Sanjeevani

Neuro-Symbolic AI for Automated CKD Triage & EMR Drafting via 1D NMR

Abhiram Radha Krishna | Kaggle MedGemma Impact Challenge

1. Problem Statement

The Silent Epidemic & The Spectroscopic Bottleneck

Chronic Kidney Disease (CKD) affects **10% of the global population**, yet early stages are entirely asymptomatic. Urinary 1D ^{1}H NMR spectroscopy can detect early metabolic shifts — depleted citrate, elevated uremic toxins (TMAO), increased lactate — before any eGFR decline is measurable. However, interpreting NMR spectra requires highly trained spectroscopists, creating a critical access bottleneck for under-resourced and rural clinics.

The Unmet Need: Why Pure LLMs Cannot Solve This

Large Language Models are text-prediction engines. They are mathematically incapable of performing spectral deconvolution on arrays of raw float values, and when prompted with raw NMR data they produce dangerous clinical hallucinations. A safe, deployable CKD screening tool therefore requires a hybrid approach: deterministic signal processing to extract verified biochemical ratios, with LLMs used strictly for communication tasks they are designed for.

Clinical Impact

- Democratises expert-level NMR analysis for rural and under-resourced hospitals
- Eliminates AI hallucination through deterministic mathematical gating before any LLM is invoked
- Reduces physician administrative burden by auto-drafting both the clinical Assessment & Plan and the patient-facing EMR message

2. Solution Architecture

Sanjeevani deploys three AI agents sequentially, each operating strictly within its domain of competence. No agent is asked to perform a task it cannot do reliably.

Stage 1 — Vision Gatekeeper (PaliGemma)

Running expensive deep learning inference on degraded samples wastes compute and risks false positives. PaliGemma (paligemma-3b-mix-224) receives a rendered image of the 2.5–4.5 ppm Anchor Region with a colour-coded ROI overlay and classifies it as 'sharp peaks' or 'flat noise'. If noise is detected, the pipeline halts immediately and alerts the technician. A secondary mathematical gate checks that the creatinine anchor amplitude exceeds a biochemically-justified threshold, providing a dual-layer QC that neither agent can bypass alone.

Stage 2 — Neuro-Symbolic Physics Engine (Custom PyTorch)

A custom Multi-Task CNN-Transformer (SanjeevaniEngine) processes the 1000-point spectrum through a shared InceptionBlock encoder and 5 independent decoder branches — one per biomarker. Independent decoders prevent channel collapse, a critical failure mode where dominant biomarkers suppress weaker ones in a shared decoder. The engine outputs both a classification mask and a regression amplitude for each of five CKD-relevant biomarkers:

- Creatinine (3.05, 4.05 ppm) — normalisation anchor
- Citrate (2.55, 2.68 ppm) — depleted in renal tubular dysfunction
- Lactate (1.33 ppm) — elevated in renal hypoxia
- TMAO (3.26 ppm) — uremic toxin elevated in CKD
- Taurine (3.30, 3.42 ppm) — depleted in late-stage CKD

The verified ratios feed a transparent neuro-symbolic risk score:

```
Risk Score = (Citrate Deficit × 1.5) + (TMAO Ratio × 1.0) +
(Lactate Ratio × 0.5)  |  Threshold: 1.2
```

This formula is fully auditable — every coefficient has a clinical justification and can be reviewed or overridden by a physician.

Stage 3 — Clinical Co-Pilot (MedGemma)

MedGemma (medgemma-1.5-4b-it) receives only the verified mathematical ratios, never raw spectral data. It operates under two distinct prompt personas:

- Physician View: Expert nephrologist synthesising biomarker ratios with patient history into a formal Assessment & Plan, including pathophysiological reasoning and next-step recommendations.
- Patient Portal View: Strict negative constraints forbid medical jargon and chemical names. MedGemma translates the clinical note into an 8th-grade reading level message ready for 1-click export to hospital EMR systems (e.g., MyChart).

3. Technical Details & Engineering Journey

Model Architecture

- Encoder: 2× InceptionBlock (kernels 1, 3, 5, 7) + MaxPool1D → 64-channel representation at 250-point resolution
- Sequence model: 2-layer TransformerEncoder (d_model=64, nhead=4, dim_feedforward=256)
- Decoder: 5 independent branches (ConvTranspose1d 64→32 + BatchNorm + LeakyReLU → classification head + regression head)

- Training: 10,000 synthetic spectra, 25 epochs, Adam lr=0.0005, BCEWithLogitsLoss + MSELoss with per-channel weights [1.0, 2.0, 3.0, 5.0, 2.0]
- Infrastructure: Hugging Face Transformers, native float16, dual T4 GPUs, Gradio tabbed clinical dashboard

Synthetic Data Generator (CKDMetabolomeGenerator)

All training data is generated from first principles using pseudo-Voigt profiles (60% Lorentzian / 40% Gaussian mixture), which correctly model NMR lineshapes including Lorentzian tails. Each training spectrum includes:

- Biologically-overlapping class distributions (healthy/sick share a 'grey zone' rather than clean separation)
- Phase distortion via Hilbert transform with random phase angle
- 5–15 random decoy peaks outside per-biomarker exclusion zones, amplitude ≤ 0.15
- 50% probability of clean flat baseline vs. sinusoidal distortion
- High-frequency machine static (σ = 0.02) simulating Bruker electronic noise
- 10% catastrophic QC failure injection (pure Gaussian noise)

Performance Progression

The table below documents the iterative development. The accuracy decrease from V1 to Final is intentional — it reflects replacing an inflated closed-world metric with a more honest model that does not collapse weaker biomarker channels.

Configuration	**Accuracy**	**Sensitivity**	**Specificity**
Baseline (shared decoder)	93.0%	96.6%	87.8%
V2: Independent decoders + decoys	91.0%	96.1%	85.7%
Final: + channel-weighted loss	89.5%	100.0%	76.9%

4. Engineering Challenges Solved

The following table documents the four major failure modes encountered and resolved during development. Each represents a genuine technical contribution, not a trivial bug fix.

Issue	**Root cause**	**Fix applied**
TMAO/Lactate channel collapse	Shared decoder prioritised dominant channels; weak signals yielded dead ReLU	5 independent decoder branches with per-channel BatchNorm and weighted loss (TMAO 5×, Lactate 3×)

Zero specificity after decoy injection	Water peak (amp 5–15) set noise floor at 0.3, erasing citrate/lactate signals before the model	Removed water peak (simulating clinical suppression); added dynamic noise floor (2% of global max)
Transfer test false activations	Decoy peaks placed inside biomarker exclusion zones via broad (2.9–4.2 ppm) window	Narrowed exclusion zones to per-biomarker windows; reduced decoy amplitude ceiling to 0.15
Real-data baseline elevation	MTBLS1 spectra exhibit broad macromolecular baseline floor absent from synthetic training data	Identified as V2 priority: Spinach J-coupling simulator + baseline augmentation

5. Real-World Transfer Evaluation (MTBLS1)

To assess clinical readiness beyond synthetic validation, we performed a zero-shot transfer evaluation using real experimental urine NMR data from the publicly available MTBLS1 dataset (European Bioinformatics Institute, MetaboLights). A single patient FID was downloaded, processed with nmrglue (digital filter removal, FFT, ACME autophasing), interpolated to 1000 points, and apodized (σ = 1.5) before inference.

Domain Gap Analysis

Comparison of a synthetic training spectrum with the MTBLS1 experimental spectrum revealed two distinct domain gaps:

Gap 1 — Amplitude Distribution Mismatch

In the synthetic generator, creatinine is always the dominant peak (amplitude 0.8–1.2). In real urine, creatinine competes with glucose, hippurate, citrate and other metabolites and is frequently not the dominant feature. After dynamic scaling, the creatinine signal at 3.05 ppm was suppressed below the model's learned detection threshold. The model outputs zero for the creatinine channel on real data.

Gap 2 — Elevated Baseline Floor

The MTBLS1 spectrum exhibits a broad macromolecular background sitting at approximately 0.25–0.35 normalised intensity across the entire 0.5–4.5 ppm region. The synthetic generator produces near-zero baseline in 50% of samples. This floor compresses all real peak amplitudes relative to the training distribution and is the primary cause of failed biomarker detection on real data.

Note on J-Coupling

Contrary to an earlier assessment, creatinine's two resonances at 3.05 and 4.05 ppm are both singlets — the CH_3 and CH_2 groups are separated by a nitrogen atom that breaks the J-coupling pathway. The lineshape mismatch is therefore attributable to the baseline floor and amplitude distribution issues above, not to multiplet structure.

Proposed V2 Fixes

- Baseline augmentation: Add a random broad Lorentzian background (amplitude 0.1–0.4, width 0.5–2.0 ppm) in 50% of training samples to simulate macromolecular baseline
- Amplitude distribution calibration: Study MTBLS1 relative peak heights and update creatinine amplitude range from [0.8–1.2] to the real urinary distribution [0.2–1.5]
- Additional metabolite simulation: Add glucose (3.4–3.9 ppm multiplets) and hippurate (7.8 ppm, outside current mask) as decoy peaks to improve model robustness
- Spinach integration: Replace pseudo-Voigt peaks with quantum spin Hamiltonian simulated multiplets for true J-coupled training spectra

6. Honest Limitations

Synthetic Validation Only

All quantitative metrics (91% accuracy, 100% sensitivity) are derived from synthetic test data generated by the same family of distributions used for training. These numbers demonstrate that the architecture works correctly, but they do not represent clinical performance. Real-world accuracy is unknown until the domain gap is resolved.

TMAO Channel Intermittency

TMAO (3.26 ppm) sits 0.04 ppm from taurine (3.30 ppm). Despite independent decoders and 5× channel weighting, TMAO detection is intermittent on borderline cases where taurine amplitude is similar. The current risk score partially compensates through citrate deficit weighting, but TMAO reliability requires further architectural work.

MedGemma Consistency on Zero-Valued Channels

When TMAO or lactate channels output near-zero (non-detection rather than true low values), MedGemma interprets the zero as a normal measured value rather than a missing measurement. This produces statements like 'TMAO is normal' when the correct statement is 'TMAO was not detected.' Channel reliability flags should be added to the MedGemma prompt in V2.

7. Conclusion

Sanjeevani demonstrates that a hybrid neuro-symbolic pipeline — where LLMs are restricted to communication tasks and deterministic physics handles signal extraction — is a viable architecture for clinical NMR screening. The project produced four concrete technical contributions: (1) an independent-decoder multi-task architecture that prevents channel collapse in dense biomarker clusters, (2) a per-channel weighted loss scheme that forces the model to learn weak signals it would otherwise abandon, (3) a dual-layer QC gate combining VLM visual classification with biochemical anchor detection, and (4) a rigorous domain gap analysis that quantifies exactly why zero-shot real-data transfer fails and prescribes specific V2 fixes.

The 93% synthetic accuracy of the naive baseline dropped to 89.5% in the final model — not because the system regressed, but because we identified and removed artificial inflation caused by closed-world evaluation, shared decoder dominance, and training distribution mismatch. A model that fails honestly and explicably is more valuable for clinical AI development than one that succeeds silently on its own test set.

Reproducibility

All code, model weights, and the Gradio dashboard are available in the submitted Kaggle notebook. The MTBLS1 sample used for transfer evaluation is publicly available at https://www.ebi.ac.uk/metabolights/MTBLS1.