A Appendix

Below in A.1 is the prompt used to generate this paper in a single shot, subject to a small number of minor improvements prior to submission. Placeholders inside { } braces are replaced with specific statistics and values from the selected feature disease combination webpage. We also include the AutoInterp reasoning trace in A.2.

A.1 Prompt used to generate paper in one shot

Your task is to write the main body of a LaTex paper to be submitted to the Agents4Science 2025 conference at Stanford. It is structured as an up-to-8 page paper. Write it like an expert academic, and in passive voice. Be comprehensive in writing the paper, the methodology and specific details are what makes this a strong paper. We're looking to get excellent reviews from NeurIPS-style reviewers.

Below I provide you information on the paper you are writing, including links to figures you created, and then you will think about how to write the paper before writing it out in full in one shot. Also, add citations provided using their numbers.

—- Our overall project is structured as follows. This will help you write the Abstract, Introduction and Methods: Uncovering Scientific Knowledge from Multimodal Medical Data through Foundation Models and Mechanistic Interpretability

Background and Motivation Modern healthcare generates a vast and complex array of multimodal data — CT imaging studies, radiology reports, laboratory results, clinical notes, and structured EHR data — all of which contain latent information about disease mechanisms and patient outcomes. Foundation models trained across these modalities have achieved exceptional diagnostic and prognostic performance, uncovering subtle relationships that elude conventional statistical approaches.

However, despite their accuracy, these models remain difficult to interpret. Techniques such as saliency maps or gradient-based attributions offer limited insights into what the models actually learn, particularly when applied across multiple data types. The challenge is not only to explain predictions but to expose the internal representations and mechanisms that connect to real biomedical phenomena.

This work explores a path toward mechanistic interpretability: understanding and restructuring the internal feature spaces of large multimodal foundation models through sparse autoencoders (SAEs). The goal is to convert opaque, high-dimensional representations into sparse, interpretable concept spaces that make model reasoning visible and scientifically meaningful.

Concept and Approach Foundation models capture intricate, multidimensional patterns that encode relationships across imaging, clinical text, and structured health data. These representations are powerful but dense and difficult to parse. Sparse autoencoders provide a systematic way to restructure these representations: by enforcing sparsity, they isolate the key activations that explain most of the variance in the data. Each sparse feature can then correspond to a clinically meaningful concept — for example, a specific imaging phenotype or laboratory pattern.

Prior work outside medicine has shown that such sparse representations can reveal interpretable internal concepts hidden within large models. Translating this approach to the medical domain can provide a way to map model features to physiological and pathological mechanisms. The hypothesis is that SAEs trained on foundation model embeddings will preserve predictive performance while exposing interpretable and clinically relevant structure.

Overview of Method This project leverages a large, IRB-approved multimodal dataset from Stanford Health Care (SHC) and Stanford Health Care Tri-Valley (SHC-TV), which includes CT scans, radiology reports, clinical documentation, laboratory results, ICD codes, and longitudinal outcome data for more than 45,000 patients.

Multimodal Representations Foundation model embeddings will be extracted from both text and imaging data. Large language models such as Gemini and Qwen will be used to embed clinical reports and notes, while Merlin, a multimodal medical foundation model developed by our group, will be used to obtain CT-based image embeddings. These representations will form the input space for subsequent sparse autoencoder training.

Sparse Autoencoder Training Sparse autoencoders will be trained to reorganize the dense foundation model features into sparse, interpretable concept spaces. We will implement and compare architectures such as Top-K SAEs and Matryoshka SAEs, evaluating how effectively they produce disentangled, meaningful concepts that preserve the informational richness of the original embeddings.

Automated Concept Interpretation To assign meaning to each sparse feature, we will construct an automated interpretability pipeline. For every SAE-derived concept, we will identify samples that strongly activate it and use large language or vision–language models to describe shared attributes and propose semantic labels. These candidate concept labels will be iteratively refined and validated on held-out samples to ensure reliability and clinical coherence.

Performance and Utility Assessment The resulting sparse concept representations will be evaluated for their ability to support clinically relevant predictive tasks. We will compare diagnostic and prognostic performance between the sparse concepts and the original foundation model embeddings, and assess whether accurate predictions can be achieved using a smaller, more interpretable subset of sparse features.

Future Directions Once validated, this framework will serve as a foundation for deeper scientific discovery. Sparse, interpretable features derived from foundation models could be used to explore mechanistic hypotheses about disease progression, treatment response, and outcome prediction. For example, newly identified latent concepts may correspond to previously unrecognized imaging biomarkers or physiological signatures that warrant clinical study.

Future work will expand to additional modalities such as genomics and longitudinal EHR data, and integrate causal inference tools to test whether identified concepts represent mechanistic pathways rather than statistical correlations. Further development will also focus on creating clinician-facing interfaces for interactive exploration of the sparse concept space, supporting hypothesis generation directly from model-derived insights.

In the long term, this approach aims to reposition AI in medicine — from systems that merely predict outcomes to systems that reveal structure, generate hypotheses, and advance understanding of human health and disease. By combining large-scale multimodal data from institutions like Stanford Health Care with mechanistic interpretability methods, we move toward an era where foundation models contribute not only to clinical decision-making but to genuine biomedical discovery.

References 1. Cunningham, H., et al. Sparse Autoencoders Find Highly Interpretable Features in Language Models. 2023. arXiv:2309.08600 DOI: 10.48550/arXiv.2309.08600. 2. Van Veen, D., et al., Adapted large language models can outperform medical experts in clinical text summarization. Nat Med, 2024. 30(4): p. 1134-1142. 3. Blankemeier, L., et al., Merlin: A Vision Language Foundation Model for 3D Computed Tomography. Res Sq, 2024. 4. Chen, Z., et al., Chexagent: Towards a foundation model for chest x-ray interpretation. arXiv preprint arXiv:2401.12208, 2024. 5. Gao, L., et al. Scaling and evaluating sparse autoencoders. 2024. arXiv:2406.04093 DOI:10.48550/arXiv.2406.04093. 6. Bussmann, B., et al. Learning Multi-Level Features with Matryoshka Sparse Autoencoders. 2025. arXiv:2503.17547 DOI: 10.48550/arXiv.2503.17547. 7. Templeton, A., et al. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. 2024.

—- Method specifics: Specifically, for this first iteration of the project, we use 25,334 CT scans and train four different Top-K SAEs using sparsity K=5, 10, 20 and 40. We use N=8192 dictionary size. This is a Matryoshka Top-K SAE which uses nested dictionary sizes of [128, 512, 2048 and 8192].

For AutoInterpretation we use LLMs, specifically Gemini 2.5 Pro using the same API. We provide them the anonymized radiology report findings, which describe the CT images in text by human experts. In particular, we provide 20 samples which highly activate the feature to the model, and ask it to summarize the common properties of these samples. Then, we test this interpretation on a held-out set of 20 other highly activating samples. This results in a percentage of generalization of the interpretation.

—- Results specifics: I have created a website which lists all the discovered SAE features and links them to prognostic risk for disease. We are currently on one of the subpages which links feature {feature_name} to disease {disease}. In total, given all SAE alive features across configurations, and their combinations with each disease, we have generated {number_of_hypotheses} hypotheses,

and therefore {number_of_hypotheses} LaTex papers such as this one. For this submission we have selected the generated paper for the hypothesis that feature {feature_name} is highly predictive of onset to {disease} within 0.5-7.5 years. Below we list arrays of statistics that back this up:

{statistics}

This feature was AutoInterpreted, and was given the following interpretation:

{interpretation}

We also compare to the best risk factors that have been discovered through human efforts. These are listed below. Please provide a comparison to each of these in the paper:

{best_established_risk_factors}

This approach, that we can automatically generate a LaTex paper for every SAE feature x disease combination, is also part of the Methods of this paper. On the website, there is a button I just pressed called "Generate LaTex paper" which led to this prompt being created, and then passed to the Gemini 2.5 Pro API. Describe in the paper how we autofill these statistics and figures straight from the website. Also comment how the generation of this entire paper is done in a single shot (make this clear from the abstract).

For the results, you should create tables where necessary based on the provided statistics. We also have several figures: - 'figures/sae_r2.jpg': Original SAE fit statistics, with $R\hat{2}$ on the original feature space reconstruction vs. the sparsity parameter K and number of alive features. The plot shows Matryoshka SAEs in orange and standard Top-K SAEs in blue. It shows a strong pareto improvement of the Matryoshka SAEs over the Top-K SAEs. - 'figures/sae_downstream.jpg': SAE downstream performance, which takes each SAE and uses logistic regressions to predict the binary onset of conditions. - 'figures/feature_discoverer_webapp.jpg': Shows a screenshot of the webpage for feature_name linked to disease, with the 'Generate LaTex paper' button circled in red. - 'figures/feature_disease_plot.jpg': Shows a histogram of feature_names activation bins on the xaxis, and the proportion of individuals in each bin which develop disease. - 'figures/best_odds_ratio_per_sae.jpg': 'Shows the best highest odds ratio feature yielded per SAE configuration for disease.' - 'figures/comparisons_to_human_features.jpg': Shows a barplot of odds ratios for the best established human risk factors (see the statistics for a list) and feature_name for progression to disease.

You may group figures together if you deem it helpful.

—- Discussion and Conclusion sections: Firstly, draw your own conclusions from the statistics provided for this feature and its comparison to existing risk factors. Write a discussion which summarize the strengths of this approach, the fully automatic nature of it, and the excitement around what it can discover.

Also within the discussion, you should mention the limitations of the approach, and how it could be improved, and future work based on the 'overall project structure' outlined in the beginning.

The following conclusion section should be a shorter, punchier version of that.

—- Formatting: For describing the feature and disease names, put 'b́efore any '_' and wrap it in texttt. Finally the prompt used to generate this paper and AutoInterp reasoning trace are listed in the Appendix under ref{appendix:paper_gen_prompt} and ref{appendix:llm_autointerp_reasoning}.

A.2 LLM AutoInterp reasoning

This feature was AutoInterpreted, and was given the following interpretation:

My reasoning process to determine the neuron's detected feature is as follows:

1. **Initial Scan of Reports for Overarching Themes:** I began by reading through all 15 reports to identify any high-level, recurring concepts. Two themes immediately stood out: (1) evidence of prior surgery or current medical intervention, and (2) signs of active inflammation or malignancy. * **Surgery/Intervention:** This was evident from frequent mentions of "surgically absent" organs (especially the gallbladder), "postsurgical changes," "transplant," and the presence of numerous medical devices like catheters, drains, stents, and LVADs. Reports 1, 3, 4, 6, 7, 8, 10, 12, 13, and 14 are strong examples of this. * **Inflammation/Malignancy:** This theme was clear from

descriptions of "inflammation," "peritonitis," "diverticulitis," "phlegmon," "necrotic lymph nodes," and "infiltrative soft tissue." Reports 2, 4, 6, 8, 11, and 15 are strong examples.

2. **Identifying the Common Visual Denominator:** A neuron detects visual patterns, not abstract medical concepts. My next step was to determine the common visual appearance that underlies both "surgery" and "inflammation" on a CT scan. * Surgery inherently disrupts normal anatomy. Incisions create scars, resections leave altered tissue planes, and post-operative healing often involves fluid collections (seromas, hematomas) and inflammatory reactions in the surrounding tissues. * Inflammation (like diverticulitis or peritonitis) and infiltration by tumors also disrupt normal anatomy. The key visual sign for this on CT is often "fat stranding"—a hazy, ill-defined increase in the density of the normally black abdominal fat—as well as fluid collections and abnormal soft tissue masses.

3. **Hypothesis Formulation and Testing:** I hypothesized that the core feature is the visual manifestation of this disruption. I considered several specific phrasings: * *Hypothesis A: "Indwelling medical devices."* This was strong for many reports (e.g., 3, 14) but failed to explain the high activation for reports describing only inflammation (15) or a necrotic mass (2). * *Hypothesis B: "Abnormal fluid collections."* This was present in many cases but was explicitly absent in several high-ranking reports, including the 1 report ("no free fluid"). * *Hypothesis C: "Disruption of abdominal fascial planes."* This was a strong conceptual candidate. Both surgery and inflammation blur the sharp lines that normally define anatomical structures. This concept successfully explained all 15 cases. For example, a surgical incision (1), an infiltrative mass (2), a hematoma with stranding (3), peritonitis (4), and diverticulitis (15) all visually disrupt or blur fascial planes.

4. **Refining the Description for Precision:** The concept of "disruption of fascial planes" is correct, but I needed to translate it into a more specific, descriptive radiological term that captures the primary visual evidence. The most consistent visual signs of this disruption across the reports are abnormal densities within the normally dark fat of the abdomen. This includes the hazy pattern of "stranding" and the more solid appearance of "soft tissue density" (from fluid, phlegmon, scar, or tumor). Therefore, I combined these observations into a single, comprehensive feature description.

5. **Final Verification:** I re-checked my final proposed feature—"Abnormal soft tissue density and stranding within mesenteric, peritoneal, or retroperitoneal fat"—against the highest-ranking reports. * Report 1: "post surgical changes" and "surgical clips" represent abnormal densities in the abdominal wall and retroperitoneal fat. * Report 2: "necrotic ... lymph nodes and soft tissue implants" are abnormal soft tissue densities in the pelvic/retroperitoneal fat. * Report 3: "hematoma" and "stranding" are a perfect match. * Report 4: "peritonitis" and "fluid collection" match the description. * Report 15 (an inflammatory example): "pericolic fat stranding" is a direct match. This description successfully unifies the varied findings (postsurgical changes, inflammation, fluid collections, and masses) into a single, coherent visual feature that a neuron could plausibly detect. It adheres to the prompt's constraints by being specific, avoiding "or," and focusing on appearance rather than pathology.

* This feature represents abnormal soft tissue density and stranding within mesenteric, peritoneal, or retroperitoneal fat.