Combinatorial complexity and dynamical restriction of network flows in signal transduction

J.R. Faeder, M.L. Blinov, B. Goldstein and W.S. Hlavacek

Abstract: The activities and interactions of proteins that govern the cellular response to a signal generate a multitude of protein phosphorylation states and heterogeneous protein complexes. Here, using a computational model that accounts for 307 molecular species implied by specified interactions of four proteins involved in signalling by the immunoreceptor FcεRI, we determine the relative importance of molecular species that can be generated during signalling, chemical transitions among these species, and reaction paths that lead to activation of the protein tyrosine kinase (PTK) Syk. By all of these measures and over two- and ten-fold ranges of model parameters — rate constants and initial concentrations — only a small portion of the biochemical network is active. The spectrum of active complexes, however, can be shifted dramatically, even by a change in the concentration of a single protein, which suggests that the network can produce qualitatively different responses under different cellular conditions and in response to different inputs. Reduced models that reproduce predictions of the full model for a particular set of parameters lose their predictive capacity when parameters are varied over two-fold ranges.

1 Introduction

Cell signalling, the biochemical process through which cells sense and respond to their environment, involves an array of proteins which include receptors, kinases, and adaptors, components of proteins such as sites of phosphorylation, and other biomolecules [1]. Early signalling events triggered by receptors in eukaryotic cells usually involve the formation of heterogeneous protein complexes in the vicinity of the cell membrane [2–4]. This process of complex formation is complicated because a typical signalling protein contains multiple sites that may be modified (e.g. phosphorylated) and that have the potential to bind other proteins or lipids. In addition, the modification or binding state of a protein can regulate its binding and enzymatic activities. Thus, signalling can generate a combinatorially large number of protein states and complexes with different potentials to generate further signals [4–8]. For example, a protein that contains 10 amino acid residues subject to the activities of kinases and phosphatases theoretically has $2^{10} = 1024$ states of phosphorylation. If the protein forms homodimers, the number of distinct complexes, or chemical species, is 524 800, a number that might exceed the total amount of this protein in the cell. For an assembly of n proteins, the number of chemical species is on the order of $\prod_{i=1}^{n} s_i$, where s_i is the number of possible states of protein i in the assembly. Thus, the number of chemical species in a system depends exponentially on the number of interactions in the system and may be quite large even when few interactions are involved. For example, a model of early signalling events mediated by the immune recognition receptor FcεRI includes 354 distinct chemical species and 3680 unidirectional reactions, but these species and reactions arise from consideration of the interactions among only a ligand and three signalling proteins—the multimeric receptor, FcεRI, and two protein tyrosine kinases (PTKs), Lyn and Syk [9]. Similar models of early events in signalling through the epidermal growth factor receptor (EGFR) also involving only a handful of proteins contain hundreds to thousands of distinct chemical species [8, 10, 11]. This combinatorial complexity has been largely ignored by both experimentalists and modellers and is a major barrier to predictive understanding of signal transduction.

Experimental resolution of protein states and complexes is usually limited to a small number of sites and interactions, but rapidly advancing proteomic technologies are likely to provide a wealth of more detailed information about signalling complexes in the near future [12–16]. A number of studies already confirm that a diverse range of molecular complexes arise during signal transduction [17–20]. Because the full spectrum of protein states and complexes is difficult to enumerate, let alone understand, computational modelling will play an important role in interpreting such data and assessing the functional significance of specific interactions and complexes [8]. Key questions to be addressed include whether networks favour the formation of specific complexes from the multitude of potential complexes, and, if so, how these favoured complexes affect signalling outcomes.

Few biochemical network models of signalling developed so far encompass the breadth of states and complexes required to address these questions. Instead, most models, given a particular set of proteins and interactions, make additional (usually implicit) assumptions that exclude the vast majority of possible species from consideration. An example is the model of EGFR signalling that was

IEE online no. 20045031

doi: 10.1049/sb:20045031

Paper first received 1st November 2004 and in final revised form 19th January 2005

The authors are with the Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos National Laboratory, Mail Stop K710, Los Alamos, New Mexico 87545, USA

E-mail: faeder@lanl.gov

developed by Kholodenko *et al.* [21] and extended by several other groups (for example [22, 23]). The original model includes six proteins and tracks 25 species, but lifting implicit assumptions in the model raises the number to hundreds or thousands of species, depending on mechanistic assumptions, even without the introduction of new rate constants or other parameters [8]. While such models have provided valuable insights into signalling mechanisms, they are not suitable for addressing the questions of whether and how signalling networks favour specific complexes, which requires models that consider the full spectrum of possible species.

Here, we analyse the specificity of complex formation in a network model for early events in signalling by the high-affinity receptor for IgE antibody (FcεRI), a key initiator of allergic reactions [24]. The model has been shown to make accurate predictions of a number of experimental observations [9, 25]. Here, we characterise the distribution of network activity in terms of individual species, reactions, and reaction sequences or paths. We then examine how the spread of network activity is affected when model parameters are randomly varied, which corresponds to changing the initial state of the cell that is receiving the signal. We also explore the possibility of developing an accurate reduced model by removing non-essential species from the reaction network. The results indicate that while only a small fraction of complexes, reactions, and paths is active for a particular cellular state, which elements are active depends strongly on the initial state of the cell. Thus, to capture the full range of signalling behaviours, a model must account for many more molecular complexes than just those that are favoured in any particular cellular state.

2 Methods

Network model. The network model analysed in this study was developed in earlier work [9] and is summarised here. The model includes just four components (Fig. 1*a*): the FcεRI receptor; a bivalent ligand that binds to a single site on FcεRI; the protein tyrosine kinases Lyn and Syk. But, in a vivid illustration of combinatorial complexity, it encompasses 307 species coupled through a biochemical network of 2326 unidirectional reactions (These numbers are smaller than the figures of 354 species and 3680 reactions given in [9] because some species and reactions of the full model are inaccessible when ligand binding is irreversible, which is the case for the IgE dimer). As shown in Fig. 1*a*, the receptor is modelled as three distinct subunits, the primarily extracellular α subunit that binds to the ligand, and the primarily cytoplasmic β and γ_2 subunits that contain immunoreceptor tyrosine-based activation motifs (ITAMs), which upon phosphorylation bind to the SH2 domains of Lyn and Syk, respectively. Lyn also associates with the unphosphorylated β subunit through an interaction involving its N-terminal unique domain [26]. A series of events (Fig. 1*b*) couples binding of the ligand, a covalently cross-linked dimer of IgE antibodies [27], to activation of Syk [28, 29], which is required for downstream signalling events and cellular responses, such as calcium mobilisation and release of histamine from mast cells [30, 31]. Ligand-receptor binding induces dimerisation of receptors, which permits Lyn that is weakly associated with a receptor to phosphorylate the ITAMs of the *trans* receptor in the dimer, leading to the recruitment of additional Lyn and Syk. Syk in dimers can be transphosphorylated on its linker region tyrosines by Lyn or on its kinase activation loop tyrosines by Syk. Phosphorylation of Syk's activation loop tyrosines is critical for all downstream signalling, while phosphorylation of Syk's linker region tyrosines has both positive and negative effects on Syk activity and downstream events.

The simplicity of this picture hides the complexity of the underlying biochemical network. Figure 1*c* displays one of a multitude of possible sequences of individual reaction steps starting from an unmodified receptor and leading to a dimer of receptors containing fully-phosphorylated Syk. At each step along this path many alternative branches are possible, as indicated by the highlighted state in Fig. 1*c* and quantified by the distribution in the number of reactions a species containing a dimer of receptors can undergo (Fig. 1*d*).

Although the simple description of early signalling events shown in Fig. 1*b* hides the underlying size of the chemical reaction network, the network itself is in fact simpler than its size would indicate. The combinatorial explosion of species and reactions described in the Introduction arises because chemical transformations occurring at a particular site on a protein are generally assumed to be independent of the modification state of other sites within the same protein or protein complex. For example, four states of the β subunit of FcεRI are possible (unphosphorylated and unbound, unphosphorylated and bound to Lyn, phosphorylated and unbound, phosphorylated and bound to Lyn) and six states are possible for the γ_2 subunit. There are thus 24 possible modifications states for the cytosolic portion of a receptor, and $24 \times (24 + 1)/2 = 300$ modification states for a dimer when all sites can be modified independently, as in our model. While assumptions of site independence produce large networks, they also permit a relatively small number of parameters to characterise the rates of the reactions that can occur. For example, the model assumes that the rate at which Lyn binds to an unphosphorylated and unbound β subunit of FcεRI (called Constitutive Lyn binding in Table 1) is independent of the binding state of the γ_2 subunit of that receptor or, whether the receptor is contained within a receptor dimer. As a result, there are 144 different reactions involving constitutive Lyn binding (Table 1), but all utilise the same rate constant. Thus, although the total number of reactions in the model is large for combinatorial reasons, the number of reaction types (or classes) is relatively small, and the number of parameters in the model is comparable to the number of protein sites, not the number of chemical species or reactions.

Ultimately, assumptions of minimal interactions among sites must be tested by experiments, but given that there is scant information about how the different components and interactions within a complex or protein affect the further transformations, they provide a basis for developing reasonable initial models. We have recently developed modelling software called BioNetGen that permits a user to create large network models by writing a relatively small number of reaction rules that generate the chemical species and reactions [32, 33].

The reaction classes that are included in the current model are listed in Table 1. The reaction rules used to generate the network along with the default values of the component concentrations and rate parameters that characterise the rat basophilic leukaemia cell line RBL-2H3 are given in Fig. 1 and Table 1 of [9]. The BioNetGen software package was used to construct the model based on these rules and parameters, and to perform calculations [32]. The model and the software are available at http://cellsignaling.lanl.gov

The model is parameterised by the initial concentrations of the four components and 17 chemical rate constants (This number is smaller than the 21 rate constants given in [9] because the two ligand dissociation reactions have zero rate and Syk association and dissociation are taken to be

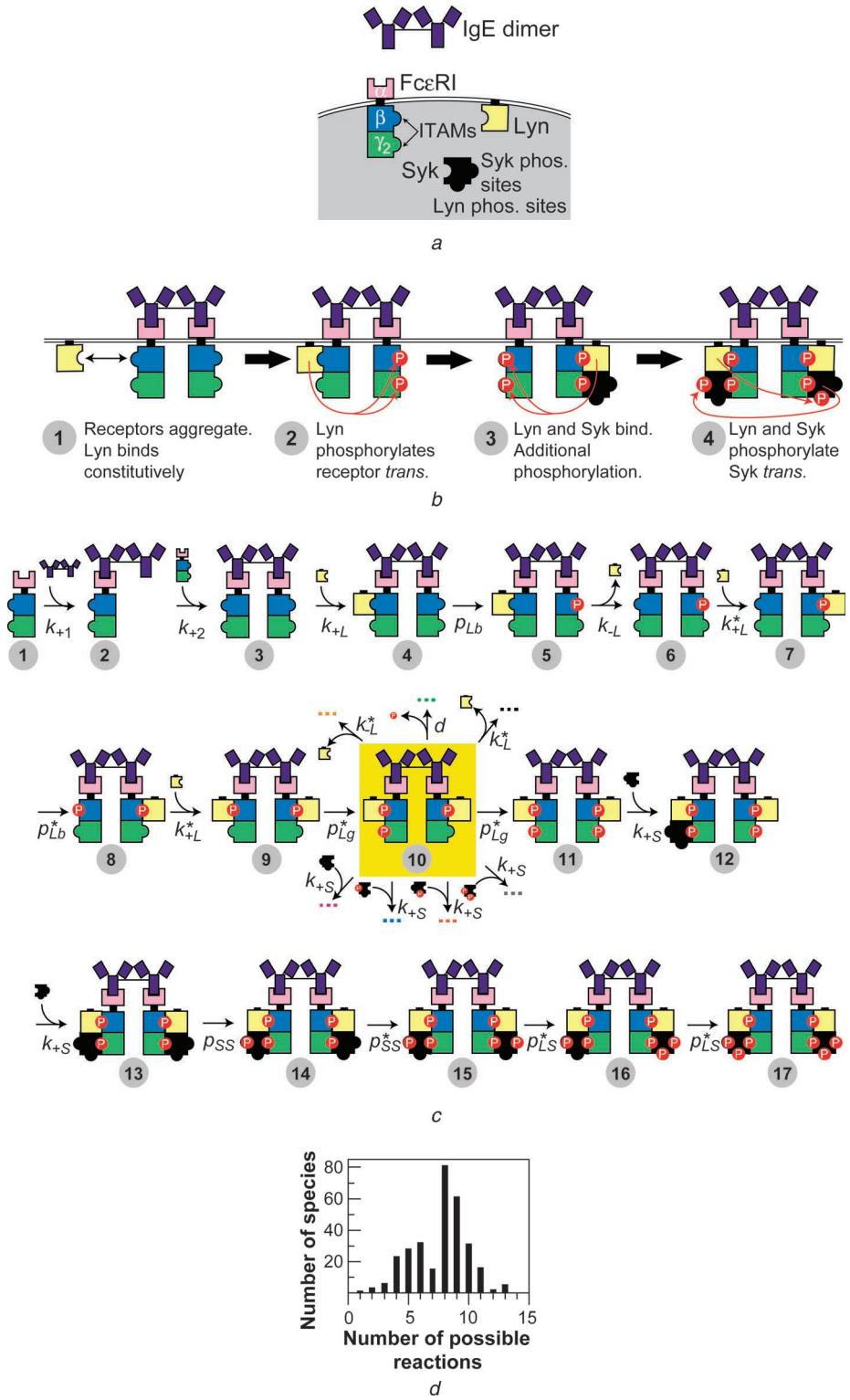

Fig. 1 *Model for early events in signal transduction through the FcεRI receptor. (a) The four basic components of the model—a bivalent ligand, the FcεRI receptor, and the kinases Lyn and Syk. Covalently cross-linked IgE dimers are bivalent ligands that bind and aggregate receptors irreversibly on the timescale considered in the model. The receptor is composed of three distinct subunits, the extracellular α subunit that binds the Fc portion of IgE with 1:1 stoichiometry, and the cytoplasmic β and γ_2 subunits that contain immunoreceptor tyrosine-based activation motifs (ITAMs), which upon phosphorylation bind to the SH2 domains of Lyn and Syk, respectively. Lyn also associates with the unphosphorylated β subunit through an interaction involving its N-terminal unique domain. (b) Coarse description of the events leading to Syk activation in the model. (c) A sequence of reactions in the model that generate the receptor dimer complex with the highest stoichiometry of binding partners and phosphorylation. This path is one of the multitude of paths that exist in the model because of the large number of branches that exist at each step. (d) The distribution of the number of possible reactions that species containing a dimer of receptors can undergo. There are 300 such species in the model*

independent of Syk's phosphorylation state). A detailed description of how the parameters are determined is provided in [9]. Most of the parameters have been determined either directly, through measurement of cellular protein levels or affinities for protein-protein interactions, or indirectly, by fitting a subset of the parameters to experimental time courses of protein phosphorylation and dephosphorylation. Requiring the model to match certain qualitative observations allowed constraints to be placed on the remaining parameters, such as the rates of

Table 1: Distribution of reaction rates for the RBL-2H3 parameter set 100 s after stimulation with 10 nM IgE dimer

Reaction class[a]	Rate constant[a]	Number of reactions	Number of important reactions	Relative rate (% of total)	
				All reactions in class	Top reaction in class
Ligand binding	k_{+1}	2	1	0.03	0.03
Receptor aggregation	k_{+2}	4	1	0.03	0.03
Constitutive Lyn binding					
Association	k_{+L}	146	3	8.58	6.55
Dissociation	k_{-L}	146	6	8.59	6.55
Lyn recruitment					
Association	k^*_{+L}	144	19	0.11	0.03
Dissociation	k^*_{-L}	144	26	0.10	0.04
Syk recruitment					
Association	k_{+S}	384	20	0.27	0.06
Dissociation	k_{-S}	384	35	0.26	0.09
Phosphorylation					
Lyn → β ITAM	$p_{L\beta}$	36	5	1.70	1.14
Lyn* → β ITAM	$p^*_{L\beta}$	36	9	4.10	1.63
Lyn → γ ITAM	$p_{L\gamma}$	24	4	0.08	0.04
Lyn* → γ ITAM	$p^*_{L\gamma}$	24	7	0.18	0.08
Lyn → Syk	p_{LS}	48	8	0.35	0.14
Lyn* → Syk	p^*_{LS}	48	12	13.42	6.12
Syk → Syk	p_{SS}	64	11	2.02	0.63
Syk* → Syk	p^*_{SS}	64	10	19.17	6.10
Dephosphorylation	d	628	53	41.00	6.07
Total	17 rate constants	2326	230	100.00	35.33

[a]Complete definitions of reaction classes and rate constants are in [9]

intracomplex phosphorylation, that are difficult to measure or assess.

Time courses. Elementary mass action kinetics give rise to a system of coupled ordinary differential equations (ODEs) that describe the time evolution of the species concentrations following the addition of ligand. These ODEs are solved numerically using the stiff solver CVODE [34], which is called by BioNetGen.

Distribution of network activity. We adopt a simple measure to determine the identity and number of active elements (species, reactions, or paths) in the network: the smallest set of elements that cumulatively account for a prescribed fraction of the total concentration (for species) or flux (for reactions and paths). This set is determined by rank ordering the elements by relative concentrations or flux from highest to lowest and dropping the remaining elements from the list when the cumulative sum of the first n elements crosses the cut-off fraction. These first n elements are considered active. The choice of cut-off is arbitrary, but for a uniform distribution over the network elements, the fraction of network elements that are active equals the cut-off value. When the fraction of active elements is much smaller than the cut-off value, the distribution can be considered skewed. An example of such a skewed distribution that is typical of our results is that only about 7% of the possible species containing activated Syk account for more than 95% of the activated Syk concentration.

Syk activation paths. We define an activation path as a sequence of reaction events by which a molecular component of the model is transformed from an inactive state into an active one. Here, we analyse the paths that transform an unphosphorylated Syk molecule in the cytosol into an autophosphorylated Syk molecule associated with a receptor dimer complex (Syk^*). As described in more detail in Appendix, Section 7.1, we use a deterministic algorithm to enumerate paths as a function of the path length and a stochastic algorithm to compute the relative contribution of each path to the rate of Syk^* production.

Parameter set ensembles. To determine the possible effect of the initial cellular state on the distribution of network activity, we generated two ensembles of 5000 randomly scaled sets of parameter values, referred to as the 2x and 10x ensembles.

Each new parameter set is produced by scaling each of the parameters in the original model, rate constants and concentrations, by an amount x^p, where p is a uniformly distributed random variable on the interval $[-1,1]$ chosen separately for each parameter. The ligand concentration is 1 nM in the unscaled parameter set, but is varied along with the other parameters in the scaled parameters sets. Two parameters, the forward rate constant for ligand-receptor binding (k_{+1}) and the forward rate constant for receptor cross-linking (k_{+2}) were not varied. Thus, the input signal in the scaled parameter sets varies only through the variation of the total ligand concentration.

The ensembles are labelled by their x value, $x = 2$ or $x = 10$. For each new parameter set generated, the time evolution of the 307 chemical species is obtained as described above. A fixed time of 100 s was chosen for sampling the distributions of activated Syk and reactive fluxes. Variation of the parameters affects the time required to achieve steady state, but the sampling time of 100 s generally occurs during the transient phase of signalling when species concentrations are changing rapidly. For example, in the unscaled parameter

set the level of Syk activation at 100 s is about 20% of its steady state level (Syk activation increases monotonically with time, as shown in Fig. 2). Over the full range of ligand concentrations sampled in the 10x ensemble, Syk activation (calculated without scaling the remaining parameters) at 100 s ranges from 2% of its steady-state value at 0.1 nM to 70% of its steady-state value at 10 nM. Increasing the sampling time to 1000 or 10 000 s was found to have a negligible effect on the ensemble results (shown in Fig. 4). A later sampling time of 1000 s was chosen for the activation path distribution to ensure the accuracy of the path sampling method (see Appendix, Section 7.1), the validity of which depends on steady-state conditions.

Model Reduction. We used an optimisation procedure based on deleting species from the full network model to find the smallest network that will reproduce the time course of the full model for a set of observed quantities to within a specified error. When a species is deleted from the network, all reactions associated with that species are also removed, but none of the remaining reactions or reaction rate constants are changed. The objective function used to test the fitness of a reduced model is the root-mean squared (RMS) of the relative error computed over all quantities and time points. The six quantities, which correspond to observable properties that either have been or could be measured for this system are FcεRIβ ITAM phosphorylation, FcεRIγ ITAM phosphorylation, Syk linker region phosphorylation, Syk kinase activation loop phosphorylation, association of Lyn with the unphosphorylated FcεRIβ subunit, and association of Lyn with the phosphorylated FcεRIβ ITAM measured at 1,10,100, and 1000 s after addition of ligand. Details of the optimisation algorithm are presented in the Appendix, Section 7.2.

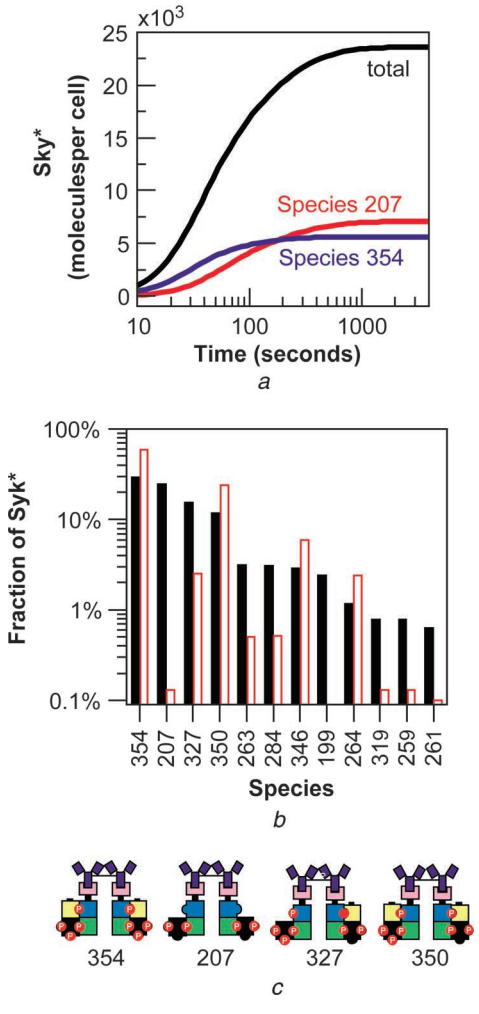

Fig. 2 *Predicted distribution of activated Syk (Syk*) after introduction of IgE dimer (10 nM) at time t = 0 s. Calculations were performed using the BioNetGen software package [32] using parameter estimates for the RBL-2H3 cell line [9], except as noted below. (a) Time courses for the total amount of Syk* (black curve) and the amount of Syk* in each of the two species containing the most Syk* at t = 100 s (red and blue curves). (b) Rank ordered distribution of Syk* at t = 100 s (black bars) and when the Lyn concentration is increased ten-fold (red bars). The 12 complexes indicated account for more than 95% of the Syk* at t = 100 s, and five of these account for 95% of the mass when the Lyn concentration is increased ten-fold. The indices used to refer to complexes are defined at our web site (http://cellsignaling.lanl.gov). Species 354, 350, 346, and 264 each contain two bound Lyn molecules; Species 207 and 199 contain no bound Lyn; and Species 327, 263, 284, 319, 259, and 261 contain one bound Lyn molecule. (c) Illustration of the four species containing the most Syk* at t = 100 s*

3 Results

In order to characterise the spread of activity in the reaction network, we consider three distributions: the distribution of activated Syk among chemical species, the distribution of reactive flux among reactions in the same class, and the distribution of frequency among paths that lead to activated Syk. These distributions are obtained for a default set of parameters that characterise the rat basophilic leukaemia cell line (RBL-2H3) [9], for the default set with the Lyn concentration increased ten-fold, and finally for ensembles of parameter sets in which the default values are randomly varied over two-fold and ten-fold ranges.

Distribution of activated Syk (Syk*). A Syk molecule that is bound to a receptor and has been phosphorylated by a second Syk is considered to be activated. The 164 species that contain Syk* represent chemically distinct output channels of the signalling model. We find that only a few of these channels dominate the distribution of Syk* at all times following addition of ligand. The two most populated species, 354 and 207, contain more than 50% of the Syk* (Fig. 2*a*), and 12 species contain more than 95% of the Syk* (Fig. 2*b*, black bars). Although relatively few Syk* species are populated, the composition of these species is heterogeneous (Fig. 2*c*), varying in the amount of associated Lyn and in the level of Lyn-mediated phosphorylation of Syk. For example, Species 354 contains two Lyn molecules and two Lyn-phosphorylated Syk* molecules, whereas Species 207 contains no Lyn and neither of its two Syk* molecules is Lyn-phosphorylated. This heterogeneity may have functional consequences, because Lyn and Lyn-phosphorylated Syk contain binding sites for signalling molecules [35–39] including Cbl, the p85 subunit of phosphatidylinositol-3′ kinase, and phospholipase Cγ. As a result, molecules associated with Lyn-containing and Lyn-deficient Syk* species can differ and the different signalling complexes have the potential to trigger distinct downstream signalling events.

The predicted distribution of Syk* changes during the response to stimulation (Fig. 2*a*). The Lyn-containing complex, 354, exhibits faster initial kinetics than the Lyn-deficient complex, 207, but as receptor phosphorylation increases, the pool of free Lyn available to bind receptors is depleted [40], and 207 replaces 354 as the most abundant form of Syk*. Thus, the temporal redistribution of Syk* could have functional consequences if co-localisation of Lyn and Syk has a strong effect on downstream signals.

The predicted distribution of Syk* also depends on the initial state of a cell. As illustrated in Fig. 2*b*, the distribution of Syk* can be shifted by a change in

Table 2: Number of possible paths and frequency of observed Syk activation paths as a function of path length. The number of observed paths and fraction of the total activation flux accounted for by paths of a given length are determined by stochastic sampling of 10^7 successful activation events at steady state, when all FcεRI are aggregated into dimers

Path length	Number of possible paths	Number of observed paths	Fraction of activation flux
2	64	64	38.2%
3	384	287	26.3%
4	2,056	773	12.6%
5	14,068	1,434	4.8%
6	108,728	1,831	4.4%
7	845,800	2,026	4.3%
8	6,301,796	2,204	3.3%
9	44,621,932	3,081	2.1%
10	300,913,268	4,206	1.3%
Total	352,808,096	15,906	97.3%

the concentration of a single component. Increasing the concentration of Lyn ten-fold causes a redistribution of Syk* into Lyn-containing complexes (Fig. 2*b*, red bars). The effect on Lyn-deficient states can be quite large: for example, the fraction of Syk* in Species 207 drops by more than a factor of 1000. Thus, a cellular response that depends on co-localisation of Lyn and Syk could be upregulated (downregulated) by increasing (decreasing) the expression of Lyn. Unfortunately, without including additional components in the model, it is difficult to predict how co-localisation would affect activity. For example, Lyn-containing Syk* complexes might upregulate Syk-dependent responses because Lyn binds the regulatory subunit of phosphatidylinositol-3′ kinase (PI3K) [35], whose catalytic activity creates plasma membrane binding sites for a number of known Syk substrates [39]. On the other hand, Lyn-containing Syk* complexes might downregulate Syk-dependent responses because Lyn phosphorylation of Syk on Tyr-317 creates a binding site for the ubiquitin ligase Cbl, which marks Syk for degradation and may block the direct binding of PLC-γ to other phosphotyrosine residues on Syk [38]

Distribution of reaction rates. Another way to measure the importance of network elements is to examine rates of individual chemical reactions. As described above, the model is constructed by lumping together similar chemical transformations into classes described by a single rate constant [9]. For example, the rate constant for Lyn binding to the phosphorylated β ITAM (k^*_{+L}) is independent of whether Lyn or Syk is bound to any of the other sites within a receptor aggregate and is used to characterise 144 distinct chemical reactions. Since the rate of each reaction in the model is given by the product of the rate constant and the concentrations of the chemical species involved, the distribution of reaction rates within the same class mirrors the distribution of complexes that can participate in the reaction class. Just as the 164 Syk*-containing complexes represent alternative output channels of the model, the multiple reactions within each class represent alternative conduits of flow. The 17 reaction classes considered in our analysis and a breakdown of their rate distributions for the default parameter set are given in Table 1. The number of important reactions within each class (defined, as above, by a 95% cut-off) is always a small fraction of the total number of reactions within a class. Cumulatively, only about 10% of the reactions in the network are characterised as important. A similarly narrow distribution of reaction rates is observed when the Lyn concentration is increased ten-fold (results not shown).

Distribution of activation paths. Our final measure of network activity is the steady-state distribution of flux among reaction paths from inactive to activated Syk. Such a path is a non-repeating ordered sequence of reactions that transforms unphosphorylated cytosolic Syk into Syk*. The number of theoretically possible activation paths grows exponentially as a function of path length and far exceeds the number of molecules in the system (Table 2), but only 12 paths account for 50% of the total activation flux and ~1000 paths account for 95%. The top two paths (Fig. 3), both involve Syk binding to a receptor that is already bound to Syk. Such shortcutting paths minimise the opportunity for branching and are thus a major contributing factor to the narrow distribution of path flux. Path 54 (Fig. 3) has the highest flux among activation paths in which Syk initially binds to a complex containing no associated kinases. Activation of Syk along such paths requires additional Lyn and Syk binding events and gives rise to more branching opportunities and a greater diversity of possible paths.

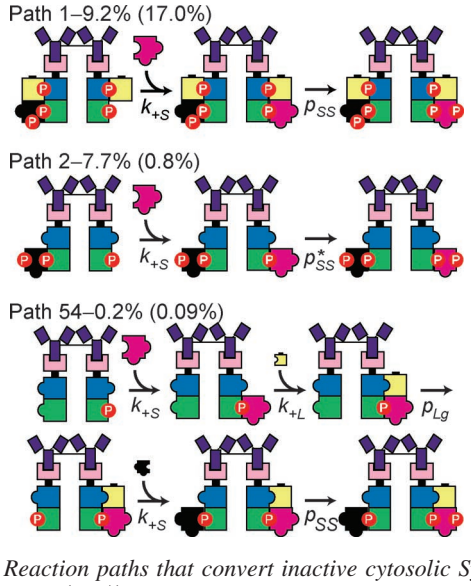

Fig. 3 *Reaction paths that convert inactive cytosolic Syk to the activated form (Syk*) under steady-state conditions. The paths are indexed by the rank of their relative flux, which is given as a percentage of the total activation flux, the rate of Syk* turnover. The relative flux of each path when the total Lyn concentration is increased ten-fold is shown in parentheses. Relative fluxes are determined from a sample of 10^7 randomly generated successful activation sequences*

Such paths, however, are relatively rare, cumulatively accounting for only 4% of the total activation flux.

Thus, most Syk activation does not follow an extended sequence of reaction events like that shown in Fig. 1*c*. The species along the top two paths of Fig. 3 also exhibit the split levels of Lyn association that were observed in the top two Syk* complexes shown in Fig. 2. Increasing the Lyn concentration ten-fold dramatically reduces the flux of activation paths (values shown in parentheses in Fig. 3) involving complexes without Lyn (Paths 2 and 54).

Variation of parameter values. To test whether a narrow distribution of network activity depends on parameterisation of the model, we examine the three measures of the activity distribution for different sets of randomly altered parameter values. The level of Syk activation varies widely among parameter sets (Fig. 4*a*), but all parameter sets yield narrow distributions of network activity in comparison to a uniform distribution into all possible Syk*-containing species, reactions, or Syk activation paths (Fig. 4*b–d*). For two-fold variations of parameters, each measure of activity is symmetrically distributed about the value characteristic of the original parameter set. For ten-fold variations of parameters, the average value of each measure decreases, although each distribution has a long tail that extends to higher values (Fig. 4*b–d*).

Systematic variation of parameter values confirms the example of Fig. 2*b*: the identity and relative contribution of important network elements can change depending on parameter values (Fig. 4*e–g*). Figure 4*e* shows how the fractional contribution of Species 354, the species containing the highest concentration of Syk* using the original parameter set, is distributed in the 2x and 10x parameter set ensembles. Species 354 contains ~ 30% of the Syk* using the original parameter set (Fig. 2*a–b*), and its fractional contribution is distributed symmetrically about this value in the 2x ensemble over a range of ~10–60%. However, in the 10x ensemble, the distribution changes substantially, with the most frequent value of the fractional contribution tending towards zero (i.e. no Syk* in this state). The fractional contributions of the Syk autophosphorylation reaction with the highest reaction rate (Fig. 4*f*) and the Syk activation path with highest relative flux (Fig. 4*g*) exhibit similar behaviour. Thus, the relative contribution of an important network element is robust to modest (two-fold) parameter variations, but larger (ten-fold) parameter variations usually cause activity to shift elsewhere in the network.

Model reduction. If a relatively small portion of the signalling network is active, one might expect that the FcεRI model could be reduced in size without changing its predictions. We tested this idea by removing species and their associated reactions to reduce the network size while

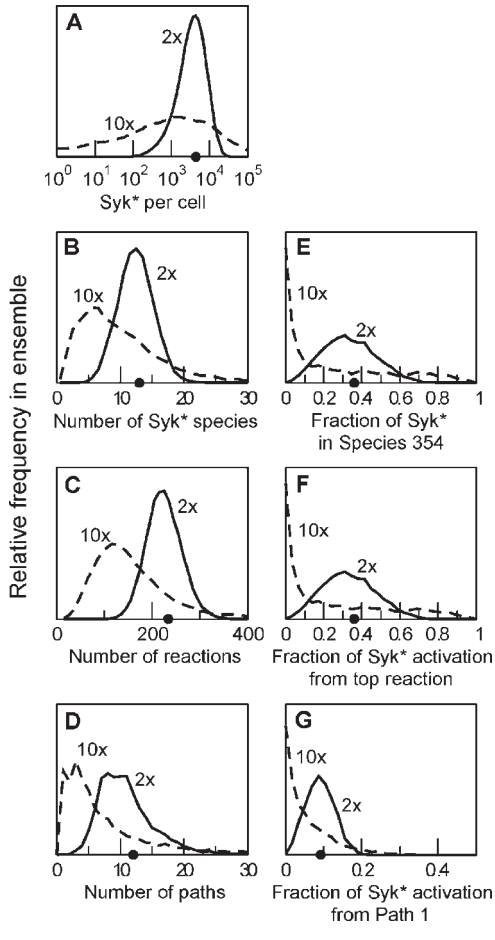

Fig. 4 *Effect of random variation of model parameter values on the distribution of network activity. The distributions of Syk* in the output species and of reaction rates grouped by rate constant are determined at t = 100 s following stimulation with IgE dimer. The distribution of activation path fluxes is sampled at t = 1000 s. For each of the following properties the panels plot relative frequency of occurrence in the 2x (solid lines) and 10x (dashed lines) ensembles: (a) Total level of Syk*. (b) Number of important Syk* species (account for more than 95% of Syk*). (c) Number of important reactions (carry more than 95% of the reaction flux in all reaction classes, as defined in Table 1). (d) Number of important Syk activation paths (carry more than 50% of the Syk activation flux, as determined from a sample of 10^5 activation events for each parameter set). (e) Fraction of Syk* contained in Species 354 (see Fig. 2c). (f) Fraction of Syk activation due to the Syk autophosphorylation reaction with the highest flux using the original parameter values. (g) Fraction of Syk activation due to Path 1 of Fig. 3. Filled circles on the x-axis indicate the value of each property calculated using the original parameter values*

Table 3: Performance of reduced models measured by the RMS error of six observables (FcεRIβ ITAM phosphorylation, FcεRIγ ITAM phosphorylation, Syk linker region phosphorylation, Syk kinase activation loop phosphorylation, low-affinity Lyn-receptor association, and high-affinity Lyn-receptor association) at three time points (*t* = 10, 100, and 1000 sec). Results are representative of at least three reduced models with the same number of nodes produced by separate optimisation runs

	Default set	2 × ensemble	10 × ensemble
44 state model			
(104 reactions)			
Mean RMS error	6.5%[a]	56%	50,000%
% sets RMS error < 10%	–	4.6%	0.2%
% sets RMS error > 50%	–	28%	77%
83 state model			
(257 reactions)			
Mean RMS error	3.0%[b]	45%	120%
% sets RMS error < 10%	–	16%	1.8%
% sets RMS error > 50%	–	22%	54%

[a]Error with 1 nM IgE dimer stimulation. Error is 10% when objective function is evaluated at conditions under which model reduction was performed (10 nM IgE dimer stimulation, objective function computed at *t* = 1, 10, 100, and 1000 s)

[b]Error with 1 nM IgE dimer stimulation. Error is 10% when objective function is evaluated at conditions under which model reduction was performed (10 nM IgE dimer stimulation, objective function computed at *t* = 1, 10, 100, and 1000 s)

minimising the error of six specified output functions in comparison to the predictions of the full model. Permitting a maximum RMS relative error of 10%, the smallest network we found contained 44 species and 104 reactions (Table 3). Although predictions of this model match those of the full model for the original parameter values, the reduced model is not predictive over a range of parameter values. Even for the 2x ensemble of altered parameter sets, the reduced model exhibits RMS errors outside the 10% tolerance in the vast majority of cases and exhibits $>50\%$ RMS error in a substantial fraction of cases. These results are insensitive to the size of the error tolerance used in model reduction (Table 3). The propensity of network activity to shift with parameter variations (Fig. 4) appears to limit the possibility of finding reduced models that apply over a broad range of cellular conditions.

4 Discussion

The protein-protein interactions of signal transduction [3], typified in the model considered here, generally imply a vast biochemical network, comprising a multitude of protein states and complexes and reactions among these. One issue that modellers of signal transduction must confront is whether this complexity affects the fundamental behaviour of the system or whether most of it may be safely ignored, as is common practice. The formulation of a simplified model amounts to assuming that a small number of states can effectively represent a multitude of potential states. One problem with such simplifying assumptions, aside from questions of accuracy, is that they limit the ability of models to predict the effect of typical experimental manipulations, such as knocking out specific sites of phosphorylation or domains of proteins.

We have attempted here to assess the role of molecular diversity in signal transduction by characterising the diversity of complexes, reactions, and activation pathways that arise in a detailed model of early signalling events in a particular pathway initiated by receptor aggregation. We find that for any given state of the cell, characterised by a particular set of model parameters, only a small fraction of the network appears to be active (Fig. 2 and Table 1), but changing the cell's state can change which elements are active. The spectrum of active complexes in the model can be shifted dramatically, even by a change in the concentration of a single protein (Fig. 2*b* and Fig. 3). Random variation of the model parameters demonstrates that the narrow distribution of network activity is a robust feature of the model (Fig. 4). The set of important network elements is generally robust to modest (two-fold) perturbations of rate constants and concentrations, and major shifts in activity require large (ten-fold) variations. It is possible to find reduced models that reproduce the behaviour of the full model for particular parameter values, but the predictions of these models are poor for perturbed cellular states (Table 3). They cannot be expected to predict accurately, for example, the effects of knocking out a particular protein domain. We conclude, therefore, that the assumptions of simplified models should be carefully validated before such extrapolations are made. The results of model reduction suggest that it will be difficult to find simplified models that are predictive over a broad range of cellular states.

One question that arises from our study is whether the topology of the network alone is sufficient to guarantee the narrow distributions of activity we observe. A simple numerical experiment demonstrates that this is not the case. Setting all four initial concentrations and 17 rate constants to unity, we find that more than 70% of the possible Syk* species are active, as compared with about 7% using the RBL cell parameters. Thus, variation in the levels of protein expression and values of rate constants are essential for producing narrow flows.

A related question is whether other large network models will also exhibit focused distributions of network activity. We have recently constructed and analysed a network model of early events in signalling through EGFR [11]. Interestingly, we find that at steady-state, the narrow distribution of active species is comparable to that of the FcεRI model, but there is a much broader transient distribution that encompasses about 30% of the possible species. The broad distribution of active species appears to arise from the roughly equal concentrations of receptor-binding proteins that produce complexes of broadly varying stoichiometry. A dramatic reduction in molecular diversity occurs at steady state because receptor-binding proteins are sequestered into a few cytosolic complexes. The limited supply of these proteins for receptor binding restricts the stoichiometry of the complexes that can form, limiting the observed diversity of species.

This study demonstrates that network dynamics alone, even in the absence of feedback or cooperative interactions, can produce highly focused flows of mass and information in a signalling network. Moreover, we have seen that these flows can be regulated by parameters such as protein expression levels and enzymatic activities. One might expect such focused flows to arise from other mechanisms, such as cooperativity, feedback, or localisation. These mechanisms may well restrict the range of complexes that form during response to a signal, but observation of limited molecular diversity among signalling complexes cannot be attributed to any particular mechanism without models that incorporate all of the potential mechanisms for limiting diversity. In particular, interpretation of proteomic data [12–16], assays of the protein phosphorylation states and complexes generated during signalling, will require models of the type analysed here to obtain mechanistic insights.

Experimental evidence for the role of differential complex formation in shaping cellular responses comes from studies of kinetic proofreading in immunoreceptor signalling (recently reviewed in [4]), which indicate that the signalling properties of a ligand are sensitive to the lifetime of ligand-receptor binding. Ligands with longer association lifetimes tend to signal more effectively because they generate 'mature' signalling complexes that carry the signal downstream, whereas shorter binding ligands produce 'frustrated' complexes that do not signal and can actually inhibit the production of mature complexes by sequestering signalling components in limited supply. Such 'antagonist' ligands have been shown to produce both altered patterns of receptor phosphorylation [41] and kinase-sequestering complexes that inhibit signalling by more strongly binding ligands [42, 43]. Both of these effects are predicted by detailed models of early signalling events [40, 44], which provide theoretical support for the ideas incorporated in simplified models of kinetic proofreading [45, 46]. In terms of the potential role that differential complex formation may play in determining and regulating signalling outcomes, these effects represent just a few possibilities. Investigating these should be a major focus of computational studies of signal transduction [6, 47, 48] in the near future.

We have shown here that the pattern of complexes formed during a response to a signal can be sensitive to

quantitative parameters that define the initial state of the cell. Because the spectrum of active complexes in our model can be shifted dramatically, even by a change in the concentration of a single protein, one function of the combinatorial complexity found in signalling systems might be to provide a mechanism for cellular decision making. Any event that changes the expression level or activity of a component of the cell could affect signal processing through a cascade involving that component, by changing the composition of signalling complexes that are generated. In this way, the complexity of signalling complexes, which until now has been merely perplexing, might turn out to be an essential element of cellular computation.

5 Acknowledgments

We dedicate this paper to the memory of Carla Wofsy. Thanks to Dan Coombs, Tony Redondo, and Henry Metzger for helpful discussions. This work was supported by grants GM35556 and RR18754 from the National Institutes of Health and by the Department of Energy through contract W-7405-ENG-36.

6 References

1 Hunter, T.: 'Signaling - 2000 and beyond', *Cell*, 2000, **100**, pp. 113–127
2 Tomlinson, M., Lin, J., and Weiss, A.: 'Lymphocytes with a complex: adapter proteins in antigen receptor signalling', *Immunol. Today*, 2000, **21**, pp. 584–591
3 Pawson, T., and Nash, P.: 'Assembly of cell regulatory systems through protein interaction domains', *Science*, 2003, **300**, pp. 445–452
4 Goldstein, B., Faeder, J.R., and Hlavacek, W.S.: 'Mathematical and computational models of immune-receptor signalling', *Nat. Rev. Immunol.*, 2004, **4**, pp. 445–456
5 Arkin, A.P.: 'Synthetic cell biology', *Curr. Opin. Biotechnol.*, 2001, **12**, pp. 638–644
6 Endy, D., and Brent, R.: 'Modelling cellular behaviour', *Nature*, 2001, **409**, pp. 391–395
7 Bray, D.: 'Molecular prodigality', *Science*, 2003, **299**, pp. 1189–1190
8 Hlavacek, W.S., Faeder, J.R., Blinov, M.L., Perelson, A.S., and Goldstein, B.: 'The complexity of complexes in signal transduction', *Biotechnol. Bioeng.*, 2003, **84**, pp. 783–794
9 Faeder, J.R. *et al.*: 'Investigation of early events in FcεRI-mediated signaling using a detailed mathematical model', *J. Immunol.*, 2003, **170**, pp. 3769–3781
10 Conzelmann, H., *et al.*: 'Reduction of mathematical models of signal transduction networks: simulation-based approach applied to EGF receptor signalling', *IEE Syst. Biol.*, 2004, **1**, pp. 159–169
11 Blinov, M.L., Faeder, J.R., Goldstein, B., and Hlavacek, W.S.: 'A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity', submitted
12 Aebersold, R., and Mann, M.: 'Mass spectrometry-based proteomics', *Nature*, 2003, **422**, pp. 198–207
13 Mann, M., and Jensen, O.N.: 'Proteomic analysis of post-translational modifications', *Nat. Biotechnol.*, 2003, **21**, pp. 255–261
14 Meyer, T., and Teruel, M.N.: 'Fluorescence imaging of signaling networks', *Trends Cell Biol.*, 2003, **13**, pp. 101–106
15 Salomon, A.R., *et al.*: 'Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry', *Proc. Natl. Acad. Sci. USA*, 2003, **100**, pp. 443–448
16 Blagoev, B., Ong, S.E., Kratchmarova, I., and Mann, M.: 'Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics', *Nat. Biotechnol.*, 2004, **22**, pp. 1139–1145
17 Pacini, S., Valensin, S., Telford, J.L., Ladbury, J., and Baldari, C.T.: 'Temporally regulated assembly of a dynamic signaling complex associated with the activated TCR', *Eur. J. Immunol.*, 2000, **30**, pp. 2620–2631
18 Wilson, B.S., Pfeiffer, J.R., and Oliver, J.M.: 'FcεRI signaling observed from the inside of the mast cell membrane', *Mol. Immunol.*, 2002, **38**, pp. 1259–1268
19 Bunnell, S.C.: 'T cell receptor ligation induces the formation of dynamically regulated signaling assemblies', *J. Cell Biol.*, 2002, **158**, pp. 1263–1275
20 Blagoev, B.: 'A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signalling', *Nat. Biotechnol.*, 2003, **21**, pp. 315–318
21 Kholodenko, B., Demin, O., Moehren, G., and Hoek, J.: 'Quantification of short term signaling by the epidermal growth factor receptor', *J. Biol. Chem.*, 1999, **274**, pp. 30169–30181
22 Schoeberl, B., Eichler-Jonsson, C., Gilles, E.D., and Muller, G.: 'Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors', *Nat. Biotechnol.*, 2002, **20**, pp. 370–375
23 Resat, H., Ewald, J., Dixon, D., and Wiley, H.: 'An integrated model of epidermal growth factor receptor trafficking and signal transduction', *Biophys. J.*, 2003, **85**, pp. 730–743
24 Kinet, J.P., 'The high-affinity IgE receptor (FcεRI): From physiology to pathology', *Annu. Rev. Immunol.*, 1999, **17**, pp. 931–972
25 Goldstein, B., *et al.*: 'Modeling the early signaling events mediated by FcεRI', *Mol. Immunol.*, 2002, **38**, pp. 1213–1219
26 Vonakis, B.M., Chen, H.X., HaleemSmith, H., and Metzger, H.: 'The unique domain as the site on Lyn kinase for its constitutive association with the high affinity receptor for IgE', *J. Biol. Chem.*, 1997, **272**, pp. 24072–24080
27 Segal, D.M., Taurog, J.D., and Metzger, H.: 'Dimeric Immunoglobulin-E Serves as a Unit Signal for Mast-Cell Degranulation', *Proc. Natl. Acad. Sci. USA*, 1977, **74**, pp. 2993–2997
28 Kent, U.M.: 'Dynamics of Signal-Transduction after Aggregation of Cell-Surface Receptors: Studies on the Type-I Receptor for IgE', *Proc. Natl. Acad. Sci. USA*, 1994, **91**, pp. 3087–3091
29 Wofsy, C., Kent, U.M., Mao, S.Y., Metzger, H., and Goldstein, B.: 'Kinetics of tyrosine phosphorylation when IgE dimers bind to Fcε receptors on rat basophilic leukemia-cells', *J. Biol. Chem.*, 1995, **270**, pp. 20264–20272
30 Turner, H., and Kinet, J.P.: 'Signalling through the high-affinity IgE receptor FcεRI', *Nature*, 1999, **402**, pp. B24–B30
31 Nadler, J.S., Matthews, S.A., Turner, M., and Kinet, J.P.: 'Signal transduction by the high-affinity immunoglobulin E receptor FcεRI: Coupling form to function', *Adv. Immunol.*, 2001, **76**, pp. 325–355
32 Blinov, M.L., Faeder, J.R., Goldstein, B., and Hlavacek, W.S.: 'BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains', *Bioinformatics*, 2004, **20**, pp. 3289–3291
33 Faeder, J.R., Blinov, M.L., Goldstein, B., and Hlavacek, W.S.: 'Rule-based modeling of biochemical networks', *Complexity*, in press
34 Cohen, S.D., and Hindmarsh, A.C.: 'CVODE, A stiff/nonstiff ODE solver in C', *Comput. Phys.*, 1996, **10**, pp. 138–143
35 Pleiman, C.M., Hertz, W.M., and Cambier, J.C.: 'Activation of phosphatidylinositol-3′ kinase by src-family kinase SH3 binding to the p85 subunit', *Science*, 1994, **263**, pp. 1609–1612
36 Qu, X.J. *et al.*: 'Negative regulation of FcεRI-mediated mast cell activation by a ubiquitin-protein ligase Cbl-b', *Blood*, 2004, **103**, pp. 1779–1786
37 Paolini, R. *et al.*: 'Activation of Syk tyrosine kinase is required for c-Cbl-mediated ubiquitination of FcεRI and Syk in RBL cells', *J. Biol. Chem.*, 2002, **277**, pp. 36940–36947
38 Hong, J.J., Yankee, T.M., Harrison, M.L., and Geahlen, R.L.: 'Regulation of signaling in B cells through the phosphorylation of Syk on linker region tyrosines', *J. Biol. Chem.*, 2002, **277**, pp. 31703–31714
39 Fruman, D.A., and Cantley, L.C.: 'Phosphoinositide 3-kinase in immunological systems', *Semin. Immunol.*, 2002, **14**, pp. 7–18
40 Wofsy, C., Torigoe, C., Kent, U.M., Metzger, H., and Goldstein, B.: 'Exploiting the difference between intrinsic and extrinsic kinases: Implications for regulation of signaling by immunoreceptors', *J. Immunol.*, 1997, **159**, pp. 5984–5992
41 Kersh, E.N., Shaw, A.S., and Allen, P.M.: 'Fidelity of T cell activation through multistep T cell receptor ζ phosphorylation', *Science*, 1998, **281**, pp. 572–575
42 Torigoe, C., Inman, J.K., and Metzger, H.: 'An unusual mechanism for ligand antagonism', *Science*, 1998, **281**, pp. 568–572
43 Torigoe, C., Goldstein, B., Wofsy, C., and Metzger, H.: 'Shuttling of initiating kinase between discrete aggregates of the high affinity receptor for IgE regulates the cellular response', *Proc. Natl. Acad. Sci. USA.*, 1997, **94**, pp. 1372–1377
44 Lee, K.H. *et al.*: 'The immunological synapse balances T cell receptor signaling and degradation', *Science*, 2003, **302**, pp. 1218–1222
45 McKeithan, T.W.: 'Kinetic Proofreading in T-cell receptor signal-transduction', *Proc. Natl. Acad. Sci. USA.*, 1995, **92**, pp. 5042–5046
46 Hlavacek, W.S., Redondo, A., Metzger, H., Wofsy, C., and Goldstein, B.: 'Kinetic proofreading models for cell signaling predict ways to escape kinetic proofreading', *Proc. Natl. Acad. Sci. USA.*, 2001, **98**, pp. 7295–7300
47 Kitano, H.: 'Computational systems biology', *Nature*, 2002, **420**, pp. 206–210
48 Wiley, H., Shvartsman, S., and Lauffenburger, D.: 'Computational modeling of the EGF-receptor system: a paradigm for systems biology', *Trends Cell Biol.*, 2003, **13**, pp. 43–50
49 Ahuja, R.K., Magnanti, T.L., and Orlin, J.B.: 'Network flows' (Prentice Hall, Upper Saddle River, New Jersey, 1993)
50 Gillespie, D.T.: 'A general method for numerically simulating the stochastic time evolution of coupled chemical reactions', *J. Comp. Phys.*, 1976, **22**, pp. 403–434

7 Appendix

7.1 Enumeration and sampling of activation paths

We define an activation path as a sequence of reaction events by which a molecular component of the model is transformed from an inactive state into an active one. Here, we focus on paths that transform an unphosphorylated Syk

molecule in the cytosol into in an autophosphorylated Syk molecule associated with a receptor dimer complex (Syk^*), but the methods can be easily generalised. The full reaction network is first transformed into a directed graph (a set of nodes and directional edges connecting nodes), from which activation paths are defined, enumerated, and sampled to determine relative activation fluxes.

Constructing the component activation graph. Each node in this activation graph represents a distinct state of Syk in the model. Nodes are created from the species that contain Syk; species that contain one Syk molecule give rise to one node, but species that contain multiple Syk molecules in distinct states give rise to multiple nodes. For example, in the second species of Path 1 in Fig. 3, the labelled Syk may be associated with either the left or the right receptor of the complex. Thus, to account for both possibilities, we must include two nodes in the graph for this species. The edges of the activation graph correspond to directed chemical transitions between nodes that can be carried out in a single reaction step. Edges are created from the reactions in the model that involve Syk; one edge is created for each distinct pair of reactant and product nodes arising from the reaction. Reactions that contain multiple Syk molecules give rise to multiple edges. For example, the first reaction shown in Path 1 of Fig. 3, where the labelled Syk may be either cytosolic (the purple Syk) or associated with a receptor complex (the black Syk), gives rise to two edges. The weight of an edge is given by the rate at which a molecule of the labelled Syk is transformed by the reaction. (If multiple reactions carry out the same transformation of nodes, the weight is the sum of the relevant rates). For the example given above, the weight of the edge involving transformation of the purple Syk is given by k_{+S} times the concentration of the species containing the black Syk, whereas the weight of the edge involving the transformation of the black Syk is given by k_{+S} times the concentration of free Syk in the cytosol. (Note that these weights are in general time dependent.) Reactions involving the loss of Syk from a symmetric complex can give rise to two edges from a single reactant node. If the number of Syk molecules in the complex is s, the weight of the edge for dissociation of the labelled Syk from the complex is $1/s$, and the weight of the edge for retention of the labelled Syk is $(s\text{-}1)/s$. The Syk activation graph constructed in this manner contains 420 nodes, of which 192 represent activated states of the labelled Syk, and 3644 non-zero edges (for irreversible ligand binding). The Syk activation graph is available from the authors upon request.

Formal definition of an activation path. A Syk activation path is defined as an ordered sequence of nodes of this graph, where the first node corresponds to unphosphorylated, cytosolic Syk, and the final node is the first node in the sequence in which the labelled Syk is autophosphorylated and part of a receptor dimer complex. Each pair of adjacent nodes in this sequence must be connected by an edge with non-zero weight. To simplify our analysis, we restrict the definition of a path to include only those sequences in which each node appears at most one time, to avoid cycles within paths.

Enumeration of paths. The enumeration of possible paths as a function of path length (column 2 of Table 2) is carried out using a modified form of the depth-first search [49]. Paths up to length N are enumerated as follows. A path is implemented as a stack (elements are added to and removed from the end of the list) and is initialised with a starting node corresponding to unphosphorylated, cytosolic Syk. (I) Loop over the edges originating from the final node of the path. If the final node of an edge corresponds to an active state of the labelled Syk, increment the number of paths of length n, where n is the number of nodes in the path, and continue with loop (I). If $n < N$ and the final node of the edge is not a member of the current path, add this node to the path and begin a new loop at (I). (II) When the edges from the final node in a path are exhausted, remove this node from the path and continue with loop (I) if the path still contains at least one element. The recursive looping implied by (I) is implemented using a second stack that contains a pointer to the current edge for each node in the path stack, where the edges for each node are stored in a linked list and looped over in that order.

Determination of activation flux. Sampling of paths to determine their relative contribution to the total activation flux is done using a stochastic algorithm based on Gillespie's method for computing chemical dynamics [50]. The first node in a path is unphosphorylated, cytosolic Syk. Paths are extended from the terminal node i in the path sequence by choosing the next node j randomly with probability $p_{i\to j} = w_{i\to j}/\sum_{j'} w_{i\to j'}$, where $w_{i\to j}$ is the weight of the edge taking i into j. Paths are terminated when the Syk molecule being traced is autophosphorylated (successful activation) or when it returns to the cytosol in its unmodified state. Sampling of paths continues until a specified number of activation events is recorded. Following a successful trace, the path is pruned to remove loops by iteratively removing all nodes between repeating nodes (including one instance of the repeating node) until no more repeating nodes are present in the path. The relative activation flux from a given activation path p is (# times p observed)/(# activation events). Edge weights are determined from the species concentrations at a particular sampling time and assumed to be constant. This assumption is equivalent to assuming that the relative change in the species concentrations is small over the duration of the activation events being sampled, which holds exactly under the steady-state conditions used to generate the data in Fig. 3 and Table 2. This assumption is also reasonably accurate for the $t = 1000$ s time point used in the sampling of activation paths in the parameter set ensembles, because the duration of the vast majority of activation events is at most a few seconds and species concentrations undergo small fractional changes on this timescale at a time so distant from the addition of ligand.

7.2 Algorithm for model reduction

The optimisation algorithm is as follows. The starting network is taken to be the full network. (1) At each step a move is attempted in which a species is randomly deleted from the network. If the objective function (defined in Section 2) is below the threshold value for the RMS, the deletion is accepted and the optimisation continues from (1). (2) Following a failed move, the deletion is removed from the list of possible deletions from the current network. (a) After a sequence of 50 failed deletions or when all deletions from the current network have been exhausted, two addition moves are made in which a randomly selected species that was previously deleted is added back to the network. Additions are allowed only if they do not increase the value of the objective function. (b) After a sequence of 1000 moves in which the size of the smallest network reached has not decreased, 50 addition moves are performed. Optimisation then continues from (1). The purpose of both types of node addition, which undoes the effects of past moves, is to prevent the procedure from being trapped in a local minimum. We found that the larger moves of type (b) were required to prevent most optimisation runs from becoming trapped in high-lying local minima.

We varied both parameters and procedures of this optimisation algorithm, but found that the above recipe produced the smallest reduced networks for a given value of the objective function threshold, with the smallest spread in the size of the smallest network found from different optimisation runs. For example, networks with 44 nodes that satisfied an error tolerance of 10% were found in three of 16 optimisation runs, each consisting of about 10^6 attempted moves. The range in the size of the smallest network found in these 16 runs was 44–49. Similarly, four of 16 optimisation runs with an error tolerance of 1% found reduced networks with 83 nodes, and the range in the size of the smallest network found was 83–90. Reduced models are available from the authors upon request.