

Disrupted Pathways: Generating Tunable Macromolecular Assembly Pathways

By

Koan E. M. Briggs

Submitted to the graduate degree program in Department of Physics and Astronomy and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Christopher J. Fischer, Chairperson

Eric J. Deeds

Committee members

Steven A. Hawley

Michael J. Murray

JiCong Shi

Date defended: October 15, 2018

The Dissertation Committee for Koan E. M. Briggs certifies
that this is the approved version of the following dissertation :

Disrupted Pathways:
Generating Tunable Macromolecular Assembly Pathways

Christopher J. Fischer, Chairperson

Date approved: October 15, 2018

Abstract

What follows is a pathway; a sequence of individual events, which together form a story. Yet it is still only a small part of what has come before. Biological structures also have individual stories; each composed of simple events in sequence. One story does not tell the whole, for that we must observe many stories, sample them if you will. Together, they bring understanding.

Assembly is an emergent property of many individual binding events. Through this, all of the structures that make up life are created. Understanding the regime of possibilities provides insight into both the breadth and tendencies of the system. Cells contain numerous types of individual proteins many of which come together to form larger complexes. I will begin by introducing the elementary building blocks of those protein complexes. An introductory example will provide the first perspective, it will form common ground and allow the telling of the larger story with a shared perspective. Then a case study, a real biological complex and how understanding the progression of its pathways provided insight into the states which it reached. With the elementary operations described, I will move on to laying out the landscape of possible pathways; first for a specific case and then the structure of the assembly pathways themselves. Thus, providing a novel framework for the understanding of the stochastic space of protein complex assembly. Finally, I will provide an example of how making changes in the possible assembly pathways leads to non-intuitive changes in the conclusion of the protein complexes' stories.

To the Universe; it's always gazing back...

$$2^{\aleph_0} \stackrel{?}{=} \aleph_1$$

Acknowledgements

Everyone I meet students, professors, colleagues teaches me a little more. From that I emerge as the person I am. Thank you to everyone I have met over these long years, I would literally not be the person I am today without you all. Now there is without question, a group of people who have influenced me more than others. Most of them probably know who they are, but that is not the point of this, this is the part where the sappiness should flow.

Caitlin, without your support I would be nothing like who I am today. Your love and my love for you is only strengthened by our shared curiosity. In the course of this document, you have been a wonderful sounding board, editor, and supporter, thank you. Mom, I know you know how much you mean to me, but I should say it more often. It was your and Pat's encouragement and support in returning to school that unquestionably created a turning point in my life. Thank you for the support, emotionally and financially to set me on this journey.

While it is customary to thank the committee, in my case each member has truly impacted my life for the better. Chris Fischer; your friendship, our growing mutual cynicism, celebrations of the supreme leader and eternal president, and of course your dedication to robust repeatable science have helped me become the well-adjusted person I am today. Eric Deeds; your enthusiasm, appreciation for agile automobiles, details of figure creation, and formal application of computational methods to the messy science of life. Steven Hawley; you probably don't even know how much you have influenced me, I have seen you since I was young looking out the window guiding the arm towards the Hubble Space Telescope, participating in the group which opened my mind to being a scientist as opposed to a military test pilot, and later providing insights into organizational

management. Michael Murry; your infectious enthusiasm while teaching was part of the reason I changed focus from mathematics to physics and your encouragement to build and investigate without bounds convinced me I was in the right place. JiCong “Jack” Shi; you showed me the power and beauty in using abstract mathematics and analytic equations for solving current problems, and the amount one can learn in just *two* more minutes. While, not on my committee, I would also like to thank Sergei Shandarin, for teaching me almost everything I know about the core tenets of physics, and under what situation there can be more *degrees* of freedom in the Soviet Union than on I-70.

For good or ill, I am one of those people who tend to find creative ways of doing things. The people who enjoyed the most interesting results of this trait, are Bob Curry and Kristin Rennells. Thank you for being there, helping my crazy plans, being genuinely good people, and *mostly* rolling your eyes when I wasn’t looking. I have had the good fortune to work with many lab mates over the years, you are all appreciated. Though there are two who I feel I need to take a special moment for. Dr. Allen Eastlund; I know from your acknowledgments I poked you about “how a scientist should look at problems,” but you taught me how to look at life and friendship. Dr. Gada Al-Ani; I know we disagreed over many trivial things, like display details in figures, but you showed me a level of resiliency, meticulousness, and tenacity that I still aspire to. Two of my friends also blurred the lines between my academic life and the “outside”. Cassandra Field; you have broadened my vision of interdisciplinary science and the need to look beyond the obvious. Bertrand Kotewall; you asked me when this document would be done more times than my mother, you are the kind of friend who calls even when I forget to. Well now it is *done*, but I guess you already lost the bet. You are literally the kind of person who bets on me.



Gypsy, Speedy, Mom
I wish all of you had been able to see me reach this point.

Contents

List of Figures	x
List of Tables	xiii
Acronyms & Glossary	xiv
Preface	1
0.1 On the First-Person Plural	1
0.2 On the Recapitulation of Background Material	1
1 One to One Protein Binding: A buildup exercise	2
1.1 A:B Binding	5
1.2 A:A Binding	11
1.3 Synthesis and Degradation	13
2 ISWI:NCP Binding	15
2.1 SBEs and Branching	16
2.2 ISWI	18
3 Stacked Trimer	24
3.1 Technical Description	25

3.2	Pathways	32
3.3	Analysis	33
4	Assembly Pathways	60
4.1	Trees: A Survey	61
4.2	General Assembly Pathways	63
4.3	Integer Representation of Complexes	72
4.4	Pathway Contribution	73
5	Proteasome Assembly Modifiers	75
5.1	Technical Description	76
5.2	Analysis	78
6	Conclusion	87
	References	89
	Supplementary Reading	100
A	Stacked Trimer ODEs	102

List of Figures

1.1	Species Yield in One to One Binding	8
A	(A:B Binding) Species Yield vs Time	8
B	(A:B Binding) Species Yield vs K_d	8
C	(A:A Binding) Species Yield vs Time	8
D	(A:A Binding) Species Yield vs K_d	8
1.2	(A:B Binding) Equilibrium Species Yield vs Initial Concentration	10
A	Yield of A vs Initial Concentrations	10
B	Yield of B vs Initial Concentrations	10
C	Yield of AB vs Initial Concentrations	10
2.1	ISWI:NCP Binding	23
A	Change in Anisotropy (Δr) vs Initial Concentrations	23
B	Yield of ISWI Species vs Initial Concentrations	23
3.1	Stacked Trimer Diagram & Bond Types	25
3.2	Stacked Trimer Subspecies Chart	27
3.3	Assembly Network Workflow	30
3.4	Stacked Trimer Reactions	31
3.5	Stacked Trimer Assembly Yield Overview	40

A	Heatmap of <i>in vitro</i> -like Assembly Yield	40
B	Heatmap of <i>in vivo</i> -like Assembly Yield	40
C	Assembly Yield vs Time (<i>in vitro</i> -like)	40
D	Assembly Yield vs Time (<i>in vivo</i> -like)	40
E	Assembly Yield vs Concentration (<i>in vitro</i> -like)	40
F	Assembly Yield vs Concentration (<i>in vivo</i> -like)	40
3.6	Stacked Trimer Pathway Contribution (<i>in vitro</i> -like)	42
3.7	Stacked Trimer Species Fractions over Time (<i>in vitro</i> -like)	43
3.8	Stacked Trimer Assembly Contributions over Time (<i>in vitro</i> -like)	45
3.9	Stacked Trimer Species Fractions over Concentration (<i>in vitro</i> -like)	46
3.10	Stacked Trimer Assembly Contributions over Concentration (<i>in vitro</i> -like)	48
3.11	Stacked Trimer Pathway Contribution (<i>in vivo</i> -like)	49
3.12	Stacked Trimer Species Fractions over Time (<i>in vivo</i> -like)	51
3.13	Stacked Trimer Assembly Contributions over Time (<i>in vivo</i> -like)	52
3.14	Stacked Trimer Species Fractions over Concentration (<i>in vivo</i> -like)	54
3.15	Stacked Trimer Assembly Contributions over Concentration (<i>in vivo</i> -like)	55
3.16	Effects of Synthesis and Degradation Rate on Assembly Yield	57
A	Assembly Yield vs Cell Division Period (s) @ Time: \approx One Day	57
B	Assembly Yield vs Cell Division Period (s) @ Time: 10^{11} (s)	57
C	Heatmap of <i>in vitro</i> -like Assembly Yield @ Time: \approx One Day	57
D	Heatmap of <i>in vivo</i> -like Assembly Yield @ Time: \approx One Day	57
E	Heatmap of <i>in vitro</i> -like Assembly Yield @ Time: 10^{11} (s)	57
F	Heatmap of <i>in vivo</i> -like Assembly Yield @ Time: 10^{11} (s)	57
3.17	Buried Solvent-Accessible Surface Area (BSASA) (<i>in vitro</i> -like)	58
3.18	Buried Solvent-Accessible Surface Area (BSASA) (<i>in vivo</i> -like)	59
4.1	Illustrative Examples of Stacked Trimer Assembly Pathways	74

5.1	Illustration of Proteasome, Half Proteasome and Bond Types	80
A	Proteasome Core Particle Diagram	80
B	Half-Proteasome Diagram	80
C	Proteasome Object Types	80
D	Normal and Inhibited Relationships of the Half-Proteasome	80
5.2	Effects of Assembly Modifiers at 38.3 (s)	81
5.3	Effects of Assembly Modifiers at 1.0×10^3 (s)	83
5.4	Effects of Assembly Modifiers at 3.3×10^4 (s)	84
5.5	Effects of Assembly Modifiers at 1.9×10^7 (s)	86

List of Tables

2.1	ISWI Two-Site Binding Parameters	22
3.1	Symmetries of the Stacked Trimer (D_3)	26
3.2	Default Stacked Trimer Properties	34
3.3	Assembly Pathways for Stacked Trimers	36
4.1	Tree Properties of Stacked Rings	64
4.2	Stacked Trimer Assembly Pathway Indices	69
4.3	Assembly Pathway Properties of Stacked Rings	70
4.4	Representations of the Stacked Trimer	71
5.1	# of Canonical Species and Reactions in the Modified Proteasome System	76
5.2	Bond Strengths of the Simulated Proteasome Core Particle	78

Acronyms & Glossary

activation energy The difference between the current energy of the system and the peak energy in the reaction process. There are many ways to define an activation energy depending on the energetics under consideration.

Supplementary Reading: Biological Chemical Reactions

see also reaction

Alexa Fluor™ 488 A proprietary chemical from Thermo Fisher Scientific, with excellent fluorescent properties in regards to biological applications.

see also fluorescence anisotropy & fluorophore

assembly pathway Binary tree representation of association reactions required to form the maximal structure.

see also binary tree, maximal structure, minimal structure & protein complex

association A reaction between two or more chemicals resulting in a single chemical. In the context of this document all association reactions are bimolecular.

see also binding & reaction

binary tree A directed acyclic graph where all nodes have at most one incoming edge and two outgoing edges. This document assumes all binary trees are rooted and have only a single connected component.

Supplementary Reading: Trees (Data Structures) and Trees (Graph Theory)

binding As this is a broad term, I will clarify how it will be used in this document.

Binding reactions are non-covalent chemical reactions with multi-object products.

Additionally, all binding reactions are reversible.

Supplementary Reading: Biological Chemical Reactions

see also association, dissociation & reaction

Boltzmann factor Ratio between two states of a system representable by a Boltzmann distribution. Commonly defined to be dependent only on the difference between relative energy levels of the states when in an isothermal system.

Supplementary Reading: Biological Chemical Reactions

branching When a reactant(s) has more than one mutually exclusive reaction available.

Supplementary Reading: Chemical Reaction Networks (CRN)

see also reaction

Brownian motion The random movement of a particle dominated by collisions in a fluid-like or gas-like environment.

Supplementary Reading: Biological Chemical Reactions

CRN Chemical Reaction Network

Supplementary Reading: Chemical Reaction Networks (CRN)

see also reaction, association & dissociation

DFT Depth First Traversal (Prefix Order)

All DFTs in this text refer to Prefix order.

Supplementary Reading: Trees (Data Structures)

dissociation A reaction where a single chemical converts into multiple chemicals. In the context of this document all dissociation reactions result in exactly two chemicals.

see also binding & reaction

DNA DeoxyriboNucleic Acid

EMSA Electrophoretic Mobility Shift Assay

Supplementary Reading: Experimental EMSA Protocols

fluorescence anisotropy Rotation of the effective dipole of a fluorophore between excitation and emission. Often accomplished by using linearly polarized light to excite the fluorophores, then measuring the polarization of the emitted light.

Supplementary Reading: Fluorescence & Fluorescence Anisotropy

see also fluorophore

fluorophore A chemical which is luminescent when excited by electromagnetic radiation. In the context of this document they are used as a monitoring probe that has been artificially added to the target of interest.

Supplementary Reading: Fluorescence & Fluorescence Anisotropy

see also fluorescence anisotropy

full binary tree A binary tree where every non-terminal node has exactly two children. Can also be known as a proper binary tree, but not a perfect binary tree or complete binary tree. The terms full, complete, and perfect are often misused.

see also binary tree

HP Half-Proteasome

Supplementary Reading: Proteasomes

see also proteasome, protein & protein complex

ISWI Imitation SWItch

A chromatin remodeling protein.

Supplementary Reading: Nucleosomes and Chromatin

see also DNA, NCP & protein

maximal structure A unique multipart object representing the completion of an assembly pathway.

minimal structure An object with no substructures.

NCP Nucleosome Core Particle

Supplementary Reading: Nucleosomes and Chromatin

see also DNA, protein & protein complex

proteasome In this document proteasome always refers to the 20S proteasome core particle.

The core particle is highly structurally conserved across species, but the regulatory caps are far more variable.

Supplementary Reading: Proteasomes

see also protein & protein complex

protein A macromolecular polymer, made up of small organic molecules (amino acids).

In the context of this document all proteins contain only a single chain, otherwise they are considered protein complexes.

protein complex Multiple proteins forming a higher order structure.

Supplementary Reading: Protein Symmetry

see also protein

reaction In the context of this document only chemical reactions are considered. Additionally, only unimolecular and bimolecular reactions are allowed.

Supplementary Reading: Biological Chemical Reactions

SBE Sequential Binding Event

stacked trimer A homohexameric protein complex with D_3 symmetry.

Supplementary Reading: Protein Symmetry

see also protein & protein complex

viral capsid The protein shell around a virus, often featuring non-covalently bound repeating patterns of a small number of proteins.

yield Quantity of a species divided by the maximum possible for that species, *i.e.*, the maximum is defined as the maximum value considering only conservation of molecules. This has *no* implied relation to steady state values for the system.

Preface

“When you make a thing, a thing that is new, it is so complicated making it that it is bound to be ugly. But those that make it after you, they don’t have to worry about making it. And they can make it pretty, and so everybody can like it when others make it after you.”

— Pablo Picasso

0.1 On the First-Person Plural

We, the community.

We, the union of author and reader.

We, the in-group of collaborators.

I will take responsibility for what *I* say, but sometimes *we* will be in this together.

0.2 On the Recapitulation of Background Material

Many subjects have been expounded upon in treatises both elegant and enlightening.

Generally arising from decades of dedication and revision.

Recapitulation with an original simulacrum, obfuscates the concepts.

Henceforth, I will provide suggested reading material for those occurrences.

One to One

Protein Binding

“In the end, bored by the repetitions,
I conclude my exercise by eating him.”

— Natsume Sōseki, *I am A Cat*

A buildup exercise

At its heart, life is about interacting objects. We do not think of individual proteins as alive; we see life as an emergent phenomenon of interactions. To this end, I will begin with a discussion of a simple class of interactions, a protein encountering another protein. So, how does a protein move? Theories of protein movement range from explicit descriptions of their subcomponents' wavefunctions up through averages of ensembles. I will be approximating molecular motion as spherical particles moving due to thermal fluctuations. While precision may be lost by using an approximation coarser than technically possible, we must remember life exists in only a narrow range of possible physical and chemical parameter space. This level of approximation provides general applicability, as more explicit descriptions increase dependence on properties of the specific system. Physical conditions demanding more explicit modeling exist in biological contexts, *e.g.*, DNA packed into viral capsids is suspected to exist in a glass-like state [1, 2]. On the other hand, the currently known extremophiles remain within a reasonable regime [3].

Let us start with a case where there is a single protein surrounded by other molecules which can collide with it, but that is all these other molecules do. Additionally, I consider the density of these background molecules to be something reasonable for a living system. For my purposes, this should end up looking a lot like a hard sphere moving around in a water-like fluid. The other objects in the fluid do not have to be water molecules, they just need to be something that does not break the biological fluid approximation. In a real system, there would be small organic molecules, other proteins, and of course some water molecules, just to name a few. The exact details of the environment are a matter of some debate [4–10]. The primary result of changing the background environment, in this approximation, is a change in the diffusion rate of the reference sphere.

I am describing a system in which the sphere's movement is dominated by collisions with the fluid, this naturally gives rise to Brownian motion [6, 11, 12]. At this point I will add more moving spheres. Now, if the spheres just bounce off of one another it is not going to change anything other than the movement dynamics of the reference spheres. Let us add the chance to bind together on contact. Binding at just the right strength is one of the keys to life. If everything binds together perfectly, the world eventually becomes one giant aggregate. To balance that, I am going to also let things fall apart. The need for this balance is part of why life works only in a small window of the energy.

I will now take a moment to explain why I have not mentioned energetics for the system as a whole. The cellular environment is a coupling of complex reaction networks, which itself is encapsulated within an external environment with constantly fluctuating thermodynamic properties. Direct modeling of this system is both highly complex and very coarse [13]. All of the properties extrinsic to the model end up being encapsulated in the diffusion, as a coarse approximation, while the one intrinsic property will be treated in the unbinding of the spheres. Thus, I am basically considering the fluid to be a heat bath.

Now it is time to get into some details. In many cases, binding is described by a collision frequency combined with a Boltzmann factor based on the activation energy of

the reaction [6]. While I could derive an approximation through activation energy-based collision theory, two issues would render such an approach complex. One, only patches of the sphere are bindable and those patches have alignment requirements. Two, in most real cases the objects are flexible with dynamic motions, resulting in non-spherical behaviors in the binding interaction. The net result is that derivations of the general association rate tend to differ from the empirically determined values by around a factor of $10^1 - 10^4$ [14–17]. Note, the spherical assumption is quite good for describing the motion in a uniform fluid; the problem is that when they collide, the collision alignment of non-spherical objects tends to be highly dependent on the deviations from a sphere. If the goal is to make predictions about a specific protein, I would recommend empirically determining the association rates or if that is not an option, carefully creating a collision model that accurately describes the particular systems of proteins' association rates. Since my goal is more general, I will fix the association rate used in all simulations to $10^6 \text{ (M}^{-1}\text{s}^{-1})^*$. Values in the range of $10^5 - 10^6 \text{ M}^{-1}\text{s}^{-1}$ have been cited as empirically representative of a reasonable breadth of proteins [14–19]. The formal system I am constructing is not limited by the parameters I am choosing; they are only for the concrete examples that I will provide.

OK, things can now come together, why do they fall apart? Elucidation of dissociation, requires an enumeration of the energetics that bind them together. As the world is full of ways for objects to be bound together, but only a couple are in the right energy scale for life. Let us start by removing some forces, *e.g.*, strong, weak, gravity, whose energetics are clearly outside the domain of interest for macromolecular interactions. I will now break the relevant forces into two primary categories. First, there is the electrostatic based intermolecular forces: hydrogen bonding, Van der Waals, covalent, ionic, and metallic bonds. Covalent, ionic, and metallic rarely apply, and in the special cases when they would be applicable, considerations beyond the scope of this work are needed [20]. Second, there

*Molar Concentration ($\text{M} = \text{mol/liter} = \text{mol/dm}^3$) is being used. The proper SI unit is mol/m^3 , but due to its impracticality in experimental contexts, it is rarely seen in use in the biological sciences.

is an entropy term due to the restriction in the rotational and positional freedom of a protein when bound vs unbound [18]. Finally, the details of the fluid do come into play at this point, assuming the fluid is polar, *i.e.*, water. There is an entropic cost associated with the interactions of water molecules with the protein's surface area, as the polar fluid will align itself in a shell to shield the bulk fluid from the protein [20, 21]. The average rate of protein:protein bond dissociation, is determined by the net energy of these effects [22, 23]. Association and dissociation form a dynamic balance from which life emerges. Unfortunately, the sheer complexity of a network encompassing even a single platonic human, is not only beyond the scale of this document, it is beyond the scale of my current understanding. So, let us start with something a smidge more tractable.

1.1 A:B Binding

Let us take a collection of effectively identical proteins, A, and a collection of distinctly different proteins, B, which can bind together to form into complexes of type AB. I can lay out a scheme of their association and dissociation:



Using standard mass action kinetics [24], the likelihood that A binds B is proportional to the concentration of A, the concentration of B and the association rate constant k_{on} ($\text{M}^{-1}\text{s}^{-1}$). Whereas, the likelihood of their dissociation is the concentration of the complex AB times the dissociation rate, k_{off} (s^{-1}). All reactions in this work are considered to be reversible, even if the rate of the reaction in one direction is minuscule, a dissociation rate will always be assigned. Reversible reactions are represented by the \rightleftharpoons symbol. $[\]$ represent concentration of the included symbol. Additionally, a concentration without a subscript implies time dependence. The total concentration, including both bound and unbound states, will be denoted by subscript "Total." "Steady State" concentrations will

be denoted by a subscript SS. For complete clarity, k_{on} , k_{off} are nonnegative real valued constants. It follows that for Scheme 1.1 the following equations fully describe the reaction kinetics:

$$\begin{aligned}
\frac{d[\text{AB}]}{dt} &= k_{\text{on,AB}}[\text{A}][\text{B}] - k_{\text{off,AB}}[\text{AB}] \\
\frac{d[\text{A}]}{dt} &= k_{\text{off,AB}}[\text{AB}] - k_{\text{on,AB}}[\text{A}][\text{B}] \\
\frac{d[\text{B}]}{dt} &= k_{\text{off,AB}}[\text{AB}] - k_{\text{on,AB}}[\text{A}][\text{B}] \\
[\text{A}]_{\text{Total}} &= [\text{A}] + [\text{AB}] & [\text{AB}](0) &= 0 \\
[\text{B}]_{\text{Total}} &= [\text{B}] + [\text{AB}] & [\text{A}](0) &= [\text{A}]_0 \\
& & [\text{B}](0) &= [\text{B}]_0
\end{aligned} \tag{1.2}$$

Additionally, for this case, it is possible to provide a complete closed form solution of the system [25]:

$$\begin{aligned}
K_d &:= \frac{k_{\text{off,AB}}}{k_{\text{on,AB}}} = \frac{[\text{A}]_{\text{SS}}[\text{B}]_{\text{SS}}}{[\text{AB}]_{\text{SS}}} \\
z_{\text{AB}} &= \sqrt{([\text{A}]_0 - [\text{B}]_0)^2 + K_d^2 + 2K_d([\text{A}]_0 + [\text{B}]_0)} \\
[\text{AB}] &= \frac{2[\text{A}]_0[\text{B}]_0}{[\text{A}]_0 + [\text{B}]_0 + K_d + z_{\text{AB}} \coth\left(\frac{1}{2}z_{\text{AB}}k_{\text{on}}t\right)} \\
[\text{A}] &= [\text{A}]_0 - [\text{AB}] \\
[\text{B}] &= [\text{B}]_0 - [\text{AB}]
\end{aligned} \tag{1.3}$$

Time dependent closed form solutions for reaction kinetics are available for only a few reaction patterns. Traditionally a steady state solution is considered. This can be found quite easily from the limit of the time dependent equations or by exploiting the total species constraint equations, in this case $[A]_{\text{Total}}$ and $[B]_{\text{Total}}$, which yields:

$$\left. \begin{array}{l} \frac{d[AB]}{dt} = 0 \\ \frac{d[A]}{dt} = 0 \\ \frac{d[B]}{dt} = 0 \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} K_d := \frac{k_{\text{off},AB}}{k_{\text{on},AB}} = \frac{[A]_{\text{SS}}[B]_{\text{SS}}}{[AB]_{\text{SS}}} \\ [AB]_{\text{SS}} = \frac{[A]_0 + [B]_0 + K_d - \sqrt{([A]_0 + [B]_0 + K_d)^2 - 4[A]_0[B]_0}}{2} \\ [A]_{\text{SS}} = [A]_0 - [AB]_{\text{SS}} \\ [B]_{\text{SS}} = [B]_0 - [AB]_{\text{SS}} \end{array} \right. \quad (1.4)$$

I have defined K_d (M) for three reasons. First, it is common to find k_{on} and k_{off} as a ratio. Second, it tends to be technically quite difficult to experimentally determine k_{on} and k_{off} . Third, K_d is the dominant way binding strength is described in the field. Many people argue for K_d 's importance, due to the relation, if $[A] = K_d$ then $[B] = [AB]$ thus half of B is in a bound state, note this is only achievable when $[A]_0 > [B]_0$. **Caution** needs to be exercised, as this relation is not ensured in reactions more complicated than this binary binding reaction, *e.g.*, coupled sequences of reactions and higher order reactions.

K_d is handy for comparing binding *strength* between reactions. The common form is to describe a reaction as stronger if the K_d is smaller, as this results in a higher tendency to be in the bound state. I will introduce one more term before I start discussing the results of these equations. Yield is the ratio of the current quantity of a species to the maximum possible quantity of the species, *i.e.*, in the case of $[AB]$ the maximum is the smaller of $[A]_0$ or $[B]_0$. Yield allows for a consistent trait that can be compared across multiple types of binding relations.

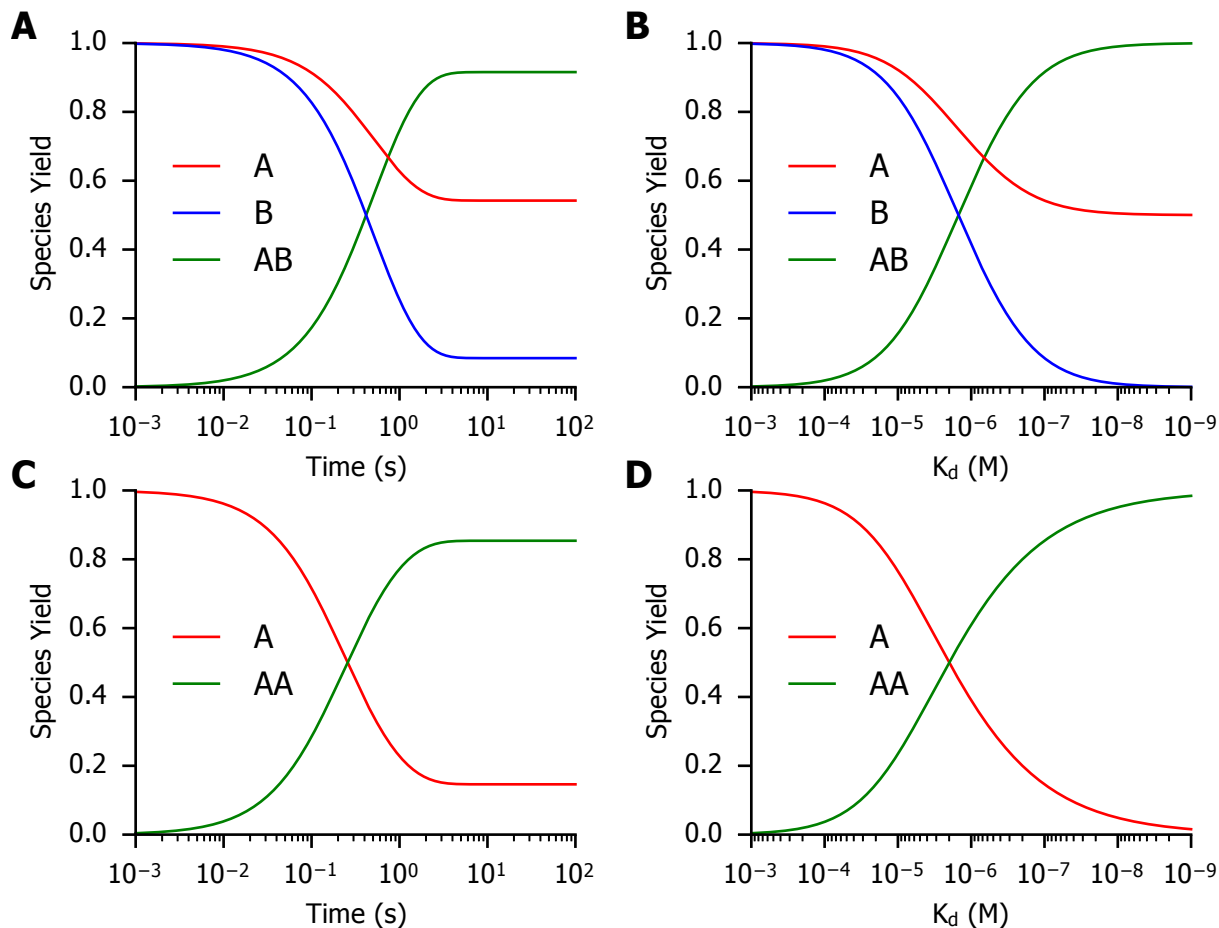


Figure 1.1: Species Yield in One to One Binding

K_d is commonly plotted from larger to smaller, as smaller values represent *stronger* binding. (A) Time dependent species yield solutions for the A:B binding relationship (Eqs. 1.3) shown with parameters $k_{\text{on}} = 10^6 \text{ M}^{-1}\text{s}^{-1}$, $K_d = 10^{-7} \text{ M}$, $[A]_0 = 2 \times 10^{-6} \text{ M}$, and $[B]_0 = 10^{-6} \text{ M}$. (B) Steady state species yield solutions for the A:B binding relationship (Eqs. 1.4) shown with parameters $[A]_0 = 2 \times 10^{-6} \text{ M}$ and $[B]_0 = 10^{-6} \text{ M}$. (C) Time dependent species yield solutions for the A:A binding relationship (Eqs. 1.9) shown with parameters $k_{\text{on}} = 10^6 \text{ M}^{-1}\text{s}^{-1}$, $k_{\text{off}} = 10^{-1} \text{ s}^{-1}$, $K_d = 10^{-7} \text{ M}$, and $[A]_0 = 2 \times 10^{-6} \text{ M}$. (D) Steady state species yield solutions for the A:A binding relationship (Eqs. 1.10) shown with $[A]_0 = 2 \times 10^{-6} \text{ M}$.

Figure 1.1A illustrates several features that will be key throughout this discourse. Equilibrium is not a complete conversion to either side of the reaction. There is always a dynamic balance between both sides of the reaction. The equilibrium level can be dominated by either K_d or by a limiting species. Figure 1.1B shows these regimes, when K_d is substantially above 10^{-8} the equilibrium is determined primarily by K_d , whereas when K_d is below 10^{-8} the reaction is limited by [B]. Clearly, the details of these regimes are dictated by the particular parameters of the reaction, but the general behaviors will always be present. All curves show a sigmoidal form; asymptotic approaches to equilibrium are common to all reversible reactions.

A technicality: I will only be considering concentration, since the species are discrete objects their behaviors at extremely low concentrations, in a finite volume, become stochastic in nature. Dynamic equilibrium is still present, so at some points there could be zero molecules of a species, but for some greater time point the species will be present. Appropriate treatment of behavior in the stochastic regime is beyond the scope of this work.

Equilibrium results, such as Fig. 1.1B and Fig. 1.2, are both simpler to calculate and experimentally easier to measure. While equilibrium results are important, the transient behavior can elucidate processes that are lost in the equilibrium. Biological systems do not always reach equilibrium, within a relevant timescale. In Chapter 3, I will introduce a model system that is still showing significant non-periodic transient behavior at 10^{12} seconds, far exceeding any known normal biological process. Thus, we must be careful in the application of steady state solutions for biological contexts. Context should always be the deciding factor in the application of simplifications.

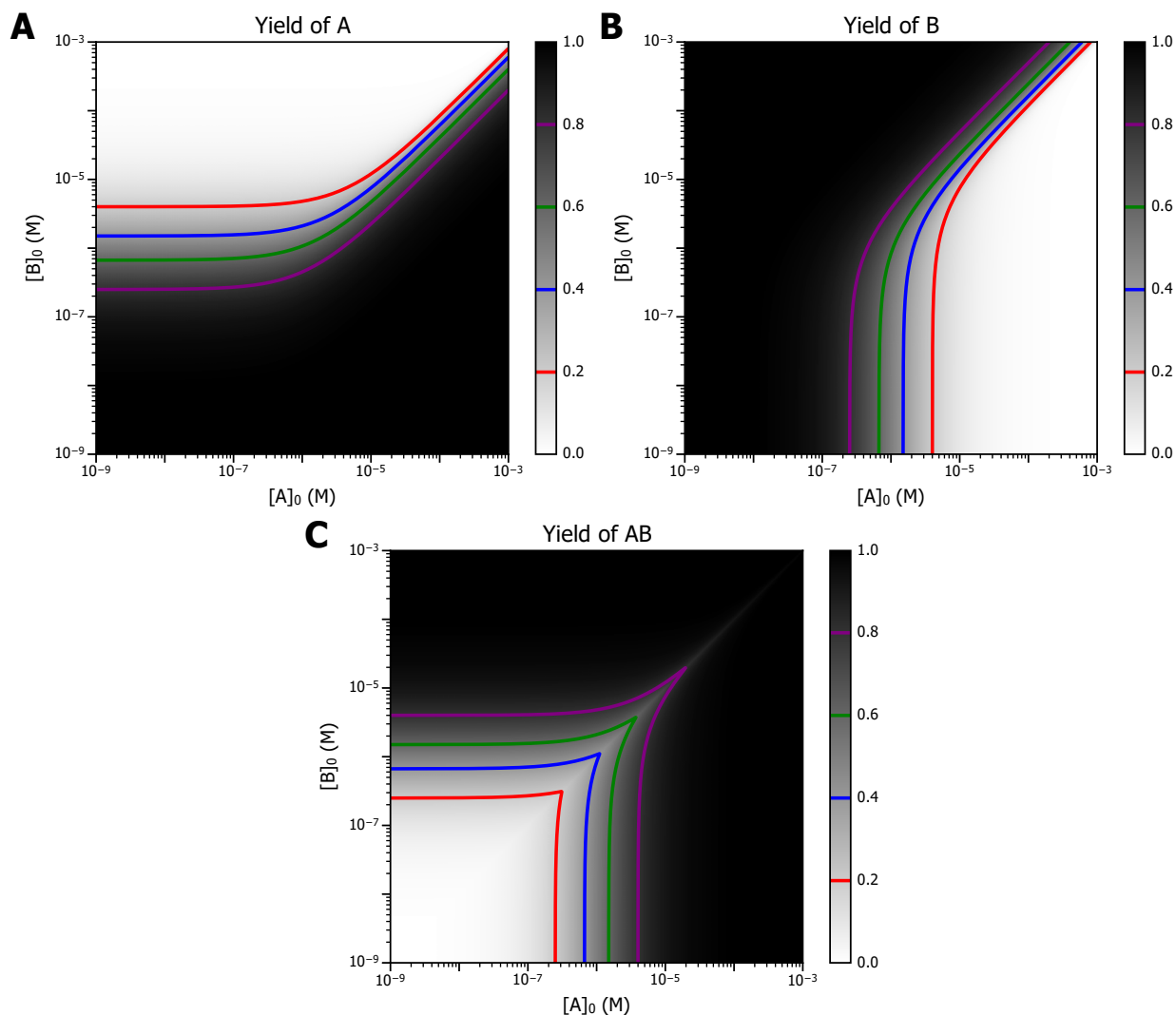


Figure 1.2: Species Yield vs Initial Concentration

Generated from Eqs. 1.4 with $K_d = 10^{-6}$ M for all plots. Black regions represent a high equilibrium yield of the species, whereas white is a depletion of the species. Contour lines demarcate several points in the transition gradient. (A) Regions where the reaction is A limited can be seen above the red contour line. (B) The system is symmetrical between A and B, thus this plot illustrates both that relationship and the region where $[A]$ and $[B]$ are below K_d . (C) Yield is normalized by the maximum possible quantity of the species. As this is calculated at every point, AB is limited by A when $[A]_0 > [B]_0$ and limited by B when $[B]_0 > [A]_0$

1.2 A:A Binding

I will now introduce a binding relation that upon first glance, appears similar to Section 1.1. These are the two elementary operations in the formalism I am presenting. It is important to understand them in detail before I begin discussing the construction of general patterns produced from these operations and the emergent properties of those patterns. As opposed to the binding of two distinct species I will now present the behavior of a self-binding species:



The symmetry of this system immediately leads to several differences in the construction of the equations describing its behavior. Unlike AB, AA contains two copies of the same protein, so the equation for $[AA]_{\text{Total}}$ becomes:

$$[A]_{\text{Total}} = [A] + 2[AA] \quad (1.6)$$

Now by differentiating both sides of Eq. 1.6, I can show a key difference in the relationship between the rates of the system [26]:

$$\begin{aligned} [A]_{\text{Total}} = [A] + 2[AA] &\Rightarrow 0 = \frac{d[A]}{dt} + 2\frac{d[AA]}{dt} \\ &\Rightarrow \begin{cases} \frac{d[AA]}{dt} = k_{\text{on,AA}}[A]^2 - k_{\text{off,AA}}[AA] \\ \frac{d[A]}{dt} = 2k_{\text{off,AA}}[AA] - 2k_{\text{on,AA}}[A]^2 \end{cases} \quad (1.7) \end{aligned}$$

Determination of this ratio can be accomplished in many ways; this method was chosen for its generality. With that, I now model reaction Scheme 1.5 with the equations:

$$\begin{aligned}
\frac{d[AA]}{dt} &= k_{\text{on},AA}[A]^2 - k_{\text{off},AA}[AA] \\
\frac{d[A]}{dt} &= 2k_{\text{off},AA}[AA] - 2k_{\text{on},AA}[A]^2 \\
[A]_{\text{Total}} &= [A] + 2[AA] \quad [AA](0) = 0 \\
&\quad [A](0) = [A]_0
\end{aligned} \tag{1.8}$$

I will now present the time dependent solution of Eqs. 1.8 [25]:

$$\begin{aligned}
K_d &:= \frac{k_{\text{off},AA}}{k_{\text{on},AA}} = \frac{[A]_{\text{SS}}^2}{[AA]_{\text{SS}}} \\
z_{AA} &= \sqrt{K_d(K_d + 8[A]_0)} \\
[AA] &= \frac{K_d + 4[A]_0 - z_{AA} \tanh\left(\frac{1}{2}z_{AA}k_{\text{on}}t + \tanh^{-1}\left(\frac{K_d + 4[A]_0}{z_{AA}}\right)\right)}{8} \\
[A] &= [A]_0 - 2[AA]
\end{aligned} \tag{1.9}$$

A selection of solutions to Eqs. 1.9 are shown in Figs. 1.1C and 1.1D. There is no substantial shift in general pattern, but the scaling has been shifted. As before, I will now derive the steady state equations for Scheme 1.5:

$$\left. \begin{aligned} \frac{d[AA]}{dt} &= 0 \\ \frac{d[A]}{dt} &= 0 \end{aligned} \right\} \Rightarrow \left\{ \begin{aligned} K_d &:= \frac{k_{\text{off},AA}}{k_{\text{on},AA}} = \frac{[A]_{\text{SS}}^2}{[AA]_{\text{SS}}} \\ [AA]_{\text{SS}} &= \frac{K_d + 4[A]_0 - \sqrt{K_d(K_d + 8[A]_0)}}{8} \\ [A]_{\text{SS}} &= [A]_0 - 2[AA]_{\text{SS}} \end{aligned} \right. \tag{1.10}$$

Note the relation, if $[A]_{\text{SS}} = K_d$ then $[A]_{\text{SS}} = [AA]_{\text{SS}}$, no longer implies half is in a bound state, it results in a third of A is in a bound state. Additionally, this relation is uniquely valid at $[A]_0 = 3K_d$. Whereas, half bound is achieved when $[A]_{\text{SS}} = \frac{1}{2}K_d$ resulting in

$[AA]_{SS} = \frac{1}{4}K_d$ and occurs when $[A]_0 = K_d$. These details are especially important when both reactions have related underlying kinetic parameters, such as k_{on} and k_{off} .

1.3 Synthesis and Degradation

In living systems, there is an introduction of species into the system and a removal of species from the system. This constant cycling is a fundamental trait of life. The equations I have presented up to this point describe systems more representative of systems we would see in an experimental setting and thus I will refer to them as *in vitro*-like. To better represent *in vivo*-like systems, I will be adding two concepts: Synthesis, Q_X ($M^{-1}s^{-1}$), to represent a constant introduction of a specific species represented by X , and Degradation, δ (s^{-1}), to represent the constant removal of all species. Synthesis is primarily accomplished by cellular functions introducing a species into the system, thus I will approximate this by a constant term in the differential of the minimal structure for the system. Degradation in living systems can take many forms, with a variety of mathematical descriptions being admitted [18]. I will be using the model where loss is dominated by the dilution of the system. Dilution is the dominant effect in rapidly dividing cells, which is also when assembly efficiency is often critical to survival. Dilution is modeled as a rate, δ (s^{-1}), proportional to concentration, for every species. I have adjusted Scheme 1.5's equations to account for an *in vivo*-like system:

$$\begin{aligned}
 \frac{d[AA]}{dt} &= k_{on,AA}[A]^2 - k_{off,AA}[AA] - \delta[AA] \\
 \frac{d[A]}{dt} &= 2k_{off,AA}[AA] - 2k_{on,AA}[A]^2 + Q_A - \delta[A] \\
 [A]_{Total,SS} &= [A]_{SS} + 2[AA]_{SS} & [AA](0) &= 0 \\
 K_d &:= \frac{k_{off,AA}}{k_{on,AA}} = \frac{[A]_{SS}^2}{[AA]_{SS}} & [A](0) &= 0 \\
 & & [A]_{Total,SS} &= \frac{Q}{\delta}
 \end{aligned} \tag{1.11}$$

In the steady state limit, $[A]_{\text{Total,SS}}$ is equivalent to $[A]_0$ from the *in vitro*-like formulation. This relationship will be used when referring to concentration in *in vivo*-like experiments to ease comparisons between the models. Care should be taken when considering $[A]_{\text{Total,SS}}$ in pre-steady state conditions. For example, consider the regime when $Q_A > t[A]_{\text{Total,SS}}$, then $[A]_{\text{Total,SS}} > [A]_{\text{Total}}$ as $[A]_{\text{Total,SS}}$ is greater than the concentration of A introduced into the system at that point.

I have now introduced the two fundamental operations of the assembly pathway formalism. In the next chapter I will be extending these operations to include chains and branching of binding events, Chapter 2. In Chapters 3 and 4, I will demonstrate how to form sets of assembly trees from these elementary operations. Then I will conclude with an example of how an alternate binding partner can disrupt the system, Chapter 5.

ISWI:NCP Binding

“Ten thousand times the web could be destroyed, and ten thousand times the spider would rebuild it. There was neither annoyance nor despair, nor any delight, just as it had been for a billion years.”

— Cíxīn Liú, *The Dark Forest*

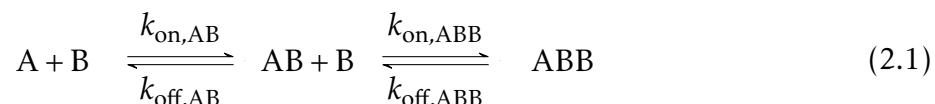
Let us consider the binding relationships introduced in Chapter 1. Envision the state before binding as an island. Then the state after binding as a different island. Now the option for binding can be thought of as a bridge between those islands with a capacity in each direction proportional to the kinetic rates (k_{on} and k_{off}). Just having a bridge does not imply that anything will traverse the bridge, there must be something on one side which would favor being on the other. This metaphor may seem familiar, as the origin of Graph Theory began with the Seven Bridges of Königsberg [27]. I will not be invoking a full graph theoretic formalism, as geometric constraints will play a major role in the determining of allowable connections, but I would like to invite the reader to apply the same principles of abstraction to the flows and states within the dynamic system. With this concept in mind, I will now begin construction of a second bridge on the first island.

To illustrate this iteration of the formalism, I will use an experimental project from Dr. Fischer’s lab as a case study. Dr. Al-Ani led the experimental work, while I focused on analysis of the results and modeling of the system [28–33]. The question on the table was “How does ISWI bind to an NCP?” Imitation SWItch (ISWI) is a large chromatin remodeling protein [34–37]. Whereas, Nucleosome Core Particles (NCPs) are highly

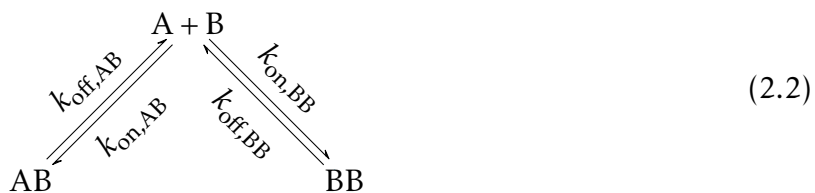
stable macromolecular complexes combining proteins and DeoxyriboNucleic Acid (DNA) [38–43]. The biological roles of both have been well covered in the literature, for this document’s purposes, the importance is their binding relationships. The ISWI:NCP system provides an excellent real-world example for the next two extensions to the formalism: Sequential Binding Events (SBEs) and Branching.

2.1 Sequential Binding Events and Branching

First, I will extend the basic models from Chapter 1 to demonstrate these extensions. SBEs admit a potential to have a sequence of binding relationships. A simple example of this would be if AB from Eqs. 1.2 could bind to B, resulting in:

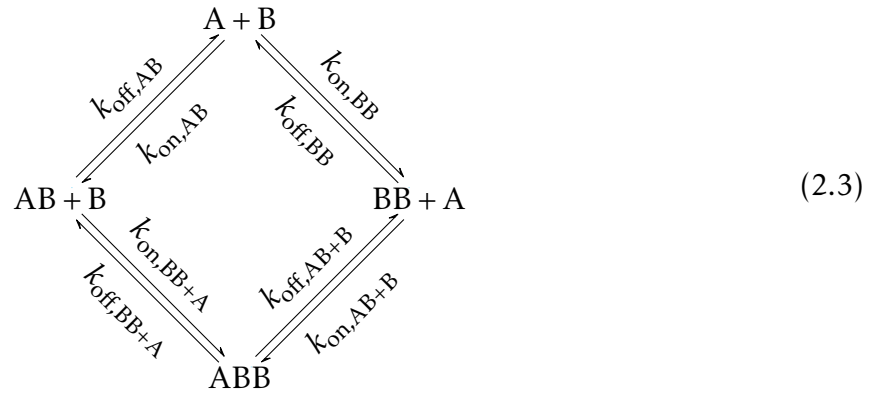


There is an important caveat to this scenario: AB must be an obligate precursor to ABB. The problems arise from potential substructures and symmetries present in the structure ABB. For example, if B can bind to B without A present then Scheme 2.1 should account for that. This issue leads directly into the second extension: branching is the option for a species to have more than one candidate for binding. For instance, branching of $A + B$ to either AB or BB, but not ABB, would look like this:



Only the branching step is shown. If ABB were admitted without the AB obligate precursor requirement, I would need to combine both extensions. For this particular example, both branches end up returning to the same state, so in effect I will be using an SBE of a branching with a reverse branch. Everything about binding in this formalism can be

thought of as symmetric with respect to path, as all reactions are reversible. Thus, a reverse branch is just a branching from the opposite point of view. Drawn in scheme form, the scheme would appear as:



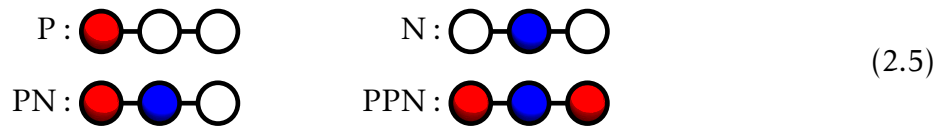
As each of the four *nodes* in Scheme 2.3 are unique, this is the simplest way to draw this reaction. Scheme 2.3 also happens to be the first time a *cycle* has appeared. Cycles provide an opportunity to use the concept of detailed balance: at equilibrium, every elementary process must be balanced by its reverse process [44, 45]. Now recall that rate constants are independent of concentration, thus for this cycle, detailed balance gives the invariant [26, 46, 47]:

$$k_{\text{on},BB} k_{\text{on},AB+B} k_{\text{off},BB+A} k_{\text{off},AB} = k_{\text{on},AB} k_{\text{on},BB+A} k_{\text{off},AB+B} k_{\text{off},BB} \quad (2.4)$$

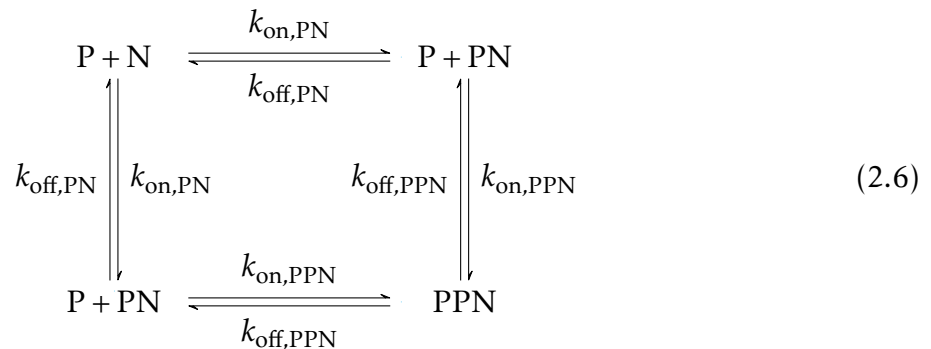
As I am not allowing rate constants to be dependent on anything in the system that changes between transient and steady state, this can be generalized to any time point in the system.

2.2 Imitation SWItch (ISWI)

Now, I will return to the ISWI:NCP system. Experimental evidence led us to a model in which a single NCP (N) has the capacity to bind two ISWIs (P). I have graphically illustrated the four possible species:

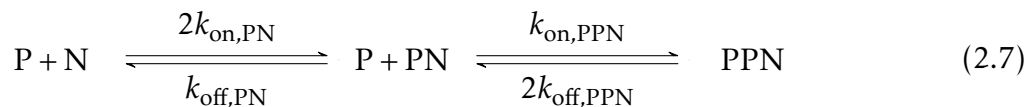


NCPs are shown in blue, ISWI is shown in red, and unfilled circles represent unoccupied binding sites. Unlike the ABB example, here we have a symmetry. *A priori* we do not know if there is a difference between binding to the two sites, hence there is only one species PN. If I proceed as before and write the binding scheme, I once again will want a branch followed by a reverse branch:



Unlike last time, two of the nodes will be equal due to the symmetry of PN. This situation is a special type of branch, as it is describing two ways to proceed to the same state. As the endpoints are the same, this can be represented by collapsing the branch into a linear sequence of binding reactions and doubling the rate constant towards the branch [18]. Remember, fully reversible reaction schemes do not include the concept of direction, so branching can happen in a way that we may perceive as reversed. Therefore, both sides of

the split can be collapsed in the same way:



There are several ways to interpret the introduction of the factors of two, introduced in the collapsing of Scheme 2.7. Here are several of the most relevant options. First, I can invoke the relation in Eqs. 1.7 as this system takes the same general form. Second, I can make a symmetry argument: there are two ways for P to bind to PN, but only one way for P to unbind. Additionally, the same is true for PPN: there are two ways for P to unbind from PPN, but only one way for P to bind. Finally, the general approach within this formalism is to consider all branches with identical endpoints to be collapsible: the links can be replaced with a single link, by adding a coefficient to rate constant equal to the number of links from before the collapsing action, *i.e.*, in this case there is a factor of two added to $k_{\text{on,PN}}$ and $k_{\text{off,PPN}}$ as they both have two links with P + PN at the other end that start from the same place.

Determination of whether to incorporate these factors, such as the factor of two preceding $k_{\text{on,PN}}$ and $k_{\text{off,PPN}}$, into the rate constants or to leave them in the equation as coefficients are a matter of what emphasis is intended. All methods are mathematically equivalent, but the resulting equations and constants can generate different implied meanings to the reader. For complete openness, my desire is to ensure the rate constants are dominated by the site-to-site binding energetics, not the combinatorics of the system. This is important in the upcoming chapters, as it allows an effective K_d to be varied in the system to represent binding site strength. Although, it would also be reasonable to include the factors in the term, as that would be more representative of what would be expected in most experimental contexts. These variations are part of what leads to terminology such as microscopic, macroscopic, step-wise. Usage of this terminology is often viewed in different subgroups as being consistent across the field, but in my experience, the wide adoption of consistent definitions is vastly overestimated. Hence, I define my terms by the

mathematical descriptions I provide, as opposed to invoking a specific name. With that clarification in mind, I present the equations governing the ISWI system:

$$\begin{aligned}
\frac{d[\text{PPN}]}{dt} &= k_{\text{on,PPN}}[\text{P}][\text{PN}] - 2k_{\text{off,PPN}}[\text{PPN}] \\
\frac{d[\text{PN}]}{dt} &= 2k_{\text{off,PPN}}[\text{PPN}] + 2k_{\text{on,PN}}[\text{P}][\text{N}] - k_{\text{off,PN}}[\text{PN}] - k_{\text{on,PPN}}[\text{P}][\text{PN}] \\
\frac{d[\text{P}]}{dt} &= k_{\text{off,PN}}[\text{PN}] + 2k_{\text{off,PPN}}[\text{PPN}] - 2k_{\text{on,PN}}[\text{P}][\text{N}] - k_{\text{on,PPN}}[\text{P}][\text{PN}] \\
\frac{d[\text{N}]}{dt} &= k_{\text{off,PN}}[\text{PN}] - 2k_{\text{on,PN}}[\text{P}][\text{N}]
\end{aligned} \tag{2.8}$$

$$\begin{aligned}
[\text{P}]_{\text{Total}} &= [\text{P}] + [\text{PN}] + 2[\text{PPN}] & \text{PPN}(0) &= 0 \\
[\text{N}]_{\text{Total}} &= [\text{N}] + [\text{PN}] + [\text{PPN}] & \text{PN}(0) &= 0 \\
K_{d,\text{PN}} &:= \frac{k_{\text{off,PN}}}{k_{\text{on,PN}}} = 2 \frac{[\text{P}]_{\text{SS}}[\text{N}]_{\text{SS}}}{[\text{PN}]_{\text{SS}}} & \text{P}(0) &= [\text{P}]_0 \\
K_{d,\text{PPN}} &:= \frac{k_{\text{off,PPN}}}{k_{\text{on,PPN}}} = \frac{1}{2} \frac{[\text{P}]_{\text{SS}}[\text{PN}]_{\text{SS}}}{[\text{PPN}]_{\text{SS}}} & \text{N}(0) &= [\text{N}]_0
\end{aligned}$$

I have now reached the point where the determination and display of the explicit solution for this and upcoming models is impractical. A closed form analytic solution for steady state conditions was determined for use in the fitting of empirical data. Displaying it here is impracticable at best and does not substantially improve understanding of the system.

As this case study comes from the experimental elucidation of the model, it is important to present the linkage between the abstract model and the experimental observables. In this case, I will only be covering the experimental evidence which is most relevant to the dynamics of the reaction network. Fluorescence anisotropy is an experimental technique, which measures the ratio of the intensities of perpendicularly polarized light [48]. The experimental technique allows for the inference of the bound state of fluorophore labeled species. For these experiments, the NCPs were labeled with Alexa Fluor™ 488. For details about this experimental system please see Al-Ani *et al.* [28, 33]. General details on the method of fluorescence anisotropy have been covered in detail by Dr. Lakowicz [48, 49].

The key property of fluorescence anisotropy, for the relevance of this work, is the ability to detect separate species in equilibrium. Results presented in this document are from steady state ensemble measurements, which result in an average fluorescence anisotropy value of [48]:

$$\bar{r} = \sum_i f_i r_i \quad (2.9)$$

Where i iterates over all fluorescent species, f_i is the fractional fluorescence intensity, and r_i is the steady state fluorescence anisotropy of the species, resulting in f_i acting as an excellent proxy for the fractional concentration. Then, by defining Δr to be $\Delta r = \bar{r} - \bar{r}_0$, where \bar{r}_0 is the measured value when $[P] = 0$ is added, I can remove the effect of $[N]$, while benefiting from \bar{r}_0 being an easy to accurately measure value. By not adding P to the initial mixture, measurements of \bar{r} due only to N can be measured under a variety of conditions and in an easily repeatable fashion. I now have a well-behaved experimental observable for determining $[PN]$ and $[PPN]$:

$$\Delta r = \frac{[PN]}{[N]_0} r_{PN} + \frac{[PPN]}{[N]_0} r_{PPN} \quad (2.10)$$

For each data point, a known concentration of P and N were added together and allowed to reach approximate equilibrium. The fluorescence anisotropy was then measured. This completes the prerequisites to determine the remaining equilibrium constants and thus elucidate the binding of ISWIs to NCPs.

The system's parameters were fit with a custom application I wrote in Python and WX, with the capabilities to perform: GUI driven, nonlinear data fitting, with Monte Carlo uncertainty propagation for modular data models. The program was created, but as colleagues completed their degrees and moved on, so did the needs of the lab, and the project was ultimately abandoned. The source code will reside for a reasonable period of time in the GitHub repository: <https://github.com/vatir/tuva> [50–58]. Figure 2.1A shows the results of that fitting, while the parameters are shown in Table 2.1.

Table 2.1: ISWI Binding Parameters with 68% Confidence Bounds

$K_{d,PN}$	2.6 ± 1.2 nM	Dissociation Constant for PN
$K_{d,PPN}$	5.0 ± 1.4 nM	Dissociation Constant for PPN
r_{PN}	0.073 ± 0.008	Fluorescence Anisotropy for PN
r_{PPN}	0.121 ± 0.003	Fluorescence Anisotropy for PPN

Table 2.1 combined with Eqs. 2.8 and 2.10 provides a complete model of the steady state binding of ISWI. $K_{d,PN}$ and $K_{d,PPN}$ are within uncertainty of each other, as I kept the combinatorial factors in the binding equations instead of the kinetic rates. This implies that the difference in strength between the binding sites and the importance of binding order is not significant for this system. Now, it is quite likely the site-specific binding effects are somewhat different, just not significant. The result does support my assumption, that for a general model we can treat the sites as independent. Though, care should be taken if small molecule binding is involved, as this assumption does not hold even for ISWI. In the same work that Dr. Al-Ani and I carried out, we determined interactions between different binding sites when a small molecule was involved [28]. For a more general treatment, see Supplementary Reading: Thermodynamics of Binding Site Interactions.

One of the many benefits of having a binding model for a system is the ability to predict the binding state of species in the mixture. Figure 2.1B shows an example of what can be predicted from knowing the initial concentrations of the species added to an experiment. Thereby, one can tailor experiments by controlling the concentration of species present. Additionally, results can be properly normalized for real concentration of the target species. For example, it would be extremely difficult to measure an interaction between something and the PN state as opposed to the PPN state without a model of this type.

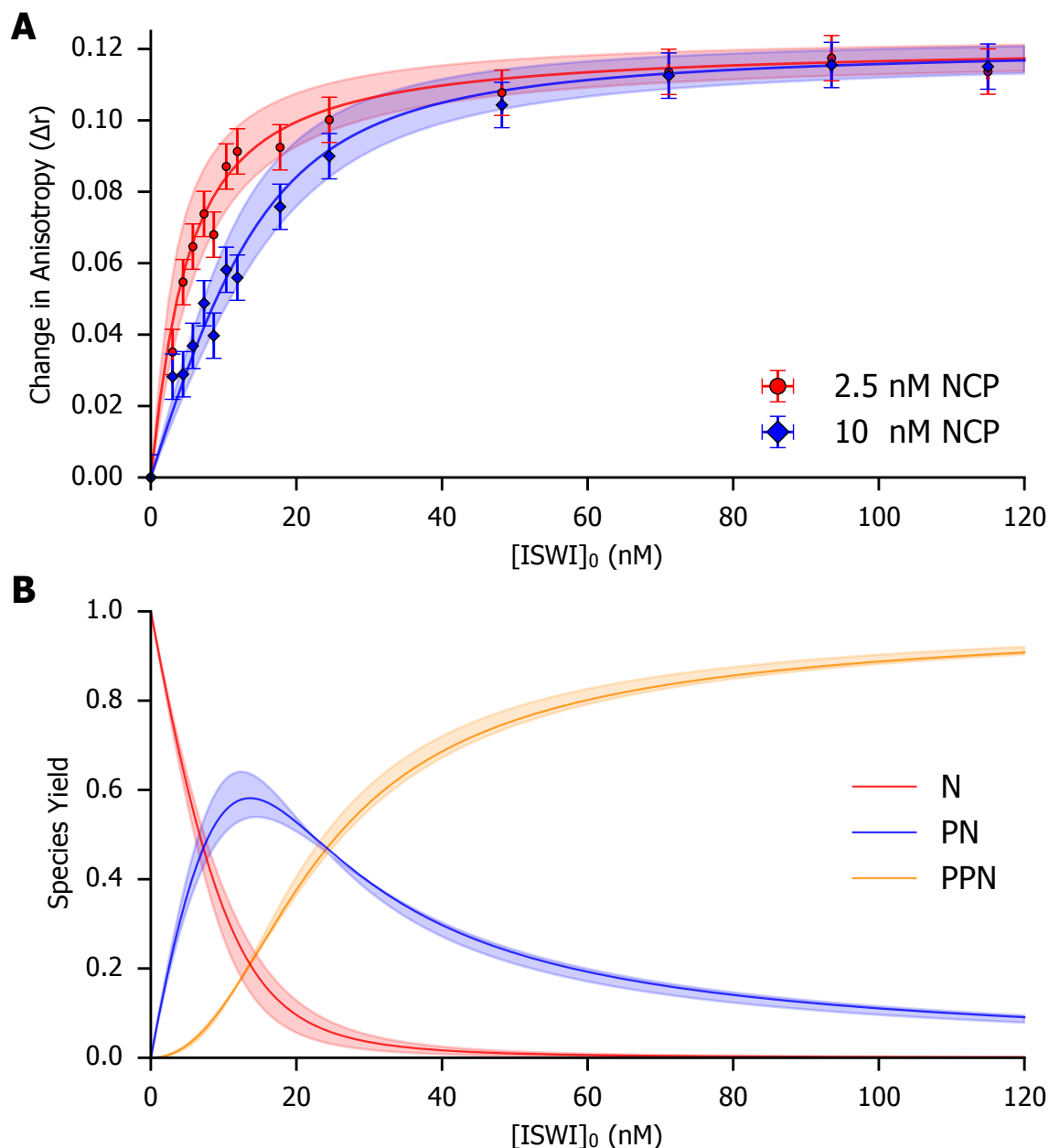


Figure 2.1: ISWI:NCP Binding

All species fractions are determined using parameters from Table 2.1. Shaded regions denote a 68% confidence range, where upper and lower bounds are derived from maximum and minimum parameter bounds, respectively. (A) Change in Anisotropy vs Initial ISWI with lines from Eq. 2.10. NCP is the starting concentration added to each experiment. Experimental results collected by Dr. Al-Ani [28]. Δr uncertainty is calculated at 68% confidence, by experimental repetition. Protein concentration uncertainty is calculated at 68% confidence by analysis of experimental apparatus, as this was determined to be the dominate source of uncertainty: error bars are smaller than data markers in most cases. (B) Species Yield vs Initial ISWI concentration, as resulting from Eqs. 2.8.

Stacked Trimer

“I like silent pictures and I always have. They are often so much more beautiful than sound pictures are. Perhaps they had to be. At any rate I wanted to restore some of this beauty. I thought of it, I remember in this way: one of techniques of modern art is simplification, and that I must therefore simplify this film.”

— Kurosawa Akira

All the elementary operations are now available. It is time to put the operations together and see what forms. By combining A:A (Section 1.2), A:B (Section 1.1), and SBEs (Section 2.1) a countably infinite class of binding relations can be formed, covering the general structure of all protein complexes. Now I need to start adding restrictions to reduce the cardinality of the space and add structure. Without that, the system is far too general to elucidate any specific system. These restrictions take two forms. First, by bounding the possible complexes allowed in the system, the cardinality is reduced to finite. Second, equivalence relations are used to form equivalence classes, *e.g.*, collapsing of identical branches as mentioned in Section 2.1. Bounding is required for most analysis methods used on these systems, whereas equivalence classes are aids to the understanding and computability of the system’s structure. As such, I will introduce several equivalence classes that I feel are widely relevant, but others are not only possible, they may be critical to the understanding of a specific system.

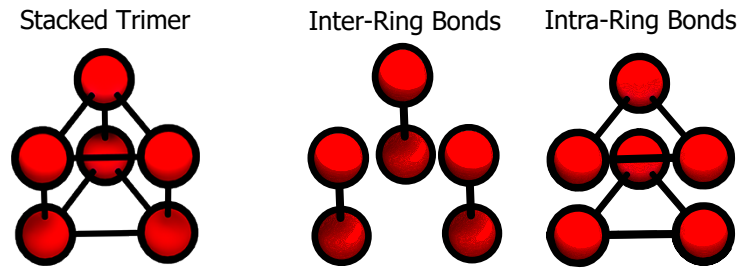


Figure 3.1: Stacked Trimer Diagram & Bond Types

3.1 Technical Description

Concepts in this chapter will be introduced through the use of the stacked trimer, illustrated in Fig. 3.1. Stacked trimers are protein complexes containing six identical proteins, homo-hexameric. Stacked trimers are symmetric under the dihedral group of order three symmetries (D_3), the same as an equilateral prism [59]. D_3 includes 3-fold rotation symmetry in the xy -plane and reflection symmetry across the z -plane (Table 3.1) [60, 61]. Subspecies of the stacked trimer can be reduced to a minimal number of classes by using the equivalence relations in D_3 . Bond types in the stacked trimer are dictated by symmetry; each protein contains three binding interfaces. Protein interfaces come about by alignment of patterns chemical relations within the binding proteins. Therefore, most binding interfaces can not be arbitrarily rotated with respect to the axis central to the bond [62]. Thus, I will impose the condition that these interfaces have the minimum conditions that still admit the structural symmetries, *i.e.*, there are three binding sites on each protein (A, B, and C) which bind (A to B) and (C to C') where C' is the reflection of site C. D_3 is the union of the subgroup cyclic group of order three (C_3) and a reflection symmetry.* Thus, I will begin by describing the protein properties induced by C_3 ; stacked trimers contain a rotationally symmetric three-member ring, hence the name stacked trimer. To form this ring I assign site A to be 120° from site B to naturally close the ring and yield the

*Protein complexes with dihedral symmetry can be bound entirely with isologous association sites or a mix of bond types as in Table 3.1, such that intra-ring bonds are heterologous and inter-ring bonds are isologous [60]. It should be noted, that the details of protein-protein binding can be very complicated in practice, so even this distinction is falling out of favor due to its oversimplification of the issue [61].

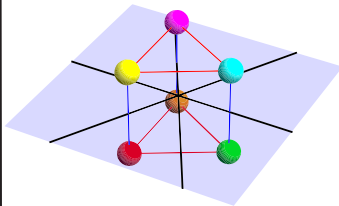
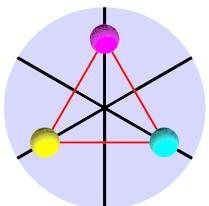
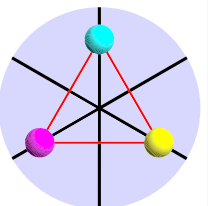
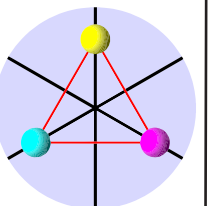
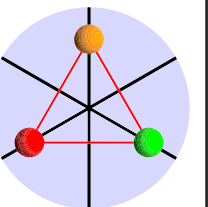
Symmetries of the Dihedral Group of Order 6 (D_3)				
Perspective	Rotation (C_1)	Rotation (C_2)	Rotation (C_3)	Reflection
				

Table 3.1: Symmetries of the Stacked Trimer (D_3)

desired C_3 subgroup. Bonds of type A and B will henceforth be referred to as *Intra-Ring* bonds, Fig. 3.1. Additionally, I assign site C to be 90° from both site A and site B. When the z-plane reflection is considered the reflection relation of C to C' becomes clear (Table 3.1). Bonds of type C and thus C' will henceforth be referred to as *Inter-Ring* bonds, Fig. 3.1.

Enumeration of substructures of the stacked trimer must be achieved before the reactions can be enumerated. The algorithmic details of how to accomplish substructure enumeration are flexible and depend on the specific structure under investigation. For conceptual clarity, I will describe the enumeration substructures in a fashion similar to how a biological system would explore the possibilities. To begin, consider a single protein: it can be represented as a substructure of the stacked trimer by considering the complementary five proteins as missing/unbound. This will be illustrated by filled and unfilled circles, as in Section 2.2. For a single protein there are six locations the protein could be placed within the full structure. D_3 symmetries provide an equivalence relation rendering all six of these equivalent, thus I have my first equivalence class. Now, for convenience, I will choose a single representative from the class to represent the class as a whole; each of the class representatives are shown in Fig. 3.2. The single protein is the minimal substructure in terms of protein count and will provide the starting point for all assembly reactions. Additionally, the minimal subunits are the only species with synthesis. Not all protein complexes have a unique minimal substructure; in Chapter 5 the system will have up to three unique minimal substructures.

Once the minimal substructure(s) have been determined, one can consider the possible structures formed by binding two minimal structures together. Yielding my first *assembly intermediate(s)*, once again they are tested for D_3 equivalences and reduced to equivalence classes. At each step all possible combinations between substructures, both by type of substructure and by symmetries of those substructures, must be tested to ensure an exhaustive listing of all possible substructures. This process is effectively what is happening in solution when the proteins are colliding with each other. This is clearly not the algorithmically optimal method of determining the substructures, as I know *a priori* the existing symmetries of the final structure, but a biological system does not have that advantage. By continuing to generate new substructures through exhaustive pairwise binding attempts between all structures known up to that point, the process will converge to only binding events resulting in the maximal structure (all possible protein positions are filled). It is important to note that not all intermediates can bind with each other, as a filled protein slot cannot overlap another filled slot in the binding partner. For the case

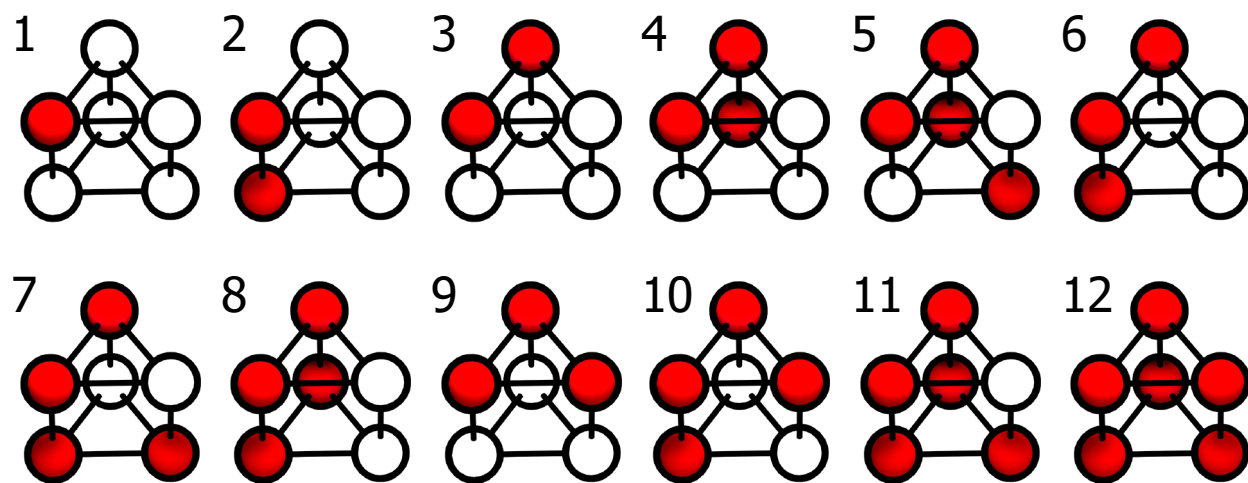


Figure 3.2: Stacked Trimer Subspecies Chart

Canonical representations of each equivalence class for the stacked trimer. Red filled circles mean a protein is present in that location, whereas empty circles represent potential locations for a protein which are not occupied in that structure. They proceed from species one, the minimal structure up through species 12 the complete stacked trimer. Additional representations are discussed in Section 4.3 and listed in Table 4.4.

of the stacked trimer this process results in 12 equivalence classes of protein complexes, including the minimal subunit and the maximal structure, all 12 are shown in Fig. 3.2.

Exhaustive enumeration also happens to determine all possible reactions in the system, one just needs to keep a tally of how many times each type of binding match is detected. As the species are described using equivalence classes, the reactions can also be grouped into binding relations between equivalence classes, but some reactions will be detected more than once between the same classes. To account for this, a combinatorial factor must be added equal to the number of times the same reaction is found. All binding relations including combinatorial factors, number of intra-ring bonds (i), and number of inter-ring bonds (j) for each relation are shown in Fig. 3.4. One goal of this approach is to look at the effects of the assembly network's layout, not just determine the result.

As such, I will introduce an effective K_d for binding relations that takes into account the differences between the species involved, but is still related to previously defined parameters. The first piece of this is to approximate all association events (k_{on}) to be equal. The second, will be choosing the K_d terms from the binding of two minimal subunits as independent variables. Now, since there are two bonds which can join the minimal subunits together, I need two K_d values: one related to the intra-ring bond ($K_{d,1}$) and the second for the inter-ring bond ($K_{d,2}$). These are defined in the same way as in previous chapters within the context of the first two reactions in Fig. 3.4. My final parameter for this effective binding strength is an adjustment accounting for the differences between complexes when more than one bond is involved in the binding relation. The term ΔG_p^0 is a description of the change in energy due to positional entropy from the effects of multiple bonds changing state in the binding relation. ΔG_p^0 has been previously described, used, and ascribed an empirical value of 9 kcal M⁻¹ [18, 63, 64]. Altogether, this yields an

effective dissociation constant of:

$$\begin{aligned}
 k_{i,j}^{\text{eff}} &= \alpha k_{\text{on}} K_{d,1}^i K_{d,2}^j e^{-(i+j-1)\Delta G_p^0/RT} \\
 \alpha &= c_0^{-i-j+1} \\
 k_{\text{on}} &= 10^6 \text{ M}^{-1} \text{ s}^{-1}
 \end{aligned}
 \tag{3.1}$$

c_0 is a reference concentration added to correct the units, the standard value is 1 M. As discussed in Chapter 1, k_{on} 's value is set to a reasonable value for most proteins. Note, if the substructure only contains one bond, $k_{i,j}^{\text{eff}}$ reduces to k_{off} for the relevant bond. Finally, $T = 300\text{K}$ is a common approximation for living systems. This results in $RT \approx 0.6 \text{ kcal M}^{-1}$. Full differential equations for the stacked trimer are in Appendix A. Their derivation followed the same general process as Sections 1.1, 1.2 and 2.2.

Synthesis of minimal subunits and degradation of all species has been included in the full equations using the same approach as Eqs. 1.11. Determination of an analytic closed form solution for stacked trimers is at best not very enlightening and likely futile. Therefore, I developed a platform using computational numerical analysis to determine solutions for results related to the stacked trimer [65, 66]. The platform performed all simulations and analysis using a Python based system for distributed processing, analysis, high performance data storage, GUI based data exploration with both local display and WebSocket based browser access [50, 54–57, 67]. The general approach is visually represented in Fig. 3.3. Source code will reside for a reasonable period of time in the GitHub repository: <https://github.com/vatir/Protein-Complex-Assembly-Pathways>.

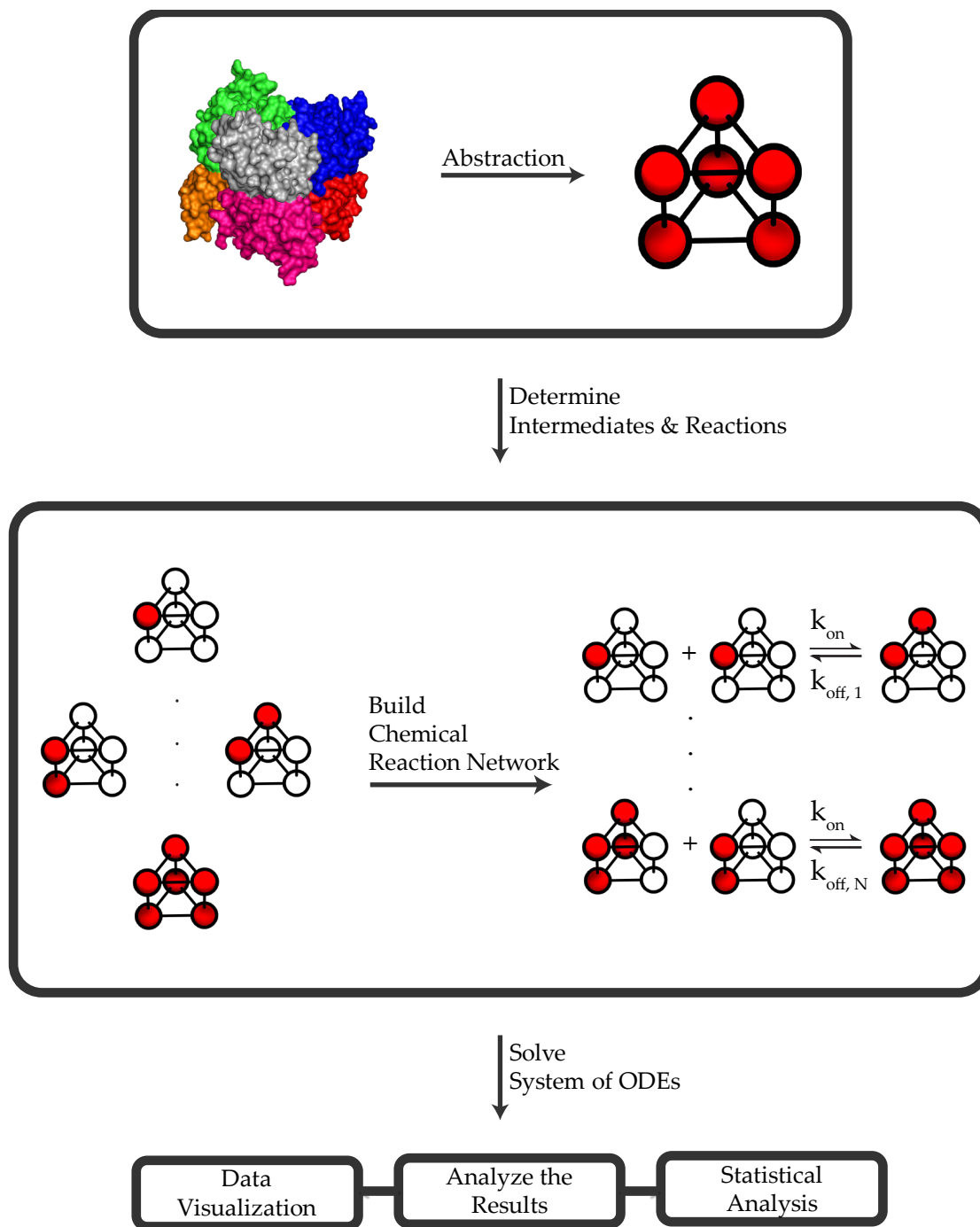


Figure 3.3: Assembly Network Workflow

General process work-flow for approaching protein complex assembly systems with my level of approximation. The relationships between the objects are important, but the experimental method of determining them is not. In this case a crystal structure is shown as it provides a visual representation of the conversion. Structure shown is PDB entry 1EHW rendered in PyMOL [68, 69].

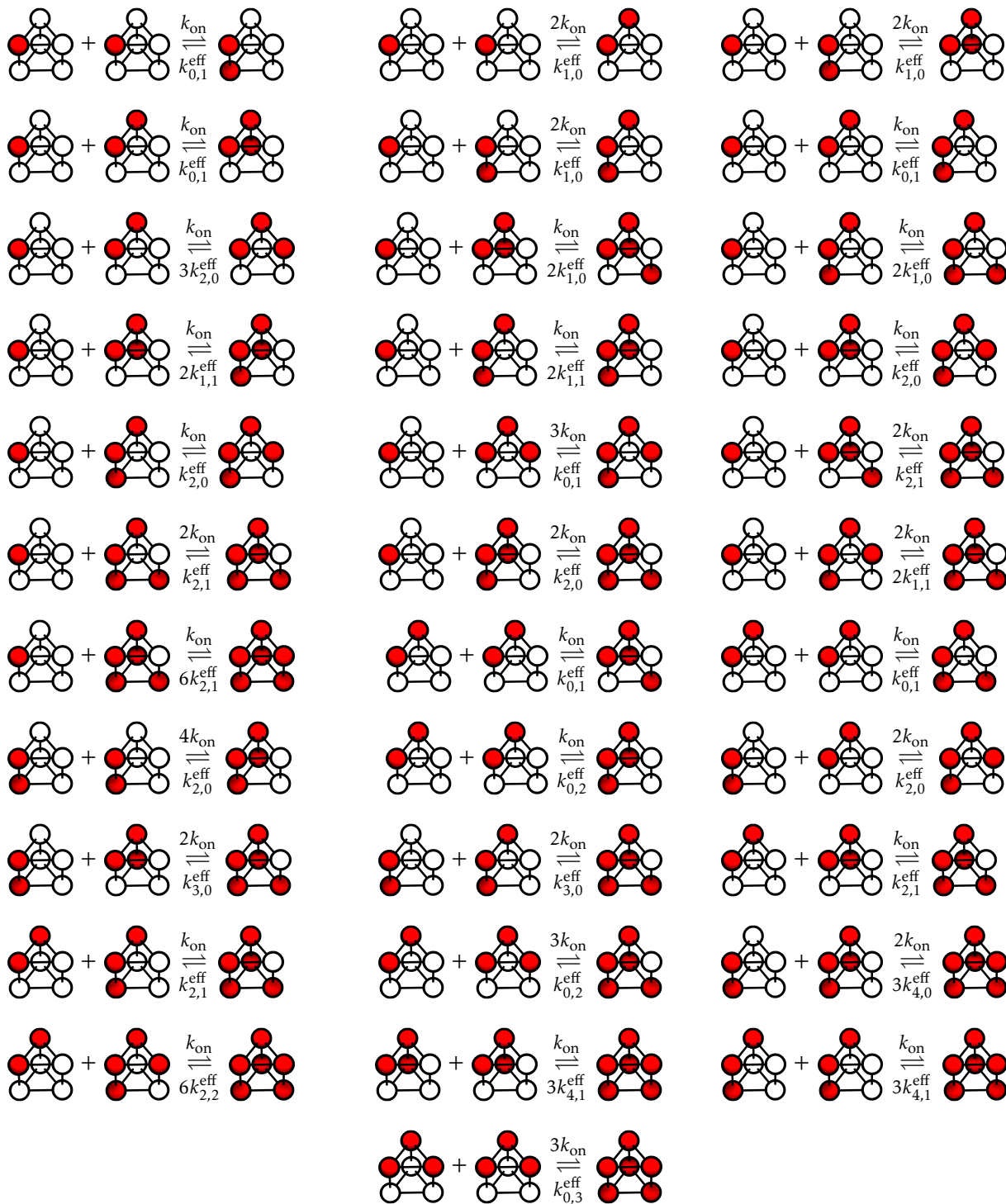


Figure 3.4: Stacked Trimer Reactions

All 34 binding relations for the stacked trimer, including combinatorial coefficients for repeated forward and reverse reactions. See Eq. 3.1 for the definition of $k_{i,j}^{\text{eff}}$, i is the number of intra-ring bonds and j is the number of inter-ring bonds.

3.2 Pathways

Elementary operations and the general formalism did not come up much in Section 3.1. Well, I do not technically need them to run simulations in time and parameter space, but the benefits to analysis are invaluable. First, I will give some perspective on the scale of data generated; a single dataset produces values for the stacked trimer's system in a dense grid of log spaced points across $100 K_{d,1}$ values \times $100 K_{d,2}$ values \times 100 Initial Concentrations \times 112 Time Points \times 12 Species, plus assorted metadata about the run to improve data governance. This yields an approximately 12 gigabyte binary storage object in the HDF5 format. Compounding this is one dataset was produced without synthesis and degradation, but there are over 20 additional datasets with an effective cell division period, the independent parameter controlling the synthesis/degradation rates. Hum, so I am now sitting on about a quarter terabyte of binary64 floating point values. How can I begin to extract patterns and meanings from this?

Boundary conditions often depict a system's potential. I know the maximal and minimal structures which the system admits, so I will start with those boundaries in place. Additionally, I constructed the system's bond patterns to preclude structures that are not substructures of the maximal structure. Those constraints close the species space. Now, I will consider the system's structural properties. The system never utilizes elementary operations other than bind self (Section 1.1) and bind other (Section 1.2). Next, branching (Section 2.1) has already been accounted for by the grouping of reactions into reaction classes and the additional of combinatorial factors. Thus, the only operation type which can connect binding operations are SBEs (Section 2.1). With those points in mind, I will consider the potential ways the stacked trimer can be built up through association. Note, the system is fully described by the network formed from the reactions in Fig. 3.4, which include the potential to cycle through both association and disassociation. My goal is to enumerate possible paths by which stacked trimers can assemble, so I will consider only association. Binary trees provide an excellent abstraction of the assembly process; they

allow a complete representation of the association steps required to form the full protein complex. Supplementary Reading: Trees (Data Structures) Within this chapter I will limit my discussion to pathways relating to the stacked trimer. For details of the general process of enumerating pathways please see Chapter 4. The reader may find the presentation clearer to assume I have exhaustively enumerated all unique assembly pathways for the stacked trimer by hand, as for a system this size it is quite possible. All 46 assembly pathways are graphically shown in Table 3.3.

At this point, I will split the development of stacked trimer assembly and assembly pathways into the remainder of the current chapter and Chapter 4, respectively. Stacked trimer assembly and assembly pathways are intertwined concepts, so while there are points of overlap between their development. Presentation of the formal underpinnings of assembly pathways, at this point, may distract the reader from the analysis of stacked trimer assembly. Several analysis methods for the stacked trimer will use both assembly pathways and Pathway Contribution, which will formally be introduced in Chapter 4 and Section 4.4. This will result in the use of equations for Pathway Contribution appearing in the analysis of the stacked trimer before they are presented. Please bear with me, but I prefer to present a coherent description of stacked trimer assembly before embroiling the reader in the background and terminology required for the formal definition of assembly pathways and Pathway Contribution. If the reader feels a strict adherence to the order of introduction with regard to technical concepts is required, please jump ahead and read Chapter 4 at this point, then return and enjoy the analysis of stacked trimer assembly.

3.3 Analysis

After writing an interactive data visualization system for the stacked trimer's Chemical Reaction Network (CRN) results, I spent quite a long time looking at all the correlations and interesting patterns the system shows. Eventually I concluded the system is complicated enough that I could find almost any pattern I wanted to if I looked long enough. The one

Table 3.2: Default Stacked Trimer Properties

Time	Concentration	Cell Division Period		Degradation Rate (δ)	
		<i>in vitro</i> -like	<i>in vivo</i> -like	<i>in vitro</i> -like	<i>in vivo</i> -like
\approx one day	4×10^{-6} (M)	∞	hour	0	2.8×10^{-4} (s^{-1})

unrelenting feature is a tendency to see higher yield in the maximal structure when the Inter-Ring $K_d <$ Intra-Ring K_d , *i.e.*, the system has a higher tendency to form Inter-Ring bonds when species capable of doing so are present. It is difficult to attribute this pattern to a specific source. The symmetry of the structure is a strong possibility, but leaves on open question: how does the symmetry result in that tendency? This is an excellent example of the importance of having a background/comparison system(s). Determining causality or meaning of the patterns found in the system becomes difficult, while it becomes worryingly easy to present subjective results, probably with very nice statistics to provide credibility due to the quantity of available data [70–72]. My direction of research after this work was preliminarily completed took a quite radical turn, so I did not follow-up this project by with development of possible reference systems. The intellectual descendant of this project is Chapter 5, which was chosen to focus on predictions for experimental projects. Associated with this transition, my focus shifted to developing and troubleshooting a series of experimental projects. With that, onto some comments on the upcoming figures.

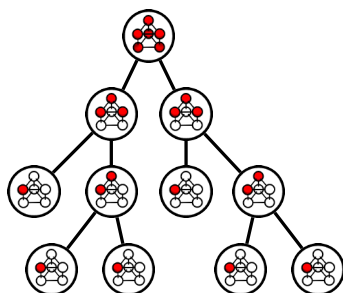
As I have mentioned, the range of the parameter space provides an expansive view of the dynamics of assembly of the stacked trimer. My goal in Figs. 3.5–3.16 is to provide a sampling of those dynamics, by showing several representative points and how the system varies as each of the parameters are changed. As I am trying to give an overview, the figures are quite dense to maximize the ability to compare several changes at once within the same overall figure. In my opinion printed 3D surface meshes often do not show data particularly well, and using them for only some of these plots would result in a non-uniform presentation. Thus, I have relied heavily on gradient based plots. In the following cases they are a top-down view with increasing height represented by darkness,

with a reference bar on each page. The visibility of some low valued results may be poorly resolved if your monitor or printer contrast is low. If a value is not represented in the figure then it was taken from the default values listed in Table 3.2.

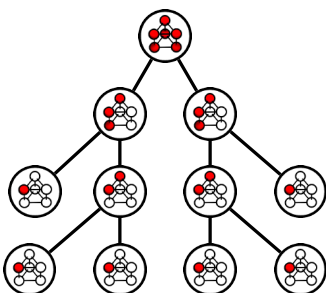
Figs. 3.17–3.18 represent the only case where non-simulated data is present for the stacked trimer. 264 stacked trimer-like structures were determined from the Protein Data Bank (PDB) [73], by writing code to parse structural data from PDBePISA [74], which is available at <https://github.com/vatir/Protein-Complex-Assembly-Pathways>. No parameter fitting related to the simulation model was used to create the BSASA plots. The relationship between Buried Solvent-Accessible Surface Area (BSASA) and K_d was used directly from Chen *et al.* [75]. Default parameters were chosen for unrelated reasons. Yet, the alignment between PDB derived experimentally determined proteins and the simulated system is quite good. Proteins were colored by domain, partly because eukaryotic proteins may tend more towards longer term systems more like the *in vitro*-like cases, whereas the bacterial proteins are more likely to be better approximated by my *in vivo*-like simulations, as the *in vivo*-like system was derived from a bacterial based model Section 1.3. It should be noted, there are many reasons why results from the PDB can be biased, such as crystal contacts, protein crystallization tendencies and experimental interest in specific systems over a uniform distribution for real proteins, just to name a few. A tendency for eukaryotic proteins to cluster near the equal strength bonding region is clear, but they still tend towards the bond strength range which aligns with the model's results. Whereas, the archaeal and bacterial proteins cover a wider range of bond strengths, there is a noticeable lack of examples with high strength intra-ring bonds while high strength inter-ring bonds are present. Therefore, these experimentally derived results are consistent with the Inter-Ring $K_d < \text{Intra-Ring } K_d$ trend.

Table 3.3: Assembly Pathways for Stacked Trimers

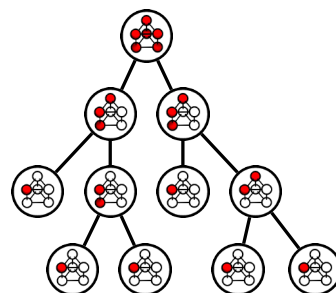
1



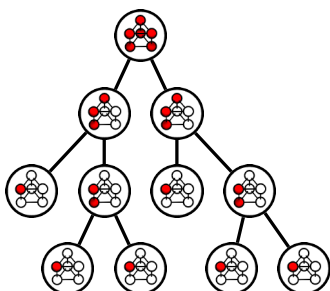
2



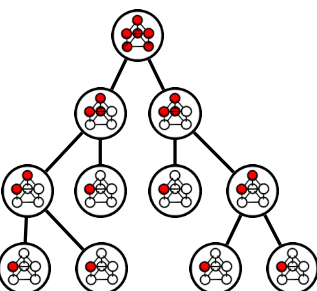
3



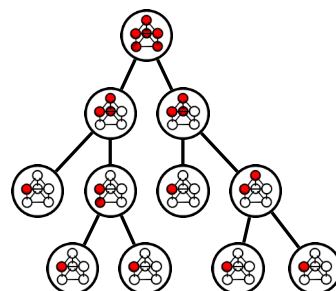
4



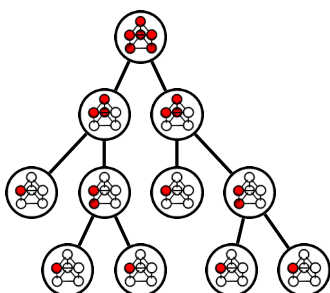
5



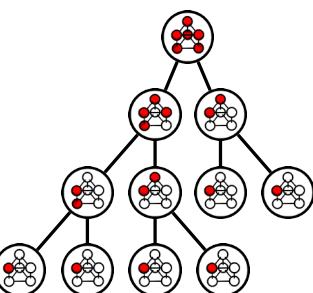
6



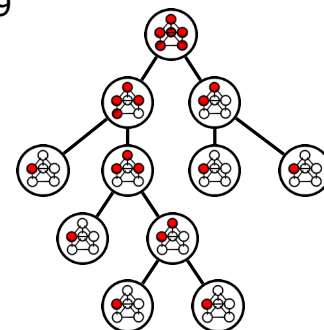
7



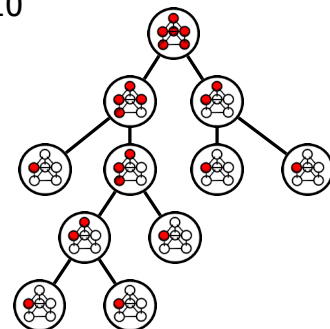
8



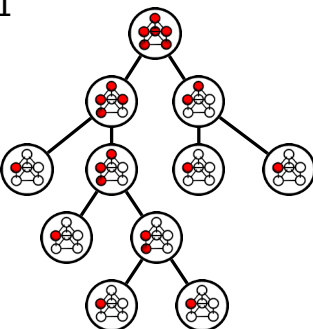
9



10



11



12

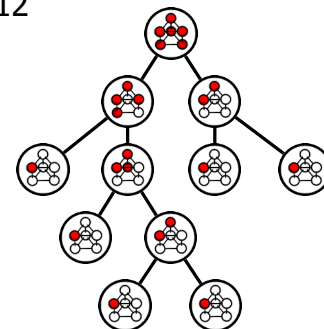
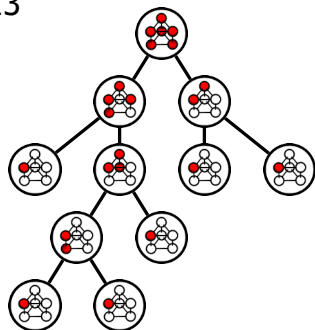
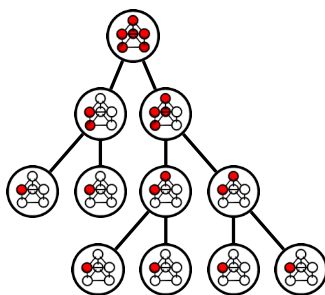


Table 3.3: Assembly Pathways for Stacked Trimers (Continued)

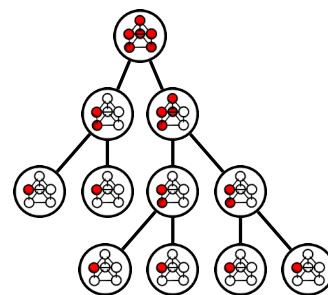
13



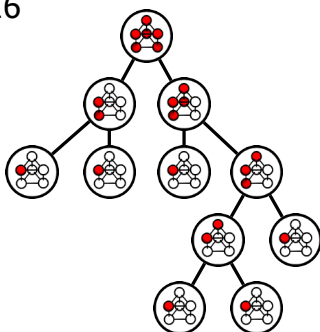
14



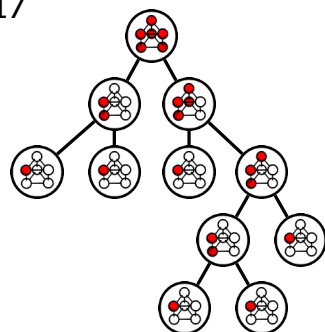
15



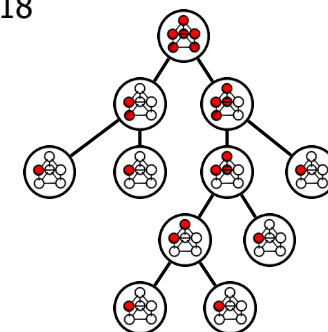
16



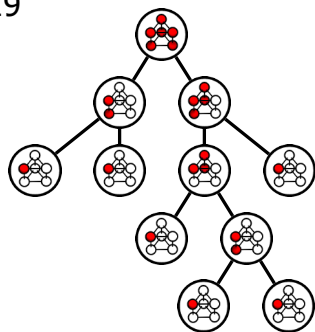
17



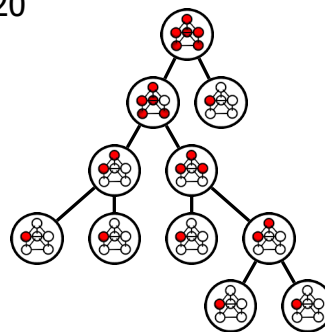
18



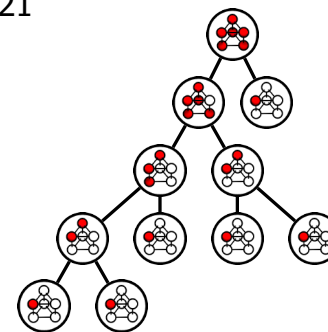
19



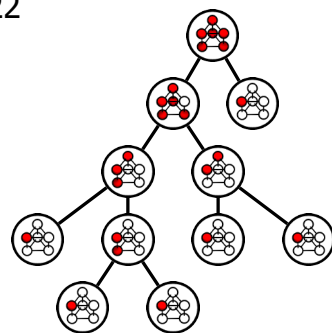
20



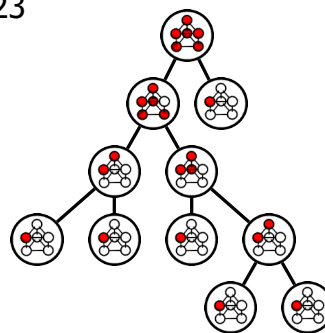
21



22



23



24

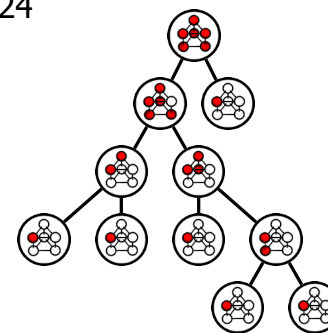
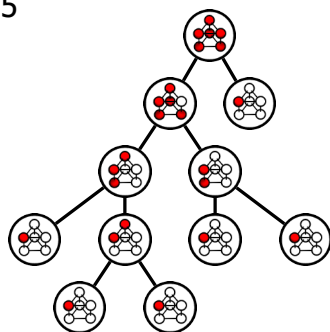
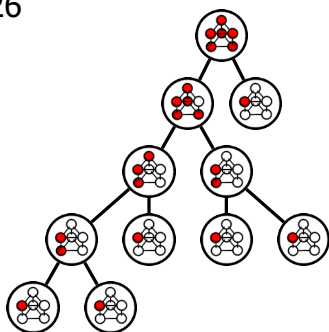


Table 3.3: Assembly Pathways for Stacked Trimers (Continued)

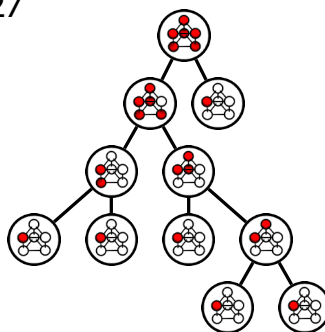
25



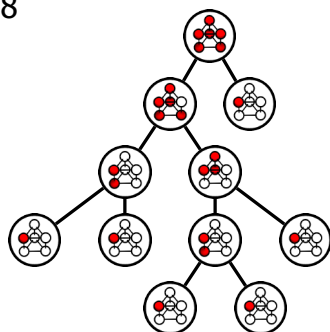
26



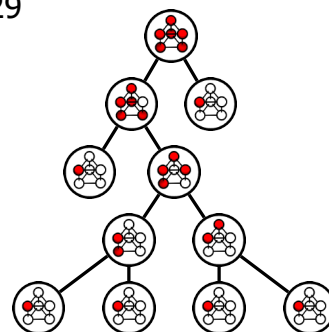
27



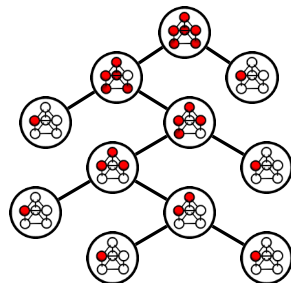
28



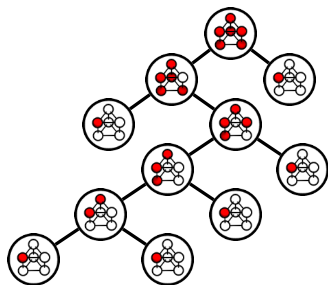
29



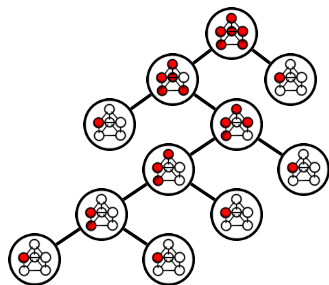
30



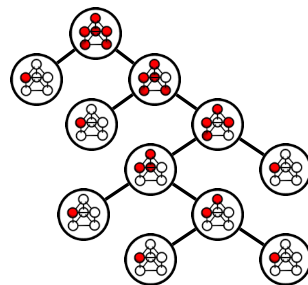
31



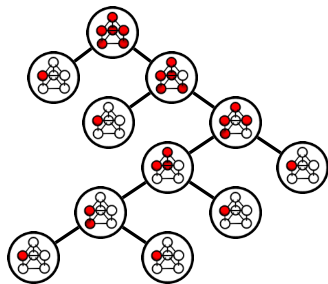
32



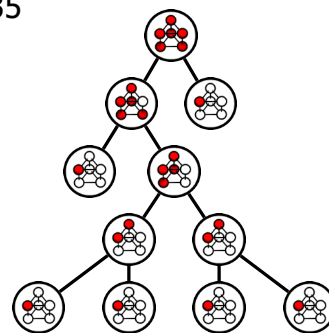
33



34



35



36

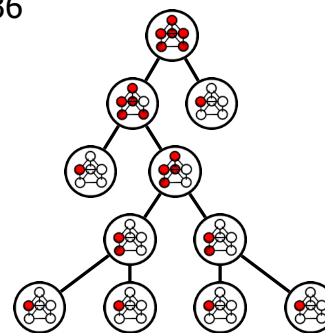
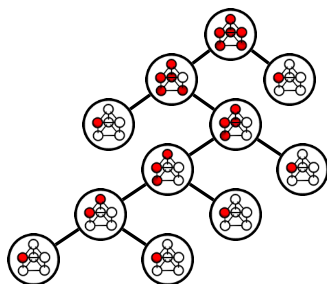
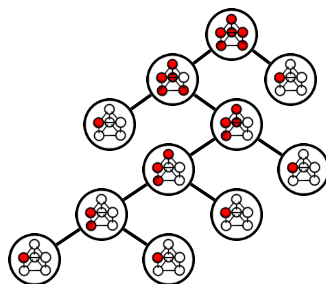


Table 3.3: Assembly Pathways for Stacked Trimers (Continued)

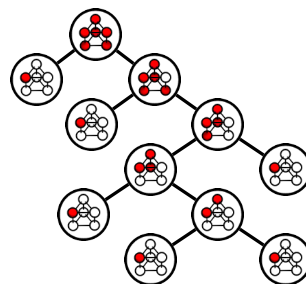
37



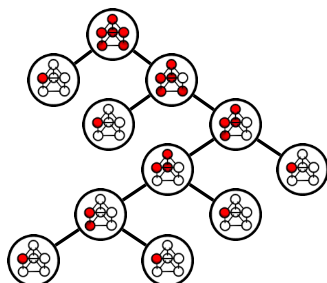
38



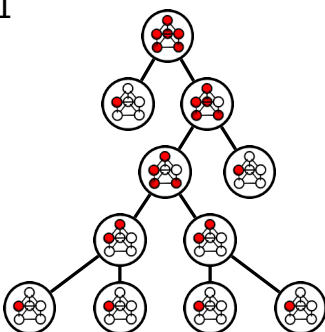
39



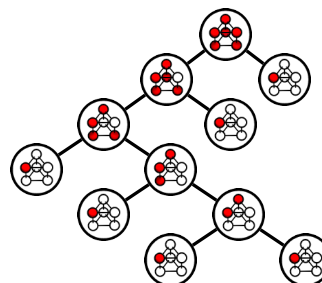
40



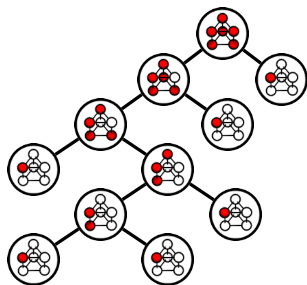
41



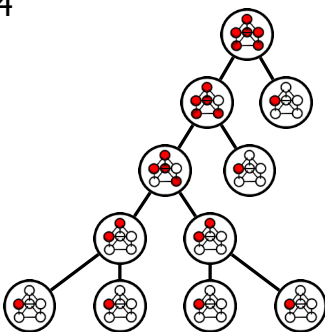
42



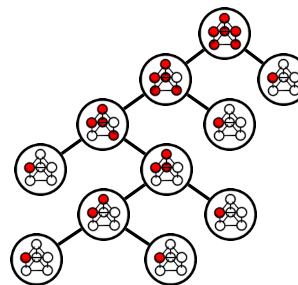
43



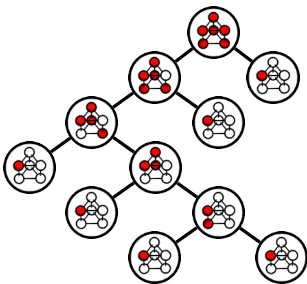
44



45



46



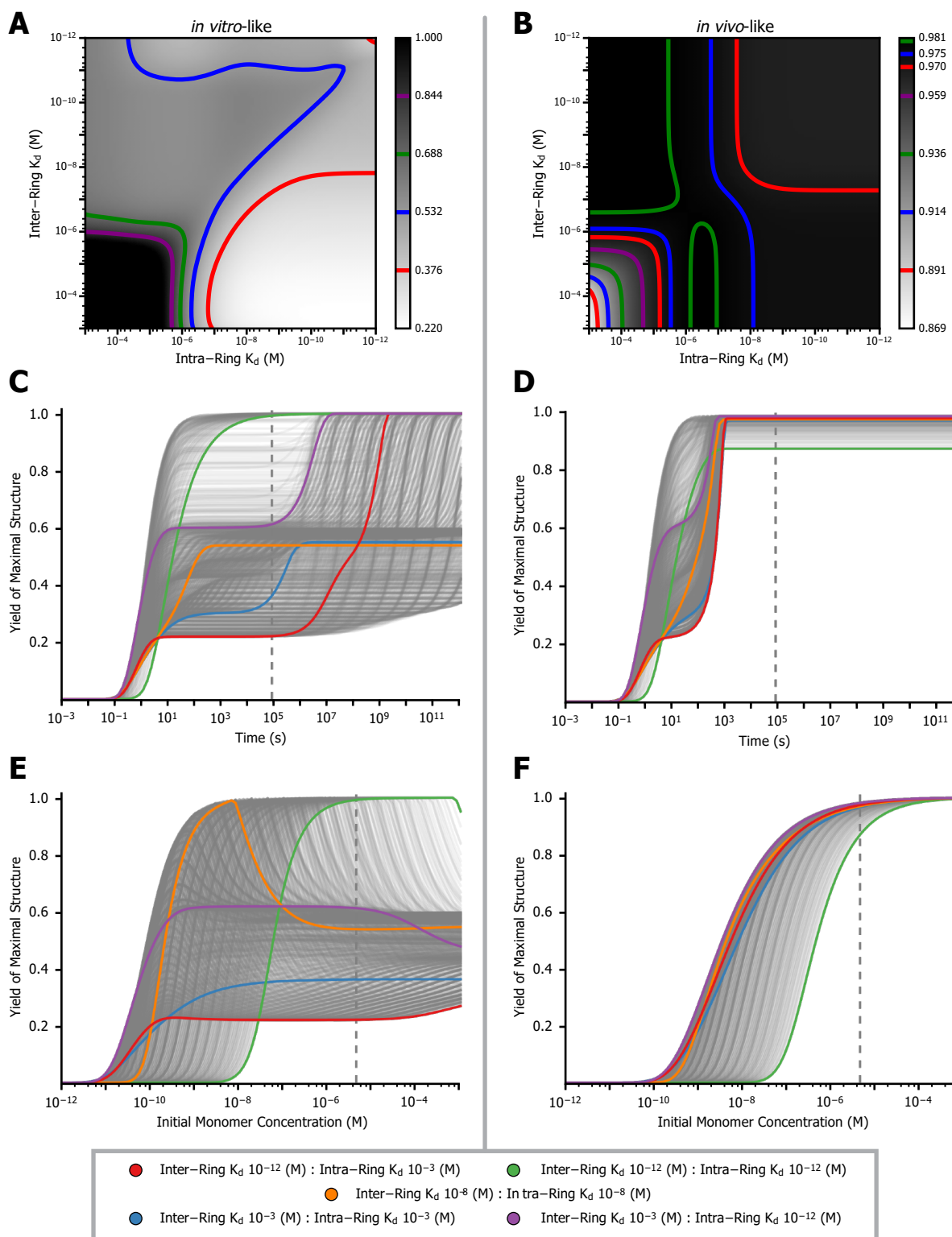


Figure 3.5: Stacked Trimer Assembly Yield Overview

Caption is on following page.

All parameters not specified are the default values enumerated in Table 3.2.

Figure 3.5: Stacked Trimer Assembly Yield Overview

Comparison between *in vitro*-like and *in vivo*-like CRN results in the left and right columns respectively. (A and B) are heatmaps with contour lines covering the intra-ring and inter-ring binding strength parameter space. Scales are normalized to the data in the plot range, so make sure to observe the indicators to the right of the plots. (B, C, E and F) display the behavior of the system along two additional parameters, time and concentration. The vertical dashed lines show the default values, which align with the values used for all other plots in the figure. Whereas colored lines are the binding strengths as defined in the legend at the base of the figure. Gray lines are 125 binding strengths uniformly distributed among the 1000 used to generate the heatmaps. The objective of this plot is to provide an overview of the range of behaviors observed in the simulated parameter regime.

Figure 3.6: Stacked Trimer Pathway Contribution (*in vitro*-like)

Subplot (C) is a modified version of Fig. 3.5A with colored markers for easy reference to the bond strengths in the other plots. Subplots (A, B, D, E and F) correspond to five different binding strengths for the bonds within the stacked trimer. Each shows the pathway contribution, calculated as described in Eq. 4.1, at the default time and concentration for the indicated binding strengths. Pathways are grouped by the final association reaction as indicated by the graphics in the top of each chart and are ordered with the same indices as in Table 3.3. Blue bars are the individual assembly pathways' contributions, whereas the red bars are the sum of each assembly pathway groups' contributions. Assembly pathways are indexed as displayed in Table 3.3.

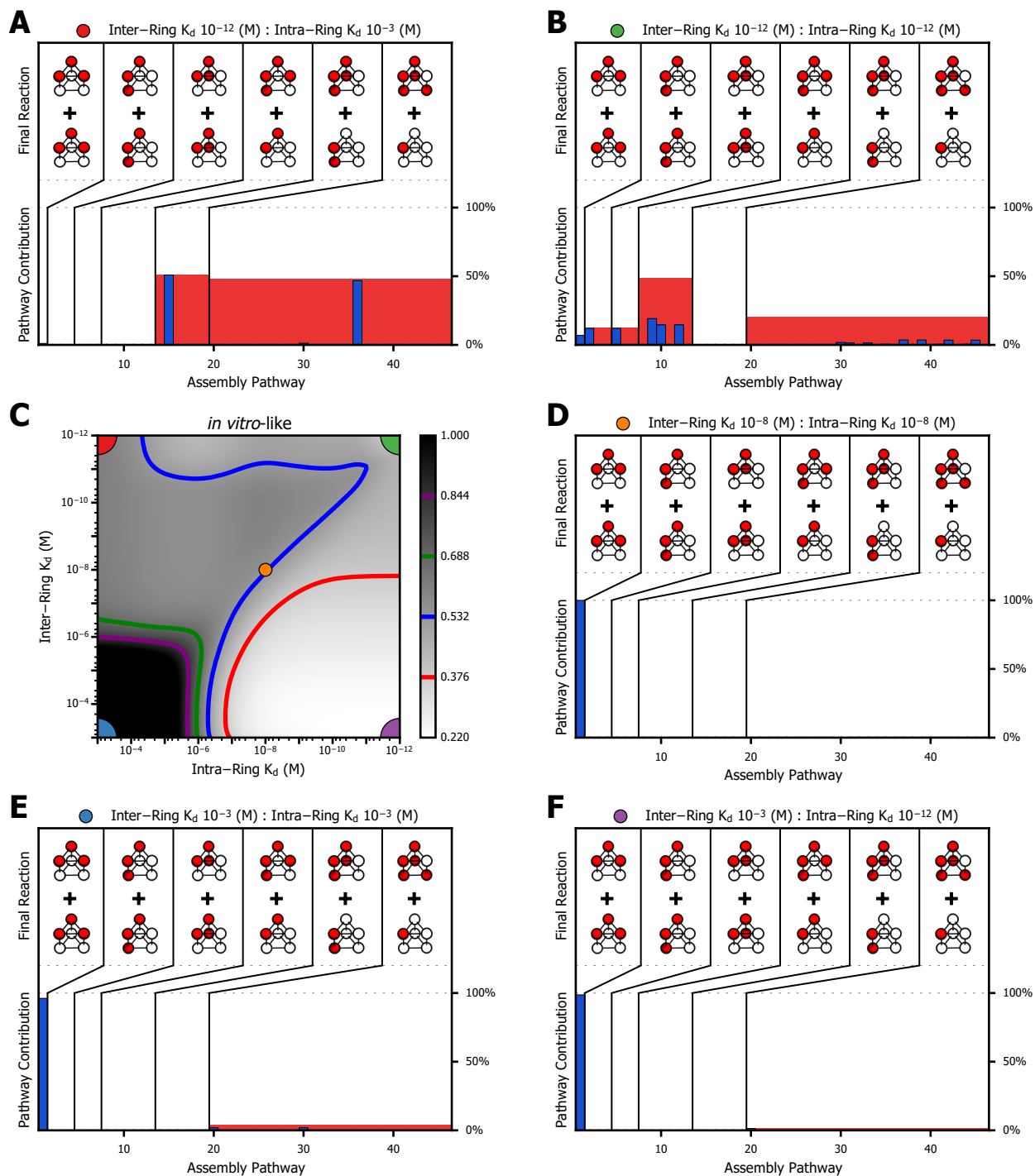


Figure 3.6: Stacked Trimer Pathway Contribution (*in vitro*-like)

Caption is on preceding page.

All parameters not specified are the default values enumerated in Table 3.2.

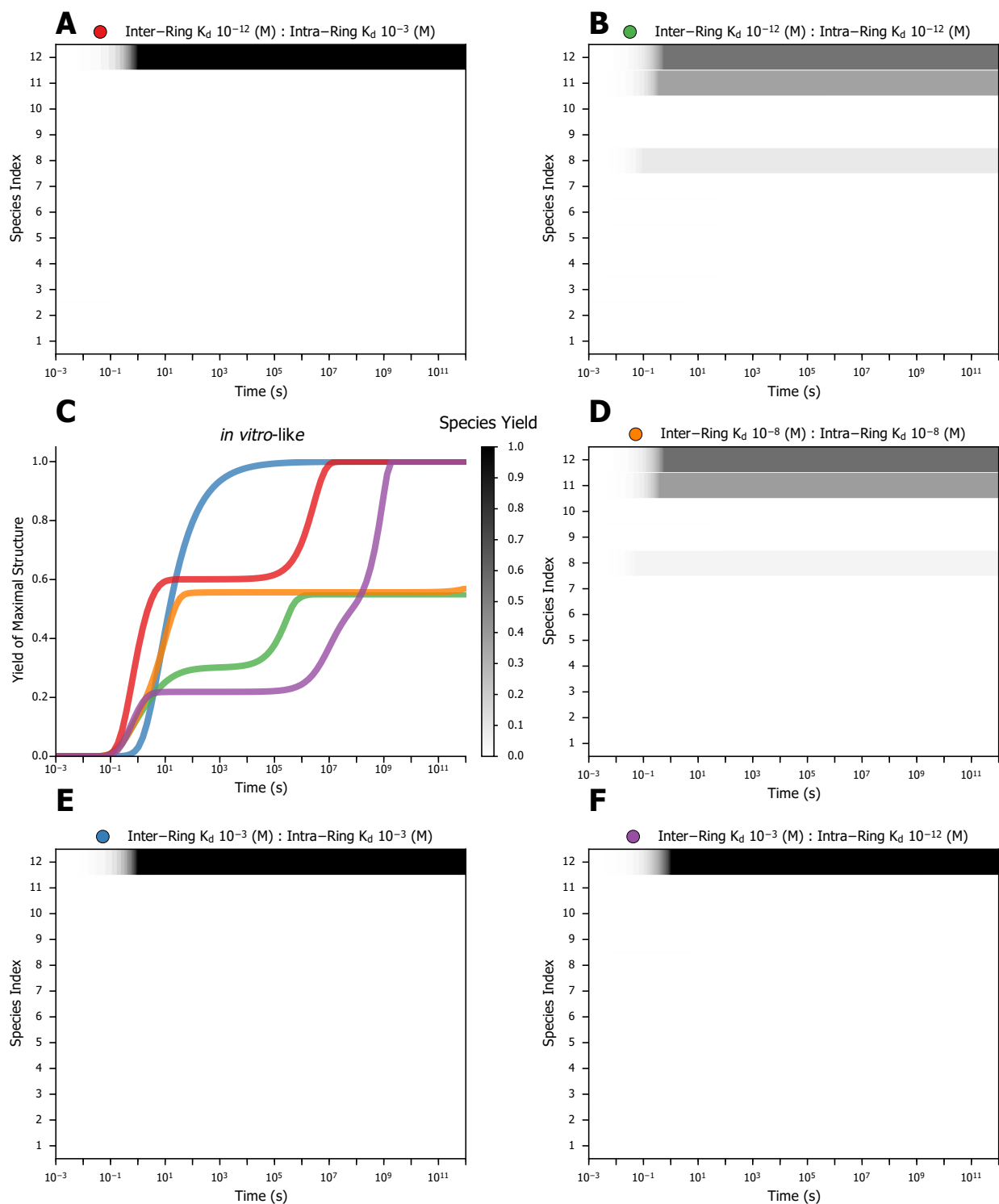


Figure 3.7: Stacked Trimer Species Fractions over Time (*in vitro-like*)

Caption is on following page.

All parameters not specified are the default values enumerated in Table 3.2.

Figure 3.7: Stacked Trimer Species Fractions over Time (*in vitro*-like)

Subplot (C) is a modified version of Fig. 3.5C with only the colored lines, for ease of reference to the other subplots in this figure. Subplots (A, B, D, E and F) are faceted heatmaps, which can be thought of as a top down view of a 3D histogram with the bar height represented by the scale in the center. Each plot uses data from the five different representative binding strengths for the bonds within the stacked trimer. Yield of each of the 12 species is shown by the horizontal gradients, as indexed in Fig. 3.2. The central “Species Yield” gradient bar indicates the intensity for all sub-figures. Yield is defined such that it is normalized vertically due to the constant maximum available protein. This may be obscured at short timescales as protein is spread out into multiple intermediates at a level which is poorly displayed by this plot type, but the plot type is clearer than a 3D histogram once biologically reasonable timescales have been reached.

Figure 3.8: Stacked Trimer Assembly Contributions over Time (*in vitro*-like)

Subplot (C) is a modified version of Fig. 3.5C with only the colored lines, for ease of reference to the other subplots in this figure. Subplots (A, B, D, E and F) are faceted heatmaps, which can be thought of as a top down view of a 3D histogram with the bar height represented by the scale in the center. Each plot uses data from the five different representative binding strengths for the bonds within the stacked trimer. Pathway contribution, calculated as described in Eq. 4.1, is shown as the simulation progresses in time. This shows how dominant pathways can switch as the system progresses. As a reminder, pathway contribution is explicitly normalized. Assembly pathways are indexed as displayed in Table 3.3.

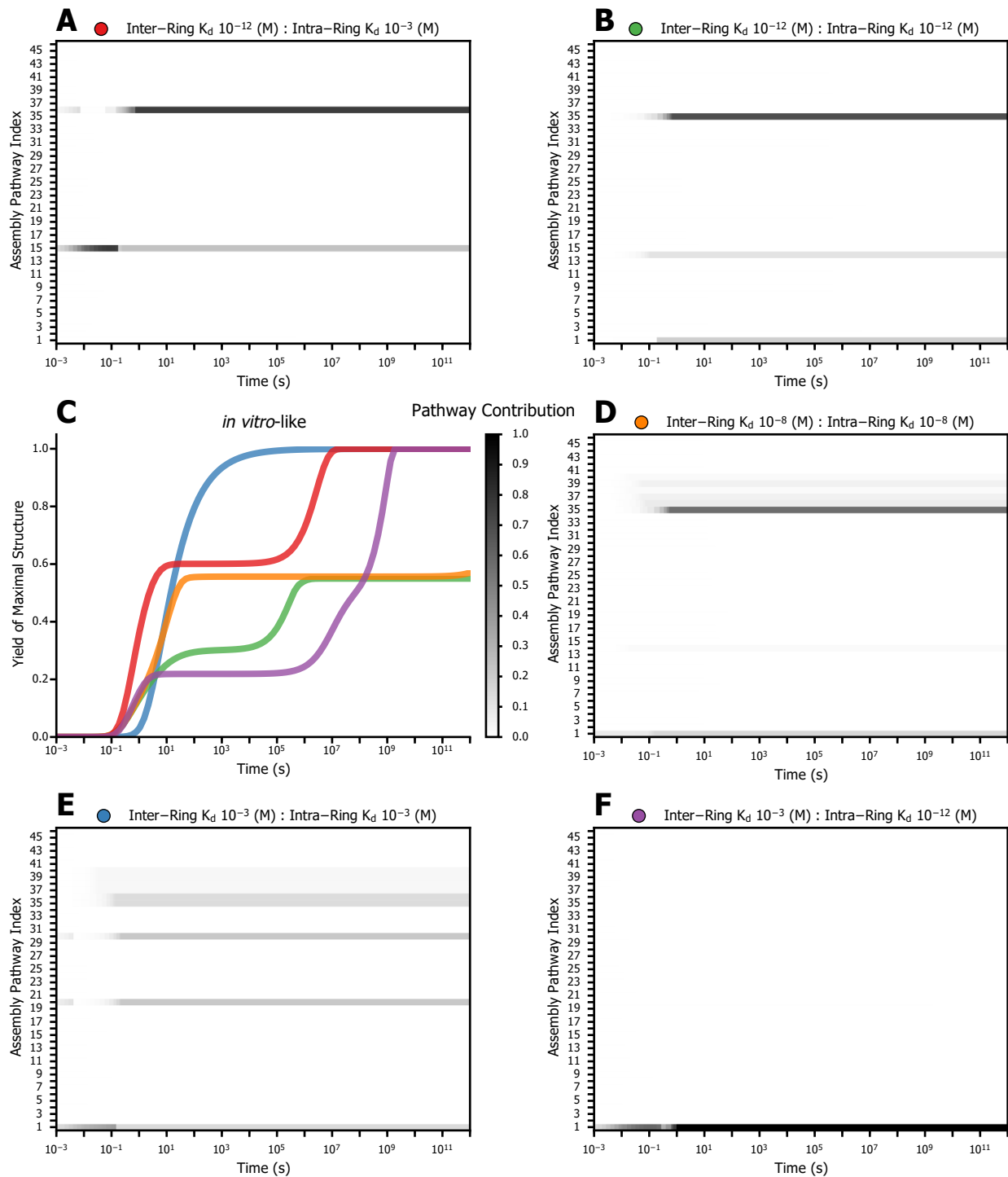


Figure 3.8: Stacked Trimer Assembly Contributions over Time (*in vitro-like*)
 Caption is on preceding page.
 All parameters not specified are the default values enumerated in Table 3.2.

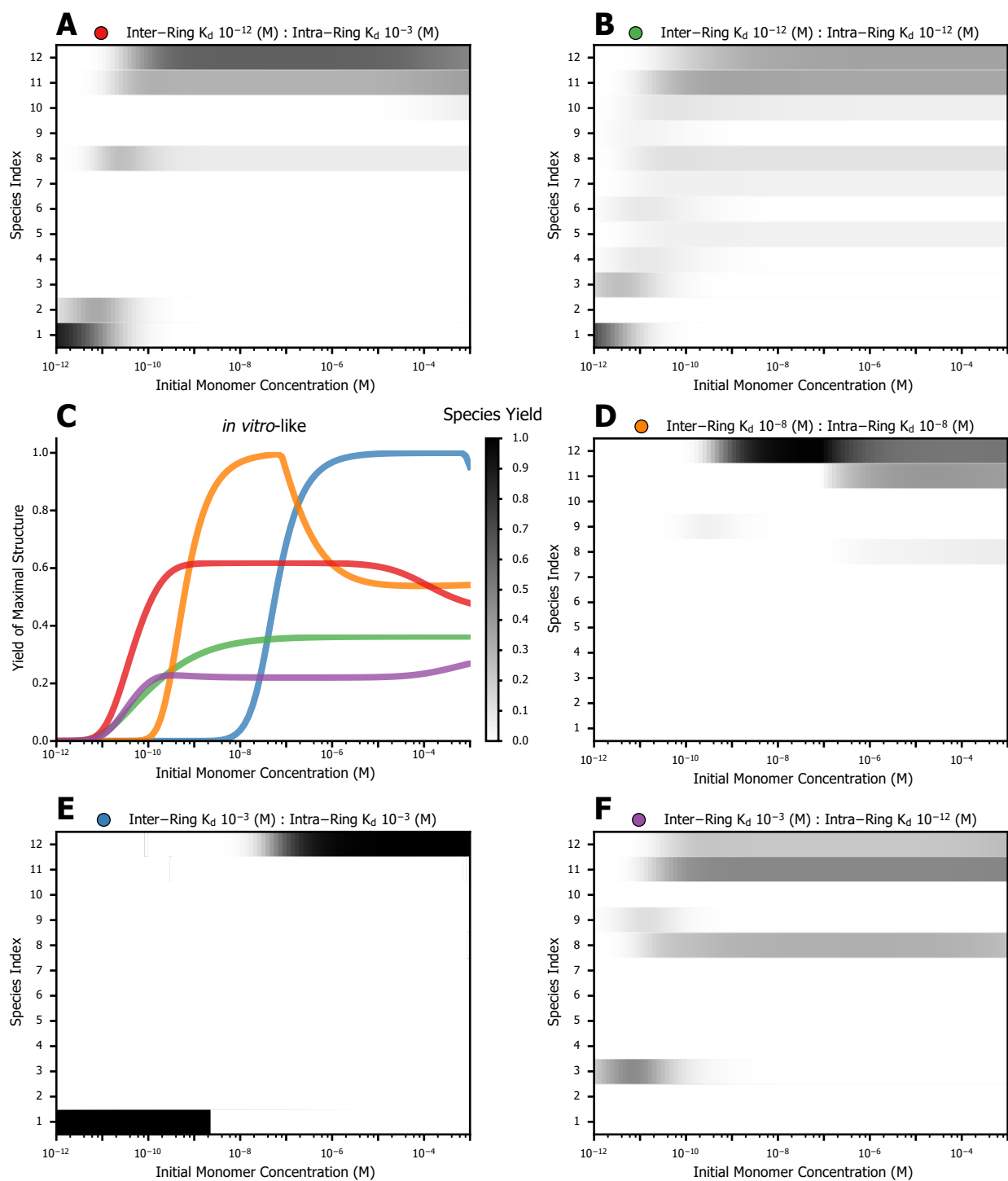


Figure 3.9: Stacked Trimer Species Fractions over Concentration (*in vitro*-like)
 Caption is on following page.
 All parameters not specified are the default values enumerated in Table 3.2.

Figure 3.9: Stacked Trimer Species Fractions over Concentration (*in vitro*-like)

Subplot (C) is a modified version of Fig. 3.5E with only the colored lines, for ease of reference to the other subplots in this figure. Subplots (A, B, D, E and F) are faceted heatmaps, which can be thought of as a top down view of a 3D histogram with the bar height represented by the scale in the center. Each plot uses data from the five different representative binding strengths for the bonds within the stacked trimer. Yield of each of the 12 species is shown by the horizontal gradients, as indexed in Fig. 3.2. The central “Species Yield” gradient bar indicates the intensity for all sub-figures. Yield is defined such that it is normalized vertically due to the constant maximum available protein.

Figure 3.10: Stacked Trimer Assembly Contributions over Concentration (*in vitro*-like)

Subplot (C) is a modified version of Fig. 3.5E with only the colored lines, for ease of reference to the other subplots in this figure. Subplots (A, B, D, E and F) are faceted heatmaps, which can be thought of as a top down view of a 3D histogram with the bar height represented by the scale in the center. Each plot uses data from the five different representative binding strengths for the bonds within the stacked trimer. Pathway contribution, calculated as described in Eq. 4.1, is shown over a range of concentrations to illustrate how initial monomer concentration can affect which pathway is utilized. As a reminder, pathway contribution is explicitly normalized. Assembly pathways are indexed as displayed in Table 3.3.

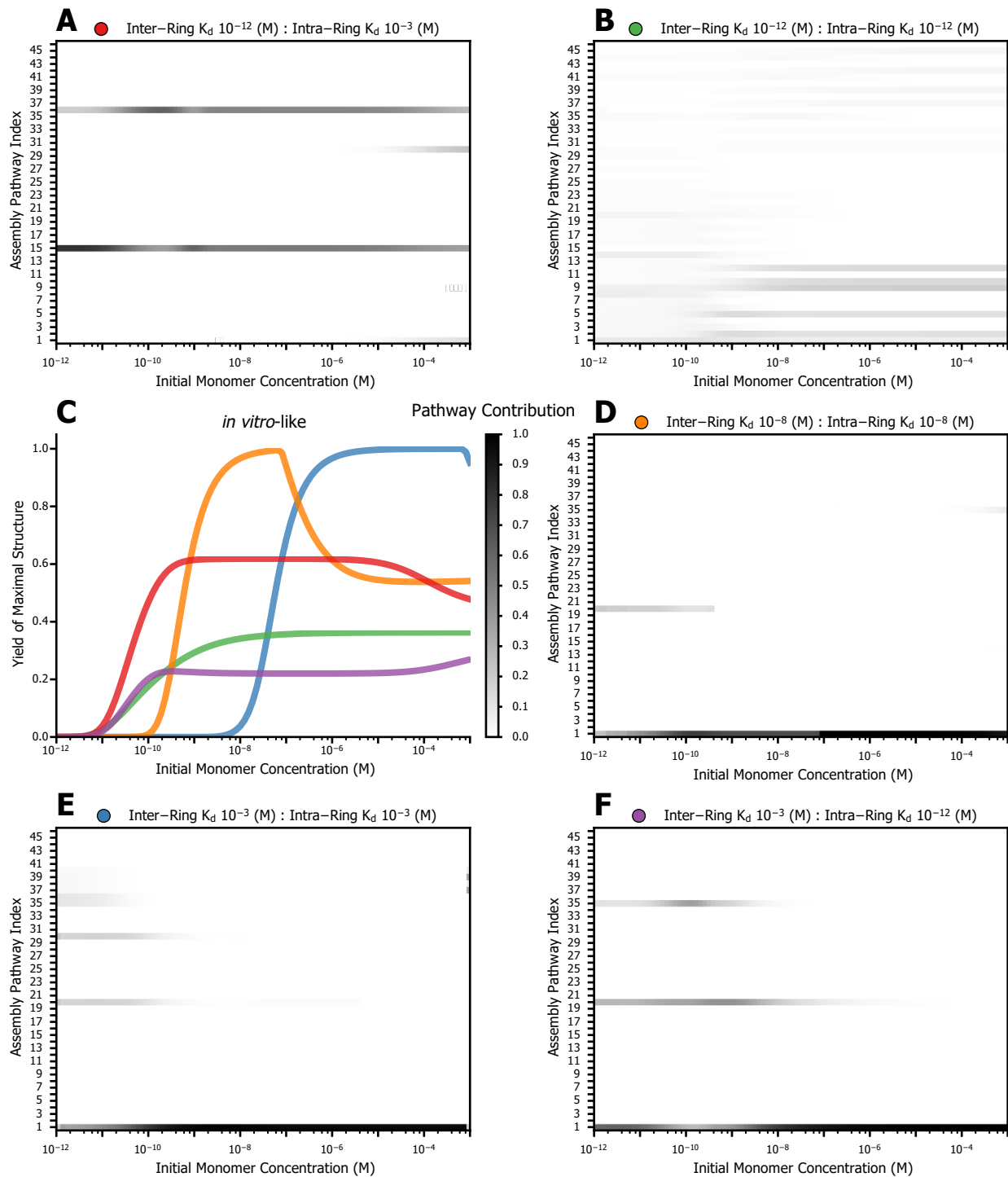


Figure 3.10: Stacked Trimer Assembly Contributions over Concentration (*in vitro-like*)
 Caption is on preceding page.
 All parameters not specified are the default values enumerated in Table 3.2.

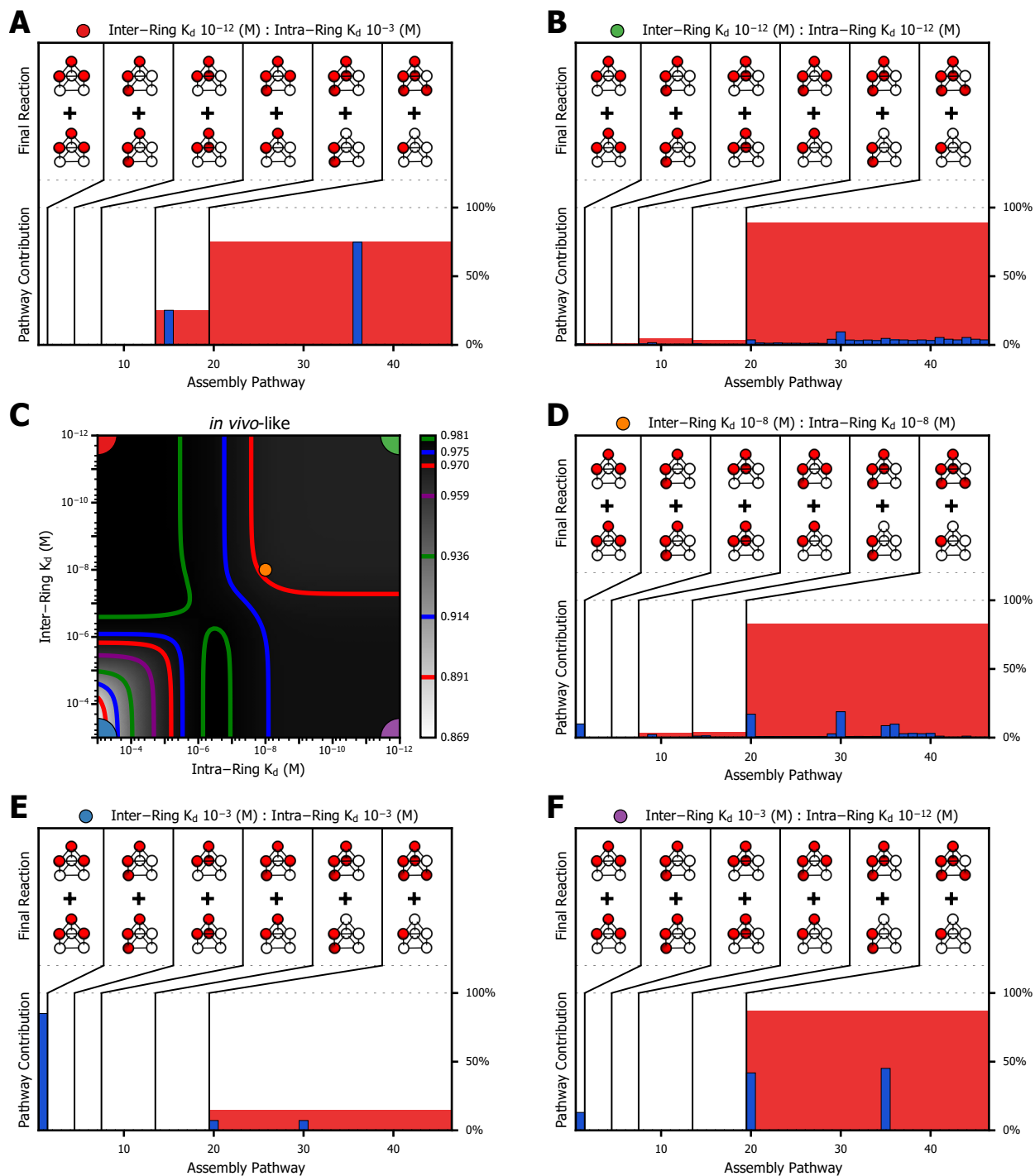


Figure 3.11: Stacked Trimer Pathway Contribution (*in vivo*-like)

Caption is on following page.

All parameters not specified are the default values enumerated in Table 3.2.

Figure 3.11: Stacked Trimer Pathway Contribution (*in vivo*-like)

Subplot (C) is the same plot as Fig. 3.5B with colored markers for easy reference to the bond strengths in the other plots. (A, B, D, E and F) correspond to five different binding strengths for the bonds within the stacked trimer. Each shows the pathway contribution, calculated as described in Eq. 4.1, at the default time and concentration for the indicated binding strengths. Pathways are grouped by the final association reaction as indicated by the graphics in the top of each chart and are ordered with the same indices as in Table 3.3. Blue bars are the individual assembly pathways' contributions, whereas the red bars are the sum of each assembly pathway groups' contributions. Assembly pathways are indexed as displayed in Table 3.3.

Figure 3.12: Stacked Trimer Species Fractions over Time (*in vivo*-like)

Subplot (C) is a modified version of Fig. 3.5D with only the colored lines, for ease of reference to the other subplots in this figure. Subplots (A, B, D, E and F) are faceted heatmaps, which can be thought of as a top down view of a 3D histogram with the bar height represented by the scale in the center. Each plot uses data from the five different representative binding strengths for the bonds within the stacked trimer. Yield of each of the 12 species is shown by the horizontal gradients, as indexed in Fig. 3.2. The central "Species Yield" gradient bar indicates the intensity for all sub-figures. As this is an *in vivo*-like simulation total protein is not strictly conserved, but in practice the effective initial monomer concentration, as described in Eqs. 1.11, can be considered constant. This may be obscured at short timescales. As protein is present in multiple intermediates at low concentration. Which, results in a light shade of gray that may display poorly on some displays. Though, this plot type is overall clearer than a 3D histogram once biologically reasonable timescales have been reached.

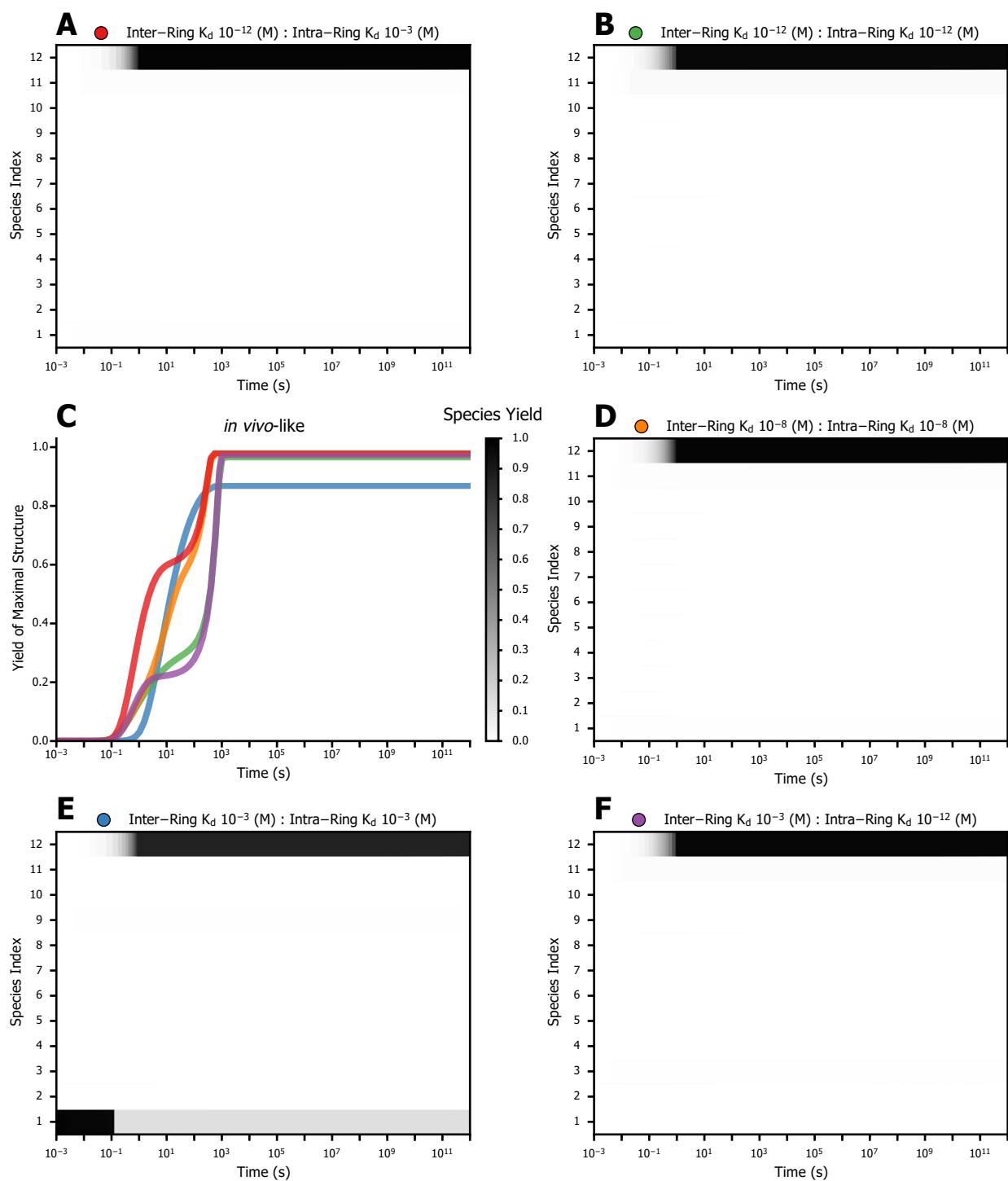


Figure 3.12: Stacked Trimer Species Fractions over Time (*in vivo*-like)
 Caption is on preceding page.
 All parameters not specified are the default values enumerated in Table 3.2.

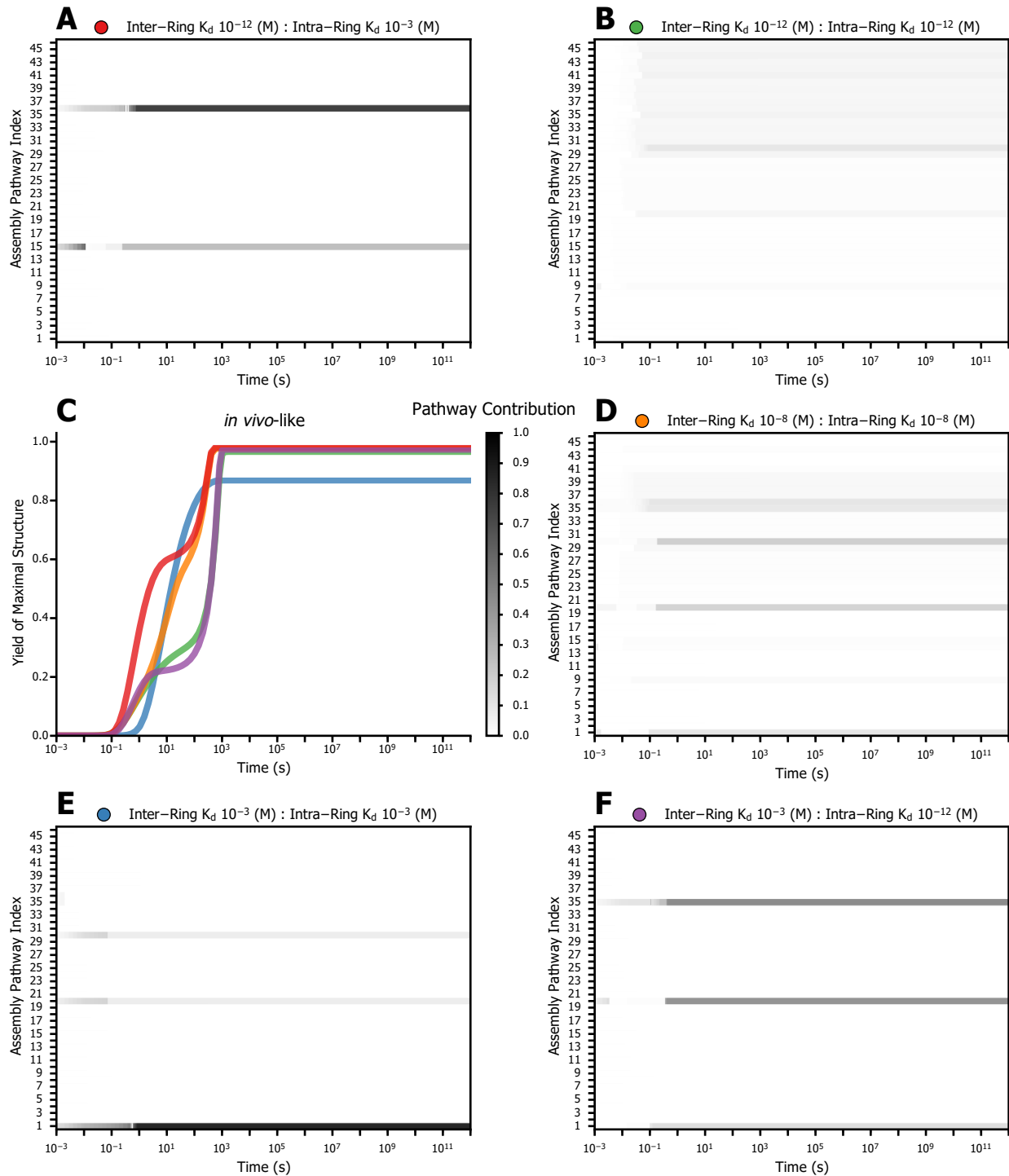


Figure 3.13: Stacked Trimer Assembly Contributions over Time (*in vivo*-like)
 Caption is on following page.
 All parameters not specified are the default values enumerated in Table 3.2.

Figure 3.13: Stacked Trimer Assembly Contributions over Time (*in vivo*-like)

Subplot (C) is a modified version of Fig. 3.5D with only the colored lines, for ease of reference to the other subplots in this figure. Subplots (A, B, D, E and F) are faceted heatmaps, which can be thought of as a top down view of a 3D histogram with the bar height represented by the scale in the center. Each plot uses data from the five different representative binding strengths for the bonds within the stacked trimer. Pathway contribution, calculated as described in Eq. 4.1, is shown as the simulation progresses in time. Showing how dominate pathways can switch as the system progresses. As this is an *in vivo*-like simulation total protein is not strictly conserved, but in practice the effective initial monomer concentration, as described in Eqs. 1.11, can be considered constant. As a reminder, pathway contribution is explicitly normalized. Assembly pathways are indexed as displayed in Table 3.3.

Figure 3.14: Stacked Trimer Species Fractions over Concentration (*in vivo*-like)

Subplot (C) is a modified version of Fig. 3.5F with only the colored lines, for ease of reference to the other subplots in this figure. Subplots (A, B, D, E and F) are faceted heatmaps, which can be thought of as a top down view of a 3D histogram with the bar height represented by the scale in the center. Each plot uses data from the five different representative binding strengths for the bonds within the stacked trimer. Yield of each of the 12 species is shown by the horizontal gradients, as indexed in Fig. 3.2. The central “Species Yield” gradient bar indicates the intensity for all sub-figures. As this is an *in vivo*-like simulation total protein is not strictly conserved, but in practice the effective initial monomer concentration, as described in Eqs. 1.11, can be considered constant. Thus, the resulting yield is normalized for practical purposes.

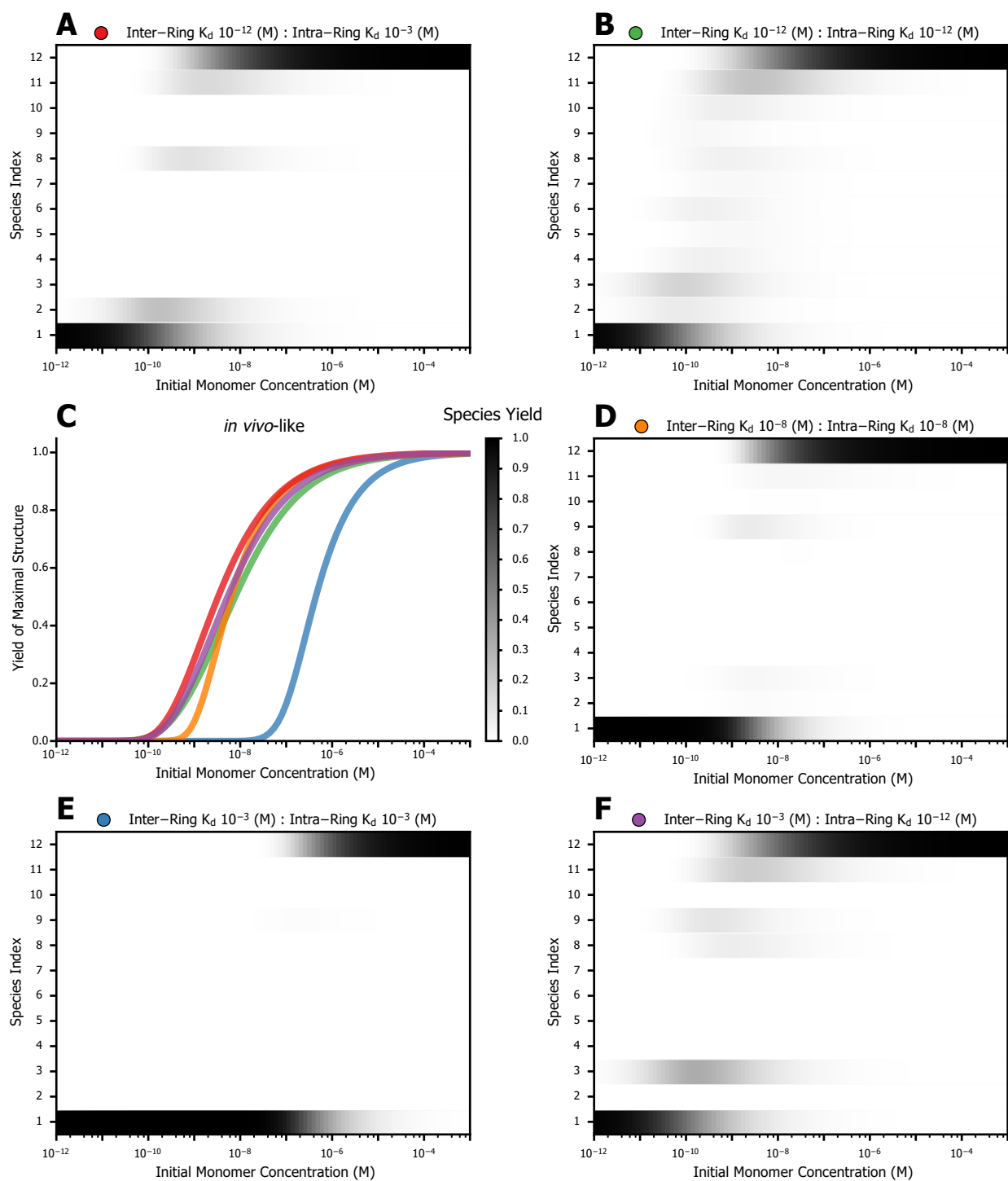


Figure 3.14: Stacked Trimer Species Fractions over Concentration (*in vivo*-like)
 Caption is on preceding page.
 All parameters not specified are the default values enumerated in Table 3.2.

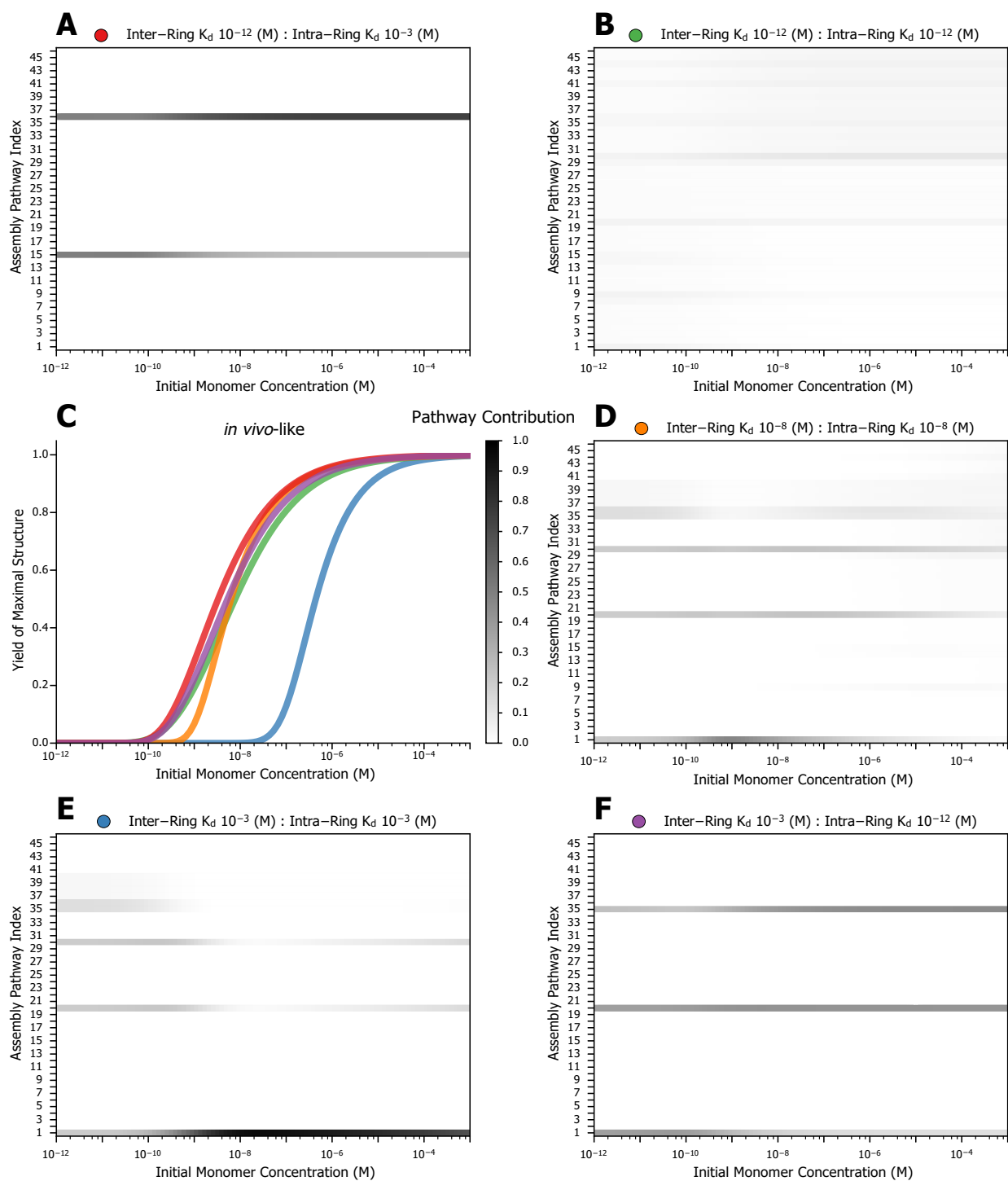


Figure 3.15: Stacked Trimer Assembly Contributions over Concentration (*in vivo*-like)
 Caption is on following page.
 All parameters not specified are the default values enumerated in Table 3.2.

Figure 3.15: Stacked Trimer Assembly Contributions over Concentration (*in vivo*-like)
 Subplot (C) is a modified version of Fig. 3.5F with only the colored lines, for ease of reference to the other subplots in this figure. Subplots (A, B, D, E and F) are faceted heatmaps, which can be thought of as a top down view of a 3D histogram with the bar height represented by the scale in the center. Each plot uses data from the five different representative binding strengths for the bonds within the stacked trimer. Pathway contribution, calculated as described in Eq. 4.1, is shown over a range of concentrations to illustrate how effective initial monomer concentration can affect which pathway is utilized. As this is an *in vivo*-like simulation total protein is not strictly conserved, but in practice the effective initial monomer concentration, as described in Eqs. 1.11, can be considered constant. Thus, the resulting yield is normalized for practical purposes. Assembly pathways are indexed as displayed in Table 3.3.

Figure 3.16: Effects of Synthesis and Degradation Rate on Assembly Yield

Comparison of the effects due to synthesis/degradation rates. Figures in the left column are generated from data at a point in time of \approx one day, whereas figures in the right column are from a time point of 10^{11} (s). (A and B) Cell Division Period is the inverse of δ (s^{-1}) when synthesis and degradation are modeled as cell growth, as discussed in Section 1.3. The CRN was simulated at 22 different δ values logarithmically spaced in the plotted range. The *in vitro*-like case was added at the equivalent of the 10^{17} (s) location. Colored lines match the binding strengths defined in Fig. 3.5's legend, while the gray lines are 125 binding strengths uniformly distributed among the simulated 1000 binding strengths. It can be clearly seen that the system's range of yields narrows considerably as the time elapsed approaches the cell division rate. Significantly larger yield ranges result from the details of the slow cell division rates/*in vitro*-like systems, compared to the more quickly dividing *in vivo*-like systems. I would also like to note that, as one might expect, for long enough cell division periods the system approaches the *in vitro*-like case, but even for the \approx one day case the approach is still significantly proceeding at 10^7 (s). For reference, that is approximately 115 days, which may or may not be meaningful in terms of biological relevance. The specific system under investigation would be the determining factor. (C and D) I have shown both *in vivo*-like (hour per division) heatmaps at the two time points to illustrate the lack of difference between them. For the parameters involved in those plots, the system is clearly at an equilibrium, as can also be seen in Fig. 3.5D. (E and F) Contrastingly, the *in vitro*-like system is affected by the change in time. These phenomena can be compared to the line plot in Fig. 3.5C. Though, the improvements in yield are still correlated to Intra-Ring $K_d >$ Inter-Ring K_d .

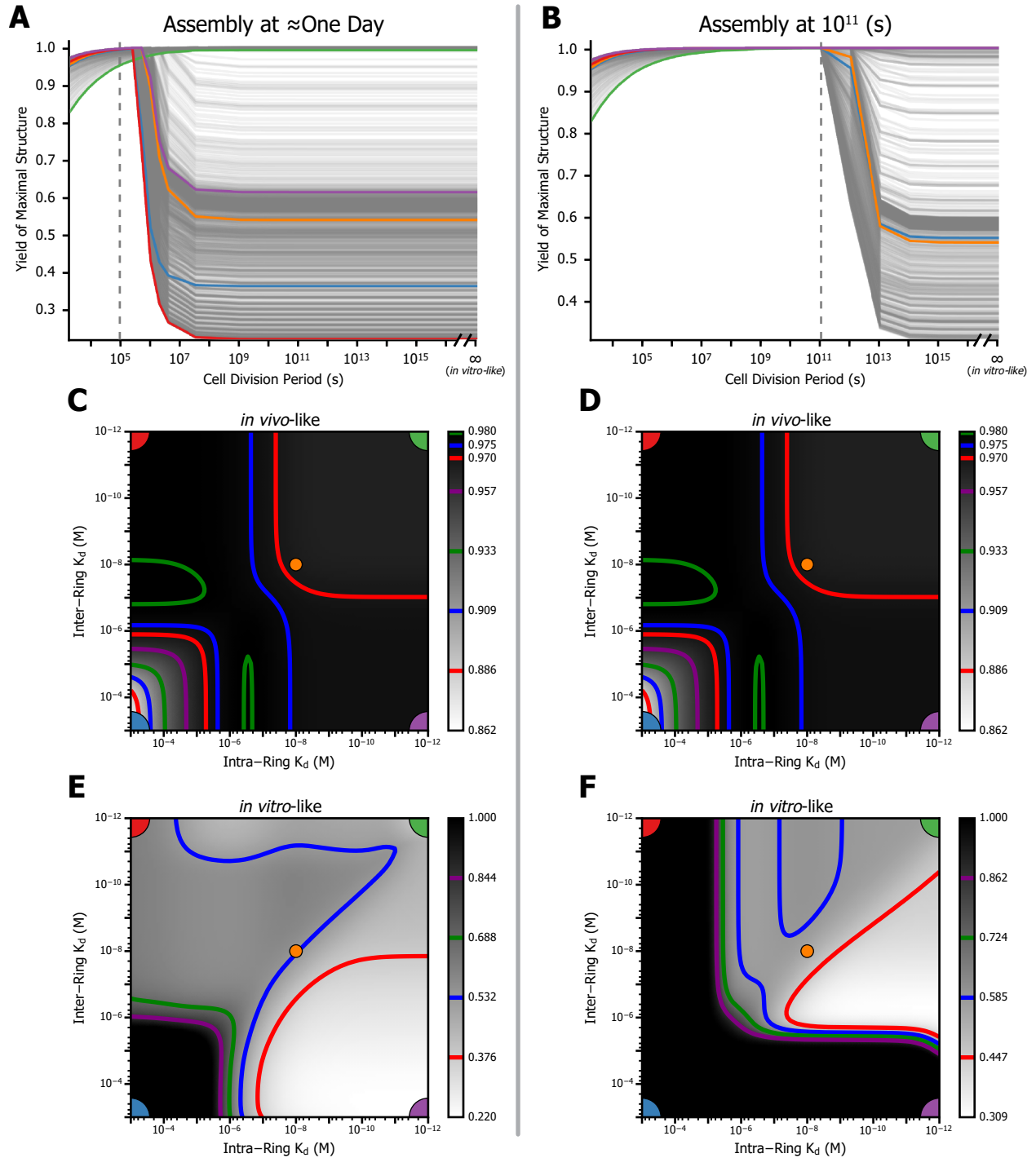


Figure 3.16: Effects of Synthesis and Degradation Rate on Assembly Yield
Caption is on preceding page.
All parameters not specified are the default values enumerated in Table 3.2.

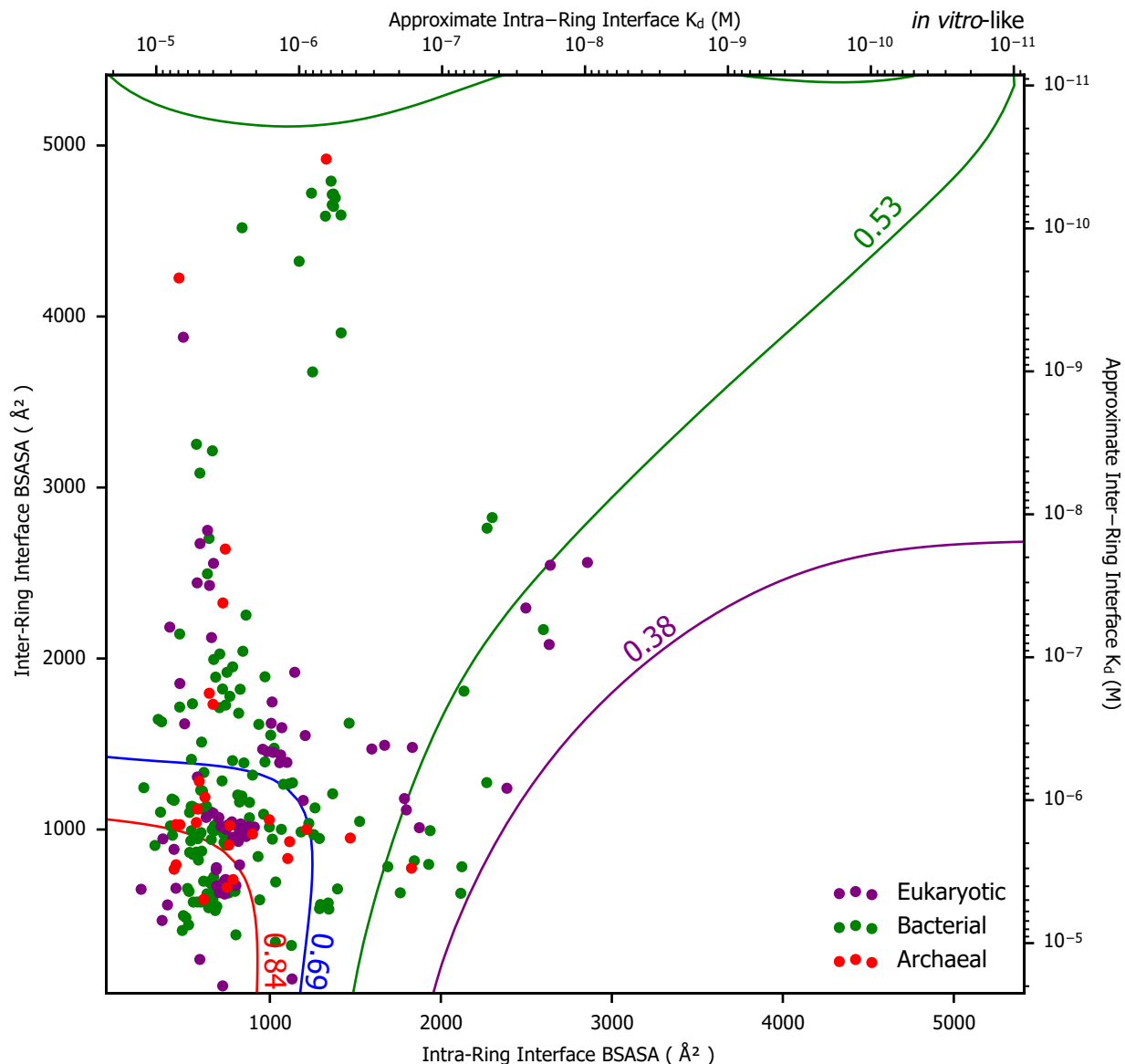


Figure 3.17: Buried Solvent-Accessible Surface Area (BSASA) (*in vitro*-like)

The goal of this figure is to convey the relationship between real proteins and the *in vitro*-like model. BSASA can be used as a rough approximation for binding strength [75]. Using parameters from Chen *et al.*, an exponential relationship is used to align the opposing axes allowing me to provide an overlay of the contour lines from Fig. 3.5A onto this plot. I determined 264 stacked trimer-like proteins from the PDB [73], by writing code to parse structural data from PDBePISA [74], available at <https://github.com/vatir/Protein-Complex-Assembly-Pathways>. Redundancies were then removed based on sequence alignment clustering using clusters made available by the PDB. Clustering of results using sequence identities of 0%, 30%, 95% and 100% resulted in no qualitative changes. 100% was chosen as it removes exact duplicates and uses the highest quality structure when duplicates are present. Protein complexes are colored by domain, as protein complexes from different domains may have different tendencies.

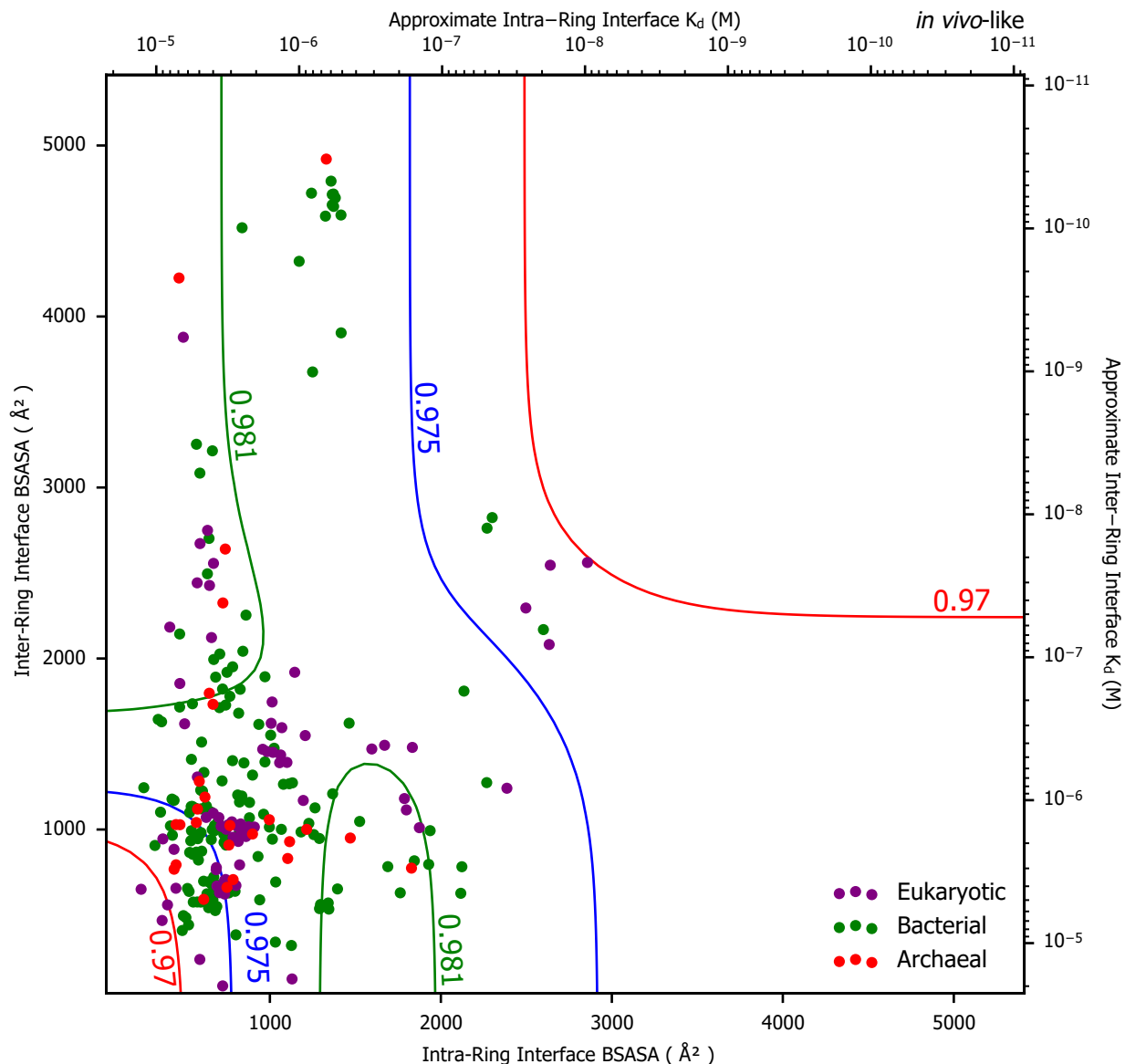


Figure 3.18: Buried Solvent-Accessible Surface Area (BSASA) (*in vivo*-like)

The goal of this figure is to convey the relationship between real proteins and the *in vivo*-like model. BSASA can be used as a rough approximation for binding strength [75]. Using parameters from Chen *et al.*, an exponential relationship is used to align the opposing axes allowing me to provide an overlay of the contour lines from Fig. 3.5B onto this plot. I determined 264 stacked trimer-like proteins from the PDB [73], by writing code to parse structural data from PDBePISA [74], available at <https://github.com/vatir/Protein-Complex-Assembly-Pathways>. Redundancies were then removed based on sequence alignment clustering using clusters made available by the PDB. Clustering of results using sequence identities of 0%, 30%, 95% and 100% resulted in no qualitative changes. 100% was chosen as it removes exact duplicates and uses the highest quality structure when duplicates are present. Protein complexes are colored by domain, as protein complexes from different domains may have different tendencies.

Assembly Pathways

“I have come to believe that the whole world is an enigma, a harmless enigma that is made terrible by our own mad attempt to interpret it as though it had an underlying truth.”

— Umberto Eco

Assembly pathways are labeled rooted binary tree-like structures, with nodes labeled by species type. The binary nature of the structure represents the binary association reactions in the CRN. Each assembly pathway is rooted by the maximal structure and all leaves represent minimal structures. The set of all assembly pathways forms a set of binary trees where each is a possible sequence of reactions capable of producing maximal structures. Since every association reaction in these systems has two reactants and one product, the binary trees will all form full binary trees. A full binary tree is defined as a binary tree where: the root node has two neighbors, leaf nodes have one, and every other node has three. This is illustrated in Table 3.3. The general binary tree structure is not the difficult part of this system to formalize, the problems arise when one tries to define uniqueness, isomorphisms, or exhaustive enumeration. On the upside, this formalism is not limited to the stacked trimer or even Stacked Rings, it works (in some cases with minor extensions) for all protein complexes.

4.1 Trees: A Survey

This section will have a wide variety of terms which may be unfamiliar to the reader. I will provide further reading suggestions and references where possible, but the key point to understand is how an assembly pathway differs from other well-known structures, not the details of all the other structures. My goal is to find common ground across disciplines and clear up potential misunderstandings arising from terminology that harbors subtle differences in meaning.

Trees are likely a familiar concept to most readers, but they are often defined differently depending on the environment in which the definition was developed. I will now cover several common and relevant definitions and how they relate to assembly pathways. I distinguish between fields of study as if they are perfectly secluded from one another, but this is purely for pedagogical benefit as attempting to describe both overlapping concepts and words tends to result in great confusion. Amusingly, this concept of term inheritance from one field to another maps rather well to the Method Resolution Order problem for multiple inheritance in object-oriented programming. Originally it was intended to be a Tree and was designed with restrictions related to Tree structures, then was generalized to a Graph structure.

Within Mathematics, there are a variety of definitions for the term “Tree,” the earliest being linked to a mixture of graph theory and enumerative combinatorics. Arthur Cayley introduced the term “Tree” in 1857 to aid in counting patterns of differential operators [76]. Its relationship to graph theory actually comes from Kirchhoff’s work in developing early spanning tree theories, which led to his *eponymous* laws [77, 78]. Cayley merged the use of the word Tree into both fields with his *eponymous* formula, which counts the number of spanning trees of a complete graph [79, 80]. In all of these approaches, and most variations following them, node labels are required to be unique. Non-unique labels are found in mathematics, but the implementation details of both the labeling and the definition of isomorphism are very specific to the topic they are being used for, *e.g.*, chemical graphs

use atomic or substructure labels but require isomerisms to maintain chemical properties. The term “Graph” was not even used in its modern form until 1878, derived from the term chemical “graphic notation,” without an implication of trees being a type of graph. This illustrates the lack of clarity in modern naming of these concepts, especially within the field of Mathematics. Overall, the “homomorphic” nature of the node/edge structure of Trees tends to be the same between Mathematical subfields, but the definition of structure preserving “isomorphic” relations is often adjusted for the specific use case [81, 82].

Meanwhile, Computer Science has defined a distinctly different, but of course similarly named collection of structures using implicit definitions which evolve as needed due to changes in the use case. This is well summed up the following quote:

“The very nature of computer representation defines an implicit ordering for any tree, so in most cases ordered trees are of greatest interest to us. We will therefore tacitly assume that *all trees we discuss are ordered, unless explicitly stated otherwise.*” - Donald Knuth (The Art of Programing)

I would describe Trees used in Computer Science as structures of convenience with definitions customized to a use case, as opposed to an overarching general mathematical formalism. It would appear to be a universal trait that trees are acyclic, but there are cases of this rule being bent or dropped, *e.g.*, functional programming language recursion trees, relational database conflict trees, and automated theorem proving with cyclical references [83, 84]. In Binary Search Trees, labels represent elements from another structure and a partial ordering is required, but uniqueness is not. Binary Search Trees are a type of Ranked tree, meaning that the child nodes have a defined left and right [81]. Overall, it is common to define Trees in Computer Science through a recursive relationship [85]. First, a Root Node is defined, then each child is defined as a disjoint subtree. The disjoint property is very useful in accelerating traversal algorithms, but is not generally required. The ability of computer representation to treat Tree-like structures as traversable collections of pointers between abstract objects is both its most powerful property and its most insidious.

4.2 General Assembly Pathways

Assembly pathways have the following properties:

- Rooted:** A node is defined as a unique starting point in the Tree.
- Hierarchical:** All child nodes have fewer objects than their parents.
- Weighted:** Edges can represent more than one CRN association path.
- Full:** Every non-leaf node has exactly two children.
- Labeled:** Each node is labeled with a canonical species.
- Non-Unique:** More than one node can be labeled with the same canonical species.
- Partially Ranked:** Left-Right child relationships are determined by label order, except when children are equal.

Assembly pathways can be enumerated without all of these traits, but uniqueness requires them all. Therefore, I proceeded by first enumerating the assembly pathways then reducing them to a unique subset of equivalence classes, much in the same way as the species enumeration in Section 3.1. For enumeration, I utilized a recursive Computer Science style definition for assembly pathways, as non-unique labeling is largely undefined in other descriptions. Most commonly, the right and left subtrees are disjoint in such definitions, but it is not strictly required.

Enumeration of assembly pathways was carried out by a method laid out in Python style pseudocode in Listing 1. Code presented in this document is intended for pedagogical purposes, as such it is not optimized for runtime or programming style elegance. Full source code used in the generation of these results will reside for a reasonable period of time in the GitHub repository: <https://github.com/vatir/Protein-Complex-Assembly-Pathways>. Methodologies in this chapter require only a listing of the canonical species and reactions, meaning they must have already been determined. The reader may notice that this algorithmic approach is an iterative form of a stack-based forest traversal. In essence, ad-

Table 4.1: Tree Properties of Stacked Rings

Ring Size	# of Node Type				Canonical	
	Terminal	Internal	Index	Total	Species	Reactions
3	6	5	4	11	12	34
4	8	7	5	15	27	145
5	10	9	6	19	59	508
6	12	11	7	23	135	1,665
7	14	13	8	27	307	5,055
8	16	15	9	31	717	14,833
9	18	17	10	35	1,682	42,079

ditional trees are generated when multi-way choices in the reaction list allow for alternate formation paths. Enumeration using this method generates a subset of possible assembly pathways reduced by canonical equivalence. Each edge in the tree can represent multiple equivalent CRN paths. The partially ranked property causes issues with distinguishing between pathways where the left child equals the right child but the subtrees are different, *e.g.*, pathway three in Table 3.3. Removal of these equivalent assembly pathways is left to the uniqueness step. Non-equal children are initially set to left or right based on the order of the integer representation as defined in Section 4.3.

Uniqueness is more difficult; while many methods will accomplish the same result as the enumeration algorithm, the uniqueness algorithm must be able to correctly group assembly pathways with the same canonical index without leaving equivalent assembly pathways with different indices. While it is potentially possible to *a priori* determine the cardinality of the equivalence classes and their respective memberships, it is debatably a more difficult task [81, 86].

I will first introduce a general method for determining canonical indices, then I will comment on how to improve it into a minimum description and assembly pathway storage system. To begin, one can describe a ranked rooted tree by the result of a traversal algorithm. I will be using a Depth First Traversal (Prefix Order) (DFT). If an assembly pathway was definitely ranked or had labels which fulfilled the requirements for a Binary

Search Tree, this would be a brief undertaking, but in many ways, it is a rather unique and challenging form of binary tree. Assembly pathways can have subtrees which are equal and only become apparent after traversing multiple steps below a given right-left split, and the ability of the children involved in the right-left split to be equal or non-equal only makes it more problematic.

In the first simple form of the canonical indexing algorithm I will approach the problem from an exhaustive standpoint. Effectively, every right-left split will be considered a potential point where the children could be swapped to generate an equivalent assembly pathway. I refer to this as partially ranked because not every reversal creates a new assembly pathway, but some do. Listing 2 provides a pseudocode description of a function which takes in an assembly pathway and returns all possible unique DFTs as a set. Then all prospective assembly pathways are compared and any intersection between sets of canonical indices groups them together into an equivalence class. This method is computationally expensive, but does catch several pesky patterns which start showing up in structures larger than a stacked trimer.

```

PathwaysWithTasks = Ordereddict()
    # Order preserving dictionary
    # Keys : All Unfinished Pathways, Values : Todo deque for the Graph
CurrentPathway = Pathway()
    # Instantiate first empty pathway object
CurrentPathway.add_root_node(RootNodeName)
    # Add starting node
PathwaysWithTasks[CurrentPathway] = deque([RootNodeName, ])
    # Add first root node to todo stack
CompletedPathways = list()
    # List for holding completed Assembly Pathways

while len(PathwaysWithTasks) > 0:
    # Run while pathway(s) have remaining task(s)
    CurrentPathway = PathwaysWithTasks.keys()[0]
    # Get first Pathway with tasks remaining
    CurrentNode = PathwaysWithTasks.values()[0].pop()
    # Pull next task off of the stack for the CurrentPathway
    ReactionClasses = reactants_lookup(CurrentNode)
    # All possible reactions resulting in CurrentNode as a list of 2-tuples

    if len(ReactionClasses) > 1:
        # If more than one way to form the CurrentNode save a copy of the Pathway and current todo stack
        OriginalPathway = copy(CurrentPathway)
        OriginalDeque = copy(PathwaysWithTasks[CurrentPathway])

        # Create copy of OriginalPathway since there is more than one way to create CurrentNode
        for Index, Children in enumerate(ReactionClasses):
            if Index > 0:
                CurrentPathway = copy(OriginalPathway)
                PathwaysWithTasks[CurrentPathway] = copy(OriginalDeque)

            # Add Children to CurrentNode with unique nodenames
            NodeName1 = unique_node_name(Children[0])
            CurrentPathway.add_node(NodeName1)
            CurrentPathway.add_edge(CurrentNode, NodeName1)
            NodeName2 = unique_node_name(Children[1])
            CurrentPathway.add_node(NodeName2)
            CurrentPathway.add_edge(CurrentNode, NodeName2)

            # Add non-terminal nodes to the todo stack for CurrentPathway
            if not_a_monomer(NodeName1):
                PathwaysWithTasks[CurrentPathway].append(NodeName1)
            if not_a_monomer(NodeName2):
                PathwaysWithTasks[CurrentPathway].append(NodeName2)

        # If Pathway is completed, remove from PathwaysWithTasks and add to CompletedPathways
        if empty(PathwaysWithTasks[CurrentPathway]):
            CompletedPathways.append(CurrentPathway)
            del PathwaysWithTasks[CurrentPathway]

```

Listing 1: Assembly Pathway Enumeration

```

def CanonicalIndex(Pathway, RootNode):
    BranchingNumber = 0
    # Number of possible Right/Left swaps
    for Node in Pathway.nodes():
        if is_not_monomer(Node):
            BranchingNumber += 1
            # Pathway can branch up to the number of non-monomeric nodes
    Combinations = all_combinations(BranchingNumber)
    # Returns all possible sequences of "Left" and "Right" of length BranchingNumber
    TraversalCombinations = set()
    # Holds all unique traversals
    for Combination in Combinations:
        Combination = deque(Combination)
        Tasks = deque()
        Tasks.append(RootNode)
        # Starts each traversal with the RootNode
        PreorderDFSTraversal = []
        # Holds current traversal
        while len(Tasks) > 0:
            CurrentNode = Tasks.pop()
            # Traversal order is controlled by the order of the Tasks stack
            PreorderDFSTraversal.append(cannonical_species_representation(CurrentNode))
            # Adds the CurrentNode to the PreorderDFSTraversal in order based on the Tasks stack
            if is_not_monomer(CurrentNode):
                # Monomers do not have children
                ChildNodes = get_children(Pathway, CurrentNode)
                # Gets children of the CurrentNode in species index order
                BranchDirection = Combination.pop()
                if BranchDirection == "Left":
                    Tasks.append(ChildNodes[0])
                    Tasks.append(ChildNodes[1])
                elif BranchDirection == "Right":
                    Tasks.append(ChildNodes[1])
                    Tasks.append(ChildNodes[0])
                # Adds Tasks to the stack based on traversal order Combination
            TraversalCombinations.add(tuple(PreorderDFSTraversal))
            # Adds the current completed traversal as a hashable object to the set
    return TraversalCombinations

```

Listing 2: Unique Pathway Determination

Several simplifications can be made to the canonical indexing process. Let us start by assuming there exists a minimal canonical index which represents each equivalence class. Then, if a representation can be used to uniquely regenerate all unique pathways and cannot be reduced any further, it is a minimal canonical index. All assembly pathways are rooted by the maximal structure, thus the first entry of every index will be the same, so it can be dropped without loss of generality. Therefore, minimal will be defined as a tuple containing one less entry than the number of internal nodes, see Table 4.1. All intermediate species must be named to define an assembly pathway, hence the minimal length definition.

The representative index will be chosen by selecting the maximal DFT from the generated options. By using the number of monomers in each species, it becomes unnecessary to enumerate leaf nodes and retain information about branch points in the index sequences. Thus, the pathway can be regenerated using only the monomer counts of the species, the ring size and a tuple of non-leaf entries. This has been directly tested by comparison to the original enumerations up through ring sizes of seven. If a pathway has no nodes with equal children then the enumeration algorithm will not generate more than one member of the equivalence class, *i.e.*, if an assembly pathway is ranked (all children have a predetermined left-right relationship) then it does not need to be checked for duplicates. I wrote the accelerated algorithm in Cython with OpenMP based threading [87]. Enumeration of the assembly pathways for the stacked trimer are shown in Table 4.3. Results of all pathways generated can be found in Table 4.3.

Table 4.2: Stacked Trimer Assembly Pathway Indices

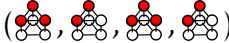

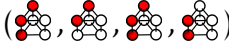
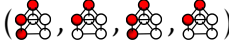

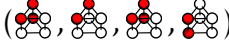

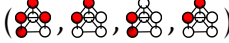


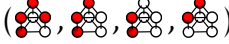

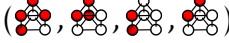



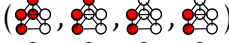
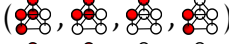
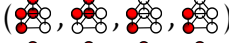


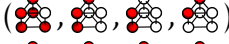

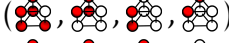

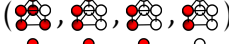

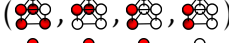



Pathway	Pathway Index by Species Representation		
	Image	Number	Integer Representation
1		(9, 3, 9, 3)	(248, 176, 248, 176)
2		(6, 2, 6, 2)	(244, 164, 244, 164)
3		(6, 3, 6, 2)	(244, 176, 244, 164)
4		(6, 3, 6, 3)	(244, 176, 244, 176)
5		(4, 2, 4, 2)	(242, 164, 242, 164)
6		(4, 3, 4, 2)	(242, 176, 242, 164)
7		(4, 3, 4, 3)	(242, 176, 242, 176)
8		(10, 3, 2, 3)	(316, 176, 164, 176)
9		(10, 9, 3, 3)	(316, 248, 176, 176)
10		(10, 6, 3, 3)	(316, 244, 176, 176)
11		(10, 6, 2, 3)	(316, 244, 164, 176)
12		(10, 4, 3, 3)	(316, 242, 176, 176)
13		(10, 4, 2, 3)	(316, 242, 164, 176)
14		(8, 3, 3, 2)	(310, 176, 176, 164)
15		(8, 2, 2, 2)	(310, 164, 164, 164)
16		(8, 6, 3, 2)	(310, 244, 176, 164)
17		(8, 6, 2, 2)	(310, 244, 164, 164)
18		(8, 4, 3, 2)	(310, 242, 176, 164)
19		(8, 4, 2, 2)	(310, 242, 164, 164)
20		(11, 9, 3, 3)	(382, 248, 176, 176)
21		(11, 6, 3, 3)	(382, 244, 176, 176)
22		(11, 6, 2, 3)	(382, 244, 164, 176)
23		(11, 4, 3, 3)	(382, 242, 176, 176)
24		(11, 4, 2, 3)	(382, 242, 164, 176)
25		(11, 6, 3, 2)	(382, 244, 176, 164)
26		(11, 6, 2, 2)	(382, 244, 164, 164)
27		(11, 4, 3, 2)	(382, 242, 176, 164)
28		(11, 4, 2, 2)	(382, 242, 164, 164)
29		(11, 10, 3, 2)	(382, 316, 176, 164)
30		(11, 10, 9, 3)	(382, 316, 248, 176)
31		(11, 10, 4, 3)	(382, 316, 242, 176)

Table 4.2: Stacked Trimer Assembly Pathway Indices (Continued)

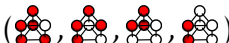
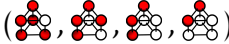


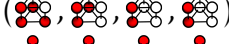


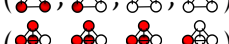

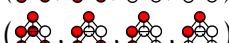

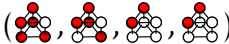
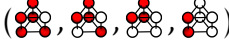
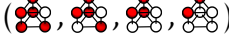













Pathway	Pathway Index by Species Representation		
	Image	Number	Integer Representation
32		(11, 10, 4, 2)	(382, 316, 242, 164)
33		(11, 10, 6, 3)	(382, 316, 244, 176)
34		(11, 10, 6, 2)	(382, 316, 244, 164)
35		(11, 8, 3, 3)	(382, 310, 176, 176)
36		(11, 8, 2, 2)	(382, 310, 164, 164)
37		(11, 8, 6, 3)	(382, 310, 244, 176)
38		(11, 8, 6, 2)	(382, 310, 244, 164)
39		(11, 8, 4, 3)	(382, 310, 242, 176)
40		(11, 8, 4, 2)	(382, 310, 242, 164)
41		(11, 7, 3, 3)	(382, 309, 176, 176)
42		(11, 7, 6, 3)	(382, 309, 244, 176)
43		(11, 7, 6, 2)	(382, 309, 244, 164)
44		(11, 5, 3, 3)	(382, 307, 176, 176)
45		(11, 5, 4, 2)	(382, 307, 242, 164)
46		(11, 5, 4, 3)	(382, 307, 242, 176)

Table 4.3: Assembly Pathway Properties of Stacked Rings

Ring (Size)	Assembly Pathways				Final Step (Add Monomer)
	Enumerated	Unique	Duplicate	% Duplicates	
3	48	46	2	4.17%	59%
4	1,034	982	52	5.03%	62%
5	29,402	28,319	1,083	3.68%	65%
6	1,057,462	1,031,629	25,833	2.44%	68%
7	44,169,144	43,438,639	730,505	1.65%	71%
8	2,142,675,188	2,118,125,722	24,549,466	1.14%	74%

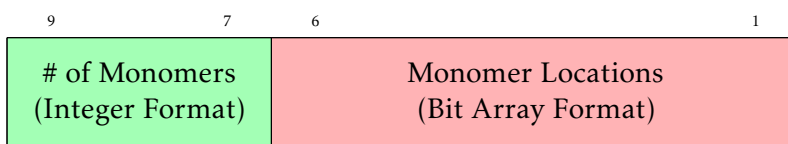
Table 4.4: Representations of the Stacked Trimer

Index	Species		Monomer Count		Monomer Location	
	Image	Integer	Integer	Bit Array	Integer	Bit Array
1		96	1	0,0,1	32	1,0,0;0,0,0
2		164	2	0,1,0	36	1,0,0;1,0,0
3		176	2	0,1,0	48	1,1,0;0,0,0
4		242	3	0,1,1	50	1,1,0;0,1,0
5		307	4	1,0,0	51	1,1,0;0,1,1
6		244	3	0,1,1	52	1,1,0;1,0,0
7		309	4	1,0,0	53	1,1,0;1,0,1
8		310	4	1,0,0	54	1,1,0;1,1,0
9		248	3	0,1,1	56	1,1,1;0,0,0
10		316	4	1,0,0	60	1,1,1;1,0,0
11		382	5	1,0,1	62	1,1,1;1,1,0
12		447	6	1,1,0	63	1,1,1;1,1,1

4.3 Integer Representation of Complexes

Since I am using a computational description of assembly pathways, I will provide a computer intelligible description of a species. One way to describe any protein complex is to assign each potential object in the structure an entry in a bit array. I will alternate between using the term object and the term monomer in this section, where object refers to any entity that can be a relevant substructure in a protein complex, and monomer will refer to objects within the stacked trimer as it is solely formed of protein monomers. Assembly pathways can be constructed for any definable protein complex, which may contain more than one type of protein monomer, small organic molecules, DNA or many other sub-objects. Table 4.4 contains an example of monomer locations for the stacked trimer. The alignment of locations to the visual representation of the species is only for human convenience; they can be arranged in any order as long as it is exhaustive.

The overall species index is ordered in terms of the integer representation of the monomer location bit array. While this was part of the original implementation, it was later realized that choosing an order which provides information to algorithmic processes is useful. Assembly pathways only allow for association reactions, so every child node must have strictly fewer objects than its parent. Therefore, maintaining an object count in the most significant bits of the bit array generates an order such that if a species has a smaller representation then it contains equal or fewer objects than other species. In the case of the stacked trimer the following representation is used:



Generalizing the bit array representation to arbitrary structures is accomplished by repeating the object location sequence once for each type of object that can be found in the protein complex, as follows:

# of Objects (Integer Format)	Object Type N (Bit Array Format)	...	Object Type 1 (Bit Array Format)
----------------------------------	-------------------------------------	-----	-------------------------------------

I have found that using a single representation for all tasks is non-optimal. Using just the object location bit array is useful in several stages for generating the systems, whereas the full representation is best for definitive identification of the species. As such, it is a better property to sort by when generating the arbitrary species index, which is the most compact.

4.4 Pathway Contribution

Assembly pathways describe how a protein complex can assemble, but do they contribute equally to that assembly? In Section 3.1, I covered equivalence classes of reactions. Just as assembly pathways describe many ways for an assembly pathway to form a maximal structure, there is often more than one way for two children to form a parent. I relate each internal node in an assembly pathway to an association reaction, by choosing an association reaction by using the child nodes' labels. The cardinality of the reaction is the number of ways the reaction could be created between non-canonical species, *i.e.*, the number of possible ways for the reaction to occur in the physical system, as opposed to my equivalence classes canonical representations. Along this line, I consider what the relative fractional contribution to the generation of a species is from all possible reactions which result in that species. Following this logic, I define the following:

Association reaction with product P: A_P

Binary reactants of A_P : R_{1,A_P}, R_{2,A_P}

Association rate of A_P : k_{on,A_P}

Class of reactions resulting in product P: C_P

Association reaction equivalence class cardinality: K_{A_P}

$$\text{Reaction Usage Factor: } RU(A_p) = \frac{K_{A_p} k_{\text{on},A_p} [R_{1,A_p}] [R_{2,A_p}]}{\sum_{\alpha \in C_p} K_{\alpha} k_{\text{on},\alpha} [R_{1,\alpha}] [R_{2,\alpha}]}$$

All nodes in assembly pathway (i): N_i

$$\text{Normalization factor for Pathway Contribution: } PC_{\text{Norm}} = \sum_{i=1}^{|PC|} \prod_{n \in N_i} RU(n) \quad (4.1)$$

$$\text{Pathway Contribution for assembly pathway } (i): PC_i = \frac{1}{PC_{\text{Norm}}} \prod_{n \in N_i} RU(n)$$

This results in a concept of relative pathway contribution by each assembly pathway. All systems in this document use the same k_{on} for all reactions, so they cancel in the reaction usage factor whereas K_{A_p} is effectively an edge weighting factor in the assembly pathways. It is important to note that not all assembly pathways are equally important even if the concentrations of the species are idealized away. Figure 4.1 shows four of the important assembly pathways of the stacked trimer.

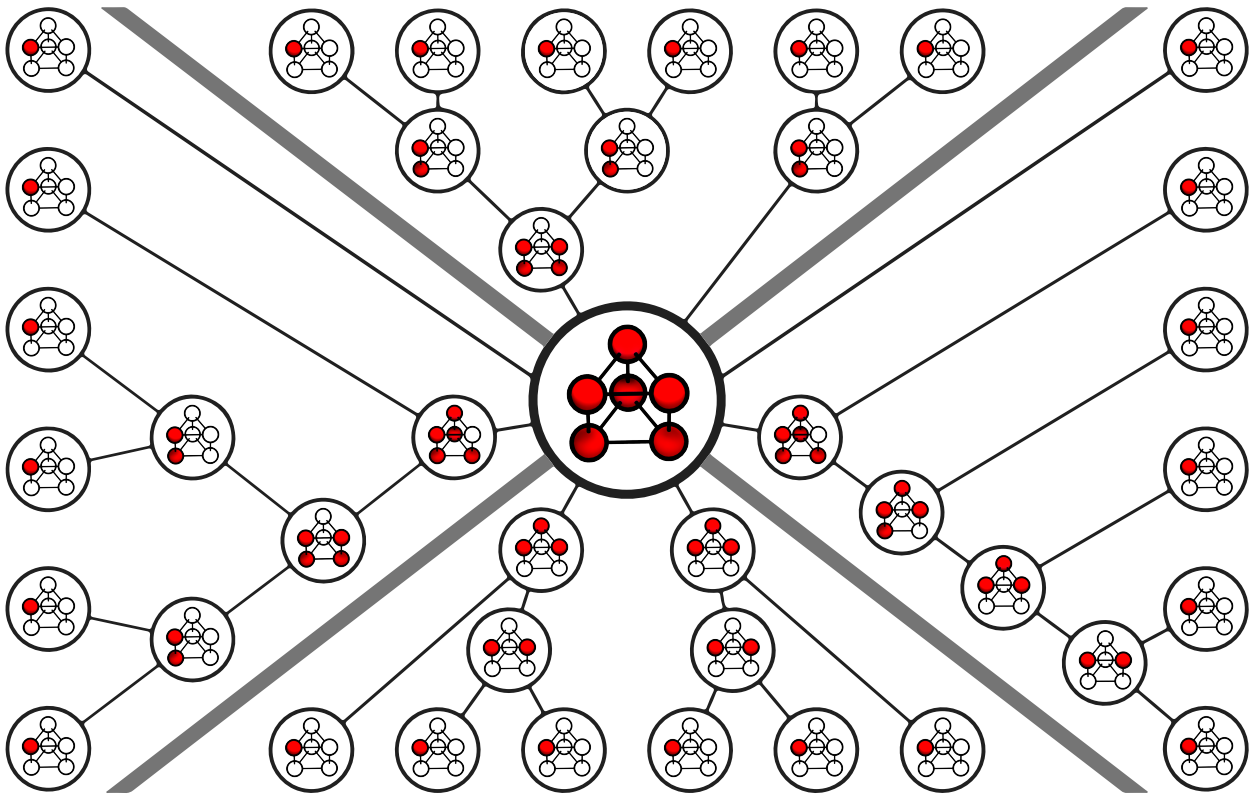


Figure 4.1: Illustrative Examples of Stacked Trimer Assembly Pathways

Proteasome Assembly Modifiers

“The Information Age offers much to mankind, and I would like to think that we will rise to the challenges it presents. *But it is vital to remember that information — in the sense of raw data — is not knowledge, that knowledge is not wisdom, and that wisdom is not foresight.* But information is the first essential step to all of these.”

— Arthur C. Clarke

The final abstraction I will be adding is a replacement operation. Think about the stacked trimer, then consider the possibility that a non-protein object(s) may be able to bind to into the structure. This alternate object could have different binding sites than monomeric proteins, *e.g.*, DNA:ISWI binding as in Chapter 2. Alternately, it could bind in place of a protein monomer thus precluding a protein at that location, potentially with a different binding strength. If that object admits bonds that are not included in the original final structure, then new assembly pathways and alternate final structures would be added. While that would be an interesting path to follow, its abstraction tends to result in a combinatorial explosion of possibilities, rendering analysis of the system difficult to reconcile with real systems. Therefore, for this work I will only be allowing replacements that can occlude a binding site. As they prevent “normal” binding, they are traditionally referred to as *inhibitors*. I mention “traditionally” as they do inhibit binding, but they do not necessarily inhibit assembly. For this work, I will consider only inhibitors that

Table 5.1: # of Canonical Species and Reactions in the Modified Proteasome System

Inhibitor Class	# of Species	# of Reactions
None	876	18,150
α - α	2,623	51,459
β - β	2,623	51,459
α - β	4,600	88,884
β - α	4,600	88,884

bind to a site already available in the structure. However, the inhibitors will not satisfy the same bonds as the original monomeric protein. This results in the binding kinetics of the inhibitors following an A:B style of interaction as in Section 1.1, whereas the protein follows kinetics as outlined in Chapter 3. Inhibitors are represented well by this construction, as they tend to be far smaller than proteins, but large enough to interfere with the normal protein:protein binding at that site.

I will use proteasome core particle formation as a case study for this abstraction, see Fig. 5.1A. To be clear, I am only referring to the 20S proteasome complex made up of 28 proteins in a four-layer barrel configuration. The regulatory caps are not being considered for this work. Also, α and β subunits are considered to only come in one type each. Similar to ISWI, the proteasome is a critical part of many biological systems, including both humans and human pathogens [88–93]. The most relevant detail about proteasomes, for this discussion, is that both over and under abundance of proteasomes have been related to human pathologies [94–101]. Hence, regulation of proteasome assembly has promise in several important fields of human health.

5.1 Technical Description

Proteasomes very likely form by assembly of Half-Proteasomes (HP), see Fig. 5.1B, followed by dimerization into the final complex [102]. Hence the CRN will be based on assembly of the HP with a final reaction completing the proteasome. Using the same core formalism

as the stacked trimer, I extended a previously existing codebase for generating canonical species and reaction classes to include inhibitor binding. This allows for the modeling of the HP, then the final dimerization step is added as a later step. Just as in the stacked trimer the proteasome is considered to have dihedral symmetry, in this case it will be D_7 . Note, that the HP only has C_7 due to α and β not being interchangeable. These symmetries lead to six natural classes of bonds, shown in Figs. 5.1C and 5.1D, four of which seemed to be the best candidates for targeting with small molecule non-covalent bonding inhibitors. I will not be considering inhibitors for the $\beta^\ddagger-\beta^\ddagger$, which bridges the two halves of the completed proteasome. There is evidence that the HP dimerization interfacing process is dominated by interactions which would be poorly represented by an inhibitor of this type. Figure 5.1C does refer to a $\beta-\beta$ class inhibitor: specifically it refers to an inhibitor interfering with the intra-ring $\beta-\beta$ bond. In both the $\beta-\beta$ and $\alpha-\alpha$ classes the system is effectively symmetric so a clarification of bond type is unnecessary. Please note, in the cases of $\alpha-\beta$ and $\beta-\alpha$ the physical interface to which the inhibitor is bound does matter to the system, hence the definition of two separate classes for those bonds. A collaborating research group, Dr. Karanicolas's Lab, has the capability to identify drug candidate compounds from analysis of the binding site of a protein's crystal structure. The choice to simulate $\alpha-\beta$ and $\beta-\alpha$ but not $\alpha'-\beta'$ was based on the expedience of determining if and which bonds might be suitable for them to target.

Unlike the stacked trimer software development process where I was able to develop a robust simulation and analysis suite, the inherited codebase for this system was based around running quickly, with minimal output and requiring substantial human intervention. My stacked trimer system turned out to have poor memory scaling when extended to systems with species and reaction counts in the range of the proteasome system, see Table 5.1. As the project initially was very time constrained, I identified the critical development priorities to be extending the species and reaction generation engine to include the inhibitors and a lightweight command and control structure for Extracting, Transforming

and Loading (ETL) the resultant data into a workable structure. Unfortunately, only the final species fraction vs time is retained, so in-depth analysis of the type in Chapter 3 was unavailable. I also had to rely on the original codebase for generating stacked rings with a different protein in each ring and a set of simulation parameters for the system without inhibitors, see Table 5.2. The parameters were determined by fitting experimental data prior to my involvement with Dr. Deeds’s lab. I should say that for the purpose of this work, the exact parameters are not especially critical, as this work focuses more on the emergent results of modifying those interactions. Additionally, all simulations were performed with initial concentrations of 4×10^{-6} Molar α and β , with *in vitro*-like conditions.

5.2 Analysis

Inhibition of protein-protein interactions can lead to increased yield of protein complexes. This is without question the most important result of this analysis. Figs. 5.2–5.5 clearly show inhibitor classes α – β and β – α can both increase and decrease yield, depending on inhibitor binding strength and concentration. Under some conditions, localized concentration in a biological system could be used to simultaneously modify protein complex assembly by inhibiting in some cases and stimulating in others. While, these phenomena

Table 5.2: Bond Strengths of the Simulated Proteasome Core Particle

Bond Class	k_{on} ($\text{M}^{-1}\text{s}^{-1}$)	K_d (M)
α – α	3,500	8.9×10^{-3}
β – β	3,500	1.9×10^{-2}
α – β	3,500	5.6×10^{-5}
α' – β'	3,500	3.4×10^{-3}
β^{\ddagger} – β^{\ddagger}	3,500	$\approx 0^{\ddagger}$

[‡]Half-Proteasome dimerization in this model is considered an irreversible process. Recall from Chapter 1, $K_d = k_{\text{off}}/k_{\text{on}}$ thus if $k_{\text{off}} \approx 0$ then K_d is ≈ 0 .

could be exploited for research and clinical benefit. Serious questions are raised about the widespread application of inhibitory compounds discovered by methods that do not elucidate functional understanding, *i.e.*, a compound discovered by high throughput screening, tested in animal models and carefully administered in clinical trials, could have the opposite effect in real world conditions when the effective concentration is lowered or increased. Hence, I use the term *Assembly Modifiers* in relation to compounds which inhibit a specific protein-protein interaction within a protein complex. As can be seen in the structure of HP, Fig. 5.1B, the system has a high level of symmetry in the $\alpha\text{-}\alpha$ / $\beta\text{-}\beta$ and $\alpha\text{-}\beta$ / $\beta\text{-}\alpha$ classes. While not equal due to bond strength differences, this is to be expected.

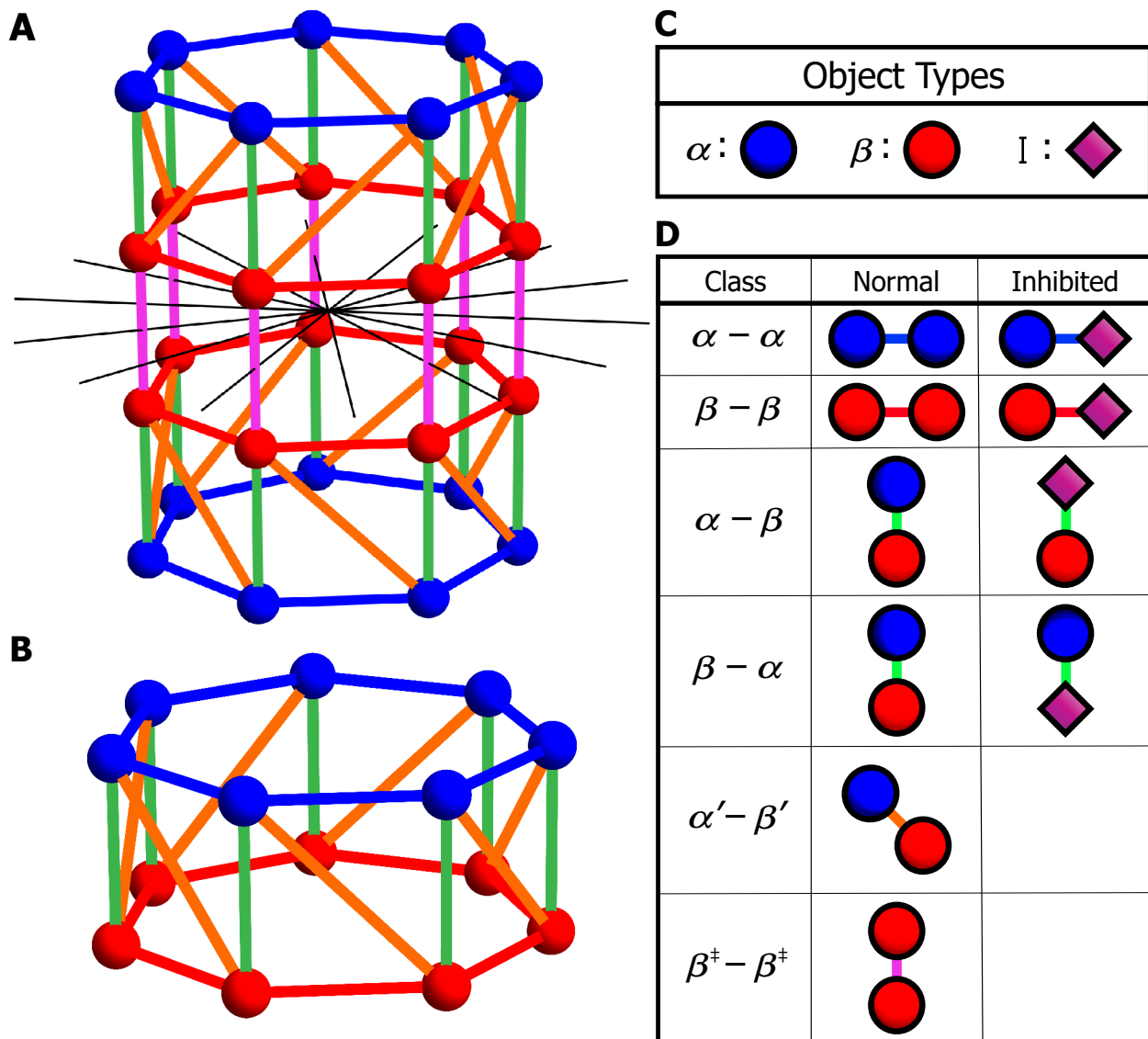


Figure 5.1: Illustration of Proteasome, Half Proteasome and Bond Types

(A) Diagram of the full proteasome core particle, showing the seven rotational offsets of the D_7 symmetry. The magenta bonds between the β -rings are involved only in the Half-Proteasome (HP) dimerization process in this model. That is also why there is not an inhibitor class targeting the interface, as discussed in Section 5.1. (B) Half-Proteasome (HP) diagram, note that it lacks the D_7 symmetry but retains C_7 . (C) Table of objects showing the alpha (α) and beta (β) protein representations as well as the inhibitor (I). (D) Table of bond types found within the Half-Proteasome (HP), each showing both the normal bond as found in (B) and the inhibited binding relationship, if there is one.

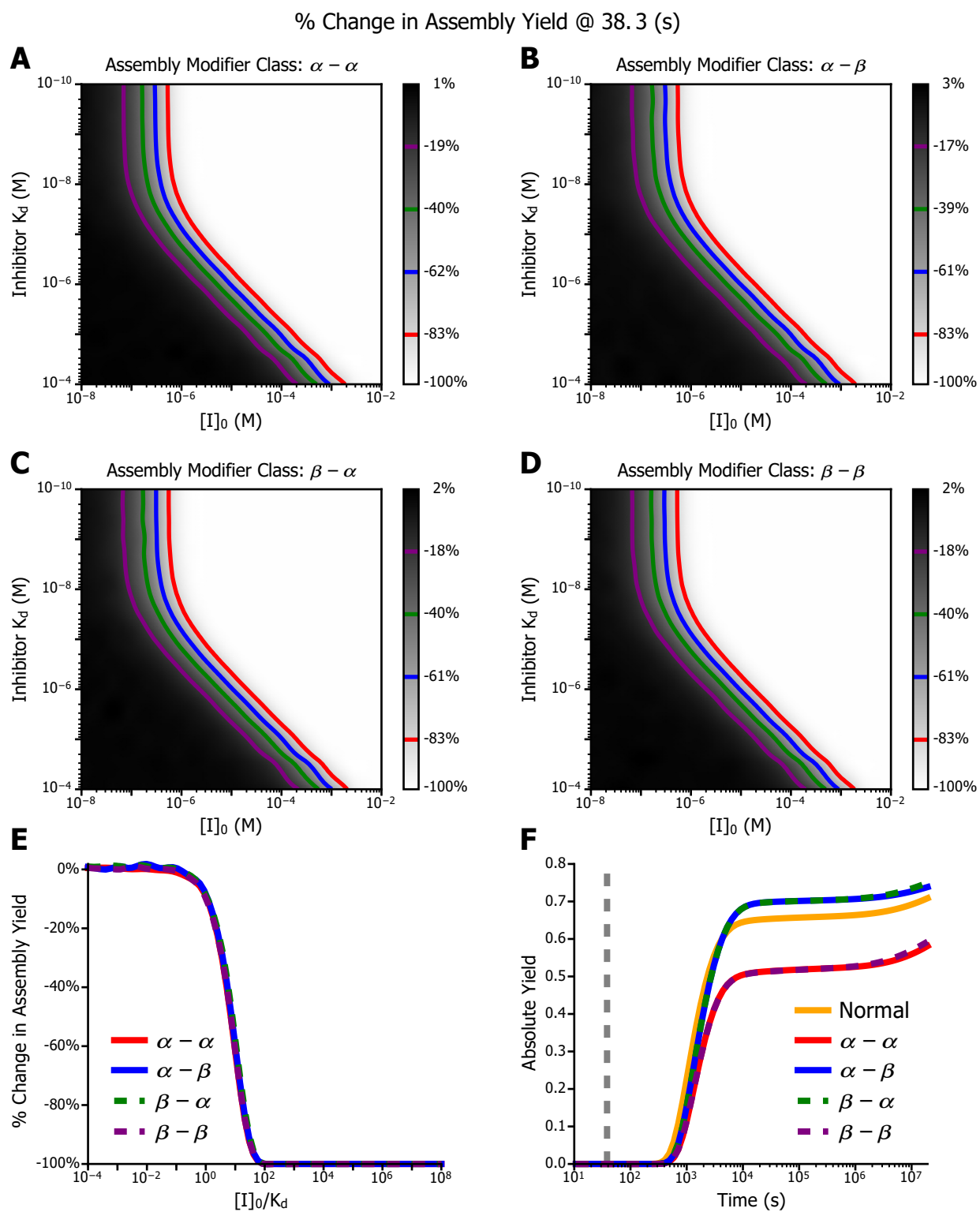


Figure 5.2: Effects of Assembly Modifiers at 38.3 (s)
Caption is on following page.

Figure 5.2: Effects of Assembly Modifiers at 38.3 (s)

All simulations are *in vitro*-like with 4×10^{-6} molar starting concentrations of both proteins, α and β . (A, B, C and D) Heatmaps with contour lines showing the percent change in assembly yield of the completed proteasome over a range of inhibitor binding strengths and inhibitor concentrations. (E) Values are taken from the same data as the heatmap with the corresponding assembly modifier class. Data is from the diagonal line starting at the lower left and going to the top right, *i.e.*, the trace of the transpose if you consider the data a matrix. This layout allows for easy comparisons between the modifier classes over a range of relevant parameters. (F) Yield of the systems vs time, note that this is the only plot in this figure that is an absolute yield as opposed to a relative change. The “Normal” line in this plot is the reference line against which all relative changes are determined. The vertical dashed gray line represents the point in time at which (A, B, C, D and E) are drawn from.

Figure 5.3: Effects of Assembly Modifiers at 1.0×10^3 (s)

All simulations are *in vitro*-like with 4×10^{-6} molar starting concentrations of both proteins, α and β . (A, B, C and D) Heatmaps with contour lines showing the percent change in assembly yield of the completed proteasome over a range of inhibitor binding strengths and inhibitor concentrations. (E) Values are taken from the same data as the heatmap with the corresponding assembly modifier class. Data is from the diagonal line starting at the lower left and going to the top right, *i.e.*, the trace of the transpose if you consider the data a matrix. This layout allows for easy comparisons between the modifier classes over a range of relevant parameters. (F) Yield of the systems vs time, note that this is the only plot in this figure that is an absolute yield as opposed to a relative change. The “Normal” line in this plot is the reference line against which all relative changes are determined. The vertical dashed gray line represents the point in time at which (A, B, C, D and E) are drawn from.

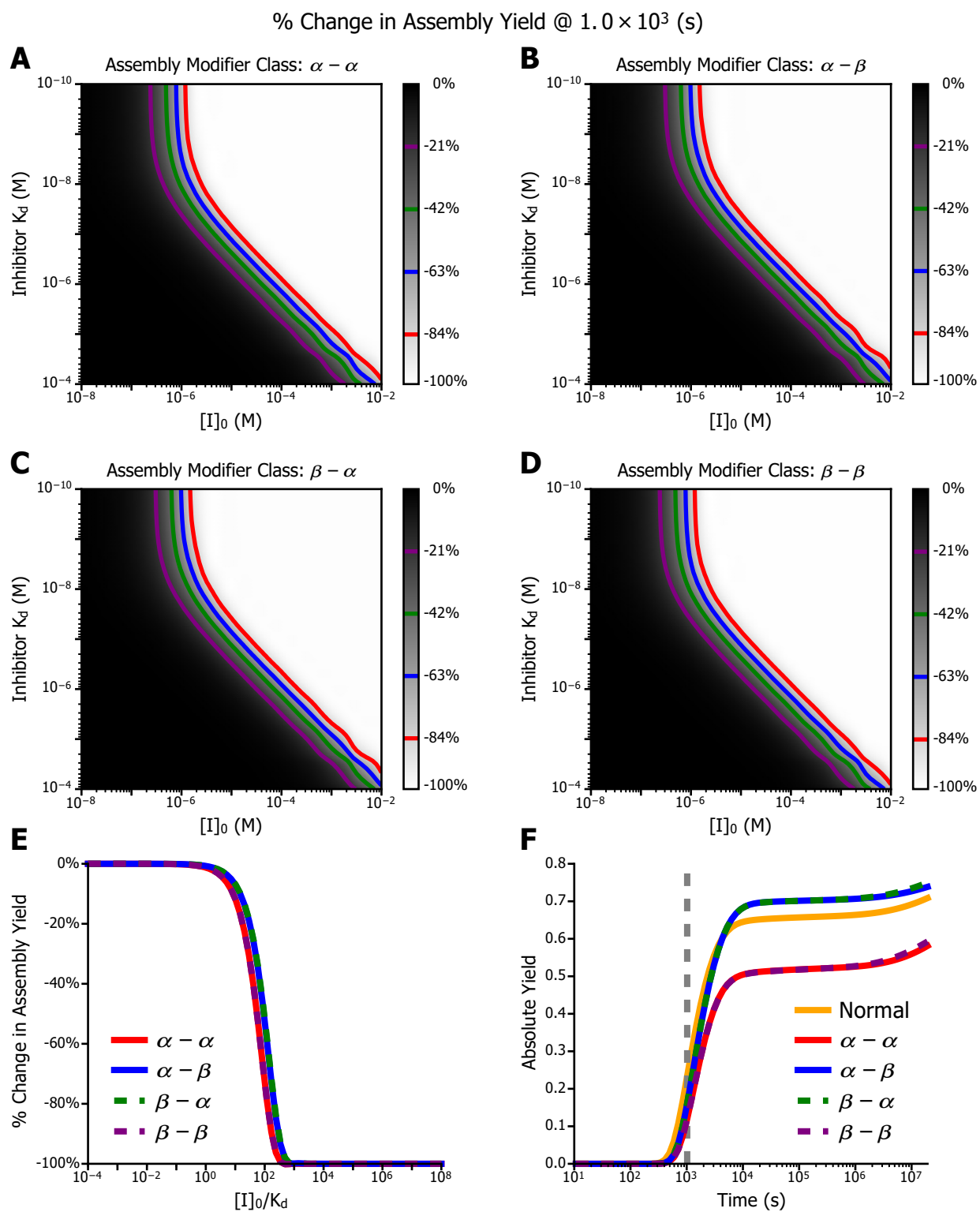


Figure 5.3: Effects of Assembly Modifiers at 1.0×10^3 (s)
Caption is on preceding page.

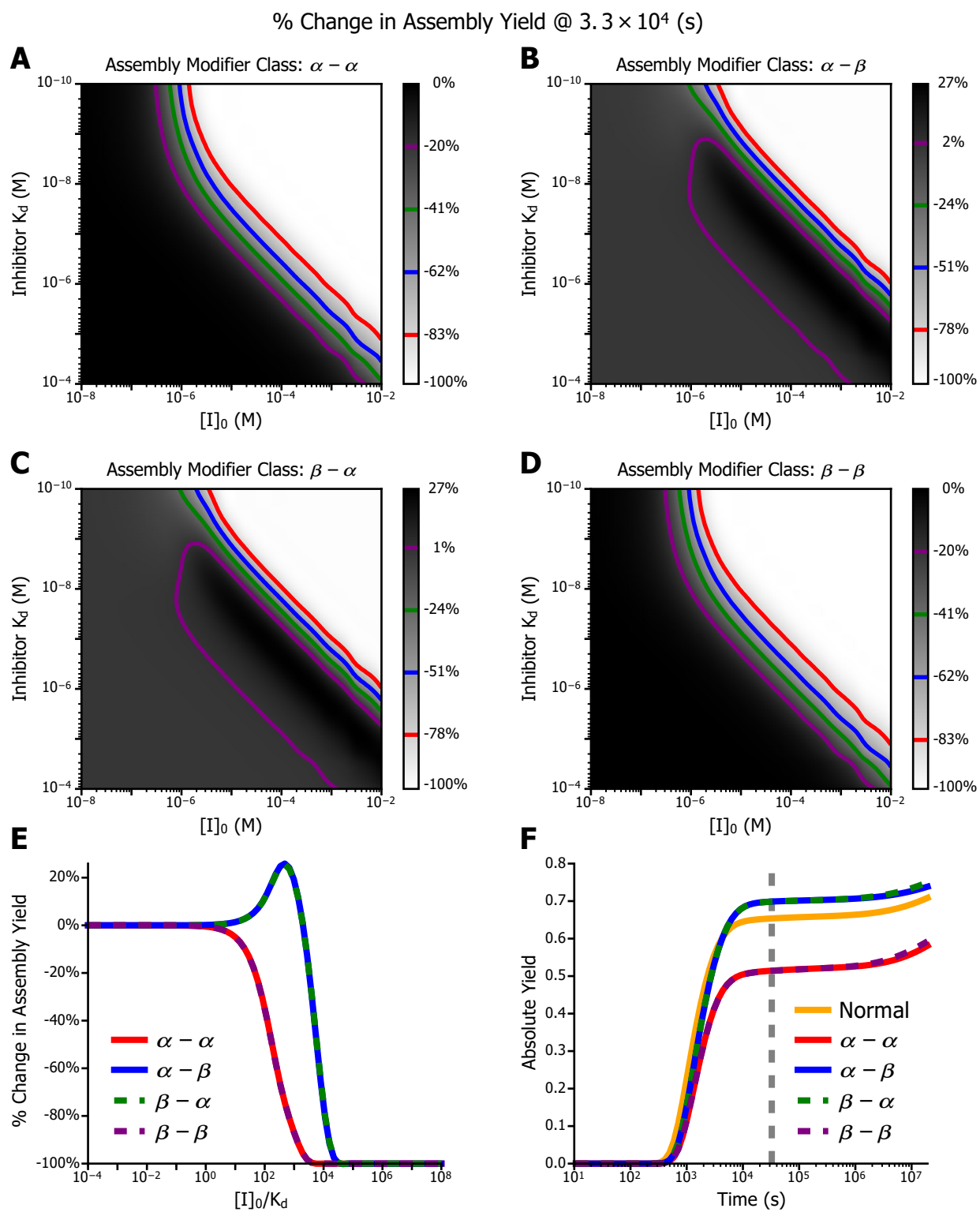


Figure 5.4: Effects of Assembly Modifiers at 3.3×10^4 (s)
Caption is on following page.

Figure 5.4: Effects of Assembly Modifiers at 3.3×10^4 (s)

All simulations are *in vitro*-like with 4×10^{-6} molar starting concentrations of both proteins, α and β . (A, B, C and D) Heatmaps with contour lines showing the percent change in assembly yield of the completed proteasome over a range of inhibitor binding strengths and inhibitor concentrations. (E) Values are taken from the same data as the heatmap with the corresponding assembly modifier class. Data is from the diagonal line starting at the lower left and going to the top right, *i.e.*, the trace of the transpose if you consider the data a matrix. This layout allows for easy comparisons between the modifier classes over a range of relevant parameters. (F) Yield of the systems vs time, note that this is the only plot in this figure that is an absolute yield as opposed to a relative change. The “Normal” line in this plot is the reference line against which all relative changes are determined. The vertical dashed gray line represents the point in time at which (A, B, C, D and E) are drawn from.

Figure 5.5: Effects of Assembly Modifiers at 1.9×10^7 (s)

All simulations are *in vitro*-like with 4×10^{-6} molar starting concentrations of both proteins, α and β . (A, B, C and D) Heatmaps with contour lines showing the percent change in assembly yield of the completed proteasome over a range of inhibitor binding strengths and inhibitor concentrations. (E) Values are taken from the same data as the heatmap with the corresponding assembly modifier class. Data is from the diagonal line starting at the lower left and going to the top right, *i.e.*, the trace of the transpose if you consider the data a matrix. This layout allows for easy comparisons between the modifier classes over a range of relevant parameters. (F) Yield of the systems vs time, note that this is the only plot in this figure that is an absolute yield as opposed to a relative change. The “Normal” line in this plot is the reference line against which all relative changes are determined. The vertical dashed gray line represents the point in time at which (A, B, C, D and E) are drawn from.

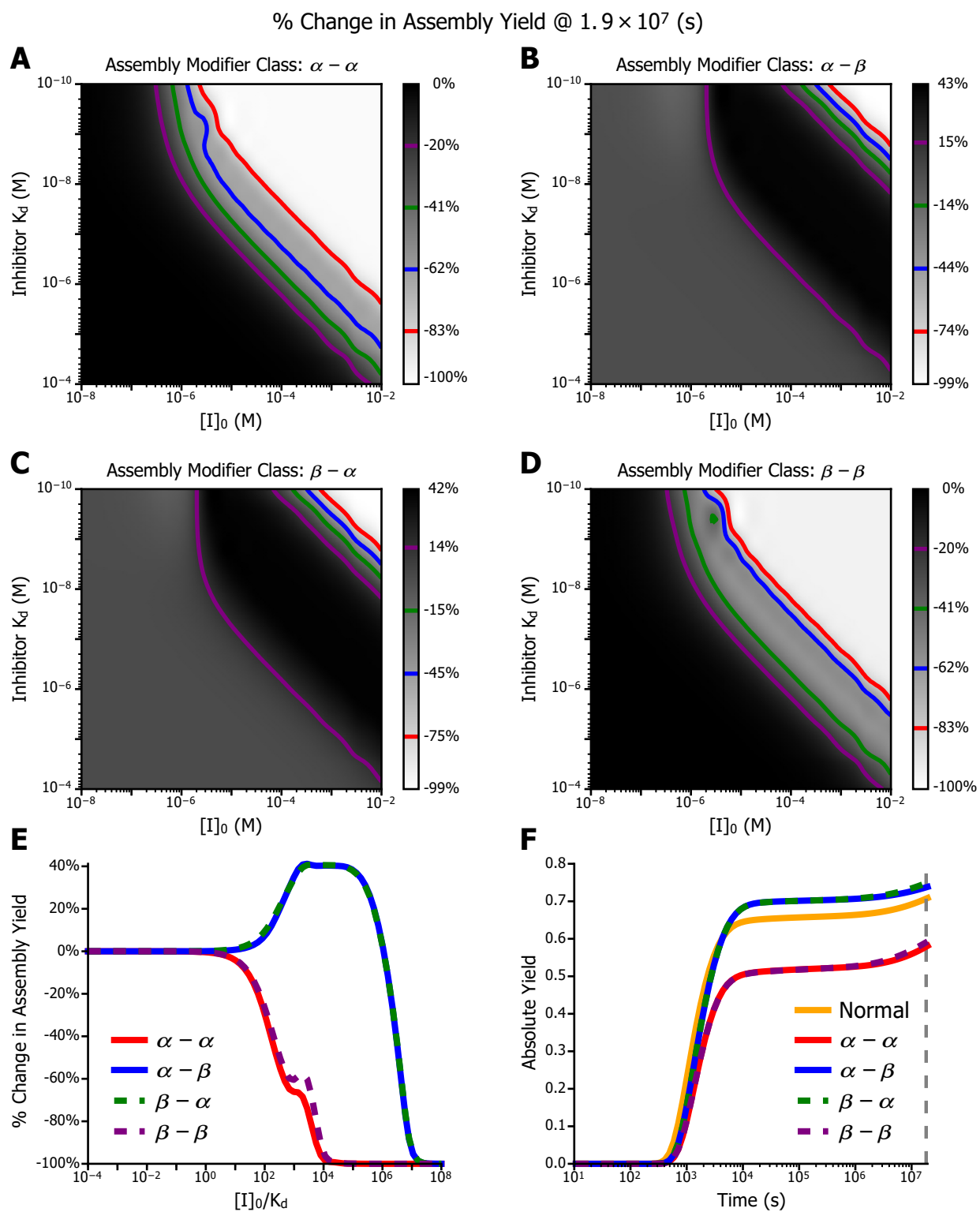


Figure 5.5: Effects of Assembly Modifiers at 1.9×10^7 (s)
Caption is on preceding page.

Conclusion

“Who knows who you are. . .

A person is a novel: you don’t know how it will end until the very last page. Otherwise, it wouldn’t be worth reading to the very end. . .”

— Yevgeny Zamyatin, *We*

Details of ISWI:DNA binding have been *elucidated*.

Numerical results of the stacked trimer have been *parsed*.

Combinations of assembly pathways have been *calculated*.

Pathways have been *disrupted*.

This document has reached its maximal structure.

The future holds many stories. One has already begun, a story of experimental analysis of proteasome assembly modifiers is in progress. Small molecules have been identified and are being tested, but that is someone else’s story to tell.

Generation of the canonical species and reactions for arbitrary protein complexes is another path into the future. Several fascinating challenges exist in the efficient determination of binding partners within an arbitrary protein complex. Personally, I find the static determination of these systems combinatorics fascinating. Enumerative combinatorics of Protein assembly can get tricky, as a composition of equivalence relations is often at play. For a single assembly pathway, the displayed pathway is representative of an equivalence class. Thus, there is a cardinality to that function to be considered. While that is fairly

straightforward, do not forget that each node is a member of another set of equivalence classes. Additionally, each reaction pair is from another set, but in that case the reactions draw their members from the species space as well. In Chapter 4, I mentioned the Method Resolution Order problem for multiple inheritance in object-oriented programming. The difficulties within that problem are not unrelated to the issue of determining the inheritance of equivalence classes in protein assembly. While writing this document I came across the field of *Combinatorial Species* [86]. The field is specially equipped to dig into the details of nested equivalences, such as those found in protein assembly.

Improvements in transmission electron cryomicroscopy (CryoEM) over the last 5-10 years have brought the possibility of direct observation of intermediate structures of the proteasome into the realm of possibility [103, 104]. Of course, challenges remain in the measurement of heterogeneous mixtures. Understanding the relative species concentrations in a protein complex assembly experiment is usually critical to optimizing data collection. One of the many ways this work may be the beginning of another story, is through its improvements in determining experimental conditions.

References

- (1) Liu, T., Sae-Ueng, U., Li, D., Lander, G. C., Zuo, X., Jönsson, B., Rau, D., Shefer, I., and Evilevitch, A. (2014). Solid-to-fluid-like DNA transition in viruses facilitates infection. *Proceedings of the National Academy of Sciences of the United States of America* 111, 14675–80.
- (2) Odijk, T. H. E. O. (2004). Statics and dynamics of condensed DNA within phages and globules.
- (3) Rothschild, L. J., and Mancinelli, R. L. (2001). Life in extreme environments. *Nature* 409, 1092–1101.
- (4) Phillip, Y., Harel, M., Khait, R., Qin, S., Zhou, H. X., and Schreiber, G. (2012). Contrasting factors on the kinetic path to protein complex formation diminish the effects of crowding agents. *Biophysical Journal* 103, 1011–1019.
- (5) Lee, B., LeDuc, P. R., and Schwartz, R. (2012). Three-Dimensional Stochastic Off-Lattice Model of Binding Chemistry in Crowded Environments. *PLoS ONE* 7, ed. by Flower, D. R., e30131.
- (6) Zhou, H.-X. (2010). Rate theories for biologists. *Quarterly reviews of biophysics* 43.
- (7) Andrews, S. S., Addy, N. J., Brent, R., and Arkin, A. P. (2010). Detailed simulations of cell biology with Smoldyn 2.1. *PLoS Computational Biology* 6.

- (8) Schuck, P. (2013). Analytical ultracentrifugation as a tool for studying protein interactions. *Biophysical Reviews* 5, 159–171.
- (9) Joo, C., Balci, H., Ishitsuka, Y., Buranachai, C., and Ha, T. (2008). Advances in Single-Molecule Fluorescence Methods for Molecular Biology. *Annual Review of Biochemistry* 77, 51–76.
- (10) Levy, E. D., and Teichmann, S. (2013). Structural, evolutionary, and assembly principles of protein oligomerization. *Progress in molecular biology and translational science* 117, 25–51.
- (11) Blomberg, C., *Physics of Life: The Physicist's Road to Biology*, 2007, p 436.
- (12) Andrews, S. S., and Bray, D. (2004). Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. *Physical Biology* 1, 137–151.
- (13) Tomita, M. (2001). Whole-cell simulation: a grand challenge of the 21st century. *Trends in Biotechnology* 19, 205–210.
- (14) Vijayakumar, M., Wong, K. Y., Schreiber, G., Fersht, A. R., Szabo, A., and Zhou, H. X. (1998). Electrostatic enhancement of diffusion-controlled protein-protein association: comparison of theory and experiment on barnase and barstar. *Journal of molecular biology* 278, 1015–24.
- (15) Northrup, S. H., and Erickson, H. P. (1992). Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proceedings of the National Academy of Sciences of the United States of America* 89, 3338–42.
- (16) Schlosshauer, M., and Baker, D. (2004). Realistic protein-protein association rates from a simple diffusional model neglecting long-range interactions, free energy barriers, and landscape ruggedness. *Protein Science* 13, 1660–1669.
- (17) Alsallaq, R., and Zhou, H.-X. (2007). Prediction of Protein-Protein Association Rates from a Transition-State Theory. *Structure* 15, 215–224.

- (18) Deeds, E. J., Bachman, J. A., and Fontana, W. (2012). Optimizing ring assembly reveals the strength of weak interactions. *Proceedings of the National Academy of Sciences* 109, 2348–2353.
- (19) Camacho, C. J., Kimura, S. R., DeLisi, C., and Vajda, S. (2000). Kinetics of desolvation-mediated protein-protein binding. *Biophysical Journal* 78, 1094–1105.
- (20) Nelson, D. L., and Cox, M. M., *Lehninger principles of biochemistry*, 6th; W.H. Freeman: New York, 2012, p 1328.
- (21) Sun, Z., Yan, Y. N., Yang, M., and Zhang, J. Z. H. (2017). Interaction entropy for protein-protein binding. *The Journal of Chemical Physics* 146, 124124.
- (22) Kastiris, P. L., and Bonvin, A. M. J. J. (2012). On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface* 10.
- (23) Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., Ji, X.-L., and Liu, S.-Q. (2016). Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. *International journal of molecular sciences* 17.
- (24) Chellaboina, V., Bhat, S. P., Haddad, W. M., and Bernstein, D. S. (2009). Modeling and Analysis of Mass-Action Kinetics: Nonnegativity, realizability, reducibility, and semistability. *IEEE Control Systems* 29, 60–78.
- (25) Wolfram Research Inc. Mathematica, Version 11.0., Champaign, IL, 2016.
- (26) Keener, J., and Sneyd, J., *Mathematical Physiology I: Cellular Physiology*; 4, 2013; Vol. 82, p 476.
- (27) Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 8, 128–140.

- (28) Al-Ani, G., Briggs, K., Malik, S. S., Conner, M., Azuma, Y., and Fischer, C. J. (2014). Quantitative determination of binding of ISWI to nucleosomes and DNA shows allosteric regulation of DNA binding by nucleotides. *Biochemistry* 53, 4334–4345.
- (29) Al-Ani, G., Malik, S. S., Eastlund, A., Briggs, K., and Fischer, C. J. (2014). ISWI remodels nucleosomes through a random walk. *Biochemistry* 53, 4346–4357.
- (30) Briggs, K., and Fischer, C. J. (2014). All motors have to decide is what to do with the DNA that is given them. *Biomolecular Concepts* 5, 383–395.
- (31) Morgan, A. M., LeGresley, S. E., Briggs, K., Al-Ani, G., and Fischer, C. J. (2018). Effects of nucleosome stability on remodeler-catalyzed repositioning. *Physical Review E* 97, 032422.
- (32) LeGresley, S. E., Briggs, K., and Fischer, C. J. (2018). Molecular motor translocation kinetics: Application of Monte Carlo computer simulations to determine microscopic kinetic parameters. *Biosystems* 168, 8–25.
- (33) Briggs, K., Al-Ani, G., Eastlund, A., and Fischer, C. J. In *Methods in molecular biology* (Clifton, N.J.) Humana Press, New York, NY: 2018; Vol. 1805, pp 333–347.
- (34) Clapier, C. R., and Cairns, B. R. (2009). The Biology of Chromatin Remodeling Complexes. *Annual Review of Biochemistry* 78, 273–304.
- (35) Demeret, C., Bocquet, S., Lemaître, J. M., Françon, P., Méchali, M., Lemaître, J. M., Françon, P., and Méchali, M. (2002). Expression of ISWI and its binding to chromatin during the cell cycle and early development. *Journal of Structural Biology* 140, 57–66.
- (36) Corona, D. F., Längst, G., Clapier, C. R., Bonte, E. J., Ferrari, S., Tamkun, J. W., and Becker, P. B. (1999). ISWI Is an ATP-Dependent Nucleosome Remodeling Factor. *Molecular Cell* 3, 239–245.
- (37) Mellor, J. (2006). Imitation switch complexes. *Ernst Schering Research Foundation Workshop*, 61–87.

- (38) Harp, J. (2000). X-ray structure of the nucleosome core particle at 2.5 Å resolution. *Section D: Biological*, 1513–1534.
- (39) Dyer, P. N., Edayathumangalam, R. S., White, C. L., Bao, Y., Chakravarthy, S., Muthurajan, U. M., and Luger, K. (2004). Reconstitution of Nucleosome Core Particles from Recombinant Histones and DNA. *Methods in Enzymology* 375, 23–44.
- (40) Pederson, T. (2000). Half a century of "the nuclear matrix". *Molecular biology of the cell* 11, 799–805.
- (41) Ramakrishnan, V. (1997). Histone Structure and the Organization of the Nucleosome. *Annual Review of Biophysics and Biomolecular Structure* 26, 83–112.
- (42) Jenuwein, T., and Allis, C. D. (2001). Translating the histone code. *Science (New York, N.Y.)* 293, 1074–80.
- (43) Cairns, B. R. (2001). Emerging roles for chromatin remodeling in cancer biology. *Trends in cell biology* 11, S15–S21.
- (44) Boltzmann, L. (1887). Über einige Fragen der Kinetische Gastheorie. *Wiener Berichte* 96, 891–918.
- (45) Lewis, G. N. (1925). A New Principle of Equilibrium. *Proceedings of the National Academy of Sciences* 11, 179–183.
- (46) Colquhoun, D., Dowdland, K. A., Beato, M., and Plested, A. J. (2004). How to impose microscopic reversibility in complex reaction mechanisms. *Biophysical Journal* 86, 3510–3518.
- (47) Rubinow, S. I., *Introduction to mathematical biology*; Dover Publications: 1975, p 386.
- (48) Lakowicz, J. R., *Principles of fluorescence spectroscopy*, 3rd; Springer: 2006, pp 1–954.
- (49) Lakowicz, J. R., Gryczynski, I., Gryczynski, Z., and Dattelbaum, J. D. (1999). Anisotropy-based sensing with reference fluorophores. *Analytical Biochemistry* 267, 397–405.

- (50) van Rossum, G. Python 2.7., 2011.
- (51) Stonebraker, M. PostgreSQL 9.1., 2011.
- (52) Kroshko, D. OpenOpt: Free scientific-engineering software for mathematical modeling and optimization., 2007.
- (53) Smart, J., Zeitlin, V., Csomor, S., Roebing, R., et al. wxWidgets., 2011.
- (54) Jones, E., Oliphant, T., Peterson, P., et al. SciPy: Open Source Scientific Tools for Python., 2001.
- (55) Pilgrim, M., *Dive Into Python*, 2004.
- (56) Oliphant, T. E., *Guide to NumPy*, 2nd, 2010; Vol. 1, p 378.
- (57) Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering* 9, 99–104.
- (58) Rappin, N., and Dunn, R., *wxPython in action*; Manning: 2006, p 552.
- (59) Arfken, G. B. (B., Weber, H.-J., and Harris, F. E., *Mathematical methods for physicists*. Academic: 2012.
- (60) Bergendahl, L. T., and Marsh, J. A. (2017). Functional determinants of protein assembly into homomeric complexes. *Scientific reports* 7, 4932.
- (61) Goodsell, D. S., and Olson, A. J. (2000). Structural Symmetry and Protein Function. *Annual Review of Biophysics and Biomolecular Structure* 29, 105–153.
- (62) Mou, Y., Huang, P.-S., Hsu, F.-C., Huang, S.-J., and Mayo, S. L. (2015). Computational design and experimental verification of a symmetric protein homodimer. *Proceedings of the National Academy of Sciences* 112, 10714–10719.
- (63) Saiz, L., and Vilar, J. M. G. (2006). Stochastic dynamics of macromolecular-assembly networks. *Molecular Systems Biology* 2, 2006.0024.

- (64) Minh, D. D. L., Bui, J. M., Chang, C. E., Jain, T., Swanson, J. M. J., and McCammon, J. A. (2005). The entropic cost of protein-protein association: A case study on acetylcholinesterase binding to fasciculin-2. *Biophysical Journal* 89, 25–27.
- (65) A. C. Hindmarsh ODEPACK, A Systematized Collection of ODE Solvers., 1983.
- (66) Petzold, L. (1983). Automatic Selection of Methods for Solving Stiff and Nonstiff Systems of Ordinary Differential Equations. *SIAM Journal on Scientific and Statistical Computing* 4, 136–148.
- (67) Hierarchical Data Format, version 5., 1997.
- (68) The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
- (69) Milon, L., Meyer, P., Chiadmi, M., Munier, A., Johansson, M., Karlsson, A., Lascu, I., Capeau, J., Janin, J., and Lacombe, M. L. (2000). The human nm23-H4 gene product is a mitochondrial nucleoside diphosphate kinase. *Journal of Biological Chemistry* 275, 14264–14272.
- (70) Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology* 13, e1002106.
- (71) Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine* 2, e124.
- (72) Motulsky, H. J. (2014). Common misconceptions about data analysis and statistics. *The Journal of pharmacology and experimental therapeutics* 351, 200–5.
- (73) Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research* 28, 235–242.
- (74) Krissinel, E., and Henrick, K. (2007). Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology* 372, 774–797.

- (75) Chen, J., Sawyer, N., and Regan, L. (2013). Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Science* 22, 510–515.
- (76) Cayley, A. (1857). On the analytical forms called trees. *American Journal of Mathematics* 13, 172–176.
- (77) Kirchhoff, G. (1847). Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird. *Annalen der Physik und Chemie* 148, 497–508.
- (78) Kirby, E. C., Mallion, R. B., Pollak, P., and Skrzyński, P. J. (2016). What Kirchhoff Actually did Concerning Spanning Trees in Electrical Networks and its Relationship to Modern Graph-Theoretical Work. *Croatica Chemica Acta* 89, 403–417.
- (79) Cayley, A. (1889). A theorem on trees. *Quart. J. Pure Appl. Math* 23, 376–378.
- (80) Biggs, N., Lloyd, E. K., and Wilson, R. J., *Graph Theory, 1736-1936*; Clarendon Press: 1986.
- (81) Murtagh, F. (1984). Counting dendrograms: A survey. *Discrete Applied Mathematics* 7, 191–199.
- (82) Whitney, H. (1932). Congruent Graphs and the Connectivity of Graphs. *American Journal of Mathematics* 54, 150.
- (83) Hamana, M. (2010). Initial Algebra Semantics for Cyclic Sharing Tree Structures. *Logical Methods in Computer Science* 6, ed. by Curien, P.-L., 1–19.
- (84) Nilsson, M., and Tanaka, H. In *Third International Conference on Logic Programming*, London, 1986, pp 593–599.
- (85) Knuth, D. E., *The art of computer programming: fundamental algorithms*, 3rd, 1997.
- (86) Bergeron, F., Labelle, G., and Leroux, P. *Combinatorial Species and Tree-like Structures.*, 1998.

- (87) Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., and Smith, K. (2011). Cython: The best of both worlds. *Computing in Science and Engineering* 13, 31–39.
- (88) Lin, G., Li, D., De Carvalho, L. P. S., Deng, H., Tao, H., Vogt, G., Wu, K., Schneider, J., Chidawanyika, T., Warren, J. D., Li, H., and Nathan, C. (2009). Inhibitors selective for mycobacterial versus human proteasomes. *Nature* 461, 621–626.
- (89) Gandotra, S., Lebron, M. B., and Ehrt, S. (2010). The Mycobacterium tuberculosis Proteasome Active Site Threonine Is Essential for Persistence Yet Dispensable for Replication and Resistance to Nitric Oxide. *PLoS Pathogens* 6, ed. by Rubin, E. J., 25–26.
- (90) Gandotra, S., Schnappinger, D., Monteleone, M., Hillen, W., and Ehrt, S. (2007). In vivo gene silencing identifies the Mycobacterium tuberculosis proteasome as essential for the bacteria to persist in mice. *Nature medicine* 13, 1515–20.
- (91) Lin, G., Li, D., Chidawanyika, T., Nathan, C., and Li, H. (2010). Fellutamide B is a potent inhibitor of the Mycobacterium tuberculosis proteasome. *Archives of Biochemistry and Biophysics* 501, 214–220.
- (92) Lin, G., Tsu, C., Dick, L., Zhou, X. K., and Nathan, C. (2008). Distinct specificities of Mycobacterium tuberculosis and mammalian proteasomes for N-acetyl tripeptide substrates. *Journal of Biological Chemistry* 283, 34423–34431.
- (93) Hu, G., Lin, G., Wang, M., Dick, L., Xu, R.-M., Nathan, C., and Li, H. (2006). Structure of the Mycobacterium tuberculosis proteasome and mechanism of inhibition by a peptidyl boronate. *Molecular microbiology* 59, 1417–28.
- (94) Dahlmann, B. (2007). Role of proteasomes in disease. *BMC Biochemistry* 8, S3.
- (95) Powell, S. R. (2006). The ubiquitin-proteasome system in cardiac physiology and pathology. *American Journal of Physiology-Heart and Circulatory Physiology* 291, H1–H19.

- (96) Donohue, T. M., Cederbaum, A. I., French, S. W., Barve, S., Gao, B., and Osna, N. A. (2007). Role of the Proteasome in Ethanol-Induced Liver Pathology. *Alcoholism: Clinical and Experimental Research* 31, 1446–1459.
- (97) Wang, J., and Maldonado, M. A. (2006). The ubiquitin-proteasome system and its role in inflammatory and autoimmune diseases. *Cellular & molecular immunology* 3, 255–61.
- (98) Hegde, A. N., and Upadhyaya, S. C. (2007). The ubiquitin–proteasome pathway in health and disease of the nervous system. *Trends in Neurosciences* 30, 587–595.
- (99) Schmidt, M., and Finley, D. (2014). Regulation of proteasome activity in health and disease. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1843, 13–25.
- (100) Chatterjee, D., Bhatt, M., Butler, D., De Genst, E., Dobson, C. M., Messer, A., and Kordower, J. H. (2018). Proteasome-targeted nanobodies alleviate pathology and functional decline in an α -synuclein-based Parkinson’s disease model. *npj Parkinson’s Disease* 4, 25.
- (101) Anvar, S., ’t Hoen, P. A., Venema, A., van der Sluijs, B., van Engelen, B., Snoeck, M., Vissing, J., Trollet, C., Dickson, G., Chartier, A., Simonelig, M., van Ommen, G.-J. B., van der Maarel, S. M., and Raz, V. (2011). Deregulation of the ubiquitin-proteasome system is the predominant molecular pathology in OPMD animal models and patients. *Skeletal Muscle* 1, 15.
- (102) Kunjappu, M. J., and Hochstrasser, M. (2014). Assembly of the 20S proteasome. *Biochimica et Biophysica Acta - Molecular Cell Research* 1843, 2–12.
- (103) Campbell, M. G., Veessler, D., Cheng, A., Potter, C. S., and Carragher, B. (2015). 2.8 Å resolution reconstruction of the *Thermoplasma acidophilum* 20S proteasome using cryo-electron microscopy. *eLife* 4.

- (104) Lander, G. C., Estrin, E., Matyskiela, M. E., Bashore, C., Nogales, E., and Martin, A. (2012). Complete subunit architecture of the proteasome regulatory particle. *Nature* 482, 186–191.
- (105) Voit, E. O. (2013). Biochemical Systems Theory: A Review. *ISRN Biomathematics*.
- (106) Feinberg, M. Lectures on chemical reaction networks., 1979.
- (107) Wegscheider, R. (1901). Über simultane Gleichgewichte und die Beziehungen zwischen Thermodynamik und Reaktionskinetik homogener Systeme. *Monatshefte für Chemie* 22, 849–906.
- (108) Ho, T. C. (2008). Kinetic Modeling of Large-Scale Reaction Systems. *Catalysis Reviews* 50, 287–378.
- (109) Hellman, L. M., and Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature Protocols*.
- (110) Olins, D. E., and Olins, A. L. (2003). Chromatin history: Our view from the bridge. *Nature Reviews Molecular Cell Biology* 4, 809–814.
- (111) Cera, E. D., *Thermodynamic Theory of Site-Specific Binding Processes in Biological Macromolecules*; Cambridge University Press: Cambridge, 1995.
- (112) Rosen, K., *Discrete mathematics and its applications*, 2012.
- (113) Bollobás, B., *Graph Theory*; Graduate Texts in Mathematics, 1979.
- (114) Bollobás, B., *Modern Graph Theory*; Graduate Texts in Mathematics, 1998.

Supplementary Reading

Biochemical Systems Theory (BST)

Voit, E. O. (2013). Biochemical Systems Theory: A Review. *ISRN Biomathematics*
A recent overview of the state of BST. Its reference list is quite comprehensive. BST models are usually more general than the models in this work, but many concepts are shared and useful for understanding why the models I have presented are structured the way they are.

Biological Chemical Reactions

Zhou, H.-X. (2010). Rate theories for biologists. *Quarterly reviews of biophysics* 43

Chemical Reaction Networks (CRN)

Feinberg, M. Lectures on chemical reaction networks., 1979

CRN Theories have their roots in early studies into complex chemical reactions [107]. The formal mathematics are well presented in Feinberg's lectures. Methodologies and generalizations have continued since spreading into highly diverse fields. One should consider looking into seemingly unrelated fields for similarities to a biological project. Just one of many examples, Petroleum Engineering has encountered and found solutions to many of the same issues as biochemical systems [108].

Experimental EMSA Protocols

Hellman, L. M., and Fried, M. G. (2007). Electrophoretic mobility shift assay

(EMSA) for detecting protein-nucleic acid interactions. *Nature Protocols*

Fluorescence & Fluorescence Anisotropy

Lakowicz, J. R., *Principles of fluorescence spectroscopy*, 3rd; Springer: 2006, pp 1–954

Nucleosomes and Chromatin

Olins, D. E., and Olins, A. L. (2003). Chromatin history: Our view from the bridge. *Nature Reviews Molecular Cell Biology* 4, 809–814

Proteasomes

Kunjappu, M. J., and Hochstrasser, M. (2014). Assembly of the 20S proteasome. *Biochimica et Biophysica Acta - Molecular Cell Research* 1843, 2–12

Protein Symmetry

Goodsell, D. S., and Olson, A. J. (2000). Structural Symmetry and Protein Function. *Annual Review of Biophysics and Biomolecular Structure* 29, 105–153

Thermodynamics of Binding Site Interactions

Cera, E. D., *Thermodynamic Theory of Site-Specific Binding Processes in Biological Macromolecules*; Cambridge University Press: Cambridge, 1995

Trees (Data Structures)

Rosen, K., *Discrete mathematics and its applications*, 2012

A general introduction to Computer Science’s mathematical approaches and basic structures. Chapter 11 (Trees) covers tree traversal.

Knuth, D. E., *The art of computer programming: fundamental algorithms*, 3rd, 1997

A classic for understanding how computers represent and manipulate informational constructs.

Trees (Graph Theory)

Bollobás, B., *Graph Theory*; Graduate Texts in Mathematics, 1979

Bollobás, B., *Modern Graph Theory*; Graduate Texts in Mathematics, 1998

Stacked Trimer ODEs

$$k_{i,j}^{\text{eff}} = \alpha k_{\text{on}} K_{d,1}^i K_{d,2}^j e^{-(i+j-1)\Delta G_p^0/RT}$$

$$\alpha = c_0^{-i-j+1}$$

$$k_{\text{on}} = 10^6 \text{ M}^{-1}\text{s}^{-1}$$

$$\begin{aligned} \frac{d[S_1]}{dt} = & k_{\text{on}}(-6[S_1]^2 - 4[S_1][S_2] - 3[S_1][S_3] - 3[S_1][S_4] - 2[S_1][S_5] - 3[S_1][S_6] - 2[S_1][S_7] \\ & - 2[S_1][S_8] - 3[S_1][S_9] - 2[S_1][S_{10}] - [S_1][S_{11}]) + 2[S_2]k_{0,1}^{\text{eff}} + 2[S_3]k_{1,0}^{\text{eff}} \\ & + [S_4]k_{1,0}^{\text{eff}} + [S_4]k_{0,1}^{\text{eff}} + 2[S_5]k_{1,0}^{\text{eff}} + [S_6]k_{1,0}^{\text{eff}} + [S_6]k_{0,1}^{\text{eff}} + 2[S_7]k_{1,0}^{\text{eff}} + 4[S_8]k_{1,1}^{\text{eff}} \\ & + 3[S_9]k_{2,0}^{\text{eff}} + 2[S_{10}]k_{2,0}^{\text{eff}} + [S_{10}]k_{0,1}^{\text{eff}} + 2[S_{11}]k_{2,1}^{\text{eff}} + [S_{11}]k_{2,0}^{\text{eff}} + 2[S_{11}]k_{1,1}^{\text{eff}} \\ & + 6[S_{12}]k_{2,1}^{\text{eff}} - \delta[S_1] + Q \end{aligned}$$

$$\begin{aligned} \frac{d[S_2]}{dt} = & k_{\text{on}}([S_1]^2 - 4[S_1][S_2] - 8[S_2]^2 - 2[S_2][S_3] - 2[S_2][S_4] - 2[S_2][S_6] - 2[S_2][S_8]) \\ & - [S_2]k_{0,1}^{\text{eff}} + [S_4]k_{1,0}^{\text{eff}} + [S_6]k_{1,0}^{\text{eff}} + 2[S_8]k_{2,0}^{\text{eff}} + [S_{10}]k_{2,0}^{\text{eff}} + 2[S_{11}]k_{3,0}^{\text{eff}} \\ & + 3[S_{12}]k_{4,0}^{\text{eff}} - \delta[S_2] \end{aligned}$$

$$\begin{aligned} \frac{d[S_3]}{dt} = & k_{\text{on}}(2[S_1]^2 - 3[S_1][S_3] - 2[S_2][S_3] - 6[S_3]^2 - [S_3][S_4] - [S_3][S_6] - 3[S_3][S_9] \\ & - [S_3][S_{10}]) - [S_3]k_{1,0}^{\text{eff}} + [S_4]k_{0,1}^{\text{eff}} + 2[S_5]k_{0,1}^{\text{eff}} + [S_6]k_{0,1}^{\text{eff}} + 2[S_7]k_{0,1}^{\text{eff}} \\ & + 2[S_8]k_{0,2}^{\text{eff}} + 3[S_9]k_{2,0}^{\text{eff}} + [S_{10}]k_{2,0}^{\text{eff}} + 2[S_{11}]k_{2,1}^{\text{eff}} + [S_{11}]k_{0,2}^{\text{eff}} + 6[S_{12}]k_{2,2}^{\text{eff}} - \delta[S_3] \end{aligned}$$

$$\begin{aligned}\frac{d[S_4]}{dt} = & k_{\text{on}}(2[S_1][S_2] + [S_1][S_3] - 3[S_1][S_4] - 2[S_2][S_4] - [S_3][S_4] - 2[S_4]^2) - [S_4]k_{1,0}^{\text{eff}} \\ & - [S_4]k_{0,1}^{\text{eff}} + 2[S_5]k_{1,0}^{\text{eff}} + 2[S_8]k_{1,1}^{\text{eff}} + [S_{10}]k_{2,0}^{\text{eff}} + [S_{11}]k_{3,0}^{\text{eff}} + [S_{11}]k_{2,1}^{\text{eff}} \\ & + 6[S_{12}]k_{4,1}^{\text{eff}} - \delta[S_4]\end{aligned}$$

$$\frac{d[S_5]}{dt} = k_{\text{on}}([S_1][S_4] - 2[S_1][S_5] + [S_3]^2) - 2[S_5]k_{1,0}^{\text{eff}} - [S_5]k_{0,1}^{\text{eff}} + [S_{11}]k_{2,1}^{\text{eff}} - \delta[S_5]$$

$$\begin{aligned}\frac{d[S_6]}{dt} = & k_{\text{on}}(2[S_1][S_2] + [S_1][S_3] - 3[S_1][S_6] - 2[S_2][S_6] - [S_3][S_6] - 2[S_6]^2) - [S_6]k_{1,0}^{\text{eff}} \\ & - [S_6]k_{0,1}^{\text{eff}} + 2[S_7]k_{1,0}^{\text{eff}} + 2[S_8]k_{1,1}^{\text{eff}} + [S_{10}]k_{2,0}^{\text{eff}} + [S_{11}]k_{3,0}^{\text{eff}} + [S_{11}]k_{2,1}^{\text{eff}} \\ & + 6[S_{12}]k_{4,1}^{\text{eff}} - \delta[S_6]\end{aligned}$$

$$\frac{d[S_7]}{dt} = k_{\text{on}}([S_1][S_6] - 2[S_1][S_7] + [S_3]^2) - 2[S_7]k_{1,0}^{\text{eff}} - [S_7]k_{0,1}^{\text{eff}} + [S_{11}]k_{2,1}^{\text{eff}} - \delta[S_7]$$

$$\begin{aligned}\frac{d[S_8]}{dt} = & k_{\text{on}}([S_1][S_4] + [S_1][S_6] - 2[S_1][S_8] + 4[S_2]^2 - 2[S_2][S_8] + [S_3]^2) - 4[S_8]k_{1,1}^{\text{eff}} \\ & - [S_8]k_{2,0}^{\text{eff}} - [S_8]k_{0,2}^{\text{eff}} + [S_{11}]k_{2,0}^{\text{eff}} + 3[S_{12}]k_{4,0}^{\text{eff}} - \delta[S_8]\end{aligned}$$

$$\begin{aligned}\frac{d[S_9]}{dt} = & k_{\text{on}}([S_1][S_3] - 3[S_1][S_9] - 3[S_3][S_9] - 6[S_9]^2) - 3[S_9]k_{2,0}^{\text{eff}} + [S_{10}]k_{0,1}^{\text{eff}} \\ & + [S_{11}]k_{0,2}^{\text{eff}} + 2[S_{12}]k_{0,3}^{\text{eff}} - \delta[S_9]\end{aligned}$$

$$\begin{aligned}\frac{d[S_{10}]}{dt} = & k_{\text{on}}([S_1][S_4] + [S_1][S_6] + 3[S_1][S_9] - 2[S_1][S_{10}] + 2[S_2][S_3] - [S_3][S_{10}]) \\ & - 3[S_{10}]k_{2,0}^{\text{eff}} - [S_{10}]k_{0,1}^{\text{eff}} + 2[S_{11}]k_{1,1}^{\text{eff}} + 6[S_{12}]k_{2,2}^{\text{eff}} - \delta[S_{10}]\end{aligned}$$

$$\begin{aligned}\frac{d[S_{11}]}{dt} = & k_{\text{on}}(2[S_1][S_5] + 2[S_1][S_7] + 2[S_1][S_8] + 2[S_1][S_{10}] - [S_1][S_{11}] + 2[S_2][S_4] \\ & + 2[S_2][S_6] + [S_3][S_4] + [S_3][S_6] + 3[S_3][S_9]) - 4[S_{11}]k_{2,1}^{\text{eff}} - [S_{11}]k_{2,0}^{\text{eff}} \\ & - 2[S_{11}]k_{1,1}^{\text{eff}} - 2[S_{11}]k_{3,0}^{\text{eff}} - [S_{11}]k_{0,2}^{\text{eff}} + 6[S_{12}]k_{2,1}^{\text{eff}} - \delta[S_{11}]\end{aligned}$$

$$\begin{aligned}\frac{d[S_{12}]}{dt} = & k_{\text{on}}([S_1][S_{11}] + 2[S_2][S_8] + [S_3][S_{10}] + [S_4]^2 + [S_6]^2 + 3[S_9]^2) - 6[S_{12}]k_{2,1}^{\text{eff}} \\ & - 3[S_{12}]k_{4,0}^{\text{eff}} - 6[S_{12}]k_{2,2}^{\text{eff}} - 6[S_{12}]k_{4,1}^{\text{eff}} - [S_{12}]k_{0,3}^{\text{eff}} - \delta[S_{12}]\end{aligned}$$