Sparse Autoencoder Interpretability of the METAGENE-1 Genomic Foundation Model

[1]

Mannat Vikramaditya Jain
Columbia University

Peyton Jackson
Columbia University

Bridget Liu
Columbia University

Ciaran Ramasastry Walsh
Columbia University

Astrid Teo
Columbia University

With
Apart Research

Abstract

Large biological sequence models can support metagenomic surveillance, but their internal representations remain difficult to interpret. We apply sparse autoencoders (SAEs) to METAGENE-1, a 7B-parameter metagenomic foundation model, to test whether its residual-stream activations contain sparse, biologically meaningful features associated with a binary pathogen/non-pathogen label. We extracted residual activations from selected transformer layers, trained BatchTopK SAEs with 32,768 latent features, and analyzed mean-pooled sequence-level SAE features. Layer-32 SAE features strongly encoded the binary target: a logistic-regression probe achieved 94.55% held-out accuracy and AUROC 0.9874, and the signal transferred to a second sequencing delivery with AUROC 0.9840. Feature-level enrichment analysis identified 16,519 pathogen-enriched latents at FDR < 0.01, and BLAST validation of top-activating sequences revealed organism-specific detectors, including high-confidence latents for Human astrovirus, Norovirus GI, and Norovirus GII. Multi-layer analysis showed that comparable pathogen/non-pathogen separability was already present by layer 8 and persisted through layers 16, 24, and 32. Together, these results show that SAE decompositions of METAGENE-1 recover sparse features that are both predictive and partially biologically interpretable, suggesting a path toward auditable metagenomic classifiers based on foundation-model internals.

1 Introduction

Recent advances in protein language models (pLMs) have shown that the transformer architecture can learn rich biological representations from sequence alone [12, 5]. Models trained on large-scale biological data have achieved strong performance on a range of downstream tasks, suggesting that they capture meaningful statistical and structural regularities in biological sequence space [12, 5, 10]. At the same time, the internal mechanisms by which these models represent and use biological information remain poorly understood.

In natural language processing, recent progress in mechanistic interpretability has provided tools for studying internal model representations in greater detail. Sparse autoencoders (SAEs), feature attribution, activation patching, and automated feature labeling have enabled researchers to decompose neural activations into more interpretable latent features and to study how these features influence model behavior [2, 14, 6, 11]. A central motivation for this work is the problem of polysemanticity: individual neurons often encode multiple unrelated concepts through superposition, making them difficult to interpret directly [4]. SAEs address this by learning a sparse latent dictionary that can recover more interpretable features from dense activations.

[1] Research conducted at the AIxBio Hackathon, April 2026

Interpreting the function of biological models is especially important in settings where model predictions may influence scientific or public-health decisions [3, 9]. In metagenomic surveillance, for example, a model may detect patterns associated with human-infecting or pathogenic viral sequences, but without interpretability, it is difficult to know whether the model is using biologically meaningful signals or exploiting spurious statistical correlations and tokenization artifacts [16, 15]. This matters for both scientific understanding and trustworthiness in high-stakes biosecurity settings.

Recent work has begun to apply mechanistic interpretability to biological sequence models. In protein language models, sparse autoencoders have been used to identify interpretable features associated with protein families, structural motifs, and other biologically meaningful properties [13, 1, 7]. However, prior SAE studies focus primarily on protein language models such as ESM-2, while comparable analyses of large-scale metagenomic foundation models remain limited.

We address this gap by applying SAEs to METAGENE-1 and evaluating whether learned features associate with human-infecting or pathogenic viral labels. We train a sparse autoencoder on residual stream activations of METAGENE-1 and analyze whether the resulting sparse latent features encode biologically meaningful signals associated with the target label. We seek to learn how the model's learned representations can be decomposed into features that are statistically associated with pathogenic versus non-pathogenic sequences and that admit preliminary biological interpretation.

Our main contributions are:

1. We apply sparse autoencoder-based mechanistic interpretability to METAGENE-1, extending SAE analysis from protein language models to a large metagenomic foundation model.
2. We show that SAE features extracted from METAGENE-1 strongly encode a binary pathogenic versus non-pathogenic viral label, using differential feature analysis, linear probing, and low-dimensional visualization.
3. We identify class-associated latent features and candidate sequence motifs that provide a foundation for downstream biological interpretation of METAGENE-1's internal representations.

2 Related Work

2.1 Sparse Autoencoders

LLMs are believed to encode multiple features per neuron through superposition, leading to polysemanticity [4]. Bricken et al. introduced the sparse autoencoder as a means to allow for more effective analysis of feature representation [2]. The sparse autoencoder framework was further scaled by Templeton et al. [14], while Gao et al. used k-sparse autoencoders to improve reconstruction and reduce dead latents [6].

Paulo et al. introduced a framework for evaluating and classifying features of SAEs using LLMs [11].

2.2 Interpretability for protein language models

Previous work on pLM interpretability has examined perturbation-based analysis [8] as well as sparse feature discovery in protein language models [13, 1, 7].

Simon and Zou trained sparse autoencoders on the residual stream of a pLM, ESM-2 [13]. Similarly, Adams et al. identified various family- and function-specific features using SAE activations, compiled in the visualizer InterProt [1]. Related recent work has also demonstrated that sparse autoencoders can recover biologically interpretable features from protein language model representations [7]. Taken together, this signifies a potential for interpretability work using SAEs to provide biological insight including pathogen identification.

Prior SAE studies focus primarily on protein language models such as ESM-2, while comparable analyses of large-scale metagenomic foundation models remain limited. We address this by applying SAEs to METAGENE-1, described below, and evaluating whether learned features associate with human-infecting viral labels.

3 Methods

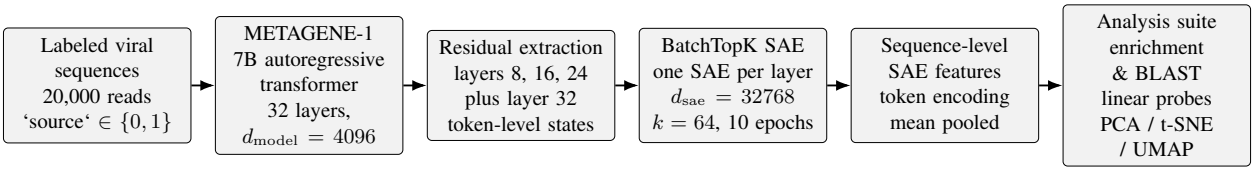

Figure 1: Overview of the analysis pipeline. Labeled viral sequences, where 0/1 denotes non-pathogen/pathogen, are passed through METAGENE-1, token-level residual activations are extracted from selected transformer layers, and a separate BatchTopK sparse autoencoder is trained at each layer. The trained autoencoders are then used to produce sequence-level sparse feature vectors for downstream enrichment analysis, linear probing, clustering, and visualization.

3.1 Model and dataset

We study METAGENE-1, a 7B-parameter autoregressive transformer for metagenomic sequence modeling [9]. The model uses a decoder-only architecture with 32 transformer layers, hidden size 4096, intermediate size 11008, 32 attention heads, RMSNorm, rotary positional embeddings, and a 1024-token BPE vocabulary over biological sequence data [9]. The model is treated as a fixed pretrained foundation model throughout.

Our primary SAE training and feature analyses used the labeled `human_virus_class1` derived from data in the METAGENE-1 Hugging Face dataset, containing 20,000 sequences balanced across the binary `source` label. We also used a second labeled delivery of similar origin, `human_virus_class2`, for cross-delivery validation. In these files, the `class` field denotes the dataset split or delivery (`class-1` or `class-2`), while `source` is the binary prediction target used in our analyses. The class-1 labeled file contains 20,000 sequences with balanced binary labels in the `source` field: 10,000 labeled `0` and 10,000 labeled `1`. We treat `source` as the binary non-pathogenic versus pathogenic target, respectively. The `class` field is not the prediction target; it denotes the dataset split or delivery. For cross-delivery validation, we also used `human_virus_class2_labeled.jsonl`, which contains another 20,000 sequences with the same balanced `source` labels and `class` uniformly set to `class-2`. In the labeled class-1 and class-2 files, sequence lengths range from approximately 100 to 291 nucleotides.

3.2 Residual-stream extraction and layer selection

We extracted METAGENE-1 residual-stream activations using the repository's extraction pipeline. Input sequences were uppercased, whitespace was stripped, and any character outside the allowed alphabet `ACGTUN` was replaced with `N`. Sequences with more than 5% invalid characters were excluded. Extraction used a maximum sequence length of 512 and processed up to 20,000 reads from each of `human_virus_class1.jsonl` and `human_virus_class2.jsonl`.

The original SAE baseline was trained on layer 32, the final transformer layer. We then extended the same pipeline to layers 8, 16, and 24 in order to sample early, middle, and late computation within the 32-layer network while keeping extraction and storage manageable. This lets us ask whether pathogen-related information only appears late or is already linearly accessible in earlier residual streams. Mean-pooled token-level residual activations were stored as float32 vectors.

3.3 Sparse autoencoder architecture and training

We trained BatchTopK sparse autoencoders on METAGENE-1 residual-stream activations. Each SAE took $d_{\text{model}} = 4096$-dimensional residual vectors as input and used an expansion factor of 8, giving $d_{\text{sae}} = 32{,}768$ sparse latents. We set the BatchTopK sparsity target to $k = 64$, so that approximately 64 latents were active per token on average, while allowing the number of active latents to vary across individual tokens. The decoder was trained to reconstruct the normalized residual activation from the sparse code, with decoder feature vectors constrained to unit norm.

All layer-specific SAEs used the same training configuration: Adam optimization[1], learning rate 2×10^{-4}, batch size 4096 token activations, 10 epochs, expansion factor 8, and $k = 64$. Activations were normalized by the mean training-vector ℓ_2 norm. We also used an auxiliary dead-feature loss to provide gradient signal to inactive features during training.[2]

We fully analyzed one layer-32 SAE and three additional SAEs trained at layers 8, 16, and 24. The organism-labeling, probe, SAE health-check, sequence-embedding, latent-clustering, and cross-delivery analyses were run on the layer-32 SAE, while the same per-layer analysis procedure was additionally applied to the layer-8, layer-16, and layer-24 SAEs for the depth comparison.

We verified SAE training quality before downstream analysis. The layer-32 BatchTopK SAE trained stably: reconstruction loss decreased rapidly and then plateaued near 5×10^{-5} MSE, while the dead-feature fraction rose transiently during early optimization and declined by the end of training, consistent with the auxiliary dead-feature loss preventing persistent feature collapse. The earlier-layer SAEs trained with the same architecture and hyperparameters had higher final reconstruction losses, approximately 1.2×10^{-4} for layers 8 and 16 and 1.3×10^{-4} for layer 24, suggesting that these residual streams were less compressible under the same TopK bottleneck, though we found that feature prediction strength was still maintained. After encoding, the layer-32 sequence-level feature matrix had 31,965 alive latents out of 32,768, a 2.45% dead-latent rate, and 2.72% nonzero entries across the $20{,}000 \times 32{,}768$ matrix. Although each token was encoded with TopK $k = 64$, mean pooling across token positions produced an average of 892.10 nonzero latents per sequence because different positions activated different features.

[1] $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

[2] Features that had not fired for 200 optimization steps were treated as dead and assigned an auxiliary reconstruction penalty weighted by 0.03125. The decoder bias was initialized to the mean activation vector. Decoder gradients were projected to remove components parallel to decoder directions after each step, and decoder vectors were renormalized to unit norm.

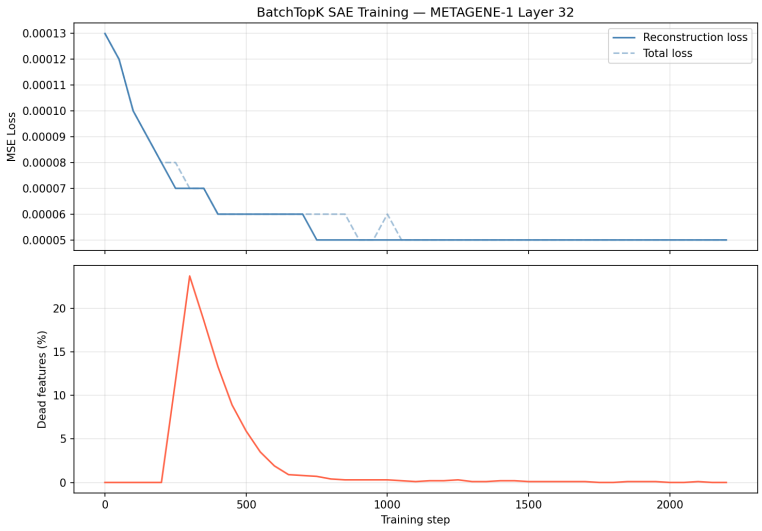

Figure 2: Training diagnostics for the layer-32 BatchTopK SAE. Reconstruction loss falls quickly and then stabilizes, while the fraction of dead features rises transiently during early competition among features and then declines by the end of training.

3.4 Sequence-level SAE features

After training, token-level activations were encoded through the SAE and converted to sequence-level feature vectors for downstream analysis. At inference time, the encoder output was sparsified with per-token TopK selection, so that each token retained exactly 64 active latent features. For each sequence, token-level latent activations were then mean-pooled across tokens to produce a single 32,768-dimensional sequence representation. These sequence-level features were saved as a matrix of shape $20{,}000 \times 32{,}768$, with row order tracked by `sequence_ids.json`.

3.5 Experimental analyses

All downstream analyses were carried out on sequence-level SAE features aligned to the binary `source` label by `sequence_id`.

For feature-level association, we compared latent activation between the two `source` labels using activation-frequency and activation-magnitude statistics, including Fisher exact tests, Wilcoxon rank-sum tests, and Benjamini–Hochberg correction for multiple testing. To test whether the enriched latents corresponded to recognizable biological entities, we selected 50 candidate pathogen-associated latents, retrieved 10 top-activating pathogen sequences per latent, and submitted all 500 sequences to BLAST.

For predictive analysis, we trained logistic-regression probes on mean-pooled sequence-level SAE features using stratified train-test splits. For geometric analysis, we reduced sequence representations with PCA followed by UMAP and used HDBSCAN to summarize cluster structure. We also analyzed latent-level organization by clustering alive latents according to their activation patterns across sequences.

To test robustness, we evaluated whether the layer-32 pathogenicity signal transferred from the class-1 delivery to the class-2 delivery and compared per-latent enrichment statistics across deliveries. For the multi-layer comparison, we applied the same core analysis pipeline to SAEs trained on layers 8, 16, and 24.

Because the layer-32 and earlier-layer probes used slightly different logistic-regression implementations, we use the earlier-layer results primarily to assess whether pathogenicity information is present across depth rather than to claim a strict ranking among all four layers.

4 Results

4.1 Sparse features recover organism-specific viral signals

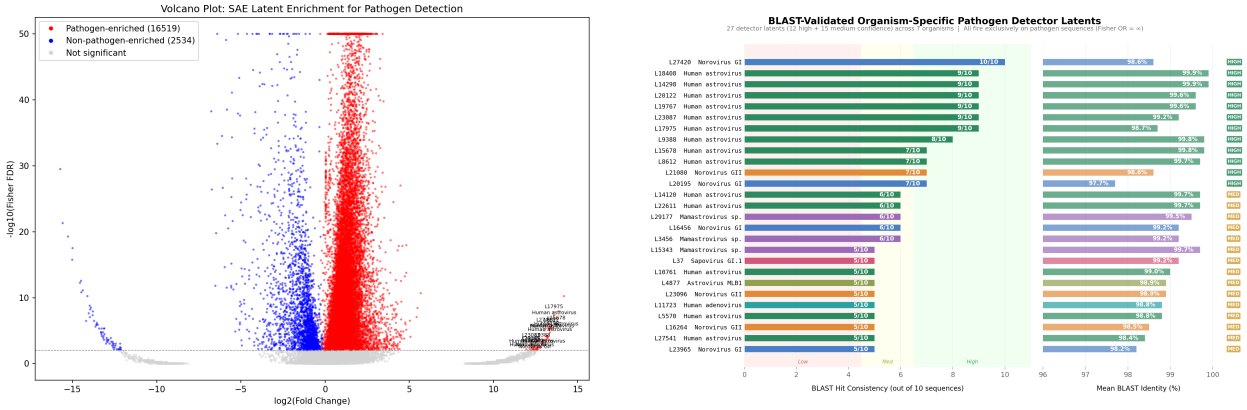

Figure 3: Layer-32 organism-detector results. Left: enrichment volcano plot over all 32,768 latents. Right: top high- and medium-confidence organism detectors validated by BLAST.

Across all 32,768 latents, Fisher tests on active versus inactive status of layer-32 latents identified 16,519 pathogen-enriched latents at FDR < 0.01 and odds ratio greater than 1, compared with 2,534 non-pathogen-enriched latents at the same threshold. This broad enrichment does not by itself prove interpretability, but it shows that many sparse features are statistically associated with the binary label.

All 500 sequences submitted to BLAST returned hits. Aggregating the top BLAST hits by latent produced 12 high-confidence organism detectors and 15 medium-confidence detectors. The high-confidence detectors corresponded to Human astrovirus (9 latents), Norovirus GI (2 latents), and Norovirus GII (1 latent), with detector-level mean percent identity ranging from 97.7% to 99.9%.

The high-confidence detectors were also highly label-specific: none of the 12 activated on non-pathogen sequences under the threshold used for the scan. Thus, the SAE features were not only predictive of the label; a subset of them mapped to externally identifiable viral organisms.

4.2 Pathogenicity is linearly decodable and geometrically separated

The sequence-level SAE feature space made the binary label highly accessible to simple linear models. On an 80/20 stratified train-test split of the 20,000 class-1 sequences, a logistic-regression probe trained on layer-32 SAE features achieved 94.55% held-out accuracy, Matthews correlation coefficient 0.8916, and AUROC 0.9874 (Figure 4, left). These metrics indicate that the mean-pooled SAE features preserve most of the information needed for pathogen versus non-pathogen classification.

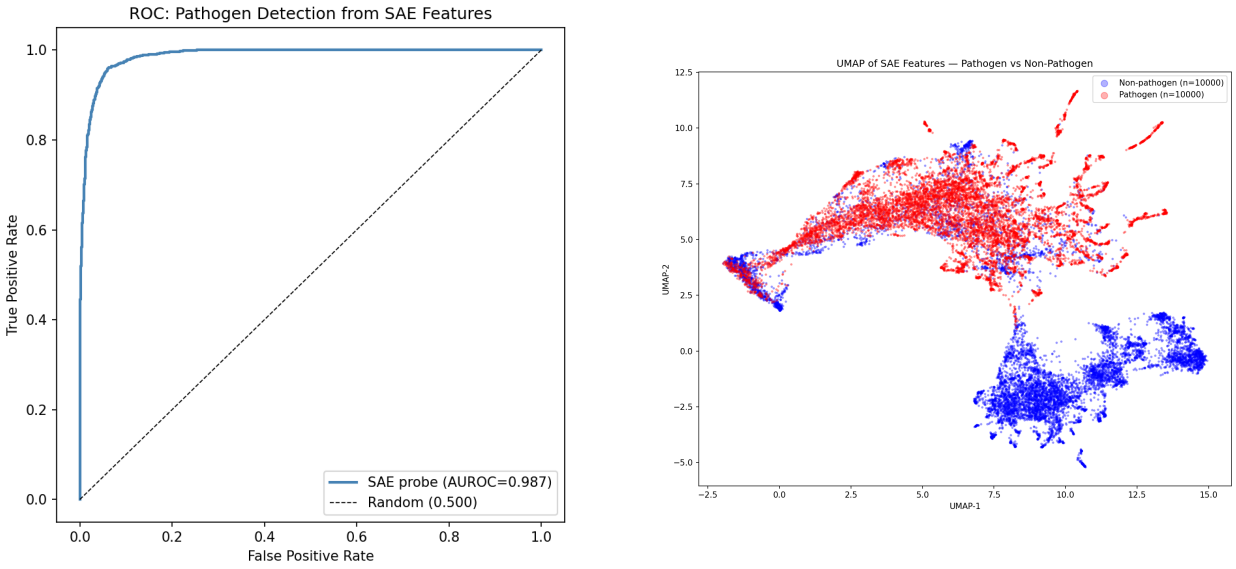

Figure 4: Pathogenicity is strongly separable in the layer-32 SAE feature space. Left: ROC curve for a logistic-regression probe trained on mean-pooled sequence-level SAE features, achieving AUROC 0.9874 on a held-out test split. Right: two-dimensional UMAP projection of the same feature matrix after PCA preprocessing, showing that pathogen and non-pathogen sequences occupy largely distinct regions of the learned sparse representation.

The probe coefficients show that this signal is distributed across the dictionary. Ranking latents by absolute coefficient magnitude, the top 2,068 latents accounted for 50% of cumulative coefficient mass, while the top 10,847 were required to reach 90%. Thus, classification performance does not come from a single dominant feature. Instead, pathogenicity is encoded by a moderately distributed set of sparse latents.

The same separation is visible geometrically. Reducing the 32,768-dimensional SAE features to 50 principal components explained 70.74% of variance. A UMAP embedding of the reduced representation separated pathogen and non-pathogen reads into distinct regions (Figure 4, right). HDBSCAN on the two-dimensional embedding produced 49 clusters plus a noise class; 26 clusters had greater than 90% pathogenic purity. The embedding and cluster statistics support the probe result that pathogenicity is a major axis of structure in the SAE feature space.

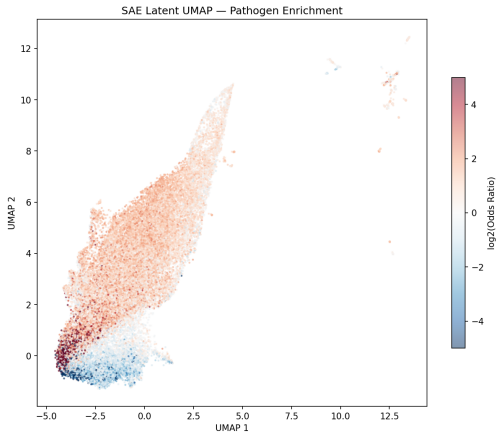

Figure 5: Latent-space UMAP for layer 32, colored by pathogen enrichment. Pathogen-associated features do not form many clearly separated groups. Instead, enrichment changes gradually across the latent space: HDBSCAN placed 31,251 of 32,703 alive latents, or 95.5%, into one dominant cluster. This suggests that pathogen information is spread across many related features, even though some individual latents act as specific organism detectors.

The latent-level view refines the interpretation of the detector result. While some individual features are highly specific, the broader pathogen signal appears continuous and distributed across many latents rather than partitioned into sharply separated modules (Figure 5).

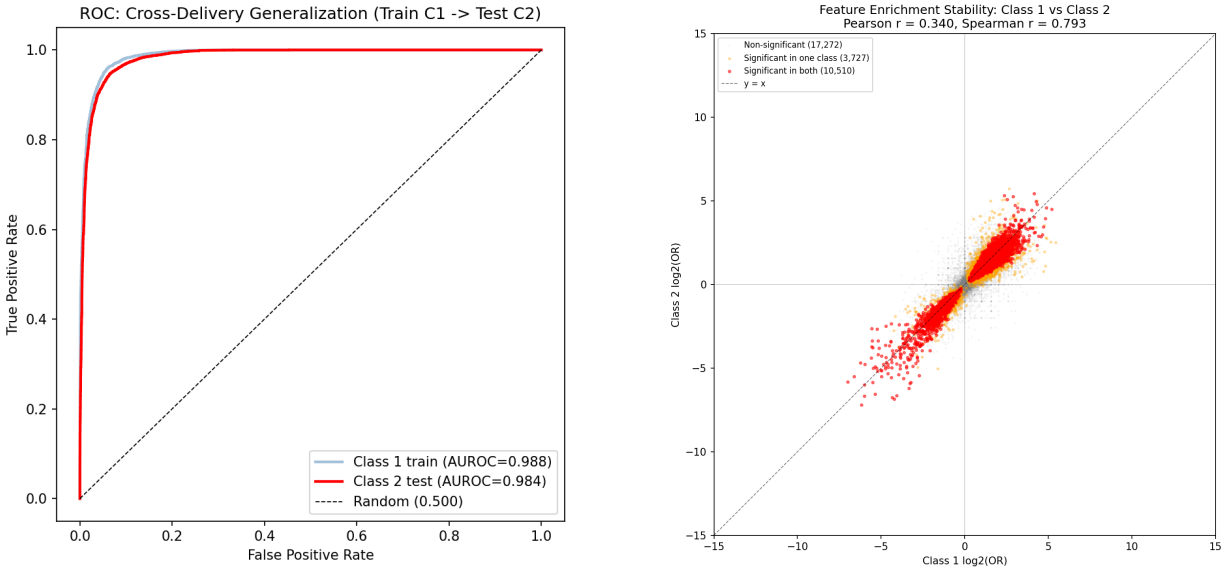

Figure 6: Cross-delivery validation. Left: the probe trained on Class 1 retains strong ROC performance on Class 2. Right: per-latent pathogen enrichment is strongly rank-preserved across deliveries, especially among significant latents.

A logistic-regression probe trained on all 20,000 Class 1 sequences achieved 94.76% accuracy and AUROC 0.9880 on Class 1, and retained 93.96% accuracy and AUROC 0.9840 when evaluated on 20,000 Class 2 sequences. Feature-level enrichment was also highly consistent across deliveries: 10,510 latents were

significant in both Class 1 and Class 2, corresponding to 83.5% of the Class 1 significant set, and enrichment scores were strongly rank-correlated across deliveries (Spearman $\rho = 0.7933$ across finite latents; $\rho = 0.8692$ among significant latents). These results suggest that the pathogenicity signal is not specific to a single sequencing delivery.

4.3 Pathogen information is present across model depth

Finally, for comparison of SAEs trained at layers 8, 16, and 24, the metrics are nearly identical across these intermediate layers[3], suggesting that the pathogen/non-pathogen distinction is available by layer 8 and remains available through layer 32. The PCA projections in Figure 7 show the same qualitative pattern: class structure is already visible in the early-layer SAE representation and persists at later depths. The supported conclusion is that strong pathogen information is already present in early and middle residual streams, while the particular SAE basis vectors used to express it evolve across layers.

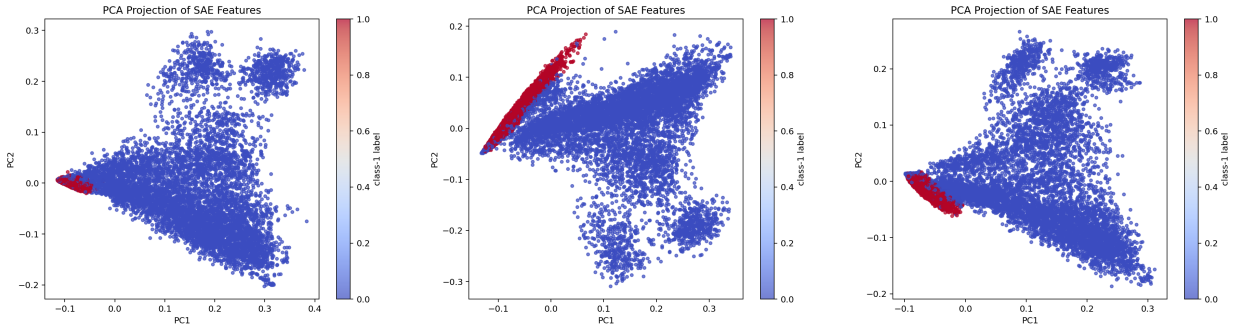

Figure 7: PCA projections of sequence-level SAE features from layers 8, 16, and 24. Class separation is already visible at layer 8 and remains comparably strong through layer 24.

We repeated the full organism-detector pipeline on the layer-16 SAE. BLAST validation of the top 50 candidate latents produced 30 high-confidence organism-specific detectors at layer 16, compared with 12 at layer 32. The layer-16 detectors covered a broader taxonomic range: Human astrovirus (13 latents), Norovirus GII (5), Sapovirus GI (4), Norovirus GI (4), Mamastrovirus sp. (3), and Astrovirus MLB1 (1), whereas layer-32 high-confidence detectors were limited to Human astrovirus (9), Norovirus GI (2), and Norovirus GII (1). All 30 had mean BLAST percent identity above 96% and hit consistency of at least 7/10. METAGENE-1 therefore builds organism-level representations by at least mid-network; the final layer appears to compress species-level information into a more distributed pathogen-versus-non-pathogen encoding.

The enrichment statistics reveal a depth-dependent shift. At layers 8 and 16, enrichment is approximately symmetric (6,566 versus 6,476 and 5,855 versus 5,740 pathogen- versus non-pathogen-enriched latents). By layer 32 the distribution is heavily skewed: 16,519 pathogen-enriched versus 2,534 non-pathogen-enriched, a ratio of 6.5:1. Early layers also concentrate class-discriminative information into fewer, more specific features (52 latents with F1 > 0.7 at layer 8, versus 4 at layer 32). This is consistent with a transition from sparse organism-level detectors in early layers to distributed pathogen encoding in the final layer.

[3]Layer 8 achieved AUROC 0.9912, AUPRC 0.9901, accuracy 0.9545, and F1 0.9552. Layer 16 achieved AUROC 0.9906, AUPRC 0.9892, accuracy 0.9540, and F1 0.9548. Layer 24 achieved AUROC 0.9914, AUPRC 0.9908, accuracy 0.9535, and F1 0.9543; the layer-32 probe reached AUROC 0.9874 and accuracy 94.55%, but it was trained with a different logistic-regression setup than the layer-8/16/24 reports

	Layer 16	Layer 32
High-confidence detectors	30	12
Medium-confidence detectors	13	15
Pathogen-enriched (FDR < 0.01)	5,855	16,519
Non-pathogen-enriched (FDR < 0.01)	5,740	2,534

Table 1: Organism-specific detectors at layers 16 and 32. High-confidence requires BLAST hit consistency $\geq 7/10$ for a single organism.

5 Discussion and Limitations

Our results demonstrate that sparse autoencoders applied to a metagenomic foundation model recover features with clear biological correspondence. Three findings stand out.

First, the SAE learns organism-specific viral detectors without any organism-level supervision. METAGENE-1 was pretrained with a next-token prediction objective on raw metagenomic reads; the SAE was trained to reconstruct residual-stream activations with no access to labels of any kind, and yet individual latents fire selectively on sequences from specific viral species—Human astrovirus, Norovirus GI and GII, Sapovirus GI, Mamastrovirus sp., and Astrovirus MLB1—with BLAST-confirmed identity above 96% and zero false-positive activations on non-pathogen sequences for the highest-confidence detectors. This parallels the finding by Adams et al. that SAE latents trained on ESM-2 residual streams correspond to specific protein families [1], and extends it from protein family membership to viral species identity in a metagenomic context.

Second, the pathogenicity signal is linearly accessible throughout the network. Probe AUROC is consistently high from layer 8 (0.9912) through layer 32 (0.9874), indicating that METAGENE-1 encodes pathogen-versus-non-pathogen information early and preserves it across all subsequent layers. This contrasts with the common expectation that more abstract or task-relevant features emerge only in later layers, and is consistent with recent observations that many downstream tasks in protein language models rely on features available in early layers [1].

Third, while character of representation changes with depth, classification performance does not. Early and middle layers maintain approximately symmetric enrichment between pathogen and non-pathogen classes and concentrate organism-level specificity into a small number of highly selective latents. This implies that pathogenicity information is already accessible early in METAGENE-1 and is preserved across depth, while later layers reorganize rather than simply create the relevant biological signal.

These findings have practical implications for metagenomic surveillance. An interpretable pathogen classifier built on SAE features could flag not only that a sample contains pathogenic material, but which specific organisms are driving the prediction without requiring a separate taxonomic classification pipeline. The organism-specific detectors we identify here could serve as the basis for a sparse, auditable early-warning system in which each detector latent corresponds to a known viral taxon. The cross-delivery validation (4.2) provides initial evidence that such a system would generalize beyond a single sequencing batch.

5.1 Limitations

All downstream analyses use mean-pooled token-level SAE activations, which collapse positional information and prevent identification of which subsequences drive each latent. We completed BLAST validation only at layers 16 and 32; the organism-detector claims at layers 8 and 24 rest on enrichment statistics alone.

We did not compare SAE feature probes against raw residual-stream probes, so we cannot quantify whether the sparse decomposition preserves or degrades classification performance relative to the base model representations. Finally, we trained all SAEs with a single hyperparameter configuration and did not sweep over sparsity level or dictionary size.

5.2 Future Work

Token-level pathogen localization, scoring each nucleotide position by the dot product of its per-token SAE activation with the probe coefficient vector, would identify which subsequences drive pathogen predictions and enable BLAST grounding to specific genes, producing heatmap visualizations analogous to InterProt's per-residue activation overlays [1]. Classifying all latents by spatial firing pattern (point, motif, periodic, whole-sequence) across layers could explain the enrichment symmetry shift we observe with depth. A multi-class species classifier trained on BLAST-derived organism labels would test whether the sparse dictionary supports fine-grained taxonomic resolution beyond binary pathogen detection.

6 Conclusion

We show that BatchTopK SAEs can be applied at multiple layers of the genomic foundation model METAGENE-1 to learn sparse features that strongly encode pathogenic versus non-pathogenic viral labels. These features support high-performing linear probes, show clear geometric separation between labeled classes, and remain informative across model depth. We further find that a subset of pathogen-associated features can be linked by BLAST to recognizable viral organisms or families, suggesting that some SAE latents capture biologically interpretable sequence signals rather than only abstract classifier information.

Other Materials

- Code repository: `https://github.com/mannatvjain/metageniuses`
- Interactive website: `https://mannatvjain.github.io/metageniuses`

Author Contributions

Astrid Teo and Ciaran Ramasastry Walsh curated and prepared data. Bridget Liu developed the SAE and feature scripts. Peyton Jackson led paper writing and assisted in both training and using the SAE, as well as performing analysis of results. Mannat Vikramaditya Jain created and led the analysis.

References

[1] Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed Alquraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 460–476, 2025.

[2] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. `https://transformer-circuits.pub/2023/monosemantic-features/index.html`, 2023. Transformer Circuits Thread.

[3] Megan B. Diamond, Aparna Keshaviah, Ana I. Bento, Otakuye Conroy-Ben, Erin M. Driver, Katherine B. Ensor, Rolf U. Halden, Loren P. Hopkins, Katrin G. Kuhn, Christine L. Moe, Eric C. Rouchka, Ted Smith, Bradley S. Stevenson, Zachary Susswein, Jason R. Vogel, Marlene K. Wolfe, Lauren B. Stadler, and Samuel V. Scarpino. Wastewater surveillance of pathogens can inform public health responses. *Nature Medicine*, 28:1992–1995, 2022.

[4] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[5] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022.

[6] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

[7] Onkar Gujral, Mihir Bafna, Eric Alm, and Bonnie Berger. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, 2025.

[8] Peter K. Koo, Antonio Majdandzic, Matthew Ploenzke, Praveen Anand, and Steffan B. Paul. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLOS Computational Biology*, 17(5):e1008925, 2021.

[9] Ollie Liu, Sami Jaghouar, Johannes Hagemann, Shangshang Wang, Jason Wiemels, Jeff Kaufman, and Willie Neiswanger. Metagene-1: Metagenomic foundation model for pandemic monitoring. *arXiv preprint arXiv:2501.02045*, 2025.

[10] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems 34*, 2021.

[11] Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*, 2024.

[12] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jianyuan Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[13] Elana Simon and James Zou. Interplm: Discovering interpretable features in protein language models via sparse autoencoders. *arXiv preprint arXiv:2412.12101*, 2024.

[14] Adly Templeton et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. `https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html`, 2024. Transformer Circuits Thread.

[15] Xiaoxiao Zhou, Zihan Wang, Jingbo Shang, and Yang E. Li. Dnamotiftokenizer: Towards biologically informed tokenization of genomic sequences. *arXiv preprint arXiv:2512.17126*, 2025. Also submitted to ICLR 2026 on OpenReview.

[16] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2024. Published as a conference paper at ICLR 2024.

LLM Usage Statement

We used Claude Code and Codex to facilitate programming, data preparation, and aid in drafting sections of the paper. All results and claims were independently verified.