Internal coarse-graining of molecular systems

Jérôme Feret[a], Vincent Danos[b], Jean Krivine[a], Russ Harmer[c], and Walter Fontana[a,1]

[a]Harvard Medical School, Boston, MA 02115; [b]University of Edinburgh, Edinburgh EH8 9Y, United Kingdom; and [c]Centre National de la Recherche Scientifique and Université Paris Diderot, 75006 Paris, France

Edited by John Ross, Stanford University, Stanford, CA, and approved February 20, 2009 (received for review October 3, 2008)

Modelers of molecular signaling networks must cope with the combinatorial explosion of protein states generated by posttranslational modifications and complex formation. Rule-based models provide a powerful alternative to approaches that require explicit enumeration of all possible molecular species of a system. Such models consist of formal rules stipulating the (partial) contexts wherein specific protein–protein interactions occur. These contexts specify molecular patterns that are usually less detailed than molecular species. Yet, the execution of rule-based dynamics requires stochastic simulation, which can be very costly. It thus appears desirable to convert a rule-based model into a reduced system of differential equations by exploiting the granularity at which rules specify interactions. We present a formal (and automated) method for constructing a coarse-grained and self-consistent dynamical system aimed at molecular patterns that are distinguishable by the dynamics of the original system as posited by the rules. The method is formally sound and never requires the execution of the rule-based model. The coarse-grained variables do not depend on the values of the rate constants appearing in the rules, and typically form a system of greatly reduced dimension that can be amenable to numerical integration and further model reduction techniques.

protein interaction networks | rule-based models | model reduction | distinguishability | information carriers

Molecular biology is spectacularly successful in disassembling cellular systems and anchoring cell-biological behaviors of staggering complexity in chemistry. This raises the challenge of reconstituting molecular systems formally, in pursuit of principles that would make their behavior more intelligible and their control more deliberate. This pursuit is as much driven by the practical need to cure disease as it reflects a desire for a theoretical perspective needed to understand the complexity of cellular phenotypes. In achieving such a perspective, we must deal with two broad problems.

First, we must be able to represent and analyze molecular interaction systems of combinatorial complexity. Although ubiquitous, such systems are perhaps most notorious in the context of cellular signaling. The posttranslational modification of proteins and their noncovalent association into transient complexes generate an astronomical number of possible molecular species that can relay signals (1). The question then becomes how to reason about system dynamics if we cannot possibly consider a differential equation for each chemical species that can appear in a system.

Second, understanding systems requires resisting the temptation of adopting the view of an outside observer. The outside view is indeed appropriate for the chemical analysis of a network, since the experimenter deliberately interacts in specific ways with the network to create measurable distinctions. Yet, the network, as a dynamical system, may not be capable of making these same distinctions. For example, an experimental technique might differentiate between SOS recruited to the membrane via GRB2 bound to SHC bound to the EGF receptor and SOS recruited via GRB2 bound to the EGF receptor directly. However, from the perspective of the EGF signaling system, such a difference might not be observable for lack of an endogenous interaction through which it could become consequential. The endogenous units of the dynamics may differ from the exogenous units of the analysis.

In an attempt at mitigating the first problem, analytical model reduction techniques eliminate variables on the basis of algebraic constraints such as conservation equations and quasi-steady-state conditions obtained mainly by exploiting separations of time and/or concentration scales (for example refs. 2 and 3). Numerical model reduction consists in integrating the kinetic rate equations of the full network and subsequently building a reduced model based on species that were observed to be significantly populated (4). Yet, all these techniques hinge on an explicit representation of the full network, which severely curtails their applicability to larger systems.

The past few years have seen the emergence of several approaches (5–8) that represent signaling systems in terms of rules stipulating conditions for specific interactions among proteins. These conditions typically specify (far) less than the full state of all proteins involved in an interaction. In this way, rules capture combinatorial complexity but avoid an explicit representation of the complete reaction network involving all possible molecular species. Yet, to explore the dynamics of a system of rules, such approaches must resort to stochastic simulations (6, 9, 10), whose event-based nature exacts a high computational cost. Ordinary differential equations (ODEs) would be highly useful for rapidly exploring system dynamics by numerical integration, but a flat-out expansion of rules into ODEs would, of course, fall victim to the combinatorial explosion. To nonetheless assemble ODEs from rules, a coarse-graining approach has been recently proposed (11–16). The idea is to convert a rule-based model into a reduced system of rate equations by identifying molecular patterns (sets of species) that act "independently" (16). We believe this approach to be promising, because it seems natural that a system described by rules might be characterized by dynamical units that are less specific than molecular species. We proceed in the same spirit, but differ significantly by seeking as variables those molecular patterns that establish the finest level of resolution at which the dynamics of the system is capable of making distinctions, thus rendering finer-grained patterns unwarranted. This we call internal coarse-graining. Moreover, our approach is formal, avoiding the limitations listed in ref. 16.

The next section surveys the language, Kappa (17), in which we cast rules of interaction. Kappa forms the basis of a substantive, formal, yet intuitive modeling framework (7, 9, 18, 19). Access to the Kappa modeling platform is provided at www.cellucidate.com.

Kappa: A Language for Molecular Biology

Kappa (17) is a formal language for defining agents (typically meant to represent proteins) as sets of sites that constitute

Author contributions: J.F. and V.D. designed research; J.F., V.D., and W.F. performed research; J.F., V.D., J.K., and R.H. contributed new reagents/analytic tools; J.F., V.D., J.K., R.H., and W.F. analyzed data; and J.F., V.D., and W.F. wrote the paper.

Conflict of interest statement: W.F. is a member of the Board of Directors of Plectix BioSystems Inc., a company that develops the Kappa modeling platform used in this research. J.F., V.D., J.K., and R.H. are consultants for Plectix BioSystems Inc.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: walter_fontana@hms.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0809908106/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0809908106

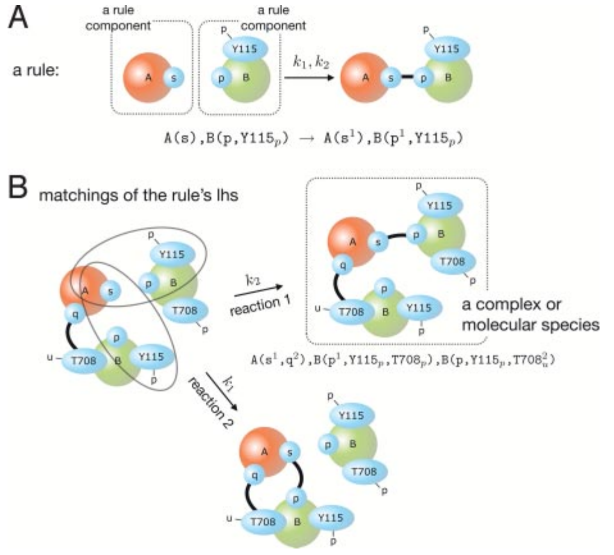

A
a rule component
a rule component
a rule:
k_1, k_2
$\mathtt{A(s),B(p,Y115_p)} \rightarrow \mathtt{A(s^1),B(p^1,Y115_p)}$
B matchings of the rule's lhs
k_2
reaction 1
a complex or molecular species
$\mathtt{A(s^1,q^2),B(p^1,Y115_p,T708_p),B(p,Y115_p,T708_u^2)}$
k_1
reaction 2

Fig. 1. Rules and reactions in Kappa. (*A*) A rule captures a high-level mechanistic statement (empirical or hypothetical) about a protein–protein interaction in terms of a rewrite directive plus rate constant(s). The left-hand side (lhs) of the rule is a pattern of partially specified agents and represents the contextual information necessary for identifying reaction instances that proceed according to the rule. The right-hand side (rhs) expresses the actions that may occur when the conditions specified on the lhs are met in a reaction mixture of Kappa agents. A maximal connected subgraph on the lhs of a rule is called a rule component. (*B*) The rule in *A* matches a combination of agents in 2 distinct ways giving rise to 2 possible reactions with different outcomes. Note that because of their local nature, Kappa rules with >1 lhs component may apply in both a unimolecular and bimolecular situation. This is why such rules are given 2 rate constants, a first-order (k_1) and a second-order (k_2) constant. In a textual representation, agents are names followed by an interface of sites delimited by parentheses. Bonds are labeled by superscripts and internal states at a site by subscripts. In the graphical rendition, internal states are indicated as labeled barbs. See *SI Appendix*, section 1 and the section *Kappa: A Language for Molecular Biology* for more details.

abstract resources for interaction, as illustrated in Fig. 1 and extensively detailed in section 1 of supporting information (SI) *Appendix*. Sites can hold an internal state, as generated through posttranslational modifications, and engage in binding relations with sites of other agents. An association of proteins is a connected (site) graph, called a complex (of agents), as shown in the box of Fig. 1*B*. The nodes of the graph are agents, but the endpoints of edges are sites, which belong to agents. Although an agent can bear many connections, a site can bear only 1.

Kappa is used to express tunable rules of interaction between proteins characterized by discrete modification and binding states. The idea of a rule, Fig. 1*A*, is to stipulate only the molecular context required for an interaction along with some rate constant(s). The left-hand side (lhs) of a rule is any site graph. Agents may mention a subset of their sites and omit states (*SI Appendix*, section 1.2). The right-hand side (rhs) exhibits the changes that occur when the lhs is matched (*SI Appendix*, section 1.4) in a mixture of agents. The difference between rhs and lhs is called the action of the rule. Sites mentioned on the lhs are said to be tested by the rule. Sites that are tested but not modified constitute the context of a rule's action. Because rules typically do not mention all of the sites and states of an agent, they keep combinatorial complexity implicit, obviating the need for eliminating it. A molecular species is a complex in which each agent occurs with a complete set of sites in definite states. We also refer to molecular species as ground-level objects. The complete set of sites defines the finest grain of resolution at which the state of an agent is known. Like rules, this set of sites can be updated to reflect new knowledge or hypotheses. Rules give rise to potentially numerous reaction instances [whose rate constants are related to the rate constant(s) of the rule]. These instances involve particular combinations of molecular species, each of which satisfies the context required for the rule to apply, see Fig. 1*B* and Fig. S4 in *SI Appendix*.

Kappa rules are both descriptions of mechanistic knowledge and executable instructions. In fact, we view Kappa as a programming language attuned to molecular signaling. Rules induce a stochastic dynamics on a mixture of agents, for which we implemented a general and efficient implicit-state version of the Doob–Gillespie algorithm (9). A Kappa model of a biological system is a concurrent computer program whose instructions are rules that asynchronously change the state of a shared store representing the reaction mixture on which the rules act. Computer programs are formal objects that can be analyzed statically. Static analysis assists in the discovery of behavioral properties of a program without running it, much like a system of differential equations can be analyzed without simulating it. Static analysis involves, for example, the inspection of causal dependencies among rules and an overapproximation of the molecular species reachable from an initial condition.

Kappa is closely related to BNGL (5), but differs from the latter in being a context-free grammar, that is, a language that expresses strictly local rules of action. The computational cost of checking whether a rule can apply to a given choice of reactants is bounded by the size of the rule's lhs and not by the reactants. This difference enables scalable simulation (9) and static analysis of the implied dynamical system (7), which plays a crucial role in the efficiency of the coarse-graining technique we describe here (see *Remarks* below and *SI Appendix*). The central role we attach to static analysis sets our framework apart from other rule-based approaches, such as BNGL (5) and "little b" (8), whose primary deliverable is the automated assembly of the full reaction network by generating all possible species and their reactions from a given set of rules. Yet, the combinatorial explosion inherent in molecular signaling makes such goals impractical and often impossible. In a pilot study of EGF signaling, we collated 71 rules representing mechanistic observations of pertinent protein–protein interactions. These rules would produce 10^{19} molecular species. Our current EGF model has grown to ≈350 rules. It thus appears more useful to forgo the expansion into an inscrutably large system of equations and, instead, apply static analysis techniques directly to the rule collection and explore the system with stochastic simulations that generate dynamical trajectories (6, 9, 10). Yet, such simulations are computationally expensive. This raises the question whether there is a system of ODEs that "corresponds" to a rule-based model, i.e. that constitutes its natural differential semantics.

From Rules to ODEs

Using a rule-based (as opposed to a reaction-based) model amounts to acknowledging that molecular species may not always be meaningful units of the dynamics. Such units should lump together species that cannot be distinguished by the dynamics arising from a given system of rules (see section 4, especially 4.2, of the *SI Appendix*). Moreover, the lumping must be self-consistent, meaning that the contribution of each rule to the rate of production or consumption of any unit should only depend on other units. In the following, we introduce 2 key properties that a suitable set of coarse-grained dynamical units—referred to as fragments (to be properly defined later)—should satisfy.

Property 1 ("No Overlap"). No fragment properly overlaps a lhs component of a rule on a modified site. This property is defining of fragments and is key (but not enough) for expressing the rate function of a fragment in terms of fragments. The reasoning is illustrated in Fig. 2. The rule *r* at the top consumes those species that match its lhs component r_{lhs}. We can think of a pattern X in terms of its extension $X^{\diamond}$, which is the set of species that match X, accounting for the many ways in which any such species might

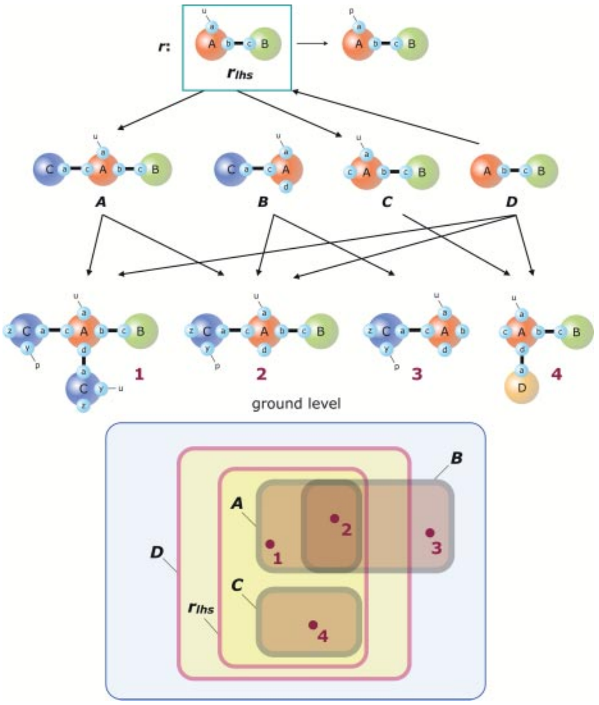


Fig. 2. Rules and fragments. The figure provides assistance in establishing criteria that define fragments, as detailed in the section *From Rules to ODEs*. The top row depicts a (unimolecular) rule whose lhs component is r_{lhs}. The third row from top shows fully specified molecular species (ground-level objects), numbered 1 to 4. The second row depicts various patterns, *A* to *D*. Arrows indicate embedding relations of one pattern (graph) into another (see *SI Appendix*, section 1.4). The rectangles at the bottom provide a schematic of relationships between sets of molecular species that match the patterns *A*–*D* and r_{lhs}. Note that *D* embeds into r_{lhs}; its matching instances are therefore a superset of those of r_{lhs}. Also, *D* does not overlap with r_{lhs} on a site that *r* modifies. Hence *r* has no effect on *D*.

match X (think symmetries). The extension $r_{\mathrm{lhs}}^{\diamond}$ of r_{lhs} is shown schematically at the bottom of Fig. 2 as a yellow area within the blue area standing for the set of all molecular species implied by the rules of a system and an initial condition. Fig. 2 provides assistance for reasoning about the suitability of a few sample patterns as potential fragments in light of Property 1. Consider pattern B. Although B does not itself match r_{lhs}, some ground-level instances of B do, such as species 2. Thus, $B^{\diamond}$ (properly) intersects $r_{\mathrm{lhs}}^{\diamond}$, which makes it impossible to express the contribution of the unimolecular rule r to the consumption rate of B in terms of B alone. Rather, we would have to know at any time the fraction of molecular species that occurs in the intersection of $B^{\diamond}$ with $r_{\mathrm{lhs}}^{\diamond}$, which is a property that requires knowing the complete reaction mixture at any time. By contrast, $A^{\diamond}$ is entirely contained within $r_{\mathrm{lhs}}^{\diamond}$. As a consequence, the firing of rule r will consume the pattern A at a rate proportional to its concentration $[A]$, defined at $t = 0$ by the number of embeddings of A in the reaction mixture. There is no need to know the reaction mixture for any subsequent time. The case of C is analogous to that of A.

It is possible to refine B into B' by adding context, such that $B'^{\diamond} \subset r_{\mathrm{lhs}}^{\diamond}$. For example, connecting agent $\texttt{A}$ at site $\texttt{b}$ to agent $\texttt{B}$ at $\texttt{c}$ yields $B' \equiv \texttt{C}\ (\texttt{a}^1),\ \texttt{A}\ (\texttt{a}_u,\ \texttt{c}^1,\ \texttt{d},\ \texttt{b}^2),\ \texttt{B}\ (\texttt{c}^2)$ with $B'^{\diamond} = B^{\diamond} \cap A^{\diamond}$. Thus, as far as rule r is concerned, patterns A, C, and B' are fragment candidates by virtue of their extensions being inside $r_{\mathrm{lhs}}^{\diamond}$. However, other rules in the system may further constrain these potential fragments. Indeed, our procedure to construct fragments depends on all rules of a given system.

Property 2 ("Orthogonality"). Fragments must partition (in the extension sense) anything that is contained within a fragment, which we refer to as a subfragment. We show later that any lhs component of a rule is a subfragment (Property 1 clarifies this only for particular components). The rate equation for a fragment affected by a rule of molecularity >1 (i.e. a rule with 2 or more lhs components) gets a contribution consisting of a monomial involving several fragments. Consider, for example, a rule of type $Z, Z' \rightarrow Z^*, Z'$, which modifies the lhs component Z into Z^*. Consider further a particular fragment $\mathcal{A}$ that is a refinement of Z and is thus consumed by the rule ($\mathcal{A}^{\diamond} \subset Z^{\diamond}$). The consumption rate of $\mathcal{A}$ will be proportional to $[\mathcal{A}]\ [Z']$. If only 1 fragment, say $\mathcal{B}$, matches the lhs component Z', then $[Z'] = [\mathcal{B}]$. However, there may be several fragments $\mathcal{B}_i$ that match Z', in which case $[Z']$ should be the sum over all $[\mathcal{B}_i]$. The only problem is that the $\mathcal{B}_i$ might have ground-level extensions $\mathcal{B}_i^{\diamond}$ that overlap, causing the naive sum to overcount. Thus, there must be a set of fragments that partitions $Z'^{\diamond}$, so that $[Z']$ can be expressed as a sum of orthogonal fragments. Property 2 does more, however: It guarantees that the concentration of any subfragment can be expressed in terms of fragment concentrations. This will be needed down the road. Properties 1 and 2 jointly ensure a self-consistent coarse-grained system whose dynamics is sound. Soundness means that computing the ground-level dynamics and then coarse-graining yields the same result as coarse-graining at the outset and then running the coarse-grained dynamics.

Note that the (possibly infinite) set of molecular species is always a trivial set of fragments enjoying Properties 1 and 2, but typically far from optimizing our criterion of "dynamical distinguishability." We can do much better without ever touching the ineffable ground-level network of species. As we show next, by proceeding directly from the rules, we construct dynamical units whose boundaries are carved out by the actions available to the system.

Constructing Coarse-Grained Fragments

In this section we implement Properties 1 and 2 by defining syntactical criteria with which we scan all rules in a model to determine which agents and sites belong to a fragment. As a test case, we apply these criteria to a rule-based model of a small section of early events in epidermal growth factor (EGF) signaling as adapted from ref. 20. These events include the binding of EGF (agent E) to the receptor (R), the subsequent dimerization of the receptor, and the eventual recruitment of SOS (O). The model consists of 39 rules r01–r39, listed in section 5.1 of *SI Appendix*. We write separate rules for binding and unbinding actions, because unbinding typically occurs under less-restrictive contexts than binding. The names of agent sites were chosen fairly arbitrarily. The biological accuracy of the published models from which we obtained the rules might be outdated, because knowledge about EGF signaling mechanisms keeps changing rapidly. Our goal here is not a particular biological insight, but a procedure of general interest. Together, the 39 rules of our test case imply 356 possible distinct molecular species. We shall see, however, that based on these rules of interaction, the system can only make 38 internal distinctions. Differential equations in these 38 variables self-consistently describe the dynamics of the system. It is very convenient to use a special map as a canvas for laying out which sites and bindings must appear together in a fragment. In ref. 7, we called this map the contact map (CM), Fig. 3*A*. The CM is generated automatically from a rule-based model and provides a summary of attainable interactions. The CM is a graph whose nodes are the agents that appear in the model. Recall that agents are sets of sites. These sites are the endpoints of edges representing possible binding interactions. Certain sites are colored to indicate that their internal state can be modified.

Syntactical Criteria for Annotating the Contact Map. We shall need the notion of a parsimonious covering, or covering for short. A covering $\mathcal{C}$ of a set $\mathcal{S}$ is a set of subsets of $\mathcal{S}$, called classes, such that (*i*) no class is empty, (*ii*) no class is a subset of another class, and (*iii*)

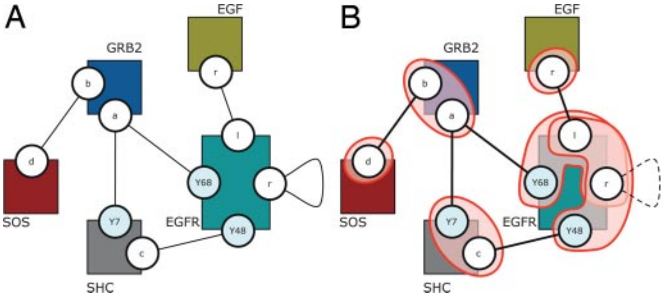


Fig. 3. The contact map. (*A*) The contact map is a graph whose nodes are the agents in the model and whose edges are possible bonds between sites. Filled circles indicate sites with modifications of state. The contact map is a fine-grained version of what is known as a protein–protein interaction (PPI) map, in that its edges end in sites of agents and not just agents. (*B*) The annotated contact map (ACM) after decoration induced by the directives Cov1–Cov3 and Edg1.

the union of all classes yields $\mathcal{S}$. A covering differs from a partition in that the elements of a covering need not be pairwise disjoint.

In preparation for building fragments, we first annotate the CM with 2 types of information obtained by applying the syntactical criteria listed below. (*i*) For each agent type A, we define a covering $\mathcal{C}$(A) of the set of its sites. (*ii*) For each edge in the CM, we define its type as either "solid" or "soft." In a second step, we assemble fragments based on the annotated CM (ACM).

The following syntactical criteria determine valid coverings for an agent and the type of a bond. We follow up with some explanatory remarks.

Cov1 (backward closure). If a rule tests a site a in an agent A and modifies a site b in the same agent, any class in $\mathcal{C}$(A) that contains b must also contain a (e.g. Fig. 4 *A* and *B*).

Cov2 (relay). If a rule tests a site a in an agent A, and A is connected by some path through a site b to an agent that is modified, any class in $\mathcal{C}$(A) that contains b must also contain a (e.g. Fig. 4*C*).

Cov3 (witness). For each agent in an unmodified lhs component, there must be a class in the agent's covering that contains all of the sites tested by the rule.

Edg1. A bond is solid if it occurs on the lhs of a rule that tests anything other than that bond.

Syntactical criteria Cov1–Cov3 and Edg1 implement Properties 1 and 2. To see this, define an overlap between 2 patterns *X* and *Y* as the set of agents and sites both mention along with a mutually compatible state. The overlap, if it exists, can be used as an instruction for gluing the patterns together, see section 2 of *SI Appendix*. Our discussion of Fig. 2 suggests that if a pattern has an overlap with a component on the lhs of a rule, and the overlap contains a site modified by the action of the rule, the pattern must be glued to the lhs component to become a fragment as far as that rule is concerned. Hence, a fragment $\mathcal{A}$ either has no overlap with the sites that are modified by the rule, or it contains a whole lhs component (*SI Appendix*, section 2). The same process—glue on overlap—is repeated for each rule and all agents, starting out with each site in its own class. Cov1 and Cov2 simply keep track of which sites of an agent must be mentioned together in a fragment as a result of this repeated glue-on-overlap. Cov3 takes care of the orthogonality property in the special case of a component required for an action but not modified by it (a "witness").

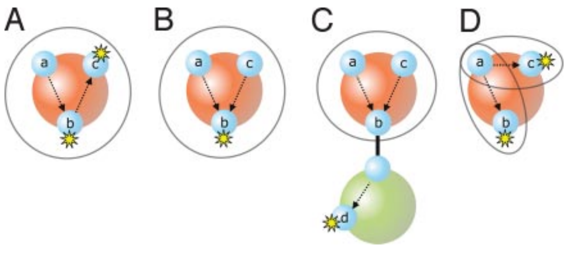


Fig. 4. Examples illustrating the syntactical criteria Cov1 and Cov2 for determining classes in the covering of an agent. See section 3 of the *SI Appendix* for further details.

The glue-on-overlap process can pull bonds into a fragment (see *B′* in the discussion of Property 1). However, not all bonds are conduits of control between the parts, say *X* and *Y*, they connect. Suppose that the only time a bond appears on the lhs of a rule is in a so-called pure dissociation rule that tests nothing except the existence of the bond that is to be broken. No rule modifying *X* or *Y* depends on that bond (or the bond would figure in a rule other than the dissociation rule). As a consequence, the fragments containing *X* can stop short of including all possible states of *Y* and vice versa. The fragments containing *X* only need to specify whether or not *X* is connected to *Y*, but they do not need to specify *Y* itself. (And vice versa.) The directive Edg1 defines those bonds that carry constraints as solid. All bonds not characterized by Edg1 can be chosen as solid or soft, and we can choose to have smaller or larger covering classes (provided they satisfy Cov1–Cov3); the fragmentation is sound either way. However, soft bonds make for smaller fragments (see next section). Our policy is to obtain small fragments by choosing covering classes that are as small as possible and considering bonds to be soft when they appear only on the rhs of a rule (bonds that are only formed) and/or on the lhs of a pure dissociation.

Fragment Assembly. To define fragments, it is convenient to extend the notion of complex with bond stubs. An agent with a bond stub is written A ($\mathtt{a}^{\mathtt{B@b}}$), which means that A's site a is bound to B's b, without, however, including agent B in the complex.

Given an ACM, a fragment $\mathcal{F}$ is a complex such that: Each agent has a set of sites that is a class, every site has an internal state if any, every site has a binding state—either free, bound, or stubbed, every stub must correspond to a soft bond in the ACM, and every bond is solid. A subfragment is a complex that embeds in a fragment.

To obtain a fragment, one starts with an agent and a site. The ACM then determines which further sites to add and which binding states (stubbed or not) are appropriate. When there is nothing more to add, one has a fragment.

As an example of this growth process, consider agent R in our rule set. According to the ACM in Fig. 3*B*, we have a choice between 2 classes. Suppose we choose class {l, r, Y48}. Next, we assign a state to each site in that class. For example, all sites are free, and Y48 is unphosphorylated. This yields fragment R ($\mathtt{Y48}_u$, l, r), which is $\mathcal{F}_{34}$ in the complete list for our example (*SI Appendix*, section 2.3). Alternatively, we might choose Y48 to be phosophorylated (fragment $\mathcal{F}_{15}$). Yet, if we choose Y48 to be also bound, then the solid link in the ACM forces agent S into the fragment, along with its site c as the link's endpoint. In turn, c forces inclusion of the class to which it belongs, {c, Y7}. Now we need to assign states to c and Y7 in agent S. For example, S ($\mathtt{Y7}_p$, $\mathtt{c}^1$), R ($\mathtt{Y48}_p^1$, l, r), which is fragment $\mathcal{F}_{04}$. A further fragment is obtained by considering site r in agent R to be bound. Site r can bind to another R agent, but the link is soft. A soft link at r does not force the inclusion of another instance of R. Instead, the bound state is only indicated with its type: S ($\mathtt{Y7}_p$, $\mathtt{c}^1$), R ($\mathtt{Y48}_p^1$, l, $\mathtt{r}^{\mathtt{R@r}}$). This fragment, however, does not show up in our list. Given our set of rules, the state in which R is dimerized at site r cannot occur if the ligand-binding site l is empty. Such a fragment is automatically eliminated from the list because a separate reachable state analysis (next section) recognizes it as inaccessible. Fragments as defined above enjoy the following properties:

Q1. No fragment strictly overlaps with a rule component on a modified site.

Q2. Any lhs component is contained in a fragment (i.e., is a subfragment).

Q3. The concentration of any subfragment can be expressed as a linear combination of fragment concentrations (Eq. **17** in *SI Appendix*).

Q4. Fragments are closed under rule actions.

Q1 is Property 1 (no overlap), whereas Q2 and Q3 imply Property 2 (orthogonality). Q4 means that fragments form a network of reactions (like species).

Q1 follows from Cov1 and 2 and Edg1; Q2 follows from Cov3 and Edg1 for nonmodified rule components, and Cov1 and 2 and Edg1 for modified ones; Q3 follows from the exhaustivity of the growth procedure for fragments, as does Q4.

Q1–3 ensure a sound translation from rules into an ODE system for fragments, as sketched next.

Assembling the Dynamical System for Fragments. The dynamical system for fragments is constructed by deriving mass action terms for the consumption and production of fragments from rules. We only sketch the reasoning here and provide a detailed account in section 6.4 of *SI Appendix*. Consider, for example, a rule of the form $Z,Z' \rightarrow Z$—Z', which binds 2 complexes Z and Z'. Based on this rule, the differential equation $d[\mathcal{F}_i]/dt$ for each fragment $\mathcal{F}_i$ that matches Z obtains a consumption term $-\gamma[\mathcal{F}_i][Z']$, where $[Z']$ is expressed as a sum of concentrations of orthogonal fragments using Q2 and 3. The factor γ depends on the rate constant of the rule and the number of ways that Z embeds into $\mathcal{F}_i$. On the production side, the kinetic terms depend on the bond type in the ACM. Consider, for example, a solid bond. A kinetic term $\gamma[\mathcal{F}_i][\mathcal{F}_j]$ is generated for the differential equation $d[\mathcal{F}_k]/dt$ of every fragment $\mathcal{F}_k$ that matches Z—Z', where $\mathcal{F}_i$ and $\mathcal{F}_j$ are fragments matching Z and Z', respectively, subject to the constraint that the match of $\mathcal{F}_k$ is the disjoint sum of the embeddings of Z and Z' into their respective fragments. If the bond in Z—Z' is soft and corresponds to a ⋯ A.a—b.B ⋯, one can replace Z—Z' with $Z^{\mathrm{B@b}}, Z'^{\mathrm{A@a}}$, because there is no information in $Z'^{\mathrm{A@a}}$ affecting $Z^{\mathrm{B@b}}$. Every fragment $\mathcal{F}_k$ matching $Z^{\mathrm{B@b}}$ gains a production term $\gamma[\mathcal{F}_i][Z']$, where $\mathcal{F}_i$ matches Z and is related to the $\mathcal{F}_k$ matching $Z^{\mathrm{B@b}}$. A similar argument applies to fragments that match $Z'^{\mathrm{A@a}}$.

The dissociation of a solid bond Z—Z' will give rise to a piece Z (and also Z') that embeds into a fragment $\mathcal{F}$. To determine the contribution of the dissociation rule to the rate of production of $\mathcal{F}$, we need the concentration of $\mathcal{F}$—Z'. However, $\mathcal{F}$—Z' is not itself a fragment but, rather, a subfragment. This is why, for our method to result in a closed system of equations, we must be able to express the concentration of a subfragment in terms of fragments (see Q3 and Property 2).

Fig. 5 was obtained by running a microscopic stochastic simulation of the early EGF test system, driven by rules r01–r39 while reporting the concentrations of fragments $\mathcal{F}_{01}$–$\mathcal{F}_{38}$. This stands as a proxy for the numerical integration of the deterministic ground-level system of 356 ODEs and the subsequent lumping of species into our 38 fragments. As a comparison, the smooth curves result from the direct numerical integration of the automatically generated ODE system for fragments.

Remarks

Reachability. Underlying several steps of our procedure is a very fast overapproximation α of the set of reachable species, deploying the framework of abstract interpretation (21) as described in ref. 19. This overapproximation comes into play at 3 junctures (*i*) The contact map reports edges and site modifications only if they are reachable by α. (*ii*) Fragments that are not reachable by α are discarded. (*iii*) The procedure for compressing rules (see below) makes use of α. In ref. 19, we characterize those special situations for which α is exact. (The present EGF example is such a case.)

Making Rules Concise. Because fragment construction proceeds by inspecting the structure of rules, it is important that rules be concise, in the sense of avoiding redundant contextual conditions (tests) on their lhs. However, what classifies as redundant depends on the remaining rules in the model. Because rules record empirical observations or hypotheses, they tend to be crafted in isolation. Consider, for the sake of illustration, rule r02 expressing the binding of ligand to receptor: $\texttt{R}\,(\texttt{l}, \texttt{r}), \texttt{E}\,(\texttt{r}) \rightarrow \texttt{R}\,(\texttt{l}^1, \texttt{r}), \texttt{E}\,(\texttt{r}^1)$. The rule mentions 2 sites, l and r, of the receptor R. Site l is the ligand (EGF)-binding site, whose state is modified by the action of r02, whereas r is the site at which the receptor dimerizes (as described in r03). Rule r02 asserts that binding of E (EGF) to R requires not just a free l, but also a free r. Given the other rules of the model, there is no reachable state of the reaction mixture in which R could dimerize before binding E. Hence, in the context of the remaining 38 rules of this model, asking for site r to be free is a redundant condition for the firing of rule r02, because a free l implies a free r. Without removing such redundancies, fragments would be more numerous and bloated by fictitious dependencies. To reduce the extent to which this happens, we preprocess a rule system with an automatic compression that removes unnecessary contextual specifications. This technique rests on the reachability overapproximation referred to in the previous paragraph. In section 5.2 of the *SI Appendix*, we list the 39 compressed rules cr01–cr39 from which the 38 fragments were derived.

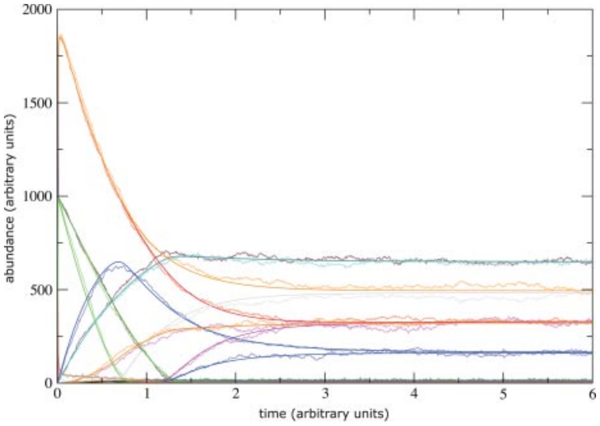


Fig. 5. Comparison between microscopic dynamics and fragment dynamics. Wiggly curves: The microscopic dynamics of the early EGFR example is executed with a Doob–Gillespie simulation (9) while reporting the coarse-grained fragment concentrations. This serves as a proxy for the deterministic microscopic dynamics. Steady curves: The output of the deterministic fragment dynamics. Still, many fragments (and many more molecular species) only acquire tiny concentration values, causing far fewer than 38 curves to be discernible by eye in this plot.

Role of Rate Constants. All ground-level reactions into which a rule expands inherit its rate constant (after accounting for possible symmetry reductions upon expansion). Beyond any specific values of rate constants, rules themselves already imply a notion of kinetic distinguishability. For example, our toy model of early EGF events posits that the phosphorylated EGFR receptor (R) binds the protein SHC (S), which would read as $\texttt{R}\,(\texttt{Y48}_p), \texttt{S}\,(\texttt{c}) \rightarrow \texttt{R}\,(\texttt{Y48}^1_p), \texttt{S}\,(\texttt{c}^1)$. Yet, such a rule does not appear in the model. Rather, the same binding action between R and S is found in 2 rules r24 and r28 that differ in their contexts. Rule r24, $\texttt{R}\,(\texttt{Y48}_p), \texttt{S}\,(\texttt{c}, \texttt{Y7}_u) \rightarrow \texttt{R}\,(\texttt{Y48}^1_p), \texttt{S}\,(\texttt{c}^1, \texttt{Y7}_u)$, specifies that site Y7 of S must be unphosphorylated and free, whereas rule r28, $\texttt{R}\,(\texttt{Y48}_p), \texttt{S}\,(\texttt{c}, \texttt{Y7}^1_p), \texttt{G}\,(\texttt{a}^1, \texttt{b}) \rightarrow \texttt{R}\,(\texttt{Y48}^2_p), \texttt{S}\,(\texttt{c}^2, \texttt{Y7}^1_p), \texttt{G}\,(\texttt{a}^1, \texttt{b})$, specifies that Y7 of S is phosphorylated and bound to G. The only reason to warrant such a distinction is an actual or hypothesized difference in the rate constants for the 2 contexts. Hence, regardless of the specific values of rate constants, positing 2 rules with different contexts for the same action affects the construction of fragments. The precise values of the rate constants of rules enter the ODEs for fragments, but they do not affect the fragments themselves, because the latter are based on the distinguishability of control flows shaped by rule contexts.

Limitations. For our coarse-graining procedure to be well defined, rules must have unique ground-level molecularity, i.e., a rule with 2 lhs components must apply only to disjoint reactants (unlike in Fig. 1). Rules whose arity does not always match the arity of their ground-level instances (molecularity mismatches) can give rise to polymerization and result in an infinite number of fragments.

Multiple occurrences of the same agent in a rule do not constitute a problem; neither does the production of an agent. The destruction of an agent poses no theoretical problem either but is costly in terms of fragment numbers—as is the BNGL "." (dot) operator.

We do not claim that our method generates a smallest set of fragments or that it is unique. In particular, our method carries a deliberate bias by defining fragments as connected patterns. As a consequence of our construction via an annotated contact map, fragments are closed under the operational semantics of Kappa, i.e., rules convert fragments into fragments (Q4). This allows us to conveniently picture a reaction network at the level of fragments. However, this is not necessary for sound coarse-graining, and alternatives remain to be explored.

We are mathematically certain that any information lost by our coarse-graining is not distinguishable by the microscopic dynamics. However, we cannot prove that all information retained in our fragments is distinguishable. One reason is that rule compression (see above) is, in general, an approximation.

Prior Art. Our method differs from prior approaches in several aspects. First, our method is formal, which makes its properties more transparent and amenable to proof. It suffers from none of the limitations listed in ref. 16, as far as deterministic dynamics is concerned. Second, our approach focuses on interaction-based distinguishability rather than "independence." In section 4 of *SI Appendix*, we provide some thoughts on independence and distinguishability that are conceptually useful for appreciating our stance but not needed for grasping our method. The similarity between the approach sketched in ref. 16 and our present work ends at directive Cov1, because control flows across bindings are treated differently. In section 8 of *SI Appendix*, we compare the outcome of our method with the manual procedure described in ref. 14.

Conclusions

Rule-based representations have been recently proposed to address the dynamics of combinatorial systems for which an expansion into the full reaction network is virtually impossible (5–7). It would be highly useful to construct a deterministic projection of rule-based dynamics for several reasons. On the practical side, rule-based models require stochastic simulations, which can be very time consuming. Although stochastic kinetics can provide insights not accessible from deterministic rate equations, the latter are useful for calibration, analysis, and judicious simplification. On the conceptual side, many of the molecular species that are, in principle, attainable by a given system seem unlikely to play a significant dynamical role, because they either are too improbable, or the dynamics of the system cannot differentiate them. The latter is already implicit in the use of rules, which specify patterns of interactions, rather than reactions between fully detailed molecular species.

We have presented a formal method for automatically generating a dynamical system of coarse-grained variables from a given set of rules, guided by a criterion of distinguishability. The method is exact in the sense that coarse-graining first and then integrating the fragment ODEs is equivalent to first integrating the network ODEs at the level of molecular species and then coarse-graining. The fact that the ground system is oftentimes ineffable because of combinatorial blow-up is of no consequence, because these patterns are constructed directly from the rules.

Our running test case was a limited model of early events in EGFR signaling (21), consisting of 39 rules that generate 356 molecular species. Our method yielded a dynamical system of 38 fragments. A pilot study on a larger section of the EGFR system (19), comprising 71 rules potentially expanding into 18,051,984,143,555,729,567 molecular species, yields 175,988 fragments, which reconnects the system to the realm of feasible ODEs.

In particular cases, fragments become independent units. (A necessary condition being that the coverings of all agents are partitions.) We call such systems "tileable." In section 4.1 of *SI Appendix*, we provide a connection between tileability and invertibility. Although exact, our coarse-graining is not invertible, in general.

It might be biologically insightful to attempt a sensitivity analysis of the fragmentation process, to determine which rules, when changed, have the biggest impact on the nature and number of fragments. Can highly consequential rules be guessed from the annotated contact map? Issues like these suggest that internal coarse-graining is not only of practical use but of theoretical import for understanding the informational architecture of molecular signaling systems.

1. Hlavacek WS, *et al.* (2006) Rules for modeling signal-transduction systems. *Science STKE* 344:re6.
2. Krüger R, Heinrich R. (2004) Model reduction and analysis of robustness for the Wnt/β-catenin signal transduction pathway. *Genome Inform* 15:138–148.
3. Ciliberto A, Capuani F, Tyson JJ (2007) Modeling networks of coupled enzymatic reactions using the total quasi-steady state approximation. *PLoS Comput Biol* 3:e45.
4. Faeder JR, Blinov ML, Goldstein B, Hlavacek WS (2005) Combinatorial complexity and dynamical restriction of network flows in signal transduction. *IEE Syst Biol* 2:5–15.
5. Blinov ML, Faeder JR, Hlavacek WS (2004) BioNetGen: Software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics* 20:3289–3292.
6. Lok L, Brent R (2005) Automatic generation of cellular reaction networks with Moleculizer 1.0. *Nat Biotechnol* 23:131–136.
7. Danos V, Feret J, Fontana W, Harmer R, Krivine J (2007) Rule-based modelling of cellular signalling. *Lecture Notes in Computer Science* (Springer, Lisboa, Portugal), Vol 4703, pp 17–41.
8. Mallavarapu A, Thomson M, Ullian B, Gunawardena J (2008) Programming with models: Modularity and abstraction provide powerful capabilities for systems biology. *J R Soc Interface* 10.1098/rsif.2008.0205.
9. Danos V, Feret J, Fontana W, Krivine J (2007) Scalable simulation of cellular signalling networks. *Lecture Notes in Computer Science* (Springer, Berlin), Vol 4807, pp 139–157.
10. Yang J, Monine MI, Faeder JR, Hlavacek WS (2008) Kinetic Monte Carlo method for rule-based modeling of biochemical networks. *Phys Rev E* 78:031910.
11. Borisov NM, Markevich NI, Hoek JB, Kholodenko BN (2006) Trading the micro-world of combinatorial complexity for the macro-world of protein interaction domains. *BioSystems* 83:152–166.
12. Conzelmann H, Saez-Rodriguez J, Sauter T, Kholodenko BN, Gilles ED (2006) A domain-oriented approach to the reduction of combinatorial complexity in signal transduction networks. *BMC Bioinformatics* 7:34.
13. Koschorreck M, Conzelmann H, Ebert S, Ederer M, Gilles ED (2007) Reduced modeling of signal transduction–a modular approach. *BMC Bioinformatics* 13:336.
14. Conzelmann H, Fey D, Gilles ED (2008) Exact model reduction of combinatorial reaction networks. *BMC Systems Biology* 2:78.
15. Conzelmann H (2008) PhD Thesis (Institut für Systemdynamik der Universität Stuttgart, Stuttgart, Germany).
16. Borisov NM, Chistopolsky AS, Faeder JR, Kholodenko BN (2008) Domain-oriented reduction of rule-based network models. *IET Syst Biol* 2:342–351.
17. Danos V, Laneve C (2004) Formal molecular biology. *Theor Comput Sci* 325: 69–110.
18. Danos V, Feret J, Fontana W, Harmer R, Krivine J (2008) Rule-based modelling, symmetries, refinements. *Lecture Notes in Bioinformatics* (Springer, Cambridge, UK), Vol 5054, pp 103–122.
19. Danos V, Feret J, Fontana W, Krivine J (2008) Abstract interpretation of cellular signalling networks. *Lecture Notes in Computer Science*. (Springer, Berlin), Vol 4905, pp 83–97.
20. Blinov ML, Faeder JR, Goldstein B, Hlavacek WS (2006) A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. *BioSystems* 83:136–151.
21. Cousot P, Cousot R (1977) *Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints* (ACM Press, New York), pp 238–252.

Supporting Information

Jérôme Feret[*], **Vincent Danos**[†], **Jean Krivine**[*] , **Russ Harmer**[‡], **and Walter Fontana**[*]

[*]Harvard Medical School, Boston, USA,[†]University of Edinburgh, Edinburgh, United Kingdom, and [‡]CNRS & Paris Diderot, Paris, France

1 Kappa

Kappa is a formal language for representing molecular objects as agents [1, 2]. These agents are decorated with sites that carry modifiable state and/or bind sites of other agents to form complexes. Empirical observations about the states of proteins that permit them to interact in specific ways are expressed as rules at a level of abstraction consistent with how molecular biologists approach networks of protein-protein interactions. Kappa-rules stand in analogy to reaction rules in organic chemistry, where aspects of molecules that are irrelevant to a chemical rearrangement are designated as "remainder" groups, Figure 1. In Kappa, irrelevant context is simply not mentioned. While chemistry has a theoretical foundation for rationalizing rules of reaction, Kappa-rules only codify observations, not why these observations might make sense to a structural biologist or biochemist. In this section, we provide a formal syntax of the language and an equivalent graphical rendering, Figure 1.

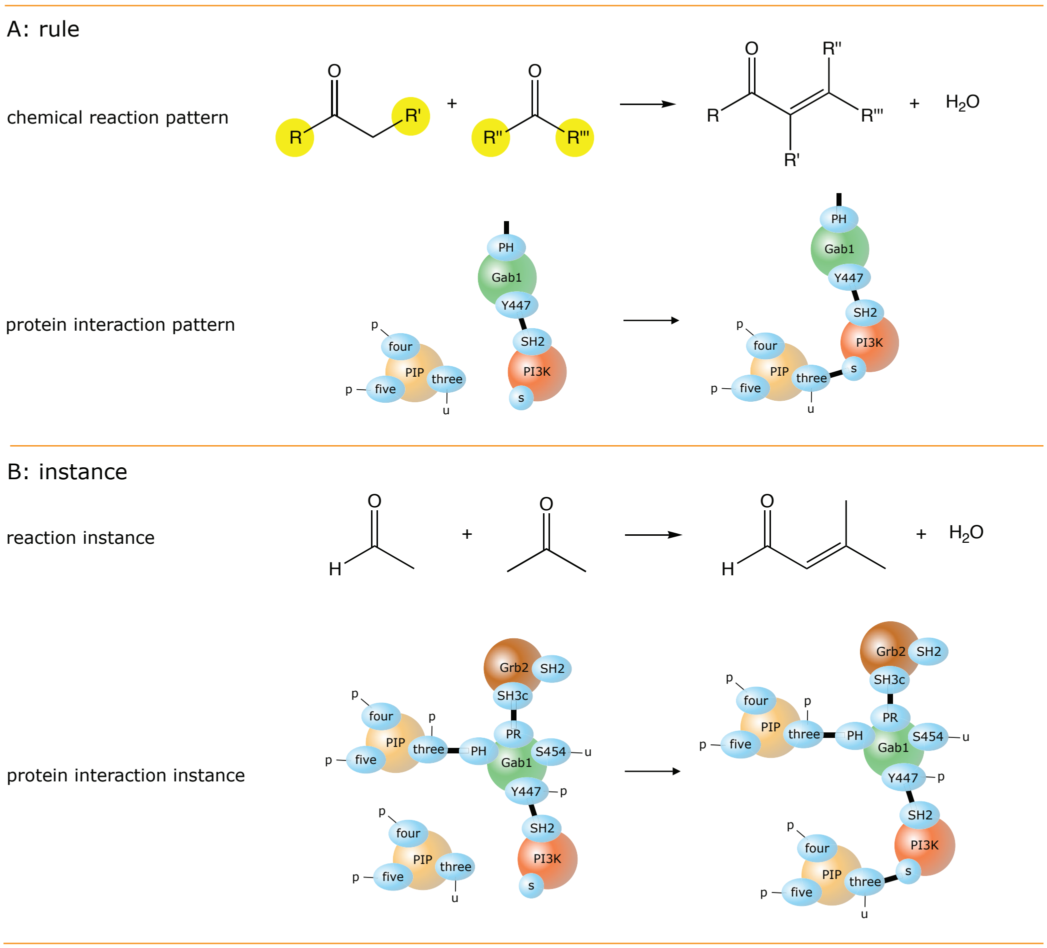


Fig. 1. Rule-based languages: the analogy between chemistry and Kappa. A: An aldol condensation is used as an example to illustrate the concept of a reaction rule in chemistry. The rule details only those molecular parts that are relevant to a particular scheme of reaction, designating unspecified context in terms of remainder groups R, R', R", and R"' (highlighted). Sometimes a generic scheme, such as in (A), requires refinement into special classes defined by different sets of contexts R to R"'. B: Upon full specification of the contexts $R = H$, $R' = H$, $R'' = CH_3$, and $R''' = CH_3$, the rule (A) becomes a reaction instance (B). Kappa proceeds in complete analogy. A rule (A) describes the context required for a local interaction to occur. Panel B shows an instance that complies with the rule depicted in A.

1.1 Agents, complexes, mixtures. We first provide a formal definition of the context-free grammar of Kappa [3, 4], followed by a few explanations for readers unfamiliar with Backus-Naur notation.

In the following, let $\mathcal{A}$ be a set of agent names, $\mathcal{S}$ a set of site names (and let $\wp(\mathcal{S})$ denote the powerset of $\mathcal{S}$), $\mathbb{V}$ a set of internal states, and $\mathbb{N}$ a set of labels. Further, let $\psi : \mathcal{A} \mapsto \wp(\mathcal{S})$ be a map that associates an agent name to a set of sites, called the agent's interface.

Definition 1.1 (Agents)**.**

(i)	agent	a	::=	$N(\sigma)$	
(ii)	agent name	N	::=	$A \in \mathcal{A}$	
(iii)	interface	σ	::=	$\varepsilon \mid s, \sigma$	
(iv)	site	s	::=	n_ι^λ	
(v)	site name	n	::=	$x \in \mathcal{S}$	
(vi)	internal state	ι	::=	ϵ	(*any state*)
				$\mid m \in \mathbb{V}$	
(vii)	binding state	λ	::=	ϵ	(*free*)
				$\mid \; -$	(*semi-link: "bound to something”*)
				$\mid \; ?$	(*unspecified: free or bound*)
				$\mid \; i \in \mathbb{N}$	(*bond label*)

Definition 1.2 (Expressions)**.**

(viii)	expression	E	::=	$\varepsilon \mid a, E$

Definition 1.3 (Well-formedness)**.**

(ix)	unique interface	the sites form a set and each site name in the scope of an agent named A must be in $\psi(A)$
(x)	agent scope	a site name can occur only once in a given interface
(xi)	binary binding	a binding state $i \in \mathbb{N}$ occurs exactly twice, if it occurs at all

Definition 1.4 (Structural equivalence)**.**

(xii)	interface	$E, A(\sigma, s', s, \sigma), E'$	$\equiv$	$E, A(\sigma, s, s', \sigma), E'$	*(site permutation)*
(xiii)	mixture	E, a, a', E'	$\equiv$	E, a', a, E'	*(agent permutation)*
(xiv)	edge labels	$i, j \in \mathbb{N} \wedge i$ not in E	$\Rightarrow$	$E[i/j] \equiv E$	*(relabeling)*

The grammatical rules (i)-(xi) define well-formed expressions in Kappa. We shall define the syntax of Kappa-rules in section 1.3. The notion of "rule-based" models refers to rules expressing actions, not to the grammatical rules defining the terms of the language.

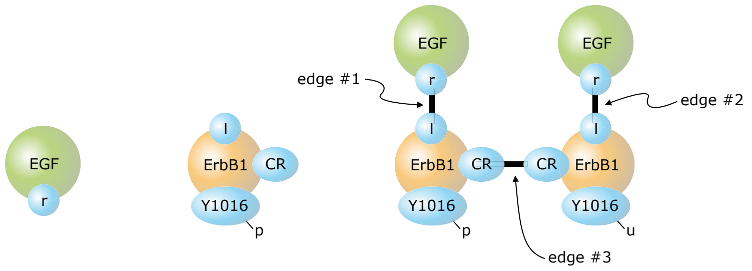
The grammatical rule (i) defines the overall syntax of an agent as consisting of a name N, taken from the set $\mathcal{A}$ (rule ii), and an interface σ. For example, we may call an agent `ErbB1`. Rule (iii) and (ix) define the interface of an agent as a finite set $\sigma = \{s_1, s_2, \ldots, s_n\}$ of *sites*. The vertical bar (|) in (iii) indicates a choice in the recursive application of the grammar when constructing agents. The rule is recursive because σ appears on both sides of the definition: a set of sites consists of a site s and a set of sites. Each time we iterate over (iii), we instantiate a different site s. The s in (iii) refers to the syntactical category "site” defined in (iv). The construction of an interface terminates by choosing the empty interface ε. The sites of an agent control the interactions it participates in. These interactions are defined by Kappa-rules, section 1.3. As indicated in (v), a site s is referred to by an arbitrary name in $\mathcal{S}$, much like an agent. According to (iv), a site carries two types of information, notated as a superscript and subscript to the site name. The subscript ι (iota) of a site refers to its *internal state*, which either assumes some definite value or is left unspecified (ϵ), as declared in (vi). In most biological interpretations, the value of an internal state indicates a post-translational modification, such as "phosphorylated", "unphosphorylated", "methylated". The superscript λ of a site refers to its *binding state*, defined in (vii). Agents may be bound to other agents at sites that belong to them. To indicate that site `l` of agent `ErbB1` is bound to site `r` of agent `EGF`, we deploy the same superscript at both sites. For example, the expression $\texttt{ErbB1(l}^2\texttt{),EGF(r}^2\texttt{)}$ indicates an agent `ErbB1` that is bound to an agent `EGF` at the sites indicated. A superscript uniquely labels a bond between two agents, as laid out in rule (xi). The superscript ϵ means that the site is unbound (free), while a subscript ϵ indicates an unspecified state (like a wild card). We typically do not write the value ϵ. For example, $A(s_\epsilon^\epsilon) \equiv A(s)$.

The object $\texttt{ErbB1(l}^2\texttt{),EGF(r}^2\texttt{)}$ is not itself an agent, because an agent has only one name by virtue of (i). In fact, `ErbB1` bound to agent `EGF` is a *complex*, which belongs to the syntactical category of *expression*, Definition 1.2. In the grammar rule (viii) for forming expressions, the symbol a refers to agents, as defined in (i)-(vii). An expression is simply a set of comma-separated agents. The syntactical category of expression thus includes the notion of a complex. For example, the expression

$$\texttt{EGF(r}^1\texttt{)} \;,\; \texttt{ErbB1(l}^1\texttt{,CR}^3\texttt{,Y1016}_p\texttt{)} \;,\; \texttt{EGF(r}^2\texttt{)} \;,\; \texttt{ErbB1(l}^2\texttt{,CR}^3\texttt{,Y1016}_u\texttt{)} \qquad \textbf{[1]}$$

denotes a complex in which two `ErbB1` agents, each bound to an `EGF` agent, have dimerized on their sites named `CR` (Figure 2).

An agent is an atomic entity, in the sense of not being decomposable into further agents. A complex is a connected graph of agents. (In chemistry, an atom would correspond to an agent in our sense, and a molecule to a complex.) An expression is more general than a complex, since Definition 1.2 does not require any bindings between agents in an expression. Figure 2 illustrates an expression (and a graphical presentation) consisting of an agent `EGF(r)`, an agent $\texttt{ErbB1(l,CR,Y1016}_p\texttt{)}$, and the complex represented in **[1]**. As defined in (viii), an expression is a graph over agents whose connected components are complexes.



$\mathtt{EGF(r)}$, $\mathtt{ErbB1(l,CR,Y1016}_p\mathtt{)}$, $\mathtt{EGF(r}^1\mathtt{)}$, $\mathtt{ErbB1(l}^1\mathtt{,CR}^3\mathtt{,Y1016}_p\mathtt{)}$, $\mathtt{EGF(r}^2\mathtt{)}$, $\mathtt{ErbB1(l}^2\mathtt{,CR}^3\mathtt{,Y1016}_u\mathtt{)}$

Fig. 2. A Kappa expression. The textual representation of a small reaction mixture containing 6 agents that are divided into three complexes (underlined) is shown at the bottom. The two complexes on the left are simple agents, while the complex on the right is made of 4 agents hanging together as shown. An equivalent graphical rendition is depicted above the textual expression, exhibiting the complexes in the same order (left to right) as in the expression below. Names of agents and sites are written inside their corresponding nodes, while internal states of sites, such as the phosphorylated state (p) of Y1016 at ErbB1, are indicated by a labeled barb coming off the corresponding site node.

An agent should be thought of as being associated with a unique interface (by virtue of the mapping ψ). As we shall see later, agents in an expression are oftentimes mentioned with only a subset of their sites. Rule (ix) ensures that these sites are elements of the agent's interface.

We would like an expression to represent the contents of a well-stirred mixture or chemical solution. To formalize this intent, we define structural equivalences between expressions, Definition 1.4. This is a standard procedure in computer science to undo the distortion in literal meaning arising from the constraints of linear text. The first two equivalences, (xii) and (xiii), erase any notion of space in the Kappa language. This is important to keep in mind, since textual (and graphical) renditions have a tendency to fool us. In particular, rule (xii) states that an interface is a set, not an ordered sequence of sites. Hence, the placement of sites in a graphical representation, such as Figure 2, has no significance. Rule (xiii) states that an expression has no spatial meaning. Every agent or complex is "equidistant" from any other, since all shuffles of an expression are equivalent. An expression, therefore, represents a well-mixed solution of molecular objects. Rule (xiv) states that we can relabel edges (bonds) as we please, provided the labels remain unique. Thus, if j is an edge label in an expression E and i is not, then we can substitute i for j in E (denoted by $E[i/j]$) without changing the meaning of E.

1.2 Patterns. An expression representing the contents of a reaction mixture typically contains complexes made of agents with a completely specified interface. In the main text we refer to these as molecular species or ground-level objects. However, it is useful to consider agents with only a partially specified interface. Recall that chemical rules, such as the one in Figure 1A, refer to partially specified molecules for the purpose of isolating an action that occurs across many reaction instances consisting of different fully specified molecules.

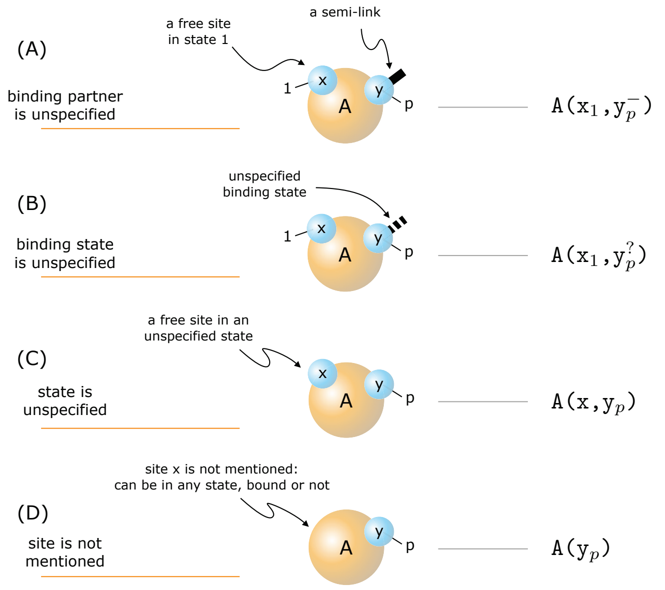

(A)
binding partner is unspecified
a free site in state 1
a semi-link
1
x
y
A
p
$\mathtt{A(x}_1\mathtt{,y}_p^-\mathtt{)}$
(B)
binding state is unspecified
unspecified binding state
1
x
y
A
p
$\mathtt{A(x}_1\mathtt{,y}_p^?\mathtt{)}$
(C)
state is unspecified
a free site in an unspecified state
x
y
A
p
$\mathtt{A(x,y}_p\mathtt{)}$
(D)
site is not mentioned
site x is not mentioned: can be in any state, bound or not
y
A
p
$\mathtt{A(y}_p\mathtt{)}$

Fig. 3. Basic Kappa patterns. A pattern consists of a partially specified agent or set of agents. Information can be omitted as follows. A: the binding partner of a site is left unspecified. B: the binding state of a site is left unspecified. C: the internal state of a site is unspecified. D: both internal and binding states of a site are left unspecified by not mentioning the site at all. See text for details.

A *pattern* is an expression consisting of partially specified agents. Figure 3 depicts the basic patterns exemplified by an agent $\mathtt{A(x,y)}$ with two sites:

1. *Unspecified binding partner.* The expression $\mathtt{A(x_{1},y_{p}^{-})}$, Figure 3A, specifies an agent in state 1 at site $\mathtt{x}$ and in state p at site $\mathtt{y}$. In addition, site $\mathtt{y}$ is bound, but we don't specify to whom. We call this a *semi-link* and indicate it by a hyphen $(-)$ instead of an edge label.
2. *Unspecified binding state.* The agent expression $\mathtt{A(x_{1},y_{p}^{?})}$, Figure 3B, is similar to the previous one, except that we do not care whether site $\mathtt{y}$ is bound. We indicate this by a question mark (?) in the bond superscript. Note that by *not* mentioning any binding state for site $\mathtt{x}$ we assert that this site is free (unbound).
3. *Unspecified internal state.* In $\mathtt{A(x,y_{p})}$, Figure 3C, we do not care about the internal state of site $\mathtt{x}$, because we omit its subscript. However, we do care that the site be free (as in all previous cases). Site $\mathtt{y}$ is in state p and free.
4. *Omitted site.* In $\mathtt{A(y_{p})}$, Figure 3D, we omit site $\mathtt{x}$ entirely, asserting that we don't care about its internal state nor its binding state. Site $\mathtt{y}$ is in state p and free.

1.3 Rules. The main use of patterns is in the definition of Kappa rules. In analogy to chemical reaction rules, a rule is a pair of expressions that are typically patterns:

$$E_{\text{left}} \longrightarrow E_{\text{right}}.$$

The pattern E_{left} defines conditions on internal states and binding states of agents that have to be satisfied for the rule to apply. Rules are applied to a mixture, that is, an expression S representing the contents of a reaction system at a given time. The basic idea is illustrated in Figure 4 for the rule

$$\begin{gathered}\mathtt{Gab1(PH^{-},Y447^{1}),PI3K(SH2^{1},s),PIP(three_{u},four_{p},five_{p})} \qquad \textbf{[2]}\\ \big\downarrow \\ \mathtt{Gab1(PH^{-},Y447^{1}),PI3K(SH2^{1},s^{2}),PIP(three_{u}^{2},four_{p},five_{p})}\end{gathered}$$

(which we write vertically for ease of placement on the page). Below the rule in Figure 4, we have sketched a hypothetical mixture. We want to identify *a* configuration of (fully specified) agents in the reaction mixture S that satisfies the pattern of reactants on the left hand side (lhs), E_{left}, of the rule. When such a configuration has been located, it is replaced by the configuration specified on the right hand side (rhs), E_{right}, of the rule. Replacement consists in updating the internal states and the binding states that are changed by the rule. The operational meaning of a match and a replacement are formalized in section 1.4 and 1.5, respectively.

We can think of a rule as an *action* that is applied to a configuration in the mixture. The action is the difference between the right hand side (rhs) and lhs of a rule. The differences may be many, such as changing several internal states and binding states at once, but they all boil down to a handful of *elementary actions* that cannot be further decomposed within the present definition of Kappa: binding, unbinding, and the change of an internal state. Kappa also allows for the creation and the removal of an agent.

Rules must obey certain constraints to be sound. Obviously, expressions E_{left} and E_{right} must be well-formed, that is, in compliance with Definitions 1.1, 1.2, and 1.3. The interpretation of a rule, however, requires a mapping of agent identities across the arrow. We must know which agents on the right of a (textual) rule correspond to which agents on its left. There are several ways of defining such a mapping. We opted for a simple convention: both sides of a rule, E_{left} and E_{right}, are compared with one another proceeding from the left of each expression. The comparison only checks agent names and interfaces, but is blind to the states of the sites mentioned. It ends at the first difference. This procedure identifies a longest left-anchored substring – a prefix – common to both expressions. (It may be empty.) The prefix now establishes a sequential correspondence between agents on the left and right hand sides of a rule. Anything after the common prefix is interpreted in terms of deletions and introductions of agents, depending on whether an agent is missing on the right or left hand side, respectively. Subtleties of the mapping rise to the user's attention only when using textual input. An example may help.

$$\mathtt{A(x^{1}),B(x^{1},y_{u})} \rightarrow \mathtt{A(x^{1}),B(x^{1},y_{p})} \qquad \big\{ \text{change state of } \mathtt{B} \qquad \textbf{[3]}$$

The common prefix in rule 3 establishes a correspondence between the agents mentioned on the left and the right. This rule states that if agent $\mathtt{A}$ is bound at site $\mathtt{x}$ to $\mathtt{B}$ at site $\mathtt{x}$ and $\mathtt{B}$ is unphosphorylated at site $\mathtt{y}$ (more precisely, "site $\mathtt{y}$ is in state u"), $\mathtt{B}$ will be phosphorylated at $\mathtt{y}$ – a common situation in signaling.

Let us now replace $\mathtt{A(x^{1}),B(x^{1},y_{p})}$ with $\mathtt{B(x^{1},y_{p}),A(x^{1})}$. By themselves, these expressions denote the same graph or complex. However, in the context of a rule, where a correspondence between agents on both sides has to be established to represent a set of actions, the structural equivalence, Definition 1.4, is suspended. The left and the right hand side of the rule have no common prefix, which triggers the addition and deletion actions:

$$\mathtt{A(x^{1}),B(x^{1},y_{u})} \rightarrow \mathtt{B(x^{1},y_{p}),A(x^{1})} \qquad \begin{cases}\text{delete the } \mathtt{A} \text{ referenced on the left}\\ \text{delete the } \mathtt{B} \text{ referenced on the left}\\ \text{add a } \mathtt{B(x,y_{p})}\\ \text{add an } \mathtt{A(x)}\\ \text{bind } \mathtt{B(x,y_{p})} \text{ at } \mathtt{x} \text{ to } \mathtt{A(x)} \text{ at } \mathtt{x}\end{cases} \qquad \textbf{[4]}$$

1.4 Pattern matching. Matching is a process that establishes whether a more detailed expression E' conforms to a less detailed expression E. To gain some intuition, consider agents first. A specification A' of an agent *conforms* to a specification A, if

(i) A' and A coincide in agent name and all site names that A mentions, and
(ii) the state values $(\iota \in \mathbb{V} \mid \epsilon)$ and binding values $(\lambda \in \mathbb{N} \mid - \mid ?)$ of each site mentioned in A, are either equal or less specific than those mentioned in A'. With regard to binding state, '?' is less specific than ϵ or '$-$', and '$-$' is less specific than a label $i \in \mathbb{N}$. With regard to internal state, ϵ is less specific than a value $\iota \in \mathbb{V}$.

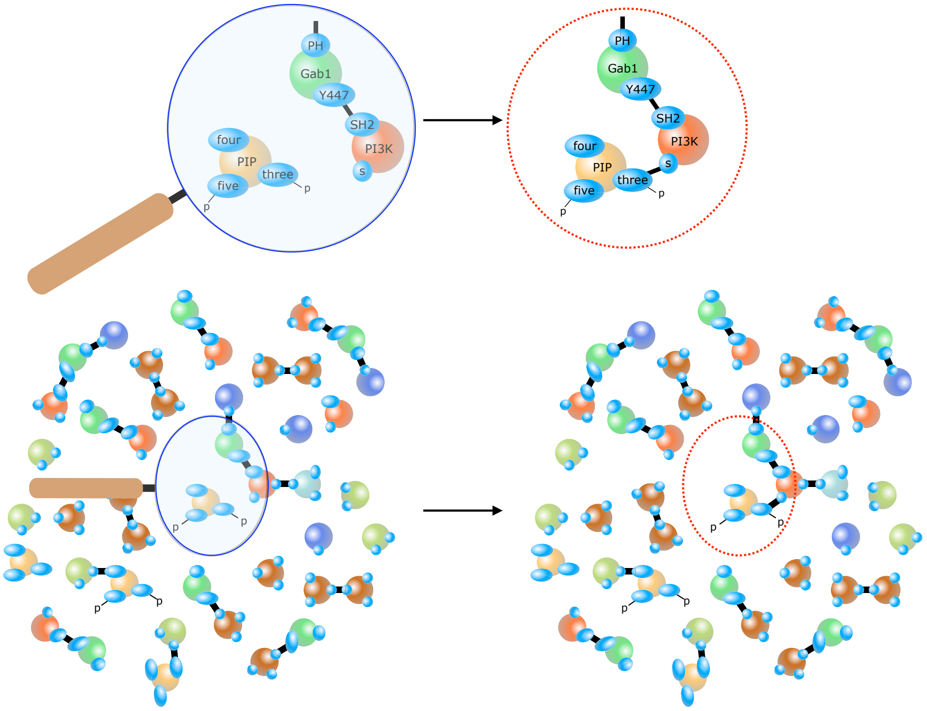


Fig. 4. Rule application in Kappa. First, a match between the pattern on the left hand side of a rule (blue lens) and the mixture (bottom) is identified. The action specified by the rule is then applied to the matching configuration, resulting in a new configuration according to the rule's right hand side (red circle). Many matchings may be possible for any given rule and many different rules may be applicable at any given moment. Rules and matchings are chosen for execution in a way that generates probabilistically correct sequences of events, following a generalization [3] of the Doob-Gillespie algorithm [5, 6] for stochastic chemical kinetics.

The concept of a match can be extended to expressions (mixtures) E' and E, by saying that E' conforms to E, written as $E' \models E$, if every agent in E' conforms to a distinct agent in E. In particular, anything conforms to an empty expression. Usually, E' is a reaction mixture, and E is the pattern on the lhs of a rule. We next formalize the notion of "being conformant" as a satisfaction relation $\models$. Symbols refer to the corresponding syntactical categories as in the agent Definition 1.1. The specificity ranking of binding states is such that '?' (unknown) subsumes 'ϵ' (free) and '$-$' (bound), and '$-$' (bound) subsumes a binding label indicating a specific bond to an agent identified in the expression. Likewise, the specificity ranking of internal states is such that 'ϵ' (unspecified) subsumes any specified state. In symbols:

$$\begin{tikzcd} & \epsilon & & \\ \text{binding state:} \quad ? \arrow[ur] \arrow[dr] & & \text{internal state:} \quad \epsilon \arrow[r] & \iota \in \mathbb{V} \\ & - \arrow[r] & \lambda \in \mathbb{N}, & \end{tikzcd} \qquad [5]$$

where the arrow means "is a superset of" (or, equivalently "is less specific than"): $x \rightarrow y \equiv x \supseteq y$. Equality applies between two λs that are identical in value. Of course, we have $\epsilon = \epsilon$ and $? = ?$ (question marks). In the following, a fraction denotes an inference from the precondition (in the numerator) to the postcondition (in the denominator), i.e., $\frac{A}{B}$ means "if A then B".

Definition 1.5 (Conforming, Matching)**.** To establish whether E' conforms to (matches) E, $E' \models E$, apply the following criteria:

(i)	site match	$n_{\iota'}^{\lambda'} \models n_{\iota}^{\lambda}$, if $\lambda' \subseteq \lambda$ and $\iota' \subseteq \iota$
(ii)	empty interface	$\sigma' \models \varnothing$
(iii)	interface	$\dfrac{s' \models s \quad \sigma' \models \sigma}{s', \sigma' \models s, \sigma}$
(iv)	agent name	$\dfrac{\sigma' \models \sigma}{N(\sigma') \models N(\sigma)}$
(v)	empty expression	$E' \models \varepsilon$
(vi)	expression	$\dfrac{a' \models a \quad E' \models E}{a', E' \models a, E}$

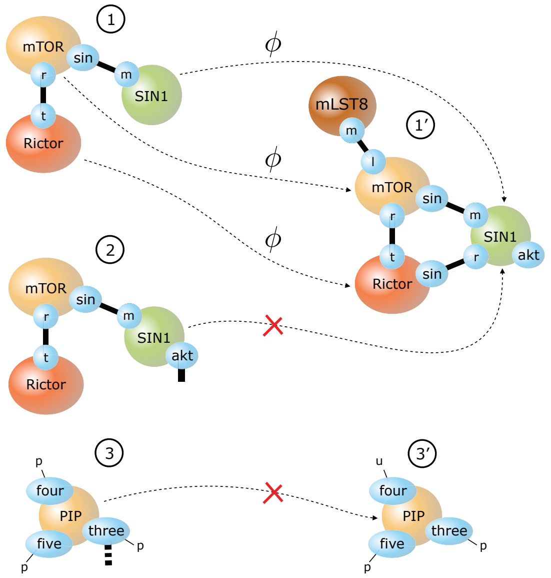


Fig. 5. Pattern matching. The embedding (fitting) of less detailed graphs on the left into more detailed graphs on the right. Graph 1 on the left embeds into complex 1' (a signaling assembly known as mTORC2, consisting of mTOR, Rictor, SIN1, and mLST8). Graph 1 agrees in all the names and states it mentions with graph 1'. Since graph 1 omits the site `l` from agent `mTOR`, `l`'s state in graph 1' is irrelevant. The label ϕ indicates the injective mapping from agents in 1 into agents in 1'; we say graph 1 embeds into graph 1', ϕ, $G_1 \lhd_\phi G'_1$. The graphs G'_1 and G_1 can be rendered in terms of their respective string expressions E'_1 and E_1. Upon arranging the strings according to ϕ, the criteria in Definition 1.5 establish that E'_1 conforms to E_1, $E'_1 \models E_1$. Graph 2, in contrast, does not fit complex 1', as the former demands that `SIN1` be bound to something at its site `akt`, but 1' specifies site `akt` to be free. In graph 3, agent `PIP` on the left does not match `PIP` in complex 3' on the right, as the latter has site `four` in an unphosphorylated state, while the former requests a phosphorylated state. There is no disagreement on site `three`, as graph 3 does not care about its binding state (the dotted line stands for a '?' in the textual representation, indicating "bound or unbound").

Definition 1.5 is understood as a relation between literal expressions (strings of text), established by stepping through the strings E' and E from left to right. However, finding a match of E' to E may necessitate the inspection of several structural equivalences of E', generated by reordering agents, interfaces, and relabeling bonds using Definition 1.4. It is not part of Definition 1.5 to produce such a reordering; rather, the reordering is an implicit input through the literal form of E'.

Embedding a graph into another graph. An expression E' can be represented by a site graph G', in which nodes are agents identified by their sequential position in the expression (see Figure 2) and edges correspond to bonds between sites as indicated in the expression. The concept of matching an expression E' to a less specific expression E, $E' \models E$, has a natural extension in the notion of embedding the graph G (corresponding to E) into the graph G' (corresponding to E'). An embedding of G into G' is equivalent to first finding an expression E'' that is structurally the same as E', $E'' \equiv E'$, which amounts to a graph isomorphism between G' and G'' (corresponding to E''), and then a matching of E' to E, which amounts to a graph inclusion from G into G'. In symbols: $E' \equiv E'' \models E$ corresponds to $G'' \overset{iso}{=} G' \xleftarrow{incl} G$. Keep in mind that we think of a smaller (less detailed) graph G as being embedded into a larger (more detailed) graph G', while a more detailed expression E' matches a less detailed expression E (the "pattern"). When dealing with graph embeddings it is more natural to simply think of the smaller graph G being relabelled (isomorphism) to "fit" the larger one. Intuitively, an embedding of G into G' (the first not larger than the second) is a process whereby we move G over G', trying to overlay G on G' such that agent types match (as well as the states of the sites mentioned in G). Several overlays may be possible, because either graph may contain multiple agents of the same type connected in the same way. Each overlay generates a distinct embedding. For example, let G be the obvious graph of $\mathtt{A}_1\mathtt{(x}^1\mathtt{)},\mathtt{A}_2\mathtt{(y}^1\mathtt{)}$ and G' the graph of $\mathtt{A}_1\mathtt{(x}^1\mathtt{)},\mathtt{A}_2\mathtt{(y}^1\mathtt{,x}^2\mathtt{)},\mathtt{A}_3\mathtt{(y}^2\mathtt{)}$. (Since isomorphisms play a role, we attach an identifier to each agent.) There are two embeddings of G into G': (1) the inclusion of G into G' and (2) a nontrivial isomorphism of G, i.e. the graph of $\mathtt{A}_2\mathtt{(x}^1\mathtt{)},\mathtt{A}_3\mathtt{(y}^1\mathtt{)}$, that also embeds into G'. Figure 5 provides a few graphical examples to fix the concepts. It is much more convenient to reason formally in terms of graphs than expressions, which is what we do in the main text and in subsequent sections. We notate the embedding of G into G' as $G \lhd_\phi G'$, where the subscript ϕ indicates a particular embedding.

1.5 Replacing a pattern. The execution of a rule $E_{\text{left}} \rightarrow E_{\text{right}}$ consists in testing whether an expression S conforms to E_{left}, $S \models E_{\text{left}}$, as defined in section 1.4, and then overwriting (updating) the matching region in S with E_{right}. Typically, the expression S represents the contents of a reaction mixture. Here we formalize what it means to overwrite an expression E_l with another expression E_r, $E_l[E_r]$. The definition of replacement below makes use of a "null"-agent $\varnothing$ for the purpose of describing agent deletion and addition. However, we have not defined a null-agent in Definition 1.1. Instead, we shall use the following convention. Let *prefix* be the longest common left-anchored substring between the lhs and the rhs in the rule *lhs* $\rightarrow$ *rhs*, as described in section 1.3. Let L (R) be the remainder of *lhs* (*rhs*) after the *prefix*, thus, *lhs* $=$ *prefix*,L and *rhs* $=$ *prefix*,R. For the replacement rules to add the agents in R and delete those in L, we pad the rule with appropriately placed null-agents, $|R|$ null agents on the left and $|L|$ on the right:

$$\textit{prefix}, L, \underbrace{\varnothing, \ldots, \varnothing}_{|R|\text{ times}} \longrightarrow \textit{prefix}, \underbrace{\varnothing, \ldots, \varnothing}_{|L|\text{ times}}, R.$$

Proper execution of replacement must avoid capturing (i.e. duplicating) bond labels that exist elsewhere in E_l. Our implementations automatically avoid capture by relabeling using Definition 1.4. Furthermore, to apply a rule with an empty lhs (production of agents) or with an empty rhs (deletion of agents), we need to extend the structural equivalences, Definition 1.4, and the matching relation, Definition 1.5, with a dummy "empty agent", $\varnothing$, that matches an empty lhs expression and that *is* an empty rhs that can be used to overwrite the deleted agent(s). Thus, Definition 1.4 is extended with $E \equiv E, \varnothing$, and Definition 1.5 with $\varnothing \vDash \varnothing$.

Definition 1.6 (Replacement)**.**

(i)	overwrite binding state	$\lambda_l[\lambda_r] = \begin{cases} \lambda_l & \text{if } \lambda_r \longrightarrow \lambda_l \text{ in } \mathbf{[5]}, \\ \lambda_r & \text{otherwise.} \end{cases}$
(ii)	overwrite internal state	$\iota_l[\iota_r] = \begin{cases} \iota_l & \text{if } \iota_r \longrightarrow \iota_l \text{ in } \mathbf{[5]}, \\ \iota_r & \text{otherwise.} \end{cases}$
(iii)	overwrite site	$n^{\lambda_l}_{\iota_l}[n^{\lambda_r}_{\iota_r}] = n^{\lambda_l[\lambda_r]}_{\iota_l[\iota_r]}$
(iv)	interface unchanged	$\sigma[\varnothing] = \sigma$
(v)	overwrite interface	$(s, \sigma)[s_r, \sigma_r] = s[s_r], \sigma[\sigma_r]$
(vi)	overwrite agent	$N(\sigma)[N(\sigma_r)] = N(\sigma[\sigma_r])$
(vii)	agent deletion	$N(\sigma)[\varnothing] = \varnothing$
(viii)	agent introduction	$\varnothing[N(\sigma_r)] = N(\sigma_r)$
(ix)	expression unchanged	$E[\varepsilon] = E$
(x)	overwrite expression	$(a, E)[a_r, E_r] = a[a_r], E[E_r]$

2 Glue-on-overlap procedure for fragments

Figure 6 illustrates an aspect of fragment construction. Consider whether pattern B can be made into a fragment with regard to rule r from Figure 2 in the main text, r: $\mathtt{A(a}_{u}\mathtt{,b}^{1}\mathtt{),B(c}^{1}\mathtt{)} \longrightarrow \mathtt{A(a}_{p}\mathtt{,b}^{1}\mathtt{),B(c}^{1}\mathtt{)}$. Recall from the discussion of Figure 2 (main text) that we need $B^\diamond \subset r^\diamond_{\mathtt{lhs}}$. While this is not the case for the B in Figure 6, B might be further specialized (refined) into B' by adding more context, after which it might qualify as a fragment for r. Refining B into B', such that $B'^\diamond \subset r^\diamond_{\mathtt{lhs}}$, amounts to glueing together B and $r_{\mathtt{lhs}}$. To do this, we first identify a glueing region. The glueing region of B and $r_{\mathtt{lhs}}$ is the set of agents and sites that both mention, along with a mutually compatible state (following the specificity ranking given in **[5]**). If there is no such state, the glueing region is empty. In the middle of the diamond in Figure 6 we see that B and $r_{\mathtt{lhs}}$ both mention agent A with site a in state u. Thus, $\mathtt{A(a}_u\mathtt{)}$ is the glueing region (shown at the top of the diamond). If B had left the internal state of site a unspecified, i.e. $B :=\mathtt{C(a}^1\mathtt{),A(a,c}^1\mathtt{,d)}$, then the glueing region with $r_{\mathtt{lhs}}$ would be $\mathtt{A(a)}$, since an unspecified internal state at site a is compatible with any specific state at that site, in particular u. In contrast, the pattern $B :=\mathtt{C(a}^1\mathtt{),A(a}_p\mathtt{,c}^1\mathtt{,d)}$ would result in an empty glueing region.

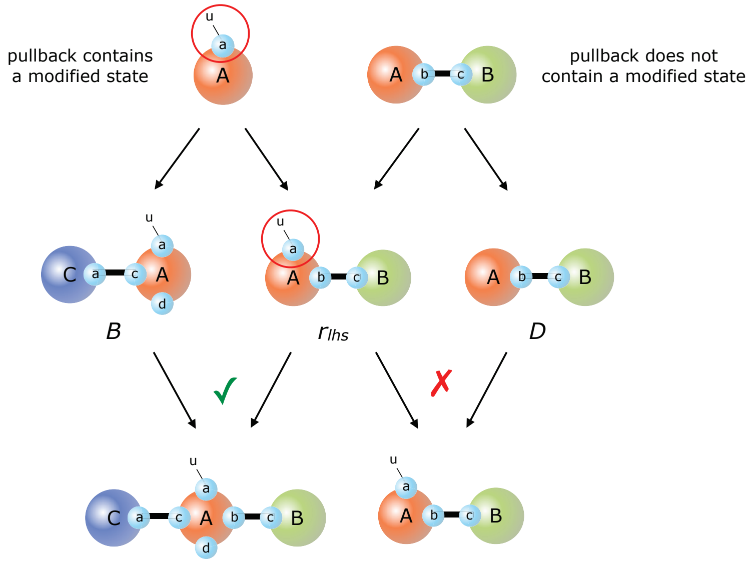


Fig. 6. Fragment construction by glue-on-overlap. Pattern B and the lhs component, $r_{\mathtt{lhs}}$, of rule r (Figure 2 of main text) have a glueing region, known in category theory as a pullback. The glueing region acts an instruction for joining both patterns into a new one (the so-called pushout), shown at the bottom of the diamond. Patterns D and $r_{\mathtt{lhs}}$ are not joined (red cross), because their pullback does not contain a site modified by r. See text for details.

We now can join B and $r_{\texttt{lhs}}$ by overlaying them on the glueing region, $\texttt{A(a}_u\texttt{)}$. This yields a new pattern that qualifies as a fragment as far as rule r is concerned. This new pattern is subject to a similar procedure with the next rule in line to be checked.

Such a refinement is not needed, however, unless a condition is met. The red circle in Figure 6 indicates the site and state modified by r, which the glueing region must contain if a pattern is to be refined into a fragment candidate. For example, pattern D in Figure 6 does have a glueing region with $r_{\texttt{lhs}}$, which does not contain the site and state modified by r. Rule r still acts on instances of D, since any ground-level instance that matches $r_{\texttt{lhs}}$ also matches D. However, since D does not care about the action of r, any instance of D transformed by r results in another instance of D, with no net effect on the concentration of D. This only means that there is no point in further refining such a pattern with regard to r. Nonetheless, a rule s might give rise to a fragment whose overlap with rule r does not contain sites and states modified by r. Naturally, that fragment's ODE will not receive any kinetic terms from r.

3 Illustration of syntactical criteria

To illustrate syntactical criteria, such as Cov1, we consider flows of control within and across agents. Consider Figure 4B in the main text and imagine two rules r_1 and r_2 both of which modify the state at site b of agent A, but r_1 conditions the modification of b on the state of site a, while r_2 conditions it on the state of c. To fix ideas, let r_1: $\texttt{A(a}_0\texttt{,b}_0\texttt{)} \longrightarrow \texttt{A(a}_0\texttt{,b}_1\texttt{)}\,@\,k_1$ and r_2: $\texttt{A(c}_1\texttt{,b}_1\texttt{)} \longrightarrow \texttt{A(c}_1\texttt{,b}_0\texttt{)}\,@\,k_2$, as shown in Figure 7. Let a ground-level species be denoted by a triplet abc reporting the state (0 or 1) at each site and let $*$ be a wildcard for expressing patterns. The lhs of r_1, $\texttt{A(a}_0\texttt{,b}_0\texttt{)}$, which is $00*$, is not a unit of the dynamics, as can be easily seen: $d[00*]/dt = d[000]/dt + d[001]/dt = -k_1[000] - k_1[001] + k_2[011] = -k_1[00*] + k_2[011]$. The reason is a configuration in $00*$ - specifically 001 - that satisfies the lhs of r_1 and the rhs of r_2. As a consequence, the rate equation of $[00*]$ receives a term from the action of r_2 (on 011). Thus, the concentration of 011 matters for describing the dynamics of the system, which means that the state at both controlling sites a and c must be known at any time. This forces both sites into the same covering class as b.

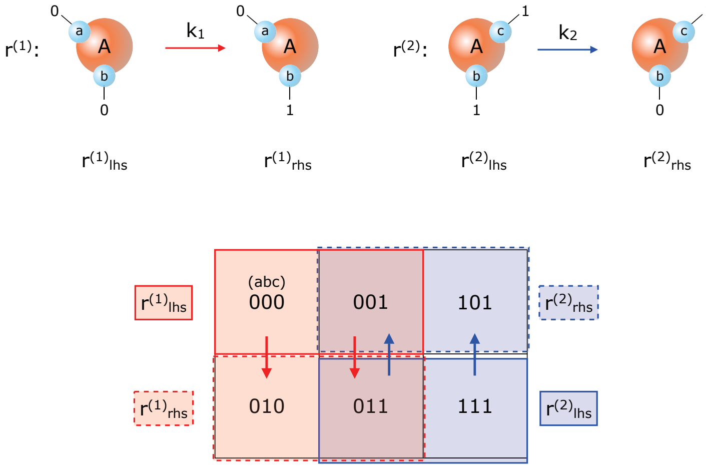


Fig. 7. Example illustrating directive Cov1 for the case shown in Figure 4B of the main text. Two rules, as described in the text, are shown at the top. The bottom illustrates the action of the rules on microconfigurations (ground-level objects). The solid (dotted) rectangles cover the configurations that match the pattern on the lhs (rhs) of each rule. See text for details.

The case of Figure 4D in the main text consists in two rules r_1 and r_2 that modify sites b and c, respectively, while both testing a condition on site a. For example, let r_1: $\texttt{A(a}_0\texttt{,b}_0\texttt{)} \longrightarrow \texttt{A(a}_0\texttt{,b}_1\texttt{)}\,@\,k_1$ and r_2: $\texttt{A(a}_0\texttt{,c}_0\texttt{)} \longrightarrow \texttt{A(a}_0\texttt{,c}_1\texttt{)}\,@\,k_2$, as shown in Figure 8. Here, too, there is a ground-level configuration, 000, whose concentration is affected by both rules, because it matches the lhs of r_1 and r_2. However, r_2 transforms 000 into 001, which still matches the lhs of r_1. In fact, this is an example of a situation in which the glueing region (section 2) between the lhs of r_1 and the lhs of r_2 does *not* contain either modified site. Hence, r_2 maps one subset of the extension of $00*$ into another without affecting the concentration of $00*$. Indeed, $00*$ is a fragment, as we can easily check: $d[00*]/dt = d[000]/dt + d[001]/dt = -k_1[000] - k_2[000] - k_1[001] + k_2[000] = -k_1[00*]$. Agent A has therefore two covering classes, $\{\texttt{a},\texttt{b}\}$ and $\{\texttt{a},\texttt{c}\}$, for a total of four fragments: $00*$, $01*$, $0{*}0$, and $0{*}1$. In general, a covering class contains the backward closure of all sites that control a particular locus of modification.

4 Independence and self-consistency

We discuss two simple examples to clarify the notions of independence [7, 8, 9] and self-consistency.

4.1 Example 1: Independence and tileable systems.

Consider a scaffold protein C(a,b) with a specific binding site for A(a) and B(b), as shown in Figure 9A. This system comprises six possible molecular species, $A, B, C, A.C, C.B, A.C.B$, where the dot indicates a bond. In this section, we use mathematical font (slanted) when referring to molecular species or patterns and their concentration variables as they appear in

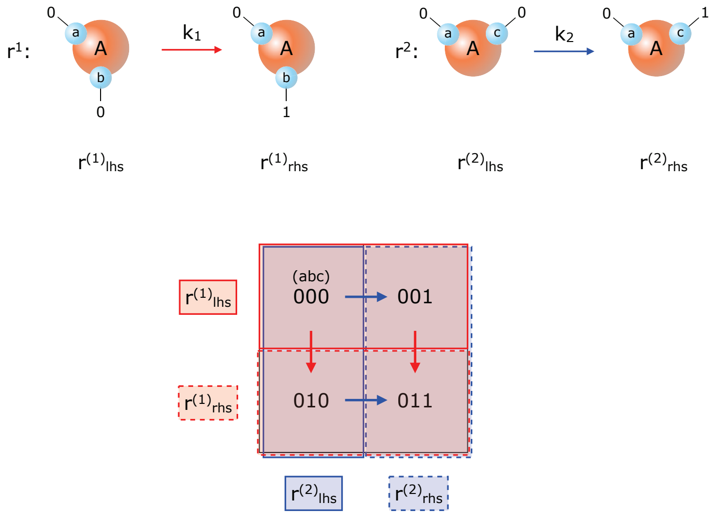


Fig. 8. Example illustrating directive Cov1 for the case shown in Figure 4D of the main text. Two rules, as described in the text, are shown at the top. The bottom illustrates the action of the rules on microconfigurations (ground-level objects). The solid (dotted) rectangles cover the configurations that match the pattern on the lhs (rhs) of each rule. See text for details.

deterministic kinetic equations. The six species are related by four association and dissociation reactions, as shown on the left:

$$\left.\begin{array}{rcl} A + C & \underset{k_{-1}^{AC}}{\overset{k_1^{A,C}}{\rightleftharpoons}} & A.C \\ A + C.B & \underset{k_{-1}^{AC}}{\overset{k_1^{A,C}}{\rightleftharpoons}} & A.C.B \end{array}\right\} \quad \mathtt{A(a),C(a)} \longleftrightarrow \mathtt{A(a^1),C(a^1)}$$

$$\left.\begin{array}{rcl} B + C & \underset{k_{-1}^{CB}}{\overset{k_1^{B,C}}{\rightleftharpoons}} & C.B \\ B + A.C & \underset{k_{-1}^{CB}}{\overset{k_1^{B,C}}{\rightleftharpoons}} & A.C.B \end{array}\right\} \quad \mathtt{B(b),C(b)} \longleftrightarrow \mathtt{B(b^1),C(b^1)} \qquad \textbf{[6]}$$

To assert that the binding sites are independent is to assert that the rate constant for the binding of A to C is the same as the rate constant for the binding of A to $C.B$, and likewise for the interactions between C and B. This kinetic indistinguishability means that only four Kappa-rules, shown on the right of **[6]**, are needed to express the eight reactions on the left. The mutual independence of the interactions between `A` and the scaffold `C` on the one hand and `B` and `C` on the other is expressed by omitting site `a` and `b`, repectively, from the corresponding interaction rules. As a consequence, the binding reaction between `A` and `C` (right arrow in the first rule) expands into two microscopic reactions on the left with identical rate constants, $k_1^{A,C}$.

The full dynamical system is given by:

$$\begin{aligned} \frac{d[A]}{dt} &= k_{-1}^{AC}\left([A.C]+[A.C.B]\right) - k_1^{A,C}[A]\left([C]+[C.B]\right) \\ \frac{d[B]}{dt} &= k_{-1}^{CB}\left([C.B]+[A.C.B]\right) - k_1^{B,C}[B]\left([C]+[A.C]\right) \\ \frac{d[C]}{dt} &= k_{-1}^{AC}[A.C] + k_{-1}^{CB}[C.B] - [C]\left([A]\,k_1^{A,C} + [B]\,k_1^{B,C}\right) \\ \frac{d[A.C]}{dt} &= k_1^{A,C}[A][C] + k_{-1}^{CB}[A.C.B] - [A.C]\left(k_{-1}^{AC} + [B]\,k_1^{B,C}\right) \\ \frac{d[C.B]}{dt} &= k_{-1}^{AC}[A.C.B] + k_1^{B,C}[B][C] - [C.B]\left(k_{-1}^{CB} + [A]\,k_1^{A,C}\right) \\ \frac{d[A.C.B]}{dt} &= k_1^{A,C}[A][C.B] + k_1^{B,C}[B][A.C] - [A.C.B]\left(k_{-1}^{AC} + k_{-1}^{CB}\right) \end{aligned} \qquad \textbf{[7]}$$

This system can be coarse-grained by conceptually splitting the centerpiece C into two fragments (Figure 9A), one containing only the A-binding site, the other only the B-binding site. This captures the fact that A and B cannot know about each other despite their interactions with a shared C. Let us denote the former fragment with $C*$ and the latter with $*C$, the asterisk indicating that we don't care about the corresponding

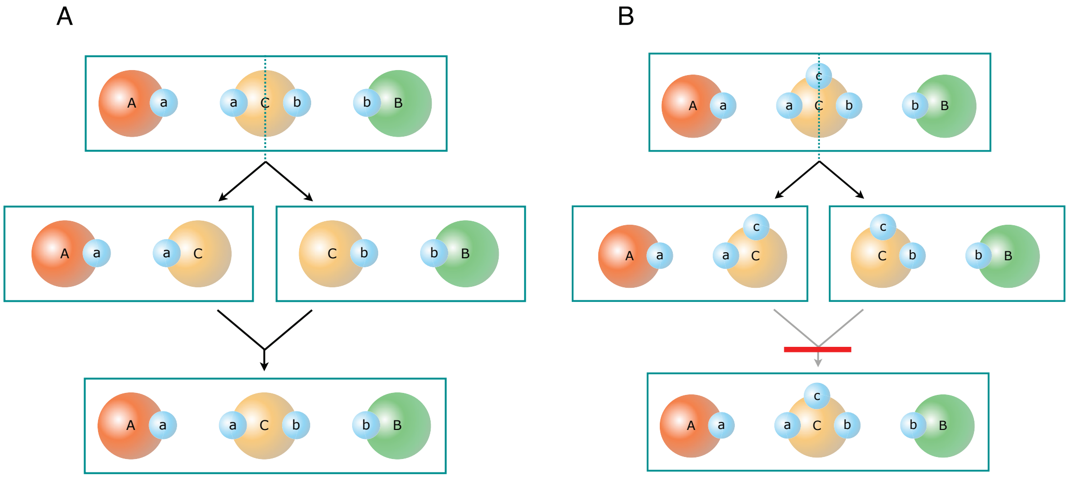


Fig. 9. Independence and self-consistency. The Figure depicts schematically Examples 1 (A) and 2 (B) described in the text. In both cases, the binding interactions on the "left" and the "right" of agent C do not influence one another. However, in case (B), agent C internally synchronizes the sites a and b through a dependency on the state of site c, but this correlation is not "readable" by the interactions that define the system. As in case (A), case (B) can be split into two self-consistently described subsystems, but they are no longer independent. The coarse-grained variables that enable the separation into subsystems cannot be used to reconstitute the microscopic dynamics.

binding site. The system then splits into two independent subsystems, self-consistently described by $\{A, C*, A.C*\}$ and $\{B, *C, *C.B\}$.

$$\begin{aligned}
&[A]\\
[C*] &\doteq [C] + [C.B]\\
[A.C*] &\doteq [A.C] + [A.C.B]\\
&[B]\\
[*C] &\doteq [C] + [A.C]\\
[*C.B] &\doteq [C.B] + [A.C.B]
\end{aligned}$$

Self-consistent means that each set of variables is closed with regard to its own dynamics:

$$\begin{aligned}
\frac{d[A]}{dt} &= k_{-1}^{AC}[A.C*] - k_1^{A,C}[A][C*]\\
\frac{d[C*]}{dt} &= k_{-1}^{AC}[A.C*] - k_1^{A,C}[A][C*]\\
\frac{d[A.C*]}{dt} &= k_1^{A,C}[A][C*] - k_{-1}^{AC}[A.C*]
\end{aligned} \qquad [8]$$

and

$$\begin{aligned}
\frac{d[B]}{dt} &= k_{-1}^{BC}[*C.B] - k_1^{B,C}[B][*C]\\
\frac{d[*C]}{dt} &= k_{-1}^{BC}[*C.B] - k_1^{B,C}[B][*C]\\
\frac{d[*C.B]}{dt} &= k_1^{B,C}[B][*C] - k_{-1}^{BC}[*C.B]
\end{aligned} \qquad [9]$$

C does not propagate any information between A and B, and it does not correlate them either. Independent then means that we can reconstruct the microscopic dynamics from the description of both subsystems, as outlined next.

Suppose we pick a C at random from the reaction mixture and observe it to be bound to a B. The conditional probability that the same C is also bound to an A is formally written as $P(A.C*|*C.B)$. If, on the other hand, we choose not to observe the B-binding site of the C we picked, that probability is $P(A.C*|*C*)$. Clearly, independence means that the two conditional probabilities are equal:

$$P(A.C*|*C.B) = P(A.C*|*C*).$$

By definition of a conditional probability, we obtain:

$$P(A.C.B) = \frac{P(A.C*)P(*C.B)}{P(*C*)}. \qquad [10]$$

These relationships are reflected by the corresponding (time-dependent) concentrations $[A.C*]$, $[*C.B]$, and $[*C*]$. Thus, equation (10) asserts

$$[A.C.B] = \frac{([A.C] + [A.C.B])([C.B] + [A.C.B])}{[C] + [A.C] + [C.B] + [A.C.B]}, \qquad [11]$$

from which it follows that

$$X \doteqdot [A.C.B][C] - [A.C][C.B] = 0. \qquad [12]$$

It is straightforward to verify that $X(t) = 0$ is an invariant of motion, provided the concentrations satisfied $X(0) = 0$:

$$\frac{dX}{dt} = -X(k_1^{A,C}[A] + k_{-1}^{AC} + k_1^{C,B}[B] + k_{-1}^{CB}), \qquad [13]$$

where $k_1^{A,C}, k_1^{C,B}$ denote association constants and k_{-1}^{AC}, k_{-1}^{CB} dissociation constants. Note that $X(t)$ will decay exponentially, if $X(0) \neq 0$.

Equation **[10]** is a manifestation of independence and can be generalized to *define* a class of systems whose fragments behave like tiles in the following sense. Two fragments $\mathcal{F}_1$ and $\mathcal{F}_2$ can be "snapped" together (possibly in more than one way), $\mathcal{F}_1 \cup \mathcal{F}_2$, if they have a non-conflicting valuation on the sites of agents both mention, $\mathcal{F}_1 \cap \mathcal{F}_2$, precisely as the glue-on-overlap in section 2. In our example, the fragments $\mathcal{F}_1 = A.C*$ and $\mathcal{F}_2 = *C.B$ can be snapped together, since they overlap in the agent name C and don't conflict in the states of the sites they mention. (The first fragment specifies the state of the A-binding site, which the second fragment ignores, and the second fragment specifies the state of C's B-binding site, which the first fragment ignores.) The overlap $\mathcal{F}_1 \cap \mathcal{F}_2$ is $*C*$.

If a self-consistent set of fragments $\mathfrak{F} = \{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ obeys the independence equation **[10]**, we can extend a fragment $\mathcal{F}_i$ into a fragment $\mathcal{F}_i \cup \mathcal{F}_j$ whose concentration is given by the product of the concentrations of $\mathcal{F}_i$ and $\mathcal{F}_j$ divided by the concentration of the snapping region, the overlap $\mathcal{F}_i \cap \mathcal{F}_j$. By extending fragments in this way, we can invert the coarse-graining. That is, we can reconstruct any molecular species that can possibly occur in the system, while in the process computing its concentration via the tiling equation **[10]**. The microscopic dynamics expressed in terms of $\mathfrak{F}$ will be exact, provided the initial condition satisfied the independence relations of the form **[12]**, otherwise it will approach the exact dynamics exponentially according to **[13]**. Typical situations, however, are more subtle, as illustrated in the next example.

4.2 Example 2: Stealth correlation. Example 2, Figure 9B, has a similar setup as Example 1, but the central scaffold `C(a,c,b)` now has three binding sites. The purpose of site `c` is to control whether sites `a` and `b` are available for binding interactions. Assume that site `c` of agent `C` has to be bound (by something) to turn on the binding capability of the other two sites. (After appropriate name changes, this corresponds to the control flow depicted in Figure 4D of the main text.) To make the point expeditiously, assume that all binding interactions are pseudo first-order because of excess `A`, `B`, and a fourth agent that binds the controller site of `C`, Let us also assume that all interactions are irreversible. This enables us to just focus on how agent `C` approaches full occupancy. Define a binding state of `C` as a triplet (acb) indicating the status of each site as either occupied, o, or unoccupied, u. As in Example 1, the binding process at site a is independent of the binding state of site b, and vice versa. The system is then described by the 5 reactions shown below on the left and whose dynamics is detailed on the right:

$$\begin{array}{c} (uuu) \xrightarrow{k^c} (uou) \\ (uou) \xrightarrow{k^l} (oou) \\ (uoo) \xrightarrow{k^l} (ooo) \\ (uou) \xrightarrow{k^r} (uoo) \\ (oou) \xrightarrow{k^r} (ooo) \end{array} \quad \text{with} \quad \begin{aligned} \frac{d[uuu]}{dt} &= -k^c[uuu] \\ \frac{d[uou]}{dt} &= -(k^l + k^r)[uou] + k^c[uuu] \\ \frac{d[uoo]}{dt} &= -k^l[uoo] + k^r[uou] \\ \frac{d[oou]}{dt} &= k^l[uou] - k^r[oou] \\ \frac{d[ooo]}{dt} &= k^l[uoo] + k^r[oou] \end{aligned} \qquad [14]$$

Considering the absence of any information propagation between sites `a` and `b`, we can define new coarse-grained variables:

$$\begin{aligned} &[uuu] \\ [uo*] &\doteqdot [uou] + [uoo] \\ [oo*] &\doteqdot [oou] + [ooo] \\ [*ou] &\doteqdot [uou] + [oou] \\ [*oo] &\doteqdot [uoo] + [ooo] \end{aligned}$$

Again, as in Example 1, the system splits into two self-consistent sets of coarse-grained variables $\{(uu*), (uo*), (oo*)\}$ and $\{(*uu), (*ou), (*oo)\}$. Each fragment now includes the central binding site c, since it determines whether a or b can undergo binding. (Thus, by construction, $[uu*] = [*uu] = [uuu]$.)

$$\begin{aligned} \frac{d[uuu]}{dt} &= -k^c[uuu] \\ \frac{d[uo*]}{dt} &= -k^l[uo*] + k^c[uuu] \\ \frac{d[oo*]}{dt} &= k^l[uo*] \end{aligned} \quad \text{and} \quad \begin{aligned} \frac{d[uuu]}{dt} &= -k^c[uuu] \\ \frac{d[*ou]}{dt} &= -k^r[*ou] + k^c[uuu] \\ \frac{d[*oo]}{dt} &= k^r[*ou] \end{aligned} \qquad [15]$$

However, unlike in Example 1, we cannot recombine the subsystems to reconstruct the microscopic dynamics via equation **[10]**. In analogy to equation **[12]**, we define $X \doteqdot [ooo][uou] - [oou][uoo]$ as a measure of independence and obtain

$$\frac{dX}{dt} = -X(k_a + k_b) + k_c[ooo][uuu], \qquad [16]$$

indicating that the two subsystems remain correlated, because the state of the controller site `c` correlates the states of `a` and `b`. While the coarse grained dynamics is still exact, it can no longer be inverted by tiling. As can be seen from Figure 10, the concentration of the fully occupied form $[ooo]$ is underestimated by assuming independence (tiling equation **[10]**), because the injection of new instances of (uou), by virtue of the first reaction in **[14]**, keeps introducing correlations between sites a and b, thus maintaining $[ooo]$ above what an observer would expect by measuring $[oo*]$, $[*oo]$, and $[*o*]$ and assuming independence. This deviation from independence ceases once the system has exhausted its reservoir of (uuu) (having converted it all to (ooo), see Figure 10), since, after all, binding at site a is independent from binding at b. The coarse-graining, which enabled us to describe the dynamics of the system in terms of two self-consistent subsystems, throws away correlation information, preventing reconstitution of the original microscopic description. From a viewpoint internal to the system, this is no loss, as the correlation cannot be observed from within the system (unless additional specific interactions are posited), and a microscopic description is therefore irrelevant. As outside observers, however, we can reason over the reaction equations globally and notice that we could measure the state of a , to give us information about the state of c, from which we could infer the state of b.

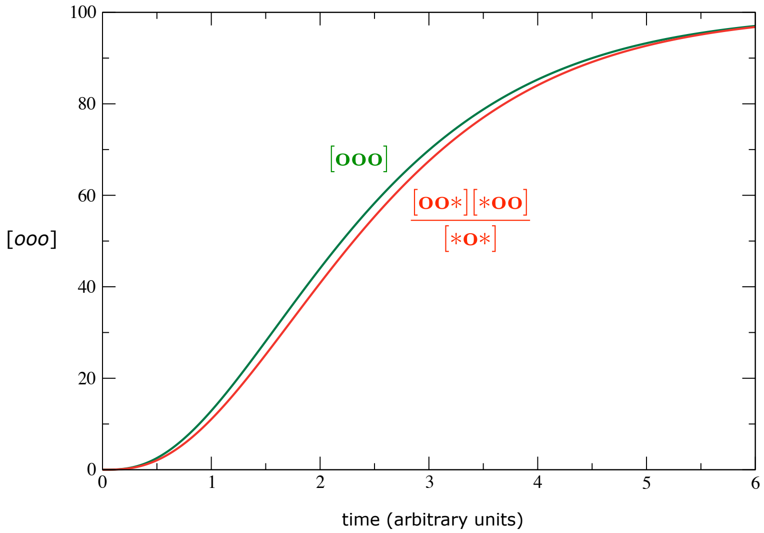


Fig. 10. Error from independence assumption. The Figure depicts the concentration dynamics of the fully occupied form of C, $[ooo]$, according to the equations **[14]** (green). The red curve shows the dynamics of $[ooo]$ when it is computed using the independence equation **[10]**, $[ooo] = [oo*][*oo]/[*o*]$. Although this relation is violated in Example 2, the utter simplicity of this scenario makes the tiling approximation still appear reasonable. This would not be the case in more complex situations.

5 The early-EGF model

We use a canonical cell signaling pathway in mammalian cells, the epidermal growth factor receptor (EGFR or ErbB) pathway, to illustrate our coarse-graining procedure. This pathway involves four receptor tyrosine kinases that interact with several extracellular ligands and each other. More than 100 proteins are involved in the production and the processing of intracellular signals induced by EGF receptor activity. The EGF system has a controling influence on cell division, cell fate and cell morphology.

The rule system below describes only a tiny set of early events (that are by no means agreed-upon) in EGFR signaling. In the present context, our objective is not to provide an accurate model of the EGFR pathway, but rather to use this relatively small example to illustrate the logic of our approach to coarse-graining in rule-based representations of complex molecular interaction systems.

In essence, a signal arrives at the cell membrane in the form of a ligand, EGF (E in our rules), which binds to the extra-cellular portion of a receptor tyrosine kinase, EGFR (named R in our system), that reaches across the membrane. This binding process is the content of rule r02 (rule r01 describes the reverse process). Upon binding E, an R becomes capable of binding to a neighbouring R, also bound to a ligand (rule r03). Receptor pairs can cross-activate one another, meaning that they mutually phosphorylate certain of their intra-cellular residues (rules r05, r07). These phosphorylated residues now serve as binding sites for a variety of proteins in the cytoplasm, such as GRB2 (named G in our system) and SHC (here named S). G can bind the phosphorylated site $\mathtt{Y68}_p$ of R, as expressed in rules r12-r15. Concurrently, G can bind SOS (agent O), while G is bound to R (r16), or bound to S (r20, r22), or standalone (r18). Using different rules for the same action is a means for differentiating between (potentially) different kinetics (or simply rate constants) depending on context. Likewise, S can engage with R, while being free (r24, r26) or bound to G (r28, r30). The contexts of G, R, and S compound combinatorially and occasionally interfere with one another. In fact, there is a competition between two "mini-pathways" for recruiting O to the membrane receptor R: via G alone or via G bound to S. The competition derives from S and R interacting with the same binding site at G, forcing a choice for any individual G, as can be clearly seen in the contact map of the main text. These pathways can be tracked automatically with procedures that will be detailed in forthcoming manuscripts (but see [10]). All rule actions are reversible, and several actions come in contextual variants (so-called refinements) reflecting differences in rate constants.

The textual exposition of the rules below can be rendered (and edited) without loss of information in a graphical format along the lines of Figure 4A.

5.1 Rules.

r01: $\mathtt{E}(\mathtt{r}^1)\,,\mathtt{R}(\mathtt{l}^1,\mathtt{r}) \longrightarrow \mathtt{E}(\mathtt{r})\,,\mathtt{R}(\mathtt{l}\,,\mathtt{r})$
r02: $\mathtt{E}(\mathtt{r})\,,\mathtt{R}(\mathtt{l},\mathtt{r}) \longrightarrow \mathtt{E}(\mathtt{r}^1)\,,\mathtt{R}(\mathtt{l}^1,\mathtt{r})$
r03: $\mathtt{E}(\mathtt{r}^2)\,,\mathtt{E}(\mathtt{r}^1)\,,\mathtt{R}(\mathtt{l}^2,\mathtt{r})\,,\mathtt{R}(\mathtt{l}^1,\mathtt{r}) \longrightarrow \mathtt{E}(\mathtt{r}^3)\,,\mathtt{E}(\mathtt{r}^2)\,,\mathtt{R}(\mathtt{l}^3,\mathtt{r}^1)\,,\mathtt{R}(\mathtt{l}^2,\mathtt{r}^1)$

r04: $E(r^{3}), E(r^{2}), R(l^{3},r^{1}), R(l^{2},r^{1}) \longrightarrow E(r^{2}), E(r^{1}), R(l^{2},r), R(l^{1},r)$

r05: $E(r^{3}), E(r^{2}), R(l^{3},r^{1}), R(Y68_{u},l^{2},r^{1}) \longrightarrow E(r^{3}), E(r^{2}), R(l^{3},r^{1}), R(Y68_{p},l^{2},r^{1})$

r06: $R(Y68_{p}) \longrightarrow R(Y68_{u})$

r07: $E(r^{3}), E(r^{2}), R(l^{3},r^{1}), R(Y48_{u},l^{2},r^{1}) \longrightarrow E(r^{3}), E(r^{2}), R(l^{3},r^{1}), R(Y48_{p},l^{2},r^{1})$

r08: $R(Y48_{p}) \longrightarrow R(Y48_{u})$

r09: $R(Y48^{1}_{p},r^{-}), S(Y7_{u},c^{1}) \longrightarrow R(Y48^{1}_{p},r^{-}), S(Y7_{p},c^{1})$

r10: $S(Y7_{p},c^{-}) \longrightarrow S(Y7_{u},c^{-})$

r11: $S(Y7_{p},c) \longrightarrow S(Y7_{u},c)$

r12: $G(a,b), R(Y68_{p}) \longrightarrow G(a^{1},b), R(Y68^{1}_{p})$

r13: $G(a^{1},b), R(Y68^{1}_{p}) \longrightarrow G(a,b), R(Y68_{p})$

r14: $G(a,b^{-}), R(Y68_{p}) \longrightarrow G(a^{1},b^{-}), R(Y68^{1}_{p})$

r15: $G(a^{1},b^{-}), R(Y68^{1}_{p}) \longrightarrow G(a,b^{-}), R(Y68_{p})$

r16: $G(a^{1},b), R(Y68^{1}_{p}), O(d) \longrightarrow G(a^{2},b^{1}), R(Y68^{2}_{p}), O(d^{1})$

r17: $G(a^{2},b^{1}), R(Y68^{2}_{p}), O(d^{1}) \longrightarrow G(a^{1},b), R(Y68^{1}_{p}), O(d)$

r18: $G(a,b), O(d) \longrightarrow G(a,b^{1}), O(d^{1})$

r19: $G(a,b^{1}), O(d^{1}) \longrightarrow G(a,b), O(d)$

r20: $G(a^{1},b), S(Y7^{1}_{p},c), O(d) \longrightarrow G(a^{2},b^{1}), S(Y7^{2}_{p},c), O(d^{1})$

r21: $G(a^{2},b^{1}), S(Y7^{2}_{p},c), O(d^{1}) \longrightarrow G(a^{1},b), S(Y7^{1}_{p},c), O(d)$

r22: $G(a^{1},b), S(Y7^{1}_{p},c^{-}), O(d) \longrightarrow G(a^{2},b^{1}), S(Y7^{2}_{p},c^{-}), O(d^{1})$

r23: $G(a^{2},b^{1}), S(Y7^{2}_{p},c^{-}), O(d^{1}) \longrightarrow G(a^{1},b), S(Y7^{1}_{p},c^{-}), O(d)$

r24: $R(Y48_{p}), S(Y7_{u},c) \longrightarrow R(Y48^{1}_{p}), S(Y7_{u},c^{1})$

r25: $R(Y48^{1}_{p}), S(Y7_{u},c^{1}) \longrightarrow R(Y48_{p}), S(Y7_{u},c)$

r26: $R(Y48_{p}), S(Y7_{p},c) \longrightarrow R(Y48^{1}_{p}), S(Y7_{p},c^{1})$

r27: $R(Y48^{1}_{p}), S(Y7_{p},c^{1}) \longrightarrow R(Y48_{p}), S(Y7_{p},c)$

r28: $G(a^{1},b), R(Y48_{p}), S(Y7^{1}_{p},c) \longrightarrow G(a^{2},b), R(Y48^{1}_{p}), S(Y7^{2}_{p},c^{1})$

r29: $G(a^{2},b), R(Y48^{1}_{p}), S(Y7^{2}_{p},c^{1}) \longrightarrow G(a^{1},b), R(Y48_{p}), S(Y7^{1}_{p},c)$

r30: $G(a^{2},b^{1}), R(Y48_{p}), S(Y7^{2}_{p},c), O(d^{1}) \longrightarrow G(a^{3},b^{2}), R(Y48^{1}_{p}), S(Y7^{3}_{p},c^{1}), O(d^{2})$

r31: $G(a^{3},b^{2}), R(Y48^{1}_{p}), S(Y7^{3}_{p},c^{1}), O(d^{2}) \longrightarrow G(a^{2},b^{1}), R(Y48_{p}), S(Y7^{2}_{p},c), O(d^{1})$

r32: $G(a,b), R(Y48^{1}_{p}), S(Y7_{p},c^{1}) \longrightarrow G(a^{2},b), R(Y48^{1}_{p}), S(Y7^{2}_{p},c^{1})$

r33: $G(a^{2},b), R(Y48^{1}_{p}), S(Y7^{2}_{p},c^{1}) \longrightarrow G(a,b), R(Y48^{1}_{p}), S(Y7_{p},c^{1})$

r34: $G(a,b), S(Y7_{p},c) \longrightarrow G(a^{1},b), S(Y7^{1}_{p},c)$

r35: $G(a^{1},b), S(Y7^{1}_{p},c) \longrightarrow G(a,b), S(Y7_{p},c)$

r36: $G(a,b^{-}), S(Y7_{p},c) \longrightarrow G(a^{1},b^{-}), S(Y7^{1}_{p},c)$

r37: $G(a^{1},b^{-}), S(Y7^{1}_{p},c) \longrightarrow G(a,b^{-}), S(Y7_{p},c)$

r38: $G(a,b^{2}), R(Y48^{1}_{p}), S(Y7_{p},c^{1}), O(d^{2}) \longrightarrow G(a^{3},b^{2}), R(Y48^{1}_{p}), S(Y7^{3}_{p},c^{1}), O(d^{2})$

r39: $G(a^{3},b^{2}), R(Y48^{1}_{p}), S(Y7^{3}_{p},c^{1}), O(d^{2}) \longrightarrow G(a,b^{2}), R(Y48^{1}_{p}), S(Y7_{p},c^{1}), O(d^{2})$

5.2 Compressed rules.

Each of the rules r01-r39 is subject to a compression procedure, as outlined in the main text. The compression of r04 (expressing the dissociation of dimerized R) into cr04 has lost a symmetry. Rule r04 has a symmetry that generates two equivalent embeddings into any concrete receptor dimer present in the reaction mixture. This symmetry is detected by the simulation algorithm [3], and causes the rate constant of r04 to be adjusted by a factor of 1/2. However, the symmetry is lost upon compression. $R(r^{-})$ on the left hand side of cr04 can still be matched in two equivalent ways by any receptor dimer in the mixture, but the simulation algorithm would forgo a division by 2, because of no detectable symmetry in the structure of cr04. This loss of symmetry is recognized by the compression process, which automatically compensates by adjusting the rate constant of cr04 to be half that of r04 (which we have arbitrarily set to 1). This is of no consequence for constructing the set of coarse-grained variables, but is required for preserving the quantitative kinetics of the rule system upon compression.

cr01: $R(l^{-},r) \longrightarrow R(l,r)$

cr02: $E(r), R(l) \longrightarrow E(r^{1}), R(l^{1})$

cr03: $R(l^{-},r), R(l^{-},r) \longrightarrow R(l^{-},r^{1}), R(l^{-},r^{1})$

cr04: $R(r^{-}) \longrightarrow R(r)$ @ 0.5

cr05: $R(Y68^{?}_{u},r^{-}) \longrightarrow R(Y68^{?}_{p},r^{-})$

cr06: $R(Y68_{p}) \longrightarrow R(Y68_{u})$

cr07: $R(Y48^{?}_{u},r^{-}) \longrightarrow R(Y48^{?}_{p},r^{-})$

cr08: $R(Y48_{p}) \longrightarrow R(Y48_{u})$

cr09: $R(Y48^{1},r^{-}), S(Y7^{?}_{u},c^{1}) \longrightarrow R(Y48^{1},r^{-}), S(Y7^{?}_{p},c^{1})$

cr10: $S(Y7_{p},c^{-}) \longrightarrow S(Y7_{u},c^{-})$

cr11: $S(Y7_{p},c) \longrightarrow S(Y7_{u},c)$

cr12: $\mathtt{G(a,b), R(Y68_{\mathit{p}}) \longrightarrow G(a^1,b), R(Y68^1_{\mathit{p}})}$

cr13: $\mathtt{G(a^1,b), R(Y68^1) \longrightarrow G(a,b), R(Y68)}$

cr14: $\mathtt{G(a,b^-), R(Y68_{\mathit{p}}) \longrightarrow G(a^1,b^-), R(Y68^1_{\mathit{p}})}$

cr15: $\mathtt{G(a^1,b^-), R(Y68^1) \longrightarrow G(a,b^-), R(Y68)}$

cr16: $\mathtt{G(a^1,b), R(Y68^1), O(d) \longrightarrow G(a^2,b^1), R(Y68^2), O(d^1)}$

cr17: $\mathtt{G(a^1,b^-), R(Y68^1) \longrightarrow G(a^1,b), R(Y68^1)}$

cr18: $\mathtt{G(a,b), O(d) \longrightarrow G(a,b^1), O(d^1)}$

cr19: $\mathtt{G(a,b^-) \longrightarrow G(a,b)}$

cr20: $\mathtt{G(a^1,b), S(Y7^1,c), O(d) \longrightarrow G(a^2,b^1), S(Y7^2,c), O(d^1)}$

cr21: $\mathtt{G(a^1,b^-), S(Y7^1,c) \longrightarrow G(a^1,b), S(Y7^1,c)}$

cr22: $\mathtt{G(a^1,b), S(Y7^1,c^-), O(d) \longrightarrow G(a^2,b^1), S(Y7^2,c^-), O(d^1)}$

cr23: $\mathtt{G(a^1,b^-), S(Y7^1,c^-) \longrightarrow G(a^1,b), S(Y7^1,c^-)}$

cr24: $\mathtt{R(Y48_{\mathit{p}}), S(Y7^?_{\mathit{u}},c) \longrightarrow R(Y48^1_{\mathit{p}}), S(Y7^?_{\mathit{u}},c^1)}$

cr25: $\mathtt{S(Y7^?_{\mathit{u}},c^-) \longrightarrow S(Y7^?_{\mathit{u}},c)}$

cr26: $\mathtt{R(Y48_{\mathit{p}}), S(Y7_{\mathit{p}},c) \longrightarrow R(Y48^1_{\mathit{p}}), S(Y7_{\mathit{p}},c^1)}$

cr27: $\mathtt{S(Y7_{\mathit{p}},c^-) \longrightarrow S(Y7_{\mathit{p}},c)}$

cr28: $\mathtt{G(a^1,b), R(Y48_{\mathit{p}}), S(Y7^1,c) \longrightarrow G(a^2,b), R(Y48^1_{\mathit{p}}), S(Y7^2,c^1)}$

cr29: $\mathtt{G(a^1,b), S(Y7^1,c^-) \longrightarrow G(a^1,b), S(Y7^1,c)}$

cr30: $\mathtt{G(a^1,b^-), R(Y48_{\mathit{p}}), S(Y7^1,c) \longrightarrow G(a^2,b^-), R(Y48^1_{\mathit{p}}), S(Y7^2,c^1)}$

cr31: $\mathtt{G(a^1,b^-), S(Y7^1,c^-) \longrightarrow G(a^1,b^-), S(Y7^1,c)}$

cr32: $\mathtt{G(a,b), S(Y7_{\mathit{p}},c^-) \longrightarrow G(a^1,b), S(Y7^1_{\mathit{p}},c^-)}$

cr33: $\mathtt{G(a^1,b), S(Y7^1,c^-) \longrightarrow G(a,b), S(Y7,c^-)}$

cr34: $\mathtt{G(a,b), S(Y7_{\mathit{p}},c) \longrightarrow G(a^1,b), S(Y7^1_{\mathit{p}},c)}$

cr35: $\mathtt{G(a^1,b), S(Y7^1,c) \longrightarrow G(a,b), S(Y7,c)}$

cr36: $\mathtt{G(a,b^-), S(Y7_{\mathit{p}},c) \longrightarrow G(a^1,b^-), S(Y7^1_{\mathit{p}},c)}$

cr37: $\mathtt{G(a^1,b^-), S(Y7^1,c) \longrightarrow G(a,b^-), S(Y7,c)}$

cr38: $\mathtt{G(a,b^-), S(Y7_{\mathit{p}},c^-) \longrightarrow G(a^1,b^-), S(Y7^1_{\mathit{p}},c^-)}$

cr39: $\mathtt{G(a^1,b^-), S(Y7^1,c^-) \longrightarrow G(a,b^-), S(Y7,c^-)}$

5.3 Fragments.

In this section we list the 38 self-consistent coarse-grained variables generated by the automatic procedure outlined in the main text and in section 6. These 38 variables form a dynamical system (shown in section 7) whose state at any time t is identical to the state attained by the microscopic dynamics (involving 356 variables) and subsequent coarse-graining.

$\mathcal{F}_1$: $\mathtt{E(r^1), R(Y48^2_{\mathit{p}},l^1,r^{R@r}), S(Y7^3_{\mathit{p}},c^2), G(a^3,b^4), O(d^4)}$

$\mathcal{F}_2$: $\mathtt{E(r^1), R(Y48^2_{\mathit{p}},l^1,r), S(Y7^3_{\mathit{p}},c^2), G(a^3,b^4), O(d^4)}$

$\mathcal{F}_3$: $\mathtt{G(a^2,b^1), S(Y7^2_{\mathit{p}},c^3), R(Y48^3_{\mathit{p}},l,r), O(d^1)}$

$\mathcal{F}_4$: $\mathtt{R(Y48^1_{\mathit{p}},l,r), S(Y7_{\mathit{p}},c^1)}$

$\mathcal{F}_5$: $\mathtt{G(a,b^1), O(d^1)}$

$\mathcal{F}_6$: $\mathtt{E(r^1), R(Y48^2_{\mathit{p}},l^1,r), S(Y7_{\mathit{p}},c^2)}$

$\mathcal{F}_7$: $\mathtt{E(r^1), R(Y48^2_{\mathit{p}},l^1,r^{R@r}), S(Y7_{\mathit{p}},c^2)}$

$\mathcal{F}_8$: $\mathtt{G(a^2,b^1), S(Y7^2_{\mathit{p}},c), O(d^1)}$

$\mathcal{F}_9$: $\mathtt{S(Y7_{\mathit{p}},c)}$

$\mathcal{F}_{10}$: $\mathtt{G(a^1,b), S(Y7^1_{\mathit{p}},c)}$

$\mathcal{F}_{11}$: $\mathtt{G(a,b)}$

$\mathcal{F}_{12}$: $\mathtt{E(r^1), R(Y48^2_{\mathit{p}},l^1,r^{R@r}), S(Y7^3_{\mathit{p}},c^2), G(a^3,b)}$

$\mathcal{F}_{13}$: $\mathtt{E(r^1), R(Y48^2_{\mathit{p}},l^1,r), S(Y7^3_{\mathit{p}},c^2), G(a^3,b)}$

$\mathcal{F}_{14}$: $\mathtt{G(a^1,b), S(Y7^1_{\mathit{p}},c^2), R(Y48^2_{\mathit{p}},l,r)}$

$\mathcal{F}_{15}$: $\mathtt{R(Y48_{\mathit{p}},l,r)}$

$\mathcal{F}_{16}$: $\mathtt{E(r^1), R(Y48_{\mathit{p}},l^1,r)}$

$\mathcal{F}_{17}$: $\mathtt{E(r^1), R(Y48_{\mathit{p}},l^1,r^{R@r})}$

$\mathcal{F}_{18}$: $\mathtt{E(r^1), R(Y48^2_{\mathit{p}},l^1,r^{R@r}), S(Y7_{\mathit{u}},c^2)}$

$\mathcal{F}_{19}$: $\mathtt{E(r^1), R(Y48^2_{\mathit{p}},l^1,r), S(Y7_{\mathit{u}},c^2)}$

$\mathcal{F}_{20}$: $\mathtt{R(Y48^1_{\mathit{p}},l,r), S(Y7_{\mathit{u}},c^1)}$

$\mathcal{F}_{21}$: $\mathtt{S(Y7_{\mathit{u}},c)}$

$\mathcal{F}_{22}$: $\mathtt{O(d)}$

$\mathcal{F}_{23}$: $\mathtt{E(r^1), R(Y68^2_{\mathit{p}},l^1,r^{R@r}), G(a^2,b^3), O(d^3)}$

$\mathcal{F}_{24}$: $\mathtt{E(r^1), R(Y68^2_{\mathit{p}},l^1,r), G(a^2,b^3), O(d^3)}$

$\mathcal{F}_{25}$: $\mathtt{G}\big(\mathtt{a}^{2},\mathtt{b}^{1}\big),\mathtt{R}\big(\mathtt{Y68}^{2}_{p},\mathtt{l},\mathtt{r}\big),\mathtt{O}\big(\mathtt{d}^{1}\big)$

$\mathcal{F}_{26}$: $\mathtt{G}\big(\mathtt{a}^{1},\mathtt{b}\big),\mathtt{R}\big(\mathtt{Y68}^{1}_{p},\mathtt{l},\mathtt{r}\big)$

$\mathcal{F}_{27}$: $\mathtt{E}\big(\mathtt{r}^{1}\big),\mathtt{R}\big(\mathtt{Y68}^{2}_{p},\mathtt{l}^{1},\mathtt{r}\big),\mathtt{G}\big(\mathtt{a}^{2},\mathtt{b}\big)$

$\mathcal{F}_{28}$: $\mathtt{E}\big(\mathtt{r}^{1}\big),\mathtt{R}\big(\mathtt{Y68}^{2}_{p},\mathtt{l}^{1},\mathtt{r}^{\mathtt{R@r}}\big),\mathtt{G}\big(\mathtt{a}^{2},\mathtt{b}\big)$

$\mathcal{F}_{29}$: $\mathtt{R}\big(\mathtt{Y68}_{p},\mathtt{l},\mathtt{r}\big)$

$\mathcal{F}_{30}$: $\mathtt{E}\big(\mathtt{r}^{1}\big),\mathtt{R}\big(\mathtt{Y68}_{p},\mathtt{l}^{1},\mathtt{r}\big)$

$\mathcal{F}_{31}$: $\mathtt{E}\big(\mathtt{r}^{1}\big),\mathtt{R}\big(\mathtt{Y68}_{p},\mathtt{l}^{1},\mathtt{r}^{\mathtt{R@r}}\big)$

$\mathcal{F}_{32}$: $\mathtt{E}\big(\mathtt{r}^{1}\big),\mathtt{R}\big(\mathtt{Y48}_{u},\mathtt{l}^{1},\mathtt{r}^{\mathtt{R@r}}\big)$

$\mathcal{F}_{33}$: $\mathtt{E}\big(\mathtt{r}^{1}\big),\mathtt{R}\big(\mathtt{Y48}_{u},\mathtt{l}^{1},\mathtt{r}\big)$

$\mathcal{F}_{34}$: $\mathtt{R}\big(\mathtt{Y48}_{u},\mathtt{l},\mathtt{r}\big)$

$\mathcal{F}_{35}$: $\mathtt{E}\big(\mathtt{r}^{1}\big),\mathtt{R}\big(\mathtt{Y68}_{u},\mathtt{l}^{1},\mathtt{r}^{\mathtt{R@r}}\big)$

$\mathcal{F}_{36}$: $\mathtt{E}\big(\mathtt{r}^{1}\big),\mathtt{R}\big(\mathtt{Y68}_{u},\mathtt{l}^{1},\mathtt{r}\big)$

$\mathcal{F}_{37}$: $\mathtt{R}\big(\mathtt{Y68}_{u},\mathtt{l},\mathtt{r}\big)$

$\mathcal{F}_{38}$: $\mathtt{E}\big(\mathtt{r}\big)$

6 Translating rules into a dynamical system for fragments

6.1 Partial Complex. A partial complex is a connected graph (a component) that occurs on either side of a rule. In our static analysis, semi-links in partial complexes are internally expanded into all possible binding partners, and labeled with a bond type of the form *partner@site*. For example, $\mathtt{R(Y48}^{1}_{p}\mathtt{),S(Y7}^{2}\mathtt{,c}^{1}\mathtt{),G(a}^{2}\mathtt{,b}^{O@d}\mathtt{)}$ is a partial complex (it is the right hand side of cr30). Thus:

Definition 6.1 (Partial complex)**.**

A partially specified complex (or partial complex for short) is a connected expression, such that

1. the set of sites shown for agent A is a subset of the interface of A
2. the internal state of a site may be omitted
3. the binding state of a site must be any of (i) free, (ii) bound, or (iii) a bond stub indicating the names of the bound agent and its binding site

We extend the concept of a match, $E' \vDash E$, Definition 1.5, or the concept of an embedding $G \lhd_\phi G'$ (see end of section 1.4), to expressions containing a stub by simply extending the specificity ranking of binding states (section 1.4) in the obvious way.

Fragments, as defined in the main text, are partial complexes, too, but whose shape is constrained by the annotated contact map (ACM). Our goal in this section is to sketch the construction of the kinetic system of differential equations describing the concentration dynamics for fragments. To this end, we must evaluate how each rule in a model contributes to the production and consumption of fragments. For a rule to be translatable into a set of reactions between fragments, we must ensure that any fragment that properly intersects a component on the lhs of a rule, and whose intersection contains a site that is modified by the action, must contain that component. This is achieved by the syntactical criteria, Cov1-Cov3 and Edg1, as explained in the main text.

6.2 Every lhs component of a rule is contained in some fragment. Directives Cov1-Cov3 ensure that there always is a class in the covering of an agent that contains at least as many sites as any occurrence of that agent on the left of a rule. The fragment growth process (see main text, subsection "Fragment assembly") then guarantees that fragments extend at least as much as any lhs component in a rule, except for pure dissociation rules. By construction, fragments do not contain bonds that are soft in the ACM, and thus do not extend components with such bonds. However, when such a component is cut at a soft bond, and the bond is replaced with two stubs, each piece can be embedded in a fragment. This reflects the fact that the system of rules cannot detect any correlations between such pieces (or the bond would be solid). The construction of fragments also ensures that when a lhs component Z contains a site that is modified, Z is contained in each fragment exhibiting that site.

6.3 Expressing the concentration of a subfragment in terms of fragment concentrations. Before proceeding we need some clarification on the notion of "concentration" in a formal rule-based approach that deals with patterns (partial complexes). What is the concentration of a partial complex in a mixture of fully specified complexes (species)?

Counting embeddings. In a stochastic setting, in which a reaction mixture x is a multi-set of molecular species, one should distinguish between two quantities for each partial complex Z: the number of embeddings ϕ (section 1.4) of Z into x, written $|\{\phi \mid Z \lhd_\phi x\}|$, and the number of embeddings corrected by the number of automorphisms of Z, $\mathit{auto}(Z) = |\{\phi' \mid Z \lhd_{\phi'} Z\}|$ (since Z may have symmetries):

$$[Z] := |\{\phi \mid Z \lhd_\phi x\}|/\mathit{auto}(Z)$$

In the case where the partial complex Z is actually a (fully specified) species, the corrected number of embeddings is the number of *occurrences* of Z in x, which yields the concentration, when divided by the volume, as in ODEs.

Given two partial complexes Z and Z' such that Z embeds into Z', the set $\{\phi \mid Z \lhd_\phi Z'\}$ of embeddings can be quotiented by the equivalence relation $\sim$, relating any pair of embeddings ϕ and ϕ' such that ϕ' can be written as $\phi \circ \sigma$ for σ an automorphism of Z'. Again, we have two choices; either we count embeddings, or embeddings up to $\sim$. We write $Z \unlhd_\phi Z'$ when ϕ is an embedding equivalence class and we have therefore:

$$|\{\phi \mid Z \lhd_\phi Z'\}| = |\{\phi \mid Z \unlhd_\phi Z'\}| \times \mathit{auto}(Z')$$

Orthogonal fragments. A subfragment is a partial complex (subsection 6.1) that embeds in a fragment. It is an important "technical object" in our method. It shows up when we compute production rates for fragments whose concentration is affected by the dissociation of a solid bond Z–Z'. Such a dissociation will give rise to a piece Z (and also Z') that might embed into a fragment $\mathcal{F}$. As the lhs component of a rule, Z–Z'

embeds into a fragment and so will Z by virtue of the fragment growth process (see main text). To determine the contribution of the dissociation rule to the production rate of $\mathcal{F}$, we need the concentration of $\mathcal{F}$–Z'. Yet, $\mathcal{F}$–Z' is not itself a fragment, but rather a subfragment. For our method to result in a closed system of equations, we must be able to express the concentration of this subfragment in terms of fragments. It turns out that this is indeed the case for any subfragment, as we show next. (We cannot, in general, extend a fragment and express its concentration using other fragments, as this would require the independence conditions, equation **[10]**, to hold. As we saw in Example 2, these conditions do not hold in general.)

To compute the concentration of a subfragment, we need to use fragments that extend it. But we must be careful about which fragments we use. A subfragment and a fragment, each identifies a set of fully specified molecular species into which they embed (i.e. their extension, as explained in the main text, section "From rules to ODEs"). The concentration of a subfragment is the sum total of the concentrations of these species (weighted by appropriate symmetry related constants). If we are to use a combination of fragment concentrations, we must ensure that the fragments used indeed partition the set of molecular species into which the subfragment expands, or we would overcount. A set of fragments that complies with this requirement is called *orthogonal*.

Let Z be a subfragment, $\mathfrak{F}$ the set of fragments, and $\mathcal{F}_1, \mathcal{F}_2 \in \mathfrak{F}$ two fragments that contain Z: $Z \lhd_\phi \mathcal{F}_1$ and $Z \lhd_{\phi'} \mathcal{F}_2$. (See section 1.4, paragraph labeled "embedding", for a definition of the embedding relation $\lhd_\phi$.) We define two fragments $\mathcal{F}_1, \mathcal{F}_2$ that contain Z as "orthogonal", when the agents on which $\mathcal{F}_1$ and $\mathcal{F}_2$ agree exhibit the same sites. Orthogonal fragments differ with regard to internal states and binding states at sites of agents they have in common, and, thus, constitute a set of patterns whose matching instances in a reaction mixture do *not* overlap. This is important for expressing the concentration of the subfragment Z in terms of fragments, since we must avoid double counting matching instances of Z in the reaction mixture. The set of fragments $\{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ from which we compute the concentration of Z should constitute a refinement of Z, in the sense of partitioning the matching instances of Z.

The formal definition of orthogonality makes use of the concept of a formal path. A formal path p is a set of symbolic instructions for navigating through a graph representing a Kappa complex. Starting at node (agent) A and following the directives provided by p will lead us to a unique target node T, which we denote by A.p. The path p is expressed as a sequence of bonds to travel. For example, A.p might result in $\mathtt{A.a.s_1.B_1.s_2\ldots.s_{2i-1}.B_i.s_{2i}\ldots..t.T}$ (with $\mathtt{s}_{2i-1}\neq\mathtt{s}_{2i}$), which is a path that goes from the originating agent A to agent $\mathtt{B_1}$ over a link between site a of A and site $\mathtt{s_1}$ of agent $\mathtt{B_1}$, and from there to agent $\mathtt{B_2}$ over a link between site $\mathtt{s_2}$ of agent $\mathtt{B_1}$ and site $\mathtt{s_3}$ of agent $\mathtt{B_2}$, and so on, to enter the target agent T through its site t. If any step some site is missing, then $A.p$ is not defined in the complex in question. An empty path means that we stay at the originating agent A.

Definition 6.2 (Orthogonal fragments)**.**

Let Z be a subfragment. Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two fragments, and ϕ, ϕ' be embedding classes such that $Z \unlhd_\phi \mathcal{F}_1$ and $Z \unlhd_{\phi'} \mathcal{F}_2$. We say $(\mathcal{F}_1, \phi)$ and $(\mathcal{F}_2, \phi')$ are Z-orthogonal, written as $(\mathcal{F}_1, \phi) \bowtie_Z (\mathcal{F}_2, \phi')$, if and only if for any agent A of Z and any path p at least one of these statements is true:

1. $\phi(A).p$ is not defined in $\mathcal{F}_1$.
2. $\phi'(A).p$ is not defined in $\mathcal{F}_2$.
3. $\phi(A).p$ and $\phi'(A).p$ are both defined in $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively, and have the same set of sites.

A fragment $\mathcal{F}_i$ may be matched in more than one way by a subfragment Z, yielding more than one embedding class ϕ. Let us collect all embedding classes mapping Z to some fragment, and define $C(Z)$ as a largest set of such embedding classes (it need not be unique) that are mutually Z-orthogonal. Write n_i for the number of embedding classes ϕ in $C(Z)$ with target $\mathcal{F}_i$, ie such that $Z \unlhd_\phi \mathcal{F}_i$.

The concentration of Z can now be expressed as:

$$[Z] = \frac{1}{|\{\phi | Z \lhd_\phi Z\}|} \sum_i n_i [\mathcal{F}_i]\, auto(\mathcal{F}_i) \qquad \mathbf{[17]}$$

Equation **[17]** formalizes our intuition that the concentration of a partial complex Z is the sum of the concentrations of the fragments that contain it, times a multiplicity counting the number of ways in which Z matches a given fragment. The only complexity comes from chosing the fragments over which the sum runs in such a way that no two fragments overlap in the set of molecular instances they match. That is what the orthogonality criterion is meant to ensure. Finally, we divide by the number of automorphisms of Z to compensate for any symmetries in Z.

6.4 Assembling the dynamical system for fragments – Version A. We assemble the system of differential equations for the fragments by determining the mass action terms that each rule type contributes. The description offered in this subsection has a deliberately "algorithmic" flavor. In the next subsection, we offer a version B that offers a more concise and abstract presentation which the reader might also find useful.

For the sake of simplifying exposition, we shall only be concerned with rules consisting of at most two components on the lhs, and whose action modifies a single internal state or a single binding state. It is straightforward to generalize this to multiple components and to multiple actions within the same rule. The k after the @-sign refers to the rate constant of the rule. Since we shall build a term for the fragment dynamics from each component on the lhs of a rule separately, the rate constant γ that enters the fragment dynamics must compensate for the number of automorphisms, *auto(lhs)*, of the lhs: $\gamma = k/$*auto(lhs)*. Thanks to **[17]** we can refer to the concentration of any subfragment, $[Z]$, and be sure we can replace with an expansion into fragment concentrations. To indicate that we are building up differential equations sequentially, we shall use the symbol $\stackrel{\pm}{=}$ as meaning "add this term to the previous ones for this equation".

$\mathbf{Z}, \mathbf{Z}' \longrightarrow \mathbf{Z}^*, \mathbf{Z}'\,@\,\mathbf{k}$

This type of rule modifies the partial complex Z.

Consumption terms. The kinetic equation of each fragment that contains Z gains a consumption term

$$\forall \mathcal{F}_i\ \forall \phi \text{ s.t. } Z \unlhd_\phi \mathcal{F}_i\ :\qquad \frac{d[\mathcal{F}_i]}{dt} \stackrel{\pm}{=} -\gamma([\mathcal{F}_i]auto(\mathcal{F}_i))([Z']auto(Z')).$$

The universal quantifier over ϕ means that the rate at which $\mathcal{F}_i$ is consumed depends on the number of ways that Z can be embedded in $\mathcal{F}_i$. The said quantification, here as well as in the subsequent cases, is over embedding classes, not over plain embeddings.

Production terms. The kinetic equation of each fragment containing Z^* gains a production term

$$\forall \mathcal{F}_k \ \forall \phi \text{ s.t. } Z^* \trianglelefteq_\phi \mathcal{F}_k \text{ and } \forall \mathcal{F}_i \text{ s.t. } Z \trianglelefteq_{\phi^\star} \mathcal{F}_i \text{ and } \mathcal{F}_k = \mathcal{F}_i^* \ : \quad \frac{d[\mathcal{F}_k]}{dt} \stackrel{+}{=} \gamma([\mathcal{F}_i]\mathit{auto}(\mathcal{F}_i))([Z']\mathit{auto}(Z')).$$

Clearly, the fragments $\mathcal{F}_k$ and $\mathcal{F}_i$ must be related by the rule action; $\mathcal{F}_k = \mathcal{F}_i^*$ means that the fragment $\mathcal{F}_k$ is obtained by applying the rule action to fragment $\mathcal{F}_i$. The notation $\phi^\star$ indexing the embedding of Z into $\mathcal{F}_i$ is meant to specify that this embedding is related to the embedding ϕ of Z^* into $\mathcal{F}_k$, since the relatedness of Z to its modified form Z^* forces not only a relatedness of $\mathcal{F}_i$ to $\mathcal{F}_k$, but also of the way these fragments extend the corresponding partial complexes Z and Z^*.

$\mathbf{Z}, \mathbf{Z}' \longrightarrow \mathbf{Z}\text{–}\mathbf{Z}' \,@\, \mathbf{k}$

This type of rule binds the partial complexes Z and Z'.

Consumption terms: The kinetic equation of each fragment that contains Z gains a consumption term

$$\forall \mathcal{F}_i \ \forall \phi \text{ s.t. } Z \trianglelefteq_\phi \mathcal{F}_i \ : \quad \frac{d[\mathcal{F}_i]}{dt} \stackrel{+}{=} -\gamma([\mathcal{F}_i]\mathit{auto}(\mathcal{F}_i))([Z']\mathit{auto}(Z')).$$

Likewise for Z':

$$\forall \mathcal{F}_i \ \forall \phi \text{ s.t. } Z' \trianglelefteq_\phi \mathcal{F}_i \ : \quad \frac{d[\mathcal{F}_i]}{dt} \stackrel{+}{=} -\gamma([\mathcal{F}_i]\mathit{auto}(\mathcal{F}_i))([Z]\mathit{auto}(Z)).$$

Production terms. On the production side, we must distinguish between solid and soft links in the ACM.

$\mathbf{Z}\text{–}\mathbf{Z}'$ solid link

$$\forall \mathcal{F}_k \ \forall \phi_1, \phi_2 \text{ s.t. } Z\text{–}Z' \trianglelefteq_{\phi_1 \uplus \phi_2} \mathcal{F}_k \text{ and } \forall \mathcal{F}_i \text{ s.t. } Z \trianglelefteq_{\phi_1^\star} \mathcal{F}_i \text{ and } \forall \mathcal{F}_j \text{ s.t. } Z' \trianglelefteq_{\phi_2^\star} \mathcal{F}_j \text{ and } \mathcal{F}_k = \mathcal{F}_i\text{–}\mathcal{F}_j \ :$$

$$\frac{d[\mathcal{F}_k]}{dt} \stackrel{+}{=} \gamma([\mathcal{F}_i]\mathit{auto}(\mathcal{F}_i))([\mathcal{F}_j]\mathit{auto}(\mathcal{F}_j)).$$

Again, we must express that the embeddings of $Z\text{–}Z'$, Z, and Z' into $\mathcal{F}_k$, $\mathcal{F}_i$, and $\mathcal{F}_j$, respectively, are related. The notation $\phi_1 \uplus \phi_2$ denotes the disjoint sum of ϕ_1 and ϕ_2: The domains of ϕ_1, $dom(\phi_1)$, and ϕ_2, $dom(\phi_2)$, have an empty intersection, and $(\phi_1 \uplus \phi_2)(x) = \phi_1(x)$ if $x \in dom(\phi_1)$ or $(\phi_1 \uplus \phi_2)(x) = \phi_2(x)$ if $x \in dom(\phi_2)$.

$\mathbf{Z}\text{–}\mathbf{Z}'$ soft link

Assume the bond to be between site a of $A \in Z$ and site b of $B \in Z'$, and let $Z^{B@b}$ and $Z'^{A@a}$ denote the partial complexes obtained from severing the bond in $Z\text{–}Z'$ and replacing it with a binding-type label.

$$\forall \mathcal{F}_k \ \forall \phi \text{ s.t. } Z^{B@b} \trianglelefteq_\phi \mathcal{F}_k \text{ and } \forall \mathcal{F}_i \text{ s.t. } Z \trianglelefteq_{\phi^\star} \mathcal{F}_i \text{ and } \mathcal{F}_k = \mathcal{F}_i^{B@b} \ : \quad \frac{d[\mathcal{F}_k]}{dt} \stackrel{+}{=} \gamma([\mathcal{F}_i]\mathit{auto}(\mathcal{F}_i))([Z']\mathit{auto}(Z')).$$

Likewise for Z':

$$\forall \mathcal{F}_k \ \forall \phi \text{ s.t. } Z'^{A@a} \trianglelefteq_\phi \mathcal{F}_k \text{ and } \forall \mathcal{F}_i \text{ s.t. } Z' \trianglelefteq_{\phi^\star} \mathcal{F}_i \text{ and } \mathcal{F}_k = \mathcal{F}_i^{A@a} \ : \quad \frac{d[\mathcal{F}_k]}{dt} \stackrel{+}{=} \gamma([\mathcal{F}_i]\mathit{auto}(\mathcal{F}_i))([Z]\mathit{auto}(Z)).$$

$\mathbf{Z}\text{–}\mathbf{Z}' \longrightarrow \mathbf{Z}, \mathbf{Z}' \,@\, \mathbf{k}$

This type of rule dissociates the partial complex $Z\text{–}Z'$.

Consumption terms. On the consumption side, we must distinguish between a solid and a soft link in the ACM.

$\mathbf{Z}\text{–}\mathbf{Z}'$ solid link

The kinetic equation of each fragment that contains $Z\text{–}Z'$ gains a consumption term

$$\forall \mathcal{F}_i \ \forall \phi \text{ s.t. } Z\text{–}Z' \trianglelefteq_\phi \mathcal{F}_i \ : \quad \frac{d[\mathcal{F}_i]}{dt} \stackrel{+}{=} -\gamma([\mathcal{F}_i]\mathit{auto}(\mathcal{F}_i)).$$

$\mathbf{Z}\text{–}\mathbf{Z}'$ soft link (By definition, this is a "pure dissociation rule".)

As above, assume the bond to be between site a of $A \in Z$ and site b of $B \in Z'$, and let $Z^{B@b}$ and $Z'^{A@a}$ denote the partial complexes obtained from severing the bond in $Z\text{–}Z'$ and replacing it with a bond type label.

$$\forall \mathcal{F}_i \ \forall \phi \text{ s.t. } Z^{B@b} \trianglelefteq_\phi \mathcal{F}_i \ : \quad \frac{d[\mathcal{F}_i]}{dt} \stackrel{+}{=} -\gamma([\mathcal{F}_i]\mathit{auto}(\mathcal{F}_i)).$$

Likewise for Z':

$$\forall \mathcal{F}_i \ \forall \phi \text{ s.t. } Z'^{A@a} \trianglelefteq_\phi \mathcal{F}_i \ : \quad \frac{d[\mathcal{F}_i]}{dt} \stackrel{+}{=} -\gamma([\mathcal{F}_i]\mathit{auto}(\mathcal{F}_i)).$$

Production terms. Here, too, we must distinguish between solid and soft links.

Z–Z′ solid link

$$\forall \mathcal{F}_i\ \forall \phi_Z \text{ s.t. } Z \trianglelefteq_{\phi_Z} \mathcal{F}_i \text{ and } \forall \mathcal{F}_k\ \forall \phi_{Z'} \text{ s.t. } \mathcal{F}_i\text{–}Z' \trianglelefteq_{\phi^\star_Z \uplus \phi_{Z'}} \mathcal{F}_k \text{ and } (\mathcal{F}_k, \phi^\star_Z \uplus \phi_{Z'}) \in C(\mathcal{F}_i\text{–}Z')\ :$$

$$\frac{d[\mathcal{F}_i]}{dt} \overset{+}{=} \gamma([\mathcal{F}_k]\mathit{auto}(\mathcal{F}_k)).$$

The above might benefit from a verbal expansion. In the reaction type $Z\text{–}Z' \rightarrow Z, Z'$, pick a fragment $\mathcal{F}_i$ that extends the partial complex Z in a particular way (that's an instance of the first two universal quantifiers). The production rate of $\mathcal{F}_i$ will be first order in a fragment $\mathcal{F}_k$ that extends the single partial complex on the left hand side of the rule, $Z\text{–}Z'$. Yet, the $\mathcal{F}_k$ in question cannot extend any old $Z\text{–}Z'$, but must extend an instance that contains the $\mathcal{F}_i$ that will emerge after the bond is split. Hence the condition that $\mathcal{F}_i\text{–}Z' \trianglelefteq_{\phi^\star_Z \uplus \phi_{Z'}} \mathcal{F}_k$. The injection map associated with this embedding must be constrained by how we chose $\mathcal{F}_i$ to extend Z. Finally, all $\mathcal{F}_k$ that contribute to the production of a given $\mathcal{F}_i$ must be mutually orthogonal with respect to $\mathcal{F}_i\text{–}Z'$, as defined in section 6.3, to avoid multiple-counting the molecular species into which the $\mathcal{F}_k$ expand.

Analogous production terms arise for fragments that extend Z' on the right hand side of the rule.

Z–Z′ soft link

As above, assume the bond to be between site a of $A \in Z$ and site b of $B \in Z'$, and let $Z^{B@b}$ and $Z'^{A@a}$ denote the partial complexes obtained from severing the bond in $Z\text{–}Z'$ and replacing it with a bond type label.

$$\forall \mathcal{F}_k\ \forall \phi \text{ s.t. } Z^{B@b} \trianglelefteq_{\phi} \mathcal{F}_k \text{ and } \forall \mathcal{F}_i \text{ s.t. } Z \trianglelefteq_{\phi^\star} \mathcal{F}_i \text{ and } \mathcal{F}_k = \mathcal{F}_i^{B@b}\ :\quad \frac{d[\mathcal{F}_i]}{dt} \overset{+}{=} \gamma([\mathcal{F}_k]\mathit{auto}(\mathcal{F}_k)).$$

Likewise for Z':

$$\forall \mathcal{F}_k\ \forall \phi \text{ s.t. } Z'^{A@a} \trianglelefteq_{\phi} \mathcal{F}_k \text{ and } \forall \mathcal{F}_i \text{ s.t. } Z' \trianglelefteq_{\phi^\star} \mathcal{F}_i \text{ and } \mathcal{F}_k = \mathcal{F}_i^{A@a}\ :\quad \frac{d[\mathcal{F}_i]}{dt} \overset{+}{=} \gamma([\mathcal{F}_k]\mathit{auto}(\mathcal{F}_k)).$$

Our implementation is actually more straightforward and uniform than this enumeration suggests, because the algorithm makes direct use of the embeddings ϕ, which remain abstract in a notation that does not exploit the structure of expressions. Version B to which to which we turn now does present things in a more uniform way.

6.5 Assembling the dynamical system for fragments – Version B (not for the faint of heart). In this section we show more abstractly how to write the system of ODEs governing the time evolution of ACM-based fragments. To simplify things and expedite the presentation, we suppose that all bonds are solid; we also suppose that rule actions involve no agent deletion or creation, i.e. all actions are reversible, and we write α^{-1} for the action inverse to α. We will also use a reinforced version of property Q1 (defined in the main text); see below. We only aim at giving a sense of the general construction, not its full development and justification, which will be detailed elsewhere.

Given a global state of the system x, a rule r, and a fragment $\mathcal{F}$, we want to express how r affects the concentration of $\mathcal{F}$. Specifically, we want to express the negative (consumption) and positive (production) terms coming from r as functions of fragment concentrations (self-consistency). The overall differential equation for $\mathcal{F}$ is then obtained by summing the contributions of all rules to $\mathcal{F}$ (see below for a complete example).

In this section it is crucial to keep in mind the distinction made in the subsection above on "Counting embeddings". As a reminder, there are two possibilities for defining the "concentration" of $\mathcal{F}$ (or any complex) in a mixture x: the number of embeddings of $\mathcal{F}$ in the reaction mixture, which we denote by $|[\mathcal{F}; x]|$, and the *discounted concentration* $|[\mathcal{F}; x]|/|[\mathcal{F}; \mathcal{F}]|$ familiar from deterministic chemical kinetics. In this section we shall use exclusively the number of embeddings and denote it with $[\mathcal{F}; x]$ (henceforth omitting the $|.|$ for cardinals), being aware that we are abusing notation (and we shall do so even more below).

We next list a few conventions about embeddings, and rules.

- Given s, x two Kappa expressions – that is: mixtures (fully or partially specified), complexes (fully or partially specified), rule components, etc. – $[s; x]$ stands for the set of embeddings of s in x;
- We will *not* use the traditional discounted concentrations, $[A; x]/[A; A]$, but rather the number of embeddings $[A; x]$, which we will sometimes write as $[A]$ when x is clear from the context.
- We will use a more convenient notation fro rules where $r = s, \alpha, k$ consists of a left hand side s, an action which can be any combination of (internal state) modification, binding and unbinding, and a rate $k > 0$. This formulation is equivalent to the more intuitive $lhs \rightarrow rhs\,@\,k$ notation that we have used so far.
- Given a rule $r = s, \alpha, k$, a mixture x, the set of events (i.e. rule applications) associated with r in x is by definition in bijective correspondence with $[s; x]$ (we distinguish here between an embedding of s into x and the event that it determines). Given $f \in [s; x]$, we write $f(\alpha) \cdot x$ for the outcome of the r-event associated with f.
- We write $s_1 \subseteq_\star s$ when s_1 is a tuple of components of s that is modified by α.
- The rule *activity* is plainly $k[s; x]$ (ie we do not divide the rate by $[s; s]$ in this context which avoids the γ of the first presentation above). This rule activity is the expected number of applications of r per time unit, i.e. the flux of r in the ODE limit.

In general the rate at which a rule $r = s, \alpha, k$ consumes/produces embeddings of $\mathcal{F}$ is:

$$\delta_r(\mathcal{F}) = k \textstyle\sum_{f \in [s;x]} \left([\mathcal{F}; f(\alpha) \cdot x] - [\mathcal{F}; x]\right)$$

Indeed, the activity of r can be written $k \sum_{f \in [s;x]} 1$ and is the expected number of applications of r per time unit. The above formula expresses the expected change in the number of embeddings of $\mathcal{F}$ due to the r-events occurring (synchronously) per time unit. Note that a rule can both consume and produce $\mathcal{F}$ in the same event. We will evaluate both contributions of r to $[F; x]$, written as $\delta_r^-(F)$ and $\delta_r^+(F)$, separately.

To evaluate the activity of a rule, we will assume, as is customary in deterministic chemical kinetics, a negligible occurrence of situations in which the embeddings of lhs components are not jointly injective. In other words, if Z and Z' are lhs rule components, we will overestimate $[Z, Z'; x]$ as $[Z; x][Z'; x]$.

Consumption terms

Choose an $f \in [s;x]$. For f to consume $\mathcal{F}$, there must be a modified lhs component s_1 whose image under f intersects an occurrence of $\mathcal{F}$ in x on modified sites. By Q1 (see main text) f factorizes as $f = \gamma(\phi + I)$, with I the identity on $s \smallsetminus s_1$, $\phi \in [s_1;\mathcal{F}]$, and $\gamma \in [\mathcal{F}, s \smallsetminus s_1; x]$.

We can summarise this factorization as follows:

$$\begin{tikzcd} s_1, s \smallsetminus s_1 \arrow[r, "\phi+I"] & \mathcal{F}, s \smallsetminus s_1 \arrow[r, "\gamma"] & x \end{tikzcd}$$

If we overapproximate $[\mathcal{F}, s \smallsetminus s_1; x]$ as $[\mathcal{F};x][s \smallsetminus s_1; x]$, we get the following bijective enumeration of the $\mathcal{F}$-consuming events associated with r:

$$\forall s_1 \subseteq_\star s,\ \forall \phi \in [s_1;\mathcal{F}] :\ \delta^-_{r,\phi}(\mathcal{F}) = k[\mathcal{F}, s \smallsetminus s_1] = k[\mathcal{F}][s \smallsetminus s_1] \qquad \textbf{[18]}$$

By Q2-3 (see main text), each contribution can be expressed as a function of fragment concentrations.

Given f, the ϕ part of the factorisation, and therefore also the γ part, is only determined up to a symmetry $\sigma \in [\mathcal{F};\mathcal{F}]$. Indeed, if $\gamma\phi = f$, then $\gamma\sigma\sigma^{-1}\phi = f$. To avoid a redundant count, it appears that one should divide the above terms by $[\mathcal{F};\mathcal{F}]$. However, each factorisation consumes $[\mathcal{F};\mathcal{F}]$ embeddings of $\mathcal{F}$ in x, and both cancel each other out. Note that we use here a stronger version of Q1, namely that the γ, ϕ decomposition is unique up to symmetries of $\mathcal{F}$ -this is only to make things simpler and is not a requirement of our algorithm.

Production terms

Choose an $f \in [s;x]$, and write $s' := \alpha \cdot s$ for the rule rhs, $x' := f(\alpha) \cdot x$ for the state obtained by triggering the r-event associated with f, and $f' \in [s';x']$ for the unique post-event embedding corresponding to f.

For f to produce an $\mathcal{F}$ there must be a modified rhs component $s'_1 \subseteq s'$ that factorises f' as $f' = \gamma'(\phi' + I)$ with $\phi' \in [s'_1;\mathcal{F}]$, for some γ' uniquely up to $[\mathcal{F};\mathcal{F}]$. We can summarize the situation as follows:

$$\begin{tikzcd} s' = s'_1, s' \smallsetminus s'_1 \arrow[r, "\phi'+I"] & \mathcal{F}, s' \smallsetminus s'_1 \arrow[r, "\gamma'"] \arrow[d, dotted, "(\phi'+I)(\alpha^{-1})"] & x' \\ s \arrow[u, dotted, "\alpha"] \arrow[r, "\phi+I"] & (\phi'+I)(\alpha^{-1})(\mathcal{F}, s' \smallsetminus s'_1) \arrow[r, "\gamma"] & x \arrow[u, dotted, "f(\alpha)"] \end{tikzcd}$$

where the choice of ϕ' uniquely determines ϕ, and therefore γ. Hence, $\mathcal{F}$-producing fs are in bijection with $s'_1 \subseteq_\star \alpha \cdot s$, $\phi' \in [s'_1;\mathcal{F}]$ up to $[\mathcal{F};\mathcal{F}]$, and $\gamma \in [(\phi' + I)(\alpha^{-1}) \cdot (\mathcal{F}, s' \smallsetminus s'_1); x]$.

Putting everything together we get a bijective enumeration of the $\mathcal{F}$-producing events associated with r:

$$\forall s'_1 \subseteq_\star \alpha \cdot s,\ \forall \phi' \in [s'_1;\mathcal{F}] :\ \delta^+_{r,\phi'}(\mathcal{F}) = k[(\phi' + I)(\alpha^{-1}) \cdot (\mathcal{F}, \alpha \cdot s \smallsetminus s'_1)] \qquad \textbf{[19]}$$

As in the consumption case, given f, the ϕ' part of the factorization is only unique up to $[\mathcal{F};\mathcal{F}]$, but since each choice creates $[\mathcal{F};\mathcal{F}]$ new embeddings of $\mathcal{F}$, the terms are correct.

From the definition of fragments (see main text) it easily follows that the components of $(\phi' + I)(\alpha^{-1}) \cdot (\mathcal{F}, \alpha \cdot s \smallsetminus s'_1)$ are subfragments. By Q2-3, this is enough to guarantee that each contribution is expressible solely in terms of fragment concentrations.

A concrete dissociation example

We can illustrate the construction of the production terms with a concrete case of a dissociation rule.

Set $s = \mathtt{A(a^1),B(b^1)}$, with $\alpha \cdot s = s' = \mathtt{A(a),B(b)}$, and:

$$\begin{aligned} \mathcal{F} &= \mathtt{A_1(a,x^1),A_2(x^1,a)} \\ x &= N * (\mathtt{A_1(a,x^1),A_2(x^1,a^2),B(b^2)}), \end{aligned}$$

where the latter expression means that the mixture x consists of N copies of the indicated expression. The indices on the $\mathtt{A}$s distinguish the two occurrences in $\mathcal{F}$. The r-events are of the form $f_i = (\mathtt{A,B} \mapsto \mathtt{A_{2i},B})$ and each produces $\mathcal{F}$ (twice). Specifically, one has $[\mathcal{F};x] = 0$, and $[\mathcal{F}; f_i(\alpha) \cdot x] = 2$. The general formula mentions a s'_1 which can only be $\mathtt{A(a)}$ in this case, so that $\alpha \cdot s \smallsetminus s'_1 = \mathtt{B(b)}$, and a ϕ' which can be either $\phi'_1 := (\mathtt{A} \mapsto \mathtt{A_1})$ or $\phi'_2 := (\mathtt{A} \mapsto \mathtt{A_2})$. By inverting the rule along both extensions $\phi'_i + I$, we get isomorphic partial complexes:

$$\begin{aligned} (\phi'_1 + I)(\alpha^{-1}) \cdot (\mathcal{F}, \mathtt{B(b)}) &= \mathtt{A_1(a^2,x^1),A_2(x^1,a),B(b^2)} \\ (\phi'_2 + I)(\alpha^{-1}) \cdot (\mathcal{F}, \mathtt{B(b)}) &= \mathtt{A_1(a,x^1),A_2(x^1,a^2),B(b^2)} \end{aligned}$$

with equal contributions $\delta^+_{r,\phi'_1}(F) = \delta^+_{r,\phi'_2}(F) = kN$, and we find that the total rate at which embeddings of $\mathcal{F}$ are produced is $2kN$, as it should.

7 The dynamical system for the early EGF model

This section shows the output generated by our automatic procedure for the rule system r01-r39 listed in section 5.1. The results have been obtained entirely by static analysis, as detailed in the main text and section 6 of this Supporting Information. The dynamical system for fragments constitutes an endogenously coarse-grained and self-consistent system that is sound with respect to the microscopic kinetics. Sound means that the outcome is identical whether one first executes the deterministic microscopic kinetics with subsequent coarse-graining or first coarse-grains with subsequent execution of the fragment dynamics. Note that the microscopic system was *never* represented explicitly (and thus never executed). It was only represented implicitly by the system of rules (which were not executed either). Because of the ability to bypass an explicit representation, the causal analysis of microscopic systems involving astronomic numbers of distinct microscopic states (molecular species) becomes possible.

The entire fragmentation of the early EGF example, beginning with the reachability analysis, followed by rule compression, fragmentation, and dynamical system generation took less than 0.2s on a 2GHz Intel Centrino Duo with 1Gb RAM. (It took 0.42s, if we include automatic LaTeX report generation.) The mass action terms for the fragment dynamics resulting from each rule, as well as the fully assembled dynamical system, are listed in the next section.

7.1 List of kinetic terms generated from each rule for each fragment.

We report the kinetic production and consumption terms for each fragment as generated by analysis of the compressed rule system of section 5.2. $\mathcal{R}^{\text{rule}}_{\text{frag}}$ denotes the kinetic terms pertinent to the dynamical equation for the fragment indicated in the subscript, and which result from the rule identified in the superscript. Thus $\mathcal{R}^{39}_{7} = \mathcal{F}_1$ means that our static analysis generates from compressed rule cr39 one unimolecular production term (involving fragment $\mathcal{F}_1$) for fragment $\mathcal{F}_7$. For the sake of a less cluttered presentation, we have set all rate constants to 1. (The right hand side of each $\mathcal{R}^{\text{rule}}_{\text{frag}}$ equation should be multiplied by the rate constant associated with the rule indicated in the superscript.)

Kinetic terms generated from rule cr39:

$\mathcal{R}^{39}_{7} = \mathcal{F}_1$
$\mathcal{R}^{39}_{6} = \mathcal{F}_2$
$\mathcal{R}^{39}_{5} = \mathcal{F}_1 + \mathcal{F}_2 + \mathcal{F}_3$
$\mathcal{R}^{39}_{4} = \mathcal{F}_3$
$\mathcal{R}^{39}_{3} = -\mathcal{F}_3$
$\mathcal{R}^{39}_{2} = -\mathcal{F}_2$
$\mathcal{R}^{39}_{1} = -\mathcal{F}_1$

Kinetic terms generated from rule cr38:

$\mathcal{R}^{38}_{7} = -\mathcal{F}_5 \cdot \mathcal{F}_7$
$\mathcal{R}^{38}_{6} = -\mathcal{F}_5 \cdot \mathcal{F}_6$
$\mathcal{R}^{38}_{5} = -\mathcal{F}_5 \cdot (\mathcal{F}_4 + \mathcal{F}_6 + \mathcal{F}_7)$
$\mathcal{R}^{38}_{4} = -\mathcal{F}_4 \cdot \mathcal{F}_5$
$\mathcal{R}^{38}_{3} = \mathcal{F}_4 \cdot \mathcal{F}_5$
$\mathcal{R}^{38}_{2} = \mathcal{F}_5 \cdot \mathcal{F}_6$
$\mathcal{R}^{38}_{1} = \mathcal{F}_5 \cdot \mathcal{F}_7$

Kinetic terms generated from rule cr37:

$\mathcal{R}^{37}_{9} = \mathcal{F}_8$
$\mathcal{R}^{37}_{8} = -\mathcal{F}_8$
$\mathcal{R}^{37}_{5} = \mathcal{F}_8$

Kinetic terms generated from rule cr36:

$\mathcal{R}^{36}_{9} = -\mathcal{F}_5 \cdot \mathcal{F}_9$
$\mathcal{R}^{36}_{8} = \mathcal{F}_5 \cdot \mathcal{F}_9$
$\mathcal{R}^{36}_{5} = -\mathcal{F}_5 \cdot \mathcal{F}_9$

Kinetic terms generated from rule cr35:

$\mathcal{R}^{35}_{11} = \mathcal{F}_{10}$
$\mathcal{R}^{35}_{10} = -\mathcal{F}_{10}$
$\mathcal{R}^{35}_{9} = \mathcal{F}_{10}$

Kinetic terms generated from rule cr34:

$\mathcal{R}^{34}_{11} = -\mathcal{F}_9 \cdot \mathcal{F}_{11}$

$\mathcal{R}_{10}^{34} = \mathcal{F}_9 \cdot \mathcal{F}_{11}$
$\mathcal{R}_{9}^{34} = -\mathcal{F}_9 \cdot \mathcal{F}_{11}$

Kinetic terms generated from rule cr33:

$\mathcal{R}_{14}^{33} = -\mathcal{F}_{14}$
$\mathcal{R}_{13}^{33} = -\mathcal{F}_{13}$
$\mathcal{R}_{12}^{33} = -\mathcal{F}_{12}$
$\mathcal{R}_{11}^{33} = \mathcal{F}_{12} + \mathcal{F}_{13} + \mathcal{F}_{14}$
$\mathcal{R}_{7}^{33} = \mathcal{F}_{12}$
$\mathcal{R}_{6}^{33} = \mathcal{F}_{13}$
$\mathcal{R}_{4}^{33} = \mathcal{F}_{14}$

Kinetic terms generated from rule cr32:

$\mathcal{R}_{14}^{32} = \mathcal{F}_4 \cdot \mathcal{F}_{11}$
$\mathcal{R}_{13}^{32} = \mathcal{F}_6 \cdot \mathcal{F}_{11}$
$\mathcal{R}_{12}^{32} = \mathcal{F}_7 \cdot \mathcal{F}_{11}$
$\mathcal{R}_{11}^{32} = -\mathcal{F}_{11} \cdot (\mathcal{F}_4 + \mathcal{F}_6 + \mathcal{F}_7)$
$\mathcal{R}_{7}^{32} = -\mathcal{F}_7 \cdot \mathcal{F}_{11}$
$\mathcal{R}_{6}^{32} = -\mathcal{F}_6 \cdot \mathcal{F}_{11}$
$\mathcal{R}_{4}^{32} = -\mathcal{F}_4 \cdot \mathcal{F}_{11}$

Kinetic terms generated from rule cr31:

$\mathcal{R}_{17}^{31} = \mathcal{F}_1$
$\mathcal{R}_{16}^{31} = \mathcal{F}_2$
$\mathcal{R}_{15}^{31} = \mathcal{F}_3$
$\mathcal{R}_{8}^{31} = \mathcal{F}_1 + \mathcal{F}_2 + \mathcal{F}_3$
$\mathcal{R}_{3}^{31} = -\mathcal{F}_3$
$\mathcal{R}_{2}^{31} = -\mathcal{F}_2$
$\mathcal{R}_{1}^{31} = -\mathcal{F}_1$

Kinetic terms generated from rule cr30:

$\mathcal{R}_{17}^{30} = -\mathcal{F}_8 \cdot \mathcal{F}_{17}$
$\mathcal{R}_{16}^{30} = -\mathcal{F}_8 \cdot \mathcal{F}_{16}$
$\mathcal{R}_{15}^{30} = -\mathcal{F}_8 \cdot \mathcal{F}_{15}$
$\mathcal{R}_{8}^{30} = -\mathcal{F}_8 \cdot (\mathcal{F}_{15} + \mathcal{F}_{16} + \mathcal{F}_{17})$
$\mathcal{R}_{3}^{30} = \mathcal{F}_8 \cdot \mathcal{F}_{15}$
$\mathcal{R}_{2}^{30} = \mathcal{F}_8 \cdot \mathcal{F}_{16}$
$\mathcal{R}_{1}^{30} = \mathcal{F}_8 \cdot \mathcal{F}_{17}$

Kinetic terms generated from rule cr29:

$\mathcal{R}_{17}^{29} = \mathcal{F}_{12}$
$\mathcal{R}_{16}^{29} = \mathcal{F}_{13}$
$\mathcal{R}_{15}^{29} = \mathcal{F}_{14}$
$\mathcal{R}_{14}^{29} = -\mathcal{F}_{14}$
$\mathcal{R}_{13}^{29} = -\mathcal{F}_{13}$
$\mathcal{R}_{12}^{29} = -\mathcal{F}_{12}$
$\mathcal{R}_{10}^{29} = \mathcal{F}_{12} + \mathcal{F}_{13} + \mathcal{F}_{14}$

Kinetic terms generated from rule cr28:

$\mathcal{R}_{17}^{28} = -\mathcal{F}_{10} \cdot \mathcal{F}_{17}$
$\mathcal{R}_{16}^{28} = -\mathcal{F}_{10} \cdot \mathcal{F}_{16}$
$\mathcal{R}_{15}^{28} = -\mathcal{F}_{10} \cdot \mathcal{F}_{15}$
$\mathcal{R}_{14}^{28} = \mathcal{F}_{10} \cdot \mathcal{F}_{15}$

$\mathcal{R}_{13}^{28} = \mathcal{F}_{10} \cdot \mathcal{F}_{16}$
$\mathcal{R}_{12}^{28} = \mathcal{F}_{10} \cdot \mathcal{F}_{17}$
$\mathcal{R}_{10}^{28} = -\mathcal{F}_{10} \cdot (\mathcal{F}_{15} + \mathcal{F}_{16} + \mathcal{F}_{17})$

Kinetic terms generated from rule cr27:

$\mathcal{R}_{17}^{27} = \mathcal{F}_{7}$
$\mathcal{R}_{16}^{27} = \mathcal{F}_{6}$
$\mathcal{R}_{15}^{27} = \mathcal{F}_{4}$
$\mathcal{R}_{9}^{27} = \mathcal{F}_{4} + \mathcal{F}_{6} + \mathcal{F}_{7}$
$\mathcal{R}_{7}^{27} = -\mathcal{F}_{7}$
$\mathcal{R}_{6}^{27} = -\mathcal{F}_{6}$
$\mathcal{R}_{4}^{27} = -\mathcal{F}_{4}$

Kinetic terms generated from rule cr26:

$\mathcal{R}_{17}^{26} = -\mathcal{F}_{9} \cdot \mathcal{F}_{17}$
$\mathcal{R}_{16}^{26} = -\mathcal{F}_{9} \cdot \mathcal{F}_{16}$
$\mathcal{R}_{15}^{26} = -\mathcal{F}_{9} \cdot \mathcal{F}_{15}$
$\mathcal{R}_{9}^{26} = -\mathcal{F}_{9} \cdot (\mathcal{F}_{15} + \mathcal{F}_{16} + \mathcal{F}_{17})$
$\mathcal{R}_{7}^{26} = \mathcal{F}_{9} \cdot \mathcal{F}_{17}$
$\mathcal{R}_{6}^{26} = \mathcal{F}_{9} \cdot \mathcal{F}_{16}$
$\mathcal{R}_{4}^{26} = \mathcal{F}_{9} \cdot \mathcal{F}_{15}$

Kinetic terms generated from rule cr25:

$\mathcal{R}_{21}^{25} = \mathcal{F}_{18} + \mathcal{F}_{19} + \mathcal{F}_{20}$
$\mathcal{R}_{20}^{25} = -\mathcal{F}_{20}$
$\mathcal{R}_{19}^{25} = -\mathcal{F}_{19}$
$\mathcal{R}_{18}^{25} = -\mathcal{F}_{18}$
$\mathcal{R}_{17}^{25} = \mathcal{F}_{18}$
$\mathcal{R}_{16}^{25} = \mathcal{F}_{19}$
$\mathcal{R}_{15}^{25} = \mathcal{F}_{20}$

Kinetic terms generated from rule cr24:

$\mathcal{R}_{21}^{24} = -\mathcal{F}_{21} \cdot (\mathcal{F}_{15} + \mathcal{F}_{16} + \mathcal{F}_{17})$
$\mathcal{R}_{20}^{24} = \mathcal{F}_{15} \cdot \mathcal{F}_{21}$
$\mathcal{R}_{19}^{24} = \mathcal{F}_{16} \cdot \mathcal{F}_{21}$
$\mathcal{R}_{18}^{24} = \mathcal{F}_{17} \cdot \mathcal{F}_{21}$
$\mathcal{R}_{17}^{24} = -\mathcal{F}_{17} \cdot \mathcal{F}_{21}$
$\mathcal{R}_{16}^{24} = -\mathcal{F}_{16} \cdot \mathcal{F}_{21}$
$\mathcal{R}_{15}^{24} = -\mathcal{F}_{15} \cdot \mathcal{F}_{21}$

Kinetic terms generated from rule cr23:

$\mathcal{R}_{22}^{23} = \mathcal{F}_{1} + \mathcal{F}_{2} + \mathcal{F}_{3}$
$\mathcal{R}_{14}^{23} = \mathcal{F}_{3}$
$\mathcal{R}_{13}^{23} = \mathcal{F}_{2}$
$\mathcal{R}_{12}^{23} = \mathcal{F}_{1}$
$\mathcal{R}_{3}^{23} = -\mathcal{F}_{3}$
$\mathcal{R}_{2}^{23} = -\mathcal{F}_{2}$
$\mathcal{R}_{1}^{23} = -\mathcal{F}_{1}$

Kinetic terms generated from rule cr22:

$\mathcal{R}_{22}^{22} = -\mathcal{F}_{22} \cdot (\mathcal{F}_{12} + \mathcal{F}_{13} + \mathcal{F}_{14})$
$\mathcal{R}_{14}^{22} = -\mathcal{F}_{14} \cdot \mathcal{F}_{22}$
$\mathcal{R}_{13}^{22} = -\mathcal{F}_{13} \cdot \mathcal{F}_{22}$

$\mathcal{R}_{12}^{22} = -\mathcal{F}_{12} \cdot \mathcal{F}_{22}$
$\mathcal{R}_{3}^{22} = \mathcal{F}_{14} \cdot \mathcal{F}_{22}$
$\mathcal{R}_{2}^{22} = \mathcal{F}_{13} \cdot \mathcal{F}_{22}$
$\mathcal{R}_{1}^{22} = \mathcal{F}_{12} \cdot \mathcal{F}_{22}$

Kinetic terms generated from rule cr21:

$\mathcal{R}_{22}^{21} = \mathcal{F}_{8}$
$\mathcal{R}_{10}^{21} = \mathcal{F}_{8}$
$\mathcal{R}_{8}^{21} = -\mathcal{F}_{8}$

Kinetic terms generated from rule cr20:

$\mathcal{R}_{22}^{20} = -\mathcal{F}_{10} \cdot \mathcal{F}_{22}$
$\mathcal{R}_{10}^{20} = -\mathcal{F}_{10} \cdot \mathcal{F}_{22}$
$\mathcal{R}_{8}^{20} = \mathcal{F}_{10} \cdot \mathcal{F}_{22}$

Kinetic terms generated from rule cr19:

$\mathcal{R}_{22}^{19} = \mathcal{F}_{5}$
$\mathcal{R}_{11}^{19} = \mathcal{F}_{5}$
$\mathcal{R}_{5}^{19} = -\mathcal{F}_{5}$

Kinetic terms generated from rule cr18:

$\mathcal{R}_{22}^{18} = -\mathcal{F}_{11} \cdot \mathcal{F}_{22}$
$\mathcal{R}_{11}^{18} = -\mathcal{F}_{11} \cdot \mathcal{F}_{22}$
$\mathcal{R}_{5}^{18} = \mathcal{F}_{11} \cdot \mathcal{F}_{22}$

Kinetic terms generated from rule cr17:

$\mathcal{R}_{28}^{17} = \mathcal{F}_{23}$
$\mathcal{R}_{27}^{17} = \mathcal{F}_{24}$
$\mathcal{R}_{26}^{17} = \mathcal{F}_{25}$
$\mathcal{R}_{25}^{17} = -\mathcal{F}_{25}$
$\mathcal{R}_{24}^{17} = -\mathcal{F}_{24}$
$\mathcal{R}_{23}^{17} = -\mathcal{F}_{23}$
$\mathcal{R}_{22}^{17} = \mathcal{F}_{23} + \mathcal{F}_{24} + \mathcal{F}_{25}$

Kinetic terms generated from rule cr16:

$\mathcal{R}_{28}^{16} = -\mathcal{F}_{22} \cdot \mathcal{F}_{28}$
$\mathcal{R}_{27}^{16} = -\mathcal{F}_{22} \cdot \mathcal{F}_{27}$
$\mathcal{R}_{26}^{16} = -\mathcal{F}_{22} \cdot \mathcal{F}_{26}$
$\mathcal{R}_{25}^{16} = \mathcal{F}_{22} \cdot \mathcal{F}_{26}$
$\mathcal{R}_{24}^{16} = \mathcal{F}_{22} \cdot \mathcal{F}_{27}$
$\mathcal{R}_{23}^{16} = \mathcal{F}_{22} \cdot \mathcal{F}_{28}$
$\mathcal{R}_{22}^{16} = -\mathcal{F}_{22} \cdot (\mathcal{F}_{26} + \mathcal{F}_{27} + \mathcal{F}_{28})$

Kinetic terms generated from rule cr15:

$\mathcal{R}_{31}^{15} = \mathcal{F}_{23}$
$\mathcal{R}_{30}^{15} = \mathcal{F}_{24}$
$\mathcal{R}_{29}^{15} = \mathcal{F}_{25}$
$\mathcal{R}_{25}^{15} = -\mathcal{F}_{25}$
$\mathcal{R}_{24}^{15} = -\mathcal{F}_{24}$
$\mathcal{R}_{23}^{15} = -\mathcal{F}_{23}$
$\mathcal{R}_{5}^{15} = \mathcal{F}_{23} + \mathcal{F}_{24} + \mathcal{F}_{25}$

Kinetic terms generated from rule cr14:

$\mathcal{R}_{31}^{14} = -\mathcal{F}_5 \cdot \mathcal{F}_{31}$
$\mathcal{R}_{30}^{14} = -\mathcal{F}_5 \cdot \mathcal{F}_{30}$
$\mathcal{R}_{29}^{14} = -\mathcal{F}_5 \cdot \mathcal{F}_{29}$
$\mathcal{R}_{25}^{14} = \mathcal{F}_5 \cdot \mathcal{F}_{29}$
$\mathcal{R}_{24}^{14} = \mathcal{F}_5 \cdot \mathcal{F}_{30}$
$\mathcal{R}_{23}^{14} = \mathcal{F}_5 \cdot \mathcal{F}_{31}$
$\mathcal{R}_{5}^{14} = -\mathcal{F}_5 \cdot (\mathcal{F}_{29} + \mathcal{F}_{30} + \mathcal{F}_{31})$

Kinetic terms generated from rule cr13:

$\mathcal{R}_{31}^{13} = \mathcal{F}_{28}$
$\mathcal{R}_{30}^{13} = \mathcal{F}_{27}$
$\mathcal{R}_{29}^{13} = \mathcal{F}_{26}$
$\mathcal{R}_{28}^{13} = -\mathcal{F}_{28}$
$\mathcal{R}_{27}^{13} = -\mathcal{F}_{27}$
$\mathcal{R}_{26}^{13} = -\mathcal{F}_{26}$
$\mathcal{R}_{11}^{13} = \mathcal{F}_{26} + \mathcal{F}_{27} + \mathcal{F}_{28}$

Kinetic terms generated from rule cr12:

$\mathcal{R}_{31}^{12} = -\mathcal{F}_{11} \cdot \mathcal{F}_{31}$
$\mathcal{R}_{30}^{12} = -\mathcal{F}_{11} \cdot \mathcal{F}_{30}$
$\mathcal{R}_{29}^{12} = -\mathcal{F}_{11} \cdot \mathcal{F}_{29}$
$\mathcal{R}_{28}^{12} = \mathcal{F}_{11} \cdot \mathcal{F}_{31}$
$\mathcal{R}_{27}^{12} = \mathcal{F}_{11} \cdot \mathcal{F}_{30}$
$\mathcal{R}_{26}^{12} = \mathcal{F}_{11} \cdot \mathcal{F}_{29}$
$\mathcal{R}_{11}^{12} = -\mathcal{F}_{11} \cdot (\mathcal{F}_{29} + \mathcal{F}_{30} + \mathcal{F}_{31})$

Kinetic terms generated from rule cr11:

$\mathcal{R}_{21}^{11} = \mathcal{F}_9$
$\mathcal{R}_{9}^{11} = -\mathcal{F}_9$

Kinetic terms generated from rule cr10:

$\mathcal{R}_{20}^{10} = \mathcal{F}_4$
$\mathcal{R}_{19}^{10} = \mathcal{F}_6$
$\mathcal{R}_{18}^{10} = \mathcal{F}_7$
$\mathcal{R}_{7}^{10} = -\mathcal{F}_7$
$\mathcal{R}_{6}^{10} = -\mathcal{F}_6$
$\mathcal{R}_{4}^{10} = -\mathcal{F}_4$

Kinetic terms generated from rule cr9:

$\mathcal{R}_{18}^{9} = -\mathcal{F}_{18}$
$\mathcal{R}_{7}^{9} = \mathcal{F}_{18}$

Kinetic terms generated from rule cr8:

$\mathcal{R}_{34}^{8} = \mathcal{F}_{15}$
$\mathcal{R}_{33}^{8} = \mathcal{F}_{16}$
$\mathcal{R}_{32}^{8} = \mathcal{F}_{17}$
$\mathcal{R}_{17}^{8} = -\mathcal{F}_{17}$
$\mathcal{R}_{16}^{8} = -\mathcal{F}_{16}$
$\mathcal{R}_{15}^{8} = -\mathcal{F}_{15}$

Kinetic terms generated from rule cr7:

$\mathcal{R}_{32}^{7} = -\mathcal{F}_{32}$

$$\mathcal{R}_{17}^{7} = \mathcal{F}_{32}$$

Kinetic terms generated from rule cr6:

$$\mathcal{R}_{37}^{6} = \mathcal{F}_{29}$$
$$\mathcal{R}_{36}^{6} = \mathcal{F}_{30}$$
$$\mathcal{R}_{35}^{6} = \mathcal{F}_{31}$$
$$\mathcal{R}_{31}^{6} = -\mathcal{F}_{31}$$
$$\mathcal{R}_{30}^{6} = -\mathcal{F}_{30}$$
$$\mathcal{R}_{29}^{6} = -\mathcal{F}_{29}$$

Kinetic terms generated from rule cr5:

$$\mathcal{R}_{35}^{5} = -\mathcal{F}_{35}$$
$$\mathcal{R}_{31}^{5} = \mathcal{F}_{35}$$

Kinetic terms generated from rule cr4:

$$\mathcal{R}_{36}^{4} = \mathcal{F}_{35}$$
$$\mathcal{R}_{35}^{4} = -\mathcal{F}_{35}$$
$$\mathcal{R}_{33}^{4} = \mathcal{F}_{32}$$
$$\mathcal{R}_{32}^{4} = -\mathcal{F}_{32}$$
$$\mathcal{R}_{31}^{4} = -\mathcal{F}_{31}$$
$$\mathcal{R}_{30}^{4} = \mathcal{F}_{31}$$
$$\mathcal{R}_{28}^{4} = -\mathcal{F}_{28}$$
$$\mathcal{R}_{27}^{4} = \mathcal{F}_{28}$$
$$\mathcal{R}_{24}^{4} = \mathcal{F}_{23}$$
$$\mathcal{R}_{23}^{4} = -\mathcal{F}_{23}$$
$$\mathcal{R}_{19}^{4} = \mathcal{F}_{18}$$
$$\mathcal{R}_{18}^{4} = -\mathcal{F}_{18}$$
$$\mathcal{R}_{17}^{4} = -\mathcal{F}_{17}$$
$$\mathcal{R}_{16}^{4} = \mathcal{F}_{17}$$
$$\mathcal{R}_{13}^{4} = \mathcal{F}_{12}$$
$$\mathcal{R}_{12}^{4} = -\mathcal{F}_{12}$$
$$\mathcal{R}_{7}^{4} = -\mathcal{F}_{7}$$
$$\mathcal{R}_{6}^{4} = \mathcal{F}_{7}$$
$$\mathcal{R}_{2}^{4} = \mathcal{F}_{1}$$
$$\mathcal{R}_{1}^{4} = -\mathcal{F}_{1}$$

Kinetic terms generated from rule cr3:

$$\mathcal{R}_{36}^{3} = -\mathcal{F}_{36} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{35}^{3} = \mathcal{F}_{36} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{33}^{3} = -\mathcal{F}_{33} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{32}^{3} = \mathcal{F}_{33} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{31}^{3} = \mathcal{F}_{30} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{30}^{3} = -\mathcal{F}_{30} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{28}^{3} = \mathcal{F}_{27} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{27}^{3} = -\mathcal{F}_{27} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{24}^{3} = -\mathcal{F}_{24} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{23}^{3} = \mathcal{F}_{24} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{19}^{3} = -\mathcal{F}_{19} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{18}^{3} = \mathcal{F}_{19} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{17}^{3} = \mathcal{F}_{16} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{16}^{3} = -\mathcal{F}_{16} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{13}^{3} = -\mathcal{F}_{13} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$
$$\mathcal{R}_{12}^{3} = \mathcal{F}_{13} \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$$

$\mathcal{R}_7^3 = \mathcal{F}_6 \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$
$\mathcal{R}_6^3 = -\mathcal{F}_6 \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$
$\mathcal{R}_2^3 = -\mathcal{F}_2 \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$
$\mathcal{R}_1^3 = \mathcal{F}_2 \cdot (\mathcal{F}_{24} + \mathcal{F}_{27} + \mathcal{F}_{30} + \mathcal{F}_{36})$

Kinetic terms generated from rule cr2:

$\mathcal{R}_{38}^2 = -\mathcal{F}_{38} \cdot (\mathcal{F}_{25} + \mathcal{F}_{26} + \mathcal{F}_{29} + \mathcal{F}_{37})$
$\mathcal{R}_{37}^2 = -\mathcal{F}_{37} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{36}^2 = \mathcal{F}_{37} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{34}^2 = -\mathcal{F}_{34} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{33}^2 = \mathcal{F}_{34} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{30}^2 = \mathcal{F}_{29} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{29}^2 = -\mathcal{F}_{29} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{27}^2 = \mathcal{F}_{26} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{26}^2 = -\mathcal{F}_{26} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{25}^2 = -\mathcal{F}_{25} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{24}^2 = \mathcal{F}_{25} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{20}^2 = -\mathcal{F}_{20} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{19}^2 = \mathcal{F}_{20} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{16}^2 = \mathcal{F}_{15} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{15}^2 = -\mathcal{F}_{15} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{14}^2 = -\mathcal{F}_{14} \cdot \mathcal{F}_{38}$
$\mathcal{R}_{13}^2 = \mathcal{F}_{14} \cdot \mathcal{F}_{38}$
$\mathcal{R}_6^2 = \mathcal{F}_4 \cdot \mathcal{F}_{38}$
$\mathcal{R}_4^2 = -\mathcal{F}_4 \cdot \mathcal{F}_{38}$
$\mathcal{R}_3^2 = -\mathcal{F}_3 \cdot \mathcal{F}_{38}$
$\mathcal{R}_2^2 = \mathcal{F}_3 \cdot \mathcal{F}_{38}$

Kinetic terms generated from rule cr1:

$\mathcal{R}_{38}^1 = \mathcal{F}_2 + \mathcal{F}_6 + \mathcal{F}_{13} + \mathcal{F}_{16} + \mathcal{F}_{19} + \mathcal{F}_{33}$
$\mathcal{R}_{37}^1 = \mathcal{F}_{36}$
$\mathcal{R}_{36}^1 = -\mathcal{F}_{36}$
$\mathcal{R}_{34}^1 = \mathcal{F}_{33}$
$\mathcal{R}_{33}^1 = -\mathcal{F}_{33}$
$\mathcal{R}_{30}^1 = -\mathcal{F}_{30}$
$\mathcal{R}_{29}^1 = \mathcal{F}_{30}$
$\mathcal{R}_{27}^1 = -\mathcal{F}_{27}$
$\mathcal{R}_{26}^1 = \mathcal{F}_{27}$
$\mathcal{R}_{25}^1 = \mathcal{F}_{24}$
$\mathcal{R}_{24}^1 = -\mathcal{F}_{24}$
$\mathcal{R}_{20}^1 = \mathcal{F}_{19}$
$\mathcal{R}_{19}^1 = -\mathcal{F}_{19}$
$\mathcal{R}_{16}^1 = -\mathcal{F}_{16}$
$\mathcal{R}_{15}^1 = \mathcal{F}_{16}$
$\mathcal{R}_{14}^1 = \mathcal{F}_{13}$
$\mathcal{R}_{13}^1 = -\mathcal{F}_{13}$
$\mathcal{R}_6^1 = -\mathcal{F}_6$
$\mathcal{R}_4^1 = \mathcal{F}_6$
$\mathcal{R}_3^1 = \mathcal{F}_2$
$\mathcal{R}_2^1 = -\mathcal{F}_2$

7.2 The dynamical system for fragments.

$$\dot{\mathcal{F}}_1 = \mathcal{R}_1^3 + \mathcal{R}_1^4 + \mathcal{R}_1^{22} + \mathcal{R}_1^{23} + \mathcal{R}_1^{30} + \mathcal{R}_1^{31} + \mathcal{R}_1^{38} + \mathcal{R}_1^{39}$$

$$\dot{\mathcal{F}}_2 = \mathcal{R}_2^1 + \mathcal{R}_2^2 + \mathcal{R}_2^3 + \mathcal{R}_2^4 + \mathcal{R}_2^{22} + \mathcal{R}_2^{23} + \mathcal{R}_2^{30} + \mathcal{R}_2^{31} + \mathcal{R}_2^{38} + \mathcal{R}_2^{39}$$

$$\dot{\mathcal{F}}_3 = \mathcal{R}_3^1 + \mathcal{R}_3^2 + \mathcal{R}_3^{22} + \mathcal{R}_3^{23} + \mathcal{R}_3^{30} + \mathcal{R}_3^{31} + \mathcal{R}_3^{38} + \mathcal{R}_3^{39}$$

$$\dot{\mathcal{F}}_4 = \mathcal{R}_4^1 + \mathcal{R}_4^2 + \mathcal{R}_4^{10} + \mathcal{R}_4^{26} + \mathcal{R}_4^{27} + \mathcal{R}_4^{32} + \mathcal{R}_4^{33} + \mathcal{R}_4^{38} + \mathcal{R}_4^{39}$$

$$\dot{\mathcal{F}}_5 = \mathcal{R}_5^{14} + \mathcal{R}_5^{15} + \mathcal{R}_5^{18} + \mathcal{R}_5^{19} + \mathcal{R}_5^{36} + \mathcal{R}_5^{37} + \mathcal{R}_5^{38} + \mathcal{R}_5^{39}$$

$$\dot{\mathcal{F}}_6 = \mathcal{R}_6^1 + \mathcal{R}_6^2 + \mathcal{R}_6^3 + \mathcal{R}_6^4 + \mathcal{R}_6^{10} + \mathcal{R}_6^{26} + \mathcal{R}_6^{27} + \mathcal{R}_6^{32} + \mathcal{R}_6^{33} + \mathcal{R}_6^{38} + \mathcal{R}_6^{39}$$

$$\dot{\mathcal{F}}_7 = \mathcal{R}_7^3 + \mathcal{R}_7^4 + \mathcal{R}_7^9 + \mathcal{R}_7^{10} + \mathcal{R}_7^{26} + \mathcal{R}_7^{27} + \mathcal{R}_7^{32} + \mathcal{R}_7^{33} + \mathcal{R}_7^{38} + \mathcal{R}_7^{39}$$

$$\dot{\mathcal{F}}_8 = \mathcal{R}_8^{20} + \mathcal{R}_8^{21} + \mathcal{R}_8^{30} + \mathcal{R}_8^{31} + \mathcal{R}_8^{36} + \mathcal{R}_8^{37}$$

$$\dot{\mathcal{F}}_9 = \mathcal{R}_9^{11} + \mathcal{R}_9^{26} + \mathcal{R}_9^{27} + \mathcal{R}_9^{34} + \mathcal{R}_9^{35} + \mathcal{R}_9^{36} + \mathcal{R}_9^{37}$$

$$\dot{\mathcal{F}}_{10} = \mathcal{R}_{10}^{20} + \mathcal{R}_{10}^{21} + \mathcal{R}_{10}^{28} + \mathcal{R}_{10}^{29} + \mathcal{R}_{10}^{34} + \mathcal{R}_{10}^{35}$$

$$\dot{\mathcal{F}}_{11} = \mathcal{R}_{11}^{12} + \mathcal{R}_{11}^{13} + \mathcal{R}_{11}^{18} + \mathcal{R}_{11}^{19} + \mathcal{R}_{11}^{32} + \mathcal{R}_{11}^{33} + \mathcal{R}_{11}^{34} + \mathcal{R}_{11}^{35}$$

$$\dot{\mathcal{F}}_{12} = \mathcal{R}_{12}^3 + \mathcal{R}_{12}^4 + \mathcal{R}_{12}^{22} + \mathcal{R}_{12}^{23} + \mathcal{R}_{12}^{28} + \mathcal{R}_{12}^{29} + \mathcal{R}_{12}^{32} + \mathcal{R}_{12}^{33}$$

$$\dot{\mathcal{F}}_{13} = \mathcal{R}_{13}^1 + \mathcal{R}_{13}^2 + \mathcal{R}_{13}^3 + \mathcal{R}_{13}^4 + \mathcal{R}_{13}^{22} + \mathcal{R}_{13}^{23} + \mathcal{R}_{13}^{28} + \mathcal{R}_{13}^{29} + \mathcal{R}_{13}^{32} + \mathcal{R}_{13}^{33}$$

$$\dot{\mathcal{F}}_{14} = \mathcal{R}_{14}^1 + \mathcal{R}_{14}^2 + \mathcal{R}_{14}^{22} + \mathcal{R}_{14}^{23} + \mathcal{R}_{14}^{28} + \mathcal{R}_{14}^{29} + \mathcal{R}_{14}^{32} + \mathcal{R}_{14}^{33}$$

$$\dot{\mathcal{F}}_{15} = \mathcal{R}_{15}^1 + \mathcal{R}_{15}^2 + \mathcal{R}_{15}^8 + \mathcal{R}_{15}^{24} + \mathcal{R}_{15}^{25} + \mathcal{R}_{15}^{26} + \mathcal{R}_{15}^{27} + \mathcal{R}_{15}^{28} + \mathcal{R}_{15}^{29} + \mathcal{R}_{15}^{30} + \mathcal{R}_{15}^{31}$$

$$\dot{\mathcal{F}}_{16} = \mathcal{R}_{16}^1 + \mathcal{R}_{16}^2 + \mathcal{R}_{16}^3 + \mathcal{R}_{16}^4 + \mathcal{R}_{16}^8 + \mathcal{R}_{16}^{24} + \mathcal{R}_{16}^{25} + \mathcal{R}_{16}^{26} + \mathcal{R}_{16}^{27} + \mathcal{R}_{16}^{28} + \mathcal{R}_{16}^{29} + \mathcal{R}_{16}^{30} + \mathcal{R}_{16}^{31}$$

$$\dot{\mathcal{F}}_{17} = \mathcal{R}_{17}^3 + \mathcal{R}_{17}^4 + \mathcal{R}_{17}^7 + \mathcal{R}_{17}^8 + \mathcal{R}_{17}^{24} + \mathcal{R}_{17}^{25} + \mathcal{R}_{17}^{26} + \mathcal{R}_{17}^{27} + \mathcal{R}_{17}^{28} + \mathcal{R}_{17}^{29} + \mathcal{R}_{17}^{30} + \mathcal{R}_{17}^{31}$$

$$\dot{\mathcal{F}}_{18} = \mathcal{R}_{18}^3 + \mathcal{R}_{18}^4 + \mathcal{R}_{18}^9 + \mathcal{R}_{18}^{10} + \mathcal{R}_{18}^{24} + \mathcal{R}_{18}^{25}$$

$$\dot{\mathcal{F}}_{19} = \mathcal{R}_{19}^1 + \mathcal{R}_{19}^2 + \mathcal{R}_{19}^3 + \mathcal{R}_{19}^4 + \mathcal{R}_{19}^{10} + \mathcal{R}_{19}^{24} + \mathcal{R}_{19}^{25}$$

$$\dot{\mathcal{F}}_{20} = \mathcal{R}_{20}^1 + \mathcal{R}_{20}^2 + \mathcal{R}_{20}^{10} + \mathcal{R}_{20}^{24} + \mathcal{R}_{20}^{25}$$

$$\dot{\mathcal{F}}_{21} = \mathcal{R}_{21}^{11} + \mathcal{R}_{21}^{24} + \mathcal{R}_{21}^{25}$$

$$\dot{\mathcal{F}}_{22} = \mathcal{R}_{22}^{16} + \mathcal{R}_{22}^{17} + \mathcal{R}_{22}^{18} + \mathcal{R}_{22}^{19} + \mathcal{R}_{22}^{20} + \mathcal{R}_{22}^{21} + \mathcal{R}_{22}^{22} + \mathcal{R}_{22}^{23}$$

$$\dot{\mathcal{F}}_{23} = \mathcal{R}_{23}^3 + \mathcal{R}_{23}^4 + \mathcal{R}_{23}^{14} + \mathcal{R}_{23}^{15} + \mathcal{R}_{23}^{16} + \mathcal{R}_{23}^{17}$$

$$\dot{\mathcal{F}}_{24} = \mathcal{R}_{24}^1 + \mathcal{R}_{24}^2 + \mathcal{R}_{24}^3 + \mathcal{R}_{24}^4 + \mathcal{R}_{24}^{14} + \mathcal{R}_{24}^{15} + \mathcal{R}_{24}^{16} + \mathcal{R}_{24}^{17}$$

$$\dot{\mathcal{F}}_{25} = \mathcal{R}_{25}^1 + \mathcal{R}_{25}^2 + \mathcal{R}_{25}^{14} + \mathcal{R}_{25}^{15} + \mathcal{R}_{25}^{16} + \mathcal{R}_{25}^{17}$$

$$\dot{\mathcal{F}}_{26} = \mathcal{R}_{26}^1 + \mathcal{R}_{26}^2 + \mathcal{R}_{26}^{12} + \mathcal{R}_{26}^{13} + \mathcal{R}_{26}^{16} + \mathcal{R}_{26}^{17}$$

$$\dot{\mathcal{F}}_{27} = \mathcal{R}_{27}^1 + \mathcal{R}_{27}^2 + \mathcal{R}_{27}^3 + \mathcal{R}_{27}^4 + \mathcal{R}_{27}^{12} + \mathcal{R}_{27}^{13} + \mathcal{R}_{27}^{16} + \mathcal{R}_{27}^{17}$$

$$\dot{\mathcal{F}}_{28} = \mathcal{R}_{28}^3 + \mathcal{R}_{28}^4 + \mathcal{R}_{28}^{12} + \mathcal{R}_{28}^{13} + \mathcal{R}_{28}^{16} + \mathcal{R}_{28}^{17}$$

$$\dot{\mathcal{F}}_{29} = \mathcal{R}_{29}^1 + \mathcal{R}_{29}^2 + \mathcal{R}_{29}^6 + \mathcal{R}_{29}^{12} + \mathcal{R}_{29}^{13} + \mathcal{R}_{29}^{14} + \mathcal{R}_{29}^{15}$$

$$\dot{\mathcal{F}}_{30} = \mathcal{R}_{30}^1 + \mathcal{R}_{30}^2 + \mathcal{R}_{30}^3 + \mathcal{R}_{30}^4 + \mathcal{R}_{30}^6 + \mathcal{R}_{30}^{12} + \mathcal{R}_{30}^{13} + \mathcal{R}_{30}^{14} + \mathcal{R}_{30}^{15}$$

$$\dot{\mathcal{F}}_{31} = \mathcal{R}_{31}^3 + \mathcal{R}_{31}^4 + \mathcal{R}_{31}^5 + \mathcal{R}_{31}^6 + \mathcal{R}_{31}^{12} + \mathcal{R}_{31}^{13} + \mathcal{R}_{31}^{14} + \mathcal{R}_{31}^{15}$$

$$\dot{\mathcal{F}}_{32} = \mathcal{R}_{32}^3 + \mathcal{R}_{32}^4 + \mathcal{R}_{32}^7 + \mathcal{R}_{32}^8$$

$$\dot{\mathcal{F}}_{33} = \mathcal{R}_{33}^1 + \mathcal{R}_{33}^2 + \mathcal{R}_{33}^3 + \mathcal{R}_{33}^4 + \mathcal{R}_{33}^8$$

$$\dot{\mathcal{F}}_{34} = \mathcal{R}_{34}^1 + \mathcal{R}_{34}^2 + \mathcal{R}_{34}^8$$

$$\dot{\mathcal{F}}_{35} = \mathcal{R}_{35}^3 + \mathcal{R}_{35}^4 + \mathcal{R}_{35}^5 + \mathcal{R}_{35}^6$$

$$\dot{\mathcal{F}}_{36} = \mathcal{R}_{36}^1 + \mathcal{R}_{36}^2 + \mathcal{R}_{36}^3 + \mathcal{R}_{36}^4 + \mathcal{R}_{36}^6$$

$$\dot{\mathcal{F}}_{37} = \mathcal{R}_{37}^1 + \mathcal{R}_{37}^2 + \mathcal{R}_{37}^6$$

$$\dot{\mathcal{F}}_{38} = \mathcal{R}_{38}^1 + \mathcal{R}_{38}^2$$

8 A comparison

For the purpose of comparison, and as a further test case, we applied our procedure to a model of crosstalk between EGF and insulin receptors, treated by Conzelmann et al. in [11] (the "CFG model"). We also applied our procedure to a simplification of the CFG-model that consists in removing contextual specifications on the lhs of dissociation rules (turning them into "pure dissociation rules" per our terminology in the main paper, section on "Syntactical criteria for annotating the contact map".) The automatically generated reports are available as separate additional information, as they comprise 86 and 39 pages, respectively.

The CFG model (table 7 of [11]) consists of 76 rules giving rise to 2899 molecular species. Conzelmann et al. report their system to comprise 5182 species. The discrepancy is a consequence of operating without a formal agent-based language. The authors of [11] have therefore no way of accounting for symmetries that might be present in molecular species. Because of extensive dimer formation, symmetries are rampant, shrinking the number of distingushable species by 44%. The method for reducing CFG as described in [11] yields 391 coarse-grained variables. Our automatic procedure yields 208 fragments. (In [11], the 391 variables are subsequently reduced to 87 by applying a strategy for detecting systemic modules, which is an alltogether different method than coarse-graining as we understand it. This method could be applied to our ODE system as well.)

Context-dependency of dissociation causes soft bonds to become solid (see directive Edg1 in the main paper). This, in turn, gives rise to larger fragments. Larger fragments often entail more numerous fragments, since a family of fragments is obtained by generating all possible state valuations on a chosen class of sites. This effect is illustrated by removing the context-dependency of dissociation in the CFG model. The resulting simplified model still consists of 76 rules and preserves the 2899 possible species, but our procedure now generates only 88 fragments.

1. Danos, V & Laneve, C. *Formal Molecular Biology* (2004) *Theoretical Computer Science* 325, 69–110.
2. Access to the Kappa modeling platform is provided at www.cellucidate.com.
3. Danos, V, Feret, J, Fontana, W, & Krivine, J. (2007) *Scalable simulation of cellular signalling networks*, Lecture Notes in Computer Science. (Springer), Vol. 4807, pp. 139–157.
4. Danos, V, Feret, J, Fontana, W, & Krivine, J. (2008) *Abstract interpretation of cellular signalling networks*, Lecture Notes in Computer Science. (Springer), Vol. 4905, pp. 83–97.
5. Doob, J. L. *Markoff chains Ð denumerable case* (1945) *Trans. Amer. Math. Soc.* 58, 455Ð–473.
6. Gillespie, D. T. *A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions* (1976) *Journal of Computational Physics* 22, 403–434.
7. Borisov, N. M, Markevich, N. I, Hoek, J. B, & Kholodenko, B. N. *Signaling through Receptors and Scaffolds: Independent Interactions Reduce Combinatorial Complexity* (2005) *Biophysical Journal* 89, 951–966.
8. Conzelmann, H, Saez-Rodriguez, J, Sauter, T, Kholodenko, B. N, & Gilles, E. D. *A domain-oriented approach to the reduction of combinatorial complexity in signal transduction networks* (2006) *BMC Bioinformatics* 7, 34.
9. Conzelmann, H. (2008) Ph.D. thesis (Institut für Systemdynamik der Universität Stuttgart).
10. Danos, V, Feret, J, Fontana, W, Harmer, R, & Krivine, J. (2007) *Rule-based modelling of cellular signalling*, Lecture Notes in Computer Science. (Springer, Lisboa, Portugal), Vol. 4703, pp. 17–41.
11. Conzelmann, H, Fey, D, & Gilles, E. D. *Exact model reduction of combinatorial reaction networks* (2008) *BMC Systems Biology* 2, 78.