Both the 'reverse engineering' of biological networks (for example, by integrating sequence data and expression data) and the analysis of their underlying design (by revealing the evolutionary mechanisms responsible for the resulting topologies) can be re-cast as problems in machine learning: learning an accurate prediction function from high-dimensional data. In the case of inferring biological networks, predicting up- or down- regulation of genes allows us to learn about the transcription factor binding sites (or 'motifs') and to generate a predictive model of transcriptional regulation. In the case of inferring evolutionary designs, quantitative, unambiguous model validation can be performed, clarifying which of several possible theoretical models of how biological networks evolve might best (or worst) describe real-world networks. In either case, by taking a machine learning approach, we statistically validate the models both on held-out data and via randomizations of the original dataset to assess statistical significance. By allowing the data to reveal which features are the most important (based on predictive power rather than overabundance relative to an assumed null model) we learn models which are both statically validated and biologically interpretable.

References: