Targeted Recovery of Weight-Space Mechanisms From Neural Networks

Antoine Vigouroux[1] Lee Sharkey[2]

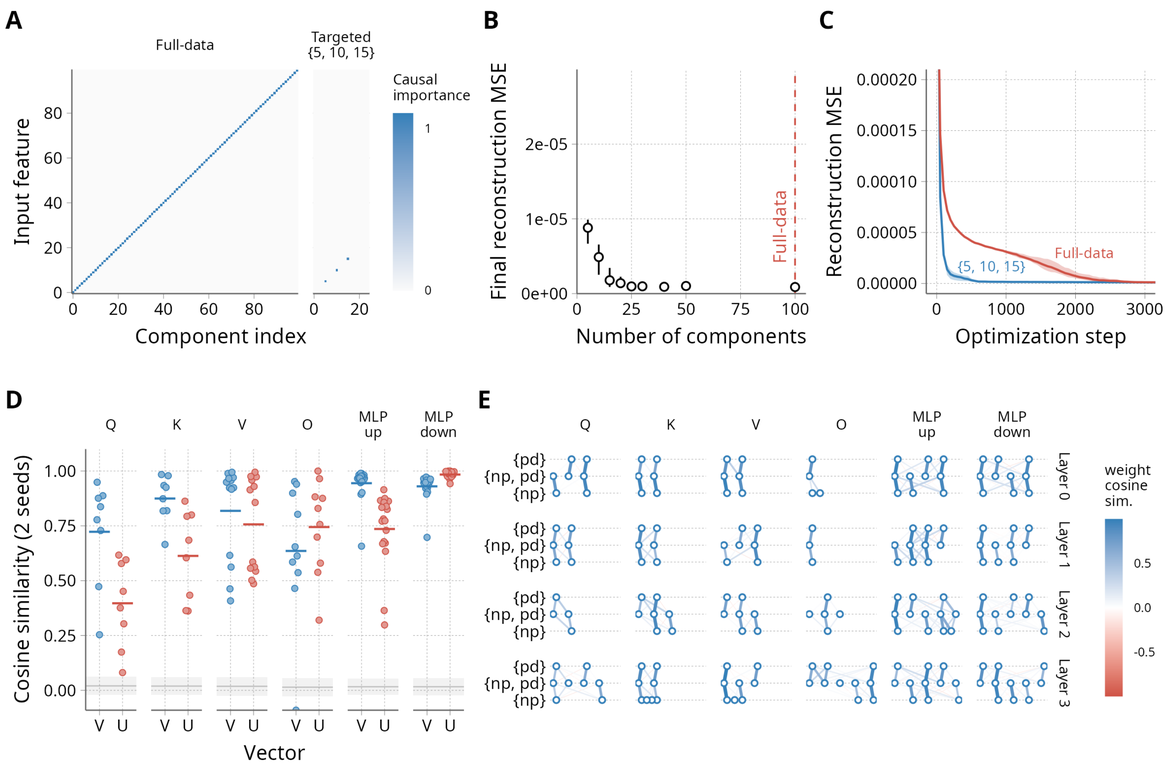
[1]MATS [2]Goodfire

ICML 2026 Mechanistic Interpretability Workshop

Abstract

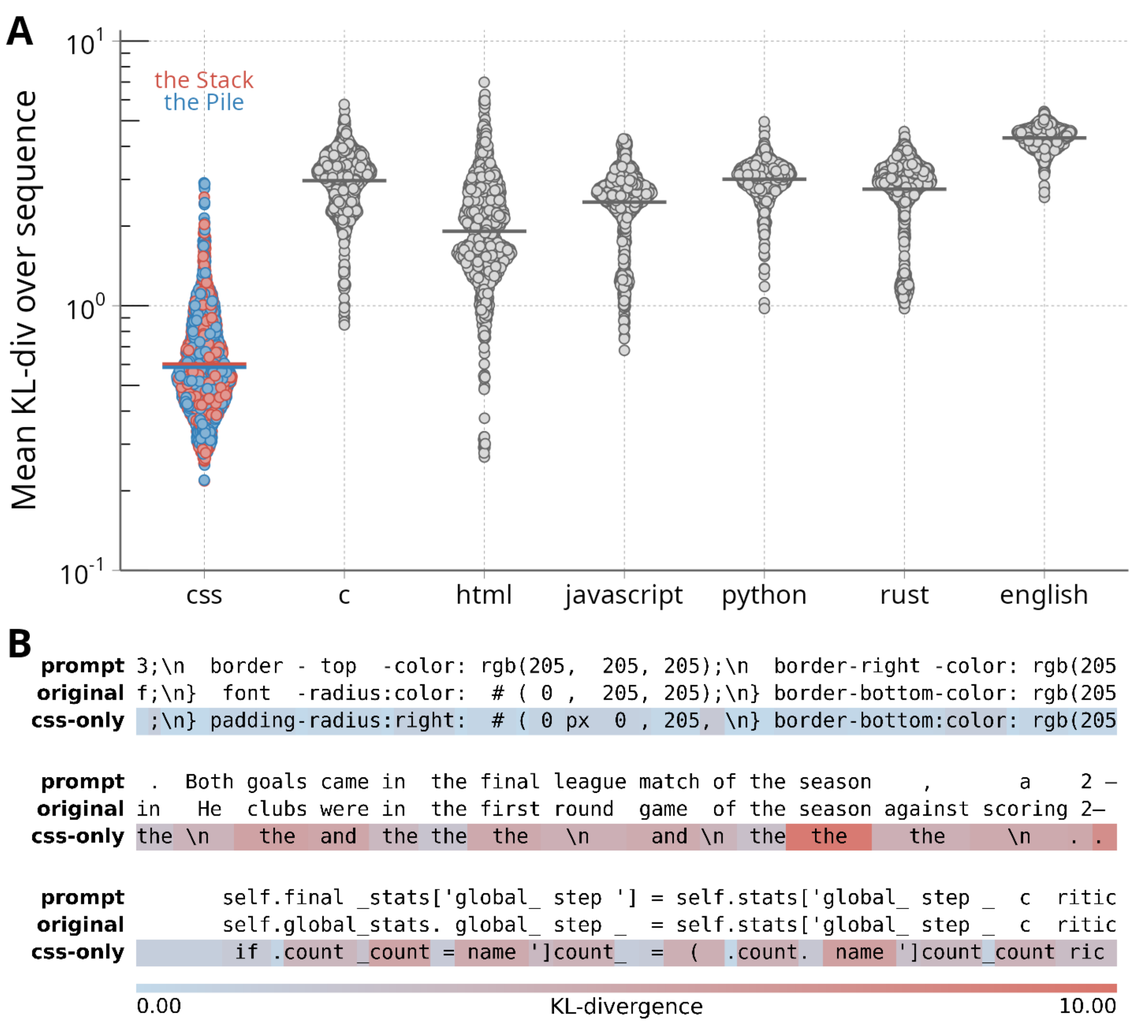
Parameter decomposition (PD) decomposes neural networks into interpretable computational components that faithfully reflect the original network's operations. However, scaling PD to large models requires vast compute, making it a costly and risky endeavor. Here we propose targeted PD (tPD), which identifies only the components that process specific inputs of interest – from isolated prompts to large subtasks – by introducing a high-rank catch-all component that handles all non-target data. We validate tPD on toy models and on transformer language models trained on The Pile, where it recovers reproducible, mechanistically faithful circuits. We extract a CSS-only submodel of a 4-block transformer using ≈7% of the FLOPs of its published decomposition, and in a 12-block transformer we surgically ablate and rewire memorized sequences, with negligible side effects on other inputs.

tPD on toy and 4-block transformer models



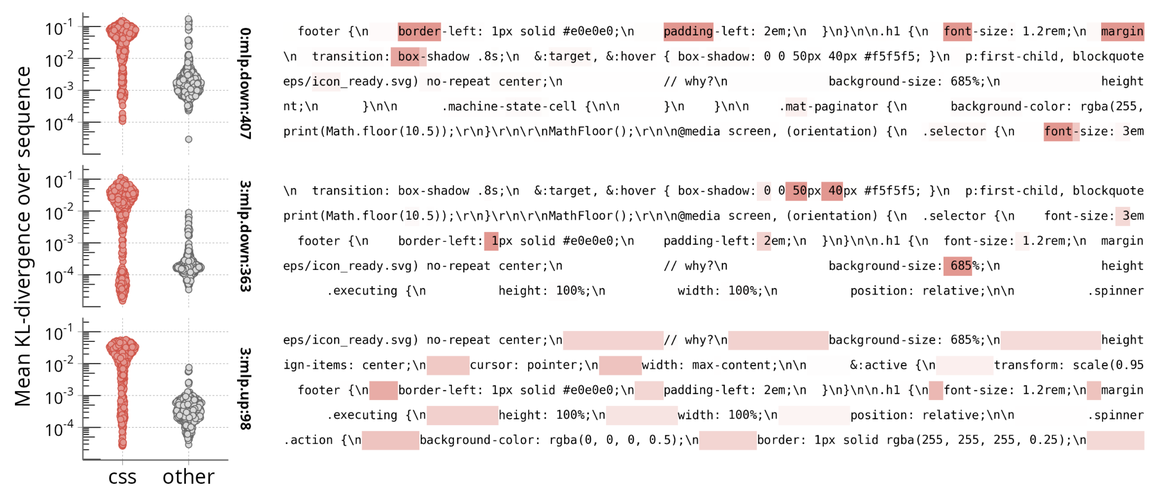
A: Causal importance patterns for the TMCC toy model's first matrix: full-data and targeted on 3 features. **B:** Mean-square error (MSE) of the reconstruction, as a proxy for decomposition success, for targeted decomposition of the TMCC toy model when a variable number of subcomponent slots are provided. **C:** MSE over training for the full-data and targeted decompositions of the TMCC. **D:** Cosine similarities of the U and V vectors between matched components across two random seeds for the joint numpy/pandas task. The horizontal bars represent the mean. The gray line and shaded area represent the mean ± standard deviation of a null distribution obtained by comparing the first decomposition to an untrained, randomly-initialized version of the second one. For comparison, full-data decompositions of the same model with different seeds (Bushnaq et al., 2026) yielded Mean Max Cosine Similarities of 0.48 and 0.51 for U and V vectors respectively. **E:** Pair-wise matches between the subcomponents obtained from decomposition on nested target sets: *import numpy as* ({np}), *import pandas as* ({pd}), or both at the same time ({np, pd}).

CSS-only submodel



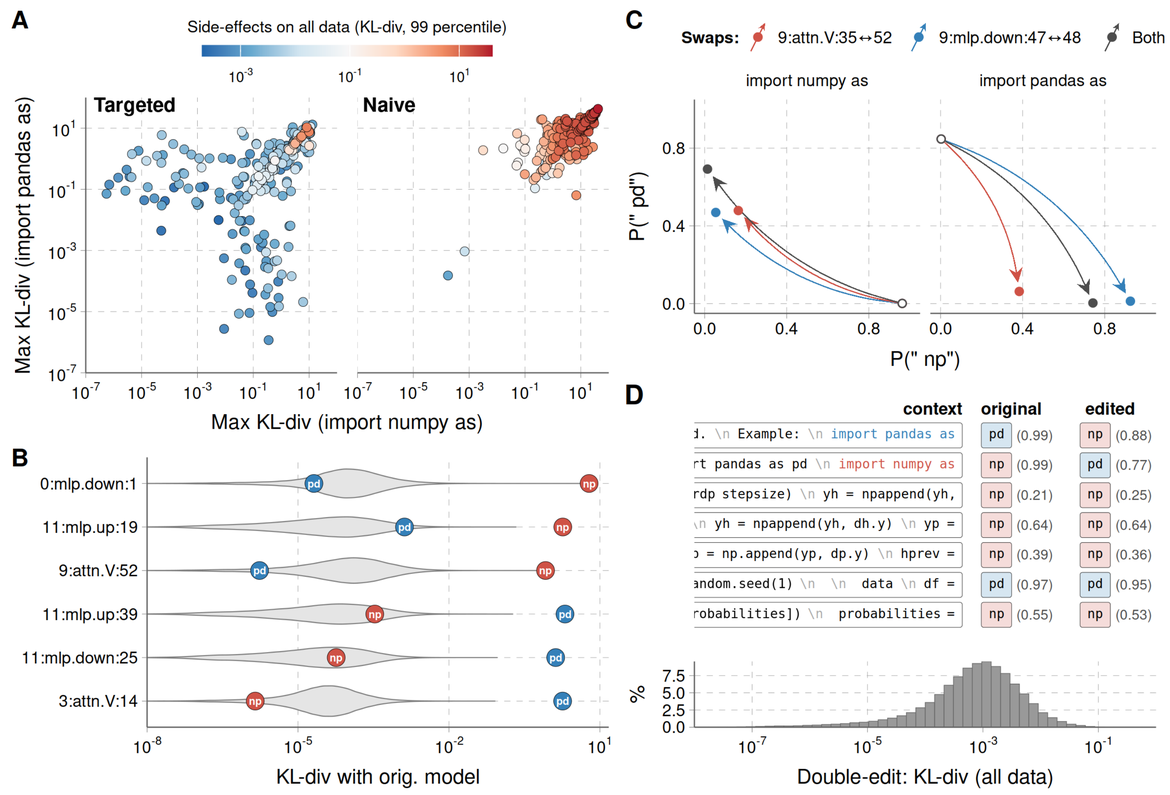
A: KL-divergence of a model made only of the subcomponents identified in the CSS-only decomposition, compared to the original model, on various languages. Lower values mean the outputs are closer to the original model's. **B:** Comparison of the original sequence, next tokens predicted by the original model, and next tokens predicted by the CSS-only model, on samples of text from CSS code, Python code, and English language. The prompt line is shifted left by one position so the real next token is aligned to the predicted next token. Highlights are the per-token KL-divergence between the original and CSS-only models clamped to $[0, 1]$.

CSS-specific subcomponents



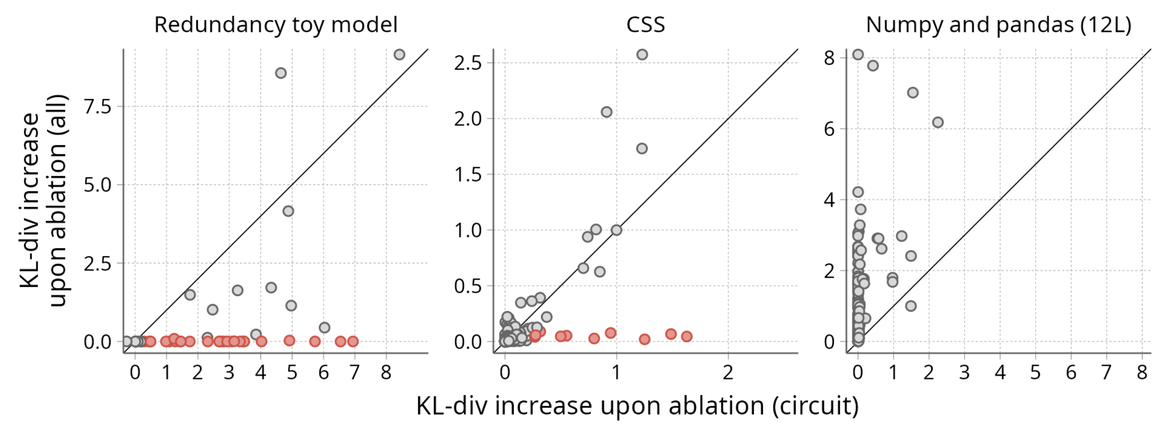
Three examples of CSS-specific subcomponents from the CSS-only decomposition. **Left:** KL-divergence compared to the original model's output when the subcomponent is individually subtracted from the model, for CSS code versus other languages. **Right:** Activation patterns of these components (displayed as the per-token KL-divergence when ablated, clamped to $[0; 1]$).

Erasing and rewiring knowledge in a 12-block transformer



A: Effect of ablating individual subcomponents obtained either with the tPD method introduced here ("targeted"), or with standard PD on the target data without non-target batches ("naive"). The x- and y-axes show the maximum KL(original‖ablated) over the two target sequences. Point color represents the 99th-percentile KL-divergence on >100k tokens of general Pile data. Higher values mean the ablated model's output deviate farther from the original outputs. **B:** Erasing the np/pd completions from a 12-block model. The "np" and "pd" points show KL(original‖ablated) on "import numpy as" and "import pandas as" respectively (higher values mean the ablation disrupts the model's behavior more on this input). Grey violins show the distribution of per-token divergences on general Pile data (lower values mean the ablation had less side effects). **C:** Effect of swapping the output directions of components from the "np" and "pd" sequences on the predicted next token probabilities. **D:** Behavior of the swapped model on sample contexts, and distribution of divergences over general Pile data (lower means less side effects).

Redundancy / completeness test



A coarse test for finding subcomponents that have a hidden redundancy among inactive/Δ components. The KL-divergence increase is defined as $\langle \mathrm{KL}(W_{\text{original}}x \| W_{\text{background without subcomponent}}x) - \mathrm{KL}(W_{\text{original}}x \| W_{\text{background}}x) \rangle$, where background is either the full model, or the "circuit" containing only active components from PD, and the angle brackets denote the average over inputs for which this subcomponent is active. Subcomponents that cause a KL increase below 0.1 on the full model and above 0.2 on the decomposed circuit are highlighted in red.