# Statistical Mechanical Framework for Predicting Cellular Response

Thesis by
Lila Forte

In Partial Fulfillment of the Requirements for the
degree of
Master of Science

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2016

# ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Tom Miller. His guidance and enthusiasm helped me to overcome many obstacles during the past two years. I am grateful for all that I have learned from him, the opportunities I was provided, and the support he has given me. I would also really like to thank Prof. Justin Bois, not only for the intellectually stimulating conversations and courses, but his tireless attention to his students' learning and support of their interests. Also, I would especially like to thank my main collaborator, Connie Wang. Her support and mentorship were invaluable throughout this process. All the work done for this thesis would not have been even close to possible without her extremely patient and insightful guidance. I always enjoyed our conversations and will miss our strolls and chats. Last, but not least I would like to thank my husband, Neal, for being a rock in his unconditional love and support in all my endeavors.

# ABSTRACT

Developments in singe-cell analysis techniques allow simultaneous high-resolution measurements of cellular component copy number and variation within a cell population. These data provide a probability distribution for all possible states of the cell, as determined by the measured component copy number per cell. We have developed a highly-flexible, theoretical statistical mechanical framework that uses single-cell cellular component data to model the evolution of the probability distribution of those components in a cell in response to an external, physical or molecular, perturbation. This framework uses Bayesian inference to compare potential functional descriptions of how the perturbation couples to the system, and to determine the uncertainty in the parameter estimations given the data. We have applied this methodology to study the impact of changes in oxygen partial pressure on the behavior of glioblastoma multiform cancer cells. We find that oxygen concentration couples not only to individual proteins, but effects the underlying effective interactions between the studied proteins as well. The underlying effective interactions were found to couple linearly to the system, indicating a simple proportional change in the protein network across oxygen concentrations. This description of the system provides improved predictive capabilities for describing the probability distribution of the measured cellular components across a wider range of perturbation conditions than previous methods. Additionally, we apply this methodology to show how it could be used to predict effects in difficult experimental perturbation regimes, identifying undruggable regimes, as well as the result of knocking our individual or combinations of proteins or protein interactions.

# TABLE OF CONTENTS

# LIST OF ILLUSTRATIONS

# LIST OF TABLES

*Chapter 1*

# INTRODUCTION

## 1.1 New Directions Using New Data Collection Methods

With advances in quantitative single-cell analysis techniques, we now have the ability to observe the steady-state probability distribution for copy number of most cellular components (e.g. proteins, mRNAs, metabolites, etc.) under specific conditions. This distribution contains information beyond population averages, describing the effective interactions between multiple cellular components, however the challenge is how to extract this information. The steady-state that can be obtained is defined by nearly constant average concentration of a cellular component over time after the application of a stimulus and given time to relax in those conditions. Following a perturbation, such as the addition of a drug or changes in resources available, cells may come to a different steady-state than before the stimulus was applied [1]. Novel technologies now exist that can measure simultaneous cellular component copy numbers to obtain the multi-dimensional steady state probability distribution over a range of perturbation conditions [2].

Stochasticity in gene expression, due to intrinsic and extrinsic noise [3], gives rise to fluctuations in protein levels and causes clonal populations of cells to exhibit phenotypic variation at steady state [1, 4, 5]. The steady-state distribution of cellular component copy number for an ensemble of genetically identical cells provides the most probable copy numbers (average), as well as fluctuations in copy number for cells in the population. The noise in expression levels of cellular components, or fluctuations in a population, have been associated with drug resistance, variation in cell-to-cell response to external stress and cell differentiation [4, 6–8]. Therefore, a clear understanding of the steady-state copy number distribution would help to predict when and what response may occur due to an external stimuli.

Master equation formulations, which use known parameters for rates of production and dilution due to cell division, have been used to predict the steady state distribution for individual cellular components for a population of cells and then checked with single cell analysis at one set of experimental conditions [9–11]. An inverted method can obtain parameters given a steady state distribution under one external condition, giving insight into the underlying dynamics of the system [10].

Therefore, understanding how the steady state distribution is coupled to a perturbation will give more information on dynamics of the system in response to that perturbation. However, since it is often experimentally intractable to take single-cell measurements of every possible perturbation condition, we instead obtain “snapshots” under specific external conditions. The challenge is to extract from those few measurements, how the steady-state distribution evolves as a function of a specific perturbation, such as changing the concentration of a drug or steadily decreasing cellular resources.

Specifically, we are interested in extracting from the few isolated experimental measurements how the copy number in a cell, that is the probability distribution of the number density, evolves as a perturbation is increased or decreased. In some families of active matter systems, thermodynamic equilibrium approaches have been used, taking a number density view of the system [12, 13]. Homogeneous steady states have been seen in active matter systems to arise from the effective generation of long-range interactions in the system, even if local density variations are ignored [14]. We therefore approach the system from a thermal equilibrium perspective, where the “free energy” of the system is relative to the number density for each cellular component in the cell. We are not concerned with the mechanistic or physical interactions of the proteins, but instead how the effective interactions and fundamental characteristics of the measured proteins, captured in the single cell data, evolve as a function of the perturbation.

## 1.2 Overview

We apply an equilibrium statistical mechanical framework to model the evolution of the probability distribution of cellular components in a cell in response to an external perturbation. We use the number density for each cellular component to define the state of the cell. Each cellular component interacts with every other measured cellular component with an associated interaction energy, providing an effective long-range interaction among the components. Additionally, the individual properties, that is the chemical potentials, for each cellular component determine the total energetics for the cell. The stable, grand canonical statistical distribution for a population of clonal cells describes the possible states of these cells in the population, as determined by their volume, temperature and internal properties or interactions, and contains information on the average and fluctuations of those cellular components. We propose simple Hamiltonian forms that describe the functional coupling of the perturbation to the steady state probability distribution, and from the data extract the best available parameterization using Bayesian inference. The functional coupling can be utilized to predict properties of the effective protein interaction network as a function of the perturbation.

### 1.3 Data: Hypoxia Effect in Cancer Cells

We apply our methodology to a data set describing the effect of hypoxia on cancer cells. In the center of most tumors, particularly in solid organ cancers, hypoxic conditions with oxygen partial pressures $\lesssim 3\%\ p_{\mathrm{O}_2}$ arise due to rapid cell growth, constriction or leaking of blood vessels, increased interstitial pressure, or edema[15, 16]. The cancer cells in these hypoxic micro environments change their metabolism through HIF and mTOR signaling pathways in order to survive [17]. Tumors with hypoxic conditions can exhibit increased proliferation, aggression, and even a decrease in response to drug therapies[18, 19].

The molecular mechanisms in this system are well known, but the quantitative effects of oxygen on these signaling networks are less clear. Heath *et. al.* performed an experimental study of the mTOR and HIF signaling networks using single-cell barcode chip (SCBC) techniques [19]. We apply our statistical mechanical framework to their data set to investigate the effect of changes in the oxygen partial pressure on protein signaling networks in glioblastoma multiform (GBM) cancer cells.

The SCBC method is a microfluidics platform that quantifies a panel of proteins from statistical numbers of single cells [2]. For this experiment, cells are loaded into chambers on a microchip, about 1 cell/microchamber. The cells on the chip are incubated at a specific oxygen concentration for 7 hr. The cells are then lysed, and secreted and intracellular proteins are captured using an antibody mixture with fluorescent probes. Copy numbers for each protein are inferred from the fluorescence intensity of each micro chamber. For this experiment, seven key functional proteins in the HIF and mTOR signaling pathways were chosen for the panel and single-cell data was collected at 21%, 3%, 2%, 1.5%, and 1% $p_{\mathrm{O}_2}$. Under each oxygen condition, ~100 single cells were analyzed, providing protein copy number data per cell for the seven measured proteins (Fig 1.1)[19].

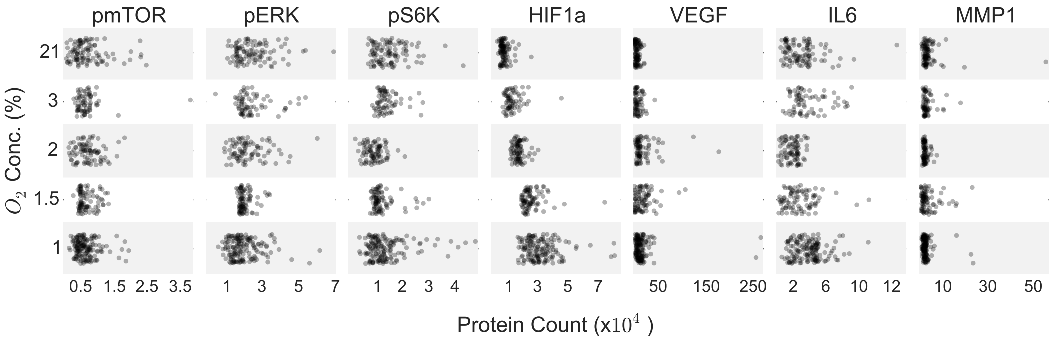

Figure 1.1: Measurements of single-cell protein data across tested oxygen concentration range. Each dot represents the copy number measured for that protein in an individual cell.

*Chapter 2*

# METHODOLOGY

## 2.1 Theory

We define a Hamiltonian for an $n$-component system, which is a function of the copy number, $N_i$, of each component, $i$ and the perturbation, $\lambda$. We consider the state of a cell can be fully described by the copy number of its components.

$$H(\{N_i\},\lambda) = H_o(\{N_i\}) + H_1(\{N_i\},\lambda) \tag{2.1}$$

The first term, $H_o$, is a reference state Hamiltonian that describes the system in absence of the perturbation, while $H_1$ contains the physics that describes how the perturbation functionally couples to the system. With no prior knowledge as to the functional form of $H_1$, this methodology provides a framework in which to extract this information from the data.

The reference state Hamiltonian for an $n$-component system is defined as

$$H_o(\{N_i\}) = \sum_{i,i\geq j}^{n} \alpha_{ij} N_i N_j - \sum_{i}^{n} \mu_i N_i \tag{2.2}$$

where $\mu_i$ is the chemical potential for component $i$ and $\alpha_{ij}$ is the effective pairwise interaction between components $i$ and $j$. Assuming a well mixed solution, such that the density of each cellular component, $\rho_i$, is uniform across the cell with a constant volume, $V$,

$$H_o(\{\rho_i\}) = \sum_{i,i\geq j}^{n} a_{ij} \rho_i \rho_j V - \sum_{i}^{n} \mu_i \rho_i V \tag{2.3}$$

where $a_{ij} = \alpha_{ij} V$, such that $a_{ij}$ has units $\frac{\epsilon V}{N^2}$ and $\mu_i$ has units $\frac{\epsilon}{N}$, and $\epsilon$ is an energy unit. The effective pairwise interaction parameters, $\{a_{ij}\}$, describe the network of protein interactions, but provide no information on mechanistic or physical interactions. The parameters in $H_o$ describe the fundamental interactions and potentials for the set of cellular components under the reference state conditions.

The functional coupling of a physical external perturbation to the system, $H_1$, such as addition of a drug or changes in cellular resources, is defined as

$$H_1(\{N_i\},\lambda) = \sum_{i,i\geq j}^{n} \frac{f_{ij}(\lambda)}{V} N_i N_j - \sum_{i}^{n} g_i(\lambda) N_i \tag{2.4}$$

where the perturbation terms, $f_{ij}(\lambda)$ and $g_i(\lambda)$, describe how the perturbation changes the effective pairwise interactions for each pair of components and chemical potentials for each cellular component, respectively. Additionally, $f_{ij}(\lambda)$ and $g_i(\lambda)$ are zero at the reference state conditions ($\lambda_o$), so that under the specified external conditions, $H(\{N_i\}, \lambda_o) = H_o(\{N_i\})$. Due to the form of the Hamiltonian, $H$, the choice of reference state conditions is mathematically arbitrary. Any choice of $\lambda_o$ provides identical information on the functional coupling to the perturbation, $H_1$, the only difference begging that the information on the fundamental interactions and potentials are for that chosen reference state, $H_o$.

We apply this framework to study the effects of hypoxic conditions on glioblastoma multiform cancer cells. We choose a set of candidate hamiltonians, $\{H_1\}$, that describe possible simple functional couplings of oxygen concentration to the system (Table 2.1). As the effective coupling of oxygen to these proteins is unknown, we begin with simple functions commonly seen in biological systems. Additionally, we include three levels of coupling: individual, pairwise, and a combination of the two. We define the reference state as normoxic conditions, 21% oxygen concentration such that the perturbations is zero at $p_{21}$, according to

$$g_i(p_{\mathrm{O}_2}) = g_i^*(p_{\mathrm{O}_2}) - g_i^*(p_{21}) \tag{2.5}$$

$$f_{ij}(p_{\mathrm{O}_2}) = f_{ij}^*(p_{\mathrm{O}_2}) - f_{ij}^*(p_{21}) \tag{2.6}$$

where $g_i^*(p_{\mathrm{O}_2})$ and $f_{ij}^*(p_{\mathrm{O}_2})$ contain the functional coupling of oxygen concentration to the system, and $g_i^*(p_{21})$ and $f_{ij}^*(p_{21})$ have the same functional forms and parameters as $g_i^*(p_{\mathrm{O}_2})$ and $f_{ij}^*(p_{\mathrm{O}_2})$, respectively, but are evaluated at only the reference oxygen concentration.

|  | **Models:** | **Functional Forms of Oxygen Coupling** |  |  |
|---|---|---|---|---|
|  | # : Name | $g_i^*\ (p_{\mathrm{O}_2})$ | $f_{ij}^*\ (p_{\mathrm{O}_2})$ | Parameters : Units |
| Individual | 1 : Lin-X | $k_i p_{\mathrm{O}_2}$ | None | $k_i : \frac{\epsilon}{N p_{\mathrm{O}_2}}$ |
|  | 2 : Exp-X | $b_i \exp(k_i p_{\mathrm{O}_2})$ | None | $k_i : \frac{1}{p_{\mathrm{O}_2}}\ \ b_i : \frac{\epsilon}{N}$ |
|  | 3 : Hill-X | $\frac{b_i}{1+(k_i/p_{\mathrm{O}_2})^{m_i}}$ | None | $k_i : p_{\mathrm{O}_2}\ \ b_i : \frac{\epsilon}{N}\ \ m_i : -$ |
|  | 4 : Logi-X | $\frac{b_i}{1+\exp[-m_i(p_{\mathrm{O}_2}-k_i)]}$ | None | $k_i : p_{\mathrm{O}_2}\ \ b_i : \frac{\epsilon}{N}\ \ m_i : \frac{1}{p_{\mathrm{O}_2}}$ |
|  | 5 : Log-X | $b_i \log(p_{\mathrm{O}_2})$ | None | $b_i : \frac{\epsilon}{N}$ |
|  | 6 : Pow-X | $b_i p_{\mathrm{O}_2}^{m_i}$ | None | $b_i : \frac{\epsilon}{N p_{\mathrm{O}_2}^{m_i}}\ \ m_i : -$ |
| Pairwise | 7 : X-Lin | None | $k_{ij} p_{\mathrm{O}_2}$ | $k_{ij} : \frac{\epsilon V}{N^2 p_{\mathrm{O}_2}}$ |
|  | 8 : X-Exp | None | $b_{ij} \exp(k_{ij} p_{\mathrm{O}_2})$ | $k_{ij} : \frac{1}{p_{\mathrm{O}_2}}\ \ b_{ij} : \frac{\epsilon V}{N^2}$ |
|  | 9 : X-Hill | None | $\frac{b_{ij}}{1+(k_{ij}/p_{\mathrm{O}_2})^{m_i}}$ | $k_{ij} : p_{\mathrm{O}_2}\ \ b_{ij} : \frac{\epsilon V}{N^2}\ \ m_{ij} : -$ |
|  | 10 : X-Logi | None | $\frac{b_{ij}}{1+\exp[-m_{ij}(p_{\mathrm{O}_2}-k_{ij})]}$ | $k_{ij} : p_{\mathrm{O}_2}\ \ b_{ij} : \frac{\epsilon V}{N^2}\ \ m_{ij} : \frac{1}{p_{\mathrm{O}_2}}$ |
|  | 11 : X-Log | None | $b_{ij} \log(p_{\mathrm{O}_2})$ | $b_{ij} : \frac{\epsilon V}{N^2}$ |
|  | 12 : X-Pow | None | $b_{ij} p_{\mathrm{O}_2}^{m_{ij}}$ | $b_{ij} : \frac{\epsilon V}{N^2 p_{\mathrm{O}_2}^{m_i}}\ \ m_{ij} : -$ |
| Combination | 13 : Lin-Lin | $k_i p_{\mathrm{O}_2}$ | $k_{ij} p_{\mathrm{O}_2}$ | see models 1 and 7 |
|  | 14 : Exp-Exp | $b_i \exp(k_i p_{\mathrm{O}_2})$ | $b_{ij} \exp(k_{ij} p_{\mathrm{O}_2})$ | see models 2 and 8 |
|  | 15 : Hill-Hill | $\frac{b_i}{1+(k_i/p_{\mathrm{O}_2})^{m_i}}$ | $\frac{b_{ij}}{1+(k_{ij}/p_{\mathrm{O}_2})^{m_i}}$ | see models 3 and 9 |
|  | 16 : Logi-Logi | $\frac{b_i}{1+\exp[-m_i(p_{\mathrm{O}_2}-k_i)]}$ | $\frac{b_{ij}}{1+\exp[-m_{ij}(p_{\mathrm{O}_2}-k_{ij})]}$ | see models 4 and 10 |
|  | 17 : Log-Log | $b_i \log(p_{\mathrm{O}_2})$ | $b_{ij} \log(p_{\mathrm{O}_2})$ | see models 5 and 11 |
|  | 18 : Pow-Pow | $b_i p_{\mathrm{O}_2}^{m_i}$ | $b_{ij} p_{\mathrm{O}_2}^{m_{ij}}$ | see models 6 and 12 |
|  | 19 : Log-Lin | $b_i \log(p_{\mathrm{O}_2})$ | $k_{ij} p_{\mathrm{O}_2}$ | see models 5 and 7 |

Table 2.1: Proposed functional forms of oxygen coupling.

The grand canonical partition function, $\Xi$, for an $n$-component system is defined as

$$\Xi(\beta, V, \{\mu_i\}, \{a_{ij}\}, \{g_i(\lambda)\}, \{f_{ij}(\lambda)\}) = \sum_{N_1=0}^{\infty} \sum_{N_2=0}^{\infty} \ldots \sum_{N_n=0}^{\infty} e^{-\beta H(\lambda, N_1, N_2, \ldots, N_n)} \tag{2.7}$$

where $\beta = 1/k_B T$, and $k_B$ is the Boltzmann constant. The discrete partition function is not analytically solvable, but we can approximate the partition function as a continuous integral over all possible $n$-component densities, which is exactly solvable (see Appendix A.1 for full derivation of the partition function). The analytic solution is

$$\Xi(\beta, V, \{\mu_i\}, \{a_{ij}\}, \{g_i(\lambda)\}, \{f_{ij}(\lambda)\}) = \left(\frac{V\sqrt{2\pi}}{2}\right)^n \frac{e^{\frac{1}{2}(\mathbf{M}+\mathbf{G})^{\mathbf{T}}(\mathbf{A}+\mathbf{F})^{-1}(\mathbf{M}+\mathbf{G})}}{\sqrt{|\mathbf{A}+\mathbf{F}|}} \tag{2.8}$$

where

$$(\mathbf{M}+\mathbf{G})^{\mathbf{T}} = \beta V \begin{pmatrix} \mu_1 + g_1(\lambda) & \mu_2 + g_2(\lambda) & \dots & \mu_n + g_n(\lambda) \end{pmatrix}$$

$$\mathbf{A}+\mathbf{F} = 2\beta V \begin{pmatrix} a_{11} + f_{11}(\lambda) & \frac{1}{2}(a_{12} + f_{12}(\lambda)) & \dots & \frac{1}{2}(a_{1n} + f_{1n}(\lambda)) \\ \frac{1}{2}(a_{21} + f_{21}(\lambda)) & a_{22} + f_{22}(\lambda) & \dots & \frac{1}{2}(a_{2n} + f_{2n}(\lambda)) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}(a_{n1} + f_{n1}(\lambda)) & \frac{1}{2}(a_{n2} + f_{n2}(\lambda)) & \dots & a_{nn} + f_{nn} \end{pmatrix}$$

such that $\mathbf{A}+\mathbf{F}$ is a symmetric matrix, for which $(\mathbf{A}+\mathbf{F})^{-1}$ and $|\mathbf{A}+\mathbf{F}|$ are the inverse and determinant, respectively.

Importantly, the analytic partition function contains all of the information about the statistical distribution for cellular component copy numbers. Two observable thermodynamic properties describing the ensemble behavior, the average, $\langle N_i \rangle$, and fluctuations, $\langle (\delta N_i)(\delta N_j) \rangle$, of the cellular component copy number, can be derived from this partition function. Evaluation of the first and second derivatives of the partition function, with respect to the conjugate thermodynamic quantity to particle number, $-\beta\mu$, provides predictions for the mean and covariance in cellular component copy number

$$\langle N_i \rangle = V \sum_j (\mu_j + g_j(\lambda))(a_{ij} + f_{ij}(\lambda))^{-1} \tag{2.9}$$

$$\langle (\delta N_i)(\delta N_j) \rangle = \sigma_{ij} = \frac{V}{\beta}(a_{ij} + f_{ij}(\lambda))^{-1} \tag{2.10}$$

where $(a_{ij} + f_{ij}(\lambda))^{-1}$ is an element of the inverse $\mathbf{A}+\mathbf{F}$ matrix (see Appendix A.2 for the $n$-component derivation). These equations show directly how the perturbation affects each observable moment of the steady state probability distribution.

Not only do Eq. 2.9 and Eq. 2.10 provide simple relations of the statistical behaviors and the perturbation, when no perturbation is applied, at the reference state, these equations reduce to

$$\langle N_i \rangle = V \sum_j \mu_j a_{ij}^{-1} \tag{2.11}$$

$$\sigma_{ij} = \frac{V}{\beta} a_{ij}^{-1} \tag{2.12}$$

where $a_{ij}^{-1}$ is an element in $\mathbf{A^{-1}}$. There is a clear 1-to-1 mapping of experimentally observed statistics (i.e. $\langle N_i \rangle$ to parameter values (i.e. $\mu_i$). These statistics, average and covariance in cellular component copy number, can be easily measured with single cell analysis techniques. Therefore, through inversion of Eq. 2.11 and Eq. 2.12, we derive the unique natural parameter values for the chosen reference state from experimental data collected under those conditions

$$\mu_i = \beta^{-1} \sum_j \langle N_j \rangle \sigma_{ij}^{-1} \tag{2.13}$$

$$a_{ii} = \frac{V}{2\beta} \sigma_{ii}^{-1} \tag{2.14}$$

$$a_{ij} = \frac{V}{\beta} \sigma_{ij}^{-1} \tag{2.15}$$

where $\sigma_{ij}^{-1}$ is an element of the inverse covariance matrix.

However, to fully understand the perturbation coupling we must obtain the parameters in the perturbation coupling terms, $g_i(\lambda)$ and $f_{ij}(\lambda)$. Due to the additional parameters present in these terms, the full Hamiltonian system is underdetermined. There is no unique analytical solutions for the parameters derived directly from experimental quantities as was conveniently possible with $H_o$. Therefore, we utilize Bayesian inference to systematically obtain best fit parameters for each candidate Hamiltonian form of coupling of the perturbation to the system, and extract from the data the proper functional coupling of the perturbation to our system.

## 2.2 Calculations

From an experimental data set, $D$, containing $n$-component copy numbers from individual cells over a range of perturbation conditions, we apply Bayesian inference to acquire the most probable functional perturbation coupling and parameterization. The probability of each candidate Hamiltonian (Table 2.1), or model, is evaluated based on the experimental data. Parallel Tempering Markov Chain Monte Carlo

(PTMCMC) is used to efficiently sample the high-dimensional parameter sets for each model.

We use PTMCMC, for each model, $M_i$, to sample the posterior distribution associated with the parameter set, $\gamma_i$, for model $i$, where $\gamma_i$ contains $\{a_{ij}\}$, $\{\mu_i\}$ and any parameters contained in the coupling terms, $\{f_{ij}(\lambda)\}$ and $\{g_i(\lambda)\}$. According to Bayes' theorem, the posterior distribution for the parameter set, $P(\gamma_i|D, M_i)$, is proportional to the likelihood, $P(D|\gamma_i, M_i)$, times the prior probability, $P(\gamma_i|M_i)$. The likelihood, or the probability of the data given $M_i$, is defined by a Boltzmann distribution

$$P(D|\gamma_i, M_i) = \frac{1}{\Xi_i(\beta, V, \gamma_i, M_i)} e^{-\beta H_i(D)} \tag{2.16}$$

where $H_i$ and $\Xi_i$ are the Hamiltonian and partition function for model $i$. For the prior probability, each of the $m$-parameters in $\gamma_i$ is considered independent. We use an uninformative, uniform prior for each individual parameter over a broad range of values

$$P(\gamma_i|M_i) = \prod_{j=1}^{m} P(\gamma_{i,j}|M_i) = \prod_{j=1}^{m} \frac{1}{\gamma_{i,j}^{max} - \gamma_{i,j}^{min}} \tag{2.17}$$

where $\gamma_{i,j}^{max}$ and $\gamma_{i,j}^{min}$ define the range allowed for parameter $\gamma_{i,j}$ (see Appendix B for full ranges of parameter values used for each model). For the sampling, parameters are initialized within this range, and we use the data from the reference state conditions and Eqs. 2.13-2.15 to find a reasonable parameter space to initialize $\{a_{ij}\}$ and $\{\mu_i\}$ and any parameters in $\{f_{ij}(\lambda)\}$ and $\{g_i(\lambda)\}$ such that the change in $\{a_{ij}\}$ and $\{\mu_i\}$ is small to keep the values in physically reasonable ranges initially.

The EMCEE package [20, 21] is used to perform the PTMCMC calculations. Each calculation is run with 20 temperatures, 2,000 walkers for 8,000-10,000 steps until convergence is reached (see Appendix B for convergence criteria discussion). We then obtain the most probable values (modes) and 95% credible regions for each parameter estimate using the marginalized distribution for the lowest temperature PTMCMC.

By sampling many temperatures, we are able to calculate the odds ratio, $O_{ij}$, which we use to compare models $i$ and $j$,

$$O_{ij} = \frac{P(M_i|D)}{P(M_j|D)} = \frac{P(M_i)P(D|M_i)}{P(M_j)P(D|M_j)} = \frac{\int d\gamma_i P(\gamma_i|M_i)P(D|\gamma_i, M_i)}{\int d\gamma_j P(\gamma_j|M_j)P(D|\gamma_j, M_j)} \tag{2.18}$$

We assume *a priori* that all models are equally likely, $P(M_i) = P(M_j)$. This form of the odds ratio results from a convenient mathematical manipulation [22], where for each temperature, $B$, sampled in the PTMCMC we calculate

$$Z_i(B) = \int d\gamma_i P(\gamma_i|M_i)[P(D|\gamma_i, M_i)]^B \tag{2.19}$$

Then we use thermodynamic integration to find

$$\ln P(D|M_i) = \ln Z_i(1) = \int_0^1 dB \langle \ln[P(D|\gamma_i, M_i)] \rangle_B \tag{2.20}$$

According to Bayes' theorem, the calculation of the odds ratio allows us to determine which tested model is most probable given the experimental data. To provide information on how close to an "ideal" model we are, that is have we tested enough models, we propose a metric to quantify the amount of perturbation captured by our perturbation terms, $\{g_i(\lambda)\}$ and $\{f_{ij}(\lambda)\}$, for each model. For this calculation, we check if the analytic reference state parameters found by solving Eqs. 2.13-2.15 for $H_o$ are found within the predicted credible regions from Bayesian analysis for those parameters. We assume that for an ideal model, where all of the perturbation effects on copy number are captured by the perturbation terms, 100% of the analytic parameters from $H_o$ will fall into the predicted credible regions. Our entire workflow to determine the functional form of the perturbation coupling is shown in Figure 2.1.

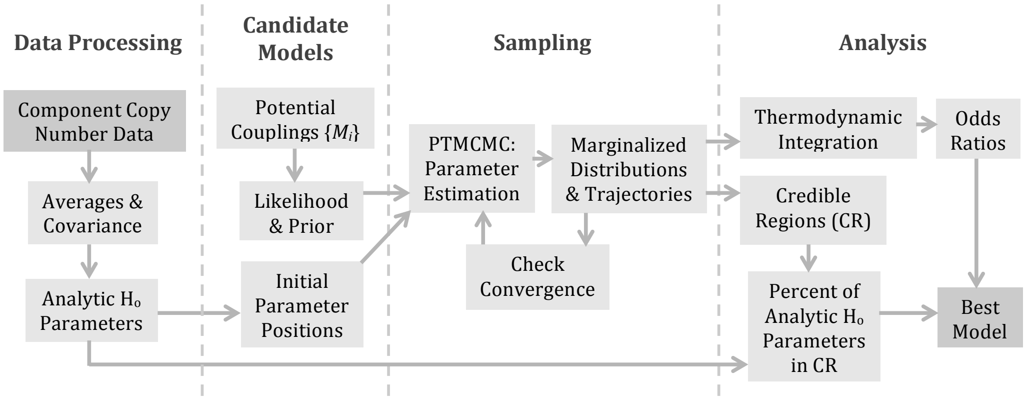
Data Processing
Candidate Models
Sampling
Analysis
Component Copy Number Data
Averages & Covariance
Analytic H₀ Parameters
Potential Couplings {$M_i$}
Likelihood & Prior
Initial Parameter Positions
PTMCMC: Parameter Estimation
Marginalized Distributions & Trajectories
Check Convergence
Thermodynamic Integration
Odds Ratios
Credible Regions (CR)
Percent of Analytic H₀ Parameters in CR
Best Model

Figure 2.1: Workflow for the application of the proposed statistical mechanical framework to extract from single cell data the best functional coupling of a perturbation to the system.

*Chapter 3*

# APPLICATION: HYPOXIA EFFECT ON CANCER CELLS

We utilize our statistical mechanical framework to analyze the effect of changing oxygen partial pressure on GBM cancer cells. We use this workflow to extract the functional coupling of oxygen to the steady state distribution of copy number for seven measured proteins from the single-cell data. This functional coupling can then be used to make predict copy number distributions under varying conditions. These predictions have several useful applications, including identification of undruggable regimes or potential drug targets in complex protein networks.

## 3.1 Selection of Perturbation Functional Coupling

To determine the functional coupling of oxygen to the steady state probability distribution for the seven measured proteins in our data set, we define 19 candidate Hamiltonians. We are interested in identifying the most simple coupling that can describe the system, so our first six Hamiltonians only have couplings of oxygen to individual proteins, i.e. $g_i(p_{O_2})$ only (Table 2.1 : $M_1 - M_6$). Biologically, this indicates that there are a set of fundamental effective protein interactions, which remain unchanged due to changes in oxygen. These models have direct connections to a previous analysis done by Heath et. al using a quantitative Le Chatelier's model[19], but contain more information about the functional form of the coupling to the individual proteins. The next six models contain couplings to the effective interactions between pairs of proteins ($f_{ij}(p_{O_2})$) only (Table 2.1 : $M_7 - M_12$). The final seven models are a set of Hamiltonians that are combinations of the pairwise and individual couplings (Table 2.1 : $M_{13} - M_{19}$). For all models, the reference state chosen is at 21% $p_{O_2}$, since this is the concentration most typically used in *in-vitro* studies.

The functional forms for these models were chosen for their simplicity or similarity to common functional forms seen in biological systems. We analyze a simple linear coupling indicating a constant proportional relationship between changes in oxygen and protein concentration. Additionally we look at three functions, exponential, logarithmic, and a general power function, indicating that the oxygen concentration affects the rate of production or decay of the proteins[23]. Finally, we analyze two threshold or switching functions, a hill and logistic function, that would indicate protein production or interactions may change at some threshold oxygen concentration. To determine the probability of each of these functional couplings, we analyze the odds ratios (only the odds ratios for the Log-Lin model are shown in Fig. 3.1a), and what percent of the perturbation is captured by each model using our afore mentioned metic (Fig. 3.1b).

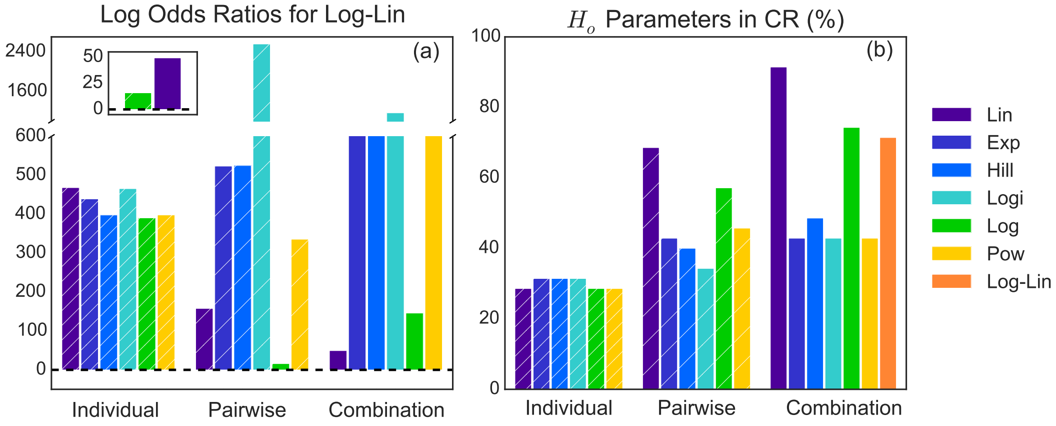

Figure 3.1: (a) Positive odds ratios (log shown) for Log-Lin model compared to all other models indicates is is the most probable model. (b) Percent of analytic $H_{\mathrm{o}}$ parameters, using 21% oxygen as the reference state, found in the predicted credible regions for those parameters for each model, indicating that Lin-Lin contains the greatest amount of the perturbation in the perturbation terms.

Overall, oxygen couplings to individual proteins only, that is with no effect on the protein interactions or network, have similar, low probabilities (Fig.3.1a) and capture the least of the perturbation (Fig.3.1b), irrespective of the functional form of the coupling. This indicates that hypoxic conditions, or oxygen perturbations, affect not only the average copy number, but the fluctuations as well. Therefore, the effective protein interaction network is likely altered under varying oxygen conditions.

In general, two functional forms, logarithmic and linear oxygen coupling are superior to the other switching, exponential or power functional couplings. This is seen both when the coupling occurs in only a pairwise manner or in a combination of pairwise and individual protein couplings. These functional forms, seen in the X-Lin, X-Log, Lin-Lin, Log-Log, and Log-Lin models, are not only most likely overall, but also capture the greatest amount of perturbation compared to all models, all containing over 50% of the perturbation. Interestingly, at low oxygen concentrations, the logarithmic functional form has a roughly linearly effect. Since much of our data is in this low oxygen range, that is four out of the five experimental conditions were $\leq 3\% p_{O_2}$, all five of these models agree that the perturbation effect is linear in low oxygen regimes.

According to the odds ratio, the Log-Linear ($M_{19}$) model is most probable given the fit and complexity of the models tested for this data (Fig.3.1a). The next best two models are X-Log ($M_{11}$) and Lin-Lin ($M_{13}$), but are significantly ($> e^{10}$) less probable. However, our metric describing the percent of the perturbation captured by a model (Fig 3.1b), indicates that the Lin-Lin model captures 91% of the perturbation, while X-Log and Log-Lin only capture 54% and 71% of the perturbation, respectively. This metric suggests that the Lin-Lin model best describes the intrinsic parameter set for the reference state, and may be close to an ideal model, which is in disagreement with the odds ratio.

Since the odds ratio is an unbiased metric, derived directly from Bayes' theorem with no approximations, the Log-Lin model is likely the best model we have tested. However, the discrepancy that arises from the Lin-Lin model capturing over 20% more of the perturbation than any other model, begs the question of what makes the Log-Lin model so much more probable. Therefore, we examine the two components that make up the odds ratio, the goodness of fit of the model to the data and the complexity of the model. The contributions of these two properties can be estimated using the likelihood ratio and the Occam factor for each pair of models, giving insight into the ordering of the odds ratios.

To estimates the posterior distribution we approximate the distribution as gaussian and compare the relative hight and width of the distributions for each model. The likelihood ratio compares the goodness of fit, or the height of the posterior, between two models. The likelihood is the probability at the maximum a posteriori (MAP). Using the most probable parameter set, the likelihood is calculated from our parameter estimation ($P(D|M_i, \gamma_i^{\text{MAP}})$). The Occam factor penalizes more complex models, those with more parameters, as well as models with less flexibility, that is models having less parameter values consistent with a given model. The Occam factor can be estimated by comparing the ratios of the widths of the priors to the widths of the posterior distributions [24],

$$\text{Occam Factor}\left(\frac{M_i}{M_j}\right) = \frac{P(\{\gamma\}|M_j)\prod_{k=1}^{n} CR(\gamma_k)_i}{P(\{\gamma\}|M_i)\prod_{k=1}^{n} CR(\gamma_k)_j} \tag{3.1}$$

For both the likelihood ratios and the Occam factor, the Log-Lin model is used as a reference since it is the most probable model overall. We only examine the top five models in this analysis.

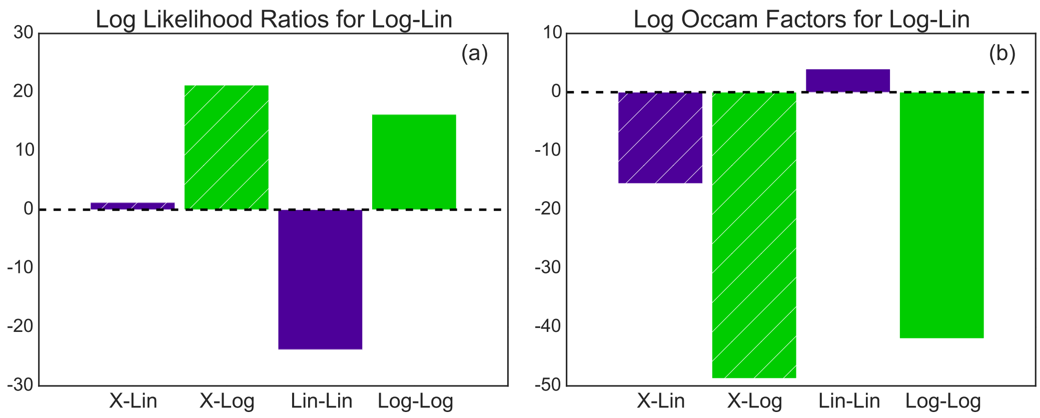

Figure 3.2: Estimated contributions to the odds ratio from (a) log likelihood ratios and (b) the Occam factor for the top five models, all in reference to the Log-Lin model ($P_{Log-Lin}/P_{M_i}$) and calculated using the MAP parameter values.

Investigation into the likelihood ratio indicates that the Lin-Lin model has a much better fit ($> e^{20}$) than the Log-Lin model, and that the X-Lin model has a nearly identical fit to the Log-Lin model. This agrees with the metic determining the percent of perturbation captured, indicating that metric is dependent on the fit of the model. Lin-Lin being the best fit explains that this model likely has less of the perturbation overflow into the intrinsic parameter set for the reference state during the parameter search, and why it has such a high percent of the perturbation captured in the proper terms. Additionally, this analysis implies that the models with linear functional coupling in the pairwise terms best fit this data set, indicating a linear relation between oxygen concentration and effective protein interactions.

However, the Occam factor analysis illustrates that the Lin-Lin model is penalized because it is more complex than the X-Lin and X-Log models, having seven more parameters. Additionally, the Occam factors suggest that the logarithmic functional form provides greater flexibility in the number of possible parameters that have the same probability in the model, increasing the overall probability of the model, as functions with this form have better Occam factors (Fig. 3.2b). It is also clear that the X-Log model is the second most probable model overall, even though it has a worse fit than the other top models, because it has fewer parameters (seven less than

Lin-Lin or Log-Lin) and the increased flexibility due to the logarithmic functional coupling.

Overall, the logarithmic terms seem to add flexibility, but decrease the fit of the model. This may be due to there being some leveling off of copy number or fluctuations at higher oxygen concentrations. Even if the lower oxygen concentrations seem to fit a more linear model, the high oxygen concentration copy numbers could limit the number of parameters in a linear model that fit the data equally well. Therefore, the Lin-Lin model, even though it has a high fit for this data set, would be too inflexible and make poor predictions for further experiments. The greater range of parameter values allowed by the logarithmic terms in the Log-Lin model indicate that this model will likely make better predictions over a greater range of oxygen concentrations.

The relative fit and flexibility of these two models can also be compared visually (Fig. 3.3). Although we see larger credible regions for the Log-Lin model, indicating greater flexibility but also an indication of a worse fit. Overall, the two models are quite similar. Without more data it would be difficult to distinguish if the Lin-Lin model is overfit for this dataset, or if it actually is a more accurate representation of natural functional coupling of oxygen to this system of proteins.

Biologically, the functional forms for both the top models, Lin-Lin and Log-Lin, indicate that the oxygen causes a linear response in the effective protein interaction network. The results also show that the pairwise interactions changing is likely a key aspect of the system's response to a perturbation. Surprisingly, even though the perturbation is linear in each effective protein pair interaction, the fluctuations and covariance change can be much more complex due to the many interacting proteins (Fig. 3.3). There is however still some ambiguity about whether the chemical potentials couple linearly or logarithmically to the oxygen concentration. More experimentation would be useful to distinguish this coupling, particularly at a few more oxygen concentrations, perhaps 10% and 80% $p_{O_2}$ to get a full range of behaviors. Due to this uncertainty, for the rest of the analysis, both the Lin-Lin and Log-Lin models will be utilized to make comparisons and predictions.

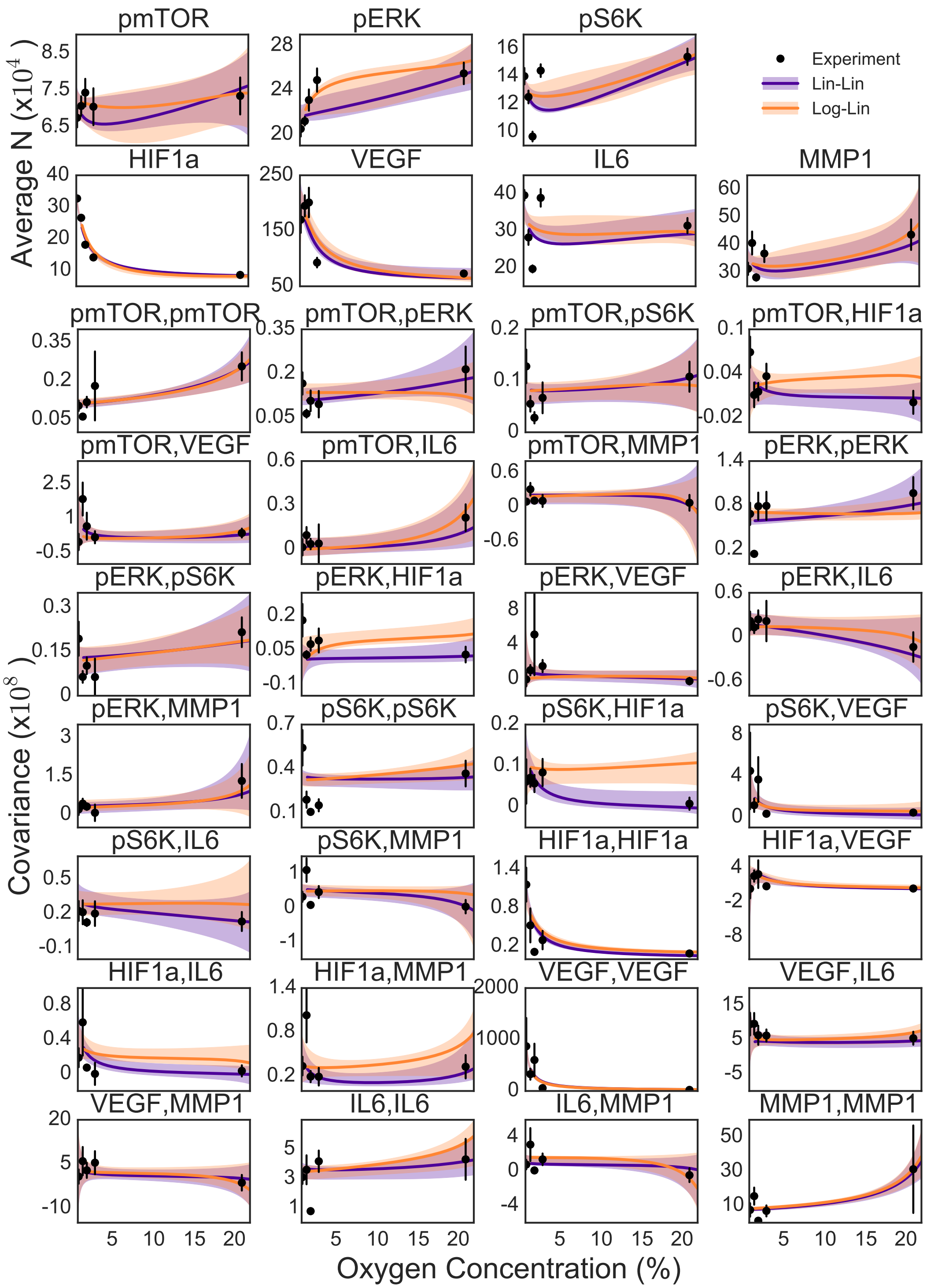

Figure 3.3: Fit of Lin-Lin and Log-Lin models to average and covariance in copy number.

### 3.2 Comparison to Previous Analysis

We compare our framework and models to a previous analysis of this data, which used a quantitative version of Le Chatelier's principle[19]. The Le Chatelier's principle model relates the change in average copy number with a change in the chemical potentials due to a change in the external conditions, but does not specify the form of the change in chemical potentials over oxygen ranges. This model assumes constant fluctuations (or covariance), which is analogous to our models with only individual protein couplings ($M_1 - M_7$). However, even with these simple models we have the ability to identify the specific functional form of the perturbation, unlike the Le Chatelier's principle model, which means we can predict copy number averages and covariances, where as the Le Chatelier's principle can only predict a change.

To analyze what information is gained by the addition of the functional form we compare our most probable model from the first six models, Log-X, to Le Chatelier's principle model. Additionally, to see what is lost by assuming the fluctuations are constant, and how that effects our predictive capabilities, we compare our best fitting models, Log-Lin and Lin-Lin, to the Le Chatelier's principle model. Since the Le Chatelier's principle model can only predict changes in average copy number and predicts no changes in covariance, we analyze the error in these predictions compared to experiment for each model overall (Fig 3.4a-b) and by protein (Fig 3.4c).

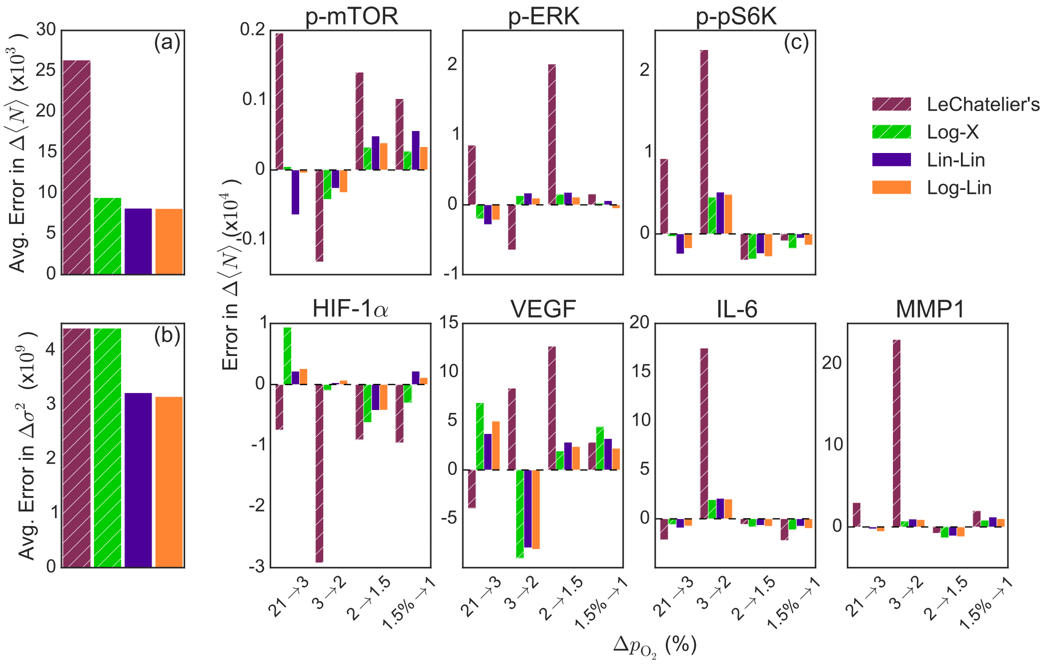

Figure 3.4: Predictions from Log-X, Lin-Lin and Log-Lin models are compared to quantitative Le Chatelier's principle model. The average error per model for predictions of the change in (a) average and (b) covariance copy number due to a change in oxygen concentration compared to experimental values. (c) Error per model for predictions in change in average copy number for each protein.

Overall, our top models are significantly more accurate, at least two times more so, than the previous Le Chatelier's principle model in predicting the change in average copy number (Fig. 3.4a). Log-X, like Le Chatelier's principle, assumes constant effective protein interactions and therefore has constant covariance, so we see the same poor prediction in covariance change (Fig.3.4b). However, the Log-X comparison shows that just knowing the functional form provides increased predictive capabilities for change in copy number. The functional form implies that the growth rates of proteins may be effected by changes in oxygen concentration in a logarithmic fashion. We again see that Lin-Lin and Log-Lin are similar, and both are more accurate at fitting both, average and covariance, than models with no pairwise coupling (Log-X or Le Chatelier's). However, Log-Lin is actually better at predicting changes in covariance than Lin-Lin, perhaps another indication of its being the most probable model, and increasing the evidence that the individual protein growth rates are affected logarithmically by oxygen change.

In looking at individual protein predictions, we do see that Le Chatelier's principle makes poor predictions for some particular ranges in oxygen, particularly from 3% to 2% $p_{O_2}$. Our models however, have more consistent error across all oxygen changes, overall fitting the data more accurately, and increasing our confidence in more expansive predictive capabilities of the model. Additionally, by looking over the individual proteins, some models, including Le Chatelier's, do better, sometimes significantly so, for one protein and very poorly for another. This could indicate that mixed models, where each protein could have a different type of individual or pairwise functional form may provide even more actuate models.

The poor predictions using Le Chatelier's principle model between 3% and 2% $p_{O_2}$ were attributed to the oxygen concentration being a strong perturbation in this regime, which was considered the cause for the disagreement with the theory[19]. Our model does not support this, and therefore see no reason to assume one concentration is a stronger or weaker perturbation, but instead the effect seen is due to the functional coupling of the oxygen concentration to the system. Additionally, using PCA analysis, the previous model indicated that a few proteins became decoupled from the network. Specifically, VEGF seemed to be unpredictable between 2% and 1.5% $p_{O_2}$, perhaps decoupling from the rest of the network. There also seemed to be a loss in mTOR signaling coordination, which lead to p-mTOR becoming unresponsive to some drug treatments[19]. However, they were unable to specifically say how the coordination shift occurs, or the degree to which this occurs. Using the changes of the chemical potentials and effective protein interaction network, we can analyze these two claims in more depth.

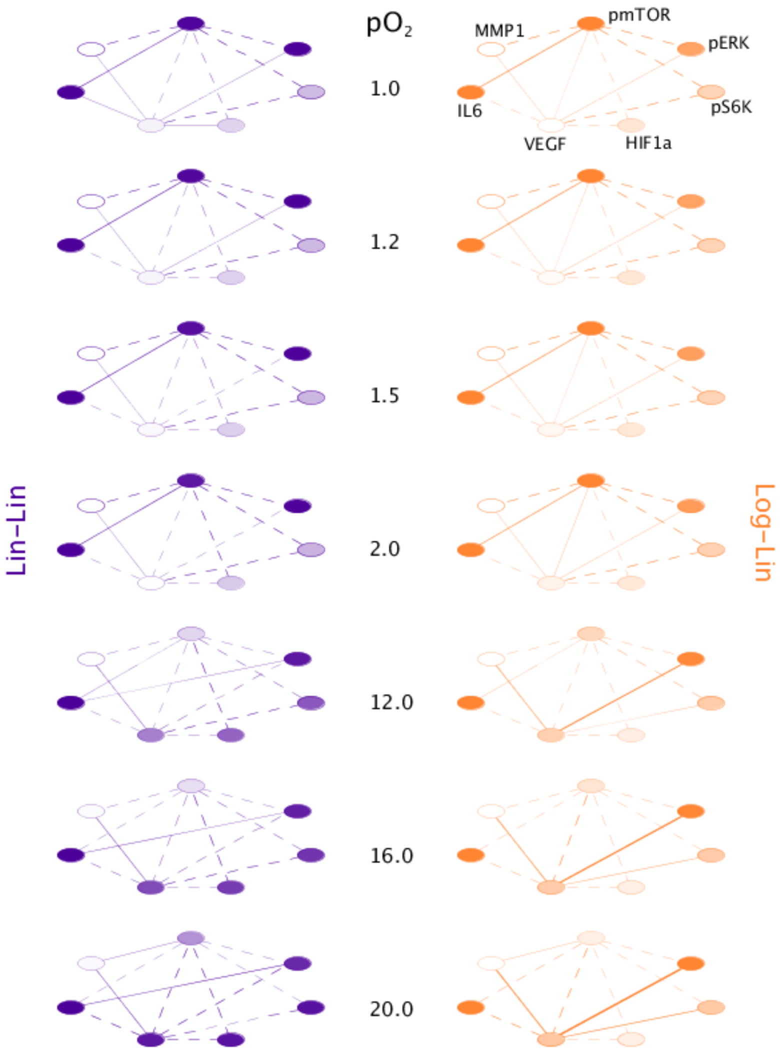

Figure 3.5: Protein networks for p-mTOR and VEGF over a range of oxygen concentrations as predicted by the Log-Lin and Lin-Lin models. Widths of connecting lines and circles represent strength of effective interaction, circle fill represents strength of chemical potential. Dashed and solid lines represent inverse and positive effective pairwise interactions, respectively.

Both of our best fit models indicate that VEGF does indeed become decoupled, and then changes sign of interaction with p-ERK, HIF-1$\alpha$, IL6 and itself, between 1% and 2% $p_{O_2}$ (Fig. 3.5). Additionally, the chemical potential of VEGF also goes to zero, though at 2% $p_{O_2}$ in the Lin-Lin model and 1% $p_{O_2}$ in the Log-Lin model. The chemical potential gives us an indication of the natural flux in copy number, and therefore, in this range there could be large fluctuations in the particle number for VEGF, since there is no natural tendency to either increase or decrease, both are possible, which is in agreement with what the previous analysis saw for this protein.

However, unlike in the Le Chatelier's principle model, p-mTOR does not seem to have a loss in its signaling network between 1.5-2% $p_{\mathrm{O}_2}$. There is a slight weakening of the interactions from p-mTOR to all other proteins as oxygen concentration increases. However the protein network only becomes significantly different for mTOR at high oxygen concentrations (12-20 $p_{\mathrm{O}_2}$), which is where *in-vitro* studies are conducted. In order to test the validity of these predictions in the protein signaling network, single-cell data at more oxygen concentrations between 21% and 3% would be useful.

From this analysis, we see that our framework provides a more specific picture of system's response to a perturbation. We are able to examine not just a change in copy number, but also specific network and individual protein characteristics, providing more insight into the biological response.

### 3.3 Predictions

Besides gaining information on the functional coupling of the perturbation to the system, this framework can also make predictions outside of the measured perturbation conditions or related to testable changes to the protein network, such as the knocking out a protein or an interaction.

Predictions outside of the perturbation regimes measured experimentally could identify interesting and useful perturbation regimes, where there may be different behaviors in average protein copy number or large changes in fluctuations. For the hypoxic system, predictions can be made for extremely low and very high oxygen concentration conditions.

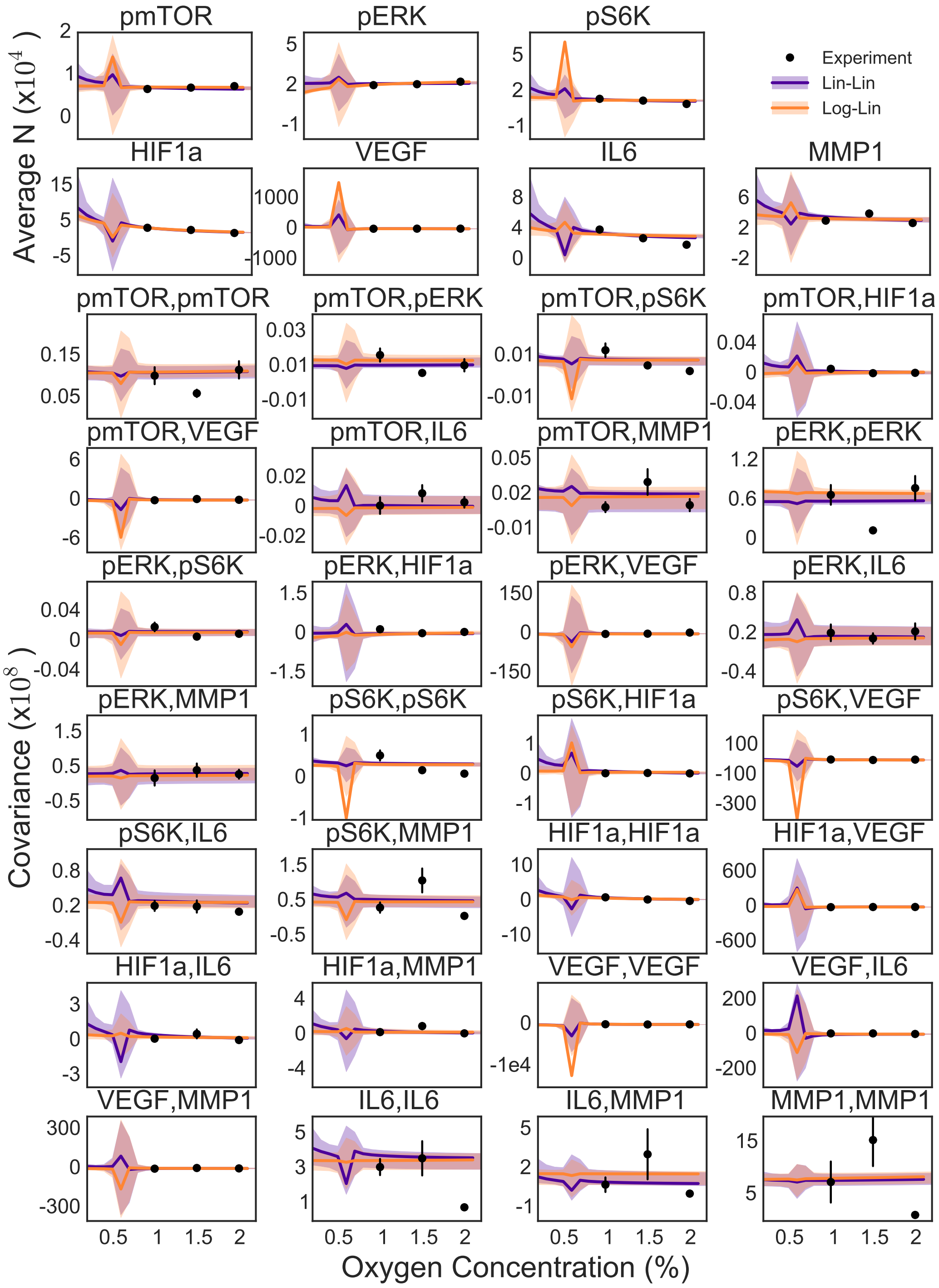

Figure 3.6: Predictions of average and covariance in copy number with the Lin-Lin and Log-Lin models over low oxygen concentrations.

At low oxygen concentrations, at about 0.6 $p_{O_2}$, we note a discontinuity in our predictions for the average copy number and covariance (Fig. 3.6). This can be attributed to a singularity in the pairwise interaction matrix, $\mathbf{A} + \mathbf{F}$, causing the determinate of the pairwise interaction matrix to go to zero. This discontinuity is found in six ($M_7$,$M_{11}$,$M_{12}$,$M_{13}$,$M_{17}$ and $M_{19}$) of the nineteen models. Three of those have linear pairwise coupling to the oxygen concentration, and the discontinuity arises between 0.6 and 0.62 $p_{\mathrm{O}_2}$. Two of them are logarithmic in the pairwise coupling terms and they both have two discontinuities, at about 0.45-0.5 and 0.84-0.87 $p_{\mathrm{O}_2}$.

This behavior could potentially be interesting, showing that the fluctuations in copy number in this regime are large and indicate a kind of phase transition, as has been proposed (although for slightly higher oxygen concentrations, about $1.5 p_{\mathrm{O}_2}$) by Heath et. al [19]. However, two checks would need to be made before this conclusion could be made. First, although we constrain the system to have bounded energetics by checking that the Hamiltonian has a minimum at each experimentally tested oxygen concentration (see full description of the constraint in Appendix B), we do not apply this constraint to all possible oxygen concentrations, in an attempt to use minimal constraints. However, these limited constraints may not be enough to get physically reasonable predictions over all oxygen concentrations. Therefore, before moving on to additional experimentation, the second check on the validity of this discontinuity, it is suggested to first apply this constraint in all oxygen regimes. This could be implemented by checking that the roots of the determinate of the $\mathbf{A} + \mathbf{F}$ matrix fall outside of the possible oxygen concentration regimes (0-100% $p_{\mathrm{O}_2}$).

This theoretical framework also provides a way to investigate the effect of changing or removing interactions between proteins. This could be extremely useful in determining how a drug that alters a single protein or protein interaction affects the complex protein network over a range of perturbation conditions. For example, in looking at the effect of hypoxia on cancer cells, it has been seen that p-mTOR inhibitors fail in some oxygen concentration regimes [25]. To look at how a mixture of drugs may be useful, we can remove one or more effective interactions and see how the system responds across oxygen concentrations (Fig. 3.7).

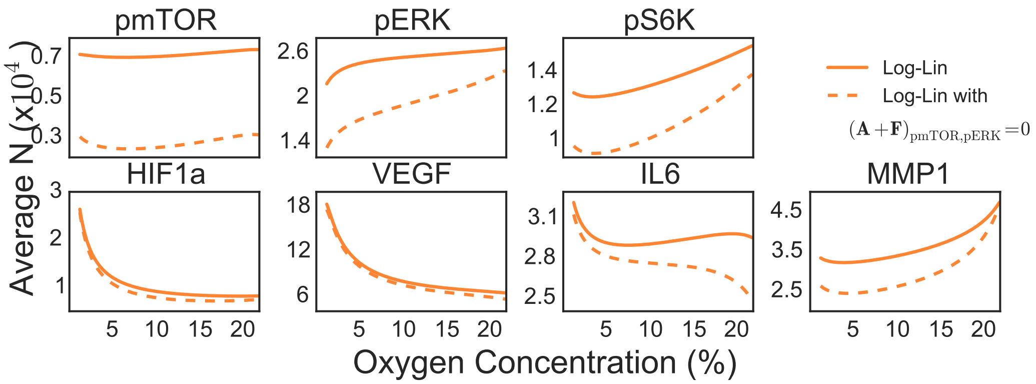

Figure 3.7: Predictions from Log-Lin model showing effects of removing removing the effective interaction between pmTOR and pERK.

To test this hypothesis we removed the effective interaction between p-mTOR and p-ERK while keeping the other interactions the same. For simplicity we only show the effects of this on the average copy number, though this would effect the covariance values as well. Changing even one interaction causes significant changes to the average copy number for some proteins (p-mTOR, p-ERK, p-S6K). This would be a helpful tool in identifying if more or other drugs were needed to remove multiple interactions to cause the effect of interest, or to be sure to keep some interactions intact (say for some important functions in healthy cells).

Our best fit model also makes useful predictions about the effect of knocking out certain genes to render a protein dysfunctional [26, 27]. In many biological systems it can be hard to predict the effect of stopping the production or function of a protein in a cell since most protein interactions are not well known and likely vary with cell conditions. However, our tool provides a way to explore the effect of removing one or more proteins on other proteins copy number (and fluctuations). This can be done by effectively causing a protein to have an average copy number of zero (Fig. 3.8).

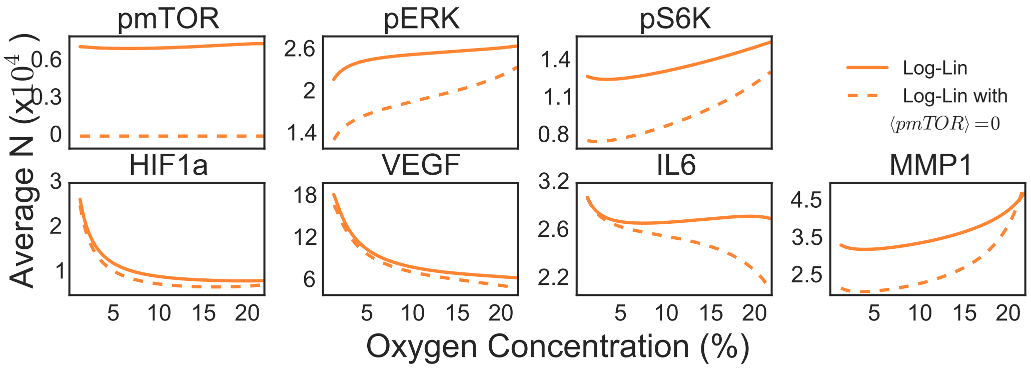

Figure 3.8: Predictions from Log-Lin model showing effects of knocking out pmTOR.

Here we knockout p-mTOR and as expected p-ERK and p-S6K are largely effected over all oxygen concentration ranges since these are known to be downstream effectors of p-mTOR. We also see some proteins, IL6 and MMP1, are affected at only one end of the oxygen concentration range which we would not have been able to predict without this analysis due to the complex interactions of these proteins. This reinforces the necessity of understanding the evolution of the steady-state distribution, since some effects manifest only under certain perturbation conditions, which would be missed if only one particular steady state was examined.

These predictive capabilities could be extremely useful in understanding how to effectively drug a protein network in cancer cells. Cancer cells have been shown to be very plastic, so changes in protein networks are frequent and not always lethal. Similarly, drug sensitivity is dependent on the conditions that the cell is under, as some cell states are more susceptible to a therapy than others. Describing how many cell states respond to the same perturbation could address this therapeutic challenge. Specifically, our model can describe both how protein interaction networks evolve, and which conditions have cell states most susceptible to a drug. Our framework provides a way that would allow multiple knockouts or interactions to be tested, like a combination drug, and the effects could be analyzed over a large range of perturbation conditions.

*Chapter 4*

# CONCLUSIONS

## Summary

We propose a simple theoretical statistical mechanical framework to model the evolution of the probability distribution of cellular components in a cell in response to an external perturbation. We describe a methodology using Bayesian inference, to extract the functional coupling of a perturbation to the system of interest from single-cell experimental data. The flexibility of the framework allows the application of this methodology to any system where simultaneous measurements of single-cell cellular component data can be collected. Any number and type of cellular components can be studied, from proteins, to mRNA, to small molecules, such as metabolites. Even without describing the mechanistic bimolecular interactions, the results of this approach extract many fundamental underlying characteristics of the system. It provides a new way to assess and understand the state of the signaling network and individual proteins under the influence of a physical or molecular perturbation.

The application of our framework to explore the effect of hypoxia in GBM cancer cells found that oxygen couples to the effective interactions between proteins, not just to individual proteins. The most probable coupling to the effective interactions was found to be linear, and the individual couplings were found to be either logarithmic or linear. These results indicate that the state of the protein interaction network changes linearly with the oxygen concentration, but that this can cause more complex behavior in the fluctuations and average copy number in a cell. We show that this methodology provides more accurate predictive capabilities than analyses like the Le Chatelier's principle method, along with a more detailed description of the effect of the perturbation.

## Future steps

We propose several future experiments that could be done to test the validity of our approach and also suggest potential systems to investigate.

Further experiments and theoretical analysis can both strengthen and expand the approach described here. SCBC experiments at other oxygen concentrations, particularly useful would be 0.1, 10 and 80 $p_{\mathrm{O}_2}$, could capture a broader range of cell behaviors by surveying the full spectrum of possible oxygen concentrations. This additional data could clarify whether the Lin-Lin model is overfitting the current data, or if the Log-Lin model is actually the best fit model. Also to investigate the observed discontinuity in fluctuations and average copy number, we could change the way in which we implement our constraints during parameter searches. If the discontinuity is still observed, then experiments should be done in this low oxygen concentration regime to check the validity of this unexpected prediction. Additionally, in the raw data there are several outliers (see Fig. 1.1) and some tests with the removal of these outliers would be a good robustness check.

To determine the need for a more complex model, we propose trying to add a theoretical 'extra' protein for a calculation and remove one of the measured proteins in another. By removing a protein we could investigate if the most probable candidate Hamiltonians order changes leading to a new coupling being predicted. If a subset of proteins are better described by a different functional form, it may not be appropriate to presume all cell components have the same functional coupling to the same perturbation. This could indicate that mixed models are perhaps necessary and should be tested. Adding an 'extra' protein would help to identify if the subset of cellular components experimentally chosen captures enough of the interactions for this network. For example, it could indicate if some protein had a high influence on this system, we might find that this influence would then be captured in the extra protein and would help to direct next experiments to include or find this molecule.

Since this theoretical framework is extremely flexible, it can easily be applied to new systems. In particular, it would be ideal to find a data set for, or experimentally test, a system that is fully contained (or as much as is possible in biology). Analogous to new methodologies created to examine electronic structures, testing is done on a well known, well studied and fully understood model systems. Ideally, finding a simple biological system in a cell, consisting of two or three cellular components and having a well known functional coupling to a perturbation would be ideal. Perhaps finding a simple two state switching system in which the proteins can be "on" or "off" may be able to provide this. Additionally, using a more simple model organism as well would be useful, perhaps *E.coli* where there are fewer factors at play and the system is more fully studied.

Once more standardized checks have been completed for the framework, larger systems should also be tested. SCBC methods can now test upwards of 50 cellular components at a time [2]. It would be interesting to investigate what useful information can be provided from understanding how a larger system responds to a perturbation. Also, a specific system that this might be useful for, is testing the effects of drug conditions on cancer cells, particularly for finding useful drug concentrations or combinations by using the convenient predictive tools this methodology provides, such as looking at protein knockouts or interactions removals.

# BIBLIOGRAPHY

(1) Vogel, C.; Marcotte, E. M. *Nat. Rev. Genet.* **2012**, *13*, 227–232.

(2) Yu, J.; Zhou, J.; Sutherland, A.; Wei, W.; Shin, Y. S.; Xue, M.; Heath, J. R. *Annu. Rev. Anal. Chem.* **2014**, *7*, 275–95.

(3) Elowitz, M. B.; Levine, A. J.; Siggia, E. D. *Science.* **2002**, *297*, 1183–1186.

(4) Maheshri, N.; Shea, E. K. O. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 413–434.

(5) Hart, Y.; Reich-Zeliger, S.; Antebi, Y. E.; Zaretsky, I.; Mayo, A. E.; Alon, U.; Friedman, N. *Cell* **2014**, *158*, 1022–32.

(6) Brock, A.; Chang, H.; Huang, S. *Nat. Rev. Genet.* **2009**, *10*, 336–342.

(7) Mettetal, J. T.; Muzzey, D.; Pedraza, J. M.; Ozbudak, E. M.; Oudenaarden, A. V. *Proc. Natl. Acad. Sci.* **2006**, *103*, 7304–7309.

(8) Hallen, M.; Li, B.; Tanouchi, Y.; Tan, C.; West, M.; You, L. *PLoS Comput. Biol.* **2011**, *7*, 1–16.

(9) Friedman, N.; Cai, L.; Xie, X. S. *Phys. Rev. Lett.* **2006**, *97*, 1–4.

(10) Cai, L.; Friedman, N.; Xie, X. S. *Nat. Lett.* **2006**, *440*, 358–362.

(11) Taniguchi, Y.; Choi, P. J.; Li, G.-w.; Chen, H.; Babu, M.; Hearn, J.; Emili, A.; Xie, X. S. *Science* **2011**, *329*, 533–539.

(12) Ocko, S. A.; Mahadevan, L. *Phys. Rev. Lett.* **2015**, *114*, 1–5.

(13) Marchetti, M. C.; Curie, M.; Curie, M. *Rev. Mod. Phys.* **2013**, *85*, 1143–1189.

(14) Toner, J. *Phys. Rev. Lett.* **2012**, *108*, 1–5.

(15) Bertout, J. A.; Patel, S. A.; Simon, M. C. *Nat. Rev. Cancer* **2008**, *8*, 967–975.

(16) Brown, J. M.; Wilson, W. R. *Nat. Rev. Cancer* **2004**, *4*, 437–447.

(17) Wouters, B. G.; Koritzinsky, M. *Nat. Rev. Cancer* **2008**, *8*, 851–864.

(18) Harris, A. L.; Radcliffe, J. *Nat. Rev. Cancer* **2002**, *2*, 38–47.

(19) Wei, W.; Shi, Q.; Remacle, F.; Qin, L.; Shackelford, D. B.; Shin, Y. S.; Mischel, P. S.; Levine, R. D.; Heath, J. R. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 1352–60.

(20) Foreman-Mackey, D.; Hogg, D. W.; Lang, D.; Goodman, J. *Publ. Astron. Soc. Pacific* **2013**, *125*, 306–312.

(21) Goodman, J.; Weare, J. *Commun. Appl. Math. Comput. Sci.* **2010**, *5*, 65–80.

(22) Goggans, P. M.; Chi, Y. In *AIP Conf. Proc.* 2004; Vol. 707, pp 59–66.

(23) Adler, M.; Mayo, A.; Alon, U. *PLoS Comput. Biol.* **2014**, *10*, 1–14.

(24) Sivia, D.; Skilling, J., *Data Analysis: A Bayesian Tutorial*, 2nd ed.; OUP Oxford: 2006, pp 79–84.

(25) Cloughesy, T. F. et al. *PLos Med.* **2008**, *5*, 139–151.

(26) Schulze, A.; Downward, J. *Nat. Cell Biol. Technol. Rev.* **2001**, *3*, 190–195.

(27) Santiago, Y.; Chan, E.; Liu, P.-q.; Orlando, S.; Zhang, L.; Urnov, F. D.; Holmes, M. C.; Guschin, D.; Waite, A.; Miller, J. C.; Rebar, E. J.; Gregory, P. D.; Klug, A.; Collingwood, T. N. *Proc. Natl. Acad. Sci.* **2008**, *105*, 5809–5814.

*Appendix A*

# DERIVATIONS USING GRAND CANONICAL ENSEMBLE

This appendix shows the derivations for $H = H_o$ for simplicity. However, analogous derivations can be completed for the full Hamiltonian when a perturbation is present, since the perturbation terms combined with the reference state terms can simply be thought of as forming 'new' chemical potentials and effective interactions at each oxygen concentration.

## A.1 Partition Function

### 1-Component System

Assuming cells are at a constant volume and temperature, but particle number (cellular component count) fluctuates, the grand canonical ensemble can be used. The grand canonical partition function is given by

$$\Xi_o = \sum_{N,E} e^{-\beta E + \beta \mu N} \tag{A.1}$$

A given microstate in the system is completely determined by the protein copy number ($N$) and is weighted according to a Boltzmann distribution. The same energetics of a given microstate ($N$) obtained from two different volume and inverse temperature values are given as follows:

$$\frac{\beta^i}{V^i} a^i N^2 + \beta^i \mu^i N = \frac{\beta^j}{V^j} a^j N^2 + \beta^j \mu^j N \tag{A.2}$$

By setting $\beta^i = \xi \beta^j$ or $V^i = \xi V^j$, where $\xi$ is an arbitrary scaling factor, and comparing each term in the above equation shows that the energetics of $i$ and $j$ microstates are equivalent to one another given appropriately rescaled parameters ($a$ and $\mu$). Therefore volume and beta are redundant and their values are arbitrary.

The partition function can be written as as sum over all discrete particle numbers since each state will have a different energy according to the particle number:

$$\Xi_o = \sum_{N} e^{-\beta H(N)} \tag{A.3}$$

The discrete partition function is not analytically solvable, however, we can approximate the partition function as a continuous integral over all possible cellular component densities

$$\Xi_o = V\int_0^{\infty} d\rho e^{-\beta a\rho^2 V+\beta\mu\rho V} \tag{A.4}$$

The exactly solvable Gaussian integral can be evaluated using the following form:

$$I = \int_{-\infty}^{\infty} e^{-\frac{1}{2}Cx^2+Dx}dx \tag{A.5}$$

where we complete the square as

$$-\frac{1}{2}Cx^2 + Dx = -\frac{1}{2}C(x-\frac{D}{C})^2 + \frac{D^2}{2C} \tag{A.6}$$

When $C = 2a$, $D = \mu$ and $x = \rho$, the partition can be solved as

$$\Xi_o = V\int_0^{\infty} d\rho e^{\beta V(-a(\rho-\frac{\mu}{2a})^2+\frac{\mu^2}{4a})} \tag{A.7}$$

$$\Xi_o = Ve^{\frac{\beta V\mu^2}{4a}}\int_0^{\infty} d\rho e^{-\beta Va\rho^2} \tag{A.8}$$

$$\Xi_o = \frac{V}{2}\sqrt{\frac{\pi}{\beta Va}}e^{\frac{\beta V\mu^2}{4a}} \tag{A.9}$$

**n-Component System**

With $n$ cellular components, the partition function can be written by integrating over the density for each component.

$$\Xi_o = V^n\int_0^{\infty} d\rho_1\int_0^{\infty} d\rho_2 \ldots\int_0^{\infty} d\rho_n e^{\beta V(-\sum_{i,i<j} a_{ij}\rho_i\rho_j+\sum_i \mu_i\rho i)} \tag{A.10}$$

As with the 1-component case, this exponential can be rewritten to complete the square with the use of matrix form as

$$\Xi_o = V^n\int_0^{\infty} d\rho_1\int_0^{\infty} d\rho_2 \ldots\int_0^{\infty} d\rho_n e^{-\frac{1}{2}\rho^{\mathbf{T}}\mathbf{A}\rho+\mathbf{M}^{\mathbf{T}}\rho} \tag{A.11}$$

where $\boldsymbol{\rho} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_n \end{pmatrix}$, $\boldsymbol{M^T} = \beta V\begin{pmatrix} \mu_1 & \mu_2 & \ldots & \mu_n \end{pmatrix}$ and $\boldsymbol{A} = 2\beta V\begin{pmatrix} a_{11} & \frac{1}{2}a_{12} & \ldots & \frac{1}{2}a_{1n} \\ \frac{1}{2}a_{21} & a_{22} & \ldots & \frac{1}{2}a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}a_{n1} & \frac{1}{2}a_{n2} & \ldots & a_{nn} \end{pmatrix}$ such that $\boldsymbol{A}$ is a symmetric matrix.

This equation can be simplified by transforming $\boldsymbol{\rho}$ to another vector $\boldsymbol{q}$ with an orthogonal matrix $\boldsymbol{S}$ with a determinant of unity:

$$\rho = \mathbf{Sq} \tag{A.12}$$

The integral is now

$$\Xi_o = V^n \int_0^\infty dq_1 \int_0^\infty dq_2 \ldots \int_0^\infty dq_n e^{-\frac{1}{2}\mathbf{q^T S^{-1} A S q + M^T S q}} \tag{A.13}$$

The matrix $\boldsymbol{S}$ is a diagonalizing matrix which means that

$$\mathbf{S}^{-1}\mathbf{AS} = \begin{pmatrix} d_1 & 0 & \ldots \\ 0 & d_2 & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix} = \mathbf{D} \tag{A.14}$$

which transforms as

$$-\frac{1}{2}\rho^{\mathbf{T}}\mathbf{A}\rho + \mathbf{M}^{\mathbf{T}}\rho \rightarrow -\frac{1}{2}\mathbf{q}^{\mathbf{T}}\mathbf{Dq} + \mathbf{M}^{\mathbf{T}}\mathbf{Sq} \tag{A.15}$$

The variables can now be separated, which expands to

$$-\frac{1}{2}(d_1 q_1^2 + d_2 q_2^2 + \ldots d_n q_n^2) + M_\alpha S_{\alpha 1} q_1 + M_\alpha S_{\alpha 2} q_2 + \ldots M_\alpha S_{\alpha n} q_n \tag{A.16}$$

where the summation of terms will be over the index $\alpha$. The square can be completed for each $\boldsymbol{q_i}$ term.

$$-\frac{1}{2}d_i(q_i - \frac{M_\alpha S_{\alpha i}}{d_i})^2 + \frac{(M_\alpha S_{\alpha i})^2}{2d_i} \tag{A.17}$$

Using the completed square the partition function can be evaluated as

$$\Xi_o = (\frac{V}{2})^n \frac{(2\pi)^{\frac{n}{2}}}{|\mathbf{A}|^{\frac{1}{2}}} e^{\frac{1}{2}\mathbf{M^T A^{-1} M}} \tag{A.18}$$

where $|\boldsymbol{A}\ |$ is the determinant of a $\boldsymbol{A}$ .

## A.2 H$_o$ Ensemble Statistics

The average, $\langle N \rangle$, and fluctuations, $\langle(\delta N)^2\rangle$, in cellular component copy number are calculated from the derivative of the partition function with respect to its conjugate variable $\beta\mu$ according to the following equations:

$$\langle N \rangle = \frac{1}{\beta}\frac{\partial \ln \Xi}{\partial \mu} \tag{A.19}$$

$$\langle(\delta N)^2\rangle = \frac{1}{\beta}\frac{\partial\langle N\rangle}{\partial\mu} \tag{A.20}$$

By choice in our candidate Hamiltonians, we have chosen to only include couplings to individual and/or pairwise protein interactions. Therefore, the third and all higher moments in each model are zero by definition. However, the data contains this information and could be included in a new candidate functional form, in which case, the third derivative with respect to $\beta\mu$ of the partition function would give the third moment.

$$\langle(N-\langle N\rangle)^3\rangle = \langle N^3\rangle - 3\langle N\rangle(\langle N^2\rangle - \langle N\rangle^2) - \langle N\rangle^3 = \frac{1}{\beta^3}\frac{\partial}{\partial\mu_i}\frac{\partial}{\partial\mu_j}\frac{\partial \ln\Xi}{\partial\mu_k} \tag{A.21}$$

**1-Component**

By evaluation of this derivative, the average number of proteins is found to be

$$\langle N\rangle = \frac{1}{\beta}\frac{\partial}{\partial\mu}\left[\frac{\beta V\mu^2}{4a} + \ln\left(\frac{V}{2}\sqrt{\frac{\pi}{\beta V a}}\right)\right] \tag{A.22}$$

$$\langle N\rangle = \frac{V\mu}{2a} \tag{A.23}$$

and the fluctuation in number of proteins is found to be

$$\langle(\delta N)^2\rangle = \frac{V}{2a\beta} \tag{A.24}$$

**2-Components**

To begin generalizing we start with a two component system, with components 1 and 2 the average, variance, and covariance would be

$$\langle N_1\rangle = \frac{V(2\mu_1 a_{22} - a_{12}\mu_B)}{4a_{11}a_{22} - a_{12}^2} \tag{A.25}$$

$$\sigma_2^2 = \langle(\delta N_1)^2\rangle = \frac{2Va_{22}}{\beta(4a_{22}a_{11} - a_{12}^2)} \tag{A.26}$$

$$\sigma_{12} = \frac{-Va_{12}}{\beta(4a_{22}a_{11} - a_{12}^2)} \tag{A.27}$$

**3-Components**

To see full generalizations we show a three component system, with components 1,2, and 3

$$[\frac{1}{2}\mathbf{M^T A^{-1} M}] = \frac{(\beta V)^2}{2|\mathbf{A}|}[\mu_1^2\mathbf{C}_{11}+\mu_2^2\mathbf{C}_{22}+\mu_3^2\mathbf{C}_{33}+2\mu_1\mu_2\mathbf{C}_{12}+2\mu_1\mu_3\mathbf{C}_{13}+2\mu_2\mu_3\mathbf{C}_{23}] \tag{A.28}$$

where **C** is the cofactor matrix of **A**. This leads to an average, variance (one shown), and covariance (one shown) of

$$\langle N_1 \rangle = \frac{\beta V^2[\mu_1 \mathbf{C}_{11} + \mu_2 \mathbf{C}_{12} + \mu_3 \mathbf{C}_{13}]}{|\mathbf{A}|} \tag{A.29}$$

$$\langle (\delta N_1)^2 \rangle = \frac{V^2 \mathbf{C}_{11}}{|\mathbf{A}|} = \frac{\beta^2 V^4 [4a_{22}a_{33} - a_{23}^2]}{|\mathbf{A}|} \tag{A.30}$$

$$\langle (\delta N_1)(\delta N_2) \rangle = \frac{V^2 \mathbf{C}_{12}}{|\mathbf{A}|} = \frac{\beta^2 V^4 [a_{23}a_{13} - 2a_{12}a_{33}]}{|\mathbf{A}|} \tag{A.31}$$

**n-Components**

To generalize the 1, 2 and 3-component systems, we can multiply by the determinate of the **A** matrix. This gives

$$\langle N_i \rangle = \frac{1}{\beta}\frac{\partial \ln \Xi}{\partial \mu_i} = \frac{1}{\beta}\frac{\partial}{\partial \mu_i}[\frac{1}{2}\mathbf{M}^{\mathbf{T}}\mathbf{A}^{-\mathbf{1}}\mathbf{M}] \tag{A.32}$$

$$\sigma_{ij} = \frac{1}{\beta^2}\frac{\partial^2 \ln \Xi}{\partial \mu_i \partial \mu_j} = \frac{1}{\beta^2}\frac{\partial}{\partial \mu_i}\frac{\partial}{\partial \mu_j}[\frac{1}{2}\mathbf{M}^{\mathbf{T}}\mathbf{A}^{-\mathbf{1}}\mathbf{M}] \tag{A.33}$$

which arrive at Eqs. 2.11 and 2.12, respectively.

*Appendix B*

# CALCULATION DETAILS

## Constraints

In order for a given candidate Hamiltonian to have a guaranteed energy minimum at some cellular component copy number we apply the second derivative test in $n$-variables. Therefore, the determinate of the Hessian of the Hamiltonian must be positive definite. The Hessian of the Hamiltonian is just the matrix $\mathbf{A}+\mathbf{F}$. Therefore any parameterization of the Hamiltonian must follow this constraint. We enforce this in our sampling algorithm by checking that for every oxygen concentration experimentally tested that the $\mathbf{A} + \mathbf{F}$ matrix is positive definite.

In order to perform parameter estimation for each model, to encode our prior knowledge we use an uninformative uniform prior for each parameter in the model. The natural parameters for the reference state, $\{a_{ij}\}$ and $\{\mu_i\}$, are very small, much less than 1, therefore for all models to encode our uncertainty we give an extremely wide range for these parameters, using a uniform prior from (-500, 500). Ranges for all other a parameters for each model can be found in Table B.1. Parameter prior ranges for these other parameters are only limited from the (-500,500) range if the units would limit the parameter values (ex. $M_3$ has a parameter with units $p_{O_2}$ and these values can only be from 0 to 100 as it is a percent), or if too wide of range of parameter values would mean essentially effect gives the same result (ex. with exponents once they become large, the effect is essentially the same so we cut it off before 500).

| Models: | | Parameter (min, max) |
|---|---|---|
| Individual | $M_1$ | $k_i(-500,500)$ |
| | $M_2$ | $k_i(-500,500)$ $b_i(-500,500)$ |
| | $M_3$ | $k_i(0,100)$ $b_i(-500,500)$ $m_i(-150,150)$ |
| | $M_4$ | $k_i(0,100)$ $b_i(-500,500)$ $m_i(-500,500)$ |
| | $M_5$ | $b_i(-500,500)$ |
| | $M_6$ | $b_i(-500,500)$ $m_i(-150,150)$ |
| Pairwise | $M_7$ | $k_{ij}(-500,500)$ |
| | $M_8$ | $k_{ij}(-500,500)$ $b_{ij}(-500,500)$ |
| | $M_9$ | $k_{ij}(0,100)$ $b_{ij}(-500,500)$ $m_{ij}(-150,150)$ |
| | $M_{10}$ | $k_{ij}(0,100)$ $b_{ij}(-500,500)$ $m_{ij}(-500,500)$ |
| | $M_{11}$ | $b_{ij}(-500,500)$ |
| | $M_{12}$ | $b_{ij}(-500,500)$ $m_{ij}(-150,150)$ |
| Combination | $M_{13}$ | see $M_1$ and $M_7$ |
| | $M_{14}$ | see $M_2$ and $M_8$ |
| | $M_{15}$ | see $M_3$ and $M_9$ |
| | $M_{16}$ | see $M_4$ and $M_{10}$ |
| | $M_{17}$ | see $M_5$ and $M_{11}$ |
| | $M_{18}$ | see $M_6$ and $M_{12}$ |
| | $M_{19}$ | see $M_5$ and $M_7$ |

Table B.1: Uniform prior ranges used in parameter estimation calculations.

**Convergence**

In order to test convergence of our PTMCMC simulations, we use three criteria, two visual and one analytical. First we analyze the log($Z_1$) error calculated in the EMCEE package that indicates the error in this value associated with only sampling 20 temperatures; practically this number looks at the error associated with having half as many samples. Therefore, if there is a small difference with half as many samples it is an indicator that convergence has been reached. In practice, as long as the error in the log-evidence is a small fraction, <0.1%, of the log-evidence, we accept that criteria as converged.

Additionally we check the trajectories of a random sample of walkers in parameter space for the last 2000 steps, which are the values we use in our calculations of the odds ratio and parameter estimations. We plot a subset of these steps and the average value. An example plot can be seen in Figure B.1. Convergence is indicated If the average value is unchanging and the walkers are sampling a small space around the average.

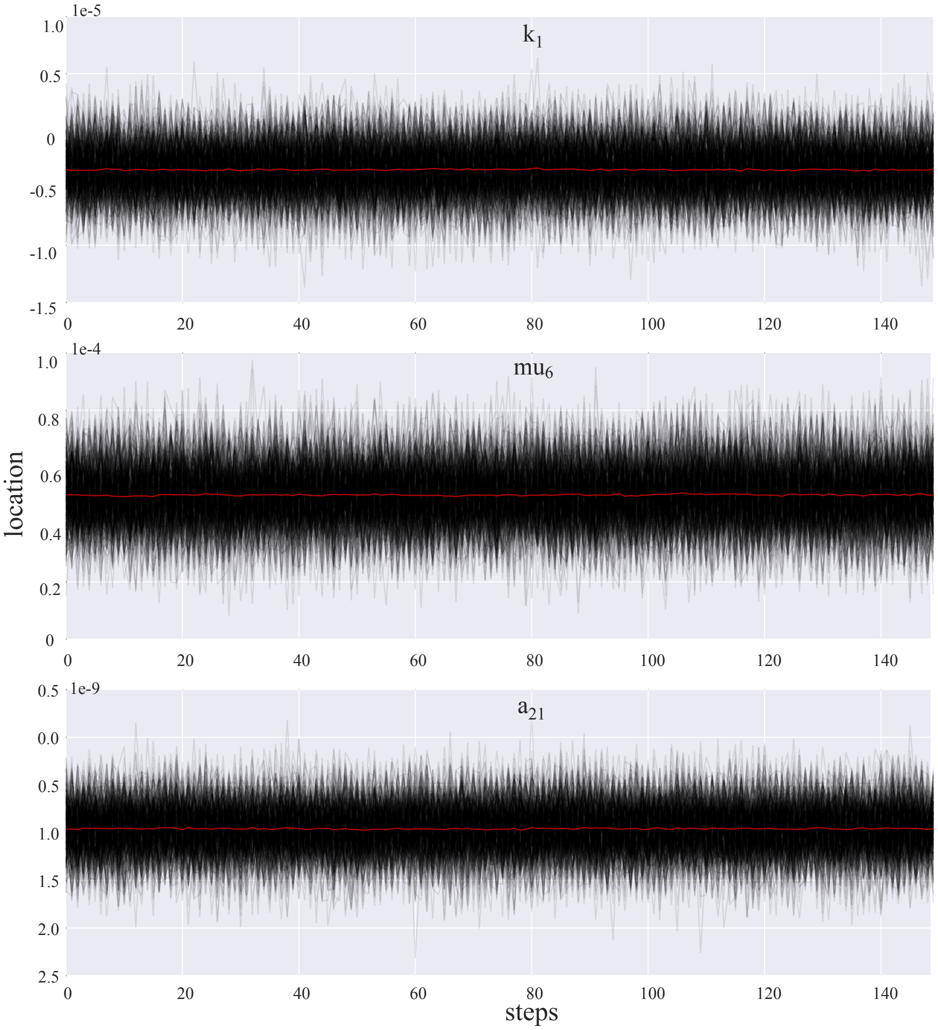

Figure B.1: Walker trajectories for 200 walkers are shown. The average value is shown in red. Three of 42 parameters are shown for the Lin-X model.

Finally, we visualize the posterior by plotting the marginalized posterior for all parameters in the model. Although this does not tell us exactly if convergence is reached, in general the distributions will reach a more cleanly shaped distribution (usually gaussian shaped if not bimodal) showing one or two clear peaks once convergence is reached. This also indicates if we might have a multi-modal posterior which would make our other convergence checks not work as well. So this final check is more useful to gain an overall picture of the posterior.

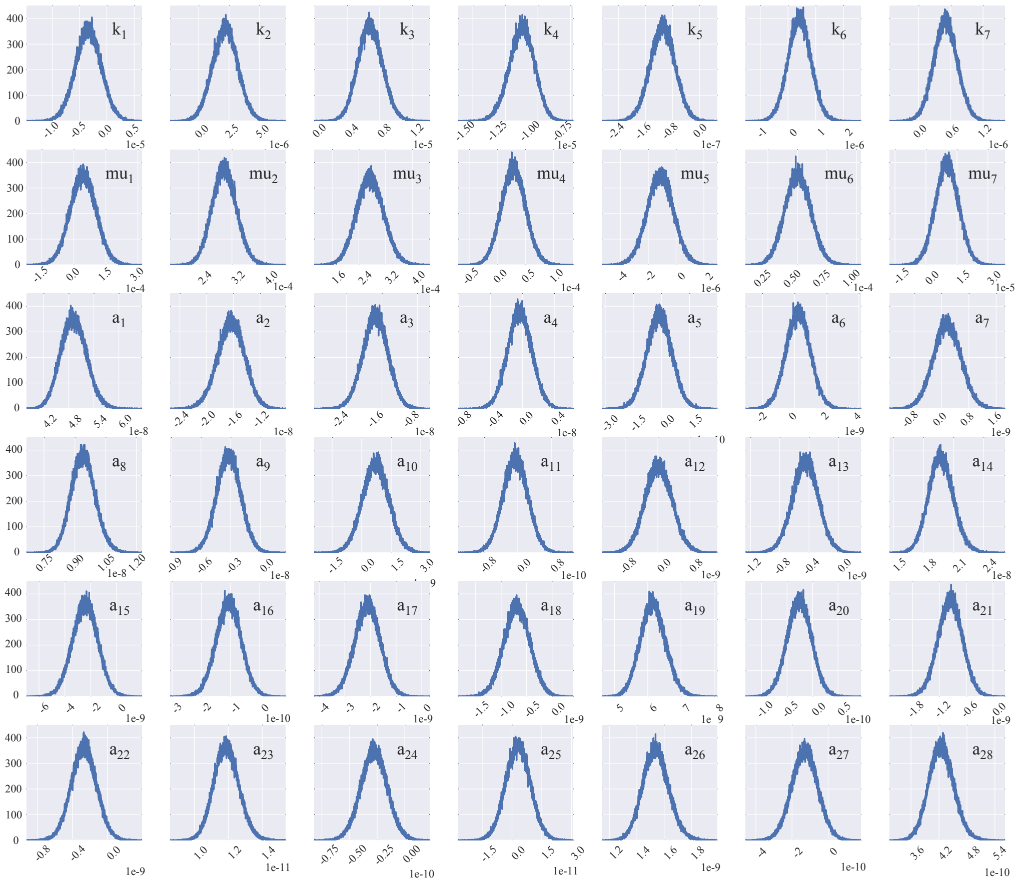

Figure B.2: Marginalized posterior for each parameter in the Lin-X model.