Supplemental Figures

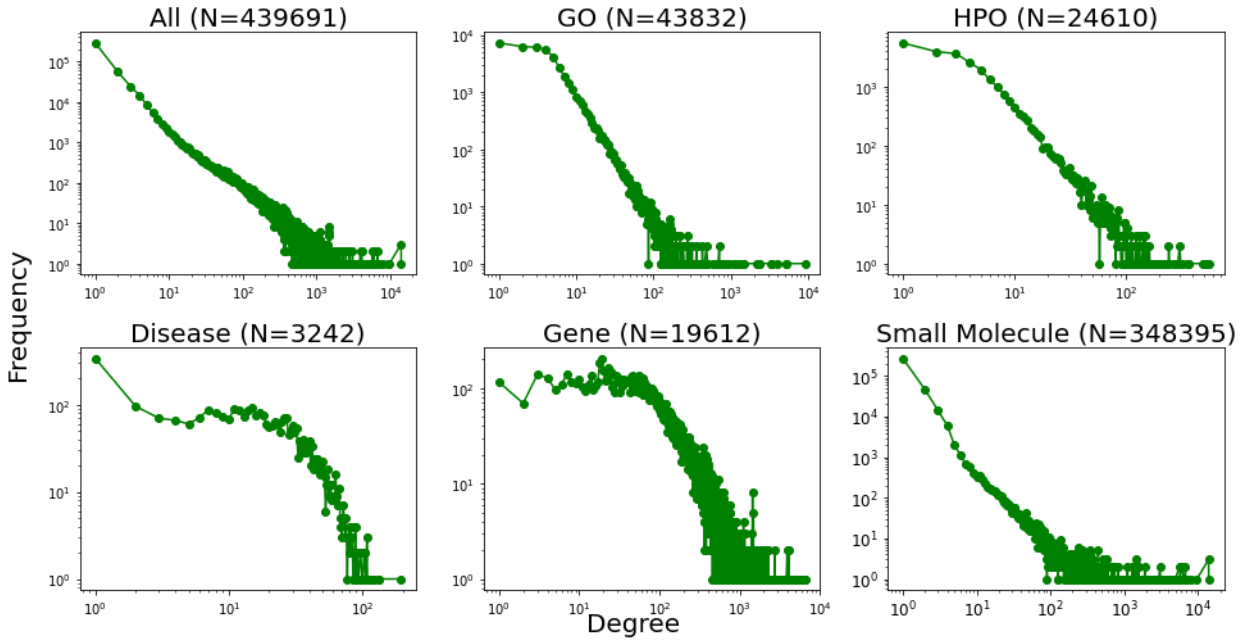

Figure S1 The degree distribution of various node types within the rare disease knowledge graph. Both y and x axes show the raw data values on a log-scale.

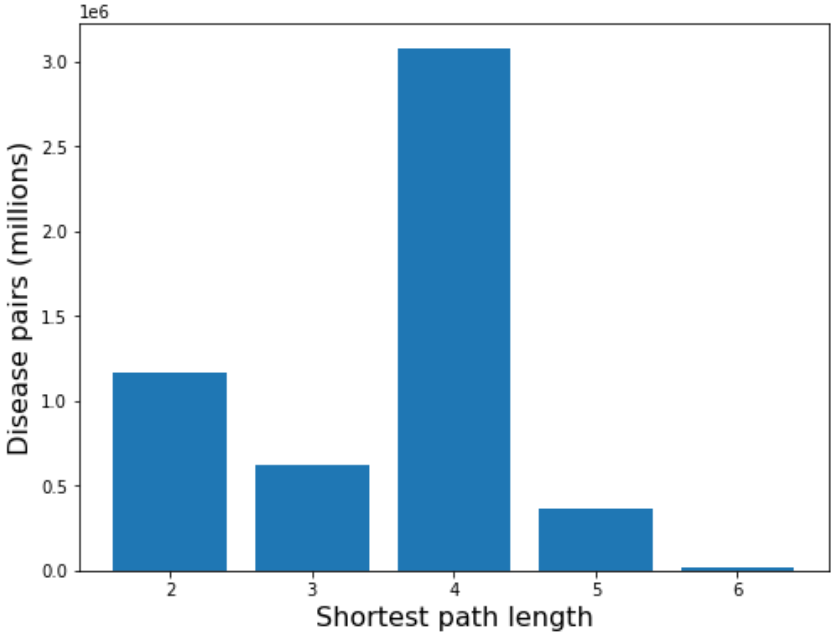

Figure S2 The distribution of shorest path lengths between all disease pairs within the rare disease knowledge graph

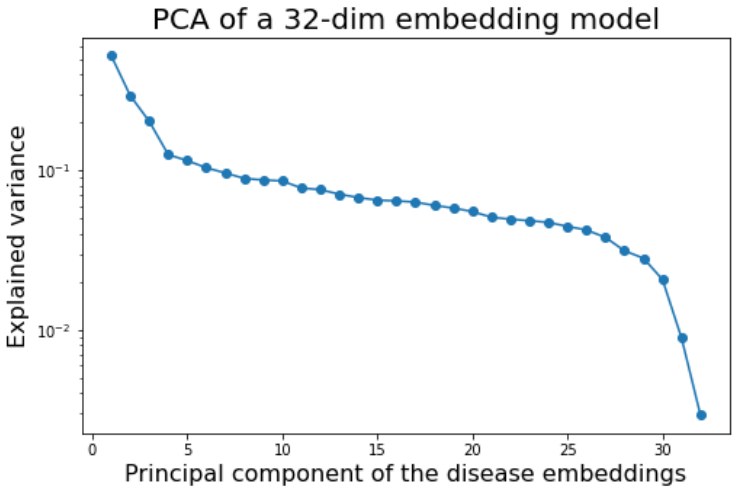

Figure S3 The variance in embedding values explained by successive principal component of the disease embedding matrix for an embedding dimension of 32.

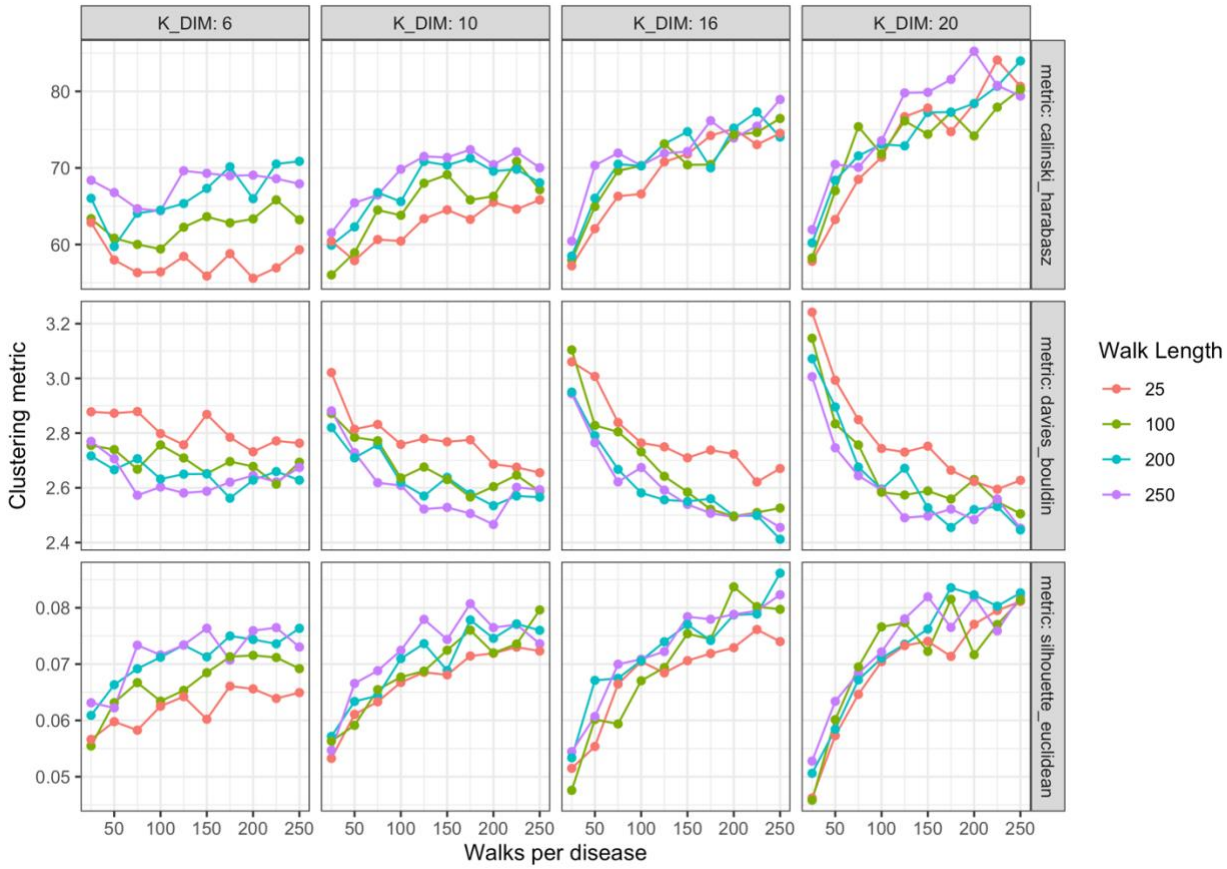

Figure S4: Internal clustering metrics as function of node embedding model hyper parameters. The colors of each line signify the length of the random walks used to construct the disease corpus. The chart columns vary the embedding context window (K_DIM). The chart rows show the three different clustering metric used, which are each plotted over the number of random walks generated per disease.

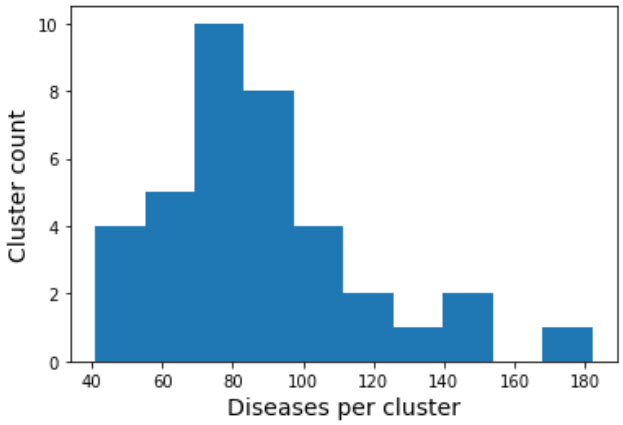

Figure S5 A histogram displaying the distribution of the number of diseases in each of the final clusters