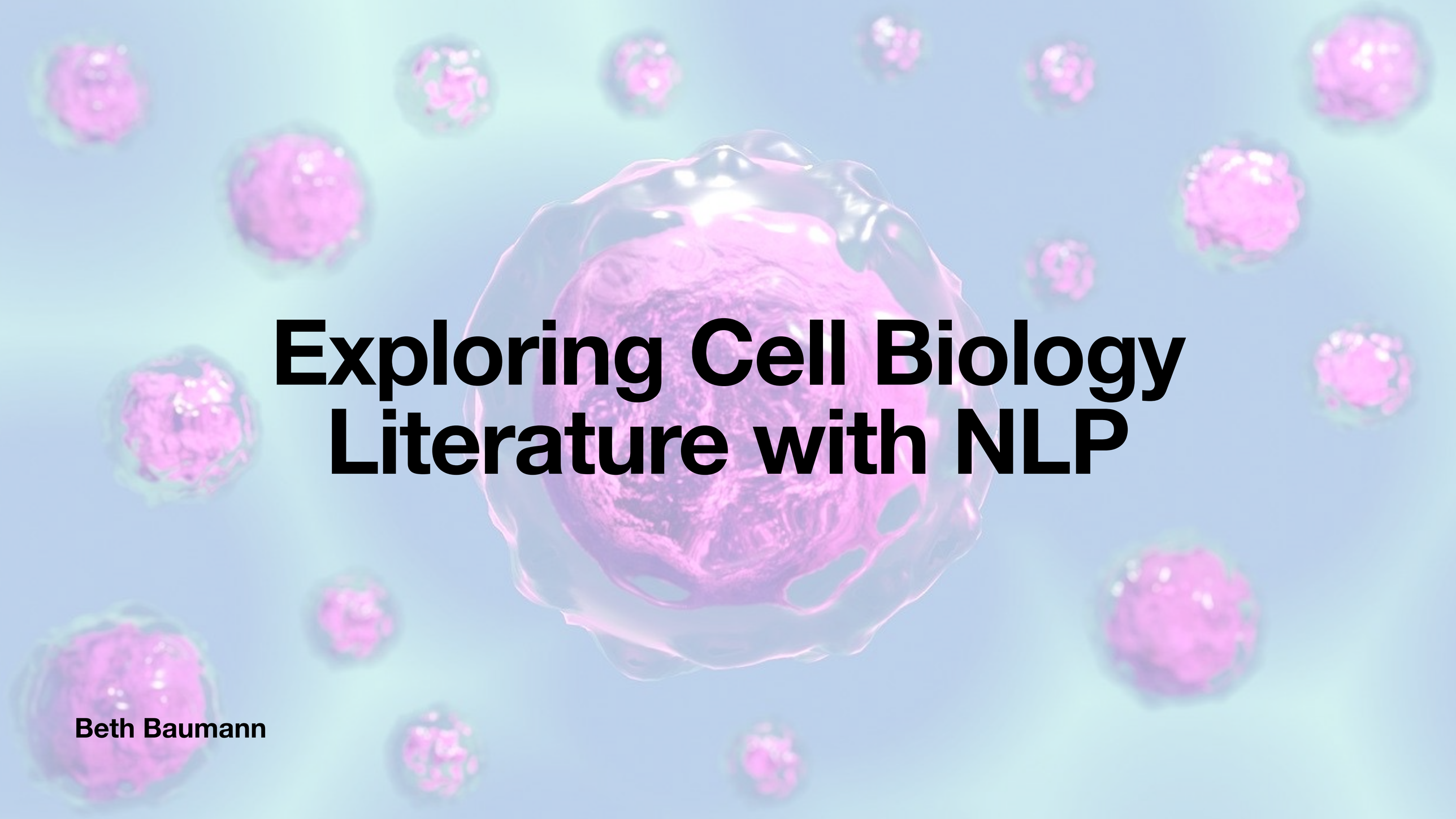
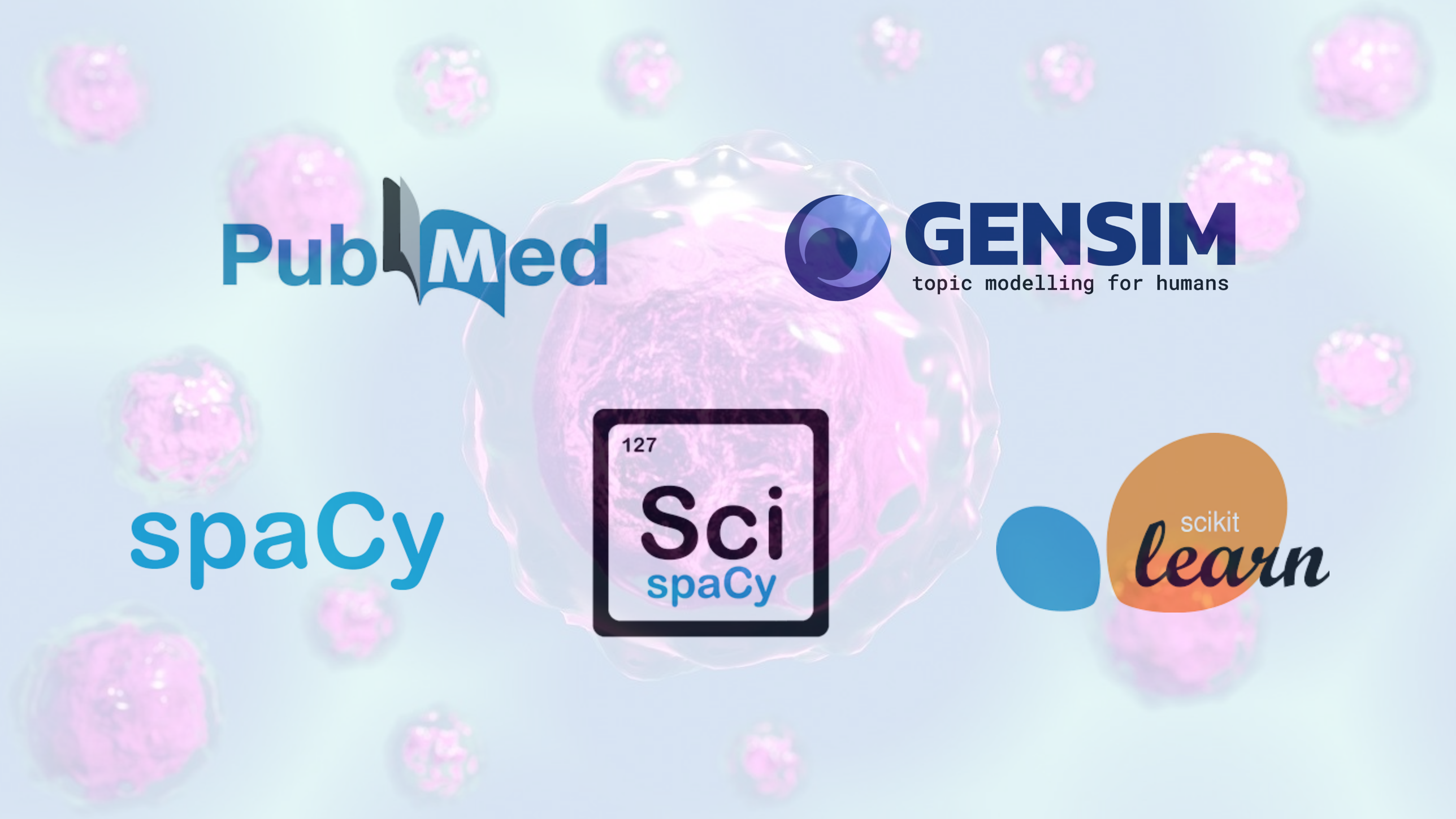


Exploring Cell Biology Literature with NLP

Beth Baumann


PubMed
GENSIM
topic modelling for humans
spaCy
127
Sci
spaCy
scikit
learn

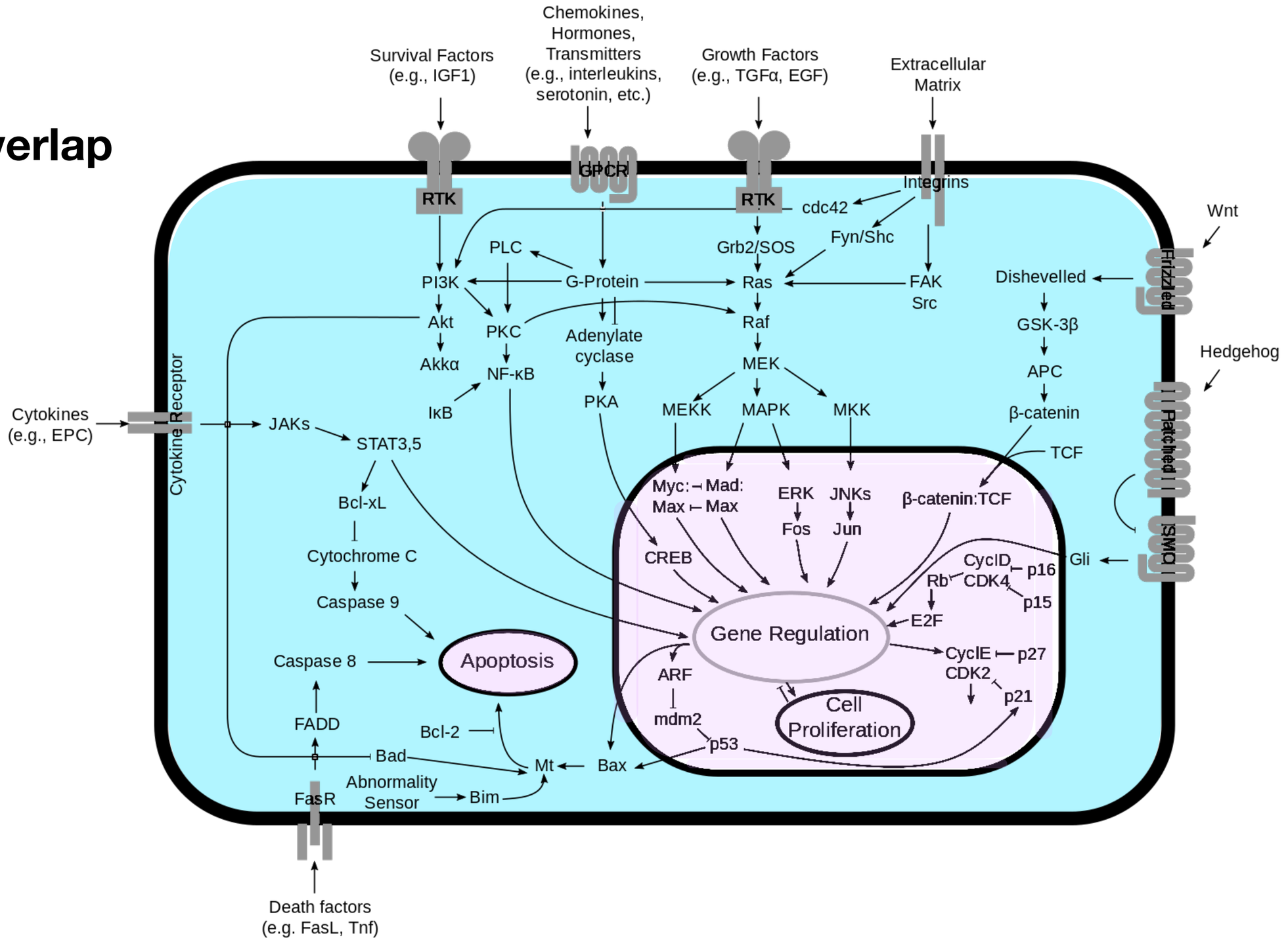
Goals of this NLP analysis

1. Identify the natural clustering of cell biology abstracts into signaling pathways

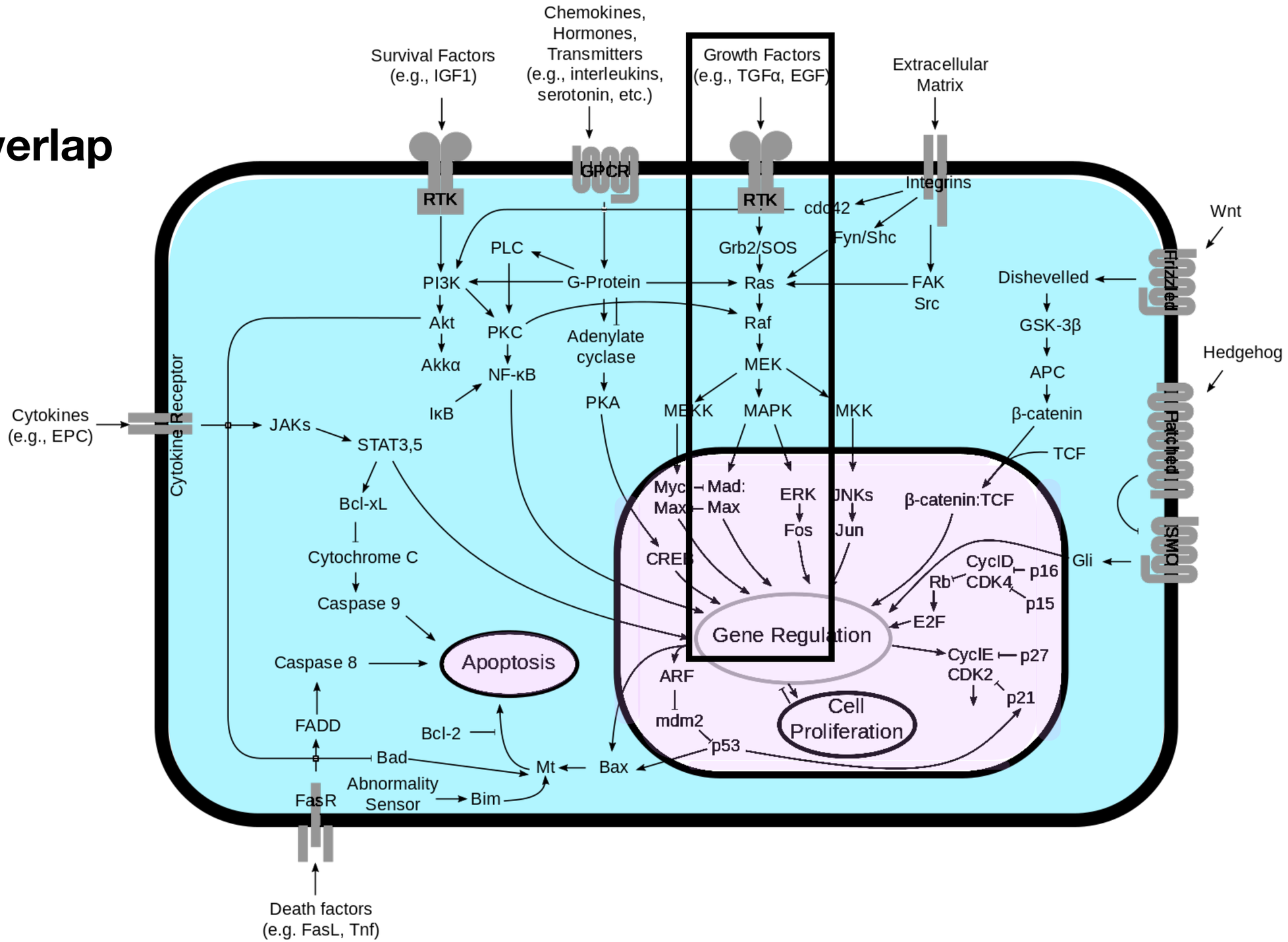
Signaling pathway: a chain or network of proteins that work together for a specific effect on the cell

2. Determine protein membership of signaling pathways
3. Explore cosine similarity of biological terms with Word2Vec

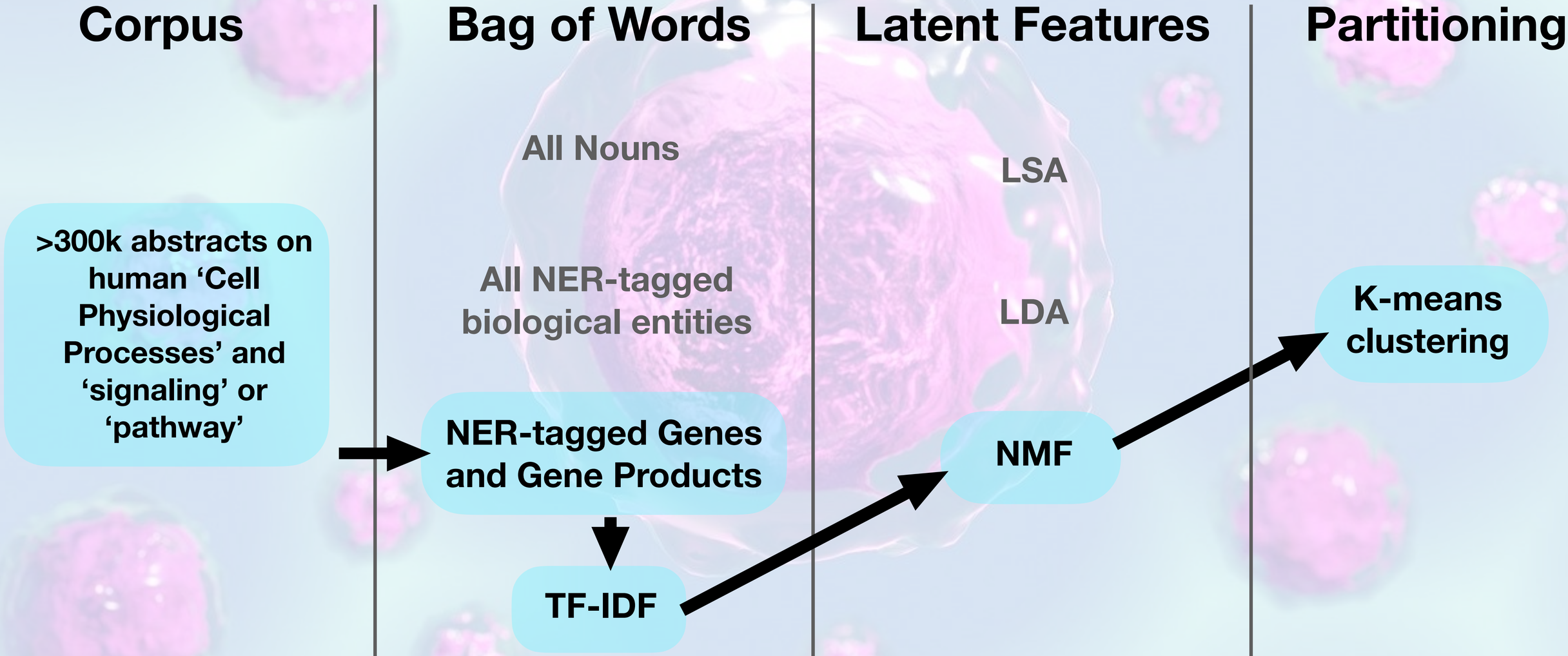
Signaling
Pathway overlap



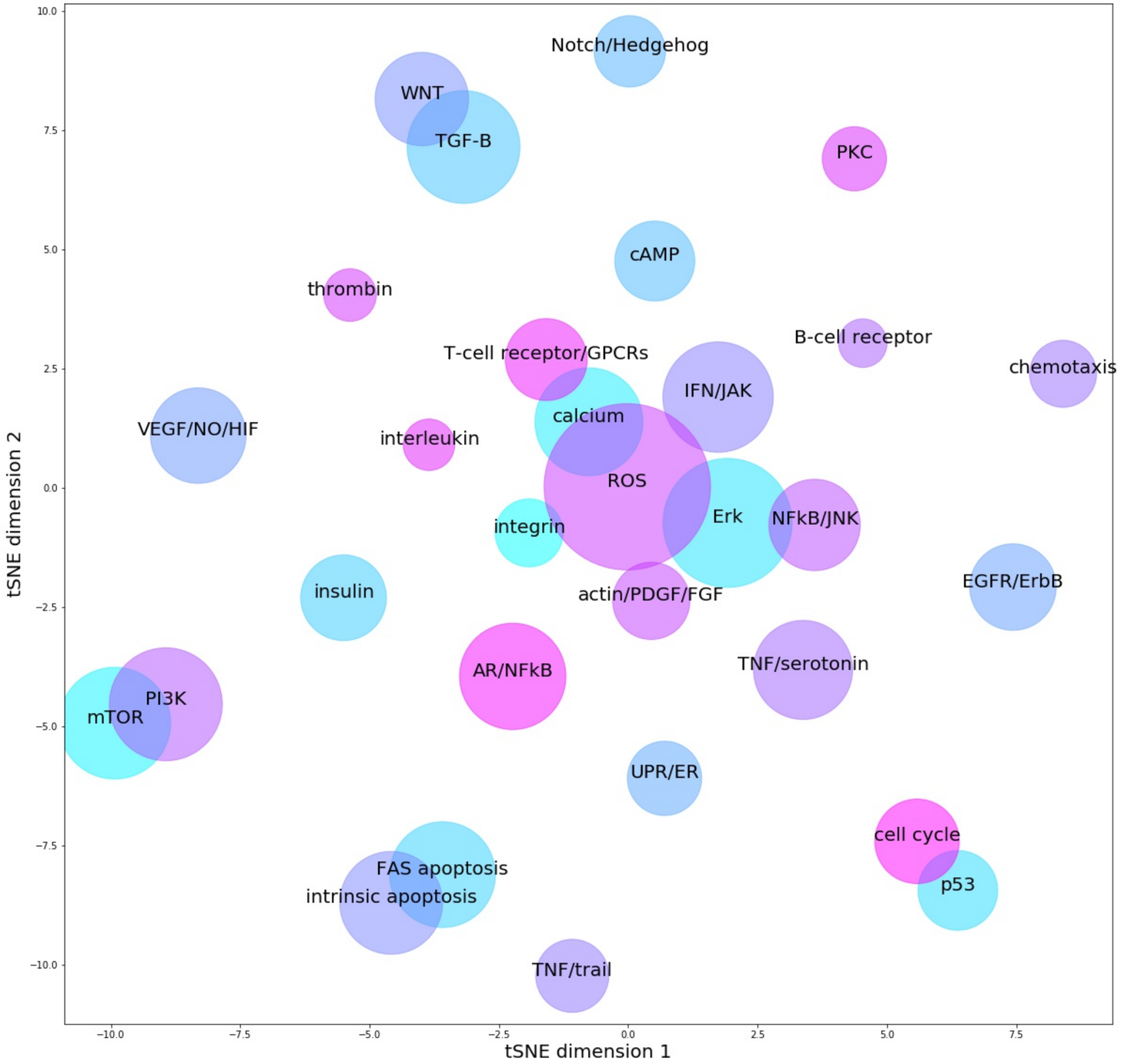
Signaling
Pathway overlap



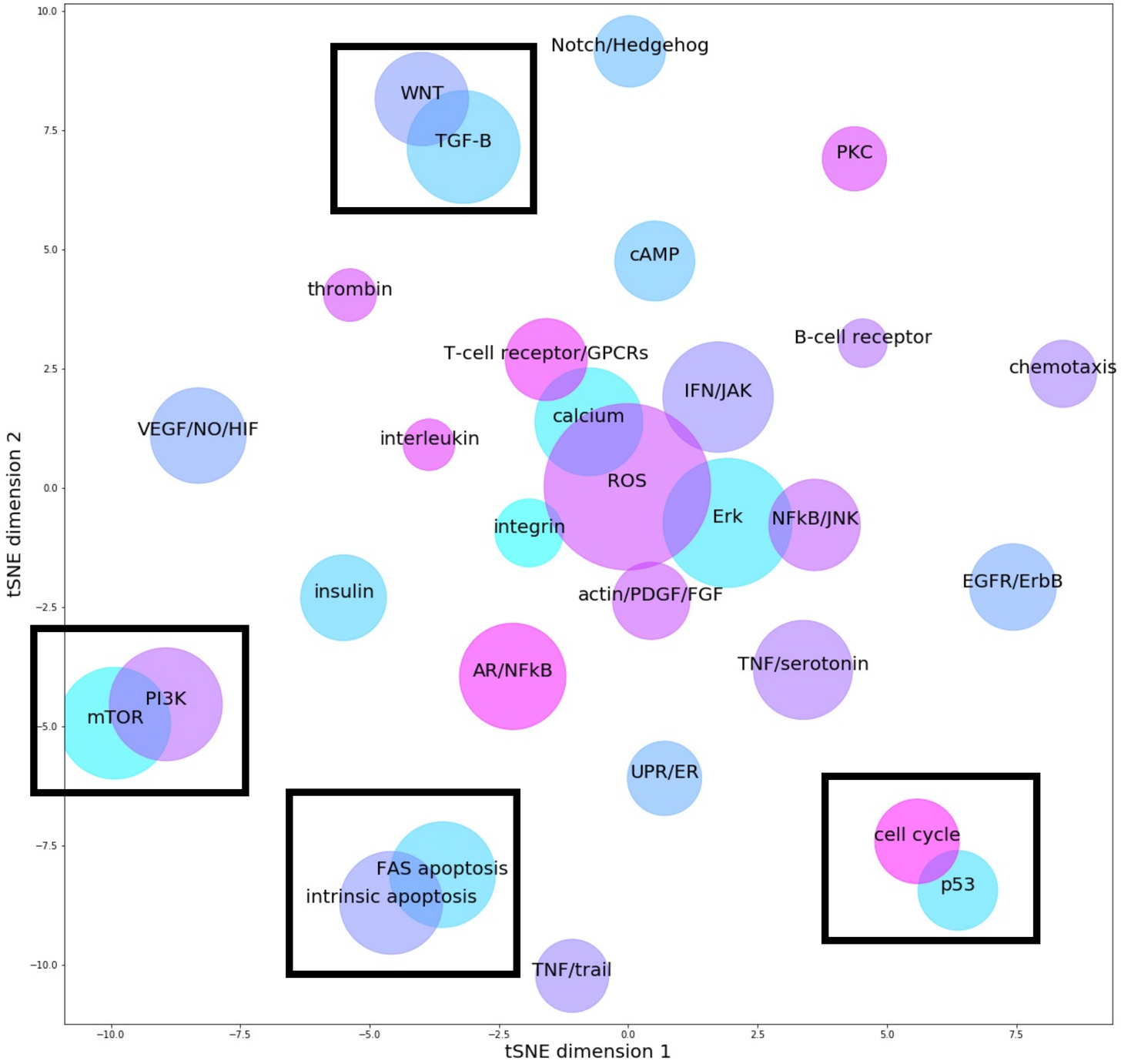
Part I: Topic Modeling with NMF



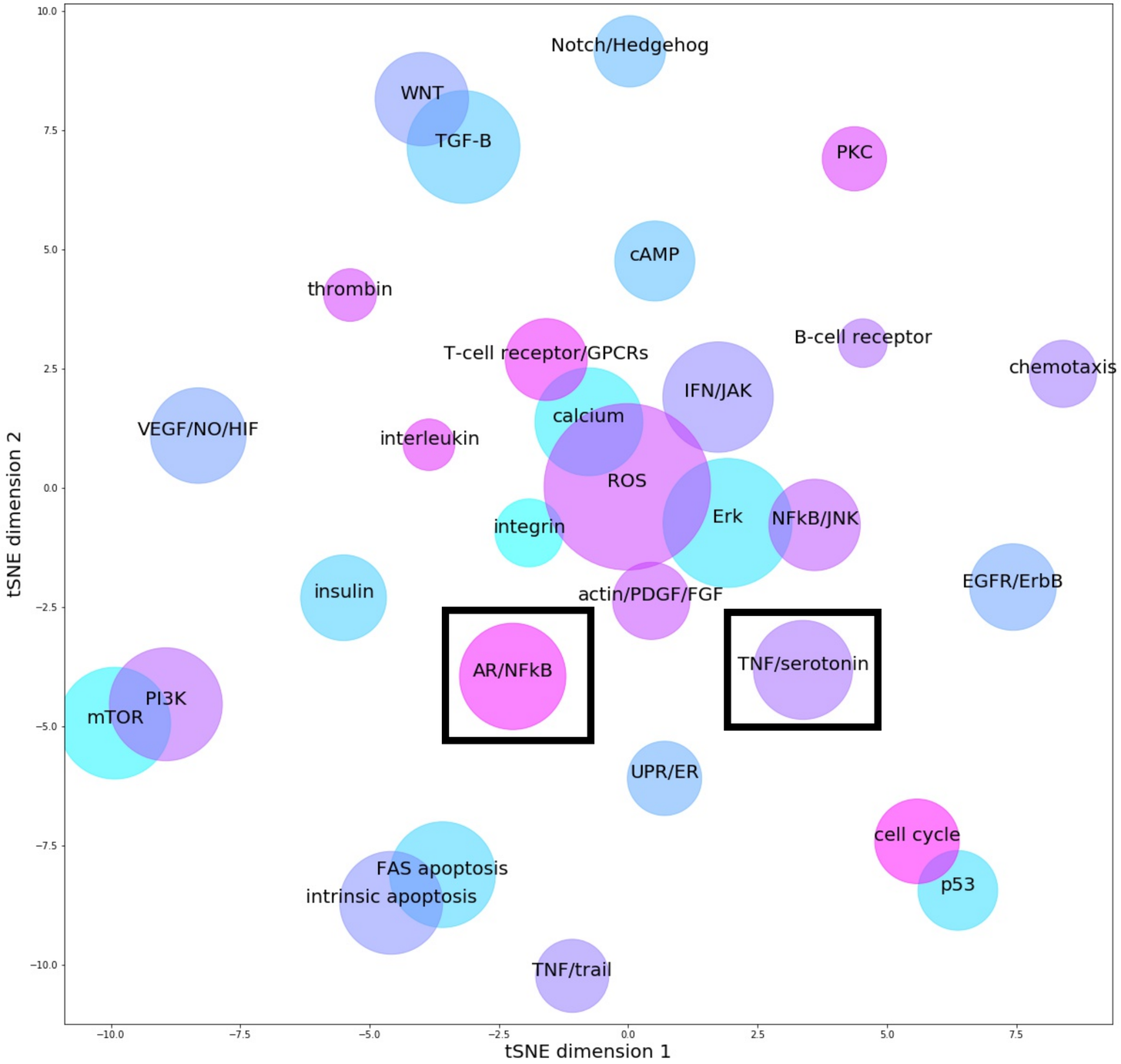
Abstracts fell into 30 K-means clusters along 25 latent topic dimensions



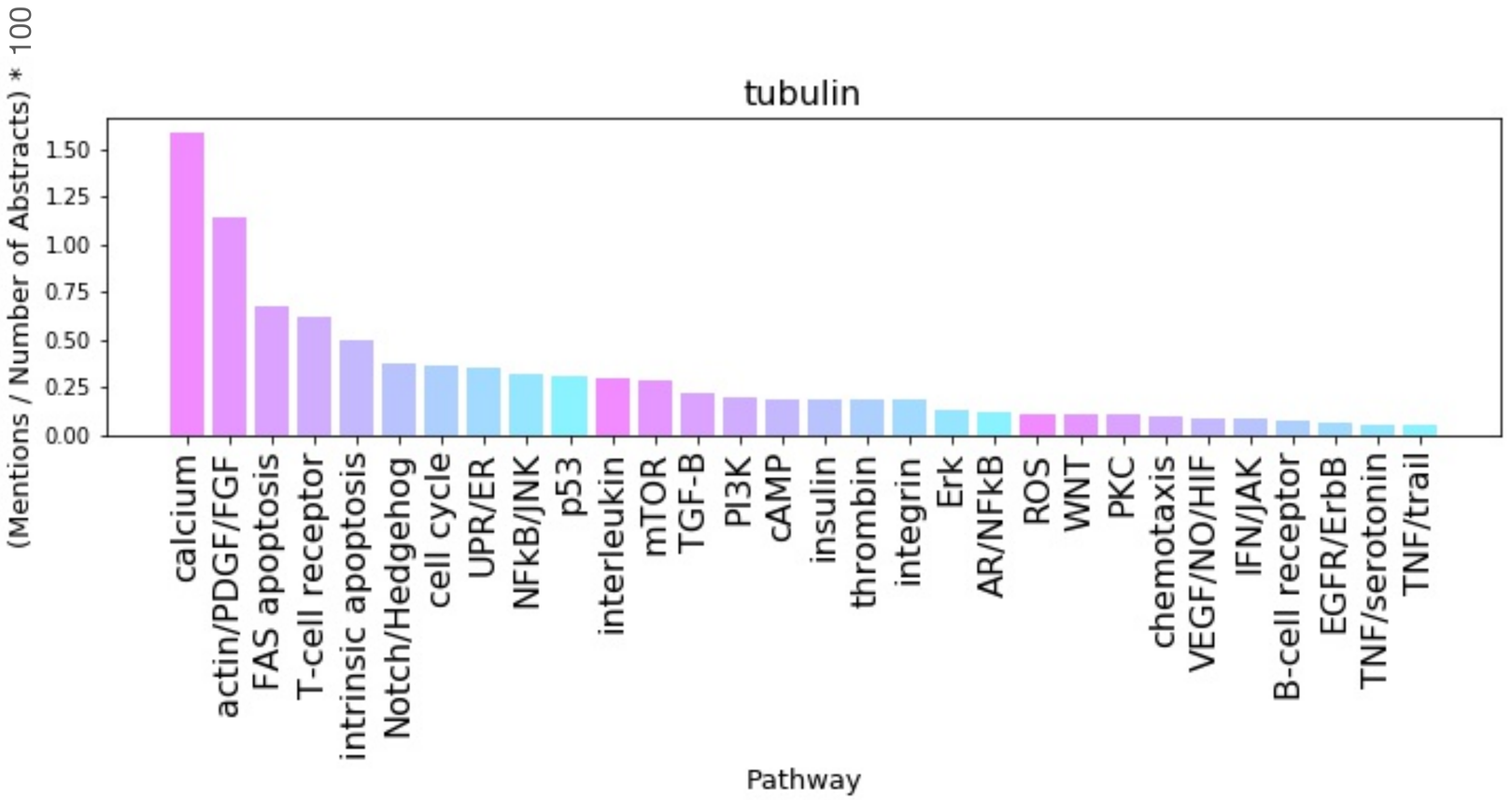
Some clusters of well-known related pathways always appear adjacent in tSNE



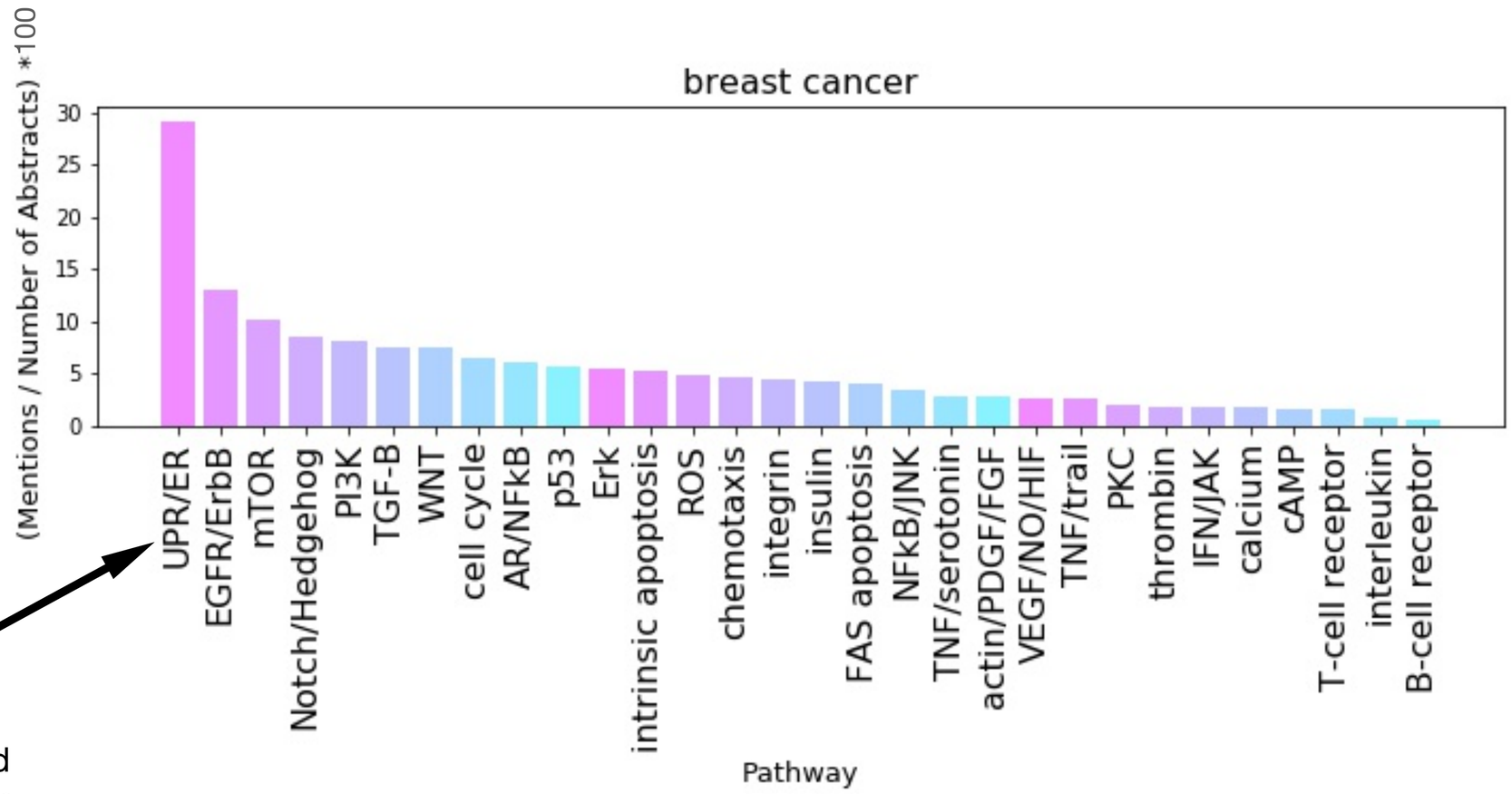
Other clusters represented pathways that are not as ‘canonical’ as the others



Looking up representation of gene tokens by cluster

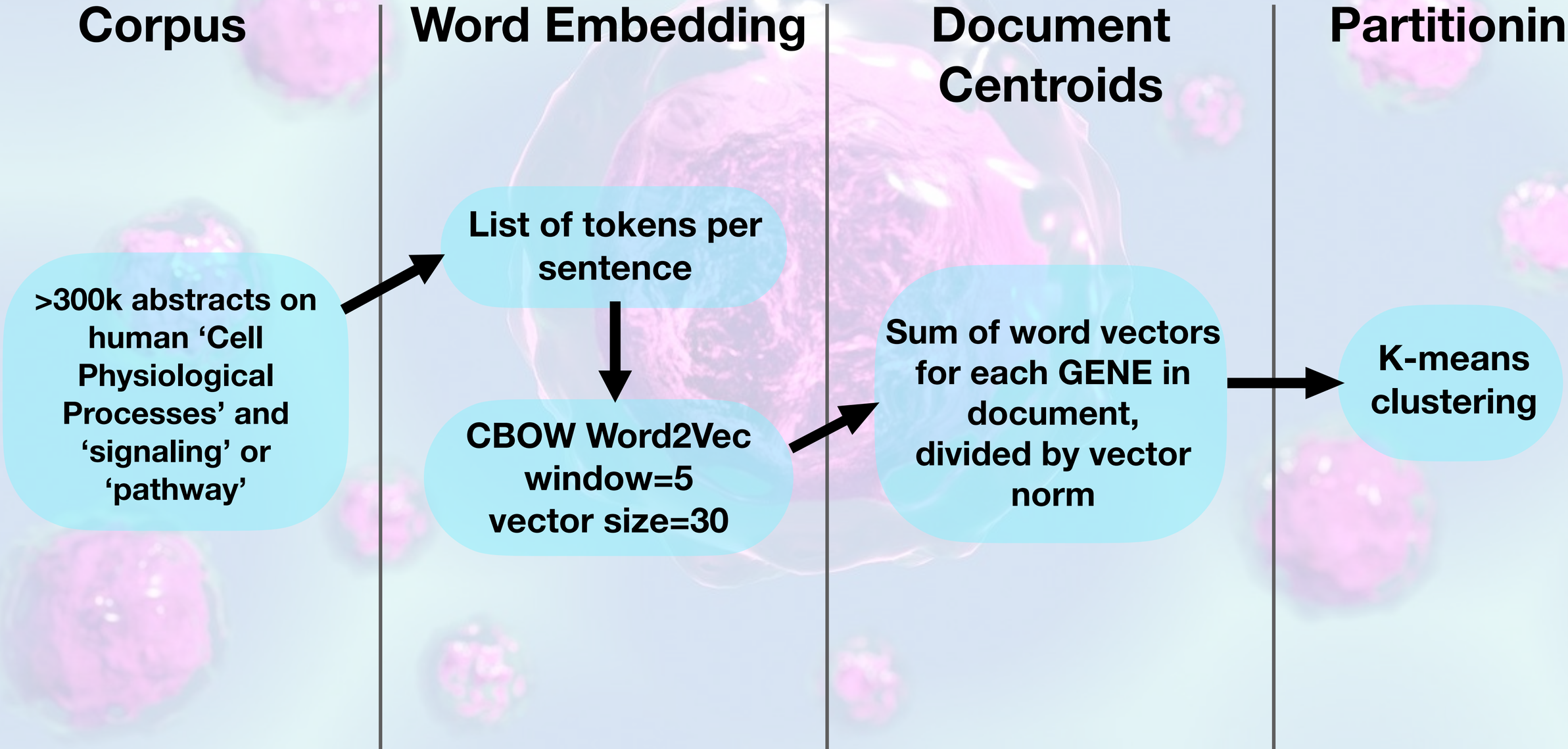


Looking up representation of other tokens by cluster



Estrogen Receptor and Endoplasmic Reticulum have same acronym

Part II: Word2Vec Embedding



Biology Analogies from Word2Vec

Activated by

Smad : Tgf-B :: STAT : JAK

Activates

cAMP : PKA :: calcium : CaMKII

Inhibited by

Nrf-2 : Keap-1 :: NF-kB : IKKa (2nd)

Interacts with

B-catenin : TCF :: Fos : JunB

Unit of

actin : F-actin :: tubulin : microtubule (2nd)

Fluid part

heart : blood :: brain : csf (2nd)

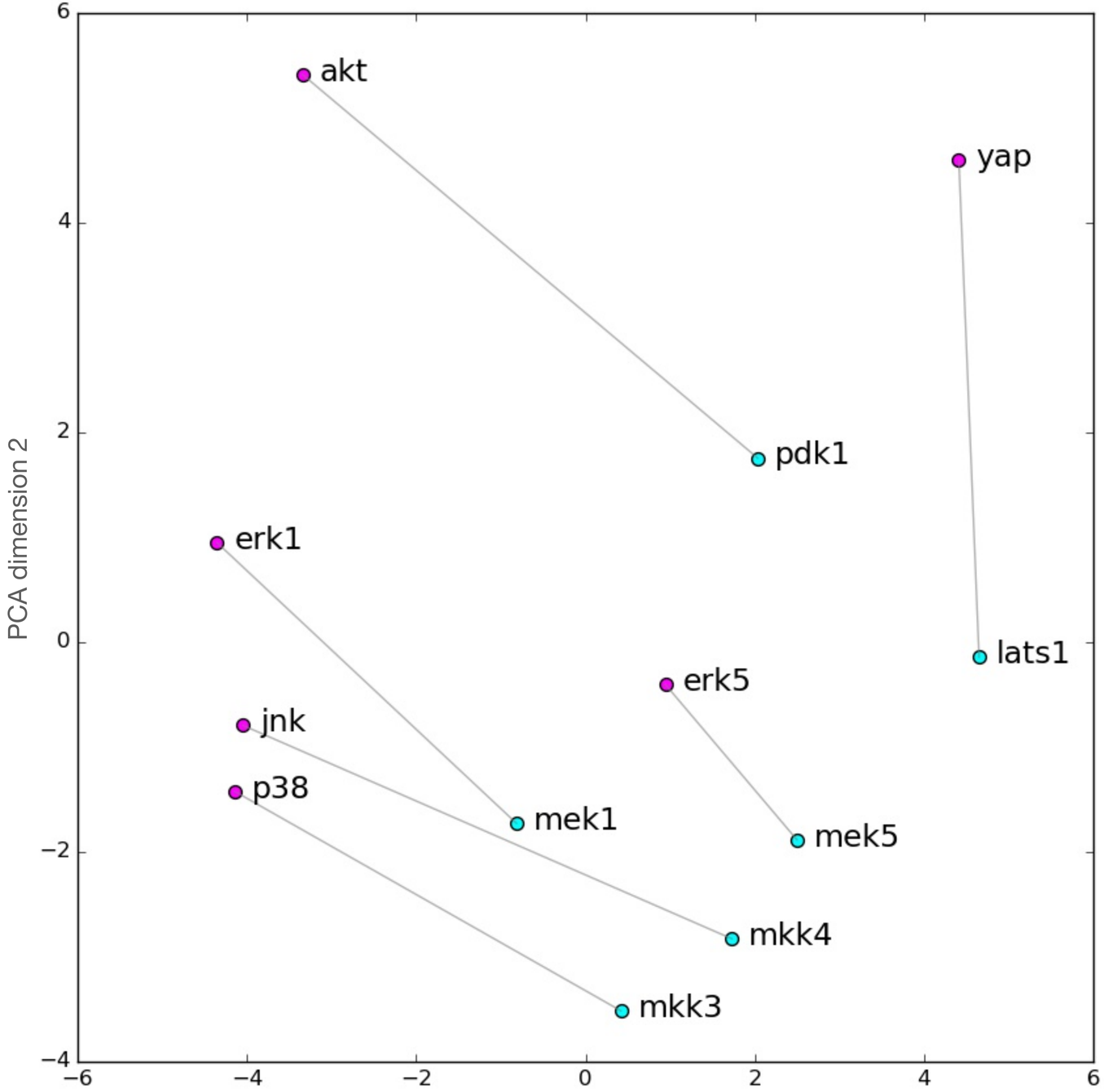
Hormone produced by

insulin : pancreas :: erythropoietin : fetal liver

Binds to receptor

glucose : GLUT4 :: glutamate : AMPAR

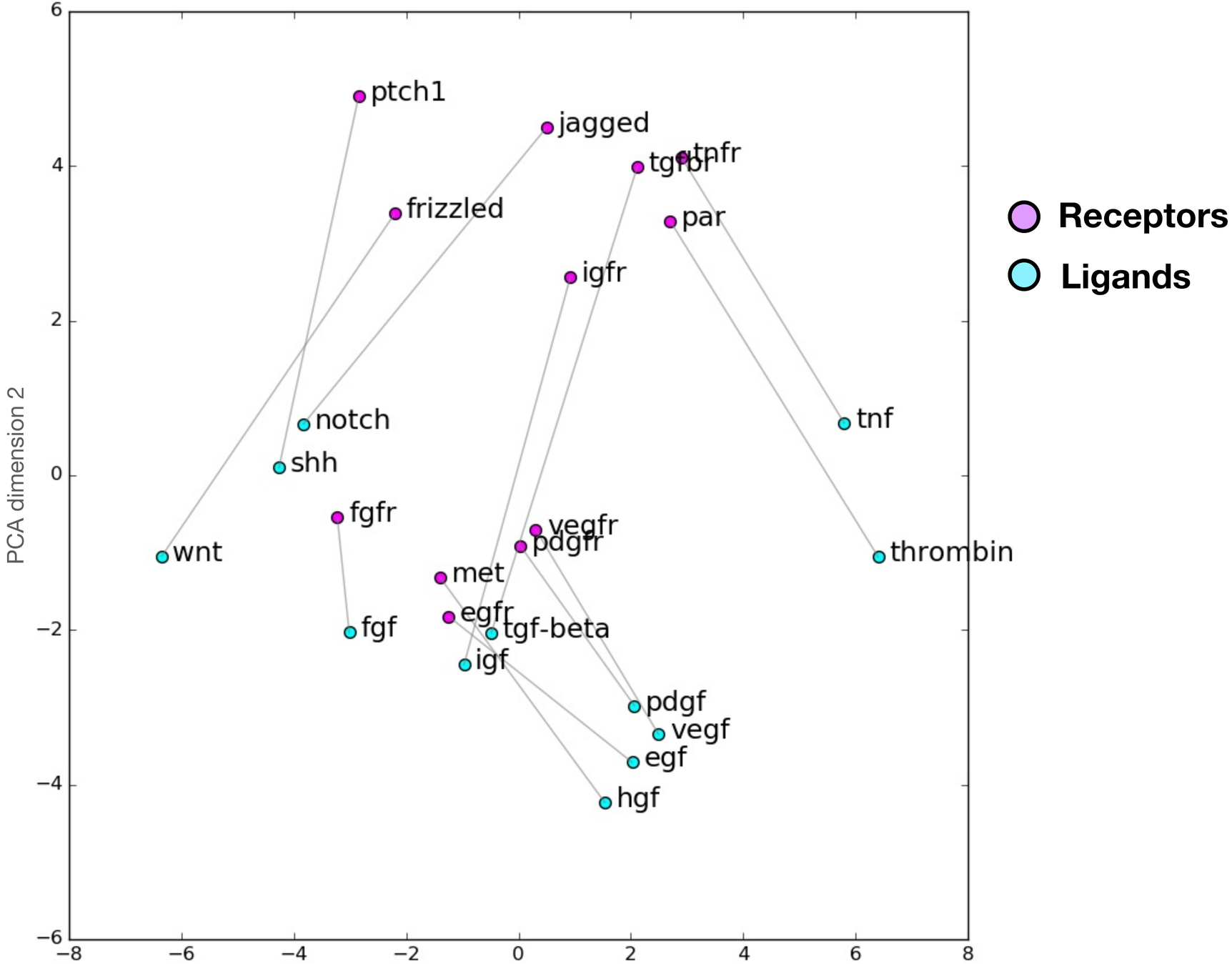
Kinase Cascade Word Embeddings



Downstream kinase

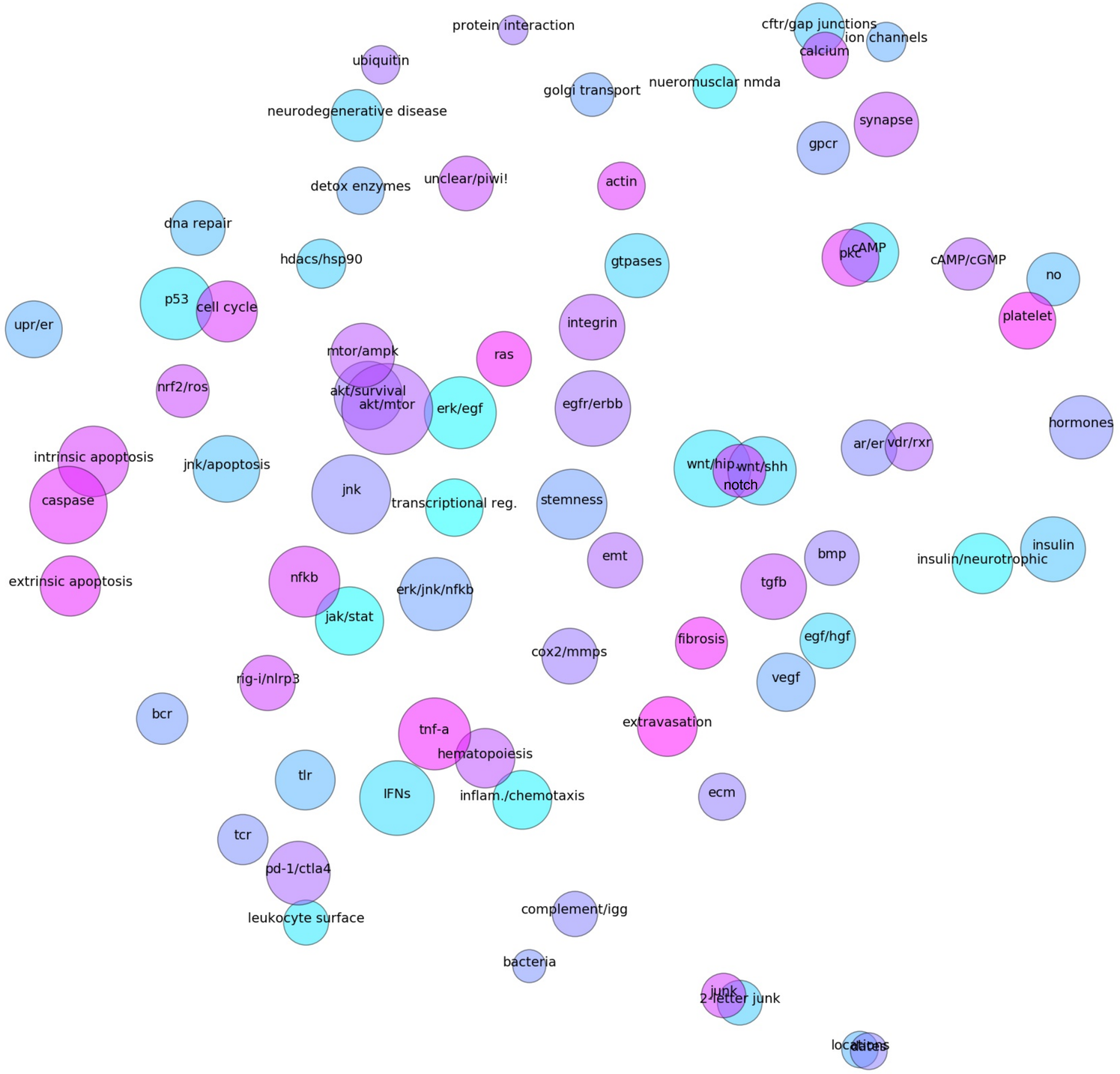
Upstream kinase

Receptor-Ligand Word Embeddings

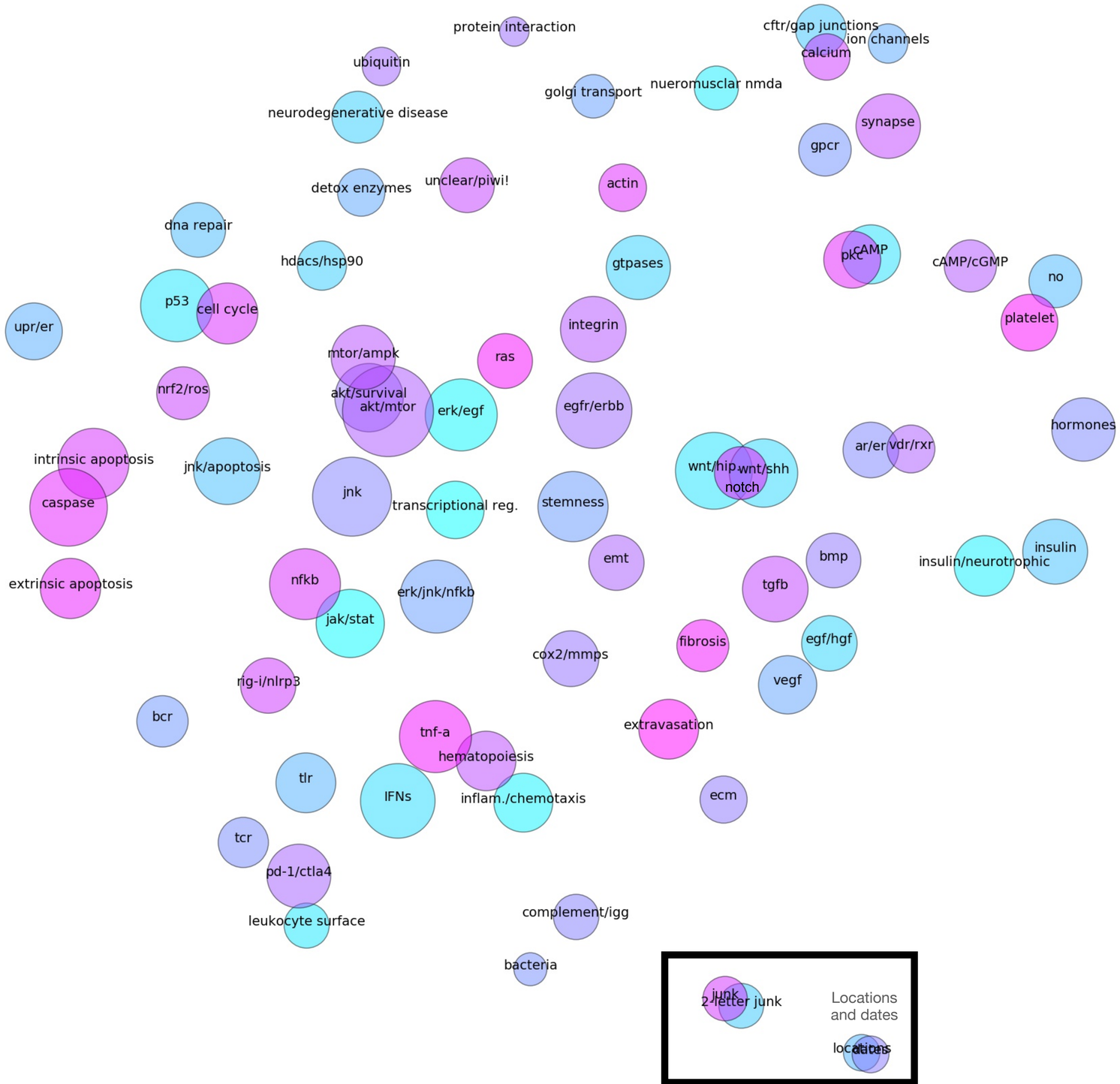


More nuanced clusters from document centroids

Improvement over NMF



Bad documents that only contained author info were filtered out



Future Directions

- App to search gene and token frequency for the Word2Vec-based clusters
- Entity linking tools to collapse gene and gene product aliases
- Sub-word embedding (learns partial word embeddings)
- Clustering techniques that allow greater variation in cluster size and shape
- Compare and upgrade to
  - BioConceptVec - CBOW word embeddings, entity linking, 30M abstracts
  - BioWordVec - Sub-word embeddings, 30M abstracts


Thank you!

Left double click to zoom in, hold for details, right double click to zoom out : Don't show again : Help
Cell-Cell communication
Nephrin family interactions
Signal regulatory protein family interactions
Hemostasis
Platelet Adhesion to exposed collagen
Kinesins
Basigin interactions
Tie2 Signaling
PECAM1 interactions
Dissolution of Fibrin Clot
Reproduction
Fertilization
Sperm Motility And Taxes
Meiotic recombination
Meiotic synapsis
DNA Repair
Fanconi Anemia Pathway
Disease
Infectious disease
Potential therapeutics for SARS
HCMV Early Events
HCMV Late Events
Leishmania infection
LTC4-CYSLTR mediated IL4 production
ADORA2B mediated anti-inflammatory cytokines production
WNT5:FZD7-mediated leishmania damping
Purinergic signaling in leishmaniasis infection
FCGR3A-mediated phagocytosis
Programmed Cell Death
Apoptosis
Regulated Necrosis
Disorders of transmembrane transporters
Diseases of metabolism
Abnormal conversion of 2-oxoglutarate to 2-hydroxyglutarate
Phenylketonuria
Cellular responses to external stimuli
Cellular responses to stress
Attenuation phase
HSF1 activation
Response of EIF2AK1 (HRI) to heme deficiency
Response of EIF2AK4 (GCN2) to amino acid deficiency
Amino acids regulate mTORC1
HSP90 chaperone cycle for steroid hormone receptors (SHR)
Detoxification of Reactive Oxygen Species
Oncogene Induced Senescence
Oxidative Stress Induced Senescence
Organelle biogenesis and maintenance
Mitochondrial biogenesis
Cilium Assembly
Signal Transduction
Signaling by NOTCH
Signaling by Hippo
MAPK family signaling cascades
Signaling by GPCR
GPCR downstream signalling
GPCR ligand binding
G alpha (12/13) signalling events
Opsins
DARPP-32 events
Olfactory Signaling Pathway
Signaling by Nuclear Receptors
RA biosynthesis pathway
Signaling by WNT
Signaling by Rho GTPases
Rho GTPase cycle
Signaling by Receptor Tyrosine Kinases
Signaling by MST1
Signaling by Leptin
Signaling by BMP
LGI-ADAM interactions
Regulation of signaling by NODAL
Formation of the cornified envelope
Nervous system development
Transcriptional regulation of granulopoiesis
Myogenesis
Transcriptional regulation of white adipocyte differentiation
Neuronal System
Vesicle-mediated transport
Membrane Trafficking
Cargo recognition for clathrin-mediated endocytosis
Translocation of SLC2A4 (GLUT4) to the plasma membrane
Regulation of gap junction activity
Circadian Clock
NR1D1 (REV-ERBA) represses gene expression
RORA activates gene expression
Gene expression (Transcription)
RNA Polymerase II Transcription
Generic Transcription Pathway
Nuclear Receptor transcription pathway
YAP1- and WWTR1 (TAZ)-stimulated gene expression
TP53 Regulates Transcription of DNA Repair Genes
TP53 Regulates Metabolic Genes
RNA Polymerase II Transcription Pre-Initiation And Promoter Opening
RNA Polymerase II Promoter Escape
RNA Polymerase II Transcription Initiation
RNA Polymerase II Pre-transcription Events
Transcriptional Regulation by VENTX
Transcriptional Regulation by E2F6
Notch-HLH transcription pathway
Mitochondrial transcription termination
DNA methylation
RNA Polymerase I Transcription Termination
Smooth Muscle Contraction
Striated Muscle Contraction
Cell Cycle
Cell Cycle, Mitotic
Chromosome Maintenance
Cell Cycle Checkpoints
G2 Phase
DNA Replication
Synthesis of DNA
Chromatin organization
Chromatin modifying enzymes
HDMs demethylate histones
HDACs deacetylate histones
RMTs methylate histone arginines
HATs acetylate histones
PKMTs methylate histone lysines
Immune System
Cytokine Signaling in Immune system
Signaling by Interleukins
Regulation of IFNG signaling
Regulation of IFNA signaling
Prolactin receptor signaling
STAT5 Activation
Innate Immune System
Neutrophil degranulation
Alpha-protein kinase 1 signaling pathway
Advanced glycosylation endproduct receptor signaling
DAP12 signaling
Adaptive Immune System
Antigen processing: Ubiquitination & Proteasome degradation
MHC class II antigen presentation
Butyrophilin (BTN) family interactions
Rap1 signalling
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
Metabolism
Metabolism of lipids
Lipid particle organization
eNOS activation
Metabolism of nucleotides
Interconversion of nucleotide di- and triphosphates
Vitamin D (calciferol) metabolism
Metabolism of vitamins and cofactors
Citric acid cycle (TCA cycle)
Pyrophosphate hydrolysis
Electron transport from NADPH to Ferredoxin
Biological oxidations
PPARA activates gene expression
Metabolism of amino acids and derivatives
Metabolism of carbohydrates
Lactose synthesis
Urea cycle
Reversible hydration of carbon dioxide
Aflatoxin activation and detoxification
Cytosolic iron-sulfur cluster assembly
Transport of small molecules
Cellular hexose transport
Intracellular oxygen transport
Processing of SMDT1
TRP channels
Ion transport by P-type ATPases
Autophagy
Metabolism of RNA
snRNP Assembly
mRNA Capping
Metabolism of proteins
Post-translational protein modification
Protein methylation
Neddylation
Protein repair
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
Extracellular matrix organization
ECM proteoglycans
FoamTree