AIM 3. Evaluating the Importance of Hit-and-Run Through the Quality of Forecasting

**Hypothesis:** Hit-and-run binding can influence downstream expression at the network level. **Rationale:** The above aims will show that transient hit-and-run binding leave physical traces such as methylation on promoters. We hypothesize that these physical traces of a transient TF-target interactions captured in cells will affect the expression of target genes *in planta*. Previous studies have ignored these transient events. **Approach:** To test whether hit-and-run events affect expression *in planta*, we will compare expression forecasting quality based on a null hypothesis which assumes that transient binding has no effect and an alternative hypothesis which assumes that transient binding could have some effect. If the null hypothesis holds, then forecasting without information about transiently bound transcription factors should have as small a mean squared error as forecasting with that information. The alternative hypothesis is that the transient information helps.

**Approach:** To perform this evaluation, we introduce a new machine learning-based genomics-level tool called OutPredict that offers a novel combination of features needed for this goals: (i) OutPredict forecasts the expression value of an unseen time point; (ii) it allows for non-linear dependencies of target genes on causal transcription factors; (iii) it incorporates steady state data, notably transient and stable binding information to bias (and hopefully improve) forecasts. We compare this method to the state-of-the-art algorithms Dynamic Genie 3 which supports (i) and (ii) but not (iii) and with Inferelator which supports [Jacopo, please fill in], and Neural Networks which supports [Jacopo please fill in] and our own Dynamic Factor Graph (DFG) approach which supports [Jacopo, please fill in]. \*Discuss whether current time-series based ML methods do not use priors

Intuitively, OutPredict achieves these three features by learning a function that maps expression values of potentially all transcription factors at time t to the expression value of each target gene at the next time point. This per-gene function is embodied in a random forest, allowing it to reflect highly non-linear relationships. OutPredict uses prior information (such as stable binding or transient binding information) to bias the choice of transcription factors in the decision trees of the random forest. Specifically, in the model for gene g, if transcription factor F is known to bind to g, then F will be more likely to be a decision node in a decision tree for g than some other transcription factor F’ for which there is no evidence of binding to g. As we will see, incorporating such prior information reduces the error in the forecasting. Further the random forest model leads to a ranking of the influence of various transcription factors on target genes, thus yielding a gene expression causal network.

**Preliminary Results:** We apply OutPredict and state-of-the-art algorithms on a previously published nitrogen time series on Arabidopsis shoots, where

[Show the results and commentary]

**Jacopo: Preliminary Results**

 Learn time series model with NxTIME shoots

 Priors: 4 TFs from PNAS

 Priors: DAP-Seq (DHS) filtered

 DAP-Seq intersect with TARGET regulated

 TARGET Regulated

 Outcome 1: Accuracy on predicting out of time points

 Compared to: DynGenie 3 (TIME) and NeuralNet

 Outcome 2: Network Inference 🡪 Most influential TF

 Roots: NxTime roots to learn networks

 Priors: 17/33 TFs with DAP seq data

 Priors: DAP-Seq (DHS) filtered

**Interpretations and Expected Outcomes:** If the stable binding data improves the forecasting (reduces the error) over no binding data and the transient data improves the forecast even further, then we will have demonstrated both that our approach to using steady state data holds promise and that transient hit-and-run relationships are causally involved in gene expression.

**Potential Problems and Alternative Approaches:** One problem with our experiments so far is that we have binding information about so few transcription factors. This problem will be greatly diminished as we acquire data about more transcription factors using the high throughput techniques outlined above. [Another experiment: Use only validated transient interactions vs all transient targets based on regulation]