Thermodynamic consistency of autocatalytic cycles

Thomas Kosc,[1] Denis Kuperberg,[2] Etienne Rajon,[1] and Sylvain Charlat[1,*]

[1] *LBBE (Laboratoire de Biométrie & Biologie Evolutive) Université Lyon 1, CNRS, Villeurbanne, France*
[2] *LIP(Laboratoire de l'Informatique du Parallélisme), ENS Lyon, CNRS, France*

Autocatalysis is seen as a potential key player in the origin of life, and perhaps more generally in the emergence of Darwinian dynamics. Building on recent formalizations of this phenomenon, we tackle the computational challenge of exhaustively detecting minimal autocatalytic cycles (autocatalytic cores) in reaction networks, and further evaluate the impact of thermodynamic constraints on their realization under mass action kinetics. We first characterize the complexity of the detection problem by proving its NP-completeness. This justifies the use of constraint solvers to list all cores in a given reaction network, and also to group them into compatible sets, composed of cores whose stoichiometric requirements are not contradictory. Crucially, we show that the introduction of thermodynamic realism does constrain the composition of these sets. Compatibility relationships among autocatalytic cores can indeed be disrupted when the reaction kinetics obey thermodynamic consistency throughout the network. On the contrary, these constraints have no impact on the realizability of isolated cores, unless upper or lower bounds are imposed on the concentrations of the reactants. Overall, by better characterizing the conditions of autocatalysis in complex reaction systems, this work brings us a step closer to assessing the contribution of this collective chemical behavior to the emergence of natural selection in the primordial soup.

Significance Statement: Describing the processes behind the origin of life requires us to better understand self-amplifying dynamics in complex chemical systems. Detecting autocatalytic cycles is a critical but challenging step in this endeavor. After characterizing the computational complexity of this problem, we investigate the impact of thermodynamic realism on autocatalysis. We demonstrate that individual cycles, regardless of thermodynamic parameters, can always be activated as long as entities may occur at any required concentration. In contrast, two cycles can become mutually incompatible due to thermodynamic constraints, and will thus never run simultaneously. These results clarify the implications of physical realism for the realization of autocatalysis.

I. INTRODUCTION

It is increasingly recognized that producing a consistent explanation for the origination of life will require us to explain how Darwinian evolution may have *gradually* emerged from a non-biological, purely physical world [1–9]. Gradually rather than suddenly, that is, without assuming that natural selection only came into play once chance alone had produced the first obvious "replicators", displaying the same heritable variance as current organisms. Under this perspective of a smooth transition from physics to biology, natural selection is hypothesized to have been active already in the "prebiotic" soup as a driver to complexity; yet in a rudimentary and currently unrecognizable fashion.

To explore this path, autocatalysis is often taken as a plausible starting point (Box 1) [6, 10–14]. Here, more specifically, we envision autocatalytic cycles as the putative elementary components of higher level systems that may engage in "increasingly Darwinian" dynamics. In doing so, we aim at keeping the best of the two traditionally opposed approaches to the origin of life: physico-chemical realism of the metabolism-first view, and evolvability of the gene-first perspective. Beyond the specifics of terrestrial life, progressing toward an articulation of Darwinian principles with physics appears to be a prerequisite to assessing their putative relevance to other physical systems [2, 3].

We build on recent theoretical and computational developments [10, 15, 16] to systematically search for minimal autocatalytic cycles (denoted autocatalytic cores, *sensu* Blokhuis et al [15]) in reaction networks and then assess their thermodynamic consistency, i.e. the impact of thermodynamic constraints on their realization, under mass action kinetics. We first prove that finding autocatalytic cores in the network is an NP-complete problem – a question that was left open in earlier work [17, 18] – and converge with other authors in using constraint solvers as a technical solution [10, 16]. We then question whether such autocatalytic cores, defined on the sole basis of the reaction network topology, can also be realized once thermodynamic constraints are introduced. To do so, we take into account the reaction kinetics that themselves depend on the Gibbs free energies and concentrations of the reactants, and the activation barriers of the reactions. We show that regardless of these physical quantities, any potential autocatalytic core may be instantiated in some region of the concentration space as long as this space is assumed unbounded. In contrast, thermodynamic constraints do restrain compatibility relationships between autocatalytic cores and will thereby impact the dynamics in complex chemical networks.

Box 1: Related work on autocatalysis

The present study takes place within a flourishing body of literature taking autocatalysis as a plausible primary component of proto-biotic or proto-Darwinian systems. Our model contrasts with those based on the RAF framework [19, 20] in that it follows a bottom-up approach to autocatalysis: rather than setting catalytic relationships between components of the system and randomly picked reactions, we let the reaction network generate (or not) these relationships, as formalized by Blokhuis et al [15]. Catalysis and autocatalysis then simply emerge in the reaction network as pathways involving

* sylvain.charlat@univ-lyon1.fr

entities that act both as reactants and products. For example, in the reactions A+C→AC, AC+B→ABC, ABC→AB + C, entity C can be simply described as a catalyst of the reaction A+B→AB.

In taking such a bottom-up angle, our framework is much related to that of several recent studies [10–12, 16, 21, 22]. Some of these have considered the implications of thermodynamic constraints and mass action kinetics on specific autocatalytic motifs [11, 12, 21]. Others have implemented tools for the exhaustive detection of autocatalysis [10, 16, 22]. Here we jointly consider these two components of the problem, i.e. exhaustive detection and thermodynamic realism.

On a more conceptual ground, we share with Baum et al [6] the view that collections of autocatalytic cycles, rather than cycles alone, might constitute the scale at which incipient heritable variations may occur.

II. FRAMEWORK AND DEFINITIONS

We analyze networks of reversible reactions governed by mass action kinetics. This typically applies to reactions that simply consist in the association of two entities and the reciprocal dissociation (e.g. $A + B \rightleftharpoons AB$). The entities are fully defined by their composition (e.g. A_2B_2 is not distinct from B_2A_2). Given a list of entities, this rule sets the list of all possible reactions, only some of which are assumed to exist to generate a particular reaction network – this is equivalent to assuming that some reactions have an infinite activation barrier and thus have a null flow.

We can then apply the formalism of Blokhuis et al [15] to identify autocatalytic motifs in such reaction networks. Intuitively, these can be conceived as cyclic subnetworks admitting a regime where each entity has a positive net production rate.

Definition 1. *An autocatalytic motif is defined as a set of entities E_C and reactions R_C such that:*

- *For each reaction $R \in R_C$, at least one entity on each side is in E_C.*
- *There exists a vector $\vec{v}$ of flows for reactions from R_C defining a regime where the total contribution of reactions from R_C is strictly positive for each entity of E_C.*

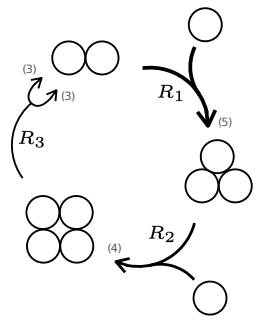
Consider for instance a reaction $A + B \rightleftharpoons AB$, with reactants A, B and product AB. Then the stoichiometry σ_A^R of A in R is -1 (or -2 if $A = B$), and σ_{AB}^R is 1. If an entity e does not appear in a reaction R, we set $\sigma_e^R = 0$.

We will note v_R the flow of the reaction at a given instant, that will be positive if the association rate is larger than the dissociation rate.

Given a motif candidate C formed of entities E_C and reactions R_C, and an entity e in E_C, we define the variation of e's concentration **due to** C as:

$$\Delta_C(e) = \sum_{R \in R_C} \sigma_e^R \cdot v_R \tag{1}$$

We define a *witness* as a choice of v_R for each $R \in R_C$, such that for each $e \in E_C$ we have $\Delta_C(e) > 0$. This can be formalized using linear algebra, following Blokhuis et al [15].



M matrix:

$$\begin{array}{c} \\ A_2 \\ A_3 \\ A_4 \end{array} \begin{array}{c} \begin{array}{ccc} R_1 & R_2 & R_3 \end{array} \\ \begin{pmatrix} -1 & 0 & 2 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \end{array}$$

Figure 1. Schematic view of a formose-like autocatalytic motif. Entities A_2, A_3 and A_4 are part of the motif while entity A_1 serves as food. The arrows indicate the net direction of each reaction, while the line width indicates their respective flows, that must be decreasing from reactions R_1 to R_3 for the cycle to run. As an example, the flows values (indicated in brackets) would produce a net increase of 1 of each entity. The right panel shows the corresponding stoichiometric matrix M. Given the represented flow vector $\vec{v} = (5, 4, 3)$, we obtain $M \cdot \vec{v} = (1, 1, 1)$, showing that $\vec{v}$ is indeed a witness.

Indeed, if M is the stoichiometric matrix restricted to E_C and R_C, then the candidate C is an autocatalytic motif if and only if there exists a witness vector $\vec{v}$ such that all coordinates of $M \cdot \vec{v}$ are strictly positive.

Example 1. *We illustrate the definition in Figure 1 with a simple formose-like cycle comprising three entities A_2, A_3 and A_4 and using entity A_1 as food, with detailed explanations provided in the caption.*

Entities appearing in R_C that are not part of E_C will be called either "food" if they are consumed or "waste" if they are produced by a reaction of R_C, taking into account the sign of the witness vector that indicates the direction of reactions. Notice that an entity may simultaneously appear as food and waste in an autocatalytic motif.

Formally, a motif is said to *contain* another one if it includes all its entities and reactions. In contrast, a motif is *minimal* if it does not contain any other, in which case it corresponds to the "autocatalytic core" from Blokhuis et al [15] and to the stoichiometric autocatalysis of Gagrani et al [16]. The present study focuses on such minimal motifs, that we denote *Potential* Autocatalytic Cores (PACs), to emphasize that they are defined on the sole basis of the reaction network topology, so that their realizability under thermodynamic constraints remains to be assessed.

Definition 2. *A Potential Autocatalytic Core (PAC) is defined as a minimal autocatalytic motif, whose identification relies solely on the reaction network topology as formalized by the stoichiometric matrix.*

Example 2. *The autocatalytic motif shown in Figure 1 is minimal, and hence constitutes a PAC.*

Notice that minimality implies that a PAC witness cannot contain null reaction flows: if this were the case, the corresponding reaction could be removed without altering autocatalysis, which contradicts the minimality criterion. Furthermore, it is shown in Blokhuis et al [15] that in a PAC, each

entity is the reactant of a unique reaction, and each reaction has a unique entity of the PAC as reactant (other reactants being food). This implies that the direction of reactions is consistent across all witnesses of a given PAC: flipping the direction of one reaction would impose flipping all the others. Therefore, each reaction of the PAC has a unique possible net direction that will be compatible with all its witnesses flow vectors. In Appendix 1, we provide a formal proof of this property that was put forward in Blokhuis et al [15, SI].

III. DETECTING POTENTIAL AUTOCATALYTIC CORES

Our goal is to enumerate all PACs in a reaction network. To this end, we first assess the complexity of this problem, in order to determine which computational tools are required for its resolution.

A. NP-completeness proof

Enumerating all PACs in a reaction network involves sequentially solving problems of the type: "Is there a PAC in the system besides those previously found?". We will show that a particular variant of this question is NP-complete. Namely, we will prove the NP-completeness of deciding whether a PAC exists that contains an entity A and takes food from a given subset F.

In this section, for simplicity, we will relax any compositionality constraint on reactions so that letters like $A, B, E \ldots$ will be shorthand for any kind of entity. Yet it would be straightforward (but less readable) to extend the construction to a strictly compositional framework.

Notably, the complexity of the autocatalysis detection problem has previously been considered by Andersen et al [17] but from a different angle. These authors have specifically shown that the following problem is NP-complete: considering a reaction network that contains a known autocatalytic core, can its resources be produced by the network? The difficulty of finding all autocatalytic cores in a reaction network, that we tackle here, has thus not yet been addressed.

In the framework of Blokhuis et al [15], it is easy to check whether a proposed set of entities and reactions constitutes an autocatalytic core. Indeed, thanks to the linear algebra formulation summarized in Section II, this problem is solved in polynomial time by Linear Programming. As will be shown, the difficulty rather lies in finding an autocatalytic core in a reaction network, among exponentially many possible candidates.

Formally, let PAC-DETECTION be the following algorithmic problem:

Definition 3 (PAC-DETECTION problem)**.**
INPUT*: A reaction system defined by entities* (E) *and reactions* (R)*, a target* $A \in E$*, and a set of allowed foods* $F \subset E$*.*
OUTPUT*: Is there a PAC containing* A *and using only foods from* F*?*

Theorem 1. *PAC-DETECTION is NP-complete.*

We provide here a brief description of the proof that is fully described in Appendix 2. Because a PAC candidate can be tested in polynomial time, PAC-DETECTION is in NP. It remains to be shown that it is NP-hard. To this end, we reduce from the well-known NP-complete problem SAT [23]. An instance of SAT asks whether an input formula on n boolean variables $x_1, \ldots, x_n$ is satisfiable, via a suitable assignation of variables with true/false values.

To perform the reduction, we associate each such formula φ with a reaction system S_φ, of size polynomial in φ, with a specified target entity A and a food set F. Reactions in S_φ are designed to mirror the structure of φ, ensuring that a PAC of the wanted form exists if and only if the formula is satisfiable. The only possible such PACs will actually directly encode satisfying assignments for φ.

This shows that PAC-DETECTION is NP-hard: a polynomial-time algorithm for PAC-DETECTION would yield a polynomial-time algorithm for SAT, via this reduction. We can conclude that PAC-DETECTION is NP-complete, since it is also in NP. Notice that our proof assumes a particular shape for constraints on the searched PAC (containing an entity A and using only foods from a set F). Assessing whether an unconstrained variant is NP-complete as well remains an open problem.

B. Implementation

The above NP-completeness result justifies the use of an SMT Solver such as Z3 to enumerate all PACs. We thus implemented this approach in C++ in the EmergeNS software [24] (more generally designed to simulate the dynamics of complex physicochemical systems and down the line to trace the physical emergence of natural selection). In practice, in a reaction system defined in EmergeNS, we ask the Z3 solver to find PAC candidates and to assess, for each candidate, the existence of a PAC witness, i.e. a reaction flow vector $\vec{v}$ yielding strictly positive net production rates for all entities of the candidate. We then remove the detected PAC from the search space and repeat the process until no new PAC is found.

As remarked in Section III A, verifying whether a given candidate is indeed a PAC (by assessing the existence of a PAC witness) is a linear programming problem, and can thus be achieved efficiently, i.e. in polynomial time. The need for Z3 comes from the search for PAC candidates in an exponential space of possible subsets of entities and reactions.

IV. PAC CONSISTENCY UNDER THERMODYNAMIC CONSTRAINTS

The kinetics of a reaction network must obey the second law of thermodynamics, a constraint that is not considered in the PAC definition. Indeed, this definition solely relies on the existence of a witness $\vec{v}$ of reaction flows, that may or may not be compatible with thermodynamic constraints.

More precisely, once association and dissociation constants are derived from free energies and activation barriers, they cannot be freely chosen. To clarify these constraints, recall that each entity e is associated with a chemical potential μ,

which is the sum of its molar Gibbs free energy (or standard chemical potential) G and the logarithm of its activity (that we identify here to its concentration $[e]$, hence placing ourselves in the ideal solution regime)[1]:

$$\mu = G + \ln [e] \tag{2}$$

In addition, the flow of a reaction R, denoted v_R, depends on the concentrations (i.e. activities) of its entities, under mass action kinetics:

$$v_R = k_R^{(+)} \cdot \prod_{e_i \text{ reactant}} [e_i]^{-\sigma_{e_i}^R} - k_R^{(-)} \cdot \prod_{e_j \text{ product}} [e_j]^{\sigma_{e_j}^R}, \tag{3}$$

where $k_R^{(+)}$ and $k_R^{(-)}$ are respectively the forward and backward kinetic rate constants and $\sigma_{e_i}^R$ is the stoichiometry of e_i in reaction R (negative for reactants, hence explaining the presence of a minus sign in the above formula). The key point, to enforce the second law under mass action kinetics, is to relate the rate constants to Gibbs free energies using the local detailed balance condition (also known as Eyring's formula):

$$\begin{aligned} k_R^{(+)} &= \exp(-G_R^\dagger + \sum_{e_i \text{ reactant}} (-\sigma_{e_i}^R) \cdot G_i), \\ k_R^{(-)} &= \exp(-G_R^\dagger + \sum_{e_j \text{ product}} \sigma_{e_j}^R \cdot G_j), \end{aligned} \tag{4}$$

where $G_R^\dagger$ is the free energy of the intermediate state that induces an activation barrier. Injecting Equation (4) in Equation (3) leads to the following formulation that will be further used below:

$$v_R = e^{-G_R^\dagger} \cdot \left(\prod_{e_i \text{ reactant}} e^{-\sigma_{e_i}^R \cdot \mu_i} - \prod_{e_j \text{ product}} e^{\sigma_{e_j}^R \cdot \mu_j} \right). \tag{5}$$

The sign of v_R is the same as that of the affinity of the reaction $\mathcal{A}_R = -\sum_i \sigma_{e_i}^R \cdot \mu_i$, expressing how a positive rate can be achieved. Recall, however, that positive flows for all reactions of a PAC will not necessarily entail autocatalysis, as we aim more specifically for flows balanced in such a way that each entity has a net positive production rate (Definition 1). The question of addressing the thermodynamic consistency of PACs can thus be reformulated as follows: can a PAC flow witness $\vec{v}$ be realized through a concentration vector $\vec{c}$ of all entities of the system? If such a vector $\vec{c}$ exists, the PAC will be considered a thermodynamically Consistent Autocatalytic Core (CAC). Here, "thermodynamic consistency" refers to mass action kinetics, with rate constants derived from Gibbs free energies and activation barriers.

[1] In this formulation, we set $RT = 1$, which can be done without loss of generality, as this is simply equivalent to expressing Gibbs free energies in RT units. In the following, the notation R will stand for a reaction, and the gas constant will not be referred to again. Similarly, other physical constants will be ignored in order to set a dimensionless framework for simplicity. Such quantities could be reintroduced without affecting the results presented in this work.

Definition 4. *A Consistent Autocatalytic Core (CAC) is a PAC for which there exists a concentration vector $\vec{c}$ yielding reaction flows forming a PAC witness.*

The concentration vector $\vec{c}$ will then be called a *CAC witness*. Here, we reach the second key result of the present study, namely, that one can always find such a CAC witness – i.e. that a single PAC is always thermodynamically consistent – as long as the concentration space is unbounded.

Theorem 2. *Let $\vec{v}$ be a PAC witness (or any flow vector compatible with the directions of the PAC reactions). Then there exists $\lambda > 0$ and a concentration vector $\vec{c}$ such that $\lambda\vec{v}$ is the flow vector induced by $\vec{c}$.*

The general proof of this theorem is given in Appendix 3, with an example provided in Appendix 3 a. It should be noted that Theorem 2 actually has a broader scope than the problem specifically addressed in this section, since it demonstrates that *any* flow vector $\vec{v}$ matching the directions of the PAC reactions can be realized in the concentration space up to some proportionality factor $\lambda > 0$. Notably, this holds for any values of activation barriers and Gibbs free energies, and even when food and waste concentrations are fixed to any arbitrary values.

This goes beyond the well-established result, based on affinity, that a positive flow can be achieved for any reaction by adjusting concentrations. Here, the difficulty is to aim for a fixed vector $\vec{v}$ of rates, up to some coefficient λ shared among all reactions of the PAC.

We deduce the following corollary from Theorem 2:

Corollary 1. *Any PAC is a CAC.*

Proof of Corollary 1. Consider a PAC formed of entities $(e_1, \ldots, e_n)$ and reactions $(R_1, \ldots, R_n)$ (recall that according to Blokhuis et al [15], the minimality of a PAC implies that it contains as many entities as reactions). Let $\vec{v} = (v_1, \ldots v_n)$ be a PAC witness. Since the inequalities to be satisfied by a PAC witness are all linear with respect to the coordinates of $\vec{v}$ (see Equation (1)), for all $\lambda > 0$, $\lambda\vec{v}$ is a PAC witness as well. Applying Theorem 2 gives us the existence of some $\lambda > 0$ and a concentration vector $\vec{c}$ yielding flows $\lambda\vec{v}$, which is a PAC witness. Thus, $\vec{c}$ is a CAC witness, and the arbitrary PAC we started with is indeed a CAC. □

V. COMPATIBILITY AMONG AUTOCATALYTIC CORES

In this section, we investigate whether thermodynamic constraints affect compatibility relationships among cores. To do so, we first analyze compatibility among PACs, that is, we identify sets of cores that are found compatible on the basis of the reaction network topology alone, hereafter called "multiPACs". A set of PACs is a multiPAC if there exists a common witness $\vec{v}$ of reaction flows allowing all the PACs of the set to run simultaneously. In the framework of Gagrani et al [16], this corresponds to a nonempty intersection of the flow-productive cones for the different autocatalytic cores considered. Here, instead of computing explicit intersections, we will use the Z3 solver to identify nonempty intersections, and

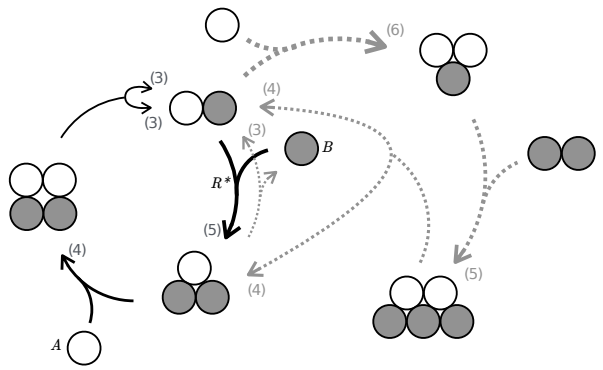


Figure 2. A simple example of two incompatible PACs. Ignoring foods and wastes and identifying A to empty circles and B to filled circles, the PAC with solid black arrows is $AB \to AB_2 \to A_2B_2 \to AB + AB$ and the PAC with gray dotted arrows is $AB_2 \to AB \to A_2B \to A_2B_3 \to AB + AB_2$. Numbers in brackets indicate the flows of PAC reactions allowing for a (local) net production of 1 of all their entities. The two PACs share the reaction labeled $R^\star$, but require it to run in opposite directions.

directly ask for a single vector $\vec{v}$ witnessing the different PACs simultaneously. To do so, we simply concatenate the requirements already defined for each individual PAC. Let us emphasize that a multiPAC witness does not guarantee global production of each entity when reactions from all PACs are jointly considered (this can be seen for example in Figure 4, where entity e_4 has a positive production rate within each PAC, although it is consumed overall).

Interestingly, we note that incompatibilities between PACs may occur for various reasons. The simplest case, illustrated in Figure 2, is when a reaction is shared between two PACs, but in opposite directions. This obviously prevents the existence of a common flow vector witness $\vec{v}$. Yet more subtle cases were also obtained with our software using randomly generated reaction networks. One example is described in Figure 3. Generally speaking, incompatibilities among PACs occur because of contradictory flow requirements, that is, when one PAC requires a reaction R_1 to run faster than a reaction R_2, while a second PAC requires the opposite.

To investigate the impact of thermodynamic constraints on compatibility relationships among cores, we now assess whether pairs of compatible PACs, witnessed by a common flow vector $\vec{v}$, also constitute pairs of compatible CACs, witnessed by a common concentration vector $\vec{c}$. As before, addressing this question using the SMT solver is straightforward: it is enough to concatenate the lists of constraints required for each of the CACs under study, and to ask if a $\vec{c}$ vector exists that simultaneously satisfies them all.

This analysis reveals that compatibility among cores inferred solely from the reaction network topology may overlook thermodynamic inconsistencies. Indeed, we find several instances of two compatible PACs making incompatible CACs. We give an example of such a behavior in Figure 4 and provide a formal proof that these two CAC are incompatible in Box 2.

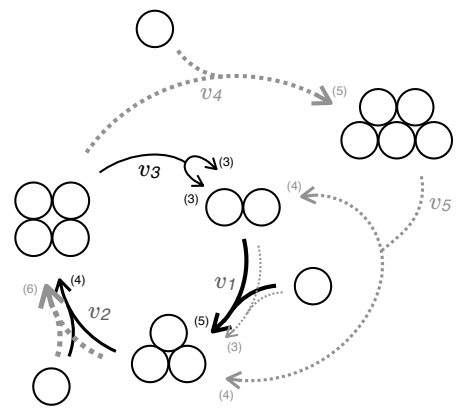


Figure 3. A more subtle example of two incompatible PACs, sharing two reactions indexed with flows v_1 and v_2. Flows allowing for unitary production of PAC entities are shown in brackets for both PACs. The inner PAC depicted with solid black arrows ($A_2 \to A_3 \to A_4 \to A_2 + A_2$) imposes the flow inequalities $v_1 > v_2 > v_3 > v_1/2$, while the gray dotted outer PAC ($A_2 \to A_3 \to A_4 \to A_5 \to A_2 + A_3$) imposes inequalities $v_1 + v_5 > v_2 > v_4 > v_5 > v_1$. This yields contradictory requirements: $v_1 > v_2$ for the first PAC and $v_1 < v_2$ for the second.

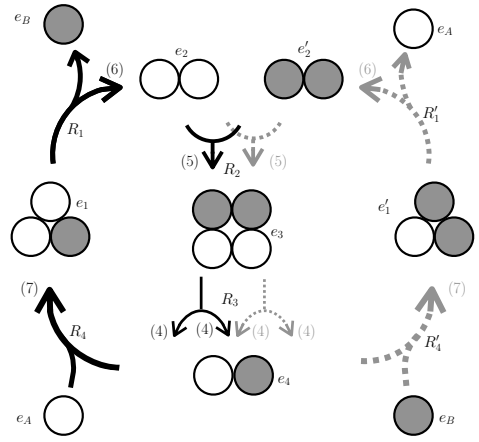


Figure 4. Two topologically compatible but thermodynamically incompatible autocatalytic cores. The two PACS ($\{R_1, R_2, R_3, R_4\}$ and $\{R_1', R_2, R_3, R_4'\}$) have two reactions in common. One can check that both cores run simultaneously with flows $v_1 = v_1' = 6$, $v_2 = 5$, $v_3 = 4$ and $v_4 = v_4' = 7$ which allows for a production rate of 1 for all entities belonging to a PAC. However, as detailed in Box 2, it can be shown that these cores cannot be instantiated simultaneously in the concentration space, i.e. they form a multiPAC but not a multiCAC.

Box 2: Thermodynamic incompatibility between CACs

On the example shown in Figure 4, we can first show that the two PACs are compatible, since we can choose a set of reaction flows that witnesses both of them simultaneously. For instance setting $v_1 = v_1' = 6$, $v_2 = 5$, $v_3 = 4$, $v_4 = v_4' = 7$ leads to a positive production rate of each entity by each PAC. More generally, this criterion is achieved for both PACs simultaneously when the following inequalities are satisfied:

$v_3 < v_2 < v_1 < v_4 < 2v_3$ and $v_3 < v_2 < v'_1 < v'_4 < 2v_3$.

We will show that these inequalities are not thermodynamically achievable, i.e. that they lead to contradictory requirements in the concentration space.

Notice that the inequalities imply that all v_i are strictly positive, because $v_3 < 2v_3$ entails $v_3 > 0$, and all other v_i are larger than v_3. As proven below, this sign constraint alone is not satisfiable: not all reactions can flow in the wanted direction.

We denote $x_i = e^{\mu_i}$ the exponential of the chemical potential of entity e_i (indexed as in Fig. 4), and $b_i = e^{-G_i^\dagger}$ where $G_i^\dagger$ is with respect to reaction R_i (similarly for R'_i). Reaction flows can thus be written as follows:

$$\begin{array}{ll} v_1 = b_1(x_1 - x_B x_2) & v_4 = b_4(x_A x_4 - x_1) \\ v'_1 = b'_1(x'_1 - x_A x'_2) & v'_4 = b'_4(x_B x_4 - x'_1) \\ v_3 = b_3(x_3 - (x_4)^2) & v_2 = b_2(x_2 x'_2 - x_3) \end{array}$$

From the positivity of all flows, we get:

$$\begin{cases} x_B x_2 < x_1 < x_A x_4 \\ x_A x'_2 < x'_1 < x_B x_4 \\ (x_4)^2 < x_3 < x_2 x'_2 \end{cases}$$

From the first line we deduce $x_B x_2 / x_4 < x_A$, and reinjecting in the second line we obtain $x_B x_2 x'_2 / x_4 < x_A x'_2 < x_B x_4$. This simplifies to $x_2 x'_2 < (x_4)^2$, which contradicts the condition of the third line. It should be noted that this proof is valid regardless of the activation barriers values, since the contradiction stems from the signs of the flows, while activation barriers only affect their amplitudes.

VI. DISCUSSION

Working toward the long-term goal of an explicit grounding of Darwinian dynamics into physical processes, we addressed in this study the implications of thermodynamic constraints on the existence and detection of autocatalytic cores given a reaction network. Our analysis builds on recent theoretical progresses made on the formalization of autocatalysis on the sole basis of the reaction network topology [15]. Under this definition, we show that the exhaustive detection of autocatalysis is an NP-complete problem. This finding fully justifies the use of constraint solvers (e.g. SMT, Integer Programming) toward which we converge with others [10, 16].

The constraints imposed by free energies and activation barriers can always be compensated by adjusting concentrations, thereby allowing any minimal autocatalytic cycle to also be thermodynamically consistent under mass action kinetics. In other words, the list of autocatalytic cores in a reaction network remains unaffected by these physical constraints, as long as concentrations are not limited by upper or lower bounds. However, as shown in Appendix 4, it should be noted that heterogeneity in free energies and activation barriers do restrict the volume of the concentration space where a core can effectively run.

These conclusions on isolated cores do not readily apply to combinations of cores. Indeed, thermodynamic realism does restrict the list of mutually compatible cores, even in an unlimited concentration space, so that topologically compatible cores can turn out incompatible. Incompatibilities between two autocatalytic cores can therefore stem from two distinct sources, namely the topology of the reaction network (PAC-incompatibility) and irreconcilable demands on concentrations (CAC-incompatibility).

A stimulating next step will be to investigate the implications of autocatalysis on the system's dynamics through time and space. Indeed, the existence of a CAC witness says little of the actual trajectories occurring in the concentration space, and on the influence of autocatalytic cores on these trajectories. Among the many autocatalytic cores that are thermodynamically achievable in a given system, which ones are actually encountered from a given starting point in the concentration space? Which ones are short-lived, long-lived, or keep running once the system has reached a steady state? Which ones run in the vicinity of this steady state, and perhaps contribute to drive the system in its direction? And finally, could autocatalysis contribute to generating more than one steady state in the concentration space? We anticipate that such multistability could enable a primordial form of heritable variation, paving the way to nascent Darwinian dynamics.

ACKNOWLEDGMENTS

We are very grateful to Nicolas Lartillot for his insightful contribution in setting up our modeling approach, and to Benjamin Kuperberg for sharing his open source GUI library OrganicUI. We also thank Romain Yvinec, Olivier Rivoire, Yann Sakref and Iris Magniez--Papillon for fruitful discussions, as well as Arnaud Mary, two anonymous reviewers and the associate editor for constructive comments on earlier versions of the manuscript. Thomas Kosc was supported by grant ANR-17-CE02-0021-01 to Sylvain Charlat.

[1] I. Budin and J. W. Szostak, Annu. Rev. Biophys. **39**, 245 (2010).
[2] N. Goldenfeld and C. Woese, Annu. Rev. Condens. Matter Phys. **2**, 375 (2011).
[3] S. Charlat, A. Ariew, P. Bourrat, M. Ferreira Ruiz, T. Heams, P. Huneman, S. Krishna, M. Lachmann, N. Lartillot, L. Le Sergeant d'Hendecourt, C. Malaterre, P. Nghe, E. Rajon, O. Rivoire, M. Smerlak, and Z. Zeravcic, Life **11**, 1051 (2021).
[4] V. Vanchurin, Y. I. Wolf, E. V. Koonin, and M. I. Katsnelson, Proc. Natl. Acad. Sci. U.S.A. **119**, e2120042119 (2022).
[5] S. Charlat, T. Heams, and O. Rivoire, in *Evolutionary Thinking Across Disciplines*, Vol. 478, edited by A. Du Crest, M. Valković, A. Ariew, H. Desmond, P. Huneman, and T. A. C. Reydon (Springer International Publishing, Cham, 2023) pp. 287–296, series Title: Synthese Library.
[6] D. A. Baum, Z. Peng, E. Dolson, E. Smith, A. M. Plum, and P. Gagrani, J. R. Soc. Interface **20**, 20230346 (2023).
[7] A. Sharma, D. Czégel, M. Lachmann, C. P. Kempes, S. I. Walker, and L. Cronin, Nature **622**, 321 (2023).
[8] L. L. J. Schoenmakers, T. A. C. Reydon, and A. Kirschning, Life **14**, 175 (2024).

[9] M. Kalambokidis and M. Travisano, Evolution **78**, 1 (2024).
[10] Z. Peng, J. Linderoth, and D. A. Baum, PLoS Comput. Biol. **18**, e1010498 (2022).
[11] Y. Liu and D. J. Sumpter, J. Biol. Chem. **293**, 18854 (2018).
[12] S. Sarkar and J. L. England, Phys. Rev. E **100**, 022414 (2019).
[13] M. Eigen and P. Schuster, Die Naturwissenschaften **64**, 541 (1977).
[14] D. A. Baum, BioScience **65**, 678 (2015).
[15] A. Blokhuis, D. Lacoste, and P. Nghe, Proc. Natl. Acad. Sci. U.S.A. **117**, 25230 (2020).
[16] P. Gagrani, V. Blanco, E. Smith, and D. Baum, J. Math. Chem. **62**, 1012 (2024).
[17] J. L. Andersen, C. Flamm, D. Merkle, and P. F. Stadler, J. Syst. Chem. **3**, 1 (2012).
[18] O. Weller-Davies, M. Steel, and J. Hein, Math. Biosci. **325**, 108365 (2020).
[19] W. Hordijk and M. Steel, Biosystems **152**, 1 (2017).
[20] S. A. Kauffman, J. Theor. Biol. **119**, 1 (1986).
[21] A. Despons, Y. De Decker, and D. Lacoste, Commun. Phys. **7**, 224 (2024).
[22] A. Arya, J. Ray, S. Sharma, R. Cruz Simbron, A. Lozano, H. B. Smith, J. L. Andersen, H. Chen, M. Meringer, and H. J. Cleaves, Chem. Sci. **13**, 4838 (2022).
[23] S. A. Cook, in *Proc. Annu. ACM Symp. Theory Comput.*, STOC '71 (1971) p. 151–158.
[24] D. Kuperberg, Emergens, `https://github.com/deniskup/EmergeNS` (2024).

APPENDIX

1. Uniqueness of reaction directions in PAC witnesses

Recall that a PAC in our framework is defined as an autocatalytic core from Blokhuis et al [15]. We will use some known properties of such cores:

Lemma 1 (from Blokhuis et al SI [15]). *Any autocatalytic core is of the following form:*

- *entities $e_1, \dots, e_n$,*
- *reactions $R_1, \dots, R_n$ (i.e. same number as entities),*
- *for each $i \in [1, n]$, the sole reactant of R_i among $e_1, \dots, e_n$ is e_i,*
- *the directed graph G with vertices $[1, n]$ and edges $\{(i, j) \mid e_j$ appears as product in $R_i\}$ is strongly connected.*

We show that this characterization implies that the direction of any reaction cannot be reversed in PAC witnesses. This was mentioned in [15, SI], with the sentence: "we choose the convention that any productive vector γ is positive, up to taking the opposite for some columns of M", but we will prove this here for completeness.

First of all, if a reaction R_i has two distinct products e_j, e_k, then the reverse reaction has two distinct reactants among the PAC entities, and is therefore incompatible with the third item of Lemma 1, so such a reaction cannot be reversed.

We now assume for contradiction that some reaction may be reversed in a PAC witness. Let us call R_{i_1} this reaction, with original reactant e_{i_1} and product e_{i_2}, with $i_1 \neq i_2$. Our assumption is that there exists another PAC witness for the same PAC, where R_{i_1} runs in reverse, that is with reactant e_{i_2} and product e_{i_1}. Since e_{i_2} can only appear once as reactant according to Lemma 1, this forces us to reverse reaction R_{i_2} as well. The former product e_{i_3} of this reaction becomes a reactant, forcing us to reverse reaction R_{i_3}. Continuing this process, one of these events will eventually occur:

1. We encounter a reaction R_{i_t} having two distinct products. Since we cannot reverse it, the entity e_{i_t} will appear twice as reactant, contradicting Lemma 1.

2. We encounter a reaction R_{i_t} whose unique product is e_{i_p} with $p < t$, i.e. e_{i_p} was already encountered in the process. Then, by the fourth item of Lemma 1, we must have $t = n$ and $p = 1$. That is, this process has reached all reactions and now loops back to the starting point.

Therefore, all that remains is to show that switching the direction of all reactions leads to a contradiction as well. Since all reactions have a unique product among the entities of the PAC, we are here in the case of a type I core in the terminology of [15]. We can renumber reactions so that in the original PAC, the unique product of reaction R_i is e_{i+1} for $i < n$, and e_1 for $i = n$. Let a_i be the absolute value of the stoichiometric coefficient of reactant e_i in R_i, and b_i be the stoichiometric coefficient of the product of this reaction (e_{i+1} or e_1 for $i = n$).

Let $\vec{v} = (v_1, \dots, v_n)$ be a PAC witness consistent with the original directions of reactions of the PAC, with $v_i > 0$ for all $i \in [1, n]$. Recall that as mentioned in Section II, no coordinate of $\vec{v}$ can be null, as the corresponding reaction could be removed, contradicting minimality of the PAC. What we aim to show is that there exists no other PAC witness $\vec{w} = (w_1, \dots, w_n)$, with $w_i < 0$ for each i, that would correspond to reversing the direction of all reactions.

The fact that $\vec{v}$ is a PAC witness is expressed by the following inequalities:

$$\begin{cases} b_n v_n & > & a_1 v_1 & \text{(growth of } e_1) \\ b_1 v_1 & > & a_2 v_2 & \text{(growth of } e_2) \\ & \vdots & & \\ b_{n-1} v_{n-1} & > & a_n v_n & \text{(growth of } e_n) \end{cases}$$

Multiplying these inequalities all together and simplifying by $v_1 v_2 \dots v_n$ (which is strictly positive) yields $b_1 b_2 \dots b_n > a_1 a_2 \dots a_n$.

This highlights the weight-asymmetry characterizing Type I cores as stated in [15, SI], with an additional constraint: the product of stoichiometric coefficients associated with reaction products is not only different from the product of absolute values of stoichiometric coefficients associated with reactants, but it must be strictly larger. This is enough to show that reversing the directions of all reactions is not possible, since this inequality would then be in the wrong direction. This completes the proof that the sign of flows in PAC witnesses cannot be altered.

2. Proof of Theorem 1: NP-completeness

This section is devoted to the proof of Theorem 1.

First, by the above remark, if a candidate PAC is given then it can be checked in polynomial time; thus PAC-DETECTION is in NP. To achieve the proof, it remains to be shown that it is NP-hard. We do so by reducing from the classical NP-complete problem SAT, that is, by translating the SAT problem into PAC-DETECTION.

An instance of SAT is a boolean formula φ on n variables $x_1, x_2, \ldots, x_n$. We will call literal a variable x_i or its negation $\overline{x}_i$. The formula φ is a conjunction of k clauses, i.e. of the form $\varphi = \bigwedge_{1 \leq j \leq k} c_j$, where each clause c_j is a disjunction of literals. The formula φ is satisfiable if there exists a valuation setting a boolean value for each x_i (and the opposite for $\overline{x}_i$), that allows φ to evaluate to *true*. The SAT problem, that is, asking whether an input formula φ is satisfiable, is known to be NP-complete [23].

Let us now encode SAT into PAC-DETECTION. Given an instance of SAT, i.e. a formula φ as above, we want to design a reaction system $S_\varphi = (E, R)$, together with a target entity $A \in E$ and a food set $F \subset E$, such that there is a PAC satisfying this instance of PAC-DETECTION if and only if φ is satisfiable. This would mean that a (hypothetical) polynomial-time algorithm for PAC-DETECTION would yield a polynomial-time algorithm for SAT, provided the reduction can be done in polynomial time, and produces a reaction system of polynomial size, which will be ensured here.

We choose as entities the set:

$$E = \{A, A_2, N\} \cup \{E_i, \overline{E}_i \mid 1 \leq i \leq n\} \cup \{C_j \mid 1 \leq j \leq k\}.$$

This means we have $3 + 2n + k$ entities in the system. Note that we use capital letters to distinguish entities from the variables and clauses from φ to which they refer.

For reactions, we choose the following set, where L_i always stands for either E_i or $\overline{E}_i$, e.g. the first line of the list of reactions shown below means that both reactions $A \rightleftharpoons E_1 + N$ and $A \rightleftharpoons \overline{E}_1 + N$ are present:

$$R : \begin{cases} A \rightleftharpoons L_1 + N & \\ E_i + \overline{E}_i \rightleftharpoons L_{i+1} & \text{for } 1 \leq i < n \\ E_n + \overline{E}_n \rightleftharpoons C_1 & \\ C_j + L \rightleftharpoons C_{j+1} & \text{for } 1 \leq j < k \text{ and } L \text{ literal of } c_j \\ C_k + L \rightleftharpoons A_2 & \text{for } L \text{ literal of } c_n \\ A_2 \rightleftharpoons A + A & \end{cases}$$

Finally, the target is A and the allowed foods will be $F = \{E_i, \overline{E}_i \mid 1 \leq i \leq n\}$.

The idea is that a valuation witnessing satisfiability of φ will correspond to a PAC of the form:

$$A \to L_1 \to \cdots \to L_n \to C_1 \to \cdots \to C_k \to A_2 \to A + A.$$

The crux of the construction is that reactions using C_j as reactant need one of the literals of c_j as food, that will be interpreted as the literal validating the clause c_j. In consequence, this literal must be a food, and cannot appear as one of the entities of the PAC. So the segment $L_1 \to \cdots \to L_n$ appearing in the PAC has to be formed exactly of the literals that are put to false in the valuation.

The entity N is used to force the direction of the reaction $A \to L_1$ in the PAC: since it can only be used as waste and not as food, the reaction cannot go in the opposite direction. From there, any PAC containing A and using only foods from F must be of the above form, and witnesses a valuation satisfying φ.

Example 3. [2] *Consider φ given by $c_1 = x_1 \vee x_2$, $c_2 = x_1 \vee \overline{x}_2$, and $c_3 = \overline{x}_1 \vee \overline{x}_2$. Then the only correct valuation is the one setting x_1 to true and x_2 to false. This is witnessed in the system S_φ by the PAC $A \to \overline{E}_1 \to E_2 \to C_1 \to C_2 \to C_3 \to A_2 \to A + A$, using E_1 and $\overline{E}_2$ as foods (and N as waste).*

If we add a clause $c_4 = \overline{x}_1 \vee x_2$ to φ, the formula becomes unsatisfiable and accordingly, the corresponding system S_φ does not contain a PAC satisfying the constraints.

This achieves the proof that PAC-DETECTION is NP-complete.

3. Proof that all PACs are CACs

In the following, to simplify notations, reactions will be oriented in accordance to the PAC, i.e. all coordinates of the PAC witness will be positive. A "positive flow vector" is a vector where all coordinates are positive.

a. An example

Example 4. *Let us first instantiate the proof with the PAC mentioned earlier, given by the following reactions (ignoring foods) $2e_1 \to e_2 \to e_3 \to e_1 + e_2$, depicted in Figure 5.*

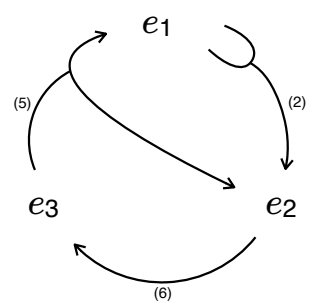


Figure 5. A PAC example of size 3 to give a first intuition of the proof.

A PAC witness is $\vec{v} = (2, 6, 5)$, ensuring a production rate of 1 for each entity. Using Equation (5), we now look for a concentration vector to realize this PAC witness with the following flows:

- $v_{R_1} = e^{-G_1^\dagger} \left(e^{2\mu_1} - e^{\mu_2}\right)$,
- $v_{R_2} = e^{-G_2^\dagger} \left(e^{\mu_2} - e^{\mu_3}\right)$,

[2] This is actually an instance of 2SAT which is a simpler problem, but it is just used here to illustrate the construction

- $v_{R_3} = e^{-G_3^\dagger}\left(e^{\mu_3} - e^{\mu_1+\mu_2}\right)$.

We will instead aim at realizing the flow vector $\lambda\vec{v}$ for some $\lambda > 0$, where $\vec{v} = (v_1, v_2, v_3)$ is the original PAC witness, e.g. $\vec{v} = (2, 6, 5)$. Letting $x_i = e^{\mu_i}$ and $w_i = v_i e^{G_i^\dagger}$ (so $w_i > 0$), the equation system can be written conveniently:

- $x_1 = \sqrt{\lambda w_1 + x_2}$
- $x_2 = \lambda w_2 + x_3$
- $x_3 = \lambda w_3 + x_1 x_2$

Substituting x_2 we obtain $x_1 = \sqrt{\lambda w_1 + \lambda w_2 + x_3}$, and finally substituting x_1 as well we obtain:

$$x_3 = \lambda w_3 + (\sqrt{\lambda w_1 + \lambda w_2 + x_3})(\lambda w_2 + x_3).$$

Let us note $g(\lambda, x_3)$ the above right-hand side expression, so that the equation becomes $x_3 = g(\lambda, x_3)$. Let us define $h(\lambda, x_3) = g(\lambda, x_3) - x_3$, so that we aim at $h(\lambda, x_3) = 0$. We will show that such a solution exists using the intermediate value theorem.
For $x_3 = 0$ and any $\lambda > 0$, we have $h(\lambda, 0) = \lambda w_3 + (\sqrt{\lambda w_1 + \lambda w_2})\lambda w_2 > 0$. Let us choose any $a \in (0, 1)$. For $x_3 = a$ and $\lambda = 0$, we have $h(0, a) = a^{3/2} - a < 0$. Because h is continuous there exists $\varepsilon > 0$ such that for all $\lambda \in (0, \varepsilon)$, $h(\lambda, a) < 0$. Let us choose $\lambda \in (0, \varepsilon)$, we know thanks to the intermediate value theorem (on x_3 with this fixed value of λ) that there exists $x_3 \in (0, a)$ such that $h(\lambda, x_3) = 0$. From this x_3 we can compute $x_2 = \lambda w_2 + x_3$, and $x_1 = \sqrt{\lambda w_1 + x_2}$. We have obtained a solution to our system by an appropriate choice of λ, x_1, x_2, x_3, thereby witnessing that the PAC is a CAC via Corollary 1.

b. *General proof*

We aim to prove Theorem 2 from Section IV by generalizing the proof exposed in the previous example to any PAC.

Notations.

Let $\vec{v}$ be a positive flow vector, we want to show that there exists $\lambda > 0$ and a concentration vector $\vec{c}$ such that $\lambda\vec{v}$ is the flow vector induced by $\vec{c}$. We will even show that this can be attained for any fixed concentration values of food and waste entities.

Let $\vec{v} = (v_1, \ldots, v_n)$ be the target flow vector, where all coordinates are strictly positive. Here n is the size of the PAC, i.e. both the number of entities $e_1, \ldots, e_n$, and reactions $R_1, \ldots, R_n$, where for each i, e_i is the sole reactant of R_i among entities of the PAC, see discussion in [15].

For an entity e_i ($1 \leq i \leq n$) with Gibbs free energy G_i, recall that its chemical potential is given by $\mu_i = G_i + \ln([e_i])$. We will be interested here in the exponential of this potential, a variable that we denote $x_i = e^{\mu_i}$.

Notice that x_i can be freely adjusted to any strictly positive value by choosing the appropriate concentration $[e_i]$, so we can turn the problem into that of finding a solution vector $\vec{x} = (x_1, \ldots, x_n)$. Analogs of components of $\vec{x}$ associated to food and waste entities are fixed to 1 for simplicity. Note that this does not involve chemostatting, but simply that a particular solution (a CAC witness) of a set of polynomial equations is searched, and that it is enough to fix food and waste variables to 1 to find such a solution. This does not mean that we impose a constant concentration of those entities over time. The proof can easily be adapted to any given values, but this will save us some notations, as food and waste entities can now be ignored in the computation of the reaction flows. One can refer to Appendix 4 for an analysis of the impact of food and waste potentials on the feasibility of PACs.

Let us recall Equation (5) giving the flow of a reaction in terms of chemical potentials:

$$v_R = e^{-G_R^\dagger} \cdot \left(\prod_{e_i \text{ reactant}} e^{-\sigma_{e_i}^R \cdot \mu_i} - \prod_{e_j \text{ product}} e^{\sigma_{e_j}^R \cdot \mu_j} \right) \tag{6}$$

In our case, given that e_i is the sole reactant of R_i (apart from possible foods), this can be rewritten for any $i \in [1, n]$:

$$v_{R_i} = e^{-G_{R_i}^\dagger}(x_i^{\sigma_{i,i}} - \prod_j x_j^{\sigma_{i,j}}) \tag{7}$$

Where $\sigma_{i,i}$ is the stoichiometry of the reactant e_i in R_i, and $\sigma_{i,j}$ is the stoichiometry of product e_j in R_i. It should be noted here that for simplicity we change the convention regarding the stoichiometry of reactants, so that all parameters are positive, including $\sigma_{i,i}$.

Let us note $w_i = v_i e^{G_{R_i}^\dagger}$. The fact that we aim for $\lambda\vec{v}$ to be the flow vector induced by $\vec{x}$ means that for all $i \in [1, n]$, we aim for $v_{R_i} = \lambda v_i$, so we must have $\lambda w_i = x_i^{\sigma_{i,i}} - \prod_j x_j^{\sigma_{i,j}}$, and finally:

$$x_i = \left(\lambda w_i + \prod_j x_j^{\sigma_{i,j}}\right)^{1/\sigma_{i,i}} \quad \text{Equation } (i) \tag{8}$$

Notice that $\sigma_{i,i} > 1$ corresponds to "inverted forks", where several e_i must be used as reactants. This can indeed occur in a PAC, as witnessed by Example 4.

Variable elimination.

Our goal is to find $\lambda > 0$ such that this system has a solution $(x_1, \ldots, x_n)$ with each $x_i > 0$. We will do so by iteratively reducing the number of variables and equations. First, we perform the following operation: as long as there exists an equation of the form $x_i = A_i$, where A_i is an expression not depending on x_i, we replace x_i by A_i in all other equations, and remove Equation (i).

We will end up with a list of equations of the form $x_i = g_i(\lambda, x_1, \ldots, x_n)$ where i ranges over some subset of $[1, n]$, and where g_i is an expression depending on some of its variables including x_i, and generated by the following grammar:

$$g := x_j \mid gg \mid (\lambda w_j + g)^{1/k}, \tag{9}$$

where j ranges over $[1, n]$, and k is a strictly positive integer. There are as many remaining equations as remaining variables. Without loss of generality, we assume that the remaining variables are $x_1, \ldots, x_p$, where $p \leq n$ is the number of remaining equations.

System rewriting as graph transformations.

Before tackling the resolution of the system per se, we relate some properties of the remaining equations to the topology of the PAC we started with. We associate the PAC with a graph $\mathcal{G}$, whose vertices are entities $e_1, \dots e_n$, and in which there is an edge $e_i \to e_j$ if e_j is among the products of the reaction using e_i as reactant.

Substituting variable x_i in the system according to Equation (i) amounts to removing vertex e_i and contracting all edges originating in this vertex, as shown in Figure 6 with $i = 5$:

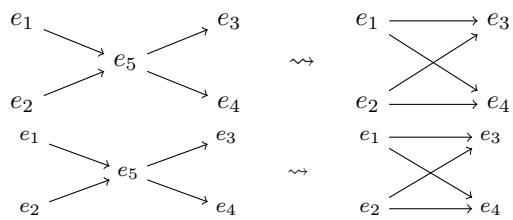


Figure 6. The effect of substituting the equation $x_5 = \lambda w_5 + x_3 x_4$ on the graph $\mathcal{G}$.

This operation can only be performed if there is no self-loop on e_i in the current graph. Therefore, the case where we can end up with only one equation on an entity x_i (i.e. the case $p = 1$) corresponds to the fact that the graph $\mathcal{G} \setminus \{x_1\}$ is acyclic.

We give in Figure 7 an example where this does not happen, no matter in which order the substitutions are done. This example corresponds to the Type V pattern from [15].

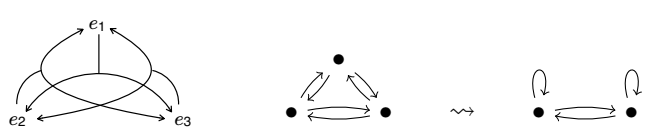


Figure 7. A 3-entity PAC, and its graph $\mathcal{G}$. After performing one substitution step, all nodes have self-loops and the substitution sequence must end, with equations of the form e.g. $x_1 = \lambda w_1 + (\lambda w_3 + x_1 x_2) x_2$.

Shape of the reduced system.

Let us now turn to the following lemma, which gives some properties of the expressions g_i reached by this process.

Lemma 2. *For all $1 \le i \le p$,*

- *g_i is always of the form $(\lambda w_i + g)^{1/k_i}$,*
- *for $\lambda = 0$, $g_i = \prod_j x_j^{m_{i,j}}$ with $m_{i,j} \ge 0$ for all j, and*
  - *if only one equation remains ($p = 1$), $m_{1,1} > 1$,*
  - *otherwise if $p > 1$, then for all $i \in [1, p]$, we have $m_{i,i} = 1$ and $m_{i,j} > 0$ for some $j \neq i$.*

Proof. The first item is a direct consequence of the operation we performed, starting from Equation (i).
The general shape of the second item is guaranteed by construction.

We now show item 2.1. Let us first give an interpretation for exponents $m_{i,j}$. Consider Equation (i) when $\lambda = 0$:

$$x_i = \prod_j x_j^{\sigma_{i,j}/\sigma_{i,i}}. \tag{10}$$

Recall that we have defined $\sigma_{i,i} > 0$ even though entity e_i is a reactant in reaction R_i. Exponents in the equation system for $\lambda = 0$ simply express the fact that in reaction R_i and starting from one unit of entity e_i, for all j one can produce $\sigma_{i,j}/\sigma_{i,i}$ units of entity e_j. The substitution algorithm is equivalent to consuming all available e_j via reaction R_j, with exponents of the products keeping track of their quantity. By iteration, if one started with entity e_i and could remove all reactions from the system (in our notation that would mean $p = 1$), one should recover that e_i can self-amplify through the reactions of the core since it allows for autocatalysis of e_i. This intuitively explains why we expect $m_{1,1} > 1$. For illustration in Example 4, the core allows the producing of 3/2 units of entity e_3 starting from one unit of it.

Let us now give a formal proof of the fact that $m_{1,1} > 1$ when $p = 1$, and let us note $m = m_{1,1}$ for concision. The argument is similar as the proof from Appendix 1. The sequence of substitutions leading to the expression g_1 amounts to describing a positive flow vector $\vec{\tau} = (\tau_1, \dots, \tau_n)$ in the following way: the flow of reaction R_1 is set to $\tau_1 = 1$, and then all available entities e_j other than e_1 are entirely consumed and substituted with the products of reaction R_j, until only m units of e_1 remain. This means that the net production of each entity is given by $M \cdot \vec{\tau} = (m - 1, 0, 0, \dots, 0)$, where M is the stoichiometric matrix associated to the PAC. Let us now consider a PAC witness $\vec{v}' = (v'_1, \dots, v'_n)$, with $v'_i > 0$ for each i, and scaled such that $v'_1 = 1$ (this is always possible since scaling preserves the PAC witness inequalities). We will show that the existence of $\vec{v}'$ implies $m > 1$. Notice that $M \cdot \vec{\tau} = (m - 1, 0, \dots, 0)$. Our goal is to show that for all $i \in [2, n]$, $v'_i < \tau_i$.

We use the graph $\mathcal{G}$ defined in the previous paragraph, and more precisely the fact that the case $p = 1$ corresponds to acyclicity of $\mathcal{G}' = \mathcal{G} \setminus \{e_1\}$. Without loss of generality, we can assume that a topological order of $\mathcal{G}'$ is given by $e_2, e_3, \dots, e_n$, that is to say there is no edge $e_i \to e_j$ with $j < i$. In particular, e_2 has no incoming edge in $\mathcal{G}'$. We compare the inflow and outflow of e_2 according to flow vectors $\vec{\tau}$ and $\vec{v}'$.

- The inflow is the same in both cases, as $\tau_1 = v'_1 = 1$ and no other incoming edge exists,
- the outflows are $\sigma_{2,2}\tau_2$ and $\sigma_{2,2}v'_2$ respectively,
- by construction, the net production rate with respect to $\vec{\tau}$ is $\sigma_{1,2}\tau_1 - \sigma_{2,2}\tau_2 = 0$, while for $\vec{v}'$, $\sigma_{1,2}v'_1 - \sigma_{2,2}v'_2 > 0$ as $\vec{v}'$ is a PAC witness.

Therefore, we have $\sigma_{2,2}v'_2 < \sigma_{1,2}v'_1 = \sigma_{1,2}\tau_1 = \sigma_{2,2}\tau_2$, so we can conclude $v'_2 < \tau_2$. We can continue by induction with $e_3, e_4, \dots$: at step i for entity e_i, the inflow is smaller or equal with flows from $\vec{v}'$ than from $\vec{\tau}$ (using induction hypothesis $v'_j < \tau_j$ for $j < i$), but the net production is positive for $\vec{v}'$ while it is 0 for $\vec{\tau}$. This allows us to conclude that for all $i \in [2, n]$, $v'_i < \tau_i$. Let us call m' the inflow of e_1 according to $\vec{v}'$, and recall that the corresponding inflow with respect to $\vec{\tau}$ is m. Since every reaction producing e_1 is strictly smaller

in $\vec{v}'$ than in $\vec{\tau}$, we have $m' < m$. However, to ensure net production of e_1 according to $\vec{v}'$, and since the outflow is $v_1 = 1$, we must have $1 < m'$. We can conclude $m > 1$ as desired.

Now if more than one equation remains, i.e. $p > 1$ (item 2.2 of the Lemma) and $m_{i,i} \neq 1$, it means that one found a subset of reactions of the cycle allowing for the net production (or consumption) of entity e_i, in contradiction with the assumption that the cycle is minimal. An illustration of this phenomenon is given in Example 5. The existence of some $j \neq i$ such that $m_{i,j} > 0$ follows from the condition that at least one other equation than the ith remains after the substitution algorithm ends. Indeed, if we could reach an equation where x_i is the only remaining variable, without having substituted all the other ones, it would mean that the PAC is not strongly connected, contradicting its minimality. □

Example 5. *We illustrate a case with $m_{i,i} \neq 1$ while two equations remain in Figure 8.A. The autocatalytic motif is not minimal, even though all entities appear only once as reactant, because it contains the core $e_2 \to e_3 \to e_1 \to e_1 + e_2$ (depicted in black arrows). Notice that the forked reaction $e_2 \to e_3 + e_4$ is not entirely in full line, because for this embedded PAC, e_4 corresponds to a waste. The equations corresponding to the full system are:*

$$\begin{cases} x_1 &= \lambda w_1 + x_2 x_3 \\ x_2 &= \lambda w_2 + x_3 x_4 \\ x_3 &= \lambda w_3 + x_1 \\ x_4 &= \lambda w_4 + x_5 \\ x_5 &= \lambda w_5 + x_3 x_4. \end{cases}$$

One can substitute x_3, then x_4 and x_2 to arrive to a system of two equations on x_1 and x_5:

$$\begin{cases} x_1 = \lambda w_1 + [\lambda w_2 + (\lambda w_3 + x_1)(\lambda w_4 + x_5)]\,(\lambda w_3 + x_1) \\ \\ x_5 = \lambda w_5 + (\lambda w_3 + x_1)(\lambda w_4 + x_5). \end{cases}$$

Taking $\lambda = 0$, one obtains $x_1 = x_1^2 x_5$ and $x_5 = x_1 x_5$, thus we have $m_{1,1} = 2$ and $m_{5,5} = 1$. The appearance of x_1^2 in the first equation betrays the existence of an inner PAC within the motif, where entity e_5 is ignored. A visual interpretation is given in Figure 8.B, which shows the reduced cycle with the two entities e_1 and e_5 obtained at the end of the sequence of substitutions. Notice that entity e_5, in order to self-amplify, must produce e_1. However, there exists a reaction path not involving entity e_5 allowing for the production of 2 units of e_1 starting with 1 unit of it (full black arrows in panel B), which is why $m_{1,1} = 2$. This indeed corresponds to the PAC shown in full black arrows in panel A.

Solving the reduced system. Turning back to the general proof of Theorem 2, it is enough to solve the remaining system of equations, which is of the form $x_i = g_i(\lambda, x_1, \ldots, x_p)$ for all $i \in [1, p]$. We can then recover the missing x_j (for $j \in [p+1, n]$) using the previous substitutions $x_j = A_j$.

Case $p = 1$. Let us first treat the easier case where $p = 1$, i.e. we have a single variable x_1 and a single equation $x_1 = g_1(\lambda, x_1)$.

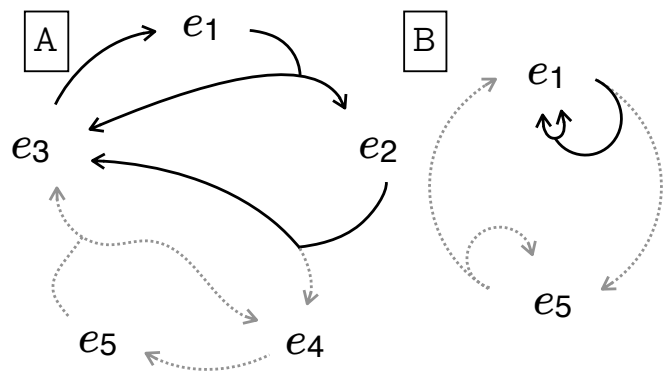


Figure 8. Non-minimal autocatalytic motif.

Lemma 3. *There exists $\varepsilon > 0$ such that for all $\lambda \in (0, \varepsilon)$, there exists $x_1 \in (0, 1)$ verifying $x_1 = g_1(\lambda, x_1)$.*

Proof. Let $h_1(\lambda, x_1) = g_1(\lambda, x_1) - x_1$. We will use the fact that h_1 is continuous, and evaluate h_1 at $x_1 = 0$ and $x_1 = a$ for some $a \in (0, 1)$ and for a suitable λ. We will conclude the proof using the intermediate value theorem.

First of all for any $\lambda > 0$ and $x_1 = 0$, we have $h_1(\lambda, 0) > 0$, since by construction g_1 is built from positive terms using functions preserving positivity, and furthermore contains a term of the form $\lambda w_1 > 0$ not depending on x_1.

Now, let us consider the case $x_1 = a$ for any $a < 1$. We have $h_1(\lambda, a) = g_1(\lambda, a) - a$. By Lemma 2, for $\lambda = 0$ we have $g_1(0, a) = a^m$ with $m > 1$. Since $a < 1$ and $m > 1$, we have $a^m < a$. Thus, $h_1(0, a) < 0$. Since the function $\lambda \mapsto h_1(\lambda, a)$ is continuous, there exists $\varepsilon > 0$ such that for all $\lambda \in (0, \varepsilon)$, we have $h_1(\lambda, a) < 0$. We can conclude (intermediate value theorem with respect to x_1) that for all $\lambda \in (0, \varepsilon)$, there exists $x_1 \in (0, a)$ such that $h_1(\lambda, x_1) = 0$.

□

Therefore, choosing any $\lambda < \varepsilon$ and applying this lemma gives a solution to the system.

Case $p > 1$. If $p > 1$, we will solve the system by constructing a sequence of partial continuous functions $f_i : \mathbb{R}^i \to \mathbb{R}$ ($1 \leq i \leq p$) and $\varepsilon > 0$ such that:

- if $x_1, \ldots, x_{i-1} \in (0, 1)$ (resp. $[0, 1)$) and $\lambda \in (0, \varepsilon)$ (resp. $[0, \varepsilon)$), then $f_i(\lambda, x_1, \ldots, x_{i-1})$ is defined, and its image lies in $(0, 1)$ (resp. $[0, 1)$),
- if there exists $\lambda, x_1, \ldots, x_p$ such that for all i, $x_i = f_i(\lambda, x_1, \ldots, x_{i-1})$, then we have a solution of the wanted equations.

We define f_i by induction, starting with f_p.

Let us consider the equation $x_p = g_p(\lambda, x_1, \ldots, x_p)$. Let $h_p(\lambda, x_1, \ldots, x_p) = g_p(\lambda, x_1, \ldots, x_p) - x_p$.

Our goal is to define f_p as a value of x_p yielding $h_p = 0$. Thus, we have to show that such a root of h_p exists.

Lemma 4. *For any $x_1, \dots, x_{p-1} \in (0,1)$, there exists ε_p such that for all $\lambda \in (0, \varepsilon_p)$, there exists $x_p \in (0,1)$ verifying $h_p(\lambda, x_1, \dots, x_{p-1}, x_p) = 0$.*

Proof. As before, we will use the fact that h_p is continuous (by construction of g_p), and look at the values of h_p for $x_p = 0$ and $x_p = a < 1$.

First of all for $x_p = 0$, for any $\lambda, x_1, \dots, x_{p-1} > 0$, we have $h_p(\lambda, x_1, \dots, x_{p-1}, 0) > 0$, by construction of g_p.

Now, let us consider the case $x_p = a$ for some arbitrary $a < 1$. We have $h_p(\lambda, x_1, \dots, x_{p-1}, a) = g_p(\lambda, x_1, \dots, x_{p-1}, a) - a$. For $\lambda = 0$, we have $g_p(0, x_1, \dots, x_{p-1}, a) = a \prod_{j \neq p} x_j^{m_{p,j}}$ by Lemma 2.

By Lemma 2, we have in addition $m_{p,j} > 0$ for some $j \neq p$. Thus, for $x_1, \dots, x_{p-1} < 1$, we have $h_p(0, x_1, \dots, x_{p-1}, 1) = a(\prod_{j \neq p} x_j^{m_{p,j}} - 1) < 0$. Since h_p is continuous, there exists ε_p such that for all $\lambda \in (0, \varepsilon_p)$, we have $h_p(\lambda, x_1, \dots, x_{p-1}, a) < 0$. By the intermediate value theorem, for all $\lambda \in (0, \varepsilon_p)$, there exists $x_p \in (0, a)$ such that $h_p(\lambda, x_1, \dots, x_{p-1}, x_p) = 0$. Here we could actually have taken $a = 1$, but we give the scheme that will be used throughout the rest of the induction. □

We can thus define $f_p(\lambda, x_1, \dots, x_{p-1})$, as the least value of x_p such that $h_p(\lambda, x_1, \dots, x_{p-1}, x_p) = 0$, as soon as the conditions of the lemma are met. Moreover, f_p is continuous as well, since it is defined as the first root of a continuous function with non-zero boundary conditions.

We now get rid of equation $x_p = g_p$ by replacing x_p with $f_p(\lambda, x_1, \dots, x_{p-1})$ in all the other equations. We then have a system of the form $x_i = g_i(\lambda, x_1, \dots, x_{p-1}, f_p(\lambda, x_1, \dots, x_{p-1}))$ for all $i \in [1, p-1]$.

Notice that replacing x_p with f_p allows us to carry out the previous construction for f_{p-1}, as the only property asked of the various x_j other than the current x_{p-1} under consideration is to be in $(0,1)$, and that is the case for f_p. Notice that x_p is now a function of $\lambda, x_1, \dots, x_{p-1}$, but this is not a problem, as long as it remains in $(0,1)$ (allowing for extremal cases if some arguments are extremal as well), the construction can be carried on.

We can thus iterate this construction, and obtain a sequence of functions f_i as wanted. At each step, the ε_i can be chosen to be the minimum of the ε_{i+1} obtained at the previous step and the ε needed at the current step. This guarantees that all f_j for $j \geq i$ are defined for all $\lambda < \varepsilon_i$.

Continuing this construction, we will end up with a sequence of functions f_i such that for all $\lambda < \varepsilon$, there exists $x_1, \dots, x_p$ such that $x_i = f_i(\lambda, x_1, \dots, x_{i-1})$ for all $i \in [1, p]$. This will give us a solution to the system of equations, and thus a solution to the original problem. Indeed, we only need to choose some $\lambda < \varepsilon_1$, $x_1 = f_1(\lambda)$, $x_2 = f_2(\lambda, x_1)$, etc. to obtain a solution to the system.

As described earlier, we can infer the values of $x_{p+1}, \dots, x_n$ from $x_1, \dots x_p$ by using the previous substitutions $x_j = A_j$. From this, we can finally compute the concentration vector $\vec{c}$ such that $\lambda \vec{v}$ is the flow vector induced by $\vec{c}$.

As explained in Corollary 1, this is sufficient to show that any PAC is CAC, as we can start with a PAC witness $\vec{v}$ and apply this construction to find a CAC witness $\vec{c}$.

4. On the degree of feasibility of PACs

In this section we investigate how thermodynamic constraints affect the degree to which a PAC can be a CAC, in the sense of exploring quantitatively and analytically how such constraints, though not being able to prevent the feasibility of PACs (see Corollary 1), nonetheless affect the limitations toward instantiating a PAC in the concentration space. We will focus on a specific topology of type I PACs, following the terminology of [15]. We will consider a PAC of size n with a single fork, while reactions other than the fork may consume food and produce waste. These assumptions are for illustration purposes, and we believe the phenomena that we will describe here occur more generally. Thus, the PAC considered is of the following form, where $e_n \to e_1 + e_1$ is the forked reaction:

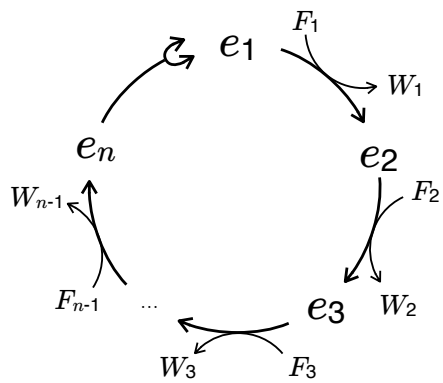


Figure 9. Illustration of a PAC of size n of type I, using the typology of [15].

Following the notations of the proof explained in Appendix 3, one would end up considering the following system of equations:

$$\begin{cases} F_1 x_1 & = & \lambda w_1 + W_1 x_2 \\ F_2 x_2 & = & \lambda w_x + W_2 x_3 \\ & \vdots & \\ F_{n-1} x_{n-1} & = & \lambda w_{n-1} + W_{n-1} x_n \\ x_n & = & \lambda w_n + x_1^2. \end{cases}$$

F_i (respectively W_i) is the exponential of the chemical potential of food (respectively waste) i. The substitution algorithm from Section 3 allows us to inject the $(i+1)^{\text{th}}$ line into the ith until one finally obtains a single equation in x_1:

$$x_1 = \lambda \left(\frac{1}{F_1} w_1 + \frac{W_1}{F_1 F_2} w_2 + \cdots + \frac{W_1 \dots W_{n-1}}{F_1 \dots F_{n-1}} w_n \right) + \frac{W_1 \dots W_{n-1}}{F_1 \dots F_{n-1}} x_1^2$$

rewritten as:

$$x_1 = g(\lambda, x_1).$$

For this particular PAC example, the dependence of λ with respect to other parameters can be made explicit:

$$\lambda = \frac{\left(1 - \frac{1}{\beta}x_1\right)x_1}{\left(\frac{1}{F_1}w_1 + \cdots + \frac{W_1 \ldots W_{n-2}}{F_1 \ldots F_{n-1}}w_{n-1} + \frac{W_1 \ldots W_{n-1}}{F_1 \ldots F_{n-1}}w_n\right)}. \quad (11)$$

It should be noted that the condition $\lambda > 0$ imposes $x_1 \in (0, \beta)$ with $\beta = \frac{F_1 \ldots F_{n-1}}{W_1 \ldots W_{n-1}}$. Notice that if all food and waste variables are set to 1, one recovers the interval of the general proof of Appendix 3 which was $(0, 1)$. Regardless, the PAC can be realized for any value of food and waste potentials.

Equation (11) shows that the solution λ has non-trivial dependencies on food, waste and PAC species chemical potentials, activation barriers of reactions and the witness $\vec{v}$. To facilitate the discussion, let us assume that the witness $\vec{v}$ is the one allowing for a unitary production of all PAC species (such a witness always exists because the stoichiometry matrix of a PAC is square and invertible, see discussion in [15, SI, proposition 2]). Hence, because the witness $\lambda\vec{v}$ can be instantiated in the concentration space, λ can be interpreted as the amplitude of the PAC, i.e. the level up to which it can effectively produce its entities, or its degree of feasibility, relative to the case of a unitary production rate of PAC entities.

Notice that both the numerator and the denominator of Equation (11) depend on terms of the form $\prod_i W_i / \prod_j F_j$. The denominator becomes smaller as the food variables grow larger relative to the waste variables. As far as the numerator is concerned, we plot in Figure 10 the function $\theta_\beta(x_1) = (1 - x_1/\beta)x_1$ for two values of β (namely $\beta = 1$ and $\beta = 2$). θ_β is an inverted parabola which is positive for $x_1 \in [0, \beta]$ and reaches it maximum at $x_1 = \beta/2$, the maximum being $\beta/4$. Hence, the numerator of Equation (11) (taken as a function of x_1) can get larger as β gets larger, i.e. as food variables get large in comparison to waste variables. We thus conclude that λ can reach higher values when the food variables are globally large in comparison to waste variables, which from now on is referred to as a favorable environment.

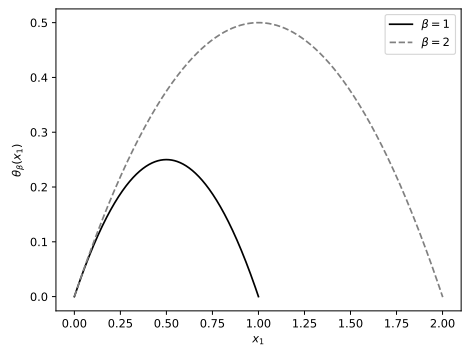


Figure 10. The numerator of Equation (11), $\theta_\beta(x_1) = (1 - x_1/\beta)x_1$ as a function of x_1 for $\beta = 1$ and $\beta = 2$.

A favorable environment can always be reached by choosing appropriate food and waste concentrations. However, it should be recalled that the variables appearing in Equation (11) are related to the free energy and concentration of entities through the relation $F_j = e^{G_{f_j}}[f_j]$ (illustrated for the jth food species).

We have seen that the PAC realization is easier when F_j's are large relative to W_i's. Recall that $W_j = e^{G_{w_j}}[w_j]$. If one waste free energy is increased by δG, then this corresponds to a multiplication of the corresponding W_j by $e^{\delta G}$, which will have to be compensated for by similar variations in the concentrations of food and/or waste entities. Thus, slight changes in the free energy values are exponentially passed on the concentration space. This highlights how PACs whose food and waste free energies are unfavorably distributed can be hard to instantiate in the concentration space in practice.

Similar conclusions can be reached regarding the concentrations of PAC entities. Notice that in Equation (11) λ depends only on x_1 in the numerator shown in Figure 10. This shows that x_1 is upper bounded, so the concentration space of entity e_1 associated to $\lambda > 0$ (i.e. the PAC running) increases its size as its free energy decreases. Notice once again that a favorable environment (i.e. large values of β) is associated with a wider interval of x_1 allowing $\lambda > 0$.

We conclude that regions of the concentration space where the PAC efficiently runs (i.e. runs with large values of λ) always exist if all concentrations can be chosen freely, but that free energies unfavorably distributed among food, waste and PAC entities will significantly restrict their volume.

On similar grounds, one can study the impact of activation barriers on the feasibility of PACs. Suppose that all reactions of the PAC get penalized via an increase of their activation barriers, such that $G_k^\dagger \leftarrow G_k^\dagger + \Delta$ with $\Delta > 0$ for all $k \in [1, n]$. This will in turn multiply the denominator of Equation (11) by a global factor e^Δ, and, as a consequence, multiply λ by $e^{-\Delta} < 1$. This can be compensated by choosing a more favorable environment, i.e. by increasing food concentrations and diminishing waste concentrations, such that β increases. Thus, the regions of the concentration space associated with large values of λ are tightened.