Mechanistic Interpretability of Antibody Language Models Using SAEs

Rebonto Haque
Department of Statistics
University of Oxford
Oxford, UK

Oliver Turnbull
Department of Statistics
University of Oxford
Oxford, UK

Nithin Parsan
Reticular
San Francisco, CA
USA

Anisha Parsan
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology (MIT)
Cambridge, MA, USA

John Yang
Reticular
San Francisco, CA
USA

Charlotte M. Deane[*]
Department of Statistics
University of Oxford
Oxford, UK

Abstract

Sparse autoencoders (SAEs) are a mechanistic interpretability technique that have been used to provide insight into learned concepts within large protein language models. Here, we employ TopK and Ordered SAEs to investigate an autoregressive antibody language model, p-IgGen, and steer its generation. We show that TopK SAEs can reveal biologically meaningful latent features, but high feature–concept correlation does not guarantee causal control over generation. In contrast, Ordered SAEs impose a hierarchical structure that reliably identifies steerable features, but at the expense of more complex and less interpretable activation patterns. These findings advance the mechanistic interpretability of domain-specific protein language models and suggest that, while TopK SAEs suffice for mapping latent features to concepts, Ordered SAEs are preferable when precise generative steering is required.

1 Introduction

Antibodies are a key part of the body's adaptive immune response and are characterised by their ability to bind to a specific antigen and subsequently neutralise it or initiate an immune response. They possess significant sequence, and therefore structural, diversity, which enables binding to virtually any target antigen [Chiu et al., 2019].

The antigen-binding domain of antibodies consists of variable heavy (VH) and variable light (VL) chains, whose binding specificity and affinity are primarily determined by six complementarity-determining regions (CDRs), three on each chain. Numerous V (variable), D (diversity), and J (joining) gene segments in the genome encode these chains, with combinatorial assembly of V, D, and J for VH and V and J for VL generating substantial sequence diversity. This diversity is further enhanced by somatic hypermutation, in which random nucleotide substitutions occur at markedly elevated rates within the rearranged V(D)J segment [Andreano and Rappuoli, 2021].

[*] Correspondence: `deane@stats.ox.ac.uk`

The ability to bind any target antigen with high specificity and affinity makes antibodies ideal candidates for drug discovery. As a result, antibody drugs hold a major and growing share of the total pharmaceutical market [Crescioli et al., 2025]. Antibody drug development pipelines need to identify candidates which bind specifically and with high affinity to the target antigen, while also being ‘developable’ [Jarasch et al., 2015]. ‘Developability’ refers to properties required for a successful drug such as immunogenicity, solubility, specificity, stability, manufacturability, and storability [Raybould and Deane, 2022].

Antibody language models have been used to optimise multiple steps of antibody-drug development pipelines from library generation [Turnbull et al., 2024] to humanisation during lead optimisation [Chinery et al., 2024]. p-IgGen is a GPT-like decoder-only model trained on antibody-sequence data, consisting of 17M parameters [Turnbull et al., 2024]. The authors released a paired model, as well as a finetuned version capable of generating diverse antibody libraries with developable properties.

The lack of interpretability of machine learning models contributes to a lack of trust in model predictions, difficulty determining whether biologically relevant features are being used to make predictions and difficulty detecting overfitting. Collectively, these pose a barrier when employing language models for drug discovery [Chen et al., 2023]. SAEs offer a promising approach to identify human-interpretable concepts learned by models and steer their generation [Chen et al., 2025, Templeton et al., 2024]. Prior works have used SAEs to understand the inner mechanisms of PLMs [Adams et al., 2025, Parsan et al., 2025, Simon and Zou, 2024], and steer model output. However, to date, SAEs have not been used to interrogate autoregressive protein or antibody-specific language models.

This work advances the interpretability of antibody language models, using SAEs to identify biologically relevant features of interest learned by p-IgGen, and predictably steer its generation. We identify antibody-specific features, such as the complementarity-determining region (CDR) identity and germline gene identity, and use them to steer p-IgGen generation for specific germline gene identities. Overall, this work shows the applicability of SAEs for incorporating rational design principles to antibody library generation, allowing the generation of libraries with desired properties such as enrichment in sequences originating from specific germlines. We show that TopK SAEs can accurately identify interpretable latents underpinning model generation, whereas Ordered SAEs can identify steerable features capable of tuning model generation.

2 Related Work

2.1 Mechanistic Interpretability

Mechanistic interpretability refers to the approach of explaining complex machine learning systems through the behaviour of their functional units [Kästner and Crook, 2024] by decomposing or reverse-engineering systems into their more elementary computations [Rai et al., 2025]. The eventual goal is to discover causal relationships between model inputs and corresponding outputs.

2.2 Sparse Autoencoders

Sparse Autoencoders (SAEs) have specifically been employed in mechanistic interpretability for feature discovery. They tackle the issue of feature superposition resulting in polysemantic neurons, where any given neuron encodes multiple, often unrelated features. SAEs tackle this problem by projecting dense neuron activations into a sparser latent space using a sparse encoder, Equation 1, whilst ensuring the latent representation can be reconstructed back into the original neuron representation by a decoder following sparsification, Equation 2.

$$z = g\big(ReLU\big(W_{\text{enc}}\, x + b_{\text{enc}}\big)\big) \tag{1}$$

$$\hat{x} = W_{\text{dec}}\, z + b_{\text{dec}} \tag{2}$$

where W are the weight matrices and b are the bias vectors, enc and dec denote the encoder and decoder respectively, x is the original hidden representation, z the latent representation, and $\hat{x}$ the reconstructed hidden representation. $ReLU$ activation is applied to the latent representation following encoding and g is a sparsification function.

2.2.1 TopK SAEs

TopK SAEs [Gao et al., 2024] limit the number of active latents to k, where $k \ll d_{\text{in}} \ll d_{\text{sae}}$. d_{in} is the input hidden dimensions, and d_{sae} is the latent or dictionary dimensions. Equation 3 shows the loss computation.

$$L(x) = \underbrace{\|x - \hat{x}\|_2^2}_{\text{Reconstruction loss}} + \underbrace{c}_{\text{Sparsity constraint}} \tag{3}$$

The $L(x)$ reconstruction loss compares the decoded representation $\hat{x}$ with the original hidden representation x. When a sparsification function is not directly applied during encoding, a separate sparsity constraint is added in loss computations, which is usually a variation of an L1 regularisation loss [Zhang et al., 2018].

2.2.2 Ordered SAEs (O-SAEs)

Ordered SAEs follow a nested SAE architecture, enabling hierarchical ordering of SAE latents. Importantly, compared to the traditional TopK SAE architecture which arbitrarily orders hierarchical latents within the dictionary space, O-SAEs enforce a strict, consistent, hierarchical ordering of latents. This is because TopK SAEs enforce sparsity within the entire dictionary space in one go, whereas O-SAEs follow a nested approach and effectively train a number of individual, nested SAEs which occupy an increasing portion of the dictionary space.

O-SAEs introduce two core components: (i) *per-index nested grouping*, and (ii) *strictly decreasing truncation weights* in order to ensure consistent ordering.

(i) For each truncation level $m \in \{1, \ldots, d_{\text{sae}}\}$, the first m rows of the encoder and decoder are isolated:

$$W_{\text{enc}}^{(m)} = [W_{\text{enc}}]_{1:m,:}, \quad W_{\text{dec}}^{(m)} = [W_{\text{dec}}]_{1:m,:} \tag{4}$$

In Eq. (4) the encoder–decoder pair $\left(W_{\text{enc}}^{(m)}, W_{\text{dec}}^{(m)}\right)$ re-uses the first m rows of the full weight matrices. Because every smaller autoencoder is a strict subset of the larger one, any latent $i \leq m$ is shared across all groups that follow. This "per-index nested grouping" forces early latents to model global structure that remains useful for every deeper stage. Per-index grouping ensures non-random sampling of dictionary sizes, unlike in Matryoshka SAEs [Bussmann et al., 2025], increasing the overall consistency of results.

(ii) Each partial reconstruction is weighted by a *monotonically decreasing* probability $p_M(m)$, so that early (low-index) features incur a higher penalty when failing to capture coarse structure. The per-truncation loss is

$$L_m(x) = p_M(m) \left\| x - W_{\text{dec}}^{(m)\top} W_{\text{enc}}^{(m)} x \right\|_2^2 \tag{5}$$

and summing over all m promotes the model to learn the most "abstract" elements first, with progressively finer details later. Combining the decreasing probability weights with nested latents further enforces ordering of identified latents and maintains a stricter hierarchy.

3 Data and Methods

We adapted TopK Sparse Autoencoders (SAEs) from the EleutherAI/sparsify GitHub repository [EleutherAI, 2025], and Ordered SAEs based on the paper by Wang et al. [Wang et al., 2025]. Training parameters were taken directly from the original repositories when available. See Appendix for full details.

We trained both TopK and Ordered SAEs on hidden layer activations of p-IgGen, using VH/VL paired sequences from the original p-IgGen training set [Turnbull et al., 2024], which contained **1,800,545** pairs from the Observed Antibody Space (OAS) database [Olsen et al., 2022, Kovaltsuk et al., 2018]. Paired VH and VL sequences were concatenated with start/end tokens and passed through p-IgGen to obtain hidden activations from all four hidden layers, with **100,000** sequences randomly subsampled for Ordered SAE training. For targeted feature selection, OAS-derived data were split into training, validation and test sets (see Appendix) to study CDR identity (whether a residue lies in a specific CDR or non-CDR region) and V/J gene identity (which germline V or J segment the sequence originated from). For CDR identity, the training matrix consisted of residue-level latent activations with 7 classes (6 CDRs plus non-CDR), while for V/J gene identity, the training matrix was the mean-pooled latent activations across residues for each sequence.

4 Results

4.1 TopK SAE-identified Features are Interpretable, Antibody-specific Concepts, but Not Steerable

4.1.1 TopK SAE latents preserve biological information following sparsification

TopK sparsification represents each token with far fewer latents than hidden neurons, raising the possibility of information loss, so we compared latents and hidden neurons on residue- and sequence-level property prediction tasks. For TopK SAEs trained on final-layer activations (layer 3) [Parsan et al., 2025], logistic regressors trained on CDR identities achieved validation accuracies of **0.99** using latent activations and **0.98** using hidden neuron activations, indicating that residue-level information is preserved. For sequence-level features, germline heavy J gene prediction using latent activations yielded a validation F1 macro score of **0.93** (reported due to class imbalance), with Table 1 showing strong precision, recall, and F1 across IGHJ classes. Together with similar findings for general protein language models [Simon and Zou, 2024, Adams et al., 2025, Parsan et al., 2025], these results suggest that SAE latents preserve key antibody information after sparsification, justifying their use for further interpretability analysis.

Table 1: Precision, recall and F1-score per IGHJ class.

Class	Precision	Recall	F1-score
IGHJ1	0.91	0.71	0.80
IGHJ2	0.94	0.93	0.94
IGHJ3	0.96	0.96	0.96
IGHJ4	0.98	0.99	0.98
IGHJ5	0.94	0.95	0.95
IGHJ6	0.98	0.97	0.98
macro average	0.95	0.92	0.93

4.1.2 TopK latent activations are visually interpretable

To investigate whether SAE latents provide an interpretable alternative to understanding model generation, we compared the activated patterns of latents and neurons correlated to properties of interest. As a baseline, we compare activations of the top correlated latents and neurons for CDRH3. Visual investigation reveals latent activations are sparse and specific to CDRH3 residues, compared to neurons which activate across the sequence without any immediately recognisable pattern (Figure 1a). This may be explained by the polysemanticity of neurons, where multiple features specific to several unrelated residues are represented by the same neuron. When investigating activation patterns, this complicates using neurons as a tool for interpretability and highlights the potential greater explainability of SAE-derived latents.

We further investigated heavy J gene activations as sequence-level concepts. Latents corresponding to heavy J gene identity activated on residues representing the concept, i.e. gene identity. In this instance, the top correlated latents were activated on the J domain of examined antibody sequences. This is interesting considering sequence-level representations are mean pools of the original residue-level representations, leading to an overall loss of positional information. Therefore, the top correlated latents also encode intrinsic positional information (Figure 1b).

To quantify the predictive properties of our identified features, we carried out an activation-threshold analysis (Table S2). Interestingly, there was a strong preference for IGHJ4 features in the final layer, with no features being identified for IGHJ1 and 5. However, we identified features correlated to the two identities in earlier layers [data not shown]. This highlighted a potential flaw in TopK SAEs; SAEs cannot consistently generate monosemantic features for all the concepts represented in the hidden state.

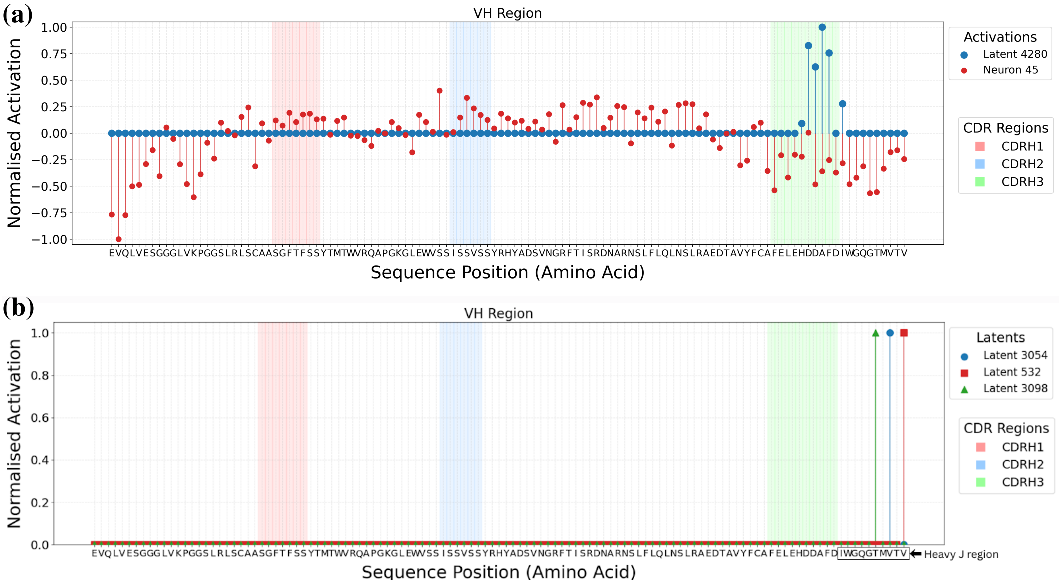

Figure 1: Latent activations (blue) and hidden neuron activations (red) for CDRH3 identity (a), and latent activations for IGHJ3 (b). The x-axis shows the amino-acid sequence of the VH region of a test antibody; the y-axis shows normalised activation. CDRs are coloured CDRH1 (red), CDRH2 (blue), and CDRH3 (green). Latent activations localise to the expected regions—CDRH3 loop and the heavy J region, whereas neuron activations are scattered across the sequence with no discernible pattern.

4.2 Ordered SAEs Identify More Steerable Features Compared to TopK SAEs

In addition to assessing how well latents predict the target concept, we also use feature steering as an indicator of feature importance [Parsan et al., 2025]: if increasing a latent consistently steers generation in a desired direction, the corresponding feature is likely important. We were unable to successfully steer on TopK latents (Figure S2), which may be attributed to known issues within these architectures such as feature splitting and absorption [Chanin et al., 2024]. We subsequently evaluated Ordered SAEs, which build a hierarchical latent space that preserves both high-level and fine-grained features [Bussmann et al., 2025], but result in a less interpretable localisation pattern (Figure S3)

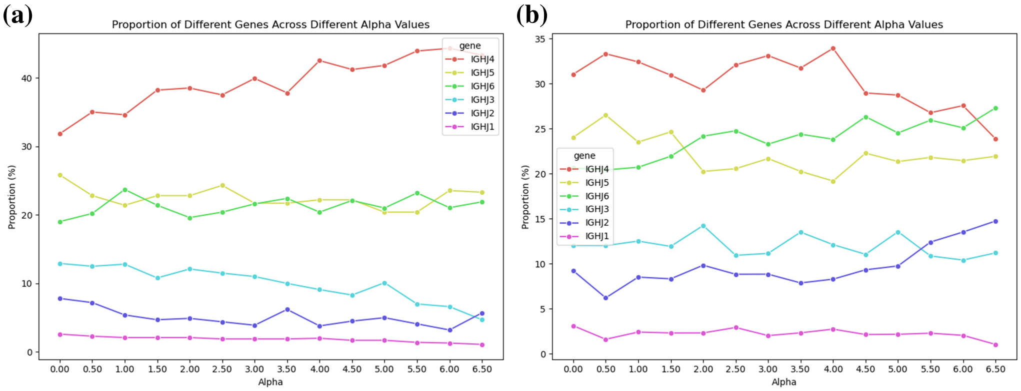

Figure 2: Results of IGHJ4 steering using Ordered latent 12 (a) and 49 (b). Y-axis shows the proportion of generated sequences. Plots are coloured by heavy J gene identity. X-axis shows the steering factor used (alpha). Results are for a library of 1000 p-IgGen-generated sequences. Latent 12—positively correlated with IGHJ4—increases IGHJ4 proportion under positive steering, whereas latent 49—negatively correlated—decreases IGHJ4 under the same steering.

We conducted a linear probe and subsequent activation-threshold analysis to identify features correlated to IGHJ4 in layer 3. Due to the implicit hierarchy in features, we ranked features with an F-score > 0.5 based on their dictionary index, with smaller indices representing higher-level features. We identified latent 12, which was positively correlated with IGHJ4, and latent 49, which was negatively correlated, and steered on these features (Figure 2).

Positively steering on latent 12 increased IGHJ4 proportion in model generation, with a Pearson's R = 0.939, p = 6.458×10^{-7} and Spearman's correlation = 0.921, p = 2.982×10^{-6}. Conversely, positively steering on latent 49 decreased IGHJ4 proportion in model generation, with a Pearson's R of -0.705, p-value = 4.89×10^{-3} and Spearman's correlation = -0.657, p = 0.0106.

5 Conclusions and Future Outlook

Sparse autoencoders offer a practical route to interrogate autoregressive antibody LMs: they surface domain-specific features, but high correlation does not guarantee causal control. In our study, Top-K SAEs often produced interpretable, residue-level signals that were weakly steerable, whereas Ordered SAEs yielded more abstract—and reliably steerable—features at the cost of intuitive localisation. Progress is currently limited by scarce labelled antibody datasets, which hampers systematic feature discovery and validation. For studies on antibody language models, we advocate rigorous steering/ablation benchmarks and scaling SAE training with curated, annotated resources (e.g., FLAb) to learn higher-level, controllable features. Realising this will enable targeted manipulation of properties such as developability and specificity, advancing rational, model-guided antibody library design.

References

Brennan Abanades, Tobias H. Olsen, Matthew I. J. Raybould, Broncio Aguilar-Sanjuan, Wing Ki Wong, Guy Georges, Alexander Bujotzek, and Charlotte M. Deane. The Patent and Literature Antibody Database (PLAbDab): an evolving reference set of functionally diverse, literature-annotated antibody sequences and structures. *Nucleic Acids Research*, 52(D1):D545–D551, January 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1056. URL https://dx.doi.org/10.1093/nar/gkad1056. Publisher: Oxford Academic.

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, February 2025. URL https://www.biorxiv.org/content/10.1101/2025.02.06.636901v1. Pages: 2025.02.06.636901 Section: New Results.

Emanuele Andreano and Rino Rappuoli. Immunodominant antibody germlines in COVID-19. *The Journal of Experimental Medicine*, 218(5):e20210281, March 2021. ISSN 0022-1007. doi: 10.1084/jem.20210281. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7933983/.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning Multi-Level Features with Matryoshka Sparse Autoencoders, March 2025. URL http://arxiv.org/abs/2503.17547. arXiv:2503.17547 [cs].

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Bloom. A is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders, September 2024. URL http://arxiv.org/abs/2409.14507. arXiv:2409.14507 [cs].

Jia-Ying Chen, Jing-Fu Wang, Yue Hu, Xin-Hui Li, Yu-Rong Qian, and Chao-Lin Song. Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review. *Frontiers in Bioengineering and Biotechnology*, 13, January 2025. ISSN 2296-4185. doi: 10.3389/fbioe.2025.1506508. URL https://www.frontiersin.orghttps://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2025.1506508/full. Publisher: Frontiers.

Wei Chen, Xuesong Liu, Sanyin Zhang, and Shilin Chen. Artificial intelligence for drug discovery: Resources, methods, and applications. *Molecular Therapy Nucleic Acids*, 31:691–702, March 2023. ISSN 2162-2531. doi: 10.1016/j.omtn.2023.02.019. URL https://www.cell.com/molecular-therapy-family/nucleic-acids/abstract/S2162-2531(23)00039-2. Publisher: Elsevier.

Lewis Chinery, Jeliazko R. Jeliazkov, and Charlotte M. Deane. Humatch - fast, gene-specific joint humanisation of antibody heavy and light chains, September 2024. URL https://www.biorxiv.org/content/10.1101/2024.09.16.613210v1. Pages: 2024.09.16.613210 Section: New Results.

Mark L. Chiu, Dennis R. Goulet, Alexey Teplyakov, and Gary L. Gilliland. Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies*, 8(4):55, December 2019. ISSN 2073-4468. doi: 10.3390/antib8040055. URL https://www.mdpi.com/2073-4468/8/4/55. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Silvia Crescioli, Kaplon , Hélène, Wang , Lin, Visweswaraiah , Jyothsna, Kapoor , Vaishali, , and Janice M. Reichert. Antibodies to watch in 2025. *mAbs*, 17(1):2443538, December 2025. ISSN 1942-0862. doi: 10.1080/19420862.2024.2443538. URL https://doi.org/10.1080/19420862.2024.2443538. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420862.2024.2443538.

Weiqi Deng, Xuefeng Niu, Ping He, Qihong Yan, Huan Liang, Yongping Wang, Lishan Ning, Zihan Lin, Yudi Zhang, Xinwei Zhao, Liqiang Feng, Linbing Qu, and Ling Chen. An allelic atlas of immunoglobulin heavy chain variable regions reveals antibody binding epitope preference resilient to SARS-CoV-2 mutation escape. *Frontiers in Immunology*, 15:1471396, January 2025. ISSN 1664-3224. doi: 10.3389/fimmu.2024.1471396. URL https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2024.1471396/full. Publisher: Frontiers.

James Dunbar and Charlotte M. Deane. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, January 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv552. URL https://doi.org/10.1093/bioinformatics/btv552.

EleutherAI. Sparsify: Sparsify transformers with saes and transcoders. https://github.com/EleutherAI/sparsify, 2025. Version 1.3.0, accessed 2025-11-30.

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, June 2024. URL http://arxiv.org/abs/2406.04093. arXiv:2406.04093 [cs] version: 1.

Alexander Jarasch, Hans Koll, Joerg T. Regula, Martin Bader, Apollon Papadimitriou, and Hubert Kettenberger. Developability Assessment During the Selection of Novel Therapeutic Antibodies. *Journal of Pharmaceutical Sciences*, 104(6):1885–1898, June 2015. ISSN 0022-3549, 1520-6017. doi: 10.1002/jps.24430. URL https://jpharmsci.org/article/S0022-3549(15)30084-8/abstract. Publisher: Elsevier.

Aleksandr Kovaltsuk, Jinwoo Leem, Sebastian Kelm, James Snowden, Charlotte M Deane, and Konrad Krawczyk. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *The Journal of Immunology*, 201(8):2502–2509, October 2018. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.1800708. URL https://academic.oup.com/jimmunol/article/201/8/2502/7953451.

Lena Kästner and Barnaby Crook. Explaining AI through mechanistic interpretability. *European Journal for Philosophy of Science*, 14(4):52, October 2024. ISSN 1879-4920. doi: 10.1007/s13194-024-00614-4. URL https://doi.org/10.1007/s13194-024-00614-4.

Marie-Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental and Comparative Immunology*, 27(1):55–77, January 2003. ISSN 0145-305X. doi: 10.1016/s0145-305x(02)00039-3.

W. Li, L. Jaroszewski, and A. Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics (Oxford, England)*, 17(3):282–283, March 2001. ISSN 1367-4803. doi: 10.1093/bioinformatics/17.3.282.

Tobias H. Olsen, Fergus Boyles, and Charlotte M. Deane. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022. ISSN 1469-896X. doi: 10.1002/pro.4205. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4205. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4205.

Nithin Parsan, David J. Yang, and John J. Yang. Towards Interpretable Protein Structure Prediction with Sparse Autoencoders, March 2025. URL http://arxiv.org/abs/2503.08764. arXiv:2503.08764 [q-bio].

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models, March 2025. URL http://arxiv.org/abs/2407.02646. arXiv:2407.02646 [cs].

Matthew I. J. Raybould and Charlotte M. Deane. The Therapeutic Antibody ProfilerTherapeutic antibody profiler (TAP) for Computational Developability Assessment. In Gunnar Houen, editor, *Therapeutic Antibodies: Methods and Protocols*, pages 115–125. Springer US, New York, NY, 2022. ISBN 978-1-0716-1450-1. doi: 10.1007/978-1-0716-1450-1_5. URL https://doi.org/10.1007/978-1-0716-1450-1_5.

Matthew I J Raybould, Aleksandr Kovaltsuk, Claire Marks, and Charlotte M Deane. CoV-AbDab: the coronavirus antibody database. *Bioinformatics*, 37(5):734–735, May 2021. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btaa739. URL https://academic.oup.com/bioinformatics/article/37/5/734/5893556.

Dominique Scaviner, Valérie Barbié, Manuel Ruiz, and Marie-Paule Lefranc. Protein Displays of the Human Immunoglobulin Heavy, Kappa and Lambda Variable and Joining Regions. *Experimental and Clinical Immunogenetics*, 16(4):234–240, November 1999. ISSN 0254-9670. doi: 10.1159/000019115. URL `https://doi.org/10.1159/000019115`.

Elana Simon and James Zou. InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders, November 2024. URL `http://arxiv.org/abs/2412.12101`. arXiv:2412.12101 [q-bio].

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html`.

Oliver M Turnbull, Dino Oglic, Rebecca Croasdale-Wood, and Charlotte M Deane. p-IgGen: a paired antibody generative language model. *Bioinformatics*, 40(11):btae659, November 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae659. URL `https://doi.org/10.1093/bioinformatics/btae659`.

Sophie L. Wang, Alex Quach, Nithin Parsan, and John Jingxuan Yang. Enforcing orderedness in saes to improve feature consistency. In *NeurIPS 2025 Workshop on Mechanistic Interpretability*, 2025. URL `https://openreview.net/forum?id=0YOBC01d0Q`. NeurIPS workshop (non-archival).

Li Zhang, Yaping Lu, Bangjun Wang, Fanzhang Li, and Zhao Zhang. Sparse Auto-encoder with Smoothed $$l_1$$Regularization. *Neural Processing Letters*, 47(3):829–839, June 2018. ISSN 1573-773X. doi: 10.1007/s11063-017-9668-5. URL `https://doi.org/10.1007/s11063-017-9668-5`.

A Methods

A.1 Sparse Autoencoder Training

A.1.1 TopK SAE

p-IgGen input dimensions $d_{\text{in}} = 768$ were projected onto a higher-dimensional latent/dictionary size d_{sae}, where $d_{\text{sae}} = d_{\text{in}} \times r = 768 \times 32 = 24{,}576$. $r = 32$ is the expansion factor. ReLU activation was applied to the projection, $z = \mathrm{ReLU}\big(W_{\text{enc}}\, x + b_{\text{enc}}\big)$, followed by a Top-$k$ sparsification with $k = 32$, retaining only the top 32 activations by magnitude. The resulting dictionary size was **24,576**. Decoder weights were initialised as the unit-normalised transpose of the encoder weights to stabilise training. Training used a batch size of 8 and Adam optimisers throughout, with a custom learning rate $\eta = \frac{2\times10^{-4}}{\sqrt{d_{\text{sae}}/16{,}384}}$.

A.1.2 Ordered SAE

Ordered Sparse Autoencoders (O-SAEs) were adopted to retain higher-level, abstract features within our latent space and hierarchically arrange the latents. In our setup, we used expansion factor $r = 8$, yielding a dictionary size $d_{\text{sae}} = d_{\text{in}} \times r = 768 \times 8 = 6{,}144$. Sparsity was again set to $k = 32$, ensuring the top 32 latents are used during reconstruction. All models were trained with Adam optimisers at a fixed learning rate $\eta = 1 \times 10^{-4}$. We chose a smaller maximum dictionary size for the O-SAEs to speed up training, effectively reducing the total number of nested SAEs being trained. Due to per-index grouping, O-SAEs need to train several nested SAEs based on the total dictionary size, whereas the regular TopK architecture only trains a single model.

A.2 Targeted Feature Identification using SAEs

A.2.1 Training data

Paired antibody sequence data were obtained from OAS, Coronavirus Antibody Database (CoV-AbDab) [Raybould et al., 2021], and the Patent and Literature Antibody Database (PLAbDab) [Abanades et al., 2024]. A total of 149,069 sequences were obtained from the respective datasets, based on their binding specificities to SARS-CoV2 RBD (binder and non-binder).

The data was clustered based on CDR sequence similarity using CD-HIT [Li et al., 2001], with a 0.8 similarity threshold on the total CDR sequence. The clusters were then randomly split into the training-validation-test set, whilst ensuring members of the same cluster were in only one of the three possible splits. The splits were further stratified based on binding specificity to SARS-CoV2 RBD.

This specific dataset was originally prepared for a separate project, and the SARS CoV2 RBD binding properties of the antibodies are not relevant for this study. Qualitatively, a dataset of equivalent size randomly sampled from OAS should produce the same results.

The following concepts were studied to identify associated latents: CDR identity, which refers to whether a given residue lies within a specific CDR region, and V/J gene identity, which refers to the germline V or J gene segment that was used to code for the final antibody sequence. For the CDR-identity, the training matrix was the latent activations for each residue. The CDR identity dataset had 7 classes (6 CDR identities and non-CDR regions). For sequence-level concepts, V/J gene identity, the training matrix was the mean pool of the residue-level latent activations in a given sequence.

A.2.2 Linear Probe

We trained a logistic regressor to act as a linear probe on the training-validation data. A logistic regressor (LR) was trained, employing 3-fold cross-validation grid search to optimise hyperparameter C. In logistic regression, C is the inverse of the regularisation strength: larger C applies less regularisation and can overfit, while smaller C applies more regularisation and can improve generalisation. Cross-validation was done during training by randomly shuffling and splitting the training data into 3 cross-validation sets. Correlation weights of all latents were stored and the latents with the top 500 positive correlation weights were used for further validation.

A.2.3 Latent Selection

The top correlated latents were further validated on the validation set. Based on the strategy by Simon and Zhou [Simon and Zou, 2024], the latent activations across the validation set were normalised using MinMax scaling; for each normalised latent, binary latent-on/latent-off labels using activation thresholds of 0.1, 0.2, 0.5, 0.8, 0.9 were applied. For each latent-concept pair, a latent was defined as an interpretable feature if its F1 score for any of the tested thresholds was greater than 0.5. At this boundary the harmonic mean guarantees both precision and recall are at least 0.5, ruling out latents that are either mostly false positives or that miss the majority of true activations.

A.2.4 Antibody sequence alignment

Antibody sequences were aligned using ANARCI [Dunbar and Deane, 2016] and the IMGT numbering [Lefranc et al., 2003].

A.3 Steering

Steering was implemented based on the strategy by Templeton et al. [Templeton et al., 2024]. Each latent can be represented by its corresponding decoder vector $d(i) = W_{\text{dec}}[i, :]$, where $d(i)$ is the decoder vector for latent i and W_{dec} is the decoder weight matrix. Steering is performed by scaling the decoder vector and adding it to the original hidden state (Equation 6).

$$h_l * \;\leftarrow\; h_l \;+\; \alpha \cdot d(i) \tag{6}$$

Here, α (Alpha) is the steering factor and h_l is the hidden state before the intervention and h_l* is the hidden state following the intervention.

A.4 Case Study of Heavy J Gene Identity for IGHJ4

The functional importance of SAEs for studying LLMs lies in their ability to interrogate specific, domain-relevant concepts rather than an undefined set of all possible ones. Here, we examine gene identity, as it directly influences antibody binding affinity and specificity [Deng et al., 2025]. Specifically, we pick IGHJ4 genes for our analysis, due to their widespread clinical significance underpinned by the fact that they are the most widely utilised J genes in our immune repertoire. That is, the majority of heavy chain antibodies within any given individual originate from the IGHJ4 germline gene. For our analysis, we decided to focus on the final layer (Table S1).

Table S1: IGHJ4 feature statistics. Latents which can be used as a binary predictor of gene identity with an F-score greater than 0.5, based on threshold-activation patterns are termed as 'features'. For each layer in p-IgGen, features corresponding to IGHJ4 were identified. The maximum F-score of all the identified features from a given layer are reported as Max F-score.

Heavy J gene	Layer	Features ($F_1 > 0.5$)	Max F_1-score
IGHJ4	Layer 0	197	0.930
	Layer 1	266	0.949
	Layer 2	189	0.930
	Layer 3	85	0.752

First, we looked at the absolute positional activations of this latent across all the sequences in our validation set which had an IGHJ4 heavy J gene, and compared it to activations on specific IMGT positions. Whilst activations on absolute positions were distributed near the end of the heavy chain corresponding to the J region, the activations on IMGT positions were more consistent and concentrated. This implies that the model does not base its activation pattern on the absolute sequence length alone, but rather the underlying sequence alignment (Figure S1).

We then chose to investigate the specific residue identities on which the latents were activated. Based on the heavy J gene sequence alignments, the top two latents activated at IMGT positions 120 and 119, which are a Q and G, respectively. These are conserved across all human IGHJ genes. The third

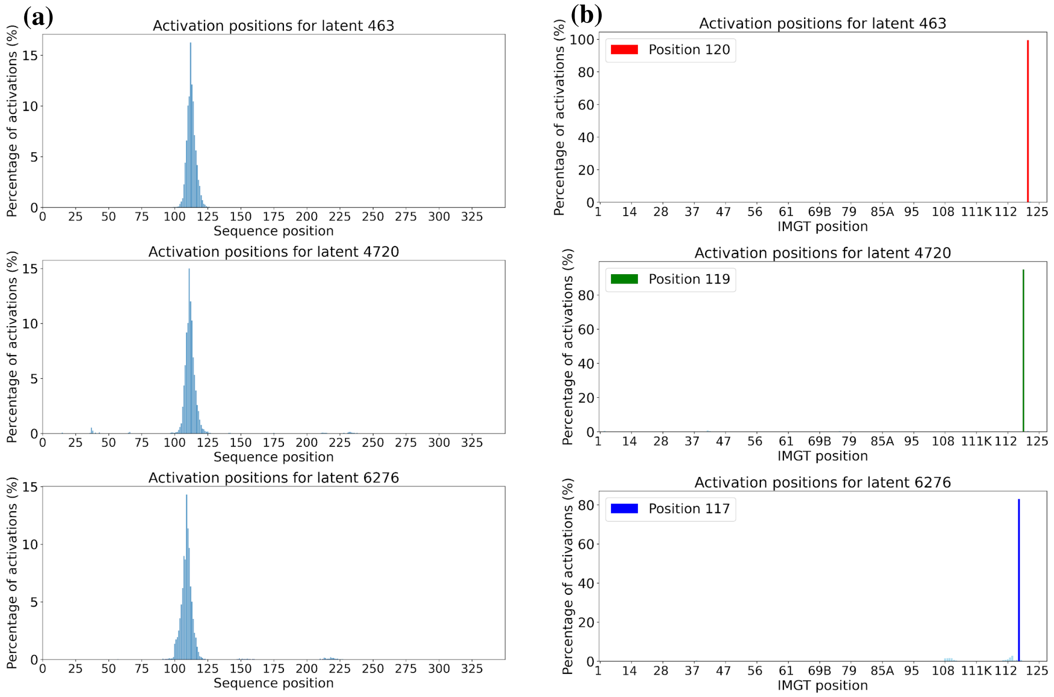

Figure S1: Comparison of absolute positional (a) and IMGT (b) activations of top three IGHJ4 latents. The sequence/IMGT positions are shown on the x-axis. For the sequence positions, the amino acid sequences were end-padded to a constant length of 350. Percentage of total activations on any given position across validation IGHJ4 sequences is shown on the y-axis. The most frequent IMGT position for activation is highlighted for each latent. Latent activations show a distribution near the end of the heavy chain when aligned based on absolute sequence position. In contrast, latents demonstrate discrete activations when aligned based on IMGT numbering.

top latent activated on Y at position 117, which is unique for IGHJ4 [Scaviner et al., 1999]. These results indicated that top latents encoded contextual information of the preceding residues.

Previous studies have highlighted how highly correlated features may be used to steer model outputs [Templeton et al., 2024, Simon and Zou, 2024]. We attempted to steer on each identified feature to investigate how it affects model generation. We positively steered on each latent, which we hypothesised should increase the proportion of IGHJ4 in generated sequences. However, steering on these latents was unpredictable and did not consistently increase IGHJ4 proportions (Figure S2).

To check if this phenomenon was somehow exclusive for IGHJ4 and layer 3, we attempted to steer across all the layers for a number of different features for various gene identities, but were unable to predictably steer model generation [data not shown]. The lack of steerability may indicate how these features individually do not contribute to the gene identity, making them informative features when used for downstream predictions, but not for biasing model output.

This may be due to feature splitting [Chanin et al., 2024] which has been reported for TopK SAEs. Feature splitting refers to the phenomenon where higher-order features are broken down into specific contextual examples. In the case of text-based language models, ’math’ may be split into ’algebra’ and ’geometry’. These phenomena arise when enforcing sparsity in a dictionary consisting of hierarchical features. In this instance, if the identified latents correspond to only single residues within the J-domain, it essentially becomes a residue-level feature as opposed to a sequence-level feature. If the feature activates on a residue specific to the gene identity, it may be a good predictor for the gene identity, but not a steerable feature. This points to the possibility that several features together confer J gene identity, and that these features are likely correlated to each other. Hence, activating one but not the others does not necessarily result in a predictable shift in model performance.

The case study on IGHJ4 indicated that identified features retain biologically relevant context information. Most highly correlated features (based on LR correlation weight and f-score) tend to be

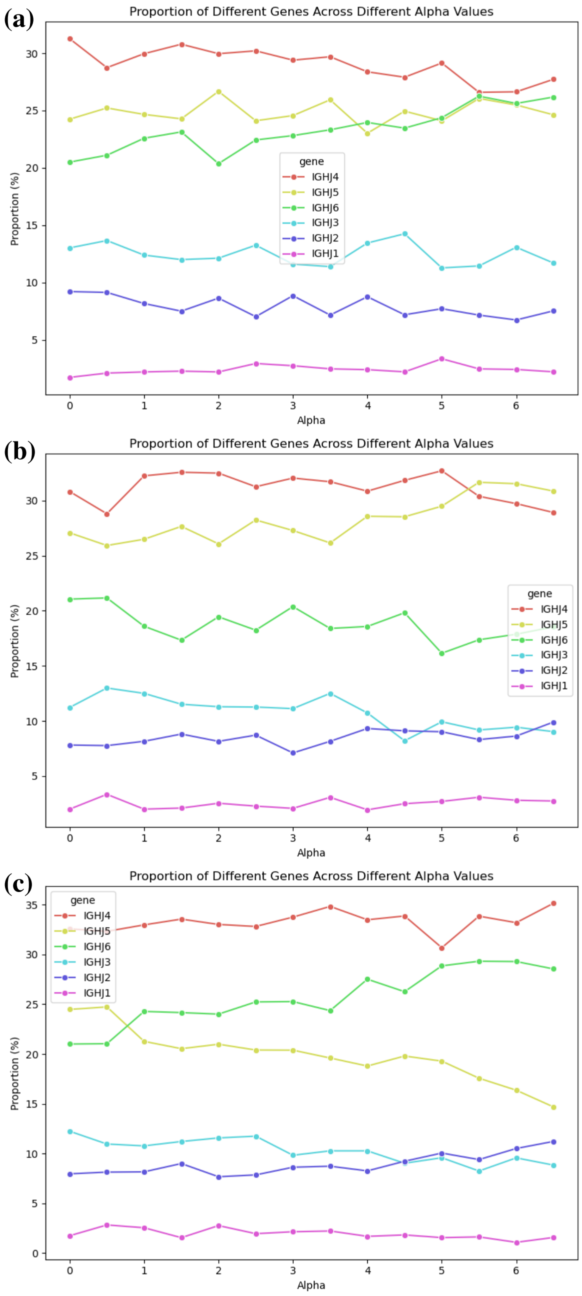

Figure S2: Results of IGHJ4 feature steering for latent 463 (a), 4720 (b), 6276 (c). Y-axis shows the proportion of generated sequences. Plots are coloured by heavy J gene identity. X-axis shows the steering factor used (alpha). Results are for a library of 1000 p-IgGen-generated sequences. For each latent tested (a-c), steering did not result in a predictable change in library composition.

residue-specific. Targeted approaches such as this cannot easily find abstract, higher-order features, assuming they are represented within the latent space to begin with. Concept-specific targeted feature identification might identify highly correlated features that are biologically informative. For instance, two of the three top features (463 and 4720) activate on conserved residues preceded by sequence motifs specific to the gene identity. The third, Latent 6276 activated on an IGHJ4-specific residue, which may explain why this feature can be used to accurately identify IGHJ4.

Highly predictive features may be correlated with other biologically informative features. To understand whether highly predictive features influence model behaviour, we tried to steer along these features to increase the proportion of IGHJ4 in generated sequences. This did not produce predictable results, making it difficult to interpret the contribution of each individual latent to model generation. Overall, TopK SAEs can identify features in targeted concept analysis which are intuitively interpretable, however, not necessarily steerable.

Table S2: Feature counts and maximum F-scores for IGHJ genes (layer 3)

Gene	Number of features	Max F-score
IGHJ1	0	0.366
IGHJ2	5	0.521
IGHJ3	12	0.866
IGHJ4	85	0.752
IGHJ5	0	0.486
IGHJ6	17	0.866

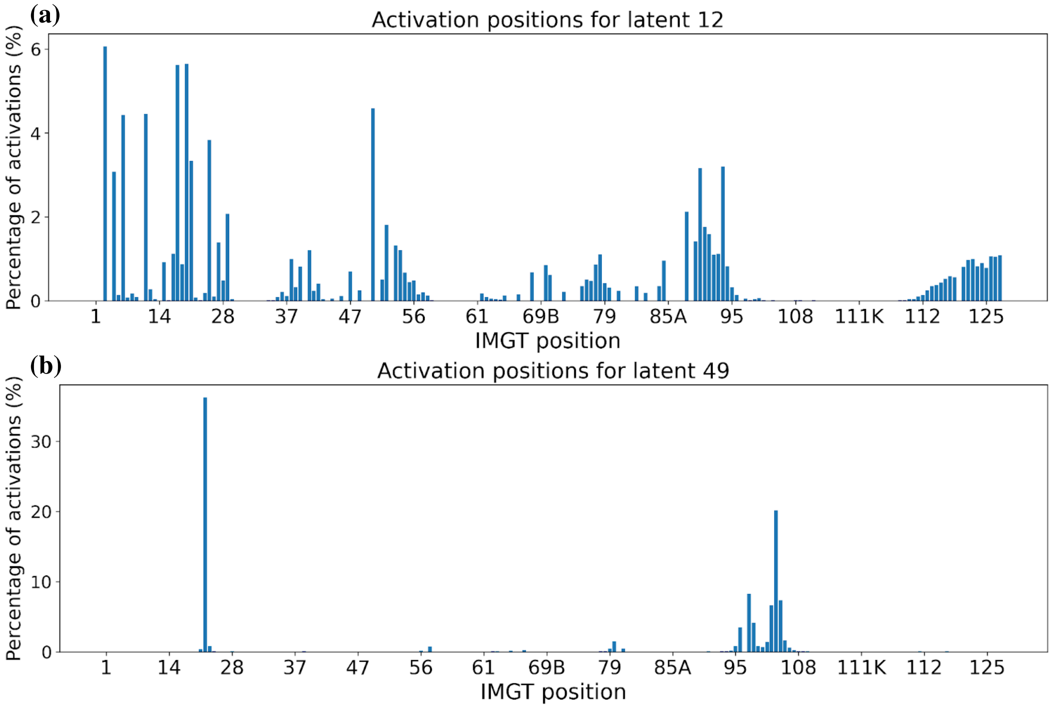

Figure S3: IMGT activations of latent 12 (a) and 49 (b). Activation patterns of both latents show scattered distribution across the range of IMGT positions.