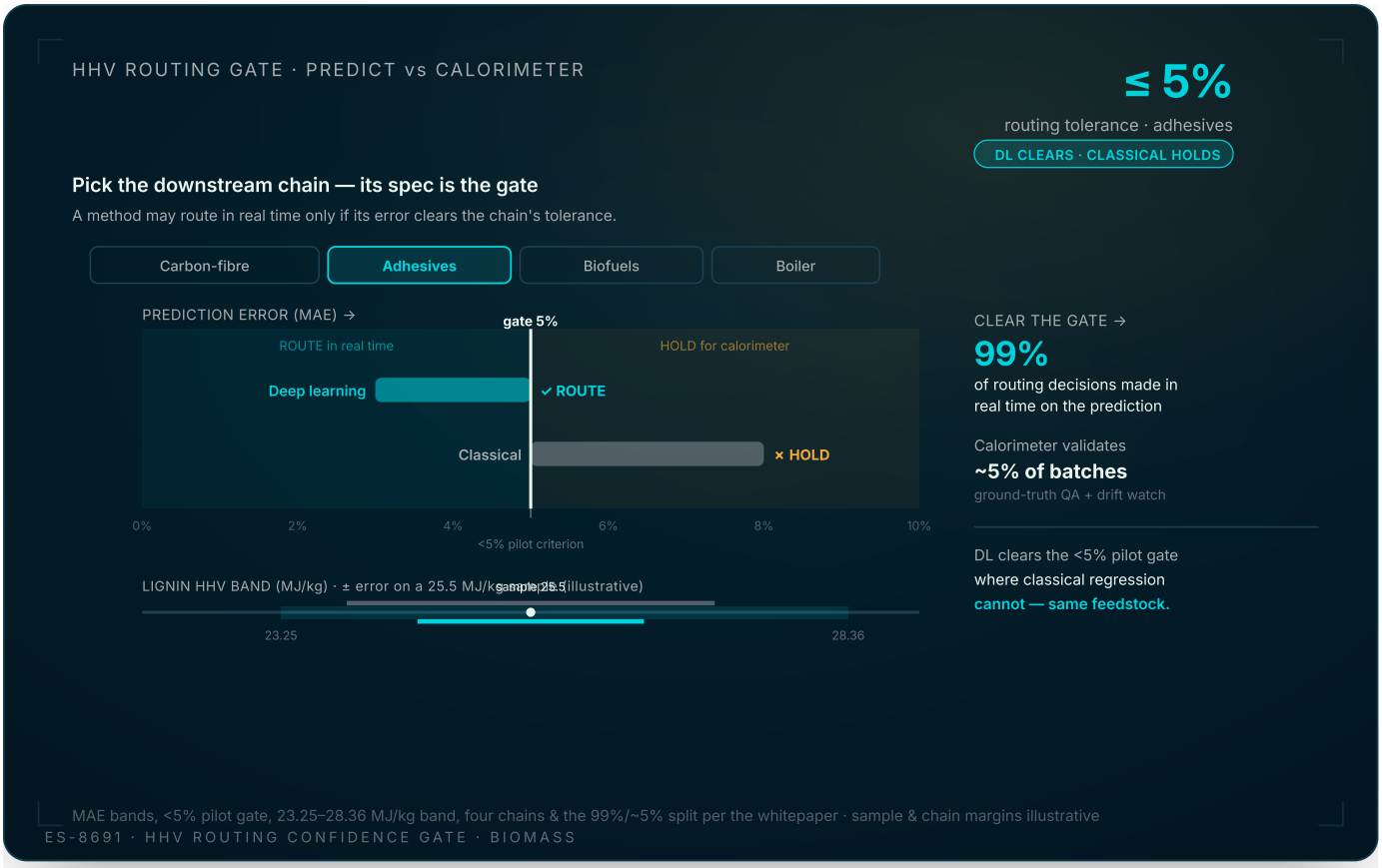
- **Reference comparison:** ~50% higher than cellulose at equivalent dryness
- **Carbon content:** ~60% (vs ~37–56% for general lignocellulosic biomass)
- **Oxygen content:** ~30% (lower oxygen → higher HHV)
- **Hydrogen content:** under 6% — implies high C/H ratio, correlated with high HHV

This last point is structural: **carbon-rich, oxygen-poor, low-hydrogen biomass is high-energy biomass.** The model's job is to predict HHV from cheaper measurements that proxy these underlying compositional realities.

Two prediction routes

Researchers have established two empirical routes for HHV prediction from cheaper analyses:

Proximate Analysis (PA) → HHV. PA measures moisture, volatile matter, fixed carbon, and ash content. These are inexpensive gravimetric measurements, available on any operating biomass line. Classical regression gets to ~5–8% MAE; modern DL closes that to ~3–5%, and crucially generalises better across feedstocks.



HHV error tolerance is the gate that decides whether a batch routes in real time or waits for the calorimeter. Each downstream chain has a spec margin; pick one and its tolerance becomes the gate. Deep learning (~3–5% MAE) clears the pilot's <5% routing gate where classical regression (~5–8% MAE) does not — and only a cleared gate lets 99% of routing decisions stop waiting for the calorimeter (which then validates ~5% of batches). The MAE bands, the <5% pilot criterion, the 23.25–28.36 MJ/kg band, the four chains and the 99%/~5% split are the whitepaper's own; the example sample and the chain spec margins are illustrative.

Ultimate Analysis (UA) → HHV. UA measures elemental composition (C, H, N, O, S). More informative but more expensive — typically requires combustion analysis or XRF/CHN gear. Best HHV models combine PA + UA features when both are available.

In MIMEA's scoping, the highest-leverage opportunity was the **temporal variant**: predict HHV concurrently with combustion-product gas concentrations (CH_4, CO, CO_2, H_2) from the time series of a gasification or pyrolysis process. This turns HHV from a static lab metric into a **control-room signal** — the operator sees predicted HHV update second-by-second as the process runs.

DEPLOYMENT SHAPE — WHAT 'GOOD' LOOKS LIKE

A paper-mill or biorefinery deployment of MIMEA-style HHV prediction looks like:

1. Process-line sensors (NIR / FTIR / mass-flow / temperature) feed the model continuously
2. Model outputs predicted HHV, ash fraction, and chemical fingerprint per batch
3. Routing actuator directs each batch to the highest-value downstream chain (carbon-fibre precursor, adhesives, biofuels, or boiler)
4. All decisions logged to an immutable audit trail for green-claims compliance

The chemistry doesn't change. The model doesn't replace a calorimeter for high-stakes contractual measurements — those still happen, just less often. What changes is that 99% of routing decisions get made in real time on the model's prediction, and the calorimeter validates a sampled subset for QA.

Hill of pain — the current biomass workflow

STEP 01

Sample collection

Operator pulls a representative sample from the process line — either by hand or via an automated sampler — and packages it for the laboratory queue.

TIME **Manual** per batch

STEP 02

Grind + prepare

Sample is ground to a fine powder and dried to a known moisture state. Inconsistent prep is the single largest source of HHV-result variance.

TIME **30–60 min** manual prep

STEP 03

Pelletise into combustion capsule

Powder is compressed into a capsule of known mass and seated in the calorimeter's combustion vessel. Tolerance: ±0.1 g.

TIME **10–15 min** per sample

STEP 04

Load into bomb calorimeter

Capsule is sealed in the oxygen-pressurised vessel. The calorimeter queue is the throughput bottleneck — 10–20 samples / day per instrument.

TIME **Sequential** queue depth bound

STEP 05

Combustion run

Oxygen-pressurised combustion ignites the sample. Temperature rise of the surrounding water bath is logged at high frequency.

TIME **4–8 hrs** wall clock

STEP 06

Temperature / pressure reading

Operator reads peak temperature, pressure rise, and combustion duration from the instrument's logged data.

TIME **~5 min** manual readout

STEP 07

HHV calculation

Manual or instrument-software calculation converts temperature rise into MJ/kg, applying calibration corrections and water-vapour latent-heat terms.

TIME **Manual** spreadsheet / instrument

STEP 08

Result to lab notebook / LIMS

Per-batch HHV is recorded in the lab system. By the time this row exists, the feedstock is already in the boiler or storage silo. No real-time routing is possible.

TIME **Annual averages** drives contracts

Each step is defensible in isolation. The cumulative latency — measured in hours per batch and quarters per fleet-wide rollout — is what kills the upside.

The pattern is familiar to anyone who has worked in a process- chemistry environment: a slow, sequential, expensive measurement loop that bottlenecks operational decision-making. Each step is defensible in isolation. The cumulative latency is what kills the upside.

Concrete pain points surfaced in the MIMEA scoping interviews:

- **Per-sample cost:** $50–200 in labour + reagents per HHV measurement
- **Per-sample time:** 4–8 hours including prep, longer if the calorimeter queue is full
- **Throughput ceiling:** typically 10–20 samples per calorimeter per day under ideal conditions; less in practice
- **No batch-level routing:** by the time HHV is known, the feedstock is already in the boiler or storage silo

- **Specialist labour:** trained operator required; the operator's time is the binding constraint on most days
- **Calibration drift:** calorimeters need recalibration every few weeks; results during the drift window are quietly noisier
- **No green-claims auditability:** the data exists in a LIMS or spreadsheet but is rarely connected to the contractual settlement that referenced it

The cost of the calorimeter itself is not the issue. The cost of the **workflow built around it** — the operator time, the queue, the six-hour wait, the inability to act on real-time signal — is what caps the value extraction from the entire biomass operation.

Faster, streamlined and reproducible — the MIMEA flow

STEP 01

Process-line sensors

NIR, FTIR, mass-flow, temperature, and moisture sensors stream the same signals the plant already collects for QC. No new instrumentation.

TIME **Sensors / sec** existing telemetry

STEP 02

MIMEA AI model

Multimodal neural surrogate trained on PA / UA / spectra / process logs. Runs cloud, on-prem, or at the edge. Same model artefact across all three.

TIME **Sub-second** inference latency

STEP 03

Predicted HHV + chemical fingerprint

Per-batch HHV in MJ/kg + a multidimensional chemical fingerprint that informs downstream routing decisions. Confidence interval propagated.

TIME **~seconds** end-to-end

STEP 04

Routing decision

Each batch is routed to the highest-yield downstream chain (carbon-fibre precursor / adhesives / biofuels / boiler) its predicted properties qualify for.

TIME **Per batch** real-time

STEP 05

Downstream chain

Routed feedstock enters the appropriate downstream process. Each chain has different yield economics — picking the right one per batch is where MIMEA's value lives.

TIME **Multi-chain** value-maximised

STEP 06

Audit-logged settlement

Every routing decision is stamped with model version + input features + prediction confidence + timestamp. Green-claim ready, regulator-defensible.

TIME **Immutable** audit trail

STEP 07

Periodic calorimeter QA

Calorimeter runs continue on ~5% of batches as ground-truth validation. Confirms model accuracy + flags drift before it matters operationally.

TIME **~5%** of batches

STEP 08

Feedback into model retraining

QA-measured ground truth feeds back into the next training cycle. Closed-loop MLOps: drift detected → model retrained → updated artefact deployed.

TIME **Continuous** closed loop

The calorimeter doesn't disappear. The workflow built around it does.

The shift is from *measure-then-decide* to *predict-and-route*. The calorimeter doesn't disappear — it remains the ground-truth instrument for QA and contractual disputes. What changes is the **frequency and purpose of its use**: from primary measurement of every batch, to periodic validation of a sampled subset.

TODAY

Bomb calorimeter required for every measurement. **4–8 hours per sample** including prep, **$50–200** in labour and reagents, throughput capped at 10–20 samples per calorimeter per day. The operator is queue-bound.

CALORIMETER-BOUND

4-8h

per-sample wallclock

$50–200 COST · 10–20 SAMPLES/DAY

TOMORROW

Marginal-cost neural surrogate inference. **Sub-second** per-sample latency, **~cents** per inference, throughput unbounded by the model. The calorimeter shifts from primary instrument to periodic-QA validator.

NEURAL-SURROGATE

~1s

per-sample wallclock

CENTS · UNBOUNDED

COMPOUNDING

The throughput unlock compounds as operators feed measured outcomes back into retraining. Each shipped batch is a labelled example. **Months → quarters → years**, the surrogate gets quietly more accurate on the operator's own feedstock.

CLOSED-LOOP

∞

cumulative learning

EVERY BATCH BECOMES A LABEL

Concrete operational changes:

- **Per-sample cost:** marginal model inference (~cents)
- **Per-sample latency:** sub-second prediction; sub-minute routing
- **Throughput:** unbounded by the model — bounded only by sensor sampling rate and downstream actuator speed
- **Batch-level routing:** every batch goes to the highest-yield downstream chain its predicted properties qualify for
- **Operator role:** shifts from per-batch measurement to exception-handling on the routing decisions
- **Calibration:** the model's drift is monitored continuously; retraining on QA data closes the loop
- **Green-claims auditability:** every routing decision is logged with the model version, input features, and prediction confidence

THE DELTA

100×

→ seconds

Faster characterisation per batch

$50–200

→ ~$0.01

Per-sample lab cost today

10–20

→ unbounded

Samples per calorimeter per day

Annual

→ per-batch

Granularity of green-claims today

Side by side — what changes, what doesn't

CONVENTIONAL METHOD

- Bomb calorimeter required for every measurement
- 4–8 hours per sample including prep
- $50–200 per-sample cost (labour + reagents)
- Throughput capped at ~10–20 samples per calorimeter per day
- Static lab metric — annual averages drive contracts
- No per-batch routing of feedstock
- Calibration drift introduces silent error windows
- Green claims rely on annual paperwork; no batch provenance
- Single-feedstock empirical formulas; long-tail samples mispredict

- Operator time is the binding constraint on throughput

MIMEA AI MODEL

- Deep learning on existing PA + UA + sensor data
- Sub-second prediction; sub-minute routing
- Marginal cost per inference (~cents)
- Throughput bounded only by sensor sampling and actuator speed
- Real-time control-room signal — every batch valued individually
- Per-batch routing to the highest-value downstream chain
- Continuous drift monitoring with retraining on QA data
- Auditable per-batch records for CBAM / SBTi / green-chain compliance
- Multi-feedstock model; transfer learning across families
- Operator shifts from per-batch measurement to routing exception-handling

The comparison is not *AI replaces calorimeter*. It is *AI replaces the workflow built around the calorimeter*. The calorimeter remains the ground-truth instrument; what changes is that 95% of operational decisions stop waiting for it and start running on the model's prediction with calorimeter-validated subsampling.

This is the same pattern that played out in well-log digitisation with VeerNet: the underlying physics didn't change; what changed was that the workflow stopped requiring a human in the inner loop. In both cases, the productivity unlock came from removing the serialisation point, not from inventing new science.

> **“Reduce uncertainty, predict and define outcomes, automate complex processes, and optimise your experts' time.”**

— THE EARTHSCAN OPERATING PRINCIPLE, APPLIED TO SUBSURFACE AND NOW TO BIOMASS

Beyond HHV — bio-inspired materials design

HHV prediction is the wedge. The bigger ceiling — and the longer research timeline — is **atomic-scale design of bio-inspired materials** using lignin as the substrate.

Lignin's structural complexity, which is its commercial weakness for combustion, is its strength for materials chemistry. The polyaromatic backbone offers anchor points for functional-group modification that yield carbon-fibre precursors, polyurethane substitutes, vanillin, phenolic resins, and a long tail of specialty chemicals. Most of these modifications have been demonstrated at lab scale. None have hit commodity-volume commercial deployment.

The bottleneck has been **the cost of evaluating the design space**. A medicinal chemist designing a small molecule has the luxury of millions of compounds catalogued with measured properties — the training corpus for ML is dense. A polymer chemist designing a lignin- derived material works in a sparse, heterogeneous space where each candidate requires expensive synthesis and characterisation to evaluate. Brute-force screening is impractical; intuition-driven design is slow.

Neural surrogates of potential energy surfaces (PES) change this equation. The pattern, validated in adjacent polymer fields:

1. **Build a training corpus** of DFT-evaluated structures across the relevant chemical space. Public datasets (Materials Project, QM9, OQMD) cover the inorganic baseline; lignin-specific computations need to be generated or partnered.
2. **Train an equivariant GNN** (NequIP-class architecture) on the corpus. Reach DFT-quality energy and force predictions at millisecond inference.

3. **Wrap the surrogate in a generative model.** Diffusion models over molecular graphs have surpassed autoregressive approaches for novel-structure generation. The generator proposes candidates; the surrogate scores them; the highest-scoring subset goes to lab synthesis.
4. **Close the loop with active learning.** Lab-synthesised candidates feed measured properties back into the training corpus. Sequential Learning (Citrine's term) selects the next batch to synthesise based on which experiments will most reduce model uncertainty.

The MIMEA scoping concluded that this path is **technically tractable but commercially patient.** A 24-month investment can produce a research-grade demonstrator across one or two functional-group modifications (carbon-fibre precursor, polyurethane substitute). A 60-month investment can produce a productionised pipeline that materially shifts the lignin valorisation industry.

By contrast, HHV prediction can ship a deployable product in 9–12 months. **MIMEA's recommendation: ship HHV first, sequence bio-inspired materials behind it.**

WHAT THIS WHITEPAPER DOES NOT CLAIM

- We do not claim a finished bio-inspired-materials product.
- We do not claim to have synthesised novel polyurethane substitutes.
- We do not claim that the MIMEA feasibility produced a working HHV prediction model in production today.
- We claim only that the technical path is well-understood, the data infrastructure is buildable, the commercial value is defensible, and the right sequencing is HHV-first.

IV

THE WAY FORWARD

Three recommendations for operators ready to move

Way forward — three recommendations

For an organisation considering where AI fits in its biomass / materials roadmap, three concrete moves in the next 6–12 months:

1. Build the data lake before the model

Almost every materials-AI project that fails does so because the data infrastructure was retrofitted after the modelling work began. The expensive part of MIMEA-style deployments is **the ingestion pipeline that consolidates fragmented PA / UA / sensor / lab-notebook data into a coherent training corpus**, with provenance, units, and lineage preserved.

Concrete first steps:

- Inventory existing data sources (LIMS, plant historians, ELN, spreadsheets in shared drives, contract-lab PDFs)
- Define a canonical schema for sample, batch, and process-condition records
- Stand up a versioned data warehouse (Snowflake / Databricks / self-hosted Postgres) with strict schema enforcement
- Backfill 12–24 months of historical data; this is the training corpus your first model uses
- Wire ingestion of new data so the corpus grows organically

Budget guideline: $200K–$500K for the data infrastructure pass, depending on existing maturity. **Spend this money before you spend any model-development money.**

2. Pilot HHV prediction first

Choose one process line at one facility. Define a tight pilot scope: predict HHV from PA + UA + sensor data, validate against weekly calorimeter measurements, route batches based on predicted yield. Six-month timeline; success criterion is HHV MAE under 5% and a demonstrated commercial value from the routing decisions.

The pilot does several things at once:

- Validates the data infrastructure under realistic load
- Produces a reference deployment that subsequent models can reuse
- Demonstrates ROI to the executive sponsor before broader rollout
- Surfaces the operational integration questions (who responds to routing exceptions, how does the model version land in production, who owns retraining cadence) at a manageable scale

3. Engage stakeholders across all three levels

The MIMEA scoping interviews surfaced that successful biomass-AI deployments need stakeholder alignment at three levels:

- **Operators** — paper mills, biorefineries, biomass power plants who own the process line and the routing decisions
- **Government** — environment ministries, energy regulators, carbon-credit certifiers who set the green-claims rules
- **Insurers / certifiers** — Bureau Veritas, SGS, TÜV, and equivalents who audit the green claims for downstream contracts

Engaging only one of these produces a research demonstrator. Engaging all three produces a deployable system whose outputs flow into contractual settlement and regulatory compliance from day one.

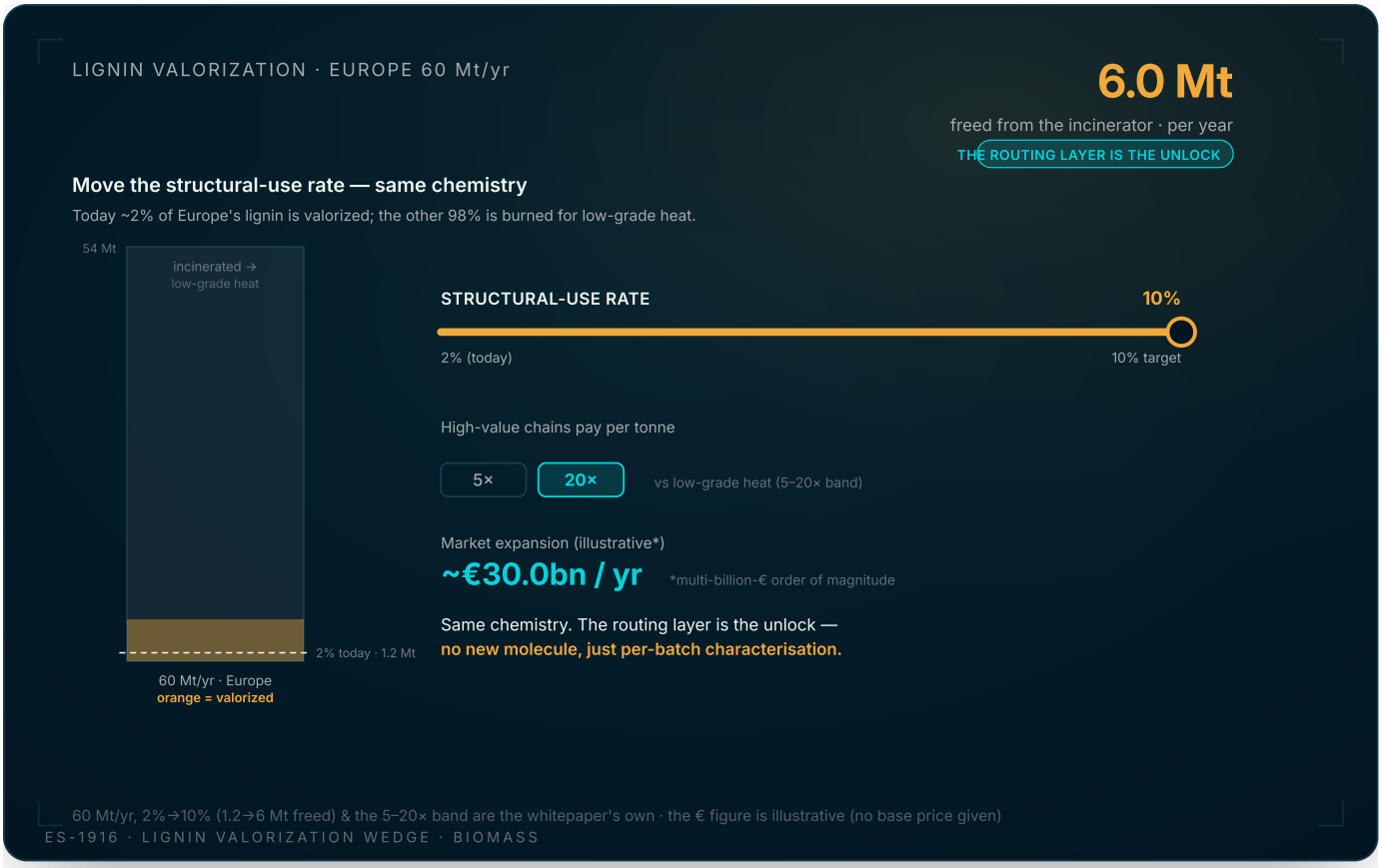
> **WHAT MIMEA'S AUTHORS WOULD DO TOMORROW**
>
> If we were starting today with a 24-month budget:
>
> - **Months 0–3:** data infrastructure pass + stakeholder discovery

- **Months 3–9:** HHV prediction pilot at one facility
- **Months 9–15:** scale HHV across 3–5 facilities; begin bio-inspired materials proof-of-concept
- **Months 15–24:** productionise HHV; ship first bio-inspired materials demonstrator

This is the sequencing the MIMEA feasibility recommended. It still holds in 2026.

What it would mean if the industry got this right

Run the numbers at a high level. Europe alone produces roughly **60 million tonnes of lignin per year** as a paper-industry waste stream. At the current ~2% structural-use rate, ~1.2 Mt/yr enters high-value chains. Push that to **10%** through better characterisation and per-batch routing, and you free **6 Mt/yr** of lignin from the incinerator into industries that pay 5–20× more per tonne.



The lignin opportunity is a routing problem, not a chemistry one. Europe makes ~60 Mt/yr of lignin; today only ~2% (1.2 Mt/yr) is structurally valorized and the rest is incinerated. Drag the structural-use rate from 2% to 10% and the freed tonnage crosses from the grey incinerator band into the high-value chains — 6 Mt/yr at 10%, into industries that pay 5–20× more per tonne, with no new chemistry. The 60 Mt total, 2%→10% (1.2→6 Mt) and the 5–20× band are the whitepaper's own; the euro figure is an illustrative order-of-magnitude (no base price is given).

That single shift, conservatively valued, is **multi-billion euro annual market expansion** for the European pulp-and-paper sector, plus equivalent value capture for the downstream chemical industries that absorb the now-priceable lignin supply. And it does so **without new chemistry** — the routing capability is what's missing, not the underlying processes.

That is what AI infrastructure does for the biomass economy. Not a moonshot. Not a new molecule. Just the missing characterisation-and- routing layer that has been deferred for fifty years because the data plumbing was hard.

It is no longer hard.

> **“The next industrial economy won't be built on harder-to-extract barrels. It will be built on harder-to-model molecules — and AI is what makes them tractable.”**

— EARTHSCAN ENERGY TRANSITION AI PRACTICE, MAY 2026

WHAT TO REMEMBER

1. Lignin is the second-most-abundant biopolymer on Earth, generated as a **100M-tonne annual waste stream**. **<2%** is sold as a structural input today.
2. The bottleneck is not chemistry, geology, or capital — it is the **prediction layer** between heterogeneous feedstock and priceable downstream input. AI is the missing layer.
3. The MIMEA flow shifts the calorimeter from primary instrument to QA validator. **4-8h per sample → ~1s**; **$50-200 → cents**; throughput unbounded by the model.
4. Equivariant GNNs (MACE, Allegro, NequIP) match DFT accuracy at **100×** the speed. The compute bottleneck has shifted entirely to data pipelines.
5. Closed-loop deployment compounds: every shipped batch becomes a labelled example for retraining. The surrogate gets quietly more accurate on the operator's own feedstock over months → quarters → years.

Get the full whitepaper

This page is the long-form summary. The complete 24-page MIMEA whitepaper includes:

- The full MIMEA feasibility methodology and decision framework
- Detailed equivariant-GNN model architecture diagrams
- A worked example of HHV prediction on a North-European hardwood feedstock dataset

- The proposed data-lake schema in full
- Stakeholder-engagement playbook for the three-level alignment
- Bibliography (35+ citations)
- Authors' notes on what we'd do differently in 2026 vs 2021

About EarthScan

EarthScan is the energy-AI brand in DeepKapha's group, alongside Real AI B.V. (research) and the broader DeepKapha consultancy. Our flagship subsurface products — **ES Raster Digitizer**, **ES W2W Correlation**, and the **VeerNet AI** research line — are deployed at major upstream operators across Europe, the Middle East, and Southeast Asia.

Our research interests extend across the energy-and-materials spectrum: subsurface AI, raster log digitisation, well-to-well correlation, multimodal seismic interpretation, and — as this whitepaper covers — biomass characterisation and bio-inspired materials design.

We collaborate with operators, researchers, and government counterparts. If you want to talk about applying any of this to your own data and processes, the conversation is one we'd genuinely enjoy having.

GLOSSARY

Active learning

Training-loop strategy where the model picks which data points to label next. In materials work it cuts the wet-lab evaluation budget by 5–10× by avoiding redundant or uninformative experiments.

DFT

Density Functional Theory — the workhorse quantum-chemistry method for computing electronic structure. Accurate but slow; neural surrogates trained on DFT outputs are 10–100× faster at near-DFT accuracy.

Equivariant

An equivariant neural network is one whose output transforms predictably when the input is rotated or translated. For atomic systems this is a hard requirement, not a nice-to-have.

GNN

Graph Neural Network — a deep-learning architecture that operates on graph-structured data like molecules. Equivariant variants (E(3)-GNN, MACE, Allegro) preserve rotational / translational symmetries, which is exactly what atomic systems require.

HHV

Higher Heating Value — total energy released per unit mass of fuel when fully combusted, including the latent heat of water vapour. The headline number used to price biomass against fossil benchmarks.

Lignin

The second-most-abundant biopolymer on Earth. Heterogeneous, branched, polyaromatic — the structural reason wood is wood. Roughly 100 million tonnes per year are generated as a paper-mill / cellulosic-biofuel waste stream.

LIMS

Laboratory Information Management System — the spreadsheet-on-steroids most labs run for sample tracking

	and result reporting. The default integration target for any AI pipeline that wants to ship to a real biorefinery.
PA / UA	Proximate analysis: moisture, volatile matter, fixed carbon, ash. Ultimate analysis: C, H, N, S, O. The two cheapest characterisation routines a biorefinery already runs.

References

The MIMEA feasibility report (Real AI B.V., 2021) drew on the following primary sources. The 2026 whitepaper version updates and extends with recent equivariant-GNN literature.

Biomass and lignin chemistry

1. Maksimuk, Y., Antonava, Z., Krouk, V., Korsakova, A., & Kursevich, V. (2021). *Higher heating value of lignin from various sources*.
2. Huang, Y. F., & Lo, S. L. (2020). *Predicting heating value of lignocellulosic biomass based on elemental analyses*. Energy.
3. Khunphakdee, P., Kokerd, S., Soanuch, C., & Chalermsinsuwan, B. (2022). *Comparative study of proximate vs ultimate analyses for biomass HHV prediction*.
4. Alejandra, B., Brizuela, M. A., Mazza, G., & Rodriguez, R. (2018). *Lignin biofuel — review of opportunities and constraints*.
5. Tao, J., et al. (2019). *Lignin valorisation across industries — a review*.
6. Sharma, V., Kaur, M., Singh, P., & Arya, S. K. (2021). *Lignin roles in plant biology*.

AI / ML for biomass and materials

1. Ghugare, S. B., Tiwary, S., Elangovan, V., & Tambe, S. S. (2014). *Prediction of higher heating value of solid biomass fuels using artificial intelligence formalisms*. BioEnergy Research, 7(2), 681–692.
2. Xing, X., Luo, J., Wang, S., Gao, X., & Fan, J. (2019). *ANN / SVM / RF for HHV prediction*.
3. Löfgren, J., et al. (2022). *Machine learning across natural science domains — a survey*.
4. Hough, B. R., Beck, D. A. C., Schwartz, D. T., & Pfaendtner, J. (2017). *Comprehensive models of biomass pyrolysis*.
5. Gu, J., et al. (2021). *Machine learning in biomass upgrading and conversion processes*.

6 Valim, I. C., Rego, A. S., Queiroz, A., et al. (2018). *AI for delignification process identification*. Computer Aided Chemical Engineering 43, 1469–1474.

7 Hiraide, K., Hirayama, K., Endo, K., & Muramatsu, M. (2021). *Application of deep learning to inverse design of phase separation structure in polymer alloy*. Computational Materials Science 190, 110278.

8 Zhai, C., Li, T., Shi, H., & Yeo, J. (2020). *Discovery and design of soft polymeric bio-inspired materials with multiscale simulations and artificial intelligence*. Journal of Materials Chemistry B, 8(31), 6562–6587.

Equivariant graph neural networks (added in 2026 update)

1 Schütt, K. T., et al. (2017). *SchNet: A continuous-filter convolutional neural network for modeling quantum interactions*.

2 Satorras, V. G., Hoogeboom, E., & Welling, M. (2021). *E(n) Equivariant Graph Neural Networks*. ICML.

3 Batzner, S., et al. (2022). *E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials*. Nature Communications.

4 Musaelian, A., et al. (2023). *Allegro: scalable equivariant interatomic potentials*. Nature Communications.

5 Batatia, I., et al. (2022). *MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields*. NeurIPS.

Market and policy

1 Global Market Insights (2020). *Lignin Market Sizing*.

2 European Union (2009). *Directive 2009/28/EC on the promotion of energy from renewable sources*.

3 IEA (2024). *Net Zero Roadmap: A Global Pathway to Keep the 1.5 °C Goal in Reach*.

4 European Commission (2023). *Carbon Border Adjustment Mechanism (CBAM) implementation guidance*.

Authors

Tarry Singh — Founder & CEO, DeepKapha / EarthScan. Two decades shipping production AI across financial services, healthcare, and energy. LinkedIn · deepkapha.com

Real AI B.V. research team — the research arm in DeepKapha's group, original authors of the 2021 MIMEA feasibility under the Dutch SNN funding programme.

Published May 2026. © DeepKapha B.V. / EarthScan. All rights reserved.

EARTHSCAN

MIMEA: AI for Biomass and Bio-Inspired Materials

Published May 2026. Generated 2026-07-22.
https://www.earthscan.io/whitepapers/mimea-ai-for-biomass-and-bio-inspired-materials
Copyright EarthScan. All rights reserved.