Mathematical Analysis of Standardizing & Angular Measures

Nathaniel Charest

US EPA

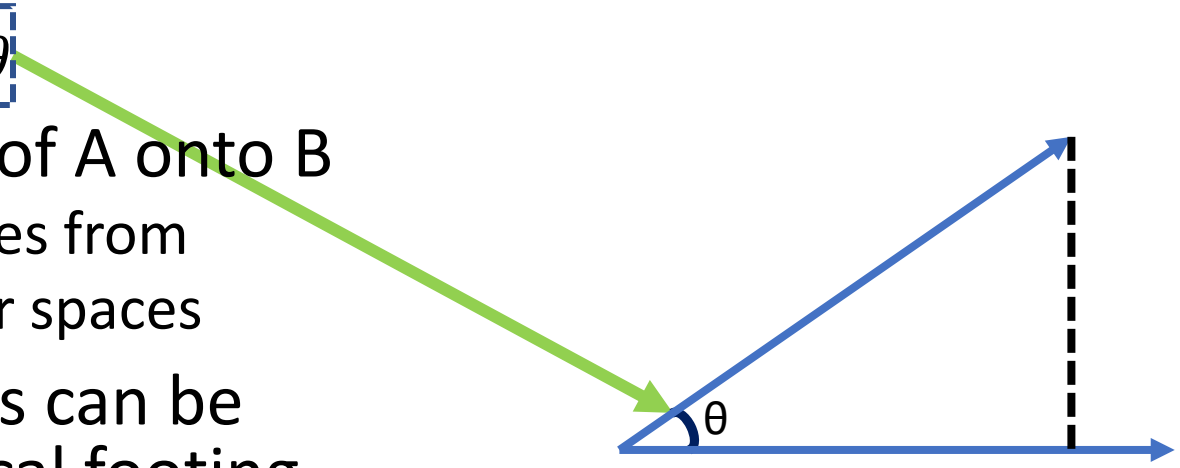
Center for Computational Toxicology and Exposure

Research Triangle Park, NC

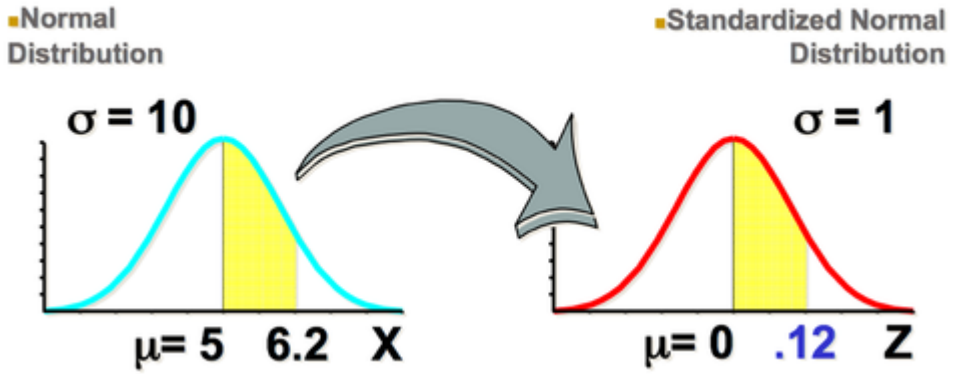
Disclaimer: The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

Cosine Similarity: An Angular Perspective

- The Dot Product
  - $\boldsymbol{A} \cdot \boldsymbol{B} = \sum A_i B_i = |A||B| \cos\theta$
- Geometrically, the projection of A onto B
  - This is where the intuition comes from
  - This intuition is based on vector spaces
- It also assumes the descriptors can be treated on similar mathematical footing, (i.e. a change, Δ, of one variable, i, is weighted as important as a change of the other variables, j)
  - It is *really* hard to reconcile measures and counts because of this



Cosine: Standardizing Changes The Space



Most of these values are positive

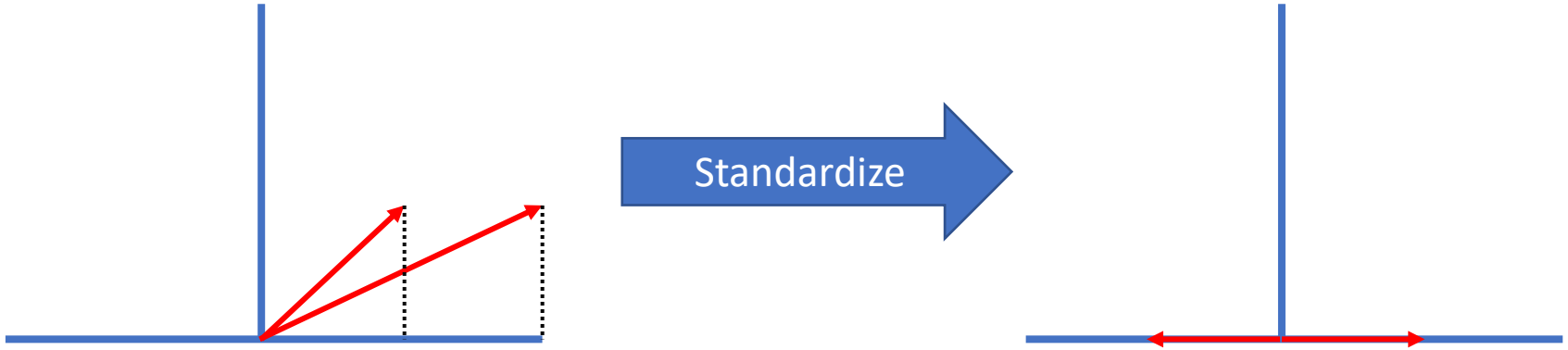
Half of these values are now negative

Standardizing was designed to make comparing normal distributions cleaner.

Cosine similarity does not compare normal distributions.

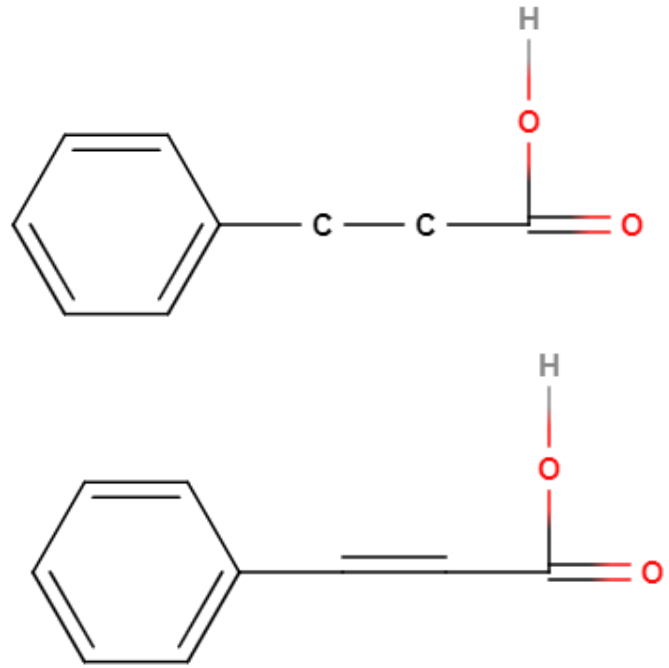
Cosine similarity compares vector representations.

Vector representations embed meaning in the positive and negative quality of their values.



Cosine: Standardizing Changes The Space

- Consider encoding:
  - (benzene fragment, carboxyl fragment, polarizability measure)
- Molecule A
  - (1, 1, 2) -> Standardize -> (0, 0, -1)
- Molecule B
  - (1, 1, 4) -> Standardize -> (0, 0, 1)
- $A \cdot B = -1$ – these molecules are opposites



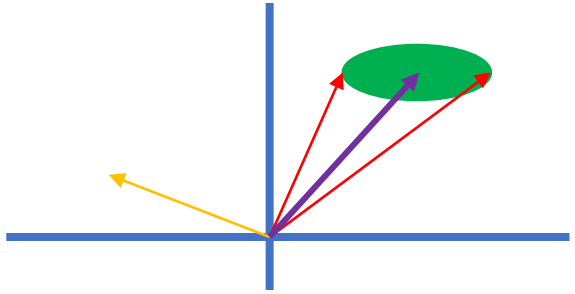
Cosine

- Standardizing warps the vector space so the training set is the entire 'world'. Watch what happens when we *don't* standardize.
- A = (1, 1, 2) B = (1, 1, 4)
- |A| = $(6)^{1/2}$, |B| = $(18)^{1/2}$, A·B = 10
- $Cosine(\theta) = \frac{10}{\sqrt{6}*\sqrt{18}} = 0.9622$ – "Not so different, you and I"
- In this case standardizing messed up the meaning of the inner product by making some values negative when their *implicit chemical meaning* is directly related to them being positive.
- Because our two very obviously related compounds were on opposite ends of the 'training space', our measure became highly subjective and intuitively nonsensical!

When Could This Matter?

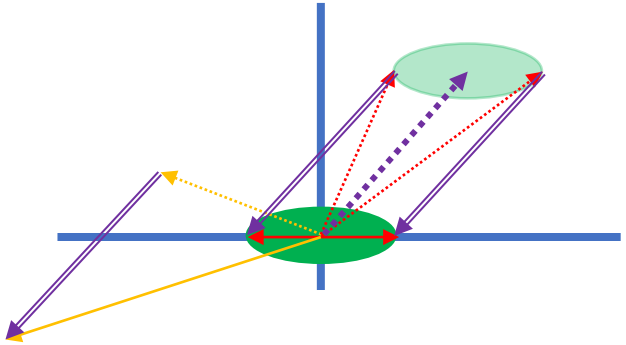
- When the coverage of your training set is low, this exacerbates the issue of comparison outside the range of coverage

The range of the training domain is reasonably small in objective chemical space.



The orange query is quite distant from the training domain. Its relationship to the training range is generally orthogonal or oblique.

The range of the training domain defines 'opposite' after standardization.
The original outlier now looks much closer to members of the training set than it did in objective space!

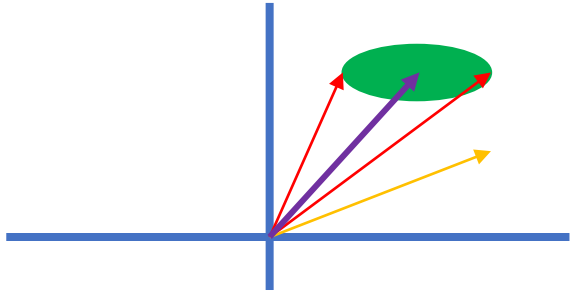


Shouldn't a robust measure of chemical similarity be objective?

When Could This Matter?

- The effects differ based on the query's original relationship to the training domain

The range of the training domain is reasonably small in objective chemical space.

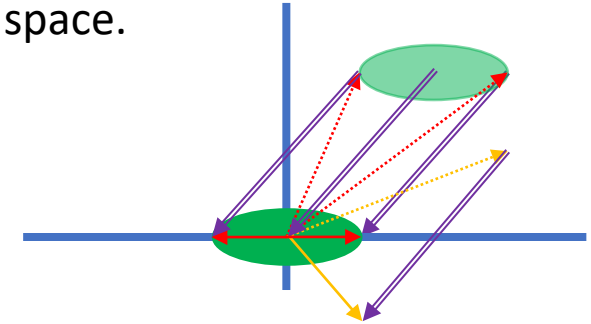


This query reads as reasonably close to the ellipsoid tangent.

Standardize

The range of the training domain defines 'opposite' after standardization.

The original outlier has gained many neighbors it did not have in objective, intuitive space.



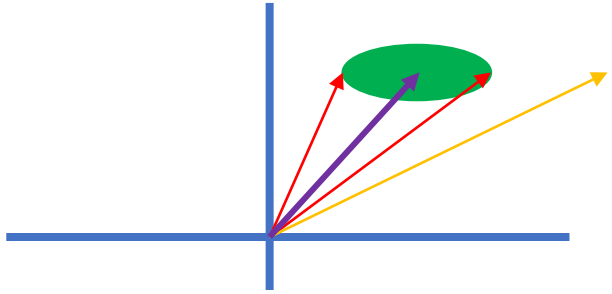
Additionally, it is further away from its original similar species.

The alteration of neighborhoods from intuitive space is <u>questionable at best</u>

When Could This Matter?

- Existing along the mean of the training set creates eyebrow raising circumstances.

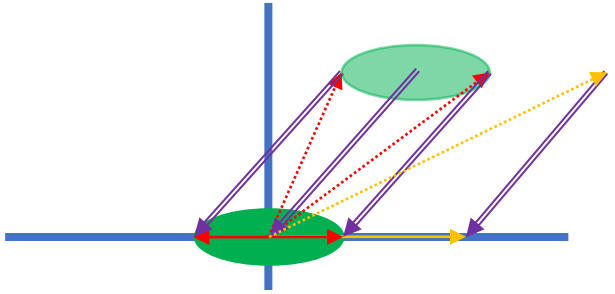
The range of the training domain is reasonably small in objective chemical space.



The orange query is an outlier along X but a mean value of Y.

The range of the training domain defines 'opposite' after standardization.

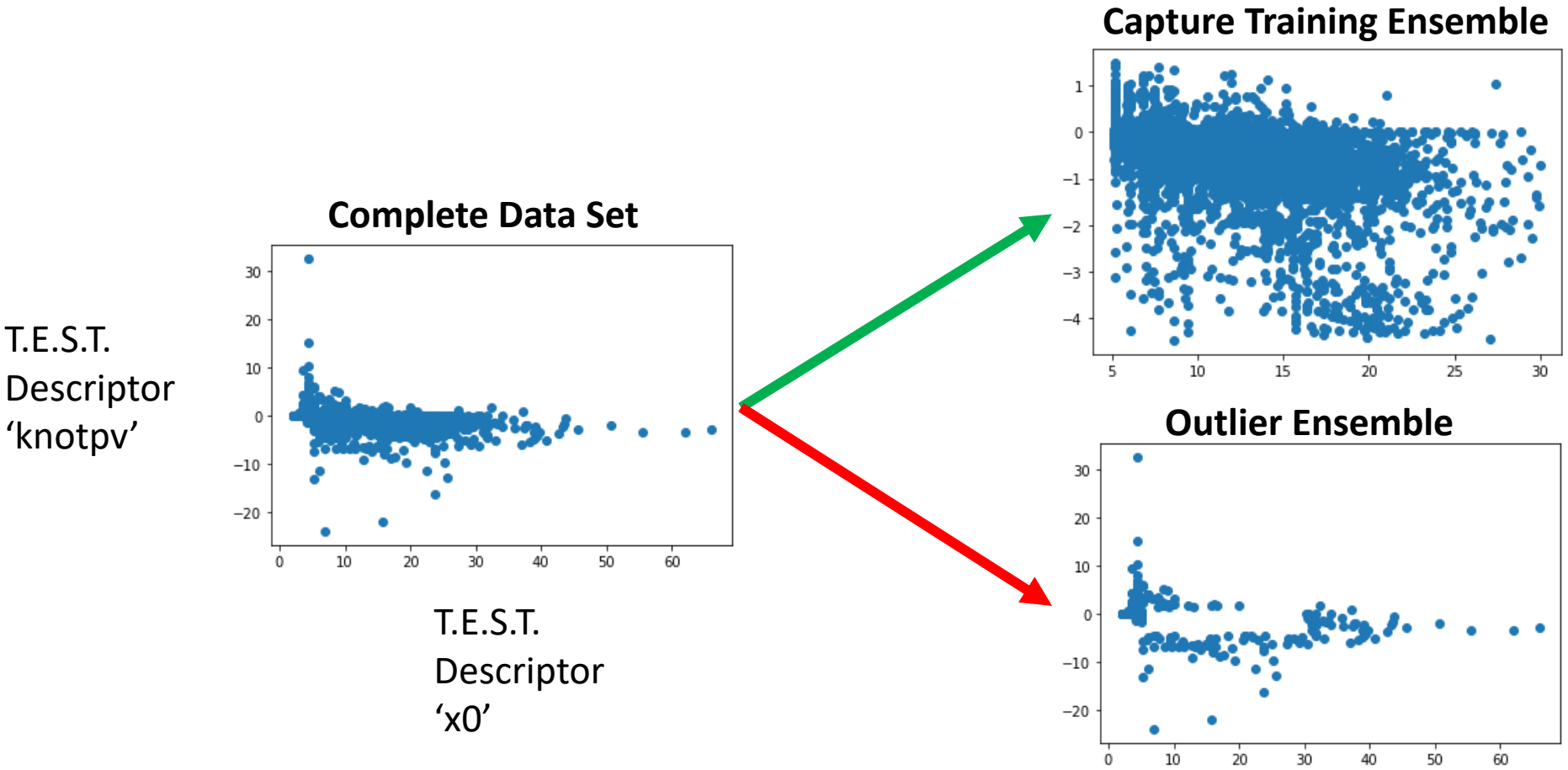
The transformation makes the outlier *identical* to the apogee of training domain.



So that's weird, but is it meaningful?

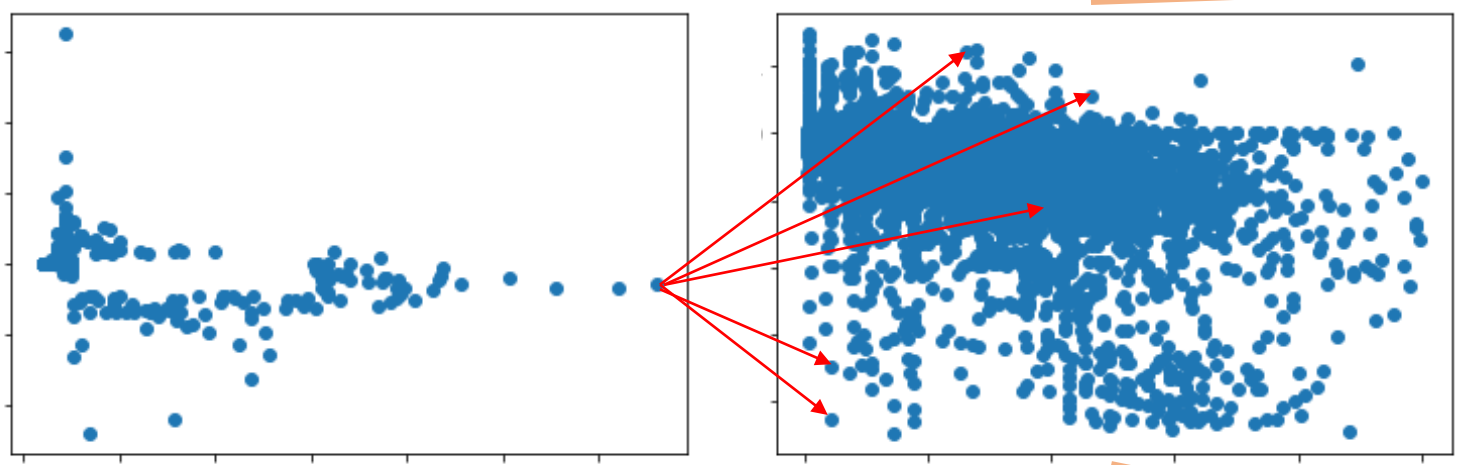
An Experiment

- Let's extract a subset domain from a real dataset. Here's data from our melting point chemical training/test set.

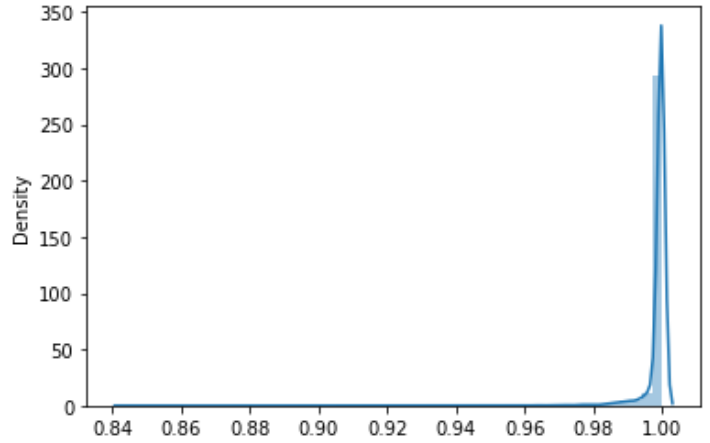


An Experiment

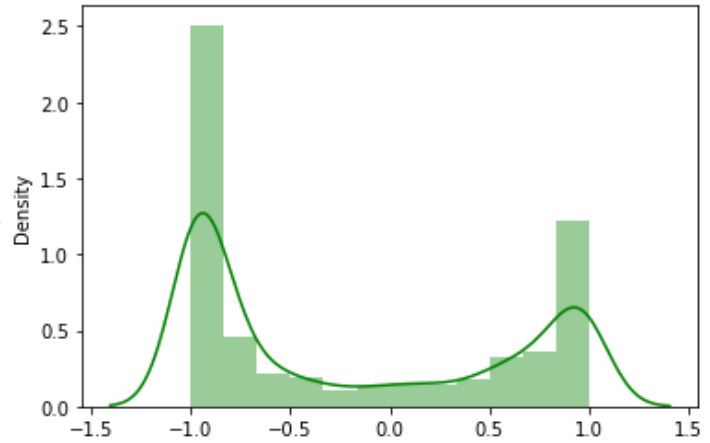
- Compute Objective Cosine vs. Standardized Cosine between outlier and train set.



Objective Cosine

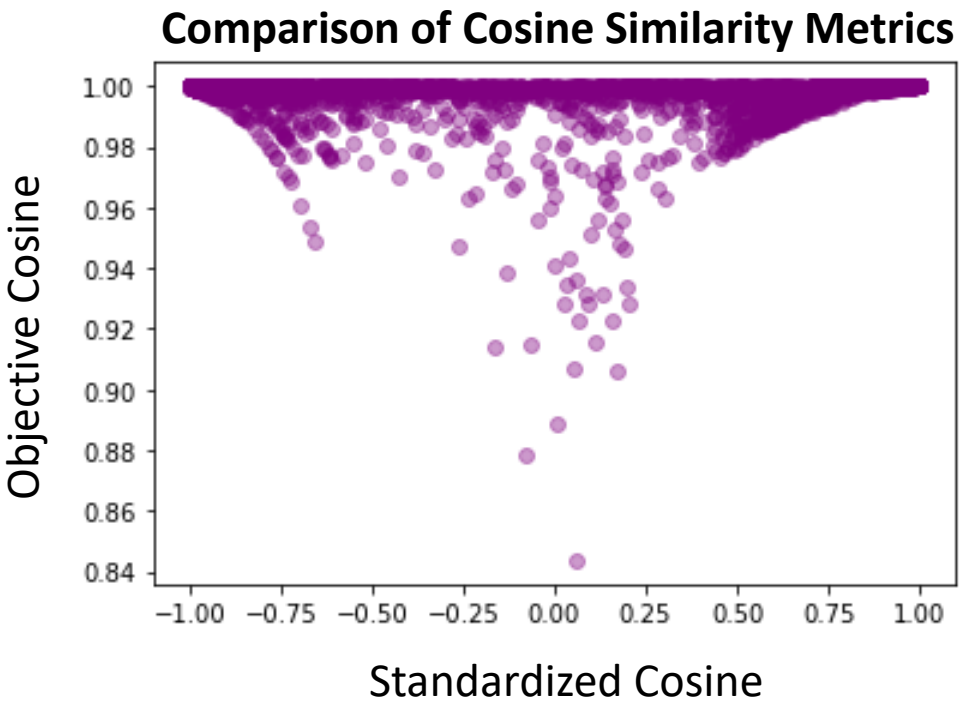


Standardized Cosine



An Experiment

- Explicitly Compare Objective Cosine & Standardized Cosine



- **There is information here**
  - But if the goal is direct comparison of how similar the chemistry or physics of species are, then this is a mess
- **Empty Mean**
  - The closer to the mean a vector was, the closer to zero is the standardized inner product for any comparison. The intuition of positive similarity is gone.

Potentially Bad News?

- Buried in all this is a worrying possibility
  - These manipulations could *improve model statistics* while *harming generalizability*
- The spatial warp *exaggerates differences within the training set*. This makes the new space easier to 'learn' and thus easier for the model to perform within it.
- If the test set was pulled at random from the training set, it likely belongs to the same region of chemical space and could benefit from the warp. The statistics might improve!
- But any future query from outside the favored domain will permanently have its objective relationship misrepresented. The applicability domain is much harder to compute or even define.
- **Bottom line:** standardizing before using cosine *builds in* a fundamental dependence on the training set. This can be considered *a form of overfitting*.