RESEARCH STATEMENT

Tracey Oellerich (toelleri@gmu.edu)

My research primarily focuses on using data-driven methods to infer dynamics. I primarily work towards constructing models related to biological pathways, although the methods developed can be applied to other types of networks with an underlying dynamical system. Furthermore, I used data-driven methods to infer symbolic forms for underlying conservation laws while keeping the amount of required data to a minimum. This can then be combined with the my method or other machine learning techniques to infer the dynamical system while maintaining the system's conservative nature. Once the dynamics are known, features such as adaptation can be tested. My work presents a scope of adaptation into a graph theory and an algebraic geometry framework, making it accessible and useful to a broader audience. I am interested in uncovering why biological networks behave the way they do and how this can be used to help inform other sciences. These results have applications in disease research, such as for cancers and drug therapies, or can be applied in other fields such as control theory. This research has lead to three publications so far: (1) inferring dynamics from biological data [1], (2) inferring conservation laws[2] and (3) extension of existing adaptation criteria [3].

Inferring Dynamics [1]

My work focuses on recovering the dynamics for biological networks from data [1], thus helping to eliminate utilizing assumptions to construct the model. My inspiration for this work began with the Sparse Identification of Nonlinear Dynamics (SINDy) method proposed by [4]. The SINDy algorithm presents a method for automating the discovery of the governing equation and takes advantage of the assumption that many dynamical systems, $\frac{d}{dt}\mathbf{x} = \mathbf{f}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, have dynamics with only a few active terms in the space of all possible right-hand side functions.

Consider time-series data $\mathbf{X} = [\mathbf{x}(t_1), \ldots, \mathbf{x}(t_N)]^T \in \mathbb{R}^{N \times n}$ harvested from experiments and assuming the structure of the dynamical system is a generalized linear model: $\mathbf{f}_k(\mathbf{x}) \approx \boldsymbol{\Theta}(\mathbf{x})\boldsymbol{\xi}_k$, $k = 1, \ldots, n$ where $\boldsymbol{\Theta}(\mathbf{x}) \in \mathbb{R}^{1 \times l}, \boldsymbol{\xi}_k \in \mathbb{R}^{l \times 1}$, l is the number of candidate nonlinear functions in $\boldsymbol{\Theta}(\mathbf{x})$, and $\boldsymbol{\xi}_k$ contains the fewest nonzero terms as possible. Nonzero entries of the sparse vector $\boldsymbol{\xi}_k$ correspond to the active terms in the resulting dynamical system. Here, $\boldsymbol{\Theta}(\mathbf{x})$ refers to the library of candidate nonlinear functions constructed from the data: $\boldsymbol{\Theta}(\mathbf{X}) = \begin{bmatrix}\mathbf{1} & \mathbf{X} & \mathbf{X}^2 & \ldots & \mathbf{X}^d & \ldots & \sin(\mathbf{X}) & \ldots\end{bmatrix}$. At this stage, SINDy uses l_1-normalized sparse regression on $\boldsymbol{\xi}_k = \arg\min_{\boldsymbol{\xi}_k'} \|\dot{\mathbf{X}}_k - \boldsymbol{\Theta}(\mathbf{X})\boldsymbol{\xi}_k'\|_2 + \lambda\|\boldsymbol{\xi}_k'\|_1$, where λ is used to enforce sparsity, to obtain the underlying dynamics. While powerful, choosing an optimal λ can prove computationally difficult.

I propose an alternative approach which avoids enforcing sparsity by taking advantage of the underlying algorithm for solving a least squares problem with non-negativity constraints. Non-negative Least Squares (NNLS)[5] proved to be the solution and responds well to noise and performed effectively even in a low data environment. Consider instead an approximation of $\mathbf{f}(\mathbf{x})$ using a generalized linear model:

$$\mathbf{f}_k(\mathbf{x}) \approx \begin{bmatrix}\boldsymbol{\Theta}(\mathbf{x}) \\ -\boldsymbol{\Theta}(\mathbf{x})\end{bmatrix} \boldsymbol{\omega}_k \tag{1}$$

where $\boldsymbol{\Theta}$ is a library of candidate nonlinear functions constructed from the data and $\boldsymbol{\omega}_k \geq 0$ contains the fewest terms as possible. Inputting the time series data and applying NNLS, the minimization problem is now:

$$\boldsymbol{\omega}_k = \underset{\boldsymbol{\omega}_k' \geq 0}{\arg\min} \left\| \begin{bmatrix}\boldsymbol{\Theta}(\mathbf{X}) \\ -\boldsymbol{\Theta}(\mathbf{X})\end{bmatrix} \boldsymbol{\omega}_k' - \dot{\mathbf{X}}_k \right\|_2 \tag{2}$$

where the top entries of $\boldsymbol{\omega}_k$ will correspond to positive coefficients in the recovered dynamics and the bottom entries are the negative.

This formalism works well for systems containing dynamics that can be written in the form $\Theta\xi$, however, it is ineffective for those containing rational functions, such as Michaelis–Menten kinetics. Both

Implicit SINDy[6] and SINDy-PI[7] alter the classic SINDy algorithm to infer rational dynamics. In similar fashion, I extend the NNLS approach to allow for dynamical systems of the form $\frac{d}{dt}x_k(t) = \frac{f_N(\mathbf{x})}{f_D(\mathbf{x})}$ where $f_N(\mathbf{x})$ and $f_D(\mathbf{x})$ represent the numerator and denominator polynomials in the state variable $\mathbf{x} \in \mathbb{R}^n$ respectfully. Now $f_{N,k}(\mathbf{x})$ and $f_{D,k}(\mathbf{x})$ can be approximated by generalized linear models: $f_{N,k} \approx \Theta_N(\mathbf{x})\omega_{N,k}$ and $f_{D,k} \approx \Theta_D(\mathbf{x})\omega_{D,k}$, where $\Theta_N(\mathbf{x}), \Theta_D(\mathbf{x})$ are the candidate function libraries and $\boldsymbol{\omega}_{N,k}, \boldsymbol{\omega}_{D,k}$ are the corresponding coefficients. Therefore:

$$\Theta_N(\mathbf{x})\omega_{N,k} - \Theta_D(\mathbf{x})\omega_{D,k}\dot{x}_k = 0 \tag{3}$$

In order to apply NNLS and avoid a null space problem, assume the coefficient associated with the $x_k\dot{x}_k$ term is 1 and thus solve:

$$x_k\dot{x}_k = \Theta_N(\mathbf{x})\omega_{N,k} - \tilde{\Theta}_D(\mathbf{x})\omega_{D,k}\dot{x}_k \tag{4}$$

where $\tilde{\Theta}_D$ is the Θ_D matrix with the column corresponding to x_k removed. Alternatively, a different coefficient can be set to 1 and the process will be similar. The optimal $\omega_{N,k}$ and $\omega_{D,k}$ can now be found using a similar method as before.

Inferring Conservation Laws [2]

While working on inferring dynamics, it became clear that many contained underlying conservation laws. Building upon the ideas proposed in [8], I have developed a robust data-driven computational framework that automates the process of identifying the number and type of the conservation law(s) while keeping the amount of required data to a minimum. As with my approach using NNLS, I consider a dynamical system $\frac{d}{dt}\mathbf{x} = \mathbf{f}(\mathbf{x})$ but now include a potential conservation law, $g(x) = C$ (if multiple exist, subscripts will be used). As before, let $\mathbf{X} = [\mathbf{x}(t_1), \ldots, \mathbf{x}(t_N)]^T \in \mathbb{R}^{N\times n}$ be experimental data collected for the system with associated derivative $\dot{\mathbf{X}}$. Rather than linearizing the dynamics, the conservation law will be rewritten as: $g(\mathbf{x}) = C \approx \Theta(\mathbf{X})\boldsymbol{\xi}$, where $\Theta(\mathbf{x})$ is a symbolic library of non-constant candidate functions, $\{\boldsymbol{\theta}_i(\mathbf{x})\}_{i,\ldots p}$. The derivative with respect to time can then be taken to obtain: $0 = \frac{d}{dt}(\Theta(\mathbf{x}))\boldsymbol{\xi} = \Gamma(\mathbf{x}, \dot{\mathbf{x}})\boldsymbol{\xi}$. Applying the data to the problem yields: $\min_{\boldsymbol{\xi}\neq\mathbf{0}} \|\Gamma(\mathbf{X}, \dot{\mathbf{X}})\boldsymbol{\xi}\|$.

The problem now equates to finding a non-trivial null space for $\Gamma(\mathbf{X}, \dot{\mathbf{X}})$. This can easily be done by computing the singular value decomposition (SVD) and identifying all right singular vectors associated with 0 or close to 0 singular values. Robustness of the proposed methodology is based on the provable stability of the singular values and singular vectors to the level of noise present in the data. While this method can determine if the system contains a conservation law in terms of library functions $\boldsymbol{\theta}_i(\mathbf{x})$, it requires the user to at least have all the correct library functions and does not eliminate the possibility of over-fitting the model. Building upon this approach, I propose an algorithmic approach which allows the user to cycle through multiple library configurations and outputs the optimal form, if one exists.

Consider a set of distinct Θ-libraries, $\Phi = \{\Theta^{(1)}, \ldots, \Theta^{(k)}\}$ where $\Theta^{(i)} \in \mathbb{R}^{p_i}$. For each $\Theta^{(i)}$, find the corresponding $\Gamma^{(i)}$-library and it's SVD, $U^{(i)}S^{(i)}(V^{(i)})^T$. Each $S^{(i)}$ contains the singular values $\sigma_1^{(i)} \leq \ldots \leq \sigma_{p_i}^{(i)}$ on the diagonal. Let j denote the index of the first singular value below a predefined cutoff such that $\sigma_j^{(i)}, \ldots, \sigma_{p_i}^{(i)} < \sigma_{\text{cutoff}}$. Let $\text{count}^{(i)} = \text{length}\begin{bmatrix}\sigma_j^{(i)} & \ldots & \sigma_{p_i}^{(i)}\end{bmatrix} = p_i - j + 1$ if there exists at least one σ_i.

Each $\delta^{(i)}$ measures the discrepancy between the singular values which approximate the matrix and those which contribute to the null space. Optimal libraries will have a large $\delta^{(i)}$, indicating that there is a clear distinction between the two sets. Small $\delta^{(i)}$ indicate that $\sigma_{j-1}^{(i)}$ could have potentially been included in the other set had the threshold value allowed it. This process is outlines in Figure 1. In [2], several benchmark examples are tested using low data and added noise.

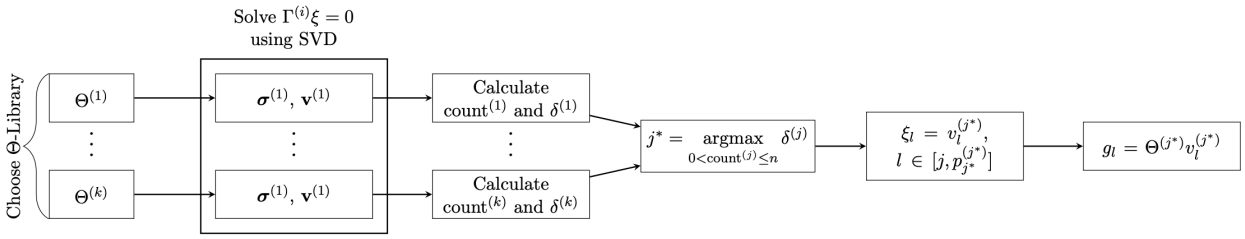

Figure 1: Flowchart detailing the process for selecting the optimal Θ-library.

Robust Perfect Adaptation [3]

Once there is a known dynamical system for a biological network, it can be analyzed and features can be identified. The second aspect of my work focuses on understanding and extending the notion of Robust Perfect Adaptation (RPA). Adaptation in the sense of asymptotic tracking of a 'set-point,' has been widely explored in the literature [9, 10] at various levels ranging from an individual cell to the whole-organism level in mammals. At the cellular level, several types of adaptation have been studied in previous works, including perfect adaptation [11], fold-change detection (FCD) [12], absolute concentration robustness [13], homeostasis [14], and robust perfect adaptation [15]. All of these concepts share adapting behavior, although they highlight certain specific types of adaptive behavior. My focus will be on understanding and extending the criteria under which a biological network will exhibit RPA. *Robust perfect adaptation (RPA)* refers to the property of a biological system to return to the same activity level following any persistent change to the incoming signal received at the input node, without the need for fine-tuning of parameters [15]. There are several reported instances of RPA occurring in biological systems, such as in bacterial chemotaxis [16, 17, 18], EGFR-regulated signaling pathways ([11, 19]), and transcription networks [20].

Consider a system with state space denoted as $\mathcal{P} \subset \mathbb{R}^N$. The state space consists of N nodes, say $\mathbf{P}(t) = [P_1(t), \ldots, P_N(t)]^T \in \mathcal{P}$, representing the interacting molecules of interest, such as proteins, RNA transcripts, genes. Let $\mathcal{U} \subset \mathbb{R}^M$ be the *input* space and $\mathbf{U}(t) = [U_1(t), U_2(t), \ldots, U_M(t)]^T \in \mathcal{U}$ be the time dependent input to the system. Consider the nonlinear dynamical system (5) for this system.

$$\frac{dP_i}{dt} = f_i(\mathbf{U}, \mathbf{P}) \qquad i = 1, ..., N \tag{5}$$

For a system to exhibit RPA, the following two conditions must be met:

$$\det(M_{IO}) = 0 \tag{6}$$

$$\det(D_{\mathbf{P}}\mathbf{F}) \neq 0 \tag{7}$$

where (6-7) is precisely the *RPA equation* and *RPA constraint* as defined in [15]. The RPA criteria defined above fails to account for systems in which the Jacobian at the network steady state is singular. To this end, I have developed criteria which extends the notion of RPA to account for a system in which $\det(D_{\mathbf{P}}\mathbf{F}) = 0$. The new criteria relies on computing a reduced SVD of the Jacobian matrix and using the Moore-Penrose pseudoinverese to solve the problem. The theorem containing this new criteria is stated and proven in [3].

Now, it just remains to identify when a Jacobain will be singular at the network steady state. I address this in [3] where several cases are proven to contain a singular Jacobian at the steady state. These cases include networks which include linear and non-linear conservation laws and network structures which are modelled by an equation involving two or more proteins, such as a molecular compound.

Future Directions

As I have worked on these problems, I have seen that there are many areas where the current work can be extended. Below I provide some future research ideas I will be exploring.

Adaptation from an Algebraic Geometry Viewpoint

Currently, the criteria for adaptation is defined in linear algebra terms; however, there is room to extend these results using an algebraic geometry approach. Future work includes extending the results for the conservation laws and special network structures to an algebraic interpretation. In recent years, many mathematicians have approached systems biology using algebraic geometry. In her paper, Dickenstein [21] provides a survey of the recent applications of algebraic geometry in the understanding of systems biology. By extending these conditions to a more algebraic interpretation, there will be a connection between networks that exhibit RPA as presented in Araujo et al. [15] and those which require the generalized RPA condition.

Comparison and Extension of Methods for Inferring Dynamics

In my work, I have both seen and employed various methods for inferring dynamics. One area of exploration is to consider the benefits and costs for each method and where one may prefer to use a specific method. Furthermore, I am exploring encoding our method into a neural-network like system which will use each new guess as the initial for the next round of optimization. Finally, I am working to reduce the amount of numerical approximation needed to initialize the algorithm by employing integrals in the optimization routine instead of approximating derivative values.

Machine Learning Approach for Discovering Conservation Laws

At present, the algorithm presented in Figure 1 is limited by what libraries the user chooses to compare. One could instead construct a global library of possible function forms and iterate over the power set, excluding the empty set and sets containing a single state variable. While this approach will more accurately infer all possible conservation laws, assuming the correct functions are in the global library, it is significantly more computationally demanding for if $\Theta^T_{\text{global}} \in \mathbb{R}^k$, $2^k - 1 - k$ sub-libraries should be considered in the algorithm. On a similar thread, for systems with multiple conservation laws originating from different sub-libraries (say one linear and one with only 3rd order terms), δ can be significantly close. Currently, user supervision would be required if there is a close decision. Future algorithms will be designed to output a set of possible libraries with sufficiently close δ values.

Modeling of Biological Pathways

The overall goal of this research is to apply it to biological data and test a protein-protein interaction network for adaptation. Thanks to collaboration with Dr. Mariaelena Pierobon at the Center for Applied Proteomics and Molecular Medicine (CAPMM), GMU, I have access to data for the mitogen-activated protein kinase (MAPK) pathway, a pathway which plays a role in various cancers when dysregulated. The MAPK signaling pathway is a key regulator of different cellular processes, such as gene expression and cellular growth, and deciphering the mechanisms of action and regulation for the MAPK pathway remains a challenge from a biological perspective.

While there is a growing interest in exploring these dynamic interactions from a biological prospective, modeling of systems such as the MAPK pathway presents multiple challenges due to a lack of populated data and the combinatorial increase in complexity when considering the full-scale reaction network. This work proposes an innovative data-driven multidisciplinary approach that combines quantitative experimental measurements of network dynamics with mathematical modeling to devise novel multi-scale pattern-oriented methods for dissecting and understanding signal transduction-based mechanisms in complex biological samples. This will push the boundaries of sparse dynamics identification methods by enhancing them with mesoscale network characterization techniques. This novel framework has a potential to find broad applicability for illuminating basic molecular mechanisms associated with different pathological processes, uncovering target-able interactions within these networks, and predicting network adaptation mechanisms and response to perturbations with applications across different fields of biomedical disciplines and mathematics.

References

[1] T. Oellerich, M. Emelianenko, M. Pierobon, and E. Baldelli., "Learning biological networks dynamic from data," *In preparation*, 2023.

[2] T. Oellerich and M. Emelianenko, "Towards robust data-driven recovery of conservation laws with limited data," *In preparation*, 2024.

[3] T. Oellerich, M. Emelianenko, L. A. Liotta, and R. P. Araujo, "Biological networks with singular Jacobians: their origins and adaptation criteria," *bioRxiv 2021.03.01.433197*, 2021.

[4] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the national academy of sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[5] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. SIAM, 1995.

[6] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Inferring biological networks by sparse identification of nonlinear dynamics," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 2, no. 1, pp. 52–63, 2016.

[7] K. Kaheman, J. N. Kutz, and S. L. Brunton, "SINDy-PI: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics," *Proceedings of the Royal Society A*, vol. 476, no. 2242, p. 20200279, 2020.

[8] E. Kaiser, J. N. Kutz, and S. L. Brunton, "Discovering conservation laws from data for control," in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 6415–6421, IEEE, 2018.

[9] R. P. Araujo and L. A. Liotta, "Design principles underlying robust adaptation of complex biochemical networks," in *Computational Modeling of Signaling Networks*, pp. 3–32, Springer, 2023.

[10] O. Hoeller, D. Gong, and O. D. Weiner, "How to understand and outwit adaptation," *Developmental cell*, vol. 28, no. 6, pp. 607–616, 2014.

[11] J. E. Ferrell, "Perfect and near-perfect adaptation in cell signaling," *Cell systems*, vol. 2, no. 2, pp. 62–67, 2016.

[12] O. Shoval, U. Alon, and E. Sontag, "Symmetry invariance for adapting biological systems," *SIAM Journal on Applied Dynamical Systems*, vol. 10, no. 3, pp. 857–886, 2011.

[13] G. Shinar and M. Feinberg, "Structural sources of robustness in biochemical reaction networks," *Science*, vol. 327, no. 5971, pp. 1389–1391, 2010.

[14] Z. F. Tang and D. R. McMillen, "Design principles for the analysis and construction of robustly homeostatic biological networks," *Journal of theoretical biology*, vol. 408, pp. 274–289, 2016.

[15] R. P. Araujo and L. A. Liotta, "The topological requirements for robust perfect adaptation in networks of any size," *Nature communications*, vol. 9, no. 1, p. 1757, 2018.

[16] T.-M. Yi, Y. Huang, M. I. Simon, and J. Doyle, "Robust perfect adaptation in bacterial chemotaxis through integral feedback control," *Proceedings of the National Academy of Sciences*, vol. 97, no. 9, pp. 4649–4653, 2000.

[17] N. Barkai and S. Leibler, "Robustness in simple biochemical networks," *Nature*, vol. 387, no. 6636, pp. 913–917, 1997.

[18] U. Alon, M. G. Surette, N. Barkai, and S. Leibler, "Robustness in bacterial chemotaxis," *Nature*, vol. 397, no. 6715, pp. 168–171, 1999.

[19] S. Sasagawa, Y.-i. Ozaki, K. Fujita, and S. Kuroda, "Prediction and validation of the distinct dynamics of transient and sustained ERK activation," *Nature cell biology*, vol. 7, no. 4, pp. 365–373, 2005.

[20] L. Goentoro, O. Shoval, M. W. Kirschner, and U. Alon, "The incoherent feedforward loop can provide fold-change detection in gene regulation," *Molecular cell*, vol. 36, no. 5, pp. 894–899, 2009.

[21] A. Dickenstein, "Biochemical reaction networks: An invitation for algebraic geometers," in *Mathematical congress of the Americas*, vol. 656, pp. 65–83, Contemp. Math, 2016.