# Statistical mechanical approaches to models with many poorly known parameters

Kevin S. Brown* and James P. Sethna
*Laboratory of Atomic and Solid State Physics (LASSP), Clark Hall, Cornell University, Ithaca, New York 14853-2501, USA*
(Received 10 January 2003; published 12 August 2003)

Models of biochemical regulation in prokaryotes and eukaryotes, typically consisting of a set of first-order nonlinear ordinary differential equations, have become increasingly popular of late. These systems have large numbers of poorly known parameters, simplified dynamics, and uncertain connectivity: three key features of a class of problems we call *sloppy models*, which are shared by many other high-dimensional multiparameter nonlinear models. We use a statistical ensemble method to study the behavior of these models, in order to extract as much useful predictive information as possible from a sloppy model, given the available data used to constrain it. We discuss numerical challenges that emerge in using the ensemble method for a large system. We characterize features of sloppy model parameter fluctuations by various spectral decompositions and find indeed that five parameters can be used to fit an elephant. We also find that model entropy is as important to the problem of *model* choice as model energy is to *parameter* choice.

DOI: 10.1103/PhysRevE.68.021904 PACS number(s): 87.16.Xa, 05.10.Ln

## I. INTRODUCTION

Multiparameter models are ubiquitous in the natural sciences, and much effort of late has gone into making such models of biological regulation in prokaryotes and eukaryotes [1–4], largely because the chemical networks controlling processes such as the cell cycle and terminal differentiation are now seen to be much more complex than previously thought, consequently reducing one's ability to understand them by intuition alone. These models conventionally consist of a large set of nonlinear ordinary differential equations (ODEs) constructed based on the kinetics of the proteins, mRNA, small molecules, etc., participating in the regulatory process. However, few rate constants for the constituent reactions have been measured in cells and one is often ignorant of absolute concentrations of signaling molecules, particularly in mammalian systems. Knowledge of the genomes of model organisms, such as *E. coli*, *Drosophila*, and *C. elegans*, while a major step forward in the large-scale generation of biological data, furnishes almost none of the information necessary to construct and evaluate such dynamical models of protein regulation. There is a famous aphorism in physics: ''Give me four parameters and I can fit an elephant. Give me five and I can wag its tail'' [5]. When one considers that these models may have parameters that number in the tens to hundreds and are only growing in size, generating meaningful and useful models of biological regulation appears even more daunting.

Most previous works using such models have employed a variety of *ad hoc* methods to attempt to deal with this problem. Some investigators have either guessed at appropriate rate values or performed ''fit-by-eye'' to selected protein activities [6]. Others have tried to fit data, using only a subset of the model parameters, which they designate to be important with some kind of sensitivity analysis [7]. In order to understand the behavior of their model when best parameters are changed, some have randomly generated rate constants to try to see what kinds of parameter sets are consistent with either quantitative or qualitative cellular data [2]. Surprisingly, despite potential pitfalls, all of these methods have yielded fruit, and we unify and place them on firm theoretical footing in our method.

Recently, independent work by another group [8] used a Monte Carlo [9,10] approach to obtain an ensemble of rate constants consistent with available time series data. Our work overlaps theirs substantially in the Cost Function and Ensembles sections, where we reference as appropriate. By considering a small transcriptional network, as opposed to our larger receptor-mediated signaling network [11], those authors avoided the numerical challenges we have resolved, and also missed the interesting emergent features of sloppy models, which is our primary focus here. We are particularly interested in some of the topological features of the energy space, which we feel to be generic to the kinds of models we discuss.

We identify three key features of current kinetic models of biological regulation.

(i) *Poorly known parameters.* As discussed above, these models tend to require a large number of poorly determined or completely unknown parameters [12].

(ii) *Simplified dynamics.* Most models are justifiably confined to a small subset of known cellular proteins, even when the process under consideration is in reality more detailed. This is a mild kind of coarse-graining that effectively ''renormalizes'' the parameters (interactions) in order to account for all the effects not explicitly considered in the model. [More severe coarse-graining makes for a much less useful model because the level of description of the model (input-output ''black boxes'' or bulk composition of components) is no longer commensurate with the level of description of current experiments (proteins and interactions).]

(iii) *Uncertain connectivity.* New proteins and interactions among known proteins continue to be discovered, making even the topology of many protein networks somewhat tentative [13].

We call models with many unknown parameters, renormalized interactions, and murky topologies *sloppy*. While on

*Corresponding author. Electronic address: ksb12@cornell.edu

 

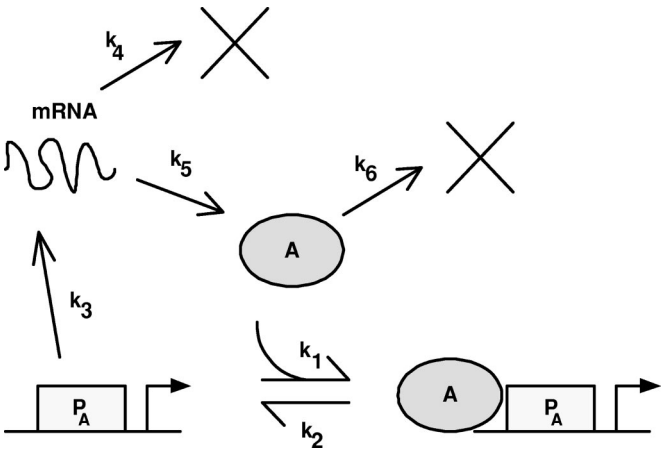


FIG. 1. Simple autoinhibitory gene circuit [18]. Our mathematical form for this model has the following four equations and six parameters: $d[P_A]/dt = k_2[P_A{:}A] - k_1[P{:}A]$, $d[P_A{:}A]/dt = -d[P_A]/dt$, $d[mRNA]/dt = k_3[P_A] - k_4[mRNA]$, and $d[A]/dt = k_5[mRNA] - k_6[A] + k_2[P_A{:}A] - k_1[P_A][A]$.

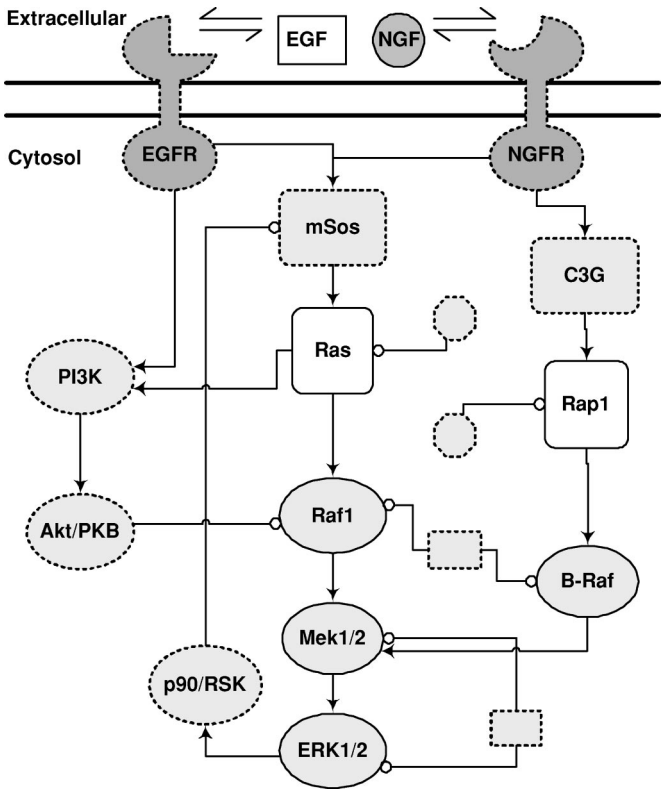


FIG. 2. Network for extracellular-regulated kinase (ERK1/2) activation by epidermal growth factor (EGF) and neuronal growth factor (NGF) in rat pheochromocytoma (PC12) cells [11]. Lines capped with an arrow represent positive stimulation and those capped with a small circle represent negative stimulation. Our mathematical form for this model has 28 first-order, nonlinear ODEs with 48 parameters, which can be found in the supplemental material to Ref. [11].

the surface pursuit of sloppy models seems hopeless, we feel with adequate care useful models of this type can be constructed and meaningful predictions can be extracted from these models. We draw on analogies from statistical mechanics to attempt to make useful biological statements even in the face of sloppiness. While we slant our presentation and applications towards biological regulation on a molecular level, our approach could be useful for modeling in other contexts: we expect that construction of models in other situations like terrestrial nutrient cycling [14] and oceanic biogeochemical cycling [15–17] share some of the features of sloppiness, particularly, in the degree of coarse-graining and the difficulty of obtaining parameter data that modelers desire. In addition, our approaches are easily generalizable, as they do not rely on a particular form for the model.

## II. ILLUSTRATIVE EXAMPLES

In order to make both the problem and our solution to it more concrete, we will consider two test problems to demonstrate our techniques, shown in Figs. 1 and 2. The first is a simple toy model of an autoinhibitory one gene circuit [18] and the second, when coupled with real cellular data, we used to try to understand aspects of differentiation in a neuronal cell line [11]. For the larger model (30 nonlinear differential equations with 48 rate constants), we consider both the real data used in Ref. [11] and two types of fake data. In the first case, which we call the ''mock'' model, data points are generated from the model, which match the real data in all respects (time, protein, fractional error) except, of course, that the model can match the data exactly. In the second case, which we will henceforth refer to as the ''perfect'' model, we go far beyond the quality and quantity of data that can currently (or in the near future) be obtained experimentally. We used the model to generate 90 data points (one each minute) for every active (phosphorylated/guanosine triphosphate-bound) chemical [19] with error bars of size one at every point, corresponding to fractional errors between 1 and $10^{-4}$ percent. In all cases, the sort of biological information we are interested in is exemplified in Fig. 3, in which both mRNA and $A$ (protein) concentration from the model of Fig. 1 are shown as a function of time. The kinds of predictions we typically make are similar to the information contained in Fig. 3 except no experimental data are initially present (see Ref. [11] for details).

## III. COST FUNCTIONS

The ensemble method (as described also in Ref. [8]) fits the model to a set of data with errors $\{(Y_i,\sigma_i)\}_1^{N_R}$, which can include both outputs of the model $y_k(t_i,\vec{\theta})$ (the concentration of chemical $k$ at time $t_i$) and (presumably poorly known) model parameters $\vec{\theta}$, which for models of protein networks contain reaction rate information and potentially initial conditions. For purposes of this study we assume we have no rate data and only time courses of chemical expression (activity) for $\{(Y_i,\sigma_i)\}$. Our starting point is a cost function of the following type, given by

$$C(\vec{\theta}) = \frac{1}{2}\sum_{i=1}^{N_R}\left(\frac{B_k y_k(t_i,\vec{\theta}) - Y_i}{\sigma_i}\right)^2 + f(\vec{y}(t,\vec{\theta})). \qquad (1)$$

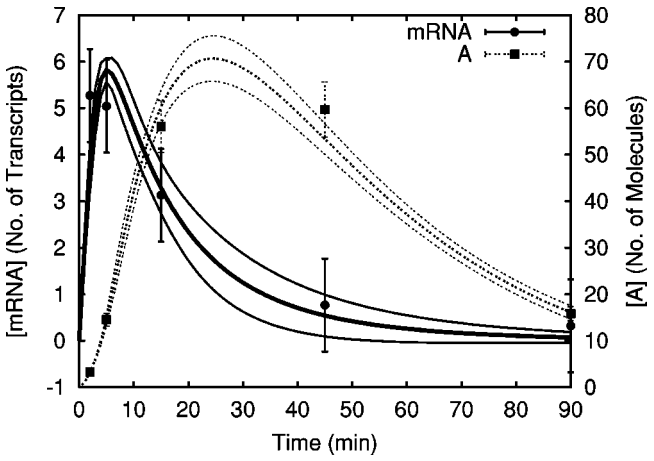


FIG. 3. Time courses for concentrations of mRNA (● and solid lines) and *A* (■ and dotted lines) from model schematized in Fig. 1. Data points are simulated, and the error bars represent *one* ensemble standard deviation, in contrast to the four standard deviations depicted in the plots of Ref. [8].

It is this cost that we associate with the energy of a statistical mechanical system. The nonleast-squares terms $f$ allow us to include somewhat fuzzy data in the model, such as inequalities on the model outputs or other general nonlinear terms. The factors $B_k$ [one for each protein (chemical) measured during a particular experiment] are inserted so we may make use of data determined only up to a multiplicative constant, as is often the case for time series of protein activities measured in cells. The $B_k$'s are fixed by demanding that the simulation outputs are rescaled to optimally match the data. Taking the derivative of Eq. (1) with respect to $B_k$ and setting it equal to 0 yields

$$B_k = \frac{\sum_{i=1}^{N_k} [y_k(t_i) Y_i / \sigma_i^2]}{\sum_{i=1}^{N_k} [y_k(t_i)/\sigma_i]^2}, \tag{2}$$

allowing us to determine $B_k$ algebraically. [In the typical case the nonlinear terms $f$ involve ratios of concentrations of the same chemical at different times, and the $B_k$'s drop out of $f$, leaving Eq. (2) unaltered.] In the models of signal transduction to which we have applied these methods, the rate constants and initial concentrations can be widely separated in magnitude, partially due to the fact that they can have a variety of different units. In order to minimize the effect of these widely separated scales and avoid exploration of unphysical negative values we deal with the logarithms of the parameters for all our calculations, rather than nondimensionalizing large sets of equations by hand. Henceforth, then, we identify $\vec{\theta}$ as a vector of logs of rate constants.

Multiple minima are the norm rather than the exception in high-dimensional nonlinear optimization problems, and we find this to be the case with the models we have considered. We find multiple minima in Eq. (1) (and hence multiple best fit parameters $\vec{\theta}^*$) by combining Monte Carlo (see below) with periodic quenches, using the Levenberg-Marquardt or conjugate gradient methods [20,21]. We then select the minimum with the lowest cost to perform our subsequent analysis. We should point out, however, that in our limited experience the different minima are qualitatively similar, especially once thermal fluctuations are allowed.

## IV. ENSEMBLES

While a set of best fit parameters can form a starting point for any analysis of a sloppy model, with relatively little time-series data and parameter information, making predictions based on a single set of parameters is at best worrisome and at worst meaningless. In order to make useful predictions, particularly, of time courses of chemical concentrations, one wants to know not only the results of the model with the optimal parameters but the model's behavior when *all* parameter combinations consistent with the available data are considered. We thus use the best fit parameters and Hessian as a starting point in generating such an ensemble in the space of parameters $\vec{\theta}$. From a practical perspective, we feel that this ensemble approach makes the model both more useful and more falsifiable. If the fluctuations about the trajectory of a particular dynamical variable are very large, then predictions about that variable's behavior will be unreliable. Conversely, new or additional data that falls far outside the ensemble deviation represents a serious failing of the model that cannot be fixed by parameter fiddling and calls for revision of the model [22].

Consider the probability $P(D|M)$ that our model with parameters $\vec{\theta}$ would generate the observed data $D=\{Y_i\}$. If we assume the $\sigma_i$ represent Gaussian random measurement errors, then

$$P(D|M(\vec{\theta})) \propto \prod_{i=1}^{N_R} \exp\left\{ -\frac{1}{2}\left( \frac{y(t_i,\vec{\theta}) - Y_i}{\sigma_i} \right)^2 \right\}.$$

If we convert the product to a sum and substitute in the definition of the cost in Eq. (1), we find

$$P(D|M(\vec{\theta})) \propto \exp[-C(\vec{\theta})/T],$$

which we can identify with a Boltzmann distribution with energy $C(\vec{\theta})$ and temperature $T=1$ in units where Boltzmann's constant is unity. In Bayesian terms, we then assume a flat prior for the parameters (the same choice made in Ref. [8]) and write, for the probability of the data producing the model, the same Boltzmann distribution as above [23]. Thus, we start at the best fit and generate a thermal ensemble in order to compute an average and standard deviation for any observable $O$ of interest,

$$\langle O \rangle = \frac{1}{N_E} \sum_{j=1}^{N_E} O_j, \tag{3}$$

$$\sigma_O = \sqrt{\langle O^2 \rangle - \langle O \rangle^2}. \tag{4}$$

Typically $O$ is a time-dependent chemical concentration, but of course any quantity one can extract from the model can be treated in such way.

The scale factors $B_k$ in Eq. (1) can also be seen as ''best fit'' parameters. For consistency, we must also consider fluctuations in these scale factors; by integrating out their fluctuations we gain an entropic correction to our cost ''energy.'' Fortunately, their quadratic contribution to the cost can be traced out of the partition function to yield a partition function only of the parameters $\vec{\theta}$, which can then be used in computing averages. Once this trace is performed, one obtains a partial free energy

$$F_p(\vec{\theta}) = C(\vec{\theta},\{B_k^0\}) - T\sum_{k=1}^{N_B} \ln\left(\sqrt{\frac{2\pi T}{a^k(\vec{\theta})}}\right), \tag{5}$$

where $\{B_k^0\}$ are the ground-state rescaling factors, $N_B$ is the number of rescaling factors, and

$$a^k(\vec{\theta}) = \sum_{j=1}^{N_k} \left(\frac{y^k(t_j,\vec{\theta})}{\sigma_j^k}\right)^2. \tag{6}$$

The free energy function in Eq. (5) is what we use for all our thermal techniques.

## V. STIFF AND SOFT DIRECTIONS

We are interested in the shape of the cost manifold as well, and we use the following approach to gain such information. Once we have obtained the best fit parameters $\vec{\theta}^*$, we compute the Hessian matrix

$$H_{ij}(\vec{\theta}^*) = \left.\frac{\partial^2 C}{\partial\theta_i \partial\theta_j}\right|_{\vec{\theta}=\vec{\theta}^*}.$$

For large systems the true second derivative matrix defined above is computationally expensive, so we also consider an approximate second-derivative matrix, which we call the Levenberg-Marquardt (LM) Hessian ($L$) because of its use in that optimization algorithm. The LM Hessian is defined as

$$L_{lm}(\vec{\theta}^*) = \sum_{i=1}^{N_R} \left.\frac{\partial r_i}{\partial\theta_l}\frac{\partial r_i}{\partial\theta_m}\right|_{\vec{\theta}=\vec{\theta}^*},$$

where $r_i = [B_k y_k(t_i,\vec{\theta}) - Y_i]/\sigma_i$ is the $i$th residual. The LM Hessian is only appropriate for least-squares problems, and one expects that the LM Hessian is a good approximation to the true Hessian when the cost at the minimum is small but agrees poorly in so-called ''large residual'' problems, in which the ground-state cost is not near 0 [21]. We wish to evaluate the utility of $L$ because it is substantially less expensive to compute than $H$, requiring only as many function evaluations as are necessary to compute the energy gradient [an $O(N)$ rather than $O(N^2)$ computation]. An eigenvector decomposition of the Hessian allows us to identify stiff (large eigenvalue) and soft (small eigenvalue) directions in parameter space. As we will show, we typically see a few very stiff and many soft directions: since one bare parameter must be varied per stiff direction, only a few need be adjusted to fit our ''elephant.''

One may naturally ask, in the spirit of the preceding section, since the best fit parameters are insufficient in adequately characterizing a sloppy model, is the best fit Hessian ($H$ or $L$) similarly unsuitable? We do indeed feel that the ensemble approach generates a better measure of model softness than $H$, since the ensemble samples the full nonlinear cost space and $H$ is a quadratic approximation to the actual shape of the cost surface, with $L$ serving as an approximation to $H$. One can construct an empirical covariance matrix $\Theta$ from the ensemble of parameters $\{\vec{\theta}_j\}_1^{N_E}$,

$$\Theta = \langle(\vec{\theta} - \langle\vec{\theta}\rangle)(\vec{\theta} - \langle\vec{\theta}\rangle)^T\rangle, \tag{7}$$

where the angle brackets denote ensemble average. An eigenvalue decomposition of this matrix (called principal component analysis (PCA) [24] in statistics) can then be inverted and information about soft and stiff modes obtained in a manner analogous to that using the Hessian, with the understanding that the PCA Hessian $P = \Theta^{-1}$. While PCA does not explicitly model cost nonlinearities—it generates an empirical Gaussian distribution for the given data—these nonlinearities will affect the shape of the resulting PCA quadratic form.

We have thus far identified four sources of soft directions in the cost Hessian.

(i) In a formally underdetermined system, there will be one exactly zero mode for every excess parameter over the number of data points.

(ii) A binding-unbinding reaction close to equilibrium will only need to constrain $K = k_u/k_b$, while the product $k_u k_b$ will be soft. Besides increased numerical stability, this is another reason why we compute $H$, $L$, and $P$ in logspace; these types of soft and stiff modes show up directly in the subsequent eigenvectors.

(iii) Another form of nontrivial soft direction is related to a type of gauge invariance found in spin glasses. Gauge invariances are associated with symmetries in the Hamiltonian or Lagrangian and they often occur when a model has more detail than nature provides. If one changes the sign of a spin in a spin glass and also changes the sign of all the bonds connecting it to its neighbors, the Hamiltonian remains unchanged. In chemical kinetic models where we only know concentrations up to an overall scaling factor, if one rescales the concentration of a chemical $C$ while simultaneously rescaling the rate constants involved in reactions connecting $C$ to others in the system, the cost is unchanged. We, therefore, expect one such gauge invariance for every chemical whose absolute concentration is unknown. These gauge invariances are broken by conservation equations and if one uses a discrete description of chemical concentrations rather than a continuous one.

(iv) Soft modes would arise at a bifurcation in parameter space, since near the bifurcation the energy surface would be very flat in the direction corresponding to the parameter (or parameters) controlling the bifurcation. We have yet to observe such a soft mode. Examples of the more interesting

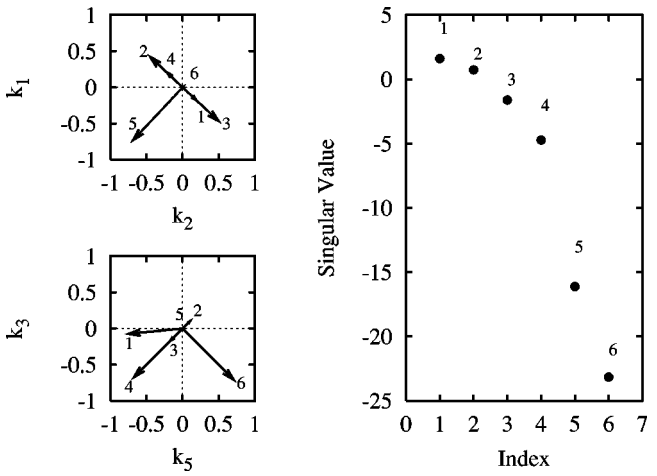


FIG. 4. Mode spectrum and eigenvector projections at a minimum for the model in Fig. 1 showing stiff and soft directions described in the text. Eigenvector-eigenvalue correspondence is indicated by the numbers 1–6. The softest mode 6 arises from a gauge invariance involving only parameters $k_3$ and $k_5$, and hence shows up as a vector with a 45° angle in the $k_5$-$k_3$ plane (lower left). The angle indicates that each rate appears to the first power in the gauge invariant quantity, which one can show algebraically given the equations accompanying Fig. 1. The $k_1$-$k_2$ plane (upper left) reveals binding-unbinding soft modes. Vector 5 is a near-equilibrium soft mode between $P_A$ and $A$, and hence appears in the third quadrant since those two rates must change together, with the same sign, to preserve $k_2/k_1$. Notice all other modes are perpendicular to 5 in this plane, indicating both the stiffness of the product and the expected orthonormality of the vectors. The large eigenvalue spectrum shown here is typical of sloppy models.

soft directions (ii) and (iii) are shown for the small autoinhibitory circuit in Fig. 4, as well as the mode spectrum.

## VI. NUMERICAL CHALLENGES

While the schemes presented above are easy to conceptualize, there are several challenges in actually performing the computations. Many of these difficulties arise because solving large systems of nonlinear ODEs with random right-hand sides can lead to all manner of numerical pathologies. For one, the ill-determined nature of sloppy problems often leads to poorly conditioned Hessians, i.e., those whose eigenvalues are widely separated in magnitude. Diagonalizing a matrix of this type can be numerically unstable, so rather than an eigenvalue decomposition we use singular value decomposition (SVD) [25]. SVD is preferable to eigenvalue decomposition in this situation because of its increased stability.

A second cause of computational woe is in the possible stiffness of the differential equations during the solution process, even if parameters giving a nonstiff set of equations are equally acceptable with respect to data description. This problem arises both in the optimization process and in the thermal sampling. We deal with this difficulty by using a technique that is analogous to a trick used both in computations and simulations of the Ising model. In the zero-field Ising model, one has a discrete $Z_2$ symmetry between the up and down states, which can make calculations of the magnetization difficult. In practice, then, one can add an infinitesimal magnetic field to break the symmetry (thus forcing the system to pick a unique ground state), perform the calculation, and then take the field to 0. Similarly, in sloppy modeling, we modify Eq. (1) to read

$$\tilde{C}(\vec{\theta}) = C(\vec{\theta}) + \frac{\gamma^2}{2} t_{ODE}^2 , \tag{8}$$

where $t_{ODE}$ is a measure of the total time required for the integration of the equations (e.g., the total number of time steps) and $\gamma \ll 1$. This term has the effect of breaking the continuous symmetry of the soft modes, which can otherwise allow wandering into regions of excessive numerical stiffness. We find that this small bias dramatically improves performance while maintaining the integrity of the results, and is almost a necessity for both the initial parameter fitting and subsequent ensemble generation steps. In fact, addition of such a term has given more that a 100-fold computational speedup without any additional computational overhead, such as that necessary for most stiff integration routines [25]. We emphasize as well that for problems we have thus far considered, the computer time cost has no effect on the properties of the solutions obtained, indicating that the stiffness is not necessary in describing the data.

There are additional problems in generating the ensemble. For one, our cost manifolds are highly asymmetric (more like cigars than spheres), and taking uniform steps in such a manifold leads to low acceptance probability. We, therefore, perform importance sampling [26] using the initial best fit Hessian to scale the parameter moves, thereby making large jumps in the soft directions and smaller jumps in the stiff ones. Doing so, however, introduces an additional complication. Extremely soft modes have step scales that can lead to numerical instability (they can be many orders of magnitude in logspace), but by not taking a step in those directions we neglect possibly vital information, since the modes that are soft with respect to available data can and do have dramatic effects on model predictions. We compromise by cutting off the scale of the move at unity, which corresponds to not allowing eigenvector movements larger than a factor of $e$ at any one step. This allows us to explore the shape of the cost basin more fully while still preserving numerical solvency, with the downside of increasing the equilibration time for the thermal Monte Carlo. We, therefore, pick trial moves

$$\Delta\theta_i = \sum_{j=1}^{N_p} \sqrt{\frac{R}{\min(\lambda_j, 1)}} V_{ij} r_j , \tag{9}$$

where $V$ is the matrix of eigenvectors of the ground-state Hessian $H_0$, $r_j$ is a Gaussian random number with zero mean and unit variance, $R \leqslant 1$ is a fixed rescaling tuned to the problem to improve the acceptance ratio, and $\lambda_j$ is an eigenvalue of $H$. We feel importance sampling of this sort is a necessity when generating ensembles for sloppy models, though one might be able to get by equally well using the Levenberg-Marquardt Hessian as well.

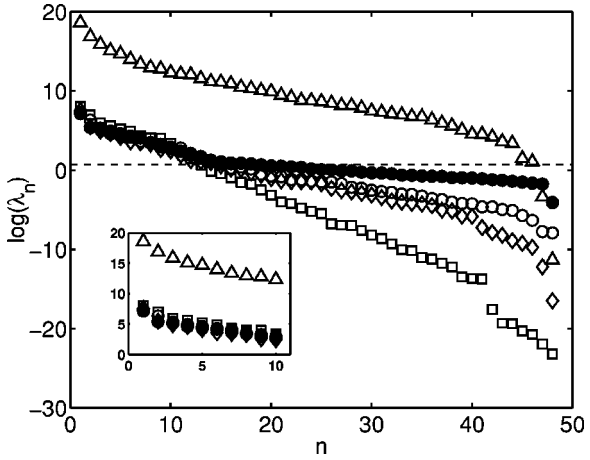


FIG. 5. Natural log of eigenvalues for the model matrices schematized in Fig. 2: real $H$ (○), real $L$ (□), real $P$ (●), mock $H$ (◇), and perfect $H$ (△). The spectrum for the perfect model's Hessian has been shifted downward by 10 to compress the axes; the dramatically different energy scale is due to the energy being extensive both with the size of the error bars and the number of data points. The dotted line is a cost significance criterion described in the text. The real $P$ spectrum was calculated from an inverse covariance matrix representing 704 independent samples.

## VII. SLOPPINESS IN THE PC12 MODEL

In order to gain some understanding of the behavior of the soft and stiff modes for the model of Fig. 2, we chose to compare mode spectra for the real model's $H$ ($H^R$), $L$ ($L^R$), and $P$ ($P^R$), as well as $H$ from the mock model ($H^M$) and $H$ from the perfect model ($H^P$). The singular value spectra of these model matrices are shown in Fig. 5. Changing an eigenvector by a factor of 2 reduces the probability of the model (likelihood that the model would reproduce any of the datasets) by a factor of $e$ if the corresponding eigenvalue lies above the dotted line drawn in the figure. First, concentrate on the stiffest few modes (enlarged in the inset) that are most important in fitting the data. Notice that the stiffest few eigenvalues agree quite well among the various models: even the perfect model's eigenvalues have the same variation up to an overall energy scale change (due to its use of much more data of high precision). How much are the eigen*vectors* shared between these model matrices? Consider the measure

$$w_i^{lm}(k) = \sum_{j=1}^{k} (\vec{v}_i^{\,l} \cdot \vec{v}_j^{\,m})^2, \tag{10}$$

which is an indicator of how much of eigenvector $i$ in model $l$ is present in the $k$ stiffest eigenvectors of model $m$, with $w^{ll}=1$. We use $k=5,6,7$ for our analyses; these are natural choices if we are interested in the tail wagging of the elephant. We find good overlap of the three stiffest modes of model $l$ with model $m$ for all $l \neq m$ pairs: $w_1(5) \geqslant 0.98$, $w_2(6) \geqslant 0.66$, and $w_3(7) \geqslant 0.61$. In fact, if one considers just the stiff-stiff squared dot product, $w_1(1) \geqslant 0.89$, indicating that the high value of $w_1(5)$ is largely due to overlap between just the stiffest mode in each model.

These results show the following. First, the calculated curvature matrix $H$ and the empirical Hessian $P$ formed by PCA are qualitatively similar. Also, $P$'s agreement with $H$ suggests that the stiffest modes are locally quadratic, though nonlinearities would appear further from the minimum. Second, the stiff modes of the true Hessian $H$ are well approximated by the Levenberg-Marquardt Hessian $L$, allowing greater computational efficiency in the sampling procedure and the possibility of a thermal variant of the Levenberg-Marquardt algorithm [27] for minimization. However, $H$ and $P$ both have the advantage of being calculable whether the cost is of least-squares type or not, which is not true for $L$. Third, the stiffest modes are essentially independent of the amount of data, the eigenvalues being related by a change of scale.

Next we consider the bulk of the spectrum, which contains what we feel to be the most unexpected features. First, the eigenvalue spacing is uniform in logspace; there is no clear cutoff between ''stiff'' and ''soft'' but rather a continuum of eigenvalues. The shapes of the spectra are also quite similar; the increased flattening in the soft part of the spectrum of $P^R$ is likely due to partial equilibration of the soft modes. The most striking feature of Fig. 5 is the mode spectrum of $H^P$, which, while shifted vertically for dimensional reasons, displays a degree of sloppiness virtually identical to the much less well-determined real and mock models. Obviously, the perfect model can give a better estimate of the model parameters; only the last eigenvalue is insignificant by our criterion (dashed line shifted downward by 10). More broadly, the fact that the eigenvalues of the perfect model have the same shape shifted upwards means that it is qualitatively similar as a fitting problem: the stiffest five eigenparameters can be used to explain and fit most of the variation in the time series curves, just as in the original models.

Considering that generating the quantity and precision of (admittedly fake) data that went into the perfect model is quite unrealistic for current and near-future biological experiments, we are lead to believe that sloppiness is an inherent feature of this problem, i.e., systemic in the energy topology of large systems of nonlinear equations coupled to data. While we believe the preponderance of sloppy modes to be related to the problem of ''separation of exponentials,'' as discussed in Ref. [28], we also feel that it could have biological implications. For example, a coordinated change along sloppy directions could alter the activity of one regulatory pathway while leaving another unchanged, thus allowing the cell to maintain certain functions while slowly evolving others. Similarly, simultaneous use of multiple mild pharmacological interventions could have a much more subtle and controlled effect than ''all-or-nothing'' inhibition of one or two protein players. The important point is that widely varying microscopic dynamics can give rise to similar protein activation phenomenology, and it is the character of this activation-deactivation cycling that the cell senses and not values of individual rate constants.

## VIII. MODEL SELECTION

We are also able to extend our thermodynamically motivated techniques to the problem of model selection, that is,

choosing one of a competing set of models for a process, given the data they are supposed to explain. In signal transduction models, a useful and persistent question one would like to answer is whether a not the addition or removal of a regulatory loop is warranted, given the available data. Based on both modeling and experiments on the PC12 system, we began to believe that the loop consisting of PI3K and Akt/PKB is essentially unnecessary in setting the dynamics of ERK [11] and we would like to have a quantitative measure of whether this is indeed so.

On first thought, the ground-state energy of the different models must enter into this problem, since the best fitting model is desirable. However, one could introduce another model with additional parameters so chosen to improve the ground-state energy, thus by extension leading one to choose a model with an infinite number of parameters in all situations, which is clear nonsense. One would like not only a measure of a model's goodness of fit but also of its complexity or entropy. Thus, the most general model selection criterion to choose between models $i$ and $j$ must be of the form

$$B_{ij}=f(F_i,F_j), \tag{11}$$

where $F=E-TS$ is the Helmholtz free energy. In Bayesian statistics, in order to choose between two models describing the same data a quantity exactly of this type, called the Bayes factor [29], is computed

$$B_{ij}=\frac{\int e^{-F_p(\vec{\theta}_i)/T}\pi_i(\vec{\theta}_i)d\vec{\theta}_i}{\int e^{-F_p(\vec{\theta}_j)/T}\pi_j(\vec{\theta}_j)d\vec{\theta}_j}, \tag{12}$$

and model $i$ is chosen if $B_{ij}>1$ and model $j$ otherwise, with strength of evidence according to the magnitude of $B_{ij}$ [30]. Except for the factors $\pi$ (about which we will have more to say shortly), Eq. (12) is a ratio of partition functions, and hence related to the difference in free energy between models $i$ and $j$.

The factors $\pi_i,\pi_j$ are called prior distributions and can incorporate biases we may have about likely values for the parameters. One might then think that in order to make a completely unbiased calculation one should just pick $\pi=1$, i.e., the uniform distribution. A difficulty with the Bayes factor arises if one chooses a prior probability distribution $\pi(\vec{\theta})$ with infinite integral such as the unform distribution. Use of such a distribution, called ''improper,'' in the computation of the partition function introduces an arbitrary multiplicative constant and makes interpretation of the Bayes factor difficult [29]. However, use of a proper $\pi(\vec{\theta})$ is not always appropriate, since this presupposes more knowledge about the parameters than one may feel is justified, and the Bayes factor can show sensitivity to the choice of the prior in systems where a relatively small amount of data is available [31]. For these reasons a simplification of Eq. (12) is often employed, which does not depend upon a prior [29]. This generates the so-called Schwartz criterion or Bayes information criterion (BIC) [32], which is an approximation to the logarithm of the Bayes factor:

$$\log B_{ij}\approx\tfrac{1}{2}(C_{j0}-C_{i0})+\tfrac{1}{2}(N_j-N_i)\log N_R,$$

$C_0$ is the ground-state cost, $N_R$ is the number of observations (data points), $N$ is the number of model parameters, and other notation is as above. The BIC is asymptotically (in $N_R$) equal to the true log Bayes factor [33], but for the problems we consider it is questionable whether we have enough data to be sufficient in the asymptotic regime.

We mention here that we have discovered other reasons, besides facilitating model selection, for picking a proper prior distribution on the parameters. We typically find that the cost space of a sloppy model looks somewhat like a golf course; a few relatively shallow holes are separated by large flat regions. Even in the stiff directions the cost can eventually flatten out far from a minimum and these flat regions, while potentially high in cost, are areas of high entropy. Thus, analogous to the thermodynamics of a lone hydrogen atom in empty space, during thermal sampling once a stiff mode escapes to these regions (or a rare thermal event causes the atom to unbind) entropy takes over and it can evaporate to infinity (in infinite time). In this work we confront this problem by (1) controlling the stepsize in the stiff directions during the Monte Carlo (as described above) and (2) running for finite time, of course. However, we recognize that this is still a problem in principle and treat it in detail elsewhere using a simpler model system (K.S.B. and J.P.S., in preparation). In short, adding a very weak Gaussian prior—corresponding to a quadratic term in the cost for the parameters—perturbs the minimum negligibly but prevents the stiff directions from evaporating.

We do not want to show any particular bias against frequentist statistical methods, and non-Bayesian criteria exist, which take into account the caveats we introduced earlier. One such popular criterion is the Akaike information criterion (AIC) [34], in which one selects the model with the smallest value of

$$A_i=2C_{i0}+2N_i. \tag{13}$$

The AIC is based on an information theoretic measure of model quality. Unlike in the Bayesian case, the fundamental quantity from which this criterion is derived [the frequentist analog of Eq. (12)] depends upon the true model and cannot be calculated, thus forcing one to go directly to the AIC or some variant thereof [35]. The BIC and AIC are point-based estimates and the true Bayes factor uses ensemble information, so we are inclined to favor the true Bayes factor when one can calculate it meaningfully. However, it remains for us to describe how to surmount the improper prior problem.

In order to arrive at a meaningful calculation of the Bayes factor without assuming too much *a priori* information about rate constant ranges and values, we propose the following method, which is a form of robust Bayesian inference [30,36]. We assume that the logarithm of each rate constant $\theta_l$ is Gaussian distributed around its best fit value $\theta^*$ with standard deviation of $M_l$, which one may interpret as a number of decades divided by log(10). To further simplify things, we assume $M_l$ is independent of $l$ (or pick $M=\max_l\{M_l\}$) and write for the $M$-corrected log Bayes factor

$$\log B_{ij}(M) = \frac{N_j - N_i}{2}\log(2\pi M^2) + \log \int d\vec{\theta}_i e^{-F_p(\vec{\theta}_i)/T - (\vec{\theta}_i - \vec{\theta}_i^*)^2/2M^2} - \log \int d\vec{\theta}_j e^{-F_p(\vec{\theta}_j)/T - (\vec{\theta}_j - \vec{\theta}_j^*)^2/2M^2}. \quad (14)$$

We now treat $M$ as a free parameter and calculate the log Bayes factor for several values of $M$ by sampling parameter space with this prior distribution, in the manner we described previously. We can then use the results from the different $M$'s to assign meaning to the calculation as follows. For example, suppose model $i$ is only favored when $M<1$. If we think knowing the rate constants to better than a factor of 2 is an unrealistic expectation for model $i$, then we can confidently choose model $j$. Similarly, if $M>50$, we only pick model $j$ if we are uncertain in our parameters by at least 21 orders of magnitude, which will generally be unrealistically large given typical guesses of, and ranges for, biological rates. On the contrary, if $M$ turns out to be a value consistent with known or guessed uncertainties in biological affinities, then we must consider the two models essentially equivalent. We should also point out that sophisticated methods exist for computing free energy differences between two systems [37], and they are particularly useful when the energy distributions for the two models are dramatically different. However, we are fortunate that in our case one model is a subset of the other and we are able to calculate $B_{ij}(M)$ from Eq. (14) directly, by converting the integrals to Riemann sums.

We used the logarithm of the Bayes factor (free energy difference), BIC, and AIC to compare the model in Fig. 2 with and without the left-hand regulatory PI3K loop. For the computation of the BIC and AIC, $N_i=59$ and $N_j=51$, while in the free energy calculation $N_i=48$ and $N_j=40$. This is because in the free energy calculation we are including the entropic contribution due to the $B$ factors [discussed previously, see Eq. (5)], while in the ground-state calculations they remain parameters, which are picked to be optimal. We find that the AIC ($A_i/A_j=0.9476$) yields a very slight preference for the larger model, and the BIC ($\log B_{ij}=6.1339$) would seem to strongly favor the larger model with the regulatory loop, though $N_R$ may be so small as to make the BIC inapplicable in this case. However, the ensemble calculation overwhelmingly favors the smaller model lacking PI3K and Akt/PKB for a variety of $M$'s, even as small as $M=0.25$. We should also point out that the exact Bayes factor calculation and our robustness calculations [11] both agree; some of the most easily varied bare parameters are associated with the PI3K loop.

How should we interpret the results of these statistical tests? In this case, there is little doubt that the PI3K loop exists in the cell—the question being tested is whether it is strongly affecting the regulation of ERK1/2. The results of the free energy calculation (that the PI3K loop is unnecessary) agree with both robustness and our intuition, since looking at ensemble time series plots with error bars show no significant differences between the model with the loop and without. Experiments show the loop doesn't matter at all [11], an even more stringent criterion. The AIC is local to the minimum, ignores model error bars in its predictions, and gives ambiguous results. The BIC is local and asymptotically correct, but we may very well be outside of its domain of validity. For these reasons, we believe the free energy difference criterion is most reliable for the three methods in this case.

## IX. CONCLUSION

We have presented a unified methodology for the construction, evaluation, and use of models with many unknown parameters, ''renormalized'' interactions, and tentative topologies, focused particularly on models of biochemical regulation in cells. Our methods draw heavily from statistical thermodynamics in order to get as much useful information out of a model as given whatever data we have available. We find many soft eigendirections in the parameter space of these models (using a variety of curvature measures of fitting cost), and show that this sloppiness is not a result of having too little dynamical data for comparison. We show that entropies and free energies of the parameter ensemble naturally arise through the statistical comparison of different models as alternative descriptions of a given data set.

## ACKNOWLEDGMENTS

The authors would like to thank C. C. Hill, R. A. Cerione, K. H. Lee, D. Schneider, C. Myers, C. J. Umrigar, M. R. Fewings, and B. Ganem for helpful discussions. We would like to thank NSF DMR-0218475 and NIH T32-GM08267 for financial support and the Cornell Theory Center for computational resources.

[1] K.C. Chen, A. Csikasz-Nagy, B. Gyorffy, J. Val, B. Novak, and J.J. Tyson, Mol. Biol. Cell **11**, 369 (2000).
[2] G. von Dassow, E. Meir, E.M. Munro, and G.M. Odell, Nature (London) **406**, 188 (2000).
[3] A.E. Smith, B.M. Slepchenko, J.C. Schaff, L.M. Loew, and I.G. Macara, Science **295**, 488 (2002).
[4] A. Hoffmann, A. Levchenko, M.L. Scott, and D. Baltimore, Science **298**, 241 (2002).
[5] We have tried to find the appropriate attribution for this quote but have been unsuccessful. Variants of the statement (differing in the number of parameters) have been attributed to C.F. Gauss, Niels Bohr, Lord Kelvin, Enrico Fermi, and Richard Feynman.
[6] B. Novak, A. Csikasz-Nagy, B. Gyorffy, K. Chen, and J.J. Tyson, Biophys. Chem. **72**, 185 (1998).
[7] B. Schoeberl, C. Eichler-Jonsson, E.D. Gilles, and G. Müller, Nat. Biotechnol. **20**, 370 (2002).
[8] D. Battogtokh, D.K. Asch, M.E. Case, J. Arnold, and H.-B. Schüttler, Proc. Natl. Acad. Sci. U.S.A. **99**, 16 904 (2002).
[9] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H.

Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[10] W.K. Hastings, Biometrika **57**, 97 (1970).
[11] K.S. Brown, C.C. Hill, G.A. Calero, K.H. Lee, J.P. Sethna, and R.A. Cerione (unpublished).
[12] J.E. Bailey, Nat. Biotechnol. **19**, 503 (2001).
[13] A.B. Vojtek and C.J. Der, J. Biol. Chem. **273**, 19925 (1998).
[14] K. Schulz, K. Beven, and B. Huwe, Soil Sci. Soc. Am. J. **63**, 1934 (1999).
[15] B.W. Frost and M.J. Kishi, Prog. Oceanogr. **43**, 317 (1999).
[16] S. Doney, Global Biogeochem. Cycles **13**, 705 (1999).
[17] M.J.R. Fasham, H.W. Ducklow, and S.M. McKelvie, J. Mar. Res. **48**, 591 (1990).
[18] C.C. Hill and J.P. Sethna (unpublished).
[19] In our real system we had no data for the PI3K loop and used methods described in this manuscript to determine if the loop should be included given the data. Thus, for the ''perfect'' model we generated no data for this loop in order to use the model as a test case for the model selection methods, as we used it as a test case for other methods. However, the perfect model is unenlightening in this respect because model selection methods in a sense tell us what we already know, i.e., that the loop was present when the data were generated and are hence necessary.
[20] D.W. Marquardt, J. Soc. Ind. Appl. Math. **11**, 431 (1963).
[21] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. (Wiley, Chichester, 1987).
[22] Our methods identify typical, but not all, members of the ensemble. New data outside the error bars may, in principle, be consistent with the existing model but restrict the parameters to a formerly insignificant subregion.
[23] The prior distribution is flat in the $B_k$'s and in the parameters (*logs* of the rate constants). Using such a deceptively simple prior (equivalent, at least for the rate constants, to a Jeffrys prior) can have more subtle and serious consequences than one might think. For more details, see the model selection section and references therein. The $B_k$'s are quadratic in the cost and can be integrated out of the partition function (see text); they play the role of uninteresting ''nuisance parameters'' in the Bayesian language.
[24] I.T. Jolliffe, *Principal Component Analysis* (Springer-Verlag, New York, 1986).
[25] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C*, 2nd ed. (Cambridge University Press, New York, 1996).
[26] M.E.J. Newman and G.T. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford University Press, New York, 1999).
[27] M.P. Nightingale and C.J. Umrigar (unpublished).
[28] C. Lanczos, *Applied Analysis* (Prentice Hall, Englewood Cliffs, NJ, 1956).
[29] L. Wasserman, J. Math. Psychol. **44**, 92 (2000).
[30] S.J. Press, *Bayesian Statistics: Principles, Models, and Applications* (Wiley, New York, 1989).
[31] R.E. Kass and A.E. Raftery, J. Am. Stat. Assoc. **90**, 773 (1995).
[32] G. Schwartz, Ann. Stat. **6**, 461 (1978).
[33] R.E. Kass and L. Wasserman, J. Am. Stat. Assoc. **90**, 928 (1995).
[34] H. Akaike, in *2nd International Symposium on Information Theory*, edited by B.N. Petrov and F. Csaki (Akademia Kiado, Budapest, 1973), pp. 267–281.
[35] H. Bozdogan, J. Math. Psychol. **44**, 62 (2000).
[36] R. McCulloch and P.E. Rossi, J. Econometr. **49**, 141 (1991).
[37] C.H. Bennett, J. Comput. Phys. **22**, 245 (1976).