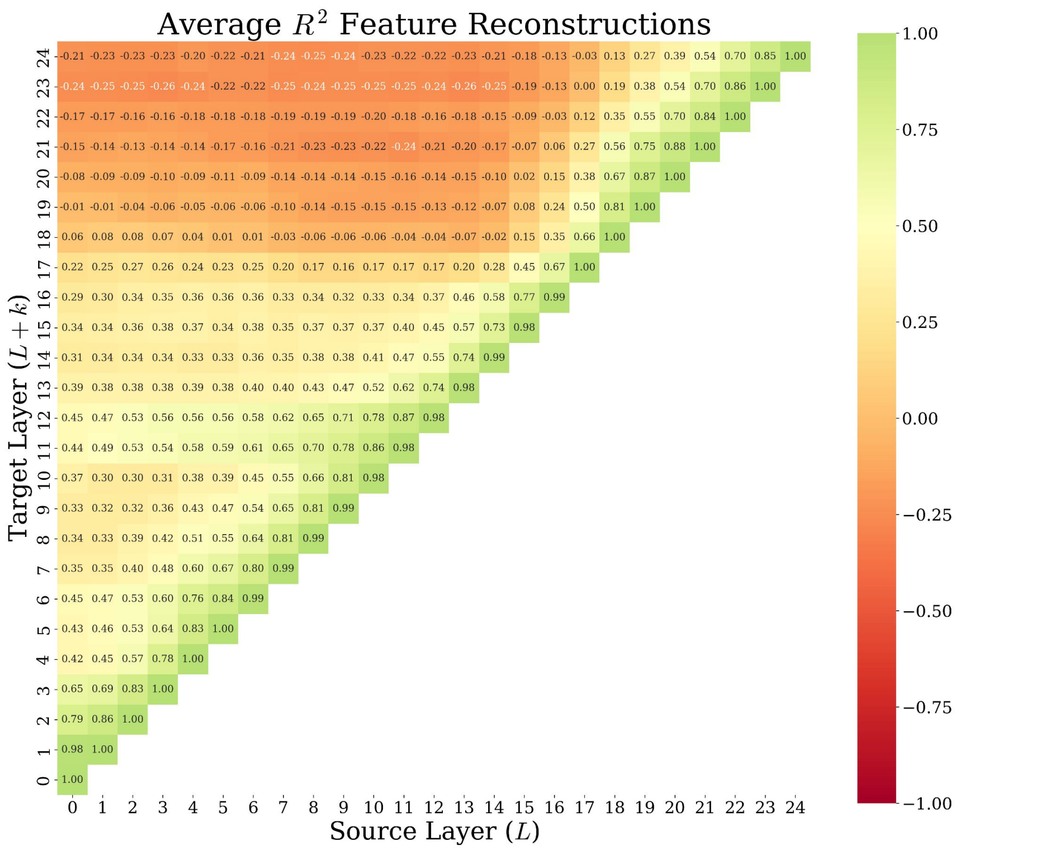
Predicting SAE Features Multiple Layers Ahead Reveals Information Flow Regimes

Activation Transport Operators

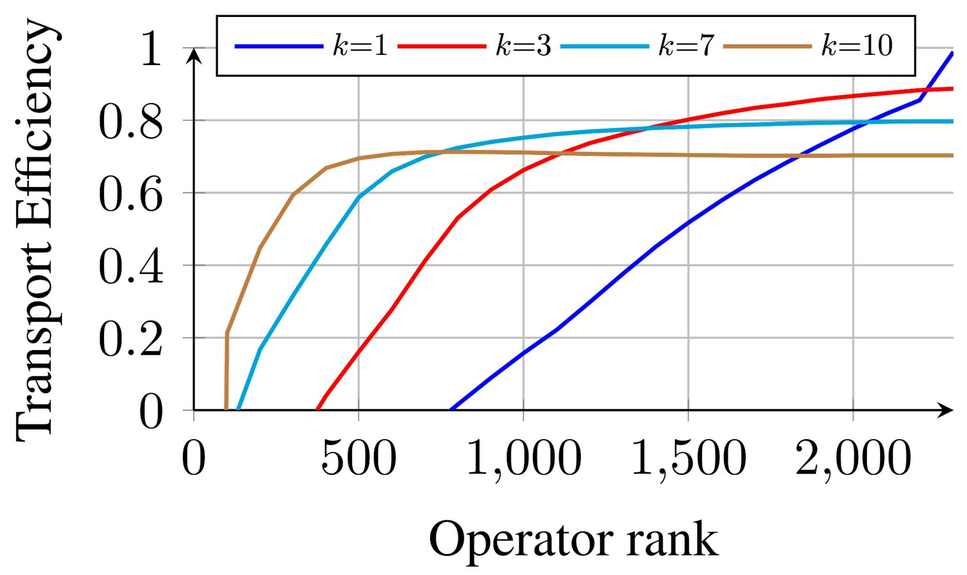
Background: The residual stream mediates communication between transformer decoder layers via *linear reads and writes* of non-linear computations. We have tools to *understand representations* and *locate behaviours*.

But can we characterise the information flow within the model?

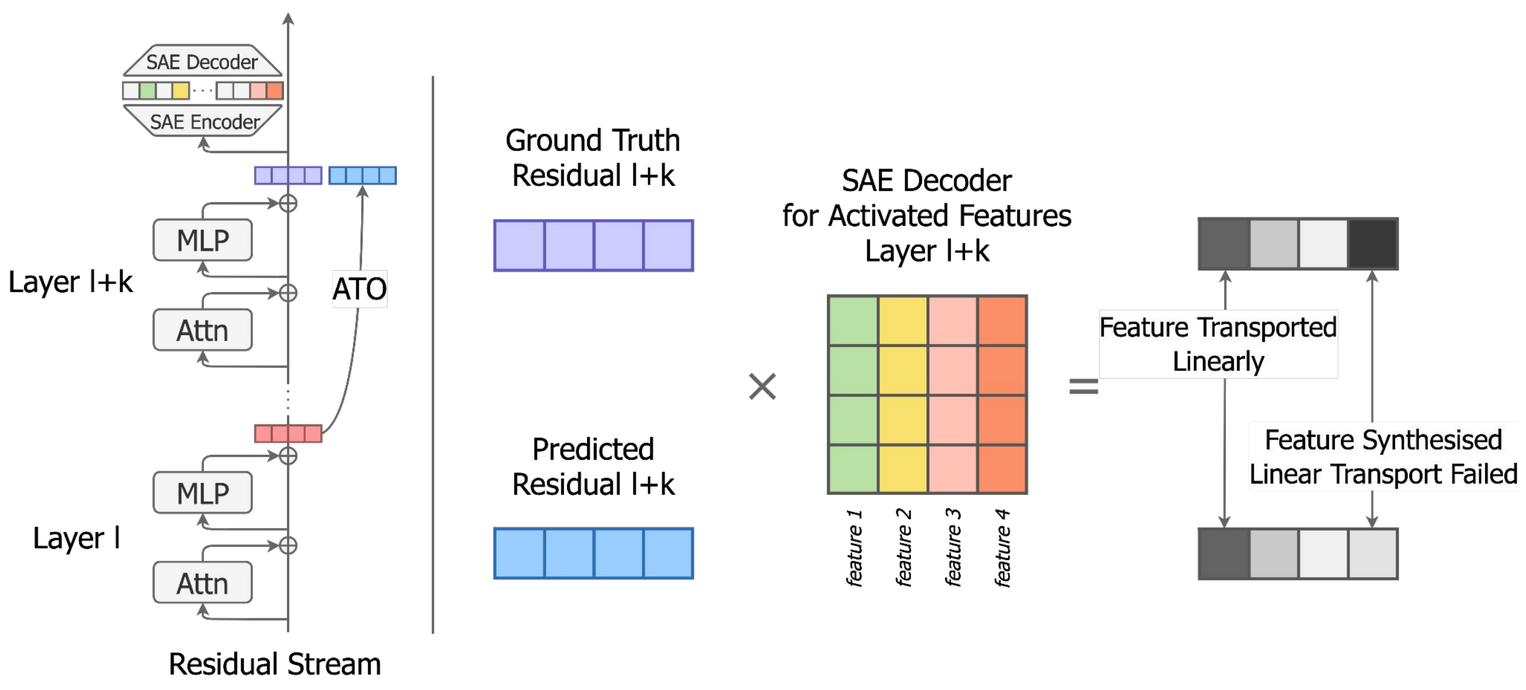
Result 1: **Linear transport deteriorates** with growing distance between layers. However, in deeper layers, we observe an **inflection point** after which the transport increases again.



Result 2: Growing operator rank for small leaps k results in a **near-linear increase of transport efficiency**. For larger leaps k, **transport efficiency plateaus early**, indicating a lower-rank mapping.



Method: Affine map from upstream residual stream to downstream SAE features



Method: Transport efficiency

$$R^2_{\text{ceiling}}(r,\, Y) \;=\; \frac{1}{d_{\text{model}}} \sum_{i=1}^{r} \rho_i^2$$

Where ρ_i^2 are the singular values of the whitened cross-covariance of upstream-downstream residual stream vectors. Then **transport efficiency** becomes:

$$\text{Eff} = \frac{\tilde{R}^2(r, \hat{Y}_T)}{R^2_{\text{ceiling}}(r, Y)}$$

Limitations:

- In this work, Linear Transport is studied only between layers for the same token position, ignoring attention-based effects. **We aim to explore cross-token transport in future works**.
- ATO does not distinguish between features that are transported from earlier layers and those that arise as their linear combinations. Furthermore, we aim to **expand our study to more models and datasets**.

PAPER

Andrzej **Szablewski*** and Marek **Masiak***

UNIVERSITY OF CAMBRIDGE

UNIVERSITY OF OXFORD

Mechanistic Interpretability Workshop, NeurIPS 2025