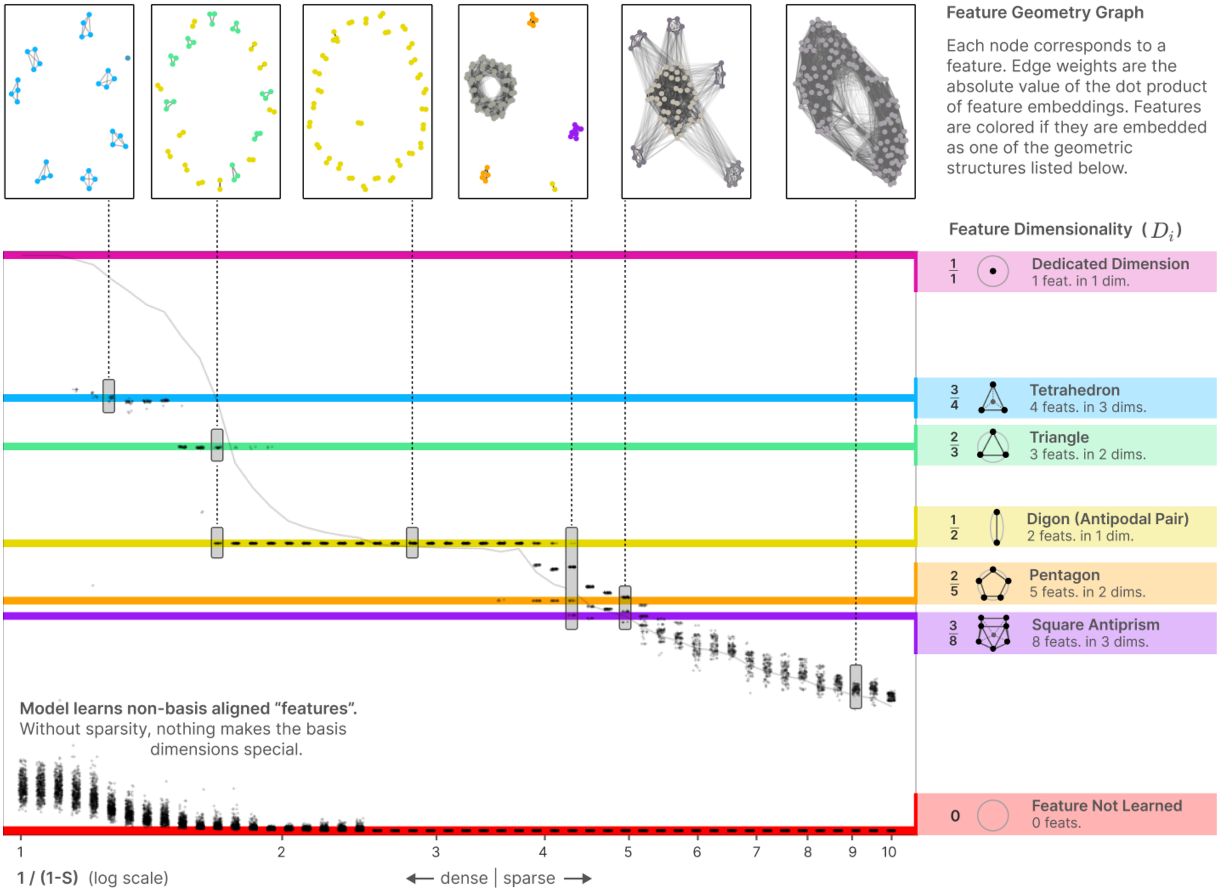

Feature Geometry Graph

Each node corresponds to a feature. Edge weights are the absolute value of the dot product of feature embeddings. Features are colored if they are embedded as one of the geometric structures listed below.

What is going on with the points clustering at specific fractions?? We'll see shortly that the model likes to create specific weight geometries and kind of jumps between the different configurations.

In the previous section, we developed a theory of superposition as a phase change. But everything on this plot between 0 (not learning a feature) and 1 (dedicating a dimension to a feature) is superposition. Superposition is what happens when features have fractional dimensionality. That is to say – superposition isn't just one thing!

How can we relate this to our original understanding of the phase change? We often think of water as only having three phases: ice, water and steam. But this is a simplification: there are actually many phases of ice, often corresponding to different crystal structures (eg. hexagonal vs cubic ice). In a vaguely similar way, neural network features seem to also have many other phases within the general category of "superposition."

WHY THESE GEOMETRIC STRUCTURES?

In the previous diagram, we found that there are distinct lines corresponding to dimensionality of: ¾ (tetrahedron), ⅔ (triangle), ½ (antipodal pair), ⅖ (pentagon), ⅜ (square antiprism), and 0 (feature not learned). We believe there would also be a 1 (dedicated dimension for a feature) line if not for the fact that basis features are indistinguishable from other directions in the dense regime.

Several of these configurations may jump out as solutions to the famous Thomson problem. (In particular, square antiprisms are much less famous than cubes and are primarily of note for their role in molecular geometry due to being a Thomson problem solution.) As we saw earlier, there is a very real sense in which our model can be understood as solving a generalized version of the Thomson problem. When our model chooses to represent a feature, the feature is embedded as a point on an m-dimensional sphere.