Extracting T cell function and differentiation characteristics from immunology literature with Snorkel and SciSpacy

Many promising cancer immunotherapy treatment protocols rely on efficient and increasingly extensive methods for manipulating human immune cells. T cells are a frequent target of the laboratory and clinical research driving the development of such protocols as they are most often the effector of the cytotoxic activity that makes these treatments so potent. However, the cytokine signaling network that drives the differentiation and function of such cells is complex and difficult to replicate on a large scale in model biological systems. Abridged versions of these networks have been established over decades of research but it remains challenging to define their global structure as the classification of T cell subtypes operating in these networks, the mechanics of their formation, and the purpose of the signaling molecules they excrete are all controversial, with a slowly expanding understanding emerging in literature over time.

To aid in the quantification of this understanding, we are developing a methodology for identifying references to well known cytokines, transcription factors, and T cell types in literature as well as classifying the relationships between the three in an attempt to determine what cytokines initiate the transcription programs that lead to various cell states in addition to the secretion profiles associated with those states. Entity recognition for this task is performed using SciSpacy and classification of the relations between these entities is based on an LSTM trained using Snorkel, where weak supervision is established through a variety of classification heuristics and distant supervision is provided via previously published immunology databases. Source code and results for this project will be maintained at https://github.com/hammerlab/t-cell-relation-extraction.