Thesis Final Presentation

April 11, 2024

Motivation

- Mechanistic interpretability requires understanding features
- Basis vectors aligned with neurons?
- Polysemantic neurons: activate across a broad range of seemingly unrelated contexts
- Toy Models of Superposition proposes an explanation
  - Predicts the geometry of features in superposition as coming from some energy minimization procedure
  - Geometry gives information about the learned correlation structure of features?
  - But lacks evidence in "real" models

Setup

- Pythia-1B-deduped, MLP layers
- Feature extraction based on "Finding Neurons in a Haystack: Case Studies with Sparse Probing"
- Hand-define several feature sets grouped on common categories of features (eg. text features, compound words, political party, occupation, etc.)
- Train logistic regression probes to predict the presence of that feature in a given token

Procedure

1. Train linear probes that test for the presence of a hand-defined feature in the activation space of a trained Transformer. The weights of these linear probes will be our feature directions.
2. Identify geometry that exists between groups of feature directions by looking at pairwise similarities between them.
3. Compare these **pairwise similarity matrices (PSMs)** to those that you would get if you were directly trying to minimize an energy function
4. Find the best match of the **target** PSM (comes from features) and **optimal** PSM (comes from the optimization procedure)

Energy Functions

- Energy functions $K(\cdot, \cdot)$:

$$\begin{aligned} \exp \text{ kernel: } & \exp\left(\frac{x_i \cdot x_j}{||x_i||_2 ||x_j||_2}\right) \\ -\log \text{ kernel: } & -\log ||x_i - x_j|| \\ \text{Riesz-}s \text{ kernel: } & \operatorname{sgn}(s) ||x_i - x_j||^{-s} \end{aligned}$$

- x_i, x_j are the two features that we are comparing
- Importantly, energy depends on norm of features for $-\log$ and Riesz

Similarity Functions

- Similarity functions $d_p(\cdot, \cdot)$:

$$\text{Cosine: } \frac{x_i \cdot x_j}{||x_i||_2 ||x_j||_2}$$

$$\text{RBF: } \exp\left(\frac{-||x_i - x_j||_2^2}{C}\right)$$

- x_i, x_j are the two features that we are comparing
- C is some constant (typically, $C = 200$)
- Used to generate the **optimal** and **target** PSMs

Feature Importance

- "Defined" in Toy Models paper
- Indicates how "useful" that feature is for the task
- More important feature $\implies$ allocated more "space"
- Important features should have low interference with other features
- Thus, associate each **optimal** feature x_i with an importance weight w_i

Joint Optimization

$$\sum_{ij} K(x_i, x_j)\tilde{w}_i\tilde{w}_j + \lambda_d \sum_{ij} (d_p(x_i, x_j) - [T_p]_{ij})^2$$

- K is an energy function
- x_i are our **optimal** features, each with dimensionality f
- d_p is a "similarity" function
- $[T_p]_{ij}$ is i, j-th entry of **target** PSM
- $\tilde{w} = \mathrm{softmax}(w)$.
- Optimize over importance weights $w \in \mathbb{R}^n$ and **optimal** features $x_i \in \mathbb{R}^f$
- Typically, $\lambda_d >> 1$

Re-optimization

- Feature importance weights w fixed (based on the value obtained in the previous step)
- x_i initialized to the output of the previous step, plus a perturbation
- Norms of x_i are restricted to be n_i, the norm of the output of the previous step
- Then, minimize the energy term and make sure the norms stay the same:

$$\sum_{ij} K(x_i, x_j)\tilde{w}_i\tilde{w}_j + \lambda_x \sum_i \mathrm{abs}(||x_i|| - n_i)$$

- This results in a final set of **optimal** features x_i that are local minima of the energy function, with some norm restrictions

Goals

1. Re-optimization should keep the distance to the PSM low $\Longrightarrow$ actual feature configuration is energy minimizing
2. If our target is a random PSM (ie. not coming from features), re-optimization should perturb it away $\Longrightarrow$ PSM minimization is different than energy minimization
3. Choice of energy function in a given layer should be robust to different feature sets
4. Learn something interpretable about structure of features

Text Features Dataset

Feature Name	Example Sentence
contains digit	style in following format?\n03.00, 02 .04\n\nif i Set \n
all digits	not differ from a normal population. From 5 to 30 months, there appeared a significant probability of intellectual
contains capital	housing in University Suites and University Apartments. More recently, Alpha Psi Lambda, a
leading capital	() is a rural locality (a village) in Sem izerye Rural Settlement, Kaduys
all capitals	LIMITED TO, THE\n* IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR\n* PURPOSE ARE DISCLAIM
contains whitespace	lines of evidence have supported the idea that capping protein blocks the barbed end of actin filaments,
has leading space	.56).The maxillary sinus was most commonly involved , followed by the nasal cavity (51%
no leading space and lower-alpha	, currentForecast = currentPageViewController.fore cast ,\nlet currentIndex = indexOf
contains all whitespace	(input.hasNextLine())\n}\n␣␣␣␣System.out.println(
is not alphanumeric	\n# yargs command completion script\n#\n # Installation : {{app_path}} completion
is not ascii	az\nKırkkaşık\nK ı rmataş\nKoçbaba\n

Feature Extraction Results - Text Features

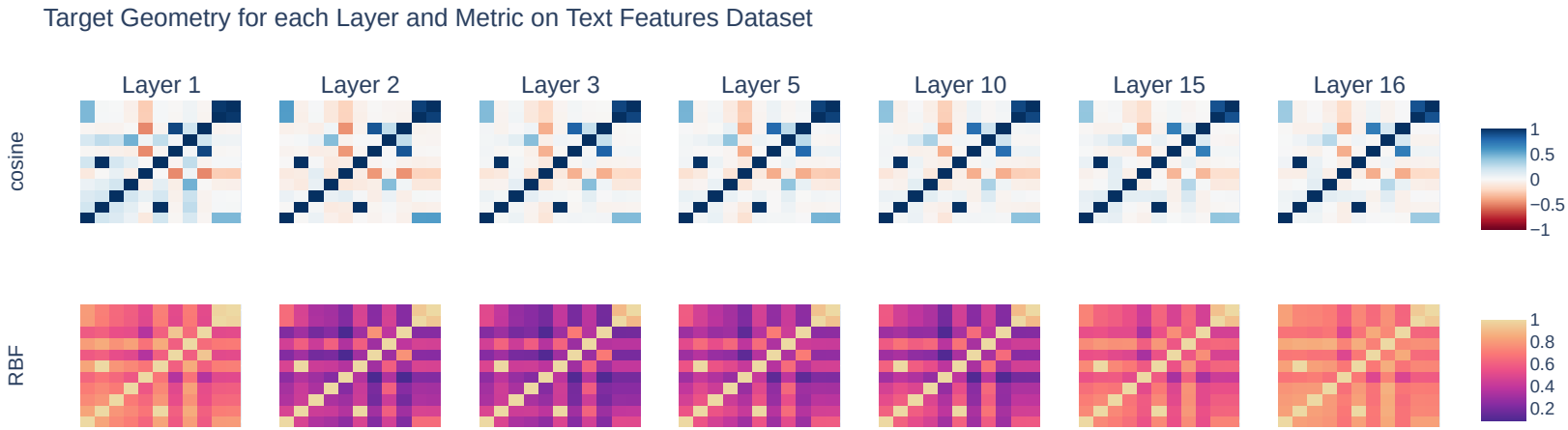


Figure: The x and y axes both correspond to different particular features. For example, the feature pair associated with x=1, y=2 here is (contains_digit, is_not_ascii).

Feature Extraction Results - Permuted Text Features

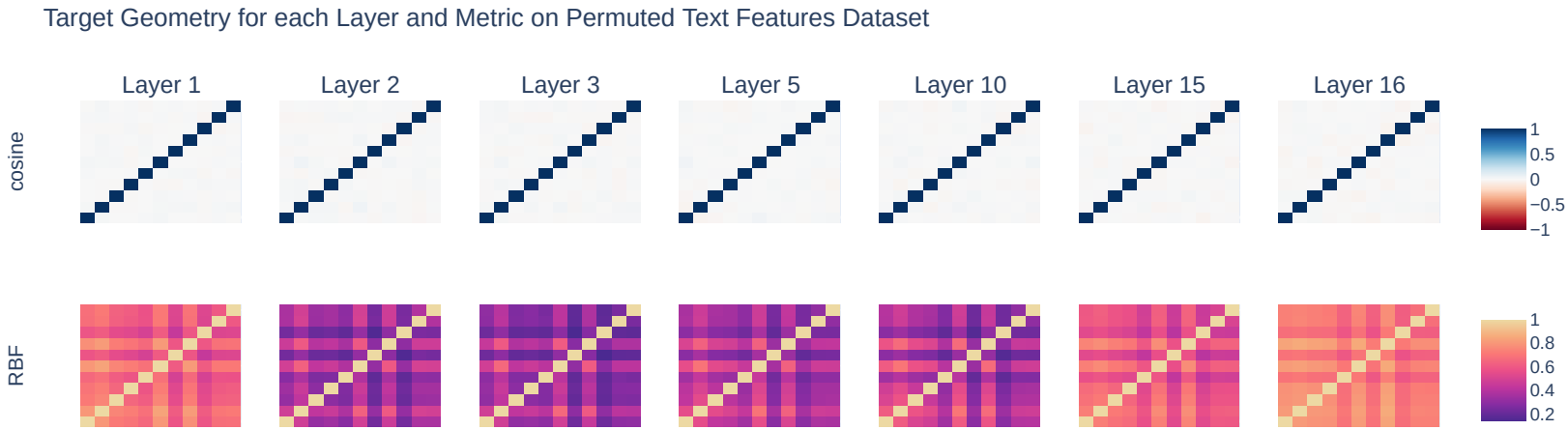


Figure: PSMs of the text features but each feature vector is randomly permuted (thus geometric structure destroyed). This makes features orthogonal (see "cosine").

Wikidata Athlete Occuptaion Dataset

Possible Value	**Positive Sentence Example**
baseball	Playing baseball professionally, sparked by reportedly meeting Babe Ruth
association football	Rutherford's colleague, Danish theorist Niels Bohr
basketball	Most asked questions of his wedding day party. Was LeBron James
ice hockey	Based on the popular comic strip by Charles M. Schulz
American football	1976 election, a Washington elector pledged to President Gerald Ford

Feature Extraction Results - Occupation Athlete Features

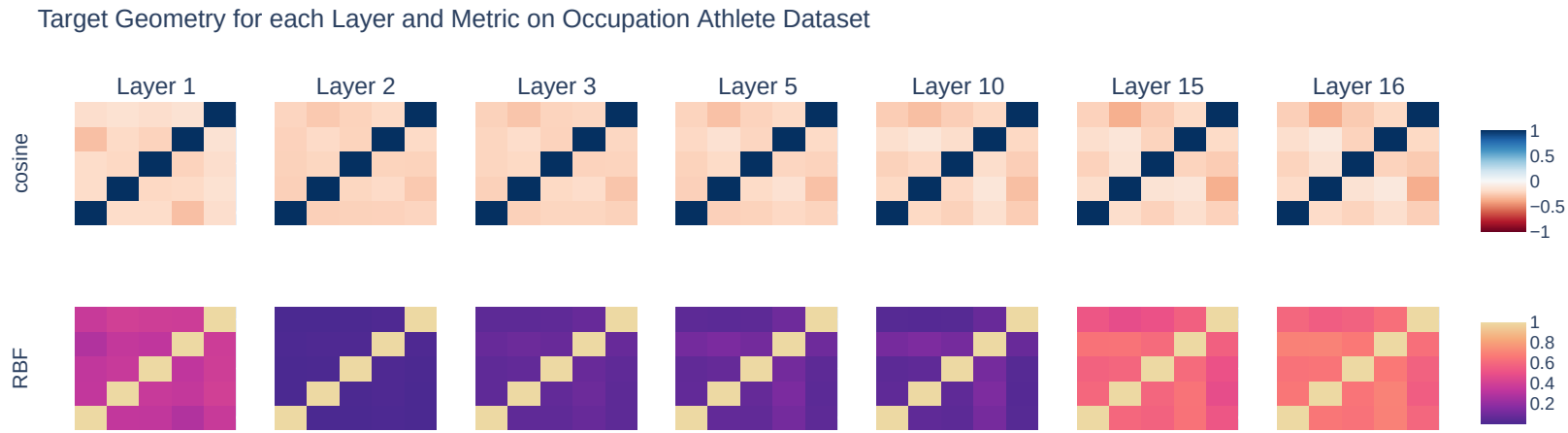


Figure: Geometric structure is similar in all layers. The changes in the overall scale in the RBF plot is due to changing magnitude of features in different layers.

Energy Minimization Experiments

- Ran procedure on both the "athlete occupation" and "text features" dataset.
- Searched over dimensionality of features, regularization parameters (λ_d, λ_x), energy functions $(-\log, \exp$, Riesz-s), and similarity functions (cosine, RBF)
- Two trials: perturbation scale 0.1, perturbation scale 0.01
- Use conjugate gradient method for joint optimization step, Powell's method for re-optimization step

10% Perturbation Results - Text Features

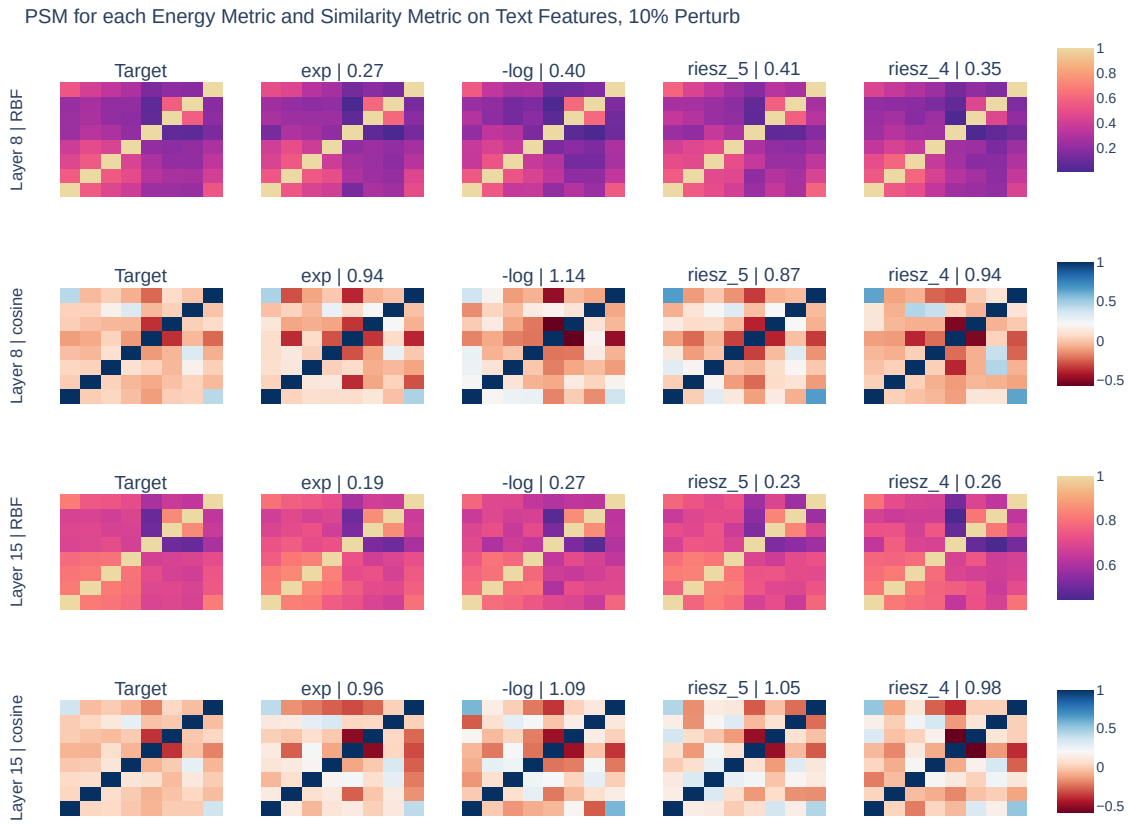


Figure: Number in title is Minkowski distance. Not comparable between similarity functions since the units are different. $\exp$ kernel performs best on RBF, and Riesz is best for cosine on layer 8, while $\exp$ is best on layer 15.

1% Perturbation Results - Text Features

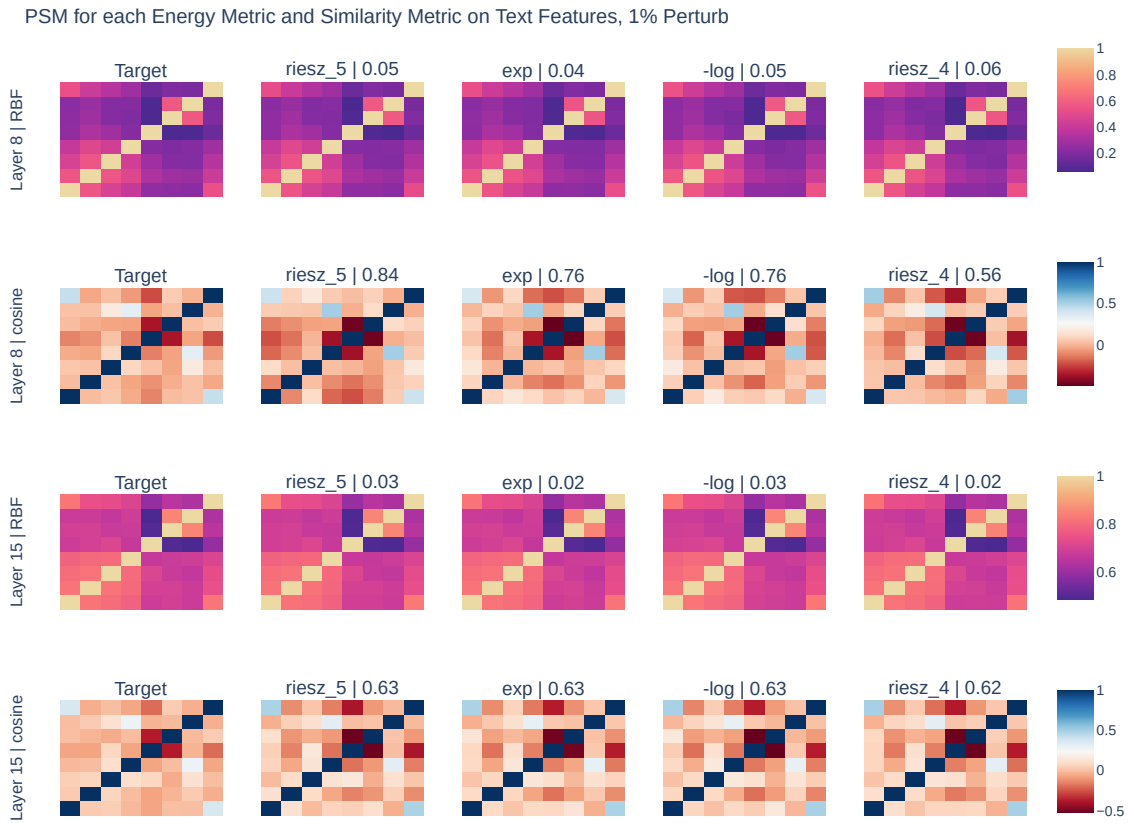


Figure: Reducing perturbation makes the match significantly better for RBF, slightly better for cosine. Indicates that the minima in the energy function could be very narrow.

10% Perturbation Results - Athlete Occupation

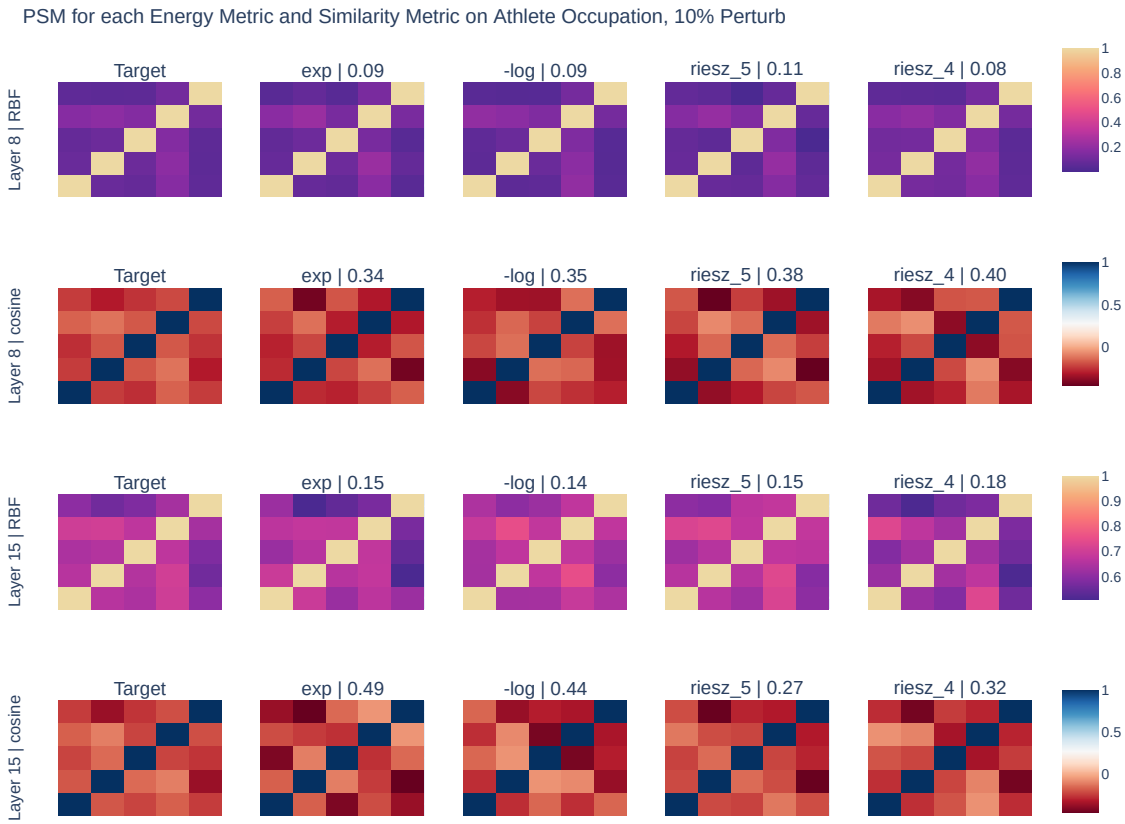


Figure: Matching is overall much better for athlete occupation. $-\log$ is now best for RBF, and for cosine layer 15 best is Riesz, and cosine layer 8 is $\exp$ best.

1% Perturbation Results - Athlete Occupation

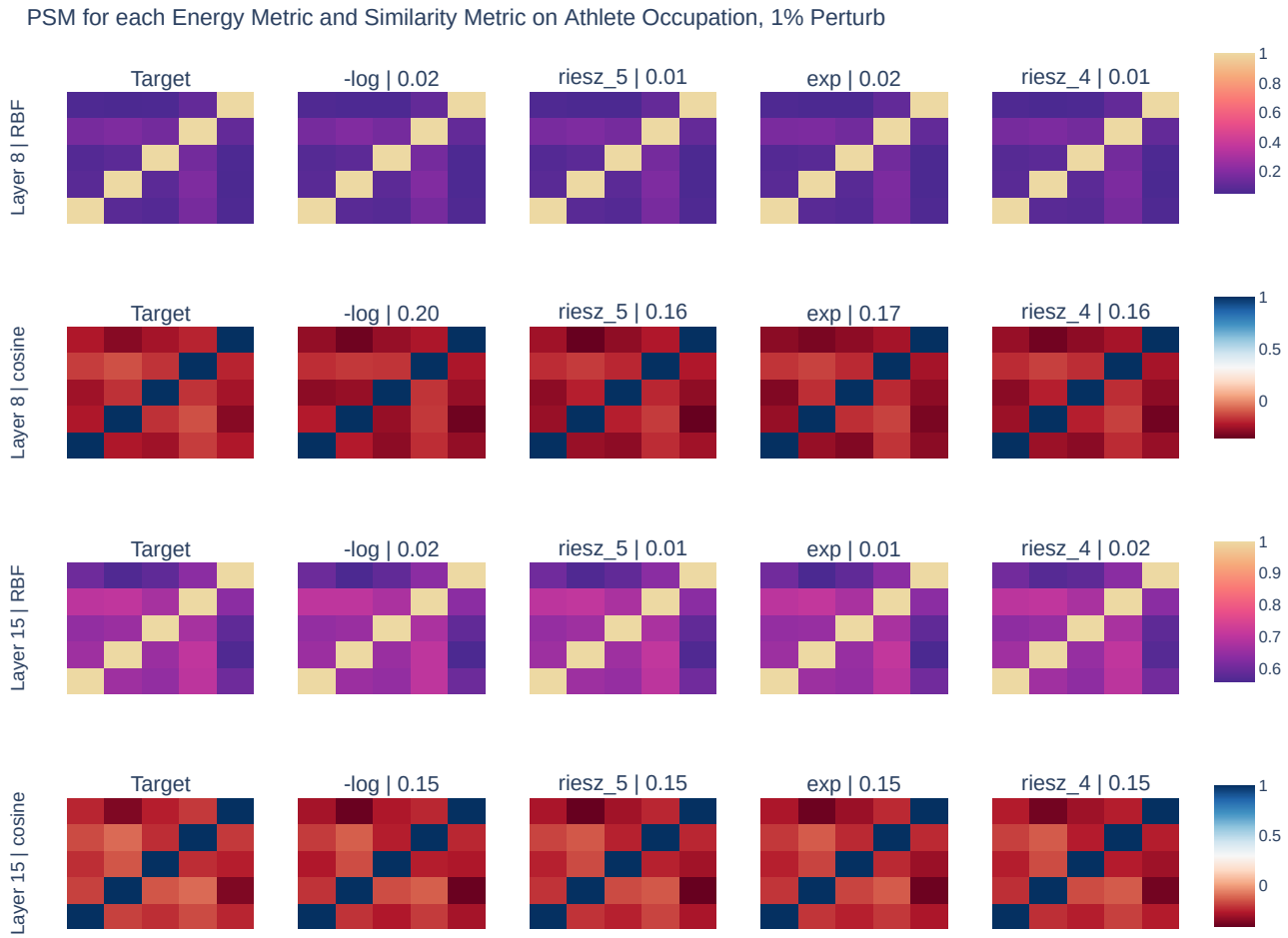
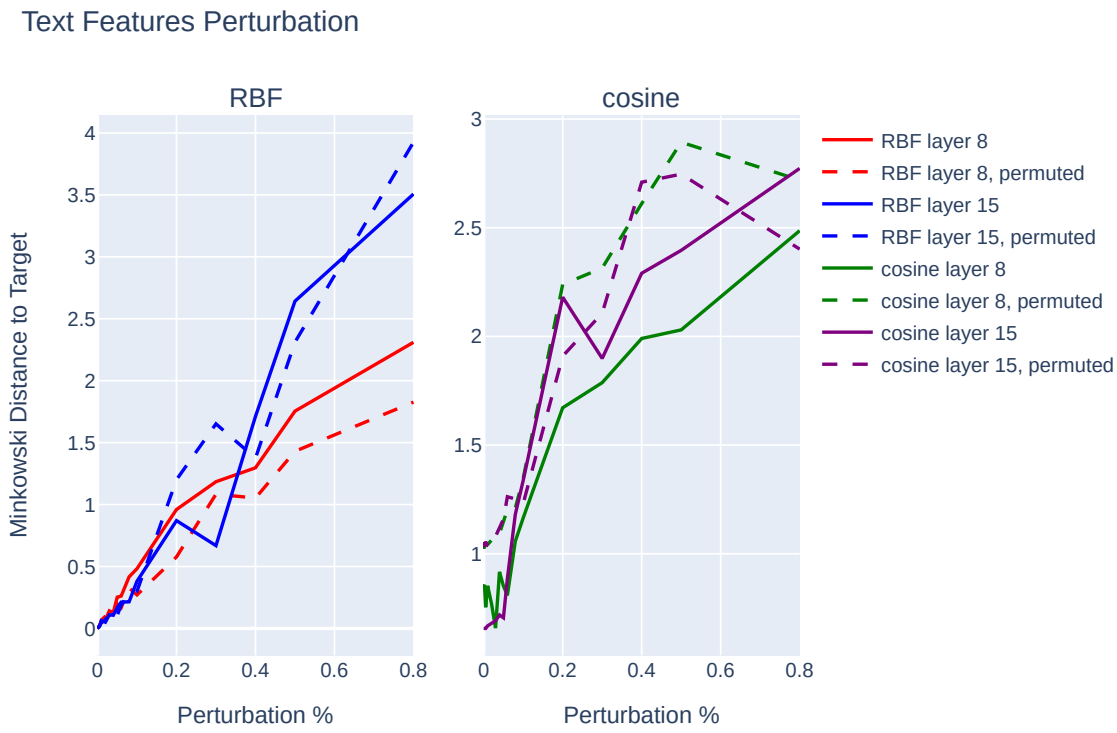


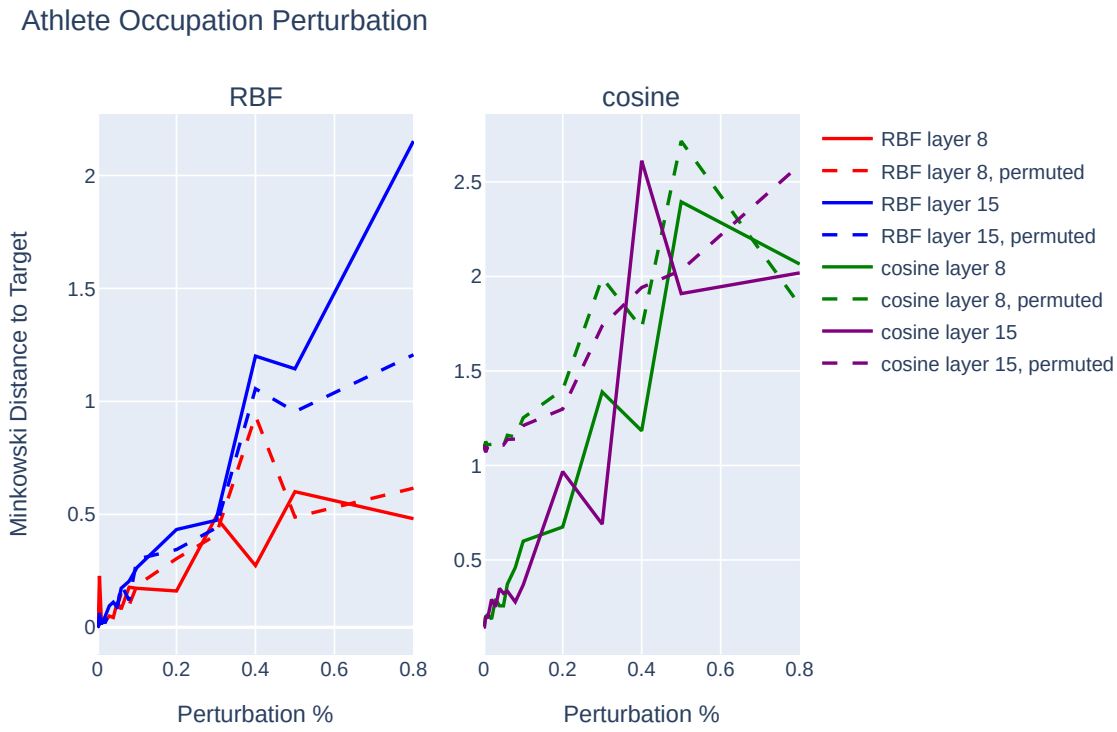
Figure: Similar to text features, minima might be narrow.

Perturbation Amount Results - Text Features



- If perturbation scale is σ, then the perturbation applied is an isotropic Gaussian with standard deviation σn_i.
- Smaller σ, smaller distance.
- Permuted does worse for cosine, but not for RBF.

Perturbation Amount Results - Athlete Occupation



- Permuted similarly does worse for cosine, but not for RBF.
- Minimizing distance to target PSM for RBF similarity $\implies$ minimizing energy
- Minimizing distance to target PSM for cosine similarity $\not\implies$ minimizing energy

Goals, revisited

1. Re-optimization should keep the distance to the PSM low
   - Yes
2. If our target is a random PSM (ie. not coming from features), re-optimization should perturb it away
   - Yes, but only for cosine
3. Choice of energy function in a given layer should be robust to different feature sets
   - No
4. Learn something interpretable about structure of features
   - Mostly no, feature importance weights were almost completely ignored in all cases (ie. set equal to each other).

Results, summary

- Simplexes for abstract features seem to exist
- Geometry is preserved throughout layers, with some caveats
  - Changing feature norms throughout layers may affect this
  - Some small differences in early/late layers
- Actual features do appear to be in local minima of energy functions
- Cosine similarity has advantage of its minima not automatically being minima of energy functions
- Feature importance weights are not used by the procedure

Future Work / Questions

- Deeper investigation into energy minimization needed
  - Very sensitive to choice of optimizer and regularization
  - Minimizing distance to PSM $\iff$ minimizing an energy function?
  - What happens for Transformers with randomized weights?
- Larger scale: more feature sets, models, etc.
- Simplexes in embedding matrices?
- Use geometry to make predictions of learned correlation structure of features?