**Explainable AI in systems biology**Piet Jones, Benjamin Garcia, Dale Pelletier, Jay Chen, Wellington Muchero, Gerald Tuskan, Daniel Jacobson
Oak Ridge National Laboratory

jonespc@ornl.gov

**Abstract**

Deep learning methods, as universal function approximators, allow us to apply predictive models to a diverse set of domains. Emphasizing explainable models, allow us to investigate feature importance. Thereby improving our understanding of the underlying system. Here we describe a number of these methods that we are currently pursuing, each with their respective strength and weaknesses. These can be used alone or in combination, thereby complimenting each other.

Factorization machines provide the means to efficiently model higher order interactions between features. These interactions can be used to uncover potential outliers in sparse data, as well as possible epistasis (multiple genes involved in a given function). Capsule networks offer a potentially explainable alternative to convolution neural networks (CNN). By propagating more information through the network a smaller dataset can be used to train the model, compared to CNNs. Encoding our data as images allows us to uncover features important in for example, protein-protein interaction. Knowledge distillation allows us to model the structure of an already learned network, using a more explainable deep learning architecture such as Bayesian Neural Network.

We apply our deep learning approaches to a number of biological data sets to improve our systems level understanding. In one dataset we investigate interactions between members of the phytobiome, elucidating potential mutualism, antagonism between and among; viruses, fungi, bacteria and the plant host system (P. trichocarpa). In another dataset we focus on understanding the gene diversity within this particular plant species, uncovering associations between these pan/core genes and various phenotypes.