**Aim 3**

**Rationale:** Aim 2 will identify the stable and transient TF-target interactions for each of the candidate *Hit-and-Run* TFs. Aim 3 will apply the *Hit-and-Run* TFs identified in Aim 2 to infer Gene Regulatory Networks (GRNs).using two new computational approaches: Network Walking (Aim 3A) and OutPredict (Aim 3B). Both direct and indirect TF-targets will be validated *in planta* in Aim 3C for their impact on NUE.

**3A. *Network Walking* - Charting a network path from TF1 direct targets in cells to indirect targets *in planta*.**

***Innovation***: A major limitation of *in planta* TF perturbation studies is the inability to distinguish direct from indirect targets without additional *in vivo* TF-target binding information, such as ChIP. By integrating TF-target interactions identified in the cell-based *TARGET* assay with *in planta* perturbation results, we are able to define a network path that connects the direct targets of TFs to indirect targets identified only *in planta*. To supplement the functionally validated TF-target edges determined using *TARGET*, we use the TF-target edges from our pruned DFG network (see Background) to link direct and indirect targets. This approach, called “Network Walking,” was recently published for two TFs known to be involved in the N-response, *TGA1* and *CRF4* (Brooks et al.). We will use the *TARGET* data for 150 TFs from Aim 1, and *in planta* overexpression results for validated *Hit-and-Run* TFs from Aim 2 to extend the *Network Walking* approach to encompass data for each of the N-responsive TFs in roots.

***Approach and Preliminary Results***: In “*Network Walking*”, the first step is to identify the direct regulated targets of some transcription facor TF1 (e.g. CRF4) using the cell-based *TARGET* assay. In the CRF4 example, the *TARGET* assay showed that CRF4 directly regulated 65 root N-responsive genes (Fig. X, yellow box), including 23 direct target genes that also respond to CRF overexpression *in planta(Brooks et al. 2019)*. The second step is to connect a path from TF1 – via a TF2 – to the TF1 indirect targets which respond only *in planta*. To connect *CRF4* to its indirect targets, we used validated TF2 direct target edges from *TARGET* assays, as well as the high-confidence TF2-target edges for 116 TFs from the pruned DFG network*(Brooks et al. 2019)*. Using this approach, we could link 87% of indirect *CRF4* targets *in planta* (158/182) back to *CRF4* through 5 direct TF2 targets of *CRF4*. To further determine which intermediate TF2s may be most important in relaying the N-signal downstream of *CRF4*, we analyzed enrichment of known motifs in *CRF4* indirect targets.

***Interpretation and Expected Outcomes***: Our recently published Network Walking approach*(Brooks et al. 2019)* integrates TF perturbation and time-series predictions from cells and *in planta* to chart a network path for a TF from direct targets to indirect targets. This can reveal how the N signal travels through transcriptional networks, and can be used to identify the partner TF2s that sustain transcription of the transient targets of a *Hit-and-Run* TF1 after it is no longer bound.

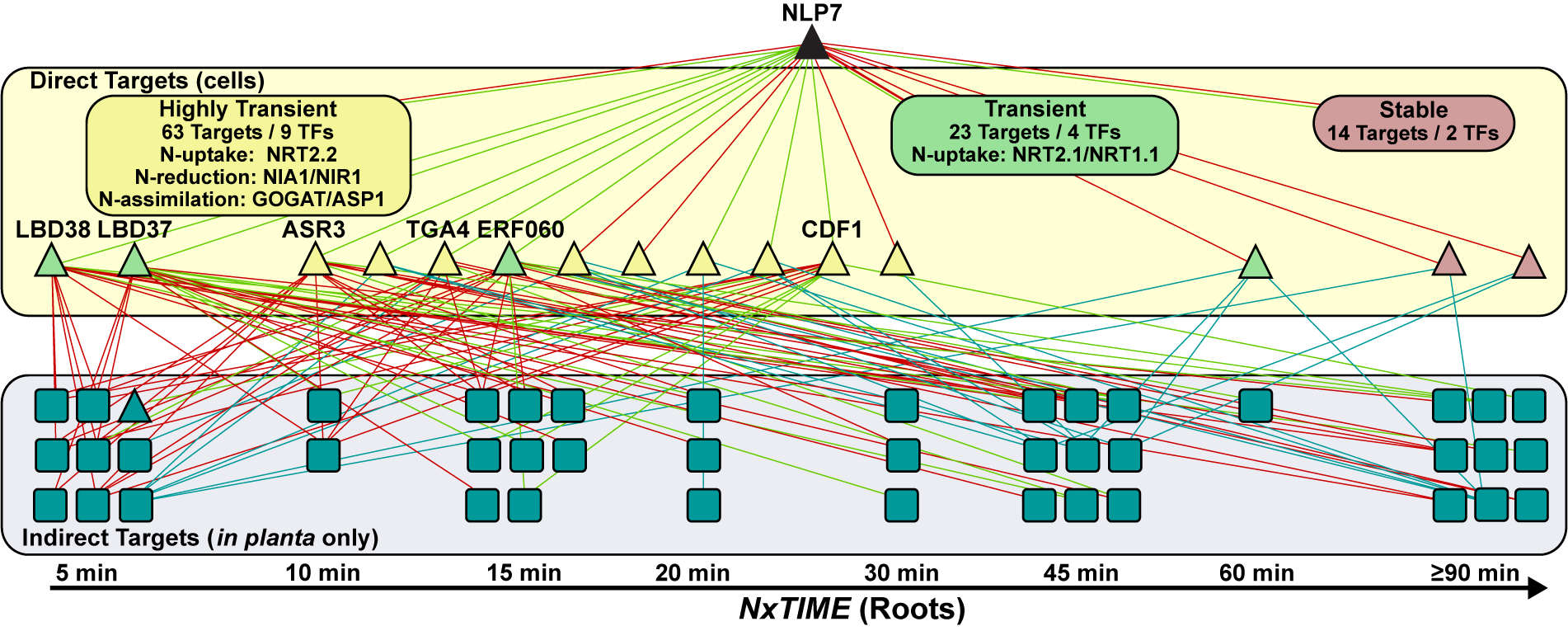


Fig. X -

**3B. Evaluating the importance of transient *Hit-and-Run* interactions through the *quality-of-Forecasting***

***Hypothesis:*** We hypothesize that transient targets of *Hit-and-Run* TFs are relevant to the N-response *in planta*, but are largely ignored due to the difficultly of capturing them using traditional ChIP and TF perturbation. We will test this hypothesis using a machine learning algorithm that we are developing (*OutPredict)* which is capable of using *in planta* time-series gene expression data and prior information on TF-target interactions to forecast gene expression values at future time-points. We will test whether the forecasting performance of *OutPredict* will be improved when stable + transient TF-targets are used, as compared to only stable TF-target interactions.

***Approach:*** Our new machine learning-based genomics-level tool called *OutPredict* offers a novel combination of features: (i) *OutPredict* forecasts the expression value of genes at an unseen time-point; (ii) the model allows for non-linear dependencies of target genes on causal transcription factors; and (iii) it incorporates prior binding (e.g. transient and/or stable TF-target interactions) information to bias the forecasts. We compare the *OutPredict* method to the state-of-the-art forecasting algorithms, such as Dynamic Genie3(Huynh-Thu and Geurts 2018), that support forecasting and non-linear relationships, but currently lack the ability to incorporate priors. Other time-based machine learning methods such as Dynamic Factor Graph(Mirowski and LeCun 2009), which we used in our previous studies*(Brooks et al. 2019; Varala et al. 2018)*, and Inferelator(Bonneau et al. 2006) are based on regularized linear regression. Neural Network approaches [ref] are also becoming increasingly popular, however these methods require much larger amounts of data.

Intuitively, *OutPredict* learns 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 or transient TF-binding) to bias the choice of transcription factors in the nodes of 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 gene g. *OutPredict* tunes the bias values (which influence “feature importance”), based on Out-of-Bag errors on the training set. The Random Forest uses bootstrap aggregation, where each new tree is trained on a bootstrap sample of the training data. The Out-of-Bag error is estimated as the average error for each training data point *pi* by evaluating predictions from the trees that do not include *pi* in their corresponding bootstrap sample. Moreover, *OutPredict* implements a gene-by-gene hyper-parameter optimization, in order to find the best set of hyper-parameters for each target gene. 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 have applied *OutPredict* to the root and shoot N-response time series transcriptome data in Arabidopsis from Varala et al.(Varala et al 2018). To evaluate the performance of our forecasting predictions, we compare the predicted expression values to the actual expression values for each gene and calculate the mean squared error (MSE) across all genes (Fig XA) We compare the MSE for predictions from three methods: (i) naïve, in which the expression of gene g is predicted to be the same as the expression of gene g at the final time point of the training set, (ii) *OutPredict*, and(iii) Dynamic Genie3 (Huynh-Thu and Geurts 2018). We found that *OutPredict* performs 34.2% better than naïve, and 61.5% better than Dynamic Genie3 (Fig. X B). Because *OutPredict* allows the incorporation of priors into the model, such as interactions validated using *TARGET* (Aims 1 and 2), we compared the forecasting performance of *OutPredict* using only the N-response time-series vs. N-response time series with the *TARGET* validated edges as priors. We found that the inclusion of priors improved the predictions for GRNs modeled on shoot N-response data compared to excluding priors (Fig XC, 9% improvement, p-value <0.05). We also found that only TF-target interactions based on direct regulated targets identified using the *TARGET* assay improved performance of forecasting compared to no priors. No significant improvement was seen when priors from TF-target interactions based on *in vitro* TF binding (filtered for open chromatin) (DAP-seq)(O'Malley et al. 2016) were used.

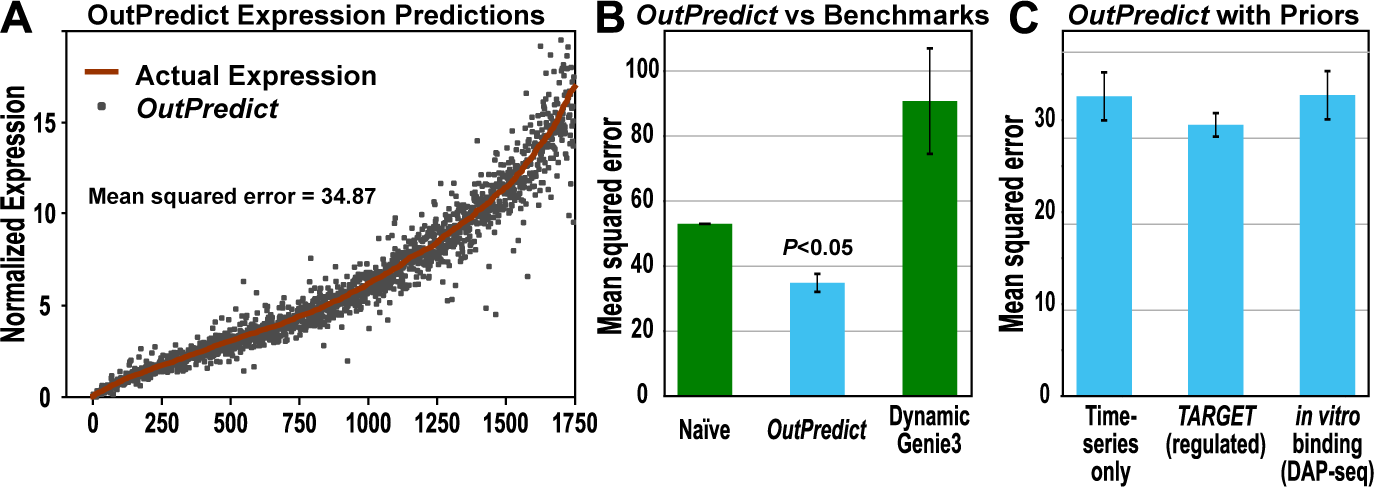


Fig. X – Comparison of time-series forecasting for shoot NxTime data A) Comparison of predicted gene expression using OutPredict (grey dots) compared to actual expression (red line) at the left-out time point. The accuracy of forecasting is measured by calculating the Mean Squared Error. B) *OutPredict* performs better (*P*<0.05) than benchmark approaches including naïve and Dynamic Genie3. C) The incorporation if priors from *TARGET* improve the performance of *OutPredict* compared to the time-series alone, whereas TF-targets identified in vitro do not help with forecasting gene expression.

**Interpretations and Expected Outcomes:** We found that *OutPredict* outperformed other state-of-the-art algorithms in the task of gene expression forecasting. This included not only Arabidopsis shoot time-series N-response data (Fig. X), but also time-series and steady-state data from bacteria, yeast, and *in silico* (DREAM). The absence of TF partners and chromatin features *in vitro* may explain why direct regulated TF-target interactions improved the forecasting predictions better than targets that are bound by a TF *in vitro*. The data from Aim 2 will allow us to test whether transient TF-target interactions which we can capture in root cells using *TARGET* are able improve the forecasting performance of *OutPredict* when used as priors. Because *OutPredict* is able to identify the TFs that are most influential for the expression of each target gene *in planta*, such TFs will be targeted for *in planta* validation in Aim 3C.

**Potential Problems and Alternative Approaches:** Currently, we have TF-regulation and TF-binding information for less than 40/258 N-responsive TFs, and a very small fraction of all ~2,000 TFs in Arabidopsis. This limits the amount of prior information available to use in *OutPredict*. This problem will be greatly diminished as we acquire data about more TFs using the *TARGET* system (Aims 1 & 2). [Dennis thinks that if we’re really tight for space, we can eliminate this part:] While *OutPredict* has thus far been better than other tools at predicting gene expression for data from several different sources, e.g. Arabidopsis, bacteria, yeast, and *in silico*, it is possible that other algorithms could perform better than *OutPredict* on another dataset. Therefore, it will be important to benchmark the performance of *OutPredict* against state-of-the-art algorithms, such as Dynamic Genie3(Huynh-Thu and Geurts 2018), for each set of data.

**Aim 3C: *In planta* validation of influential TFs predicted by Network Walking and OutPredict**

**Hypothesis:** Aims 3A and 3B will identify influential TFs that are important for mediating the regulation of the N response. We hypothesize that overexpression of these TFs will have a strong impact on NUE traits and other N-related phenotypes such as root growth, as we have seen for two TFs that we have conditionally expressed *in planta* thus far – i.e. CRF4 (Varala 2018) and TGA1 (Brooks 2019).

**Approach:** We will generate β–estradiol inducible overexpression lines (Coego et al. 2014) for 10-20 of the TFs identified using the *Network Walking* and *OutPredict* methods to be most influential on N-uptake and assimilation. These lines will be generated as described in Aim 2C. Once 2-3 stable transgenic lines are obtained for a TF, NUE phenotypes (e.g. 15N uptake, total N content), growth, and root architecture traits will be assessed for plants grown in the presence and absence of β–estradiol. If lines for a TF show a significant difference in one or more of these traits when the TF is overexpressed, further characterization, such as RNA-seq, will be used to determine how the N-signaling network is affected.

**Interpretations and Expected Outcomes:** With more than 250 transcription factors responding to N in the first 2 hours of N-exposure, methods to identify those that are the best candidates for manipulation to increase NUE is a difficult but important task. We believe that our combination of innovative experimental and computational approaches will greatly facilitate this task, as demonstrate by our preliminary results showing an N-dependent increase in biomass for lines overexpressing *TGA1* (Fig. X). This TF responds to N in both roots and shoots (Varala et al. 2018) and our *TARGET* assay and Network Walking approach revealed it regulated a significant number of N-response genes, including many TFs(Brooks et al. 2019). Identification and validation of influential TFs may one day guide strategies for generating crops that are more efficient at taking up or utilizing N, which will reduce the need for N-based fertilizers significantly impact the environment and human health.

Bonneau, Richard, David J. Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S. Baliga, and Vesteinn Thorsson. 2006. 'The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo', *Genome Biol*, 7: R36.

Brooks, M. D., J. Cirrone, A.V. Pasquino, J. M. Alvarez, J. Swift, S. Mittal, C.-L. Juang, K. Varala, R. A. Gutierrez, G. Krouk, D. Shasha, and G. M. Coruzzi. 2019. 'Network Walking charts transcriptional pathways for dynamic nitrogen signaling using validated and predicted genome-wide interactions.', *Nat Commun*.

Coego, A., E. Brizuela, P. Castillejo, S. Ruiz, C. Koncz, J. C. del Pozo, M. Pineiro, J. A. Jarillo, J. Paz-Ares, and J. Leon. 2014. 'The TRANSPLANTA collection of Arabidopsis lines: a resource for functional analysis of transcription factors based on their conditional overexpression', *Plant J*, 77: 944-53.

Huynh-Thu, Vân Anh, and Pierre Geurts. 2018. 'dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data', *Scientific reports*, 8: 3384.

Mirowski, P., and Y. LeCun. 2009. 'Dynamic Factor Graphs for Time Series Modeling', *Machine Learning and Knowledge Discovery in Databases, Pt Ii*, 5782: 128-43.

O'Malley, R. C., S. C. Huang, L. Song, M. G. Lewsey, A. Bartlett, J. R. Nery, M. Galli, A. Gallavotti, and J. R. Ecker. 2016. 'Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape', *Cell*, 166: 1598.

Varala, Kranthi, Amy Marshall-Colón, Jacopo Cirrone, Matthew D. Brooks, Angelo V. Pasquino, Sophie Léran, Shipra Mittal, Tara M. Rock, Molly B. Edwards, Grace J. Kim, Sandrine Ruffel, W. Richard McCombie, Dennis Shasha, and Gloria M. Coruzzi. 2018. 'Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants', *Proceedings of the National Academy of Sciences*, 115: 6494-99.