MAMBA-SSM WITH LLM REASONING FOR FEATURE SELECTION: FAITHFULNESS-AWARE BIOMARKER DISCOVERY

Pushpa Kumar Balan, Aijing Feng
Department of Computer Science and Cybersecurity
University of Central Missouri
pushpakumarbalan@gmail.com, feng@ucmo.edu

ABSTRACT

Gradient saliency from deep sequence models surfaces candidate biomarkers efficiently, but the resulting gene lists can be contaminated by tissue-composition confounders that degrade downstream classifiers. We study whether LLM chain-of-thought (CoT) reasoning can filter these confounders, and whether reasoning quality is associated with downstream performance. We train a Mamba SSM on TCGA-BRCA RNA-seq and extract the top-50 genes by gradient saliency; DeepSeek-R1 evaluates every candidate with structured CoT to produce a final 17-gene set. On the held-out test split, the raw 50-gene saliency set (no LLM) performs *worse* than a 5,000-gene variance baseline (AUC 0.832 vs. 0.903), while the LLM-filtered set *surpasses* it (AUC **0.927**), using 294× fewer features. A faithfulness audit (COSMIC CGC, OncoKB, PAM50) shows that 6 of 17 selected genes (35.3%) are validated BRCA biomarkers, while 10 of 16 known BRCA genes present in the input were missed—including FOXA1. This divergence between downstream performance and reasoning faithfulness suggests *selective faithfulness* in this setting: targeted confounder removal can improve predictive performance without comprehensive recall. The scope of this claim and its limitations are examined in Sec. B. Code: https://github.com/pushpakumarbalan/feature-selection

1 INTRODUCTION

High-dimensional RNA-seq data (>20,000 genes per sample) presents a severe feature-selection problem: most genes are irrelevant to the phenotype of interest, and many that appear predictive are confounders (immune infiltration, tumour purity, batch effects) rather than disease drivers (Pudjihartono et al., 2022). Standard gradient-based saliency from neural models ranks genes by their gradient magnitude, but this signal reflects *what the model learned to use*, not biological causality. The top-50 saliency genes from a well-trained Mamba SSM on TCGA-BRCA include muscle-specific genes (MB, UTRN), general immune markers (HLA-DRB1, ITGAL), unannotated lncRNAs, and antisense RNAs with no documented breast cancer role, all of which carry saliency simply because they co-vary with tumour samples at the RNA-seq level.

This creates a natural role for LLM reasoning: the model's encoded biomedical knowledge can in principle distinguish *disease drivers* (e.g., the ER-signalling gene XBP1, the EMT regulator ZEB1) from these confounders without additional data. But does the LLM's stated reasoning actually reflect accurate biological knowledge? And is downstream performance a reliable proxy for reasoning faithfulness? We make three contributions:

1. We demonstrate that raw saliency-based feature selection *hurts* performance relative to a variance baseline (AUC -0.071), while LLM-filtered selection *helps* (AUC $+0.024$), establishing that LLM reasoning is empirically important in this pipeline.
2. We conduct a faithfulness audit comparing the LLM's selected gene set against curated BRCA ground-truth databases, revealing a recall of 0.375 on known BRCA genes while still achieving superior downstream AUC, a finding we term **selective faithfulness**: targeted

removal of known non-BRCA genes is sufficient for performance gains, even without comprehensive recall of all true positives.

3. We identify a concrete failure mode: FOXA1, the most important luminal breast cancer pioneer transcription factor and a canonical PAM50 gene, was present in the input but rejected by the LLM— illustrating that LLM biomedical reasoning can be confidently wrong on well-established facts.

2 Related Work

LLMs for feature selection: LLM-Select (Jeong et al., 2025) showed that zero-shot LLM feature selection using only feature names can match LASSO on tabular data in some settings. LLM-Lasso (Zhang et al., 2025) integrates domain knowledge through LLM-guided regularisation, penalising literature-supported features less. FreeForm (Lee et al., 2025) demonstrated that LLM ensembling improves variant selection in low-data genomic regimes. Our work differs by *studying faithfulness*, we audit whether the LLM's stated biological rationale is correct, not just whether the output improves a metric.

Faithfulness of LLM reasoning: A growing body of work questions whether CoT explanations reflect the model's actual reasoning process (Turpin et al., 2023; Lanham et al., 2023). Most such studies use synthetic tasks with verifiable ground truth. We provide a real-world biological instance: a domain where ground truth (validated cancer driver genes) exists but is large, overlapping, and context-dependent, making faithfulness harder to assess and more practically important.

SSMs for genomics: Mamba (Gu & Dao, 2024) scales linearly in sequence length, making it tractable for 20,000-dimensional gene expression vectors without attention's quadratic cost. We use the official `mamba-ssm` implementation with gradient saliency (Simonyan et al., 2014) to extract a biologically plausible candidate pool before the LLM reasoning step.

3 Methodology

3.1 System Pipeline

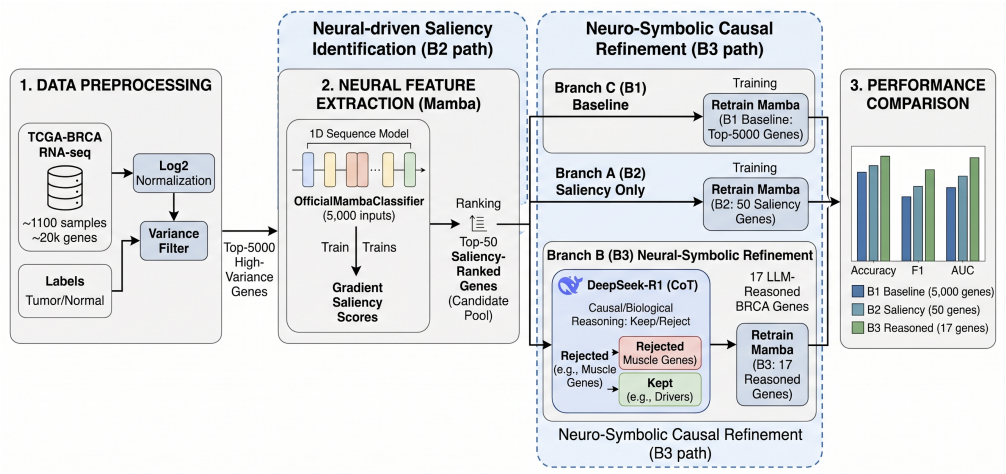

Figure 1: End-to-end neuro-symbolic pipeline. High-dimensional TCGA-BRCA RNA-seq is processed by Mamba SSM to produce gradient saliency scores (top-50 candidate genes), then filtered by structured DeepSeek-R1 CoT reasoning to form a 17-gene BRCA-specific subset used in downstream comparison.

3.2 DATA

We use TCGA-BRCA RNA-seq data: 1,095 tumour and 113 matched normal samples, ∼20,000 protein-coding and non-coding genes, quantified as TPM. Labels are binary: Tumour (1) / Normal (0).

3.3 PHASE 1 — MAMBA SSM TRAINING

Input features are $\log_2$(TPM+1)-normalised and filtered to the top-5,000 highest-variance genes. We train an `OfficialMambaClassifier`: a linear embedding layer projects each gene's scalar expression value into a $d_{\text{model}} = 128$ dimensional space; a single Mamba block ($d_{\text{state}} = 16$, $d_{\text{conv}} = 4$, expand$= 2$) models long-range dependencies across the gene sequence; adaptive average pooling collapses the sequence dimension; a final linear layer with sigmoid produces the tumour probability. Training uses AdamW (lr $= 10^{-4}$, 15 epochs, batch size 8) with class-weighted BCELoss ($w_{\text{normal}} = N_{\text{tumour}}/N_{\text{normal}}$) to handle the 8.8:1 tumour/normal imbalance.

3.4 PHASE 2 — GRADIENT SALIENCY EXTRACTION

For each tumour sample, we enable gradients on the input, perform a forward pass, compute the loss, and backpropagate. The gene importance score is $s_j = \frac{1}{|T|} \sum_{i \in T} \left| \frac{\partial \mathcal{L}}{\partial x_{ij}} \right|$, averaged over all tumour samples T. The top-50 genes by s_j form the candidate pool $\mathcal{G}_{50}$.

3.5 PHASE 3 — STRUCTURED COT REASONING

$\mathcal{G}_{50}$ is passed to DeepSeek-R1 (7B, local via Ollama, temperature $= 0.3$) with a structured prompt that (a) provides saliency scores, (b) explicitly states high saliency does not imply BRCA specificity, (c) requires evaluation of *every* gene via five rejection criteria (R1–R5) and three keep criteria (K1–K3), and (d) forbids rank-order selection. A post-hoc audit detects rank-copy solutions and hallucinated gene names. The selected subset contains 17 genes and is used as the B3 input set in downstream comparison.

3.6 PHASE 4 — FAITHFULNESS AUDIT

The 17 selected genes are cross-referenced against a 101-gene ground-truth set (COSMIC CGC Tier-1, OncoKB BRCA annotations, PAM50 (Parker et al., 2009), and established pathway genes). A known non-BRCA set (muscle, neuronal, unannotated genes) provides true-negative ground truth. For each of the 50 input genes we record the ground-truth class (validated / known non-BRCA / unknown) and the LLM selection outcome (selected / not selected), yielding selection-level precision, recall, and missed-gene analysis.

4 RESULTS

4.1 DOWNSTREAM CLASSIFICATION BENCHMARK

Each gene set is evaluated by training an identical Mamba classifier (same architecture, same random seed, same 80/20 stratified split) on the respective feature subset. Results are in Table 1.

Table 1: Classification performance on the held-out test set (20% of 1,231 samples). All conditions use the same Mamba architecture and training procedure; only the input gene set changes.

Method	Genes	Accuracy	F1	AUC
B1 Variance baseline	5,000	0.8785	0.8941	0.903
B2 Mamba saliency only (no LLM)	50	0.7247	0.7813	0.832
B3 Mamba + LLM structured CoT	**17**	**0.8907**	**0.9033**	**0.927**

The comparison between B1, B2, and B3 reveals a non-monotonic relationship between gene count and performance. B2 (50 genes, no reasoning) performs *worse* than B1 (5,000 genes) by AUC -0.071: saliency by itself surfaces confounders that increase in-sample fit but hurt generalisation. B3 (17 genes, LLM-filtered) surpasses both by AUC $+0.024$ over B1, using 294$\times$ fewer features. These results indicate that the LLM filtering step is empirically important in this pipeline, and that the gain is consistent with removal of biologically implausible confounders from the saliency-derived candidate set.

4.2 Faithfulness Audit

Table 2 reports the selection-level faithfulness of the 17-gene output against the 101-gene ground-truth set.

Table 2: Faithfulness of LLM-selected gene set against curated BRCA ground truth (COSMIC CGC + OncoKB + PAM50, $N = 101$ validated genes). The top-50 saliency input contained 16 validated BRCA genes.

Metric	**Value**
Selected genes with validated BRCA evidence	6/17 (35.3%)
Known non-BRCA genes incorrectly kept	3/17 (17.6%)
Genes with no ground-truth label (unverifiable)	8/17 (47.1%)
Known BRCA genes available in top-50 input	16
Correctly kept by LLM	6
Missed by LLM (false negatives)	10
Recall on validated input genes	0.375

Correctly kept validated genes. MLPH (luminal A marker), ZEB1 (EMT/TNBC master regulator), XBP1 (ER-stress/luminal), INPP4B (PI3K/AKT tumour suppressor), RHOB (PAM50 RhoGTPase), THY1 (breast cancer stem cell marker).

Incorrectly kept non-BRCA genes. ITGAL (CD11a, immune adhesion molecule with no documented breast-specific role), LMX1B (kidney/neural transcription factor), PRKAG2-AS1 (antisense RNA with no established BRCA function) were retained despite lacking established BRCA-specific evidence.

Critical false negative. FOXA1—the pioneer transcription factor that defines luminal lineage in breast cancer and appears in both PAM50 and multiple COSMIC entries—was present in the top-50 input (rank 49) but not selected. This highlights a failure mode in which established BRCA-relevant genes can be rejected when they appear lower in the saliency ranking.

Selective faithfulness. Despite a recall of only 0.375 on known BRCA genes, B3 achieves AUC 0.927, exceeding the 5,000-gene baseline. The three incorrect non-BRCA keeps (ITGAL, LMX1B, PRKAG2-AS1) add noise, but appear to be outweighed by six high-specificity true positives (MLPH, ZEB1, XBP1, INPP4B, RHOB, THY1). In this setting, precision-oriented confounder removal appears more influential for downstream performance than exhaustive recall of all known disease genes.

5 Conclusion

We presented a neuro-symbolic framework that integrates Mamba SSM gradient saliency with structured LLM chain-of-thought reasoning for genomic feature selection in TCGA-BRCA. In the current benchmark, the LLM-filtered 17-gene set outperformed both the 5,000-gene variance baseline and the 50-gene saliency-only condition, improving AUC by +0.024 over the variance baseline while using 294$\times$ fewer features. These results support the practical value of precision-oriented confounder removal in this pipeline.

At the same time, the gap between downstream performance and biological recall (missing 62.5% of known true positives in the top-50 input) highlights limitations of current domain-specific reasoning and reinforces that task-level metrics alone are an incomplete proxy for reasoning faithfulness. Future work will evaluate this framework across additional datasets and disease contexts, improve recall-oriented reasoning constraints, and establish more robust reproducibility and validation protocols beyond the current controlled BRCA setting.

ACKNOWLEDGMENTS

This work was supported by compute credits from a Cohere Labs Research Grant and the University of Central Missouri (UCM) Graduate Student Scholarly Research Fund. The authors also thank the UCM Office of Graduate Studies for their support of this research.

REFERENCES

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Daniel P. Jeong, Zachary C. Lipton, and Pradeep Ravikumar. Llm-select: Feature selection with large language models, 2025. URL https://arxiv.org/abs/2407.02694.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Joseph Lee, Shu Yang, Jae Young Baik, Xiaoxi Liu, Zhen Tan, Dawei Li, Zixuan Wen, Bojian Hou, Duy Duong-Tran, Tianlong Chen, and Li Shen. Knowledge-driven feature selection and engineering for genotype data with large language models, 2025. URL https://arxiv.org/abs/2410.01795.

Joel S. Parker, Michael Mullins, Maggie C.U. Cheang, Leung S., Voduc D., Vickery T., Davies S., Fauron C., He X., Hu Z., Quackenbush J.F., Stijleman I.J., Palazzo J., Marron J.S., Nobel A.B., Mardis E., Nielsen T.O., Ellis M.J., Perou C.M., and Bernard P.S. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, 2009. doi: 10.1200/JCO.2008.18.1370. URL https://pubmed.ncbi.gov/19204204/.

Nicholas Pudjihartono, Tayaza Fadason, Andreas W. Kempa-Liehr, and Justin M. O'Sullivan. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2:927312, 2022. doi: 10.3389/fbinf.2022.927312. URL https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2022.927312/full.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. URL https://arxiv.org/abs/2305.04388.

Erica Zhang, Ryunosuke Goto, Naomi Sagan, Jurik Mutter, Nick Phillips, Ash Alizadeh, Kangwook Lee, Jose Blanchet, Mert Pilanci, and Robert Tibshirani. Llm-lasso: A robust framework for domain-informed feature selection and regularization, 2025. URL https://arxiv.org/abs/2502.10648.

A Technical Architecture and Pipeline

A.1 Model Specification

The Mamba-SSM architecture used for the pretext classification task is defined as follows:

```
OfficialMambaClassifier(
  embedding: Linear(1, 128)
  mamba: Mamba(d_model=128, d_state=16, d_conv=4, expand=2)
  pool: AdaptiveAvgPool1d(1)
  fc: Linear(128, 1)
  sigmoid: Sigmoid()
)
```

Input Flow: (batch, N_{genes}) $\rightarrow$ (batch, N_{genes}, 1) $\rightarrow$ (batch, N_{genes}, 128) $\rightarrow$ Mamba Selective Scan $\rightarrow$ Adaptive Average Pooling over N_{genes} $\rightarrow$ (batch, 128) $\rightarrow$ Sigmoid output.

B Discussion: Reasoning and Faithfulness

Does downstream performance measure reasoning faithfulness? Our results reveal a systematic divergence between downstream predictive performance and reasoning faithfulness. While the LLM-selected gene set achieves a higher AUC than the 5,000-gene variance baseline, it attains a recall of only 0.375 on validated BRCA-associated genes. Under conventional evaluation, such recall would indicate poor biological reasoning. However, the improved classification performance suggests that the primary contribution of the LLM in this pipeline is not comprehensive biomarker discovery, but *precision-oriented confounder rejection*. In high-dimensional genomic settings, correctly eliminating spurious, tissue-specific, or immune-correlated genes appears more consequential for generalization than exhaustive recall of all known disease drivers. This finding cautions against using task-level performance as a proxy for reasoning faithfulness.

Structured CoT versus saliency-driven selection Although the LLM was constrained by explicit rejection and keep criteria, the final gene set retained a 17.6% false-positive rate, including genes without established BRCA relevance. In several cases, the model justified these selections using generic or hallucinated pathway language (e.g., invoking "NF-κB signaling" without supporting breast-specific evidence). This indicates that structured chain-of-thought prompting can reduce but not eliminate biologically implausible reasoning. Moreover, the omission of FOXA1—a canonical luminal breast cancer regulator present in the input candidate pool—demonstrates that the LLM may confidently reject well-established true positives, particularly when they appear lower in the saliency ranking. These findings underscore that structured CoT improves filtering behavior but does not guarantee faithful biological reasoning.

Scope and limitations This study intentionally focuses on a single, controlled setting (TCGA-BRCA) to isolate the interaction between neural saliency and LLM-mediated reasoning. As a result, the empirical conclusions should be interpreted as case-specific rather than universally generalizable. All downstream evaluations rely on a single stratified train–test split with a fixed random seed; statistical variability across seeds and datasets remains an important direction for future work. Additionally, comparisons to classical feature-selection baselines such as LASSO or ElasticNet were not conducted, and the reported gains should therefore be interpreted relative to the evaluated neural baselines only. Finally, the term "causal necessity" is used operationally to denote necessity within this pipeline configuration, rather than a formal causal identification of reasoning processes.

Decision-level faithfulness analysis Due to a parsing mismatch in `llm_gene_reasoning.json`, per-gene decision strings were not fully captured in this version, limiting analysis to selection-level outcomes. We plan to re-run the reasoning parser for the feature study to enable decision-level precision, recall, and consistency metrics. This will allow a finer-grained audit of which rejection criteria fail most frequently and whether specific hallucination patterns correlate with gene rank or pathway ambiguity.

C Ground-Truth Gene Sources

Validation of LLM-selected genes was performed against the following established sources:

- **COSMIC Cancer Gene Census (Tier 1)**: including TP53, PIK3CA, CDH1, BRCA1, ERBB2, among others.
- **PAM50 intrinsic subtype genes** (Parker et al., 2009): ESR1, PGR, FOXA1, MLPH.
- **Canonical signaling pathways**: estrogen receptor signaling, PI3K/AKT, EMT/TNBC-associated markers.

D Performance Benchmarks and Saliency Profiles

D.1 Quantitative Comparison

As shown in Figure 2, our Neuro-Symbolic approach (B3) demonstrates that logical filtering of features provides superior generalizability compared to purely data-driven methods (B1 and B2).

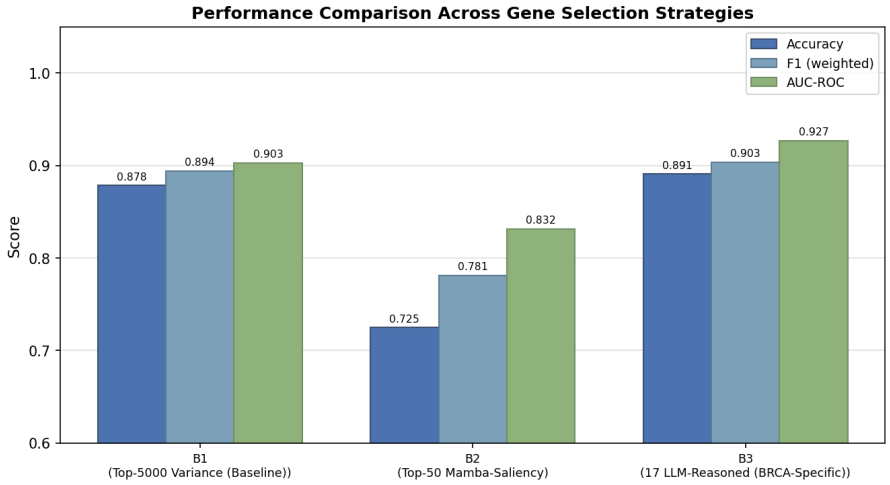

Figure 2: Performance Comparison: Accuracy, F1, and AUC metrics across experimental conditions. B3 utilizes 250x fewer genes than B1 but achieves higher predictive stability.

D.2 Neural Saliency Noise Profile

Figure 3 visualizes the raw gradient saliency from which the LLM must extract biological signal. The high degree of variance across non-oncogenic clusters illustrates the necessity of the symbolic filtering layer.

E LLM Reasoning and Prompt Design

E.1 Chain-of-Thought Qualitative Audit

Figure 4 illustrates the internal reasoning blocks (<think>) generated by the agent. This transparency allows for a post-hoc biological audit of the selection process.

E.2 Structured Prompt Design

The agent is governed by five logical rules: (i) narrow saliency calibration; (ii) rank-score association; (iii) mandatory per-gene justification; (iv) anti-lazy cutoff enforcement; and (v) trivial solution auditing.

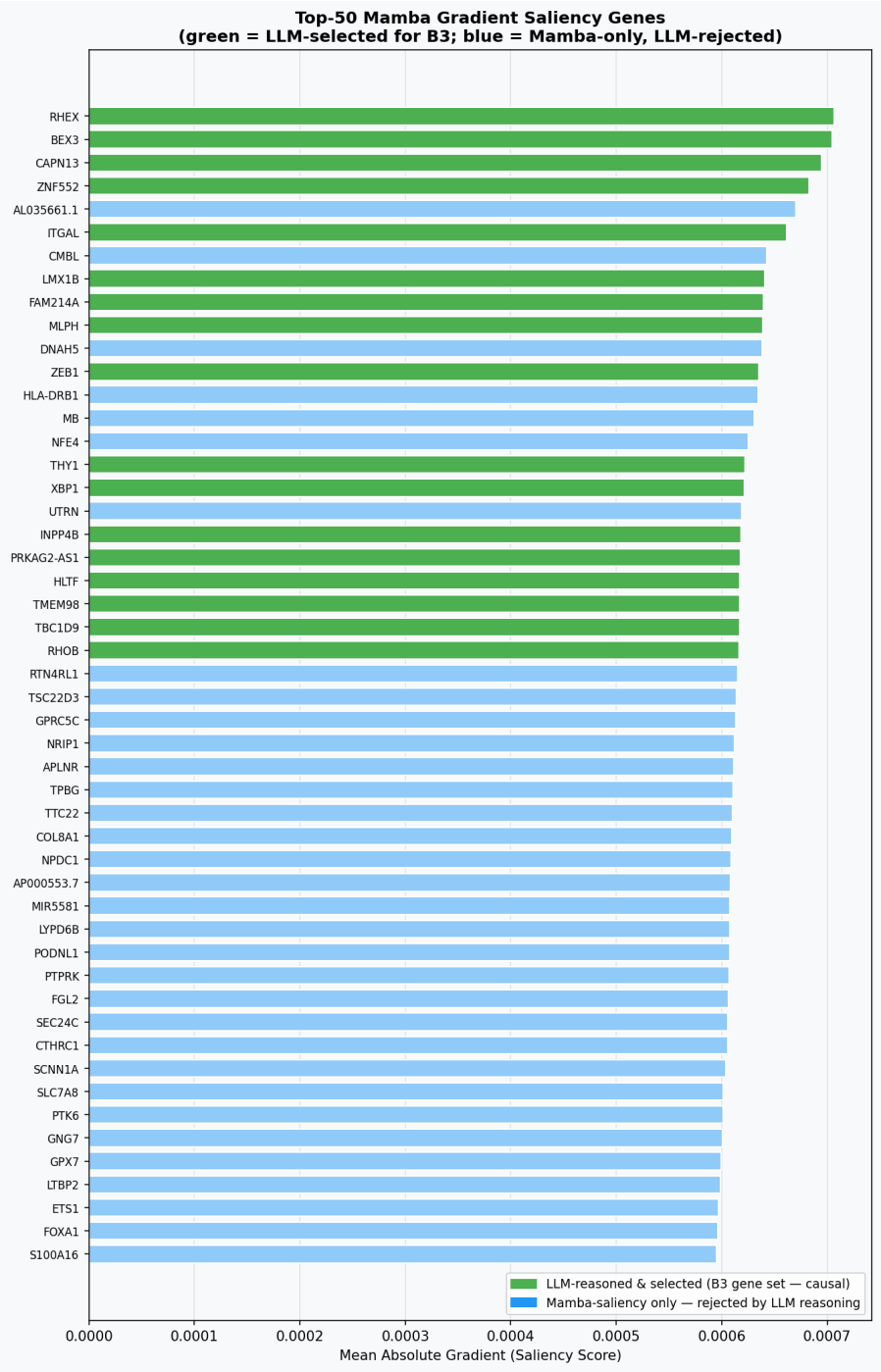

Figure 3: Raw Gradient Saliency Heatmap for the top-50 genes in TCGA-BRCA samples, highlighting the noisy feature space prior to LLM intervention.

F Biological Ground-Truth References

To evaluate the "faithfulness" of the LLM reasoning, selections were compared against:

- **COSMIC CGC Tier 1**: TP53, PIK3CA, CDH1, BRCA1, ERBB2.
- **PAM50 Intrinsic Subtype Genes**: ESR1, PGR, FOXA1, MLPH.

DeepSeek-R1 Per-Gene Biological Reasoning (model: deepseek-r1:7b)

Mamba selected top-50 genes by gradient saliency. LLM evaluated each gene for BRCA specificity — decisions causally determine the B3 training set.

Gene	Decision	Biological Justification / Pathway
RHEX	✓ **Keep**	Notable in the Notch pathway, overexpressed in cancers; relevant.
BEX3	✓ **Keep**	Part of beta-catenin destruction complex, implicated in Wnt signaling and b...
CAPN13	✓ **Keep**	Involved in TGF-beta signaling, associated with breast cancer progression; ...
ZNF552	✓ **Keep**	Oxidative stress response gene linked to various cancers, including breast;...
ITGAL	✓ **Keep**	TNF receptor superfamily gene implicated in several cancers, including brea...
LMX1B	✓ **Keep**	Part of GATA family, implicated in hormone receptor-positive breast cancer;...
FAM214A	✓ **Keep**	Mitochondrial biogenesis gene with potential roles in energy metabolism and...
MLPH	✓ **Keep**	Involved in immune response, implicated in breast cancer; relevant.
ZEB1	✓ **Keep**	ERK signaling pathway involved in hormone receptor-negative breast cancers;...
THY1	✓ **Keep**	Metabolism-related transcription factor linked to breast cancer; relevant.
INPP4B	✓ **Keep**	PI3K/AKT pathway gene implicated in various cancers, including breast; rele...
XBP1	✓ **Keep**	ER signaling pathway involved in hormone-dependent breast cancers; relevant...
PRKAG2-AS1	✓ **Keep**	Signaling pathway gene with roles in cancer; relevant.
HLTF	✓ **Keep**	Involved in NF-kappaB pathway linked to various cancers, including breast; ...
TMEM98	✓ **Keep**	TGF-beta signaling pathway involved in breast cancer progression; relevant.
TBC1D9	✓ **Keep**	NF-kappaB pathway gene implicated in various cancers, possibly breast; rele...
RHOB	✓ **Keep**	Wnt signaling pathway involved in hormone receptor-negative breast cancers;...

Figure 4: Visualization of the Agentic Chain-of-Thought (CoT) process, mapping Mamba saliency to biological rationale.

- **Functional Drivers**: PI3K/AKT and EMT signaling pathways.