TOWARDS OPEN-ENDED VISUAL SCIENTIFIC DISCOVERY WITH SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models in several scientific domains, including visual domains, learn representations that capture complex semantics from their respective fields. Despite this, most existing applications focus on pre-specified concepts or targets, excelling at confirmation but not open-ended discovery of unknown patterns. We study whether sparse autoencoders (SAEs) with heuristic feature ranking can provide a practical methodology for surfacing candidate visual features from vision foundation models for scientific discovery without concept-specific supervision. We evaluate this approach in three stages: (1) general-purpose concept rediscovery, (2) domain-specific concept rediscovery, and (3) question-driven feature ranking. On ADE20K, SAEs outperform baseline methods (k-means, PCA, SemiNMF) across most downstream metrics for concept rediscovery. On FishVista, the same approach recovers fine-grained fish anatomical features without concept supervision. Finally, on *Heliconius* butterflies, a heuristic based on the Mann-Whitney U statistic consistently surfaces features corresponding to real diagnostic keys between mimic subspecies. Together, these results show that SAEs with heuristic ranking can efficiently learn and surface semantic structure in vision foundation model representations. We view this as a methodological prerequisite for open-ended visual scientific discovery: the present experiments validate feature surfacing and ranking through rediscovery, while prospective discovery of previously unknown scientific phenomena remains future work.

1 INTRODUCTION

Scientific data now span hundreds of petabytes across domains, far outpacing what any lab can annotate or manually inspect.[1] Large neural networks trained on these datasets cover proteins (Abramson et al., 2024), weather (Lam et al., 2023; Price et al., 2025), genomics (Cui et al., 2024; Hayes et al., 2025), geospatial (Reed et al., 2023), and ecology (Gu et al., 2025). These models achieve strong predictive performance across a variety of tasks by learning compressed representations of the underlying data distribution's structure.

However, foundation models are currently limited to *confirmatory* applications where the target concepts are pre-specified (see Figure 2). Foundation models are used as frozen feature extractors or fine-tuned for specific predictions (species classification (Van Horn et al., 2021), cellular responses to perturbations (Roohani et al., 2025), 3-D protein structure (Haas et al., 2018; Pereira et al., 2021)). Because of this, the representations produced by foundation models are high-dimensional embeddings optimized for task performance. While these representations enable accurate predictions, they do not directly provide reusable scientific measurements or expose the structure the model learned for systematic investigation.

Discovery vs. Confirmation Scientific discovery often involves establishing names and descriptions of specific, previously unexplained phenomena (Kuhn & Hacking, 1970; Poincaré, 2003). A morphological feature that distinguishes subspecies but is not formally cataloged, an unexplored

[1] For example, the NIH Sequence Read Archive exposes roughly 47 PB of public sequencing data, the European Nucleotide Archive provides roughly 63 PB of immediately downloadable data, NASA Earthdata (EOSDIS) hosts more than 120 PB of open Earth science data, and a single reanalysis, ERA5, is on the order of 5 PB.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

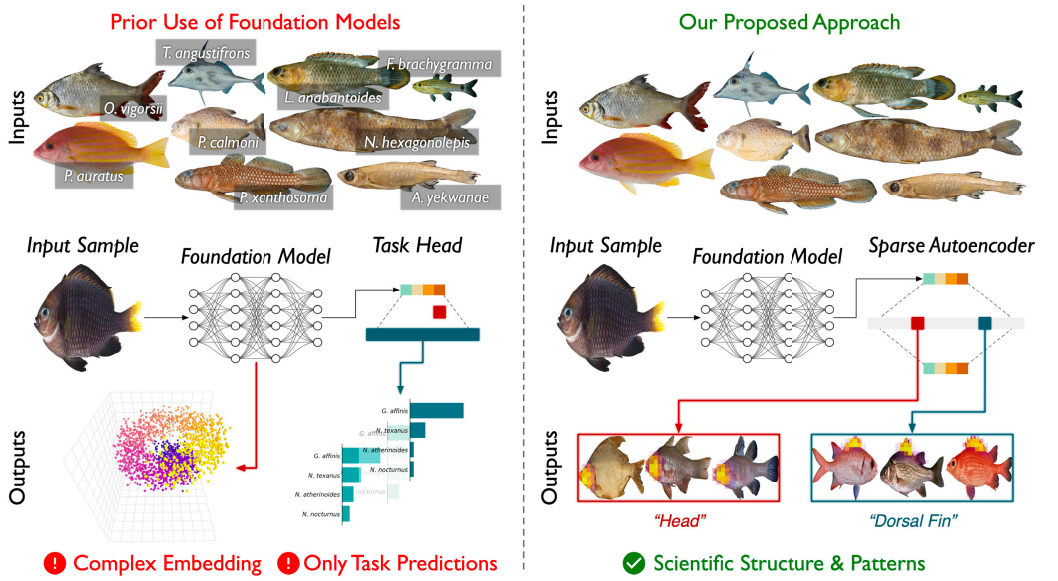


Figure 1: *Validating an instrument for open-ended feature discovery.* **Left:** Typical use of a foundation model in science: an input image is passed through a pretrained encoder and a task-specific head to yield class scores; the dense representation remains an opaque embedding, so unnamed factors are inaccessible. **Right:** Our procedure composes a foundation model with a sparse autoencoder, producing a library of interpretable semantic concepts with per-example activation maps. We find that these concepts align with and localize anatomical parts (e.g., head, dorsal fin) without seeing part-of-body labels. The learned concepts are inherently quantitative, making them easy to rank and search using heuristics.

correlation between climate variables, or a sequence motif in an understudied protein family: these cannot be found by methods that require specifying target concepts in advance. Here, we focus on "open-ended feature discovery": surfacing candidate visual features that are relevant to a specific scientific question, without specifying relevant concepts beforehand (Hughes et al., 2024). Because genuinely novel scientific findings cannot be evaluated against ground truth, we evaluate this capability through controlled rediscovery of known concepts and traits.

Interpretability methods offer a route to systematically surface the semantic structure learned by foundation models. While many existing interpretability methods excel at *confirmation* (Alain & Bengio, 2016; Kim et al., 2018; Selvaraju et al., 2017; Sundararajan et al., 2017; Simonyan et al., 2013), they cannot perform open-ended exploration to generate hypotheses (see Figure 2). In particular, methods for such open-ended discovery from *vision* foundation model representations remain underdeveloped and untested.

Our approach. We propose sparse autoencoders (SAEs) with heuristic ranking as a step towards open-ended visual feature discovery. SAEs learn an overcomplete dictionary that reconstructs model activations through sparse linear combinations, with sparsity encouraging individual features to capture semantically coherent patterns (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024; Gao et al., 2025). Unlike supervised methods, SAEs require no labels during training. Each learned feature provides a per-example activation score together with its decoding direction and exemplar evidence, enabling systematic investigation of what patterns the model represents. We then propose question-driven heuristics to rank and quickly surface relevant and possibly novel visual features. In short, we treat SAEs with heuristic feature ranking as a potential tool for open-ended visual feature discovery from foundation models and seek to rigorously evaluate that tool in controlled settings (Peng et al., 2025).

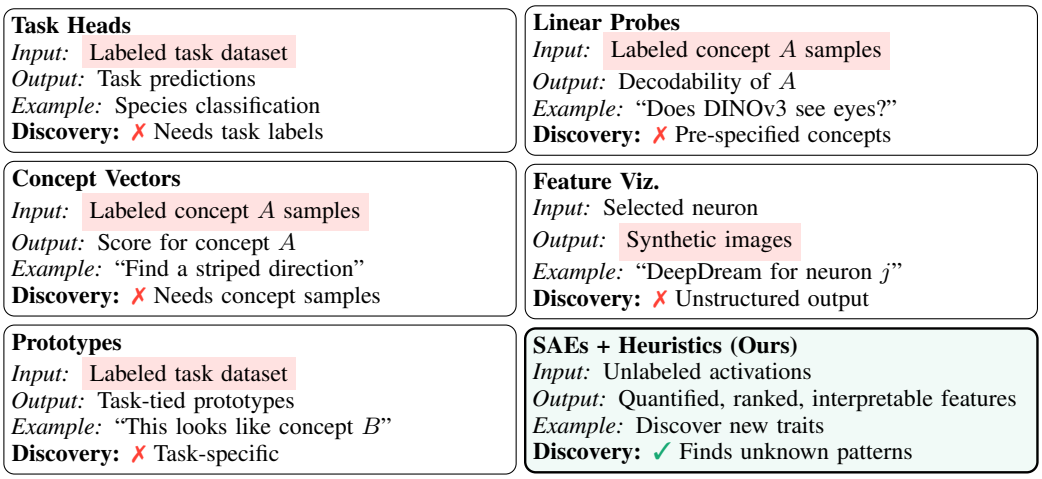


Figure 2: Comparison of methods for extracting knowledge from vision foundation models. Most approaches require researchers to specify concepts or tasks in advance.

Evidence. We evaluate this approach across three experimental settings to determine its strengths and limitations[2]:

- General-purpose rediscovery of segmentation concepts in ADE20K (Zhou et al., 2017), comparing SAEs with other feature learning baselines across vision foundation models (Siméoni et al., 2025; Bolya et al., 2025),
- Domain-specific rediscovery of segmentation concepts in FishVista (Mehrab et al., 2024a;b), comparing SAE variations and per-concept performance,
- Question-specific feature ranking in *Heliconius* butterflies (Lawrence et al., 2024), testing feature ranking for efficient expert validation.

On ADE20K, we find that SAEs outperform baseline methods (k-means clustering, PCA, and SemiNMF) in concept alignment (Section 4.1), rediscovering known segmentation concepts without supervision. On FishVista, we find that SAEs learn fine-grained anatomical features, providing measurable quantities that can be evaluated against biological annotations. Finally, we find that a modified AUC-ROC score highly ranks known diagnostic traits across several mimic subspecies pairs, relying only on species labels (Section 4.3). Our results indicate that SAEs with heuristic ranking can learn and effectively surface semantic structure from vision foundation model representations. We view this work as a prerequisite step to visual scientific discovery: establishing a methodology that can systematically recover and rank known structure from images, making search for truly novel scientific phenomena feasible.

2 Related Work

Foundation Models. Domain-specific foundation models capture rich structure and achieve state-of-the-art performance on their focal tasks, as shown in proteins (Abramson et al., 2024; Hayes et al., 2025), ecology (Stevens et al., 2024), weather (Lam et al., 2023; Price et al., 2025), and single-cell genomics (Cui et al., 2024). In both general and scientific domains, large-scale training leads to emergent abilities beyond the pre-training objective (Wei et al., 2022; Gu et al., 2025). Theory and evidence suggest that large pretrained models internalize meaningful structure (Ha & Schmidhuber, 2018; Caron et al., 2021; LeCun, 2022; Huh et al., 2024). Despite this, most scientific models are benchmarked as **task-specific predictors** rather than sources of novel concepts.

Interpretability Methods. A large body of work explains predictions without exposing a global factorization of representations. Saliency and attribution methods (e.g., gradients, Integrated Gradients, Grad-CAM) highlight input regions for *individual* decisions and have known validity pitfalls

[2]Full implementation can be found at `https://anonymous.4open.science/r/saev-25A1`

(Simonyan et al., 2013; Sundararajan et al., 2017; Selvaraju et al., 2017; Adebayo et al., 2018). Linear probes test separability of *pre-specified* targets (Alain & Bengio, 2016; Hewitt & Liang, 2019); concept methods such as TCAV/ACE require exemplar sets and so test investigator-defined ideas (Kim et al., 2018; Ghorbani et al., 2019). Prototype models (ProtoPNet, ProtoTree, ProtoPFormer) provide case-based, class-tied explanations but are trained end-to-end for specific tasks (Chen et al., 2019; Nauta et al., 2021; Xue et al., 2022). Activation maximization and "feature visualization" (DeepDream-style) optimize inputs for chosen units to aid intuition, but produce synthetic, non-natural images (Mahendran & Vedaldi, 2016; Nguyen et al., 2016; Olah et al., 2017). These approaches excel at explaining predictions or testing for known concepts (confirmation) but cannot be used to discover patterns without prior specification (Figure 2).

Sparse Autoencoders & Decomposition. Classical dictionary learning and sparse coding demonstrated that sparse, parts-based factors can emerge from natural data (Olshausen & Field, 1996; Aharon et al., 2006; Lee & Seung, 1999). Recent work applies sparse autoencoders (SAEs) to model activations, showing that sparse decompositions can partially "unmix" superposed features and yield monosemantic directions (Elhage et al., 2022; Bricken et al., 2023; Gao et al., 2025; Joseph et al., 2025) in language and vision models. SAEs address the superposition hypothesis (Elhage et al., 2022) by assigning features to directions in activation space rather than individual neurons. (Gao et al., 2025) demonstrate that k-sparse autoencoders scale effectively and improve reconstruction–sparsity tradeoffs. Furthermore, concurrent work argues for using SAEs for discovering unknown concepts rather than acting on known ones (Peng et al., 2025) and applies dictionary learning to microscopy foundation models for biological concept extraction (Donhauser et al., 2024). Beyond SAEs, concept factorization methods such as CRAFT, ACE, and related work use NMF-like decompositions to extract human-interpretable concepts and quantify their importance (Fel et al., 2023b;a; Ghorbani et al., 2019). Finally, recent analyses highlight stability and identifiability limits of SAE dictionaries across runs (Leask et al., 2025), motivating careful evaluation protocols.

Discovery. Closest in intent to our goal are (i) work advocating SAEs for uncovering *unknown* concepts rather than acting on *known* ones (Peng et al., 2025), (ii) dictionary-learning approaches that extract biological concepts from microscopy or epigenetics foundation models (Donhauser et al., 2024; Wang et al., 2026), and (iii) automated concept discovery pipelines that discover concepts and then name or validate them with limited supervision (Rao et al., 2024). We differ by offering a comprehensive analysis of SAEs for visual feature discovery, validating both on a standard semantic benchmark and an ecology case study and demonstrating how to use heuristics to enable discovery in practice. We also differ from unsupervised dense segmentation methods (LOST, TokenCut, STEGO, MaskCut) by returning global features with per-patch activations rather than per-image masks or clusters (Siméoni et al., 2021; Wang et al., 2023b; Hamilton et al., 2022; Wang et al., 2023a).

3 METHODOLOGY

Foundation Model. We use the publicly available DINOv3 ViT-L/16 checkpoint (Siméoni et al., 2025), additionally using Perception Encoder ViT-L/14 336 (Bolya et al., 2025) in ablations (Section 4.1). Images are resized with an aspect-ratio-aware transform that enforces exactly 256 patches per image (a $h \times w$ patch grid with $hw = 256$ chosen to minimize aspect-ratio distortion), instead of forcing square 256×256 inputs. We then extract patch-token activations ([CLS] and [REG] discarded).

Sparse Autoencoders (SAEs). Our dictionary learner is a ReLU SAE (Bricken et al., 2023; Templeton et al., 2024); given a d-dimensional activation vector $\mathbf{x} \in \mathbb{R}^d$ from an intermediate layer l of a transformer, an SAE maps $\mathbf{x}$ to a sparse representation $f(\mathbf{x})$ (Equations (1) and (2)) and reconstructs the original input (Equation (3)):

$$\mathbf{h} = W_{\text{enc}}(\mathbf{x} - b_{\text{dec}}) + b_{\text{enc}} \tag{1}$$

$$f(\mathbf{x}) = \text{ReLU}(\mathbf{h}) \tag{2}$$

$$\hat{\mathbf{x}} = W_{\text{dec}} f(\mathbf{x}) + b_{\text{dec}} \tag{3}$$

where $W_{\text{enc}} \in \mathbb{R}^{n \times d}$, $b_{\text{enc}} \in \mathbb{R}^n$, $W_{\text{dec}} \in \mathbb{R}^{d \times n}$ and $b_{\text{dec}} \in \mathbb{R}^d$. The training objective minimizes reconstruction error while encouraging sparsity:

$$\mathcal{L}(\theta) = ||\mathbf{x} - \hat{\mathbf{x}}||_2^2 + \lambda \mathcal{S}(f(\mathbf{x})) \tag{4}$$

where λ controls the sparsity penalty and $\mathcal{S}$ measures sparsity (L_1 norm for training, L_0 for model selection).

We train 16 384-latent SAEs for 100M patches from a single ViT layer and sweep a logarithmic grid of sparsity coefficients λ and learning rates. We summarize each trained model by its reconstruction quality (normalized MSE) and sparsity (L_0). Normalized MSE (NMSE) is MSE normalized by a mean baseline MSE:

$$\text{NMSE} = \frac{\sum_i ||\mathbf{x_i} - \hat{\mathbf{x}}_\mathbf{i}||_2^2}{\sum_i ||\mathbf{x_i} - \bar{\mathbf{x}}||_2^2} \tag{5}$$

where $\hat{\mathbf{x}}_\mathbf{i}$ is the predicted reconstruction and $\bar{\mathbf{x}}$ is the mean of all $\mathbf{x_i}$.

SAE variants. We also use TopK (Gao et al., 2025) and Matryoshka SAEs (Bussmann et al., 2025) and investigate their impact on concept rediscovery.

Matryoshka SAEs (Bussmann et al., 2025) address feature splitting (Bricken et al., 2023; Chanin et al., 2024; Leask et al., 2025) by training multiple nested dictionaries simultaneously. Each training step samples random prefixes of SAE latents and reconstructs the input using each prefix. We sample ten prefixes $\mathcal{M} = m_1, m_2, \ldots, m_{10}$ with $m_i \in \{1, ..., 16384\}$; the sum of prefix reconstruction losses replaces MSE in the objective:

$$\mathcal{L}(\theta) = \sum_{m \in \mathcal{M}} ||\mathbf{x} - \hat{\mathbf{x}}_{0:m}||_2^2 + \lambda \mathcal{S}(f(\mathbf{x})) \tag{6}$$

where $\hat{\mathbf{x}}_{0:m}$ is the reconstruction using only the first m latents. This encourages early latents to capture more general features, implicitly ordering them by decreasing reconstruction importance (analogous to PCA). At inference, prefixes are not sampled, making Matryoshka SAEs function identically to vanilla SAEs.

TopK SAEs (Gao et al., 2025) replace the ReLU activation and sparsity objective $\mathcal{S}$ with a TopK selection over latents, retaining only the k largest activations per input and zeroing the rest. This enforces a fixed sparsity level and improves feature selectivity relative to ReLU at comparable reconstruction error.

Ranking Heuristic. When applying SAEs for visual feature discovery in *Heliconius* butterflies, we focus on finding diagnostic traits between two species. For each SAE feature, we first reduce its activation map for a given image to a single scalar by taking the maximum activation value across all spatial locations (patches). We then score each SAE feature for species separation using a rank-based separability statistic built from the Mann-Whitney U-statistic (Mann & Whitney, 1947; Mason & Graham, 2002), without training a classifier. We call this metric *selectivity*: a value of 1 denotes perfect separation between subspecies, and 0 denotes chance-level separation.

$$\text{sel.} = |\frac{U}{n_1 n_2} - 0.5| * 2 \tag{7}$$

where U is the Mann-Whitney U-statistic calculated using the reduced SAE feature values and n_1, n_2 are the counts for each respective species. See Appendix C for a full definition.

4 Experiments

We evaluate the proposed method for open-ended visual discovery quality through *controlled rediscovery* of ground-truth concepts withheld from SAE training. Strong alignment on known concepts indicates that the same procedure is a promising method for surfacing meaningful structure where ground truth is unavailable.

Discovery is aided by monosemantic features: each latent should align with a single coherent concept. We measure concept alignment with 1-D logistic probes, the standard evaluation in the SAE literature (Gao et al., 2025). We report (1) probe R: loss $\mathcal{L}$ normalized by a bias-only baseline $\mathcal{L}_\pi$: $1 - (\mathcal{L}/\mathcal{L}_\pi)$, (2) mean Average Precision (mAP) across all classes, (3) Purity@k: precision of the top-k activated patches for each feature, and (4) Coverage@τ: the fraction of classes whose best-aligned feature achieves AP $\geq \tau$. We select the SAE with the best probe loss on the training split and report validation metrics. For exact definition of all metrics, see Appendix C.



(a) k-means (b) PCA (c) SemiNMF (d) SAE

Figure 3: Top images for lowest-loss probes for "person" (left) and "toilet" (right) classes for each method. **(a):** k-means does not recover the "person" class; the best "person" cluster fires on roads. **(b):** PCA learns a "person" component, but does not consistently activate on the entire person. Furthermore, PCA does not recover the "toilet" class; the best "toilet" doesn't have an obvious semantic concept. **(c):** Semi-nonnegative matrix factorization (SemiNMF) is a strong baseline and learns both "person" and "toilet" concepts. **(d):** SAEs reliably recover both "person" and "toilet" and activate on the entire object. **Takeaway**: While k-means and PCA report strong dictionary learning metrics, concepts recovered by SAEs are both more visually consistent and salient.

Table 1: SAEs vs. unsupervised baselines for rediscovering ADE20K's semantic segmentation classes from DINOv3 ViT-L/16. For each method, we choose the best layer from DINOv3 based on probe loss (Equation (10)) on ADE20K's training split and report ADE20K validation metrics. Probe R is normalized probe loss, as described in Section 4. Purity@k is precision over the top-k highest-scoring patches for a feature, where $k = 16$. Coverage@τ is the fraction of ADE20K classes with best AP $\geq \tau$, where $\tau = 0.3$. **Takeaway:** Even a vanilla ReLU SAE improves downstream concept rediscovery over traditional unsupervised methods, despite comparable reconstruction metrics.

	Source Recon. ↓		Target Recon. ↓		Downstream ↑			
Method	NMSE	L_0	NMSE	L_0	Probe R	mAP	Pur@k	Cov@τ
k-means	0.508	**1**	0.618	**1**	0.001	0.031	**0.795**	0.026
PCA	0.288	256	0.352	256	0.259	0.093	0.612	0.079
SemiNMF	0.298	256	0.363	256	0.209	0.088	0.696	0.079
SAE	**0.166**	379.2	**0.211**	379.2	**0.441**	**0.224**	0.748	**0.264**

4.1 General semantic rediscovery

We use ADE20K Zhou et al. (2017) as a controlled testbed for rediscovery: can label-free SAE features align with segmentation classes? SAEs are trained on ImageNet-1K activations only; ADE20K labels enter solely at evaluation, isolating extraction from supervision. Concretely, we extract activations on ImageNet-1K (train/val) and ADE20K (train/val) from the final layer of DINOv3 ViT-L/16.

Baseline Methods. We compare SAEs to standard unsupervised baselines: k-means, principal component analysis (PCA), and semi-nonnegative matrix factorization (SemiNMF; Ding et al., 2008). We fit all baselines on ImageNet-1K training activations. For k-means, we use k=16 384 clusters and use the nearest cluster as the predicted reconstruction with a fixed L_0=1. We fit PCA with n components $n \in \{1, 4, 16, 64, 256, 1024\}$. For SemiNMF, we use a mini-batch semi-NMF estimator with nonnegative codes and an unconstrained dictionary; codes are inferred by multiplicative updates with cached positive/negative Gram terms, while the dictionary is updated from accumulated $Z^\top Z$ and $Z^\top A$ statistics with a ridge term. For all methods, we report normalized reconstruction MSE on ImageNet-1K (validation split). See Appendix G for more details.

Results. We compare all methods on the reconstruction–sparsity tradeoff in Figure 5 (right) and Figure G5 (left), report rediscovery metrics in Table 1, and show qualitative examples in Figure 3. SAEs do not improve upon normalized MSE and sparsity, but *do* lead to better concept discovery as measured by mAP, Purity@k, and Coverage@τ when compared to other baseline methods.

Table 2: SAE variant and backbone ablation on ADE20K. Top rows compare SAE variants on DINOv3 ViT-L/16 (final layer); bottom row switches to a pretrained Perception Encoder ViT-L/14 checkpoint, using the Matryoshka+TopK SAE training protocol. Best SAE per configuration selected by training probe loss.

	Source Recon. ↓		Target Recon. ↓		Downstream ↑			
Method	NMSE	L_0	NMSE	L_0	Probe R	mAP	Pur@k	Cov@τ
ReLU	0.166	379.2	0.211	379.2	0.441	0.224	0.748	0.264
+Matryoshka	0.591	**6.5**	0.702	6.5	0.402	0.319	**0.808**	0.405
+TopK	0.207	64	0.274	64	**0.479**	**0.389**	0.788	**0.576**
PE ViT-L/14	**0.120**	128	**0.149**	128	0.123	0.110	0.557	0.092

Table 3: Matryoshka TopK SAEs for concept rediscovery on general domain (ADE20K) and ecological imagery (FishVista) under the same DINOv3 ViT-L/16 backbone. Source Recon. measures reconstruction on the SAE's training distribution; Target Recon. measures transfer to the evaluation domain.

	Source Recon. ↓		Target Recon. ↓		Downstream ↑			
Dataset	NMSE	L_0	NMSE	L_0	Probe R	mAP	Pur@k	Cov@τ
ADE20K	0.207	64	0.274	64	0.479	0.389	0.788	0.576
FishVista	0.184	16	0.184	16	0.438	0.543	0.501	0.7

Discussion. k-means and PCA are competitive on reconstruction–sparsity trade-offs, yet underperform SAEs on downstream concept-alignment metrics such as mAP, Purity@k, and Coverage@τ. This reinforces that normalized MSE and L_0 measure how well activations are compressed, not how well individual latents track semantic concepts. Prior work in evaluating SAEs for language model interpretability also finds dictionary-learning metrics insufficient (Karvonen et al., 2025).

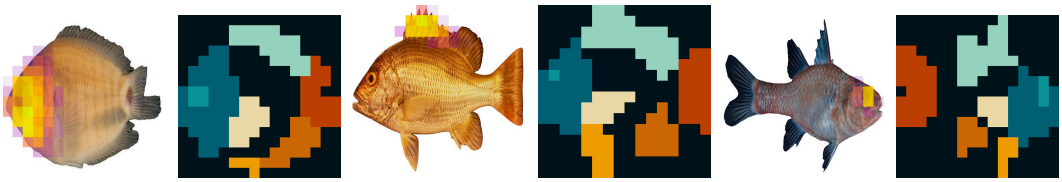
Ablations. We ablate ViT size, ViT layer, ViT family and SAE variants to understand what drives concept rediscovery on ADE20K. We find that Matryoshka SAEs with TopK perform the best on downstream metrics when trained on ImageNet-1K patch activations and evaluated on ADE20K, achieving the highest probe R, mAP, and Coverage@τ. DINOv3 substantially outperforms Perception Encoder on all rediscovery metrics (Table 2), suggesting that self-supervised objectives like DINOv3 produce representations more amenable to sparse decomposition than contrastive vision-language training. Additional ablations are found in Appendix F.

4.2 Domain-specific semantic rediscovery

To validate this approach beyond general-domain semantic concepts, we test it in ecological imagery: can vision foundation models, in combination with SAEs, rediscover the anatomy of diverse fish species? We use FishVista (Mehrab et al., 2024a;b) and the 10 annotated body-part traits to answer this question (see Figure 4 for dataset examples). We use the final DINOv3+TopK setting from Section 4.1, training SAEs on 56.3K fish images and evaluating SAE discovery on the 4.6K (disjoint) images with body part segmentations.

Results. We find that DINOv3+Matryoshka TopK SAEs reliably find domain-specific concepts; we compare results on ADE20K and FishVista in Table 3, show examples in Figure 4 and compare concept prevalence with average precision in Figure 5 (left). These results indicate that strong foundation models, in combination with SAEs, can surface features that align closely with expert-defined anatomical annotations.

Discussion. Body parts that occur in more patches are recovered with higher average precision (Figure 5, left), while rare parts are substantially harder to rediscover. This implies that scientific applications targeting rare or long-tail structures and patterns may need either more data, targeted sampling, or novel dictionary learning algorithms that learn rare structures more efficiently.



(a) **"Head" (DINOv3-16K/9)** (b) **"Dorsal fin" (DINOv3-16K/13)** (c) **"Eye" (DINOv3-16K/69)**

Figure 4: Example TopK SAE features (left) and patch-level segmentation masks (right) for **(a) "Head"**, **(b) "Dorsal fin"** and **(c) "Eye"** from DINOv3 ViT-L/16. Features were picked by minimizing cross entropy on binary classification for each body part, as described in Section 4.2 and eq. (10). **Takeaway**: TopK SAEs accurately learn scientific concepts (anatomical body parts) without labels.

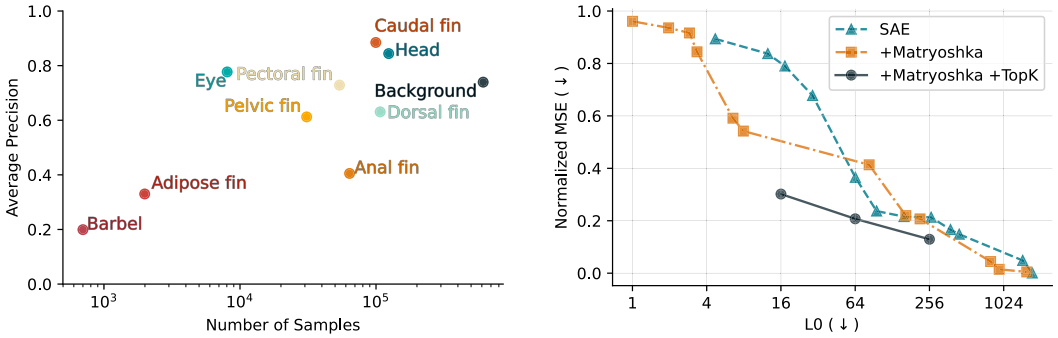
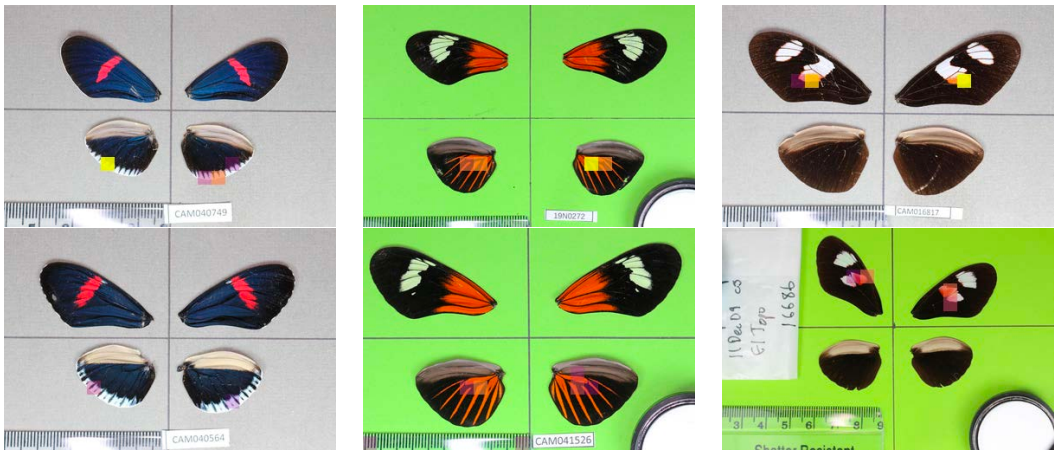


Figure 5: **(Left)** Prevalence vs rediscovery on FishVista. Each point is a body part; x-axis: number of patch-level samples in the validation split, y-axis: AP of the best-matching (lowest probe loss) TopK SAE latent. More common body parts are rediscovered more reliably. **Takeaway:** Higher concept prevalence is associated with more rediscovery. **(Right)** Reconstruction–sparsity tradeoff for SAE variants (ReLU, Matryoshka, TopK) on DINOv3 ViT-L/16 ImageNet-1K activations. TopK achieves the best reconstruction at fixed sparsity.

4.3 Question-driven feature ranking

Previously, we established that SAEs are a strong decomposition method for learning visual features. *Heliconius* butterflies are an often-studied example of mimicry and evolution (McMillan et al., 2020); understanding subspecies wing patterns and traits is of great interest to researchers (Hines et al., 2011; Jiggins, 2008; McMillan et al., 2020). Natural selection has driven striking convergence in the appearance of different *Heliconius* butterfly species, such that different species can appear almost identical (Reed et al., 2011). Nonetheless, decades of work by biologists have revealed certain key traits, such as subtle differences in color pattern, that can be used to distinguish mimics (Müller, 1879; DeVries, 1987; Jiggins, 2008; Lawrence et al., 2026). We propose using SAEs with our *selectivity* heuristic (Section 3) to identify visual features of *Heliconius* wings that may separate mimic subspecies. To test the efficacy of using this approach in practice, we test whether it can efficiently rediscover known diagnostic traits between *Heliconius erato* and *Heliconius melpomene* subspecies. We train Matryoshka TopK SAEs on DINOv3 layer 22 and 24 patch activations of 24.8K *Heliconius* images (Lawrence et al., 2024) and rank features for selectivity between three mimic pairs. We then compare the highest selectivity score across all values of k in the Pareto-optimal SAEs.

First, we find that layer 22 consistently provides higher selectivity scores, indicating that the last layer may not be optimal for all cases. Then, a co-author with domain expertise in *Heliconius* wing-patterns manually annotated the 5 highest-ranked features from layer 22 pareto-optimal SAEs for each mimic pair, determining whether each corresponded to a known diagnostic trait. We find that in all 3 mimic pairs, the Matryoshka TopK SAE features with the highest selectivity values correspond to known diagnostic traits (Figure 6, Appendix I) (Cast, 2025; Lawrence et al., 2026). Additionally, for each mimic pair, 60-100% of the five highest-ranked features by selectivity corresponded to known diagnostic traits. We provide images and annotations of the five highest-ranked features by selectivity for each mimic pair in Appendix I. Exact SAE training parameters are provided in Appendix E.



(a) **Feature 1,662** (sel.=0.96) (b) **Feature 104** (sel.=0.99) (c) **Feature 192** (sel.=0.94)

Figure 6: For a single Matryoshka TopK SAE trained on DINOv3 ViT-L/16 patch activations, the highest selectivity features for each studied *H. erato* (top)/*H. melpomene* (bottom) mimic pair. **(a) Feature 1,662** corresponds to hindwing fringe, **(b) Feature 104** corresponds to hindwing ray shape, and **(c) Feature 192** corresponds to the forewing red patch placement. Note that the highest-ranked *H. melpomene* image for **(b) Feature 104** is actually a mislabeled *H. erato*. Similar features were ranked very highly for all 3 pairs among other trained SAEs. **Takeaway**: SAE features can be easily ranked using relevant heuristics to identify the most promising discovery candidates.

Discussion. The consistently high rank of features corresponding with expert-validated diagnostic traits in *Heliconius* mimic pairs suggests that using well-designed heuristics enables quick identification of scientifically interesting features, without the need for fine-grained concept labels. Additionally, any heuristic could easily be used in place of selectivity, allowing for researchers to select for any type of visual feature that may be interesting for a specific problem.

5 Conclusion

In the settings we studied, we demonstrated that sparse autoencoders can systematically recover semantic structure from vision foundation model representations without concept supervision, using controlled rediscovery on both general-domain visual concepts and a scientific case study. On ADE20K, SAEs consistently rediscovered semantic units from DINOv3 activations trained only on ImageNet-1K, outperforming decomposition baselines (k-means, PCA & SemiNMF) on almost all concept alignment metrics. On ecological images, the same procedure surfaced anatomical structures without using segmentation or part labels during extraction. Finally, when used with simple ranking heuristics, features corresponding with known diagnostic traits were easily identified. Our results provide evidence that SAEs are a practical method for extracting interpretable features from vision foundation models without needing supervision. With this, the quantitative nature of SAE features shifts the search for interesting visual phenomena to a heuristic design problem, allowing researchers to better leverage foundation model representations and more quickly create meaningful hypotheses. Importantly, our contributions should be understood as validating SAEs as an *instrument* for open-ended visual discovery, not as delivering novel biological or physical discoveries themselves. In the future, we aim to use this approach in collaboration with domain experts to extract truly novel scientific insights from vision foundation models.

6 AI STATEMENT

In this work, we used generative AI tools to assist with editing software code, editing drafts, and reviewing existing literature. We have reviewed all AI-assisted work, including reviewing AI-generated code and manually reading any sources provided by AI when used for reviewing literature. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

7 REPRODUCIBILITY STATEMENT

In order to ensure reproducibility of this work, all training and implementation details of SAEs, metrics, and ranking heuristics are provided in Section 3, Section 4, Appendix B, Appendix C, and Appendix E. Additionally, code used for all experiments is provided at `https://anonymous.4open.science/r/saev-25A1`.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11): 4311–4322, 2006.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. Factors of transferability for a generic convnet representation. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1790–1802, 2015.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=INqBOmwIpG`.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, 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. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Michel Cast. La diversité des Heliconius, October 2025. URL `https://cliniquevetodax.com/Heliconius/Index.html`.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Philip J. DeVries. *The butterflies of Costa Rica and their natural history*. Princeton University Press, Princeton, N.J, 1987. ISBN 978-0-691-02890-3.

Chris HQ Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55, 2008.

Konstantin Donhauser, Kristina Ulicna, Gemma Elyse Moran, Aditya Ravuri, Kian Kenyon-Dean, Cian Eastwood, and Jason Hartford. Towards scientific discovery with dictionary learning: Extracting biological concepts from microscopy foundation models. *arXiv preprint arXiv:2412.16247*, 2024.

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. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems*, 36: 54805–54818, 2023a.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2711–2721, 2023b.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Jianyang Gu, Samuel Stevens, Elizabeth G Campolongo, Matthew J Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E. White, James Balhoff, Wasila Dahdul, Daniel Rubenstein, Hilmar Lapp, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP 2: Emergent properties from scaling hierarchical contrastive learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=yPC9zmkQgG.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Jürgen Haas, Alessandro Barbato, Dario Behringer, Gabriel Studer, Steven Roth, Martino Bertoni, Khaled Mostaguir, Rafal Gumienny, and Torsten Schwede. Continuous automated model evaluation (cameo) complementing the critical assessment of structure prediction in casp12. *Proteins: Structure, Function, and Bioinformatics*, 86:387–398, 2018.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SaKO6z6Hl0c.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Jingyi He, Haiyan Zhao, Ruxue Shi, Yanguang Liu, Xin Wang, Fei Sun, and Mengnan Du. SAEExplainer: Interpreting SAE Features with Activation-Guided Preference Optimization, September 2026. URL https://arxiv.org/abs/2606.08496.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Heather M. Hines, Brian A. Counterman, Riccardo Papa, Priscila A. Moura, Marcio Z. Cardoso, Mauricio Linares, James Mallet, Robert D. Reed, Chris D. Jiggins, Marcus R. Kronforst, and W. Owen McMillan. Wing patterning gene redefines the mimetic history of *Heliconius* butterflies. *Proceedings of the National Academy of Sciences*, 108(49):19666–19671, December 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1110096108. URL https://pnas.org/doi/10.1073/pnas.1110096108.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Edward Hughes, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktaschel. Open-endedness is essential for artificial superhuman intelligence. *arXiv preprint arXiv:2406.04268*, 2024.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=BH8TYy0r6u.

Chris D. Jiggins. Ecological Speciation in Mimetic Butterflies. *BioScience*, 58(6):541–548, June 2008. ISSN 1525-3244, 0006-3568. doi: 10.1641/B580610. URL http://academic.oup.com/bioscience/article/58/6/541/236012/Ecological-Speciation-in-Mimetic-Butterflies.

Sonia Joseph, Praneet Suresh, Ethan Goldfarb, Lorenz Hufe, Yossi Gandelsman, Robert Graham, Danilo Bzdok, Wojciech Samek, and Blake Aaron Richards. Steering CLIP's vision transformer with sparse autoencoders, 2025. URL https://arxiv.org/abs/2504.08729.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qrU3yNfX0d.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thomas S Kuhn and Ian Hacking. *The structure of scientific revolutions*, volume 2. University of Chicago press Chicago, 1970.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Christopher Lawrence, Elizabeth Campolongo, and Neil Rosser. Heliconius collection (cambridge butterfly), 2024. URL https://huggingface.co/datasets/imageomics/Heliconius-Collection_Cambridge-Butterfly.

Christopher Lawrence, Michelle Ramirez, Tanya Berger-Wolf, Owen McMillan, and Daniel Rubenstein. The perceptual and spatial architecture of Müllerian mimicry in *Heliconius* Butterflies, March 2026. URL http://biorxiv.org/lookup/doi/10.64898/2026.03.02.709095.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. *arXiv preprint arXiv:2502.04878*, 2025.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016.

H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, March 1947. ISSN 0003-4851. doi: 10.1214/aoms/1177730491. URL http://projecteuclid.org/euclid.aoms/1177730491.

S. J. Mason and N. E. Graham. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, 128(584):2145–2166, July 2002. ISSN 0035-9009, 1477-870X. doi: 10.1256/003590002320603584. URL https://rmets.onlinelibrary.wiley.com/doi/10.1256/003590002320603584.

W. Owen McMillan, Luca Livraghi, Carolina Concha, and Joseph J. Hanly. From Patterning Genes to Process: Unraveling the Gene Regulatory Networks That Pattern Heliconius Wings. *Frontiers in Ecology and Evolution*, 8:221, July 2020. ISSN 2296-701X. doi: 10.3389/fevo.2020.00221. URL https://www.frontiersin.org/article/10.3389/fevo.2020.00221/full.

Kazi Sajeed Mehrab, M. Maruf, Arka Daw, Harish Babu Manogaran, Abhilash Neog, Mridul Khurana, Bahadir Altintas, Yasin Bakis, Elizabeth G Campolongo, Matthew J Thompson, Xiaojun Wang, Hilmar Lapp, Wei-Lun Chao, Paula M. Mabee, Henry L. Bart Jr., Wasila Dahdul, and Anuj Karpatne. Fish-vista: A multi-purpose dataset for understanding & identification of traits from images, 2024a. URL https://arxiv.org/abs/2407.08027.

Kazi Sajeed Mehrab, M. Maruf, Arka Daw, Harish Babu Manogaran, Abhilash Neog, Mridul Khurana, Bahadir Altintas, Yasin Bakış, Elizabeth G Campolongo, Matthew J Thompson, Xiaojun Wang, Hilmar Lapp, Wei-Lun Chao, Paula M. Mabee, Henry L. Bart Jr., Wasila Dahdul, and Anuj Karpatne. Fish-vista: A multi-purpose dataset for understanding & identification of traits from images, 2024b. URL https://huggingface.co/datasets/imageomics/fish-vista.

Fritz Müller. Ituna and thyridia: a remarkable case of mimicry in butterflies. *Trans. Entomol. Soc. Lond*, 1879:20–29, 1879.

Meike Nauta, Ron Van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14933–14943, 2021.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, 29, 2016.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces, 2025. URL https://arxiv.org/abs/2504.11695.

Gonçalo Paulo and Nora Belrose. Sparse Autoencoders Trained on the Same Data Learn Different Features, 2025. URL https://arxiv.org/abs/2501.16615.

Kenny Peng, Rajiv Movva, Jon Kleinberg, Emma Pierson, and Nikhil Garg. Use sparse autoencoders to discover unknown concepts, not to act on known concepts. *arXiv preprint arXiv:2506.23845*, 2025.

Joana Pereira, Adam J Simpkin, Marcus D Hartmann, Daniel J Rigden, Ronan M Keegan, and Andrei N Lupas. High-accuracy protein structure prediction in casp14. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1687–1699, 2021.

Henri Poincaré. *Science and method*. Dover Publications, Mineola, N.Y, dover ed edition, 2003. ISBN 978-0-486-43269-4.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, pp. 444–461. Springer, 2024.

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4088–4099, 2023.

Robert D. Reed, Riccardo Papa, Arnaud Martin, Heather M. Hines, Brian A. Counterman, Carolina Pardo-Diaz, Chris D. Jiggins, Nicola L. Chamberlain, Marcus R. Kronforst, Rui Chen, Georg Halder, H. Frederik Nijhout, and W. Owen McMillan. *optix* Drives the Repeated Convergent Evolution of Butterfly Wing Pattern Mimicry. *Science*, 333(6046):1137–1141, August 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1208227. URL https://www.science.org/doi/10.1126/science.1208227.

Yusuf H Roohani, Tony J Hua, Po-Yuan Tung, Lexi R Bounds, Feiqiao B Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S Plosky, et al. Virtual cell challenge: Toward a turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A Multimodal Automated Interpretability Agent, 2024. URL https://arxiv.org/abs/2404.14394.

Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19412–19424, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

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.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12884–12893, 2021.

Nicholas Wang, Ching Fang, Mark Bissell, Archa Jain, and Daniel Balsam. Using Interpretability to Identify a Novel Class of Alzheimer's Biomarkers, 2026. URL https://www.goodfire.ai/research/interpretability-for-alzheimers-detection.

Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3124–3134, 2023a.

Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15790–15801, 2023b.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *arXiv preprint arXiv:2208.10431*, 2022.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

APPENDICES

1. Appendix A: Limitations
2. Appendix B: Training details and hyperparameters
3. Appendix C: Precise metric definitions
4. Appendix D: Additional qualitative examples
5. Appendix E: SAE training details for *Heliconius* images
6. Appendix F: Detailed ablation results
7. Appendix G: Baseline methods
8. Appendix H: Additional layer-depth ablations
9. Appendix I: Additional butterfly feature images

A LIMITATIONS

While the proposed methodology effectively learns and identifies meaningful, relevant visual features for discovery, advances are still needed to better enable discovery. First, our evaluation relies on ground-truth annotations to measure concept alignment, which limits our ability to assess discovery in truly unannotated domains. Second, previous work has demonstrated issues with SAE features, such as instability (Paulo & Belrose, 2025; Papadimitriou et al., 2025; Leask et al., 2025) and splitting or absorption (Chanin et al., 2024). With this, vision features can be difficult to interpret, and SAE features may not perfectly align with a human interpretation of them. As such, it may be necessary to include additional mechanisms for interpretability (Shaham et al., 2024; He et al., 2026) or leverage multimodal models. This approach also only provides *possibly* interesting features, and does not perform fully-automatic scientific discovery. However, we argue that visual scientific discovery necessarily requires expert validation and reasoning. Finally, while strong performance on two domains suggests the method will generalize, we leave both systematic evaluation of novel discoveries and applying SAEs to non-vision foundation models to future work.

B TRAINING DETAILS

Our code contains instructions to reproduce our exact findings.

SAE Training.

- Initialize W_{enc} and W_{dec} with Kaiming initialization (He et al., 2015) and b_{enc} and b_{dec} with all zeros.
- Train SAEs with Adam (Kingma & Ba, 2014) for 100M examples with a batch size of $16\,384$.
- Linearly warmup learning rate from 0 to the maximum over 500 steps, then use cosine decay to 0.
- Linearly warmup the sparsity coefficient λ from 0 to the maximum over all of training.
- Normalize W_{dec} to have unit norm columns at every step and remove gradients parallel to the columns to avoid interactions between Adam moments and normalization (Bricken et al., 2023).
- Sweep both learning rate and sparsity coefficient λ for each setting, using the values in Table B1.

Baseline Training. We train k-means and PCA with randomly shuffled mini-batches with a batch size of $16\,384$ using 100M examples.

Hyperparam	Sweep Values
Learning Rate	$[3 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}, 1 \times 10^{-2}, 3 \times 10^{-2}, 1 \times 10^{-1}]$
Sparsity Coeff.	$[1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}]$

Table B1: Sparse autoencoder hyperparameter sweep values.

1D Probe Training. We optimize Equation (10) for 1-dimensional logistic regression classifiers for each pair of {SAE latent i, class c} with a ridge of 1×10^{-8} and 30 steps of Newton-style optimization. We initialize w to 0 and b to the prevalence of the class c in the entire dataset.

C METRIC DEFINITIONS

We propose *selectivity* as a heuristic for identifying features corresponding to known diagnostic traits in *Heliconius* butterflies. We use a variety of metrics to evaluate SAE probe quality. We include scatter plots of our different metrics (Probe R, mAP, Purity@k and Coverage@τ) in Figure C1.

Selectivity Let $x_1, ..., x_{n_1}$ and $y_1, ..., y_{n_2}$ be SAE feature values (max per image) corresponding to 2 classes. Then, let T be the sum of the ranks of values $y_1, ..., y_{n_2}$ in the ordered sequence of x's and y's. We then have:

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T \tag{8}$$

where U is the Mann-Whitney U-statistic (Mann & Whitney, 1947). Therefore, in full we define *selectivity* as:

$$\text{sel.} = \left|\frac{n_1 n_2 + \frac{n_2(n_2+1)}{2} - T}{n_1 n_2} - 0.5\right| * 2 \tag{9}$$

Probe R. For each ground truth concept, we fit a binary logistic on each latent i's activation z_i and select the best latent per class:

$$\min_{i,w,b} \mathbb{E}\big[- y \log \sigma(w z_i + b) - (1 - y) \log(1 - \sigma(w z_i + b))\big] \tag{10}$$

Because the probe is axis-aligned by design, it favors representations whose individual axes carry semantic meaning. This is not a limitation of the evaluation but a requirement of the use case: features that cannot be understood one at a time cannot be browsed for discovery.

We record the binary cross entropy loss (see Equation (10)) for the trained probe ($\mathcal{L}$) and the loss for a bias-only probe ($\mathcal{L}_\pi$). We normalize the trained loss by the bias-only loss: $1 - (\mathcal{L}/\mathcal{L}_\pi)$.

mAP. For each class, we treat its patches as positives and all others as negatives, rank all validation patches by the corresponding probe score, and compute average precision as the area under the precision–recall curve. We report mean average precision (mAP) as the macro-average of per-class AP values, assigning AP=0 to classes without positive examples.

Purity@k. For each SAE latent i, we take the k image patches that maximally activate latent i and measure how many of the k patches have the same patch label, then normalize by k to get a score between 0 and 1. We set k=16 in our work.

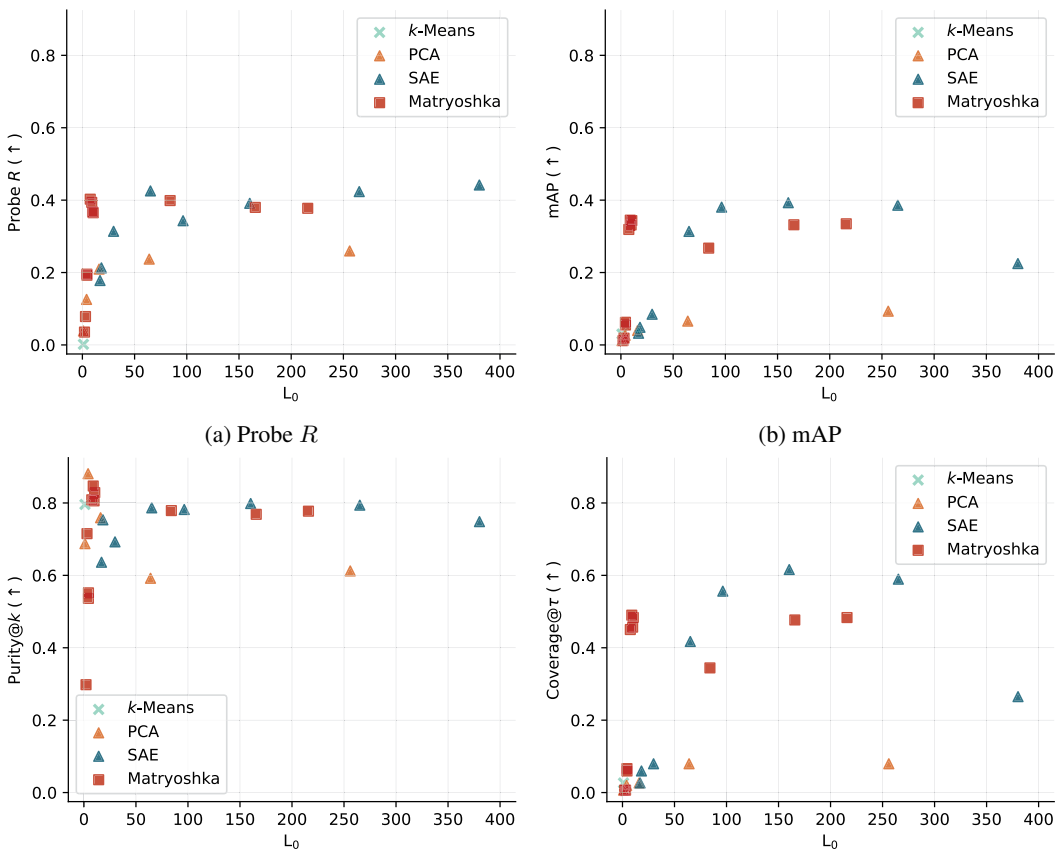
Coverage@τ. We measure the fraction of classes (151 for ADE20K, 10 for FishVista) that have an SAE latent with a minimum average precision of τ. We set τ=0.3 in our work.

D QUALITATIVE EXAMPLES

We show additional qualitative examples of both success and failure for ADE20K classes for all methods in Figure D2.

E BUTTERFLY SAE TRAINING

Our code contains instructions to reproduce our exact findings.



(a) Probe R

(b) mAP

(c) Purity@k

(d) Coverage@τ

Figure C1: We compare k-means, PCA, vanilla SAEs and Matryoshka SAEs against sparsity (L_0) and downstream metrics (Probe R, mAP, Purity@k and Coverage@τ) for ADE20K.

Butterfly Image Patch Activations

- Patch activations for layers 22 and 24 of DINOv3 ViT-L/16
- 384 patches per image, ratio-aware resized
- Only dorsal images of each subspecies pair used for calculating selectivity

SAE Training.

- Initialize W_{enc} and W_{dec} with Kaiming initialization (He et al., 2015) and b_{enc} and b_{dec} with all zeros.
- Train Matryoshka TopK SAEs with Adam (Kingma & Ba, 2014) for 100M examples with a batch size of 16 384.
- 10 prefixes sampled for each training step
- Linearly warmup learning rate from 0 to the maximum over 500 steps, then use cosine decay to 0.
- Normalize W_{dec} to have unit norm columns at every step and remove gradients parallel to the columns to avoid interactions between Adam moments and normalization (Bricken et al., 2023).
- Sweep both learning rate and k for each setting, using the values in Table E2.



(a) k-means (b) PCA (c) Vanilla SAE (d) Matryoshka SAE

Figure D2: Some additional qualitative examples for **(a)** k-means: "waterfall" (left) and "road" (right), **(b)** PCA: "sky" (left) and "pool table" (right), **(c)** vanilla SAEs: "dishwasher" (left) and "bed" (right) and **(d)** Matryoshka SAEs: "towel" (left) and "chandelier" (right).

Hyperparam	Sweep Values
Learning Rate	$[3 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}]$
k	[16, 32, 64, 128]

Table E2: Sparse autoencoder hyperparameter sweep values for SAEs trained on *Heliconius* images.

F Ablation Results

We report all metrics for the best SAE for each ViT size and layer in Table H3.

Model size. Model scale consistently correlates with downstream performance (Kaplan et al., 2020; Hoffmann et al., 2022). Prior work proposes scaling laws for SAE reconstruction based on both SAE and foundation model size (Gao et al., 2025) but does not investigate downstream performance. We assess foundation model size effects by training SAEs on patch activations from DINOv3 ViT-S/16, ViT-B/16, and ViT-L/16 checkpoints. The models vary in both total parameter count (22M, 86M and 303M, respectively) and the embedding dimension d (384, 768, and 1024, respectively). For

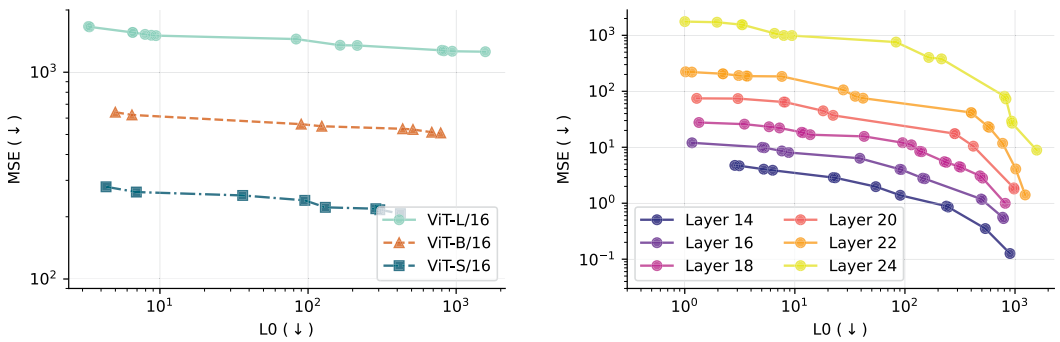


Figure F3: **(Left)** The reconstruction–sparsity tradeoff for SAEs trained on different sized ViTs. Larger backbones are harder to reconstruct at a given sparsity (ViT-L/16 > ViT-B/16 > ViT-S/16), though all improve as L_0 increases. Note that the y-axis is *raw* rather than normalized MSE. **(Right)** Layer-wise reconstruction–sparsity trade-off (DINOv3 ViT-L/16) on ImageNet-1K. Later layers consistently incur higher MSE at matched sparsity, indicating greater representational complexity.

each size, we keep the SAE width fixed at $16\,384$, train on ImageNet-1K 100M patch activations and probe the SAEs for ADE20K features.

In Figure F3 (left) we see that larger models are in fact harder to reconstruct (note that the y-axis is *raw* rather than normalized MSE), consistent with prior work in scaling SAEs for language models (Gao et al., 2025). Despite more challenging dictionary learning, we see in Appendix H that SAEs applied to larger models both recover more concepts (coverage) and are more precise (purity). All three sizes of the DINOv3 ViTs are distilled from a single, larger teacher model (Siméoni et al., 2025); we conjecture then that all three checkpoints have learned the same set of concepts. These results indicate that, at least for vision, strong foundation models in combination with SAEs can rediscover semantic concepts, which we view as a prerequisite for scientific discovery.

A potential confounder is total training FLOPs: prior work shows that additional SAE training compute (total FLOPs) leads to stronger SAEs as measured by the reconstruction–sparsity tradeoff (Gao et al., 2025). Our SAEs trained on larger embedding dimensions d use more compute during training; this could explain the stronger results for larger models. However, our SAEs trained on larger models have *worse* reconstruction–sparsity but *better* downstream concept alignment metrics.

Layer depth. Representations evolve across network depth: later layers in CNNs and ViTs tend to encode higher-level, more linearly separable semantics, which often improves transfer and dense alignment (Yosinski et al., 2014; Azizpour et al., 2015; Caron et al., 2021). However, recent results show that the final layer is not always optimal (Bolya et al., 2025). Intermediate layers can yield less entangled, more spatially localized features depending on the downstream task. We therefore evaluate six layers in the second half of the transformer for the three evaluated checkpoints: for the 24-layer ViT-L/16 model, we use layers 14, 16, 18, 20, 22, and 24 and for the 12-layer ViT-S/16 and ViT-B/16 we use the last 6 layers 7, 8, 9, 10, 11, and 12.

In Figure F3, we find that later layers are harder to reconstruct (note that the y-axis is *raw* rather than normalized MSE). We also evaluate downstream metrics (mAP) for the best SAE at each layer for all three ViT checkpoints in Figure F4. Despite more challenging reconstruction, we find that later layers lead to better recovery for all three checkpoints. Prior work in SAEs for language models finds that later layers are harder to reconstruct except for the final two layers (Gao et al., 2025). Language models use a linear embedding layer for next-token-prediction; the final layers must optimize for linear separability. In contrast, DINOv3 uses multiple, loss-specific multilayer perceptrons with additional non-linearities for pre-training (Siméoni et al., 2025); we conjecture this architectural difference leads to the difference in our results.

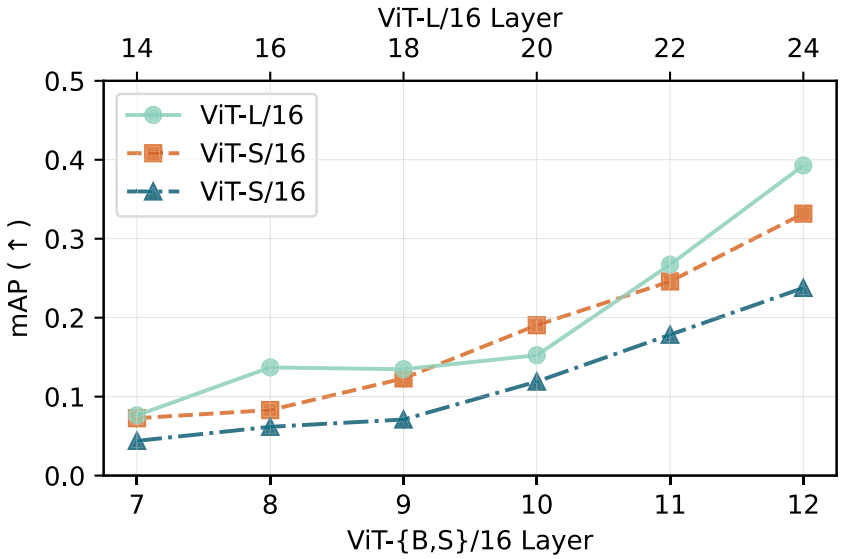


Figure F4: Matryoshka SAE probe quality as measured by mean average precision (mAP) for DINOv3 ViT-S/16, ViT-B/16 and ViT-L/16 checkpoints over different transformer layers. **Takeaway:** Despite larger ViTs and later layers being harder to reconstruct at a given sparsity (see Figure F3 for ViT size ablation), larger ViTs and later layers lead to better SAE concept alignment.

G Baseline Methods

We compare k-means, PCA, SemiNMF and SAEs along the reconstruction–sparsity tradeoff for learning to decompose ViT patch activations in Figure G5. While PCA and SemiNMF provide similar reconstruction-sparsity tradeoffs, we have demonstrated that they perform worse on downstream concept alignment.

H Additional Layer-Depth Ablations

Figure F4 reports mAP across layers; here we show the remaining metrics in Figure H6 and Table H3. All four metrics exhibit the same trend: later layers yield better concept alignment despite harder reconstruction, with larger models achieving slightly higher concept alignment in later layers.

I Additional Butterfly Feature Images

We provide the five highest-ranked features by selectivity across all pareto-optimal SAEs on layer 22 for each mimic pair, as well as their annotations: *H. erato cyrbia* and *H. melpomene cythera* (Figure I7), *H. erato lativitta* and *H. melpomene malleti* (Figure I8), and *H. erato notabilis* and *H. melpomene plesseni* (Figure I9).

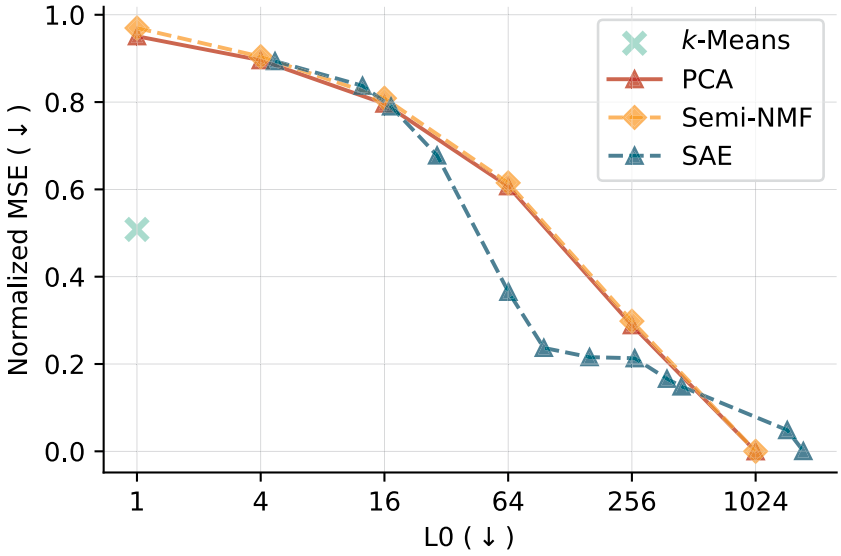
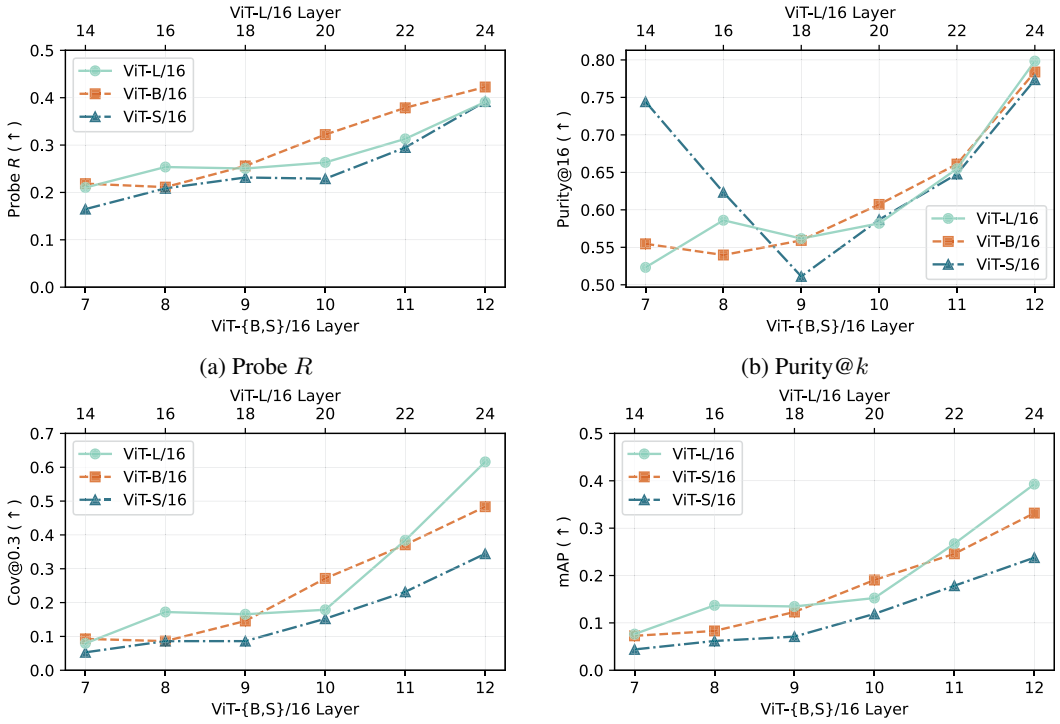


Figure G5: We compare k-means, PCA, SemiNMF and SAEs along the reconstruction–sparsity tradeoff for learning to decompose ViT patch activations. We use final layer activations from DINOv3 ViT-L/16 on ImageNet-1K; we fit all methods on the training split and measure normalized MSE (see Equation (5)) and L_0 on the validation split. For k-means, we "reconstruct" every patch with its nearest centroid; thus, L_0 is always 1. For PCA, we sweep the number of components. We sweep λ for the SAE and show the Pareto frontier of reconstruction–sparsity. **Takeaway:** Reconstruction–sparsity does not indicate an optimal method.

Table H3: Probe results for Matryoshka SAEs trained on ImageNet-1K and evaluated on ADE20K. For each layer in each size of ViT, we choose the best SAE based on training probe loss and report all dictionary learning and downstream metrics.

Model		Dict. ↓		Downstream ↑			
Size	Layer	NMSE	L_0	Probe R	mAP	Purity@k	Coverage@τ
ViT-S/16	7	0.182	70.9	0.165	0.044	0.744	0.053
ViT-S/16	8	0.143	135.1	0.209	0.062	0.623	0.086
ViT-S/16	9	0.133	159.8	0.232	0.071	0.511	0.086
ViT-S/16	10	0.299	84.1	0.251	0.099	0.626	0.093
ViT-S/16	11	0.438	23.6	0.310	0.160	0.645	0.205
ViT-S/16	12	0.294	95.1	0.434	0.203	0.742	0.245
ViT-B/16	7	0.572	19.5	0.216	0.055	0.548	0.066
ViT-B/16	8	0.497	35.4	0.240	0.061	0.516	0.073
ViT-B/16	9	0.571	10.7	0.265	0.109	0.613	0.126
ViT-B/16	10	0.431	20.1	0.322	0.191	0.607	0.272
ViT-B/16	11	0.271	47.5	0.379	0.246	0.661	0.371
ViT-B/16	12	0.359	89.9	0.441	0.312	0.812	0.430
ViT-L/16	14	0.252	90.7	0.210	0.076	0.523	0.079
ViT-L/16	16	0.336	92.4	0.266	0.117	0.575	0.152
ViT-L/16	18	0.659	11.8	0.261	0.090	0.533	0.099
ViT-L/16	20	0.471	22.2	0.263	0.152	0.582	0.179
ViT-L/16	22	0.380	35.3	0.313	0.267	0.655	0.384
ViT-L/16	24	0.703	6.5	0.403	0.319	0.809	0.450



(a) Probe R

(b) Purity@k

(c) Coverage@τ

(d) mAP (reproduced from Figure F4)

Figure H6: Layer-depth ablation for all four downstream metrics on ADE20K, for DINOv3 ViT-S/16, ViT-B/16, and ViT-L/16. The trend from Figure F4 (later layers improve concept alignment) holds across all metrics.

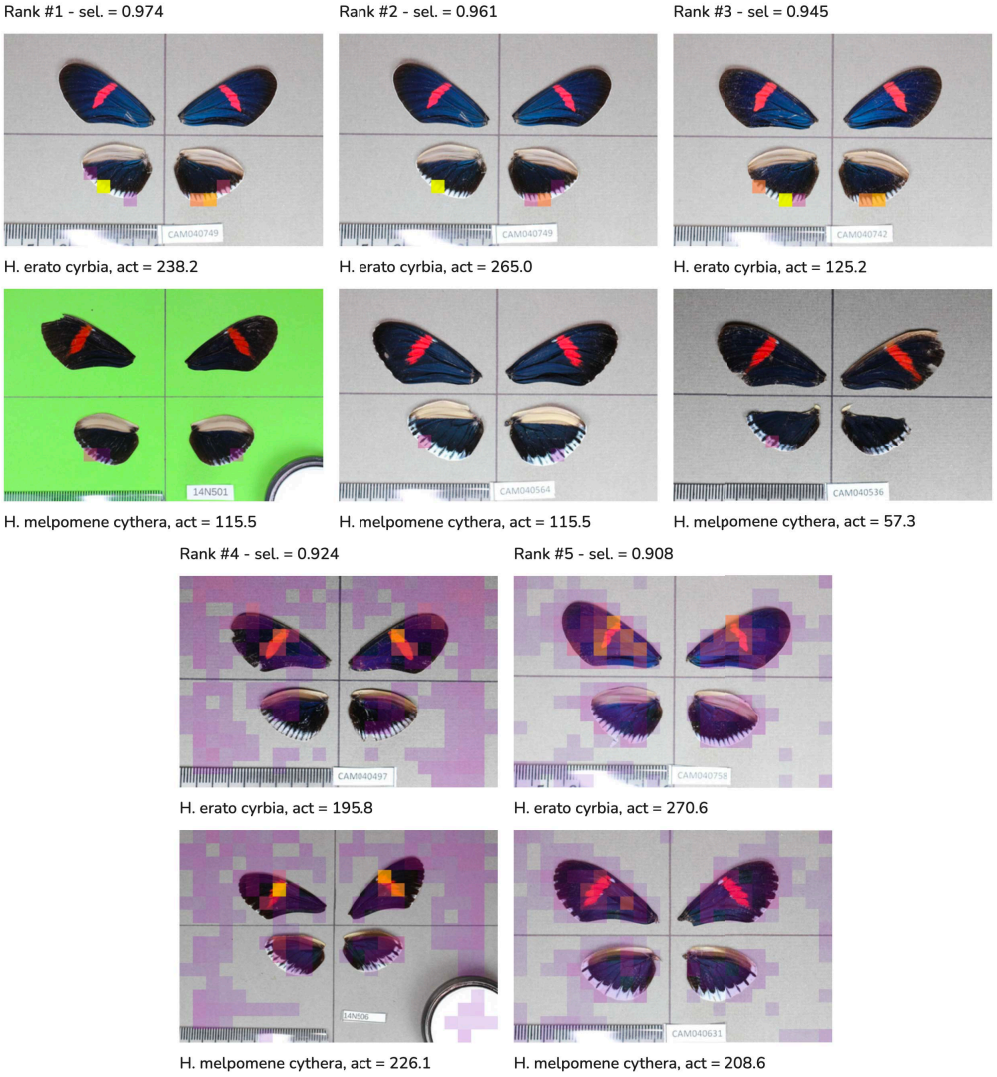


Figure I7: 5 highest selectivity features for *H. erato cyrbia*/*H. melpomene cythera*. Features 1-3 correspond to hindwing fringe in *H. erato cyrbia*, a known diagnostic trait. Features 4 and 5 may correspond to the shape of the forewing red patch, a known diagnostic trait, but these are not counted as corresponding to a known trait due to difficulty interpreting diffuse activation patterns.

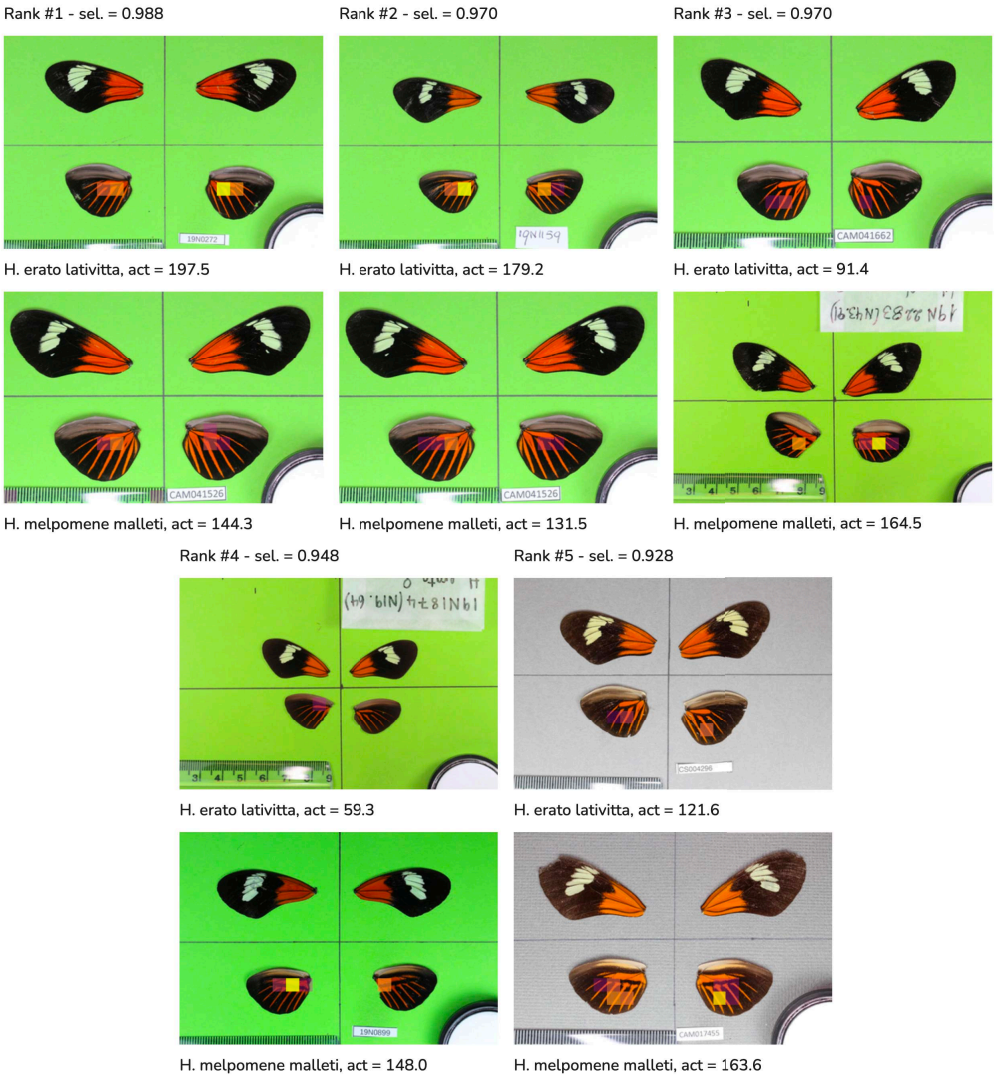


Figure I8: 5 highest selectivity features for *H. erato lativitta*/*H. melpomene malleti*. All 5 features correspond to the black bar or striping on the hindwing, a known diagnostic trait in this mimic pair.

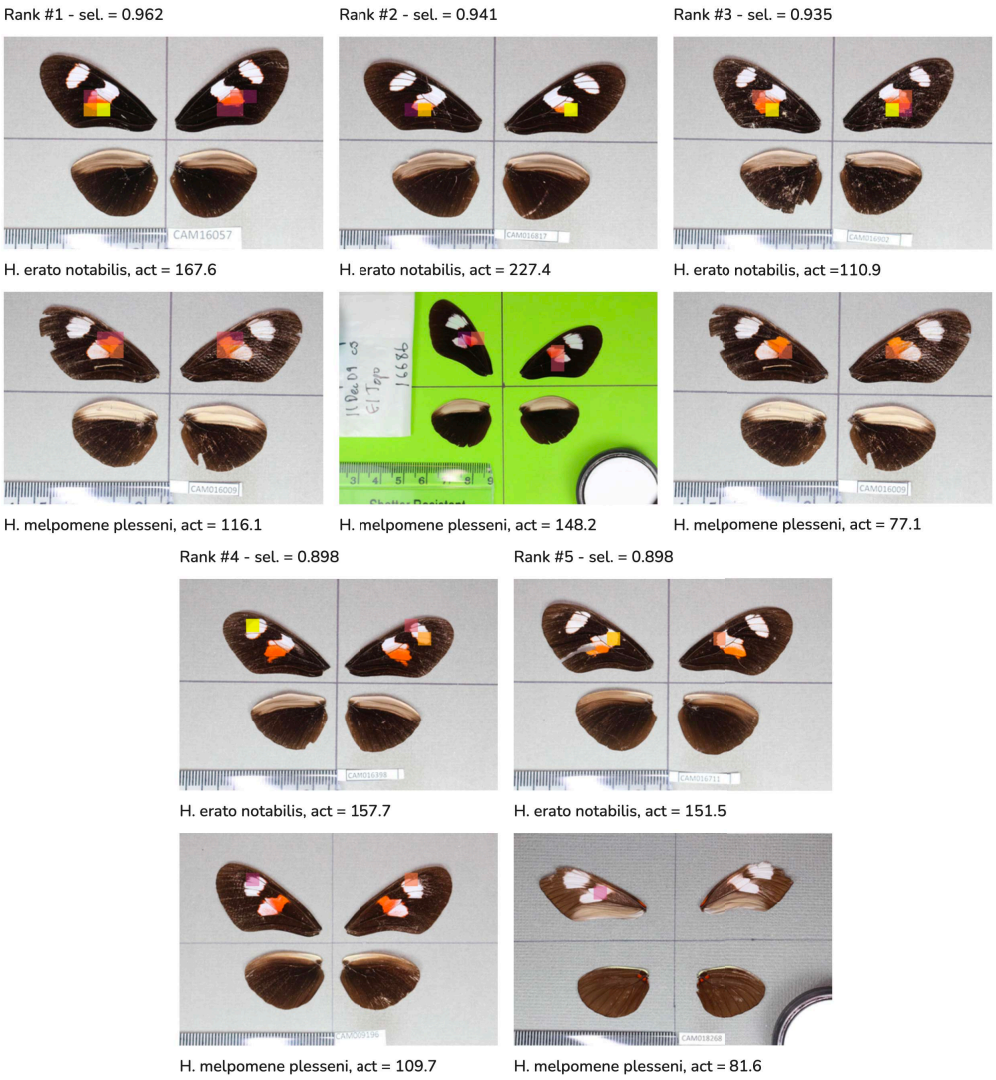


Figure I9: 5 highest selectivity features for *H. erato notabilis*/*H. melpomene plesseni*. Features 1, 2, 3, and 5 correspond to the red forewing patch placement in *H. erato notabilis*, a known diagnostic trait. Feature 4 corresponds to the upper forewing white patch in *H. erato notabilis*. No diagnostic trait for this portion of the wing is well-documented in this mimic pair.