Beddings, Vugs and Well-to-Well in One Pipeline

Carbonate reservoir characterisation has no single winning algorithm, and pretending otherwise is the most common way subsurface-AI programmes underdeliver. Fractures and bedding planes are sparse, geometric, supervised-learnable features best served by a set-prediction detection transformer; vugs are secondary-porosity blobs with no clean labels, best served by a tuned classical computer-vision pipeline; and field-scale continuity is a spatial-interpolation problem best served by geostatistics, not deep learning at all. A programme that masters one track and stops has automated a sliver of the interpretation and left the reservoir model incomplete. This whitepaper proposes a full-stack reference architecture that fuses all three, drawn from a roughly twenty-month, three-phase formation-evaluation engagement with a mid-sized Middle East carbonate operator we partnered with. We describe each track as an engineered component with its own architecture, training regime and accuracy envelope — a ResNet-10-backboned detection transformer reporting F1 of roughly 65% (fractures) and 63% (beddings) at a 3 cm depth tolerance and roughly 75% and 69% at 5 cm across fourteen wells; a multi-stage classical-CV vug quantifier validated on three wells against expert manual interpretation-software picks, running at under thirty seconds per metre versus four-to-five minutes for the prior morphological baseline; and a kriging-based well-to-well correlation layer that lifted interpretation productivity by roughly 60% and accuracy by roughly 75%. The contribution is not any single model. It is the routing-and-re-convergence pattern that carries one binary image log through three heterogeneous tracks and stitches their outputs into one reservoir-quality field model — and the argument that this fusion, not a better single detector, is what a carbonate operator actually needs to buy or build.

Tannistha Maiti
Senior AI Researcher

Quamer Nasim
ML Research Engineer

Tarry Singh
Founder & CEO

CARBONATE-RESERVOIR-CHARACTERIZATION / SUBSURFACE-AI / REFERENCE-ARCHITECTURE / FRACTURE-DETECTION / VUG-QUANTIFICATION / WELL-TO-WELL-CORRELATION

CONTENTS

A carbonate reservoir does not present one problem. It presents three, and they do not share a method. Fractures and bedding planes are sparse geometric features you can label and learn from a high-resolution image of the borehole wall. Vugs — the secondary porosity that often controls flow in a carbonate — are irregular dark blobs with no clean ground truth, a classical computer-vision problem in disguise. And the question every asset team actually asks — does this fracture set continue to the next well, does this productive zone correlate across the field — is a spatial-interpolation problem that no per-well detector can answer, because the answer lives in the rock *between* the wells. Three problems, three disciplines: supervised deep learning, tuned classical CV, and geostatistics. The mistake that quietly hollows out most subsurface-AI programmes is to master one and present it as a platform.

This whitepaper makes the opposite argument. Over roughly twenty months, across three phases, working with a mid-sized Middle East carbonate operator we partnered with, we built and handed over all three tracks — a transformer-based fracture-and-bedding detector, a classical-CV vug quantifier, and a kriging-based well-to-well correlation layer — and the more durable contribution turned out not to be any one of them. It was the architecture that fuses them: a single pipeline that routes one binary image log into three heterogeneous engineered tracks and re-converges their outputs into one reservoir-quality field model. That routing-and-re-convergence pattern is the reference architecture this paper proposes, and the case for it is simple — no single track sees the whole rock.

Why one method cannot win

It is tempting to believe that a sufficiently large neural network could learn fractures, vugs and continuity end to end. In a carbonate, with the data an operator actually has, it cannot — and understanding *why* is the entire justification for a multi-track design.

Fractures and bedding planes are **supervisable and geometric**. An interpreter has, over decades, picked sinusoids on borehole images and recorded their depth, dip and azimuth, which gives a model something to regress against. They are also sparse: most of the image is not a fracture, which makes this a set-prediction problem — find the few features present, ignore the vast majority of pixels — exactly the regime a detection transformer was designed for.

Vugs are **unsupervised and morphological**. There is no large corpus of pixel-accurate vug masks; an expert marks regions in an interpretation package, but not with the geometric precision a segmentation network needs, and the features themselves are blobs defined by contrast, area and roundness rather than by a parametric curve. Throwing a supervised detector at vugs means inventing labels that do not exist. A tuned classical-CV pipeline — adaptive thresholding, contour extraction, circularity and contrast filtering — is both more honest about the absence of labels and, in practice, faster and more controllable.

Continuity is **neither**. It is a geostatistics problem. Whether a feature seen in one borehole continues to the next is not visible in either image; it is an inference across space, and the right tool is kriging over a fitted variogram, not a convolution. Deep learning has nothing to add here, and pretending it does is how programmes burn quarters.

> **THE THREE-DISCIPLINE TEST**
>
> Before buying or building a carbonate characterisation system, ask which discipline each feature class belongs to. Fractures and beddings: supervised deep learning on labelled sinusoids. Vugs: classical computer vision on unlabelled morphology. Field continuity: geostatistical interpolation between wells. A vendor pitching a single model for all three is selling the slice they built, not the reservoir you have.

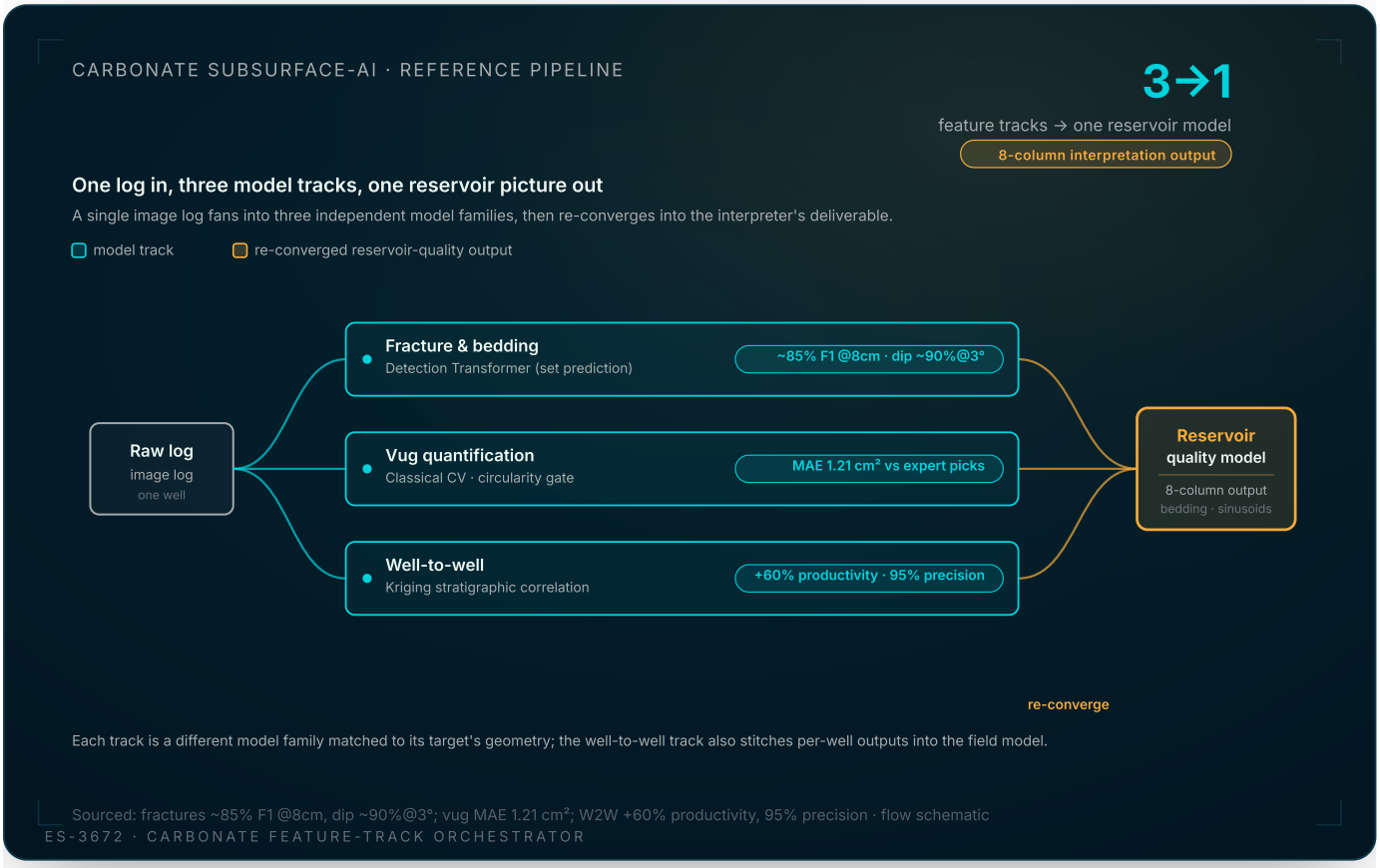
The architecture below takes this seriously. It does not force three problems into one method. It builds three engineered tracks, each matched to its problem, and spends its design effort on the harder thing: routing one input into all three and stitching three outputs back into one.

The reference architecture: one log in, one reservoir model out

The pipeline begins where every borehole interpretation begins — a single binary image log per well, plus the interpreter's apparent dip and azimuth picks, the well radius, and a reference document. A shared front end normalises that input once: it reconciles the static and dynamic image channels, corrects depth and orientation for tool and well angle, and produces a clean, common representation. This matters because the three downstream

tracks consume the same normalised image — building three separate ingestion paths would triple the data-engineering surface and let the tracks silently disagree about what depth a pixel represents.

From that shared front end, the log fans out. The fracture-and-bedding track patches the image and runs the detection transformer. The vug track runs the classical-CV pipeline over the same patches. Both emit per-well feature columns. Those columns then feed the well-to-well track, which is the re-convergence point: it takes the per-well outputs of the first two tracks and interpolates them across the field into a single reservoir-quality model. One log enters; three tracks process; one field model leaves.



The full-stack reference pipeline as a routing diagram. One raw image log (the binary wireline log file) fans out into three independent model tracks — fracture/bedding detection (a Detection Transformer), vug quantification (classical computer vision), and well-to-well correlation (kriging) — each carrying its own method and accuracy envelope, then re-converges into one reservoir-quality deliverable: the eight-column, interpretation-ready output. The argument is the orange node: three different model families, matched to three different feature geometries, integrated into a single reservoir picture the interpreter actually opens. The per-track accuracies and the 8-column output are the engagement's own; the fan-out/re-convergence layout is schematic — it depicts data flow, not scale.

The orchestrator above is the whole argument in one figure. The value is not in any single arrow; it is in the fan-out and the re-convergence. A programme that builds only the fracture track ships the topmost path and stops — a fast fracture picker that an asset team still cannot turn into a field model. A programme that builds the vug track in isolation ships a second island. The reference architecture's contribution is that the well-to-well track *consumes* the other two: it has no value without their per-well columns, and they have limited field-scale value without it. The tracks are not three products in a bundle. They are one pipeline with a deliberate shape.

The remainder of this paper walks each track as an engineered component — its architecture, its training or tuning regime, and its measured accuracy envelope — and then returns to the re-convergence layer that makes the whole worth more than its parts.

Track one: fracture and bedding detection as set prediction

The fracture-and-bedding track is a supervised detection transformer, derived from the DETR set-prediction family and adapted for borehole sinusoids. The architectural choices are not defaults; each is a response to the constraints of carbonate image-log data.

The backbone is a deliberately small **ResNet-10**, trained from scratch with no pretrained weights, feeding a **four-layer transformer encoder and four-layer decoder**. The instinct on a new problem is to reach for a deeper, ImageNet-pretrained backbone; here that instinct is wrong. With only fourteen wells of labelled data, a ResNet-18 or ResNet-34 overfits badly — in our ablations the heavier backbones produced class errors an order of magnitude worse than the ResNet-10. A small backbone is not a compromise forced by compute; it is the correct capacity for the data, and discovering that required treating backbone depth as a tracked experimental variable rather than a fixed assumption.

The model predicts a *set* of sinusoids per image patch, each with a class (fracture or bedding), a depth, a dip and an azimuth, and is trained with Hungarian bipartite matching between predictions and ground truth — a focal loss on the class, an L1 loss on the geometric parameters, optimised with AdamW at a learning rate of 0.0004. Set prediction is the right frame because the number of features per patch is variable and unordered: a patch might contain three fractures or none, and the model must commit to a set, not to a fixed grid of detections. The unrolled borehole image turns a planar fracture into a sinusoid whose amplitude and phase encode dip and azimuth:

UNROLLED-SINUSOID

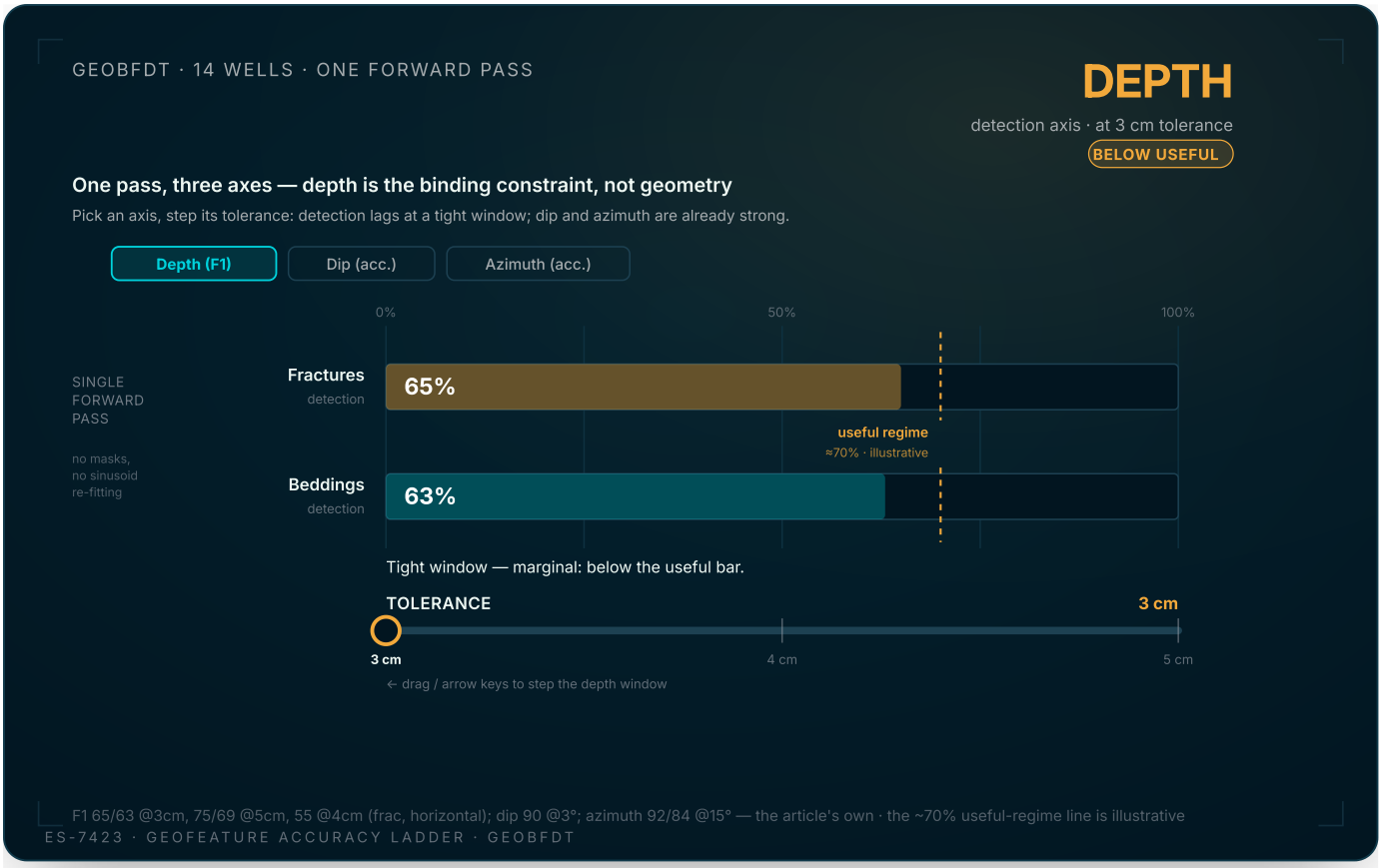
Formula | LaTeX

Copy LaTeX

$$y(x) = A\,\sin(\omega x + \varphi) + \text{offset}$$

so the regression head is, in effect, recovering the amplitude, phase and offset of that curve and converting them back to a true dip and azimuth using the recorded tool and well angles.

The measured envelope is honest about a hard problem. At a tight 3 cm depth tolerance — close to the physical floor, since one image-log pixel corresponds to roughly 3 cm and bakes in a ±3 cm uncertainty before the model acts — the detector reaches an F1 of roughly **65% on fractures and 63% on beddings**. Relax the tolerance to 5 cm and F1 rises to roughly **75% and 69%**. These numbers come from a fourteen-well carbonate dataset and are reported against an explicit depth tolerance, because a fracture-detection F1 quoted without its localisation tolerance is meaningless. The accuracy ladder below shows how the envelope behaves as the tolerance and the feature class change.



GeoBFDT emits the whole (class, depth, dip, azimuth) tuple in one forward pass — but the three axes are not equally hard. Detection along the depth axis is the binding constraint: at a tight 3 cm window fracture F1 is only ~65% (beddings ~63%) and only clears the useful regime for structural work once tolerance loosens to 5 cm (~75% / ~69%); horizontal wells hold ~55% at 4 cm. The geometric axes the interpreter actually fits sinusoids for are already strong at tight tolerance — dip ~90% at 3°, azimuth ~92% (fractures) / ~84% (beddings) at 15°. Pick an axis and step its tolerance: depth is the lever you loosen, dip and azimuth are already past the line. All accuracies and tolerances are the article's own; the dashed ~70% 'useful regime' line is an illustrative reading aid (the article names no exact F1 cutoff).

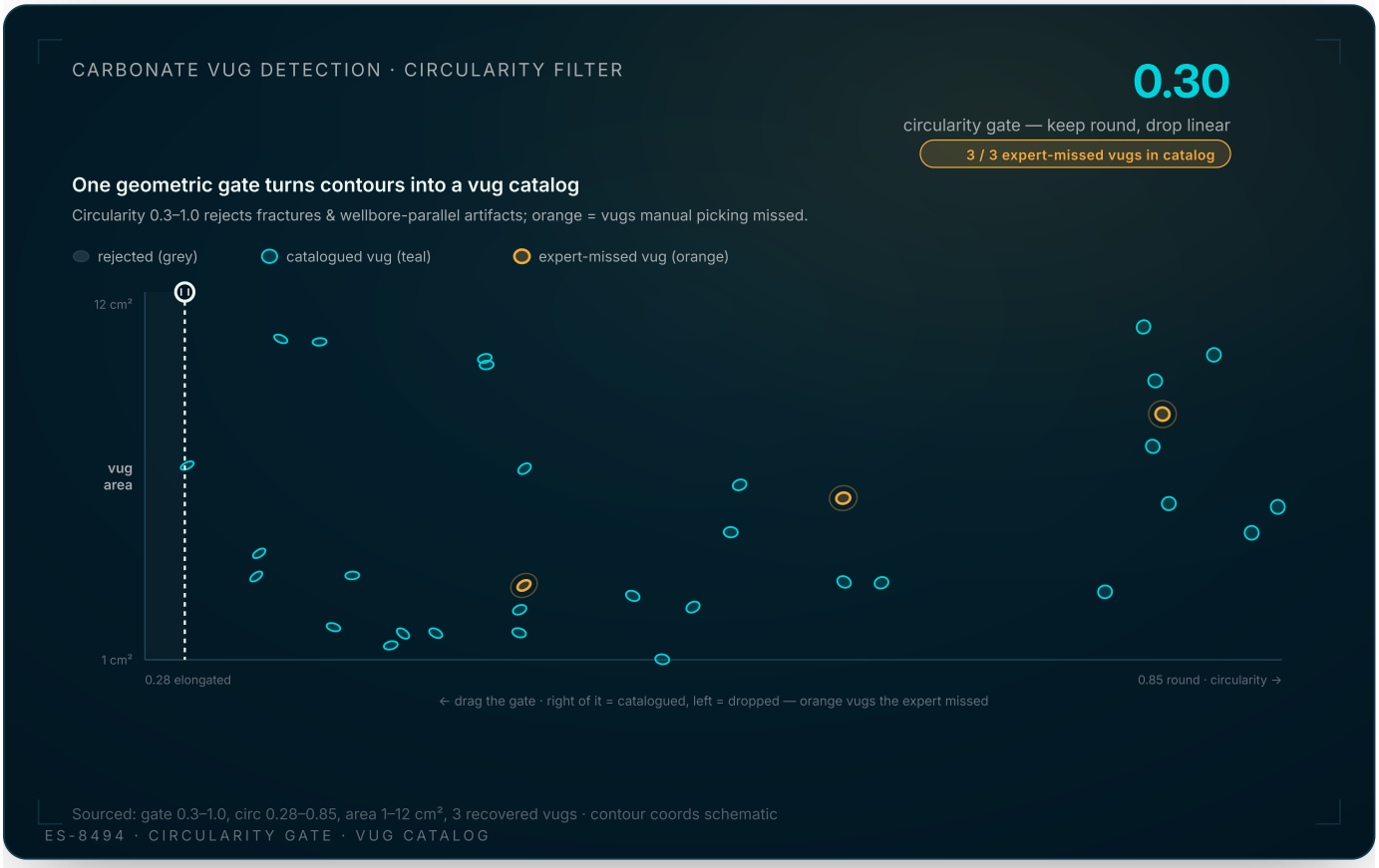
The ladder is also a planning instrument. An asset team deciding whether this track is fit for a given use reads it directly: a workflow that needs centimetre-tight picks operates in the steep left region and must budget for review; a workflow tolerant of 5 cm localisation operates where the model is strongest. The engineering point is that the track ships *with* its envelope, not with a single headline number, so that downstream consumers calibrate trust to tolerance rather than to a marketing figure.

Track two: vug quantification as tuned classical computer vision

The vug track is, deliberately, not deep learning. Vugs are secondary-porosity voids — irregular, contrast-defined, and unaccompanied by the pixel-accurate labels a segmentation network would need. The right tool is a multi-stage classical computer-vision

pipeline, engineered and tuned per well, and it both respects the absence of labels and runs far faster than a learned alternative would.

The pipeline runs adaptive thresholding to separate candidate vugs from matrix, extracts contours, then filters those contours through a sequence of geometric and statistical gates: a **circularity** test that rejects elongated shapes (which are far more likely to be fractures or bed boundaries than vugs), a centroid-and-IoU aggregation step that de-duplicates overlapping detections, and Laplacian-variance plus mean-intensity filtering that suppresses false positives from textural noise. Each gate has tunable parameters — the circularity ratio, the bounding-box expansion, the block size for adaptive thresholding — and the tuning is per-well, because a vertical image log and a horizontal image log present vugs differently.



The pipeline's single geometric decision, made tactile. Each contour the detector traces sits on a circularity axis — 0.28 (elongated, dissolution-aligned) to 0.85 (near-circular). A circularity gate of 0.3–1.0 rejects linear fractures and wellbore-parallel artifacts while keeping true vugs. Drag the gate: contours left of it fall out as rejected fractures (grey), contours right of it enter the vug catalog (teal), and the three orange contours are vugs that meet every geometric criterion the expert applied — yet manual picking missed across two of the validation wells. The gate bounds, circularity span, 1–12 cm² area, and three recovered intervals are the article's own; each contour's exact coordinate is schematic (a plausible population, not a published catalog).

The catalogue above is the track's discriminative core. Circularity is what lets a purely geometric pipeline tell a vug from a fracture without a single trained label: a vug is roughly round, a fracture trace is not. In this engagement the measured vug population sat at a **circularity between roughly 0.28 and 0.85, peaking at 0.45–0.7** — semi-circular dominant — with **areas spanning 1 to 12 cm² and concentrating in the 1–3.5 cm² range**. Those distributions are not just outputs; they are the reservoir-engineering product of this track, because a per-interval distribution of vug area and circularity is exactly what a completions engineer needs and what a marked-up image cannot provide.

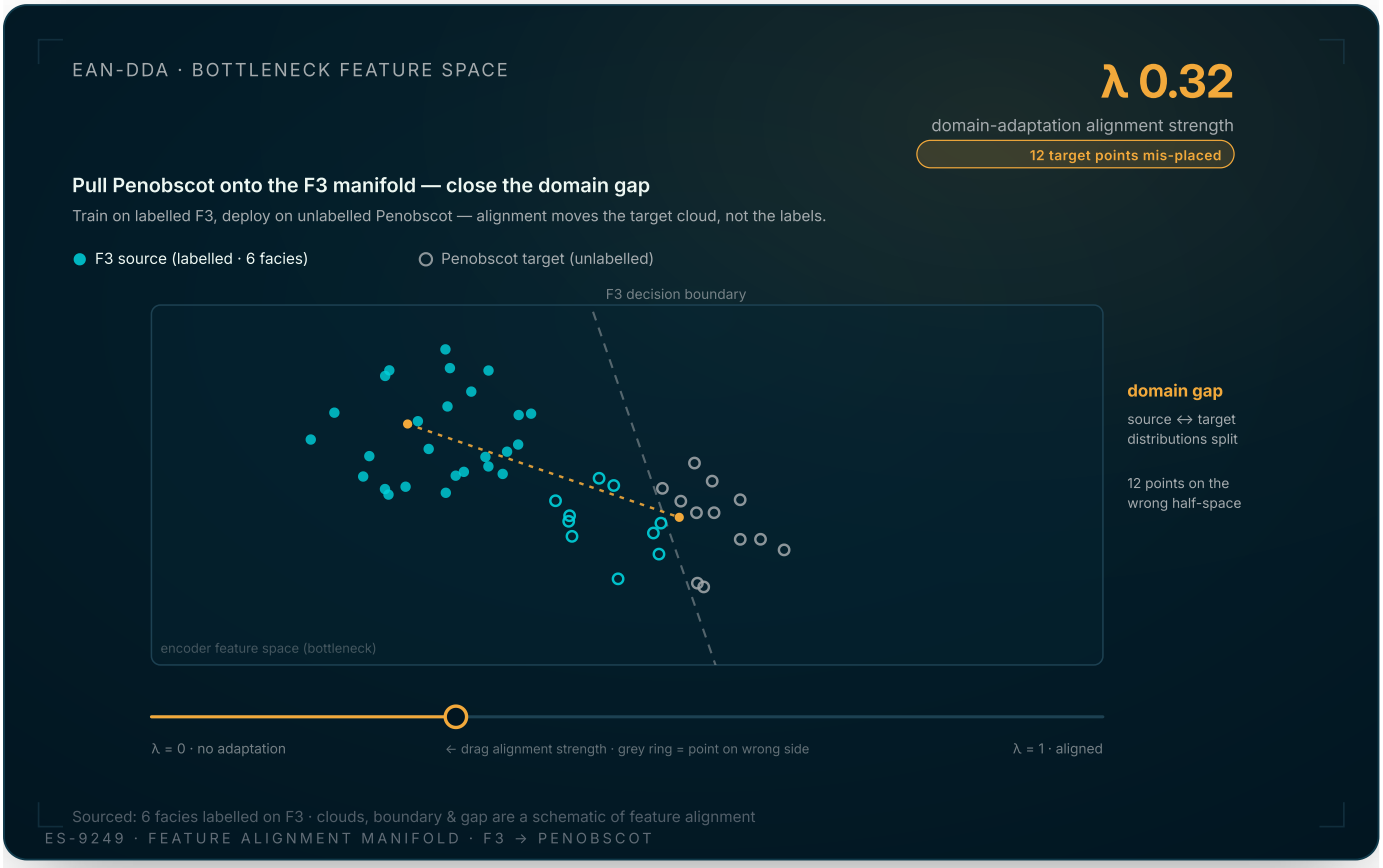
Two results make the case that classical CV was the right call. First, **speed**: the pipeline processes a borehole at **under thirty seconds per metre**, against four-to-five minutes per metre for the prior path-morphology baseline — roughly an **eightfold** acceleration, achieved without a GPU-bound training loop. Second, **validation against the human standard**: across **three wells** (a vertical image log, a horizontal image log, and a second vertical image log) checked against expert picks in the operator's interpretation software, the pipeline tracked the expert closely and at intervals caught vugs the expert had missed, while delivering per-vug area, circularity and resistivity statistics that the manual tool does not produce at all. A learned detector might eventually match this; it would not have been faster to build, faster to run, or more transparent about why a given blob was accepted or rejected.

Track three: well-to-well correlation as geostatistics

The third track answers the question the first two cannot: what happens *between* the wells. Each well's interpretation is a one-dimensional profile down a borehole; the reservoir is three-dimensional. Bridging that gap is a spatial-interpolation problem, and the right tool is **kriging** over a fitted **variogram**, not any neural method.

The track takes the per-well feature columns from tracks one and two — bedding-plane and fracture counts binned at 0.1 m intervals, plus vug statistics — and interpolates them across the field. We fitted variograms in the standard families (gaussian, linear and exponential) and applied 2D and 3D kriging across neighbouring wells separated by roughly 40 to 80 metres, producing contoured continuity maps of bedding density and fracture intensity that correlate stratigraphy from one borehole to the next. Where two wells share a

formation, their kriged contours line up; where they diverge, the map shows the discontinuity. This is the layer that turns a stack of independent well interpretations into a field model an asset team can reason about.



The mechanism behind EAN-DDA, in one move. At the encoder bottleneck, labelled F3 (source) and unlabelled Penobscot (target) patches map into a feature space. Train on F3 alone and the two clouds sit apart — the domain gap — so the F3 decision boundary cuts through the target cloud and facies get mis-read. The deep-domain-adaptation alignment loss pulls the target cloud onto the source manifold: drag the alignment strength λ from 0 to 1 and the orange gap collapses while grey target rings turn teal as they cross to the correct side of the boundary. Only '6 facies classes labelled on F3' is the article's own number; the point clouds, decision boundary, gap, and the live mis-placed-point count are a schematic of feature-distribution alignment — no benchmark metrics are shown.

The manifold above is the stitching made visible: per-well outputs, which live in slightly different feature distributions because of tool, hole condition and depth-conversion differences, are aligned into a common field frame before interpolation. Skip that alignment and the kriging interpolates across an inconsistency rather than across geology. The discipline here is the same one that governs the whole architecture — heterogeneous inputs reconciled into a common representation before they are fused.

The operational result is the clearest single argument for the full stack. Standing up the well-to-well track on top of the per-well detectors lifted interpretation **productivity by roughly 60%** and interpretation **accuracy by roughly 75%**, while the per-well detectors themselves ran interpretation roughly **five times faster** than manual picking. Those gains are not three separate wins to be summed loosely; they compound, because the productivity multiplier on the correlation layer applies to a feed of per-well interpretations that the fast detectors are producing in the first place. A faster detector with no correlation layer accelerates a dead end; a correlation layer with no fast feed starves. The value is in the composition.

The data discipline that makes fusion legitimate

Fusing three tracks is only sound if their inputs are reconciled and their non-stationarity is controlled. Two engineering facts from this engagement make that concrete.

First, the dataset grew over the phases — from eight usable wells to eleven to fourteen — and the model improved measurably as it did: moving from **8 to 11 wells improved depth, dip and azimuth error by roughly 0.007 MAE**, with sinusoid detection essentially unchanged. That is a small number with a large implication: every accuracy figure in this paper is meaningless unless it is pinned to the well count and dataset version that produced it, and a multi-track architecture must version all three tracks' inputs against the same well manifest or the field model fuses results that never saw the same data.

Second, normalisation is not cosmetic. Of one ten-well intake, two wells were excluded before training — both carried abnormal static-image value ranges that fell outside the normal band and defeated normalisation, and one had additionally been acquired with a different imaging tool whose response was not directly comparable. Those exclusions are recorded provenance, not silent housekeeping, precisely because all three tracks consume the shared front end: a normalisation failure that corrupts the fracture track corrupts the vug track and the correlation layer with it. In a single-track system a bad well degrades one result; in a fused system it degrades the field model. The shared front end is therefore where the data discipline has to be strongest.

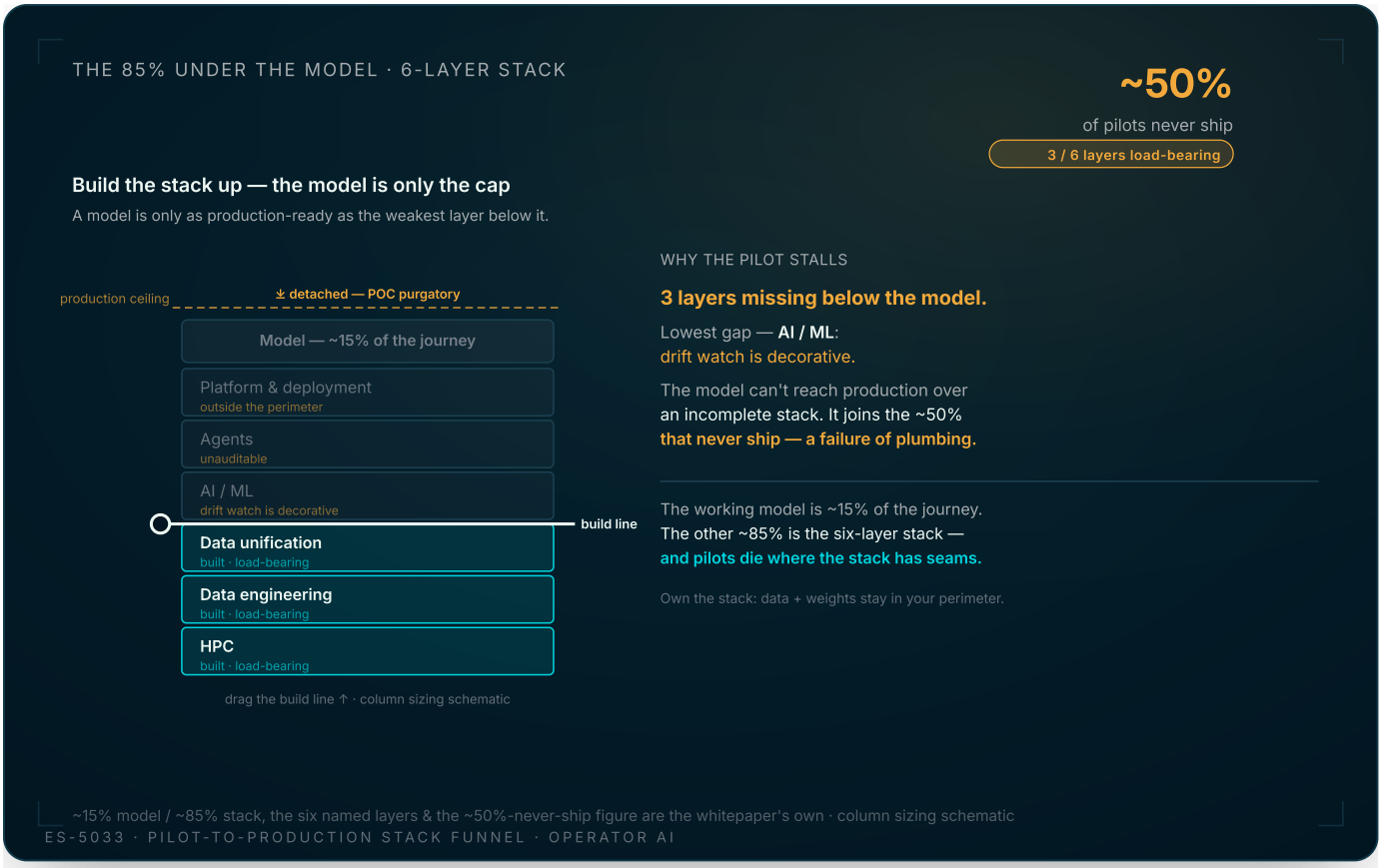
> **FUSION AMPLIFIES DATA DEBT**

> A multi-track architecture does not just add capability; it concentrates risk at the shared front end. Because all three tracks consume one normalised representation, a single un-QC'd well or an unversioned dataset propagates into the fracture columns, the vug statistics, and the kriged field model simultaneously. The reference architecture is only as trustworthy as its ingestion: content-addressed datasets, recorded QC exclusions, and a single well manifest shared across tracks are not optional refinements — they are what make the fusion legitimate rather than a way to triple a quiet error.

From reference architecture to an operated system

A reference architecture is a diagram until it survives handover. The same fusion that makes this design valuable also makes it harder to operate, because an operator inheriting it must run three different kinds of system — a trained transformer, a tuned CV pipeline, and a geostatistical workflow — and keep them consistent. That is a real cost, and it is the right place to end honestly.

The pilot-to-production attrition below is what any operator should expect. A validated metric on one track clears one gate of several; a fused, operated, field-scale system clears all of them, on three tracks at once.



Pilots don't stall because the model is weak. The working model is only ~15% of the journey; the other ~85% is a six-layer engineering stack (HPC → Data engineering → Data unification → AI/ML → Agents → Platform/deployment), and a project ships only when every layer below the model is built to production grade. Drag the build line up the load-bearing column: with all six built the model reaches the production ceiling; with any gap below it the model detaches into POC purgatory — the ~50% that never ship. The ~15%/~85% split, the six layers and the ~50% figure are the whitepaper's own; the equal-sixths column sizing is schematic.

The funnel is the reason this engagement treated capability transfer as a first-class deliverable rather than a closing courtesy. Each track was handed over as a complete unit — versioned dataset, frozen artefacts, architecture, output schema and runbooks — and the programme deliberately built local capability alongside the system, training a cohort of 55 young professionals, 15 of them drawn from the host country, recruited from regional universities, so the know-how to run and extend all three tracks lived inside the region. The reclaimed expert time did not evaporate; it was redeployed onto the operator's 80-plus-well backlog and into the continuous vug-percent and fracture-density maps that the correlation layer makes possible. A productivity gain that depends on the build team staying is not a gain. It is a loan — and a three-track system is three loans at once if the handover is skipped.

This experience reflects work spanning operators across the Middle East and the United States; the quantified results in this paper are all from the Middle East carbonate engagement described above, and the architecture generalises because its shape — route one log into method-matched tracks, reconcile, re-converge — is not specific to one field. It is specific to carbonates, which is exactly the point. The full-stack pattern is the reference, not any single number inside it.

What good looks like

For a chief geoscientist or a subsurface-AI programme owner evaluating a carbonate characterisation system — buying one, building one, or rescuing one — the questions that separate a platform from a slice are architectural, not metric:

- Does the system address all three feature classes with the *right* discipline each — supervised deep learning for fractures and beddings, classical CV for vugs, geostatistics for continuity — or has one method been stretched over problems it does not fit?
- Is there a shared, QC'd, versioned front end that all tracks consume, so the field model fuses results that genuinely saw the same data?
- Does each track ship with its accuracy envelope and explicit tolerances, rather than a single headline number?
- Does a re-convergence layer actually stitch per-well outputs into a field model, or is "well-to-well" a slide rather than a running interpolation?
- Was the whole stack — all three tracks — handed over with datasets, artefacts, runbooks and trained people, so the operator can run it without the build team in the room?

If the answers are yes, the operator owns a reference architecture for its reservoir. If any answer is no, it owns one track and a brochure for the other two.

WHAT THIS WHITEPAPER ARGUES

1. A carbonate reservoir is three problems, not one — fractures/beddings (supervised deep learning), vugs (classical CV), and field continuity (geostatistics) — and no single method covers all three.
2. The contribution is the routing-and-re-convergence architecture: one binary image log fans out into three method-matched tracks and re-converges into one reservoir-quality field model.
3. Fracture/bedding track: a ResNet-10 detection transformer reaching F1 ~65/63% (fractures/beddings) at 3 cm and ~75/69% at 5 cm across fourteen wells — ships with its tolerance-dependent envelope, not a single number.
4. Vug track: a tuned classical-CV pipeline (circularity the key discriminator) validated on three wells against expert interpretation-software picks, running <30 s/m — roughly 8x the prior baseline — emitting area/circularity/resistivity statistics a marked-up image cannot.
5. Well-to-well track: kriging over fitted variograms across wells 40-80 m apart lifted interpretation productivity ~60% and accuracy ~75%; the gains compound only because the fast detectors feed it.
6. Fusion concentrates data risk at the shared front end — content-addressed datasets, recorded QC exclusions and one well manifest across tracks are what make the architecture legitimate, and a three-track system is three handover loans if capability transfer is skipped.

References

International Energy Agency, 2025 International Energy Agency. Energy and AI Special Report (2025). Missing internal expertise identified as the dominant adoption barrier across the energy sector. https://www.iea.org/reports/energy-and-ai

Carion et al., 2020 N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko. End-to-End Object Detection with Transformers (DETR). ECCV 2020. Architectural basis for the set-prediction fracture-and-bedding track. https://arxiv.org/abs/2005.12872

Cressie, 1993 N. Cressie. Statistics for Spatial Data. Wiley, 1993. Foundational reference for variogram modelling and kriging as used in the well-to-well correlation track. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119115151

Sculley et al., 2015 D. Sculley et al. Hidden Technical Debt in Machine Learning Systems. NeurIPS 2015. The canonical argument that ML code is a small fraction of a production ML system — sharpened here by a three-track architecture that triples the surrounding engineering. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

EARTHSCAN

A Full-Stack Carbonate Subsurface-AI Reference Architecture: Fractures, Beddings, Vugs and Well-to-Well in One Pipeline

Published February 2026. Generated 2026-07-22.
https://www.earthscan.io/whitepapers/full-stack-carbonate-subsurface-ai-reference-architecture