Computing Properties of Thermodynamic Binding Networks: An Integer Programming Approach

David Haley
University of California, Davis, CA, USA

David Doty
University of California, Davis, CA, USA

Abstract

The thermodynamic binding networks (TBN) model [9] is a tool for studying engineered molecular systems. The TBN model allows one to reason about their behavior through a simplified abstraction that ignores details about molecular composition, focusing on two key determinants of a system's energetics common to *any* chemical substrate: how many molecular bonds are formed, and how many separate complexes exist in the system. We formulate as an integer program the NP-hard problem of computing *stable* (a.k.a., minimum energy) configurations of a TBN: those configurations that maximize the number of bonds and complexes. We provide open-source software [13] solving this integer program. We give empirical evidence that this approach enables dramatically faster computation of TBN stable configurations than previous approaches based on SAT solvers [3]. Furthermore, unlike SAT-based approaches, our integer programming formulation can reason about TBNs in which some molecules have unbounded counts. These improvements in turn allow us to efficiently automate verification of desired properties of practical TBNs. Finally, we show that the TBN has a natural representation with a unique Hilbert basis describing the "fundamental components" out of which *locally* minimal energy configurations are composed. This characterization helps verify correctness of not only stable configurations, but entire "kinetic pathways" in a TBN.

2012 ACM Subject Classification Theory of computation → Theory and algorithms for application domains

Keywords and phrases thermodynamic binding networks, integer programming, constraint programming

Digital Object Identifier 10.4230/LIPIcs.DNA.27.2

Supplementary Material *Software (Source Code)*: `https://github.com/drhaley/stable_tbn`

Funding Supported by NSF award 1900931 and CAREER award 1844976.

1 Introduction

Recent experimental breakthroughs in DNA nanotechnology [5] have enabled the construction of intricate molecular machinery whose complexity rivals that of biological macromolecules, even executing general-purpose algorithms [23]. A major challenge in creating synthetic DNA molecules that undergo desired chemical reactions is the occurrence of erroneous "leak" reactions [16], driven by the fact that the products of the leak reactions are more energetically favorable. A promising design principle to mitigate such errors is to build "thermodynamic robustness" into the system, ensuring that leak reactions incur an energetic cost [14, 20, 22] by logically forcing one of two unfavorable events: either many molecular bonds must break – an "enthalpic" cost – or many separate molecular complexes (called *polymers* in this paper) must simultaneously come together – an "entropic" cost.

The model of *thermodynamic binding networks* (TBNs) [9] was defined as a combinatorial abstraction of such molecules, deliberately simplifying substrate-dependent details of DNA in order to isolate the foundational energetic contributions of forming bonds and separating polymers. A TBN consists of *monomers* containing specific *binding sites*, where binding

site a can bind only to its complement a^*. A key aspect of the TBN model is the lack of geometry: a monomer is an *unordered* collection of binding sites such as $\{a, a, b^*, c\}$. A *configuration* of a TBN describes which monomers are grouped into *polymers*; bonds can only form within a polymer.[1] One can formalize the "correctness" of a TBN by requiring that its desired configuration(s) be *stable*: the configuration maximizes the number of bonds formed, a.k.a., it is *saturated*, and, among all saturated configurations, it maximizes the number of separate polymers.[2] See Figure 1 for an example. Stable configurations are meant to capture the *minimum free energy structures* of the TBN. Unfortunately, answering basic questions such as "*Is a particular TBN configuration stable?*" turn out to be NP-hard [3].

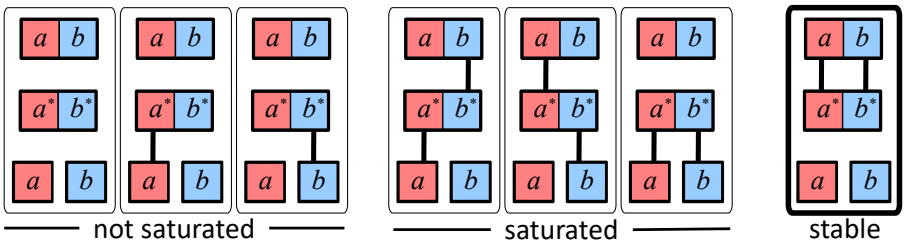

Figure 1 Example of a simple thermodynamic binding network (TBN). There are four monomers: $\mathbf{m}_1 = \{a^*, b^*\}, \mathbf{m}_2 = \{a, b\}, \mathbf{m}_3 = \{a\}, \mathbf{m}_4 = \{b\}$, with seven configurations shown: four of these configurations are saturated because they have the maximum of 2 bonds. Of these, three have 2 polymers and one has 3 polymers, making the latter the only stable configuration. Despite the suggestive lines between binding sites, the model of this paper ignores individual bonds, defining a configuration solely by how it partitions the set of monomers into polymers, assuming that a maximum number of bonds will form within each polymer. (Thus other configurations exist besides those shown, which would merge polymers shown without allowing new bonds to form.)

1.1 Our contribution

Our main contribution is a reduction that formulates the problem of finding stable configurations of a TBN as an integer program (IP). Of course, the problem, appropriately formalized, is "clearly" an NP search problem, so the mere existence of such a reduction is not particularly novel. However, our formulation is notable in three respects: 1) We carefully avoid certain symmetries (particularly those present in the existing SAT-based formulation of Breik et al. [3]), which dramatically increases the search efficiency in practice. 2) We use the optimization ability of IP solvers as a natural way to maximize the number of polymers in any saturated configuration. 3) Our formulation leads to a natural interpretation of the *Hilbert basis* [7] of a TBN as its minimal saturated polymers, which intuitively are the polymers existing in any *local* energy minimum configuration. Since highly optimized software exists for calculating Hilbert bases [1], this expands the range of TBN behaviors that can be automatically reasoned about.

[1] A primary goal of the TBN model is to establish that certain undesired configurations impose an energetic cost, i.e., to demonstrate *impossibility* results. From this perspective, the lack of geometric constraints means that such impossibility results give stronger guarantees than if geometric constraints, such as unpseudoknottedness of DNA secondary structure, were imposed on the allowable configurations. For example, "leakless" DNA strand displacement systems [20, 22] retain their leaklessness even if pseudoknotted structures are permitted.

[2] This definition captures the limiting case (often approximated in practice in DNA nanotechnology) corresponding to increasing the strength of bonds, while diluting (increasing volume), such that the ratio of binding to unbinding rate goes to infinity.

This formulation allows us to automate portions of formal reasoning about TBNs, helping verify their correctness. The TBN model abstracts away the continuous nature of real free energy into discrete integer-valued steps. In the limit of dilute solutions (bringing together polymers incurs a large energy cost) and very strong bonds (breaking a bond incurs a huge energy cost), even one integer step of energy difference is considered significant. Furthermore, in a catalytic system, the presence (or absence) of even a single polymer may have a very large impact on the system's trajectory, an indication that system analyses are very numerically sensitive, and thus their behavior is not well-described by approximation methods. For these reasons it is crucial when verifying such systems that we identify the *exact* solution to the optimization problem, rather than settling for more efficiently computable approximations (e.g., via continuous relaxation [6] or local search [18]).

1.2 Related work

Breik, Thachuk, Heule, and Soloveichik [3] characterize the computational complexity of several natural problems related to TBNs. For instance, it is NP-complete to decide whether a saturated configuration exists with a specified number of polymers, and even NP-hard to approximate within any constant factor the number of polymers in a stable configuration (i.e., the maximum in any saturated configuration).

Breik et al. also developed software using a SAT solver to produce stable configurations of a TBN. This formulation requires "labelled" monomers (where two different instances of the same monomer type are represented by separate Boolean variables), which become nodes in a graph, and polymers are realized as connected components within the graph. By labelling the monomers they become unique copies of the same monomer type; n copies of a monomer type increases the size of the search space by factor $n!$ by considering these symmetric configurations separately. Furthermore, the software explicitly explores all possible symmetries of bonding arrangements within a polymer. For instance, monomers $\{a^*, a^*\}$ and $\{a, a\}$ can bind in two different ways (the first a^* can bind either the first or second a), even though both have the same number of bonds and polymers. This over-counting of symmetric configurations prevents the software from scaling to efficiently analyze certain TBNs with large counts of monomers. Our IP formulation avoids both types of symmetry.

2 Preliminaries

2.1 Definitions

A *multiset* is an unordered collection of objects allowing duplicates (including countably infinite multiplicities), e.g., $\mathbf{v} = \{2 \cdot a, b, \infty \cdot d\}$. Equivalently, a multiset with elements from a finite set U is a vector $\mathbf{v} \in (\mathbb{N} \cup \{\infty\})^U$ describing the counts, indexed by U; in the example above, if $U = \{a, b, c, d\}$, then $\mathbf{v}(a) = 2$, $\mathbf{v}(b) = 1$, $\mathbf{v}(c) = 0$, and $\mathbf{v}(d) = \infty$. The *cardinality* of a multiset $\mathbf{v} \in \mathbb{N}^U$ is $|\mathbf{v}| = \sum_{u \in U} \mathbf{v}(u)$; a *finite multiset* $\mathbf{v}$ obeys $|\mathbf{v}| < \infty$. A *site type* is a formal symbol, such as a, representing a specific binding site on a molecule; in Figure 1 the site types are a, a^*, b, b^*. Each site type has a corresponding *complement type* denoted by a star, e.g. a^*. Complementarity is an involution: $(a^*)^* = a$. A site and its complement can form an attachment called a *bond*. We follow the convention that for any complementary pair of sites a, a^*, the total count of a^* across the whole TBN is at most that of a, i.e., the starred sites are *limiting*. A *monomer type* is a finite multiset of site types.[3]

[3] Concretely, in a DNA nanotech design, a monomer corresponds to a strand of DNA, whose sequence is logically partitioned into binding sites corresponding to long (5-20 base) regions, e.g., 5′-AAAGG-3′, intended to bind to the complementary sequence 3′-TTTCC-5′ that is part of another strand.

A *thermodynamic binding network (TBN)* is a multiset of monomer types; equivalently, a vector in $(\mathbb{N} \cup \{\infty\})^m$, if m is the number of monomer types and we have fixed some standardized ordering of them. We allow some monomer counts to be infinite in order to capture the case where some monomers are added in "large excess" over others, a common experimental approach [16,17]. A *polymer* $\mathbf{P}$ is a finite multiset of monomer types, equivalently, a vector in $\mathbb{N}^m$, where m is the number of monomer types in a standardized ordering;[4] thus $|\mathbf{P}|$ represents the number of monomers in $\mathbf{P}$. Note that despite the suggestive lines representing bonds in Figure 1, this definition does not track which pairs of complementary sites are bound within a polymer. Indeed, it is allowable in the model for a polymer to consist of monomers that cannot be connected by bonds, e.g., the polymer $\{\{a\}, \{b\}\}$ using monomers from Figure 1. In general we study configurations that *minimize* the number of merges necessary to reach them, and merges that create such polymers in general would not be useful in such a setting.

Given a TBN $\mathcal{T}$, let $S_{\mathcal{T}}$ (respectively, $S^*_{\mathcal{T}}$) be the set of unstarred (resp., starred) site types of $\mathcal{T}$. For a monomer $\mathbf{m}$ and site type $s \in S_{\mathcal{T}}$, let $\mathbf{m}(s)$ denote the count of s minus the count of s^* in $\mathbf{m}$ (intuitively, $\mathbf{m}(s)$ is the "net count" of s in $\mathbf{m}$, negative if there are more s^*.) For $s^* \in S^*_{\mathcal{T}}$ let $\mathbf{m}(s^*) = -\mathbf{m}(s)$. The *exposed sites* of a polymer $\mathbf{P}$ are a finite multiset of site types that results from removing as many (site, complement) pairs from a polymer as possible, described by the net count of sites when summed across all monomers in the polymer. For example, in the polymer $\{\{a^*, b^*, c^*\}, \{a, c\}, \{a, b, c\}, \{c, d^*\}\}$, the exposed sites are $\{a, 2 \cdot c, d^*\}$.

A *configuration* of a TBN is a partition of the monomers of the TBN into polymers; we write $|\gamma|$ to denote the number of polymers in configuration γ. A polymer is *self-saturated* if it has no exposed starred sites. A configuration is *saturated* if all of its polymers are self-saturated. Since we assume that, across the entire configuration, starred sites are limiting, this is equivalent to stipulating that the maximum possible number of bonds are formed. Write $\Gamma_{\mathcal{T}}$ to denote the set of all saturated configurations of the TBN $\mathcal{T}$. A configuration is *stable* if it is saturated and has the maximum number of non-singleton polymers among all saturated configurations.

Since the number of polymers may be infinite, we will use the equivalent notion that stable configurations are those that can be constructed by starting with the "melted" configuration whose polymers are all singletons containing only one monomer, performing the minimum number of polymer merges necessary to reach a saturated configuration. In general this quantity could be infinite, but we consider only TBNs in which a saturated configuration is reachable via a finite number of merges. For example, consider the TBN consisting of monomer types $\mathbf{t} = \{a\}$, $\mathbf{b} = \{a^*\}$, with counts $\infty \cdot \mathbf{t}$ and $2 \cdot \mathbf{b}$. The unique stable configuration has polymers $\{2 \cdot \{\mathbf{b}, \mathbf{t}\}, \infty \cdot \mathbf{t}\}$, since two merges of a $\mathbf{b}$ and a $\mathbf{t}$ are necessary and sufficient to create this configuration from the individual monomers.

2.2 Solvers

The problems addressed in this paper are NP-hard. To tackle this difficulty, we cast the problems as integer programs and use the publicly available IP solver SCIP [11].

[4] The term "polymer" is chosen to convey the concept of combining many atomic objects into a complex, but it is not necessarily a linear chain of repeated units.

We also use the open-source software OR-tools [15], which is a common front-end for SCIP [11], Gurobi [12], and a bundled constraint programming solver CP-SAT. Though we model our problems as IPs, we would also like to be able to solve for all feasible/optimal solutions rather than just one, which CP-SAT can do. This flexible front-end lets us switch seamlessly between the two types of solvers without significant alterations to the model.

We use the software package 4ti2 [1] to calculate Hilbert Bases as described in Section 4.

3 Computing stable configurations of TBNs

Section 3.1 formally defines the stable configurations problem. Section 3.2 explains our IP formulation of the problem. Section 3.3 shows empirical runtime benchmarks.

3.1 Finding stable configurations of TBNs

We consider the problem of finding the stable configurations of a TBN. Given a TBN $\mathcal{T}$, let $\Gamma_\mathcal{T}$ denote the set of all saturated configurations of $\mathcal{T}$.

Recall that a configuration $\gamma \in \Gamma_\mathcal{T}$ is defined as a partition of the monomers of $\mathcal{T}$ into polymers, so its elements $\mathbf{P} \in \gamma$ are polymers, i.e., multisets of monomers. For any $\gamma \in \Gamma_\mathcal{T}$, we define the corresponding *partial configuration* $\overline{\gamma} = \{\mathbf{P} \in \gamma : |\mathbf{P}| > 1\}$ that excludes polymers consisting of only a single monomer. Note that in the context of $\mathcal{T}$, the mapping $\gamma \mapsto \overline{\gamma}$ is one-to-one. We consider only partial configurations with finite-sized polymers. The notion of partial configuration will be useful in reasoning about TBNs with infinite monomer counts but finite size polymers, since all but finitely many monomers will be excluded from the partial configurations we consider.

Now we define the number of elementary merge operations required to reach a saturated configuration γ from the configuration of all singletons. We can calculate this directly as the difference in the counts of the monomers and polymers in the partial configuration, since each merge reduces the number of polymers by one:

$$m(\gamma) = \left(\sum_{\mathbf{P} \in \overline{\gamma}} |\mathbf{P}|\right) - |\overline{\gamma}| \tag{1}$$

We can then define the stable configurations as those saturated configurations that minimize the number of merges required to reach them from the all-singletons configuration.

$$\textsc{StableConfigs}(\mathcal{T}) = \{\gamma \in \Gamma_\mathcal{T} : (\forall \gamma' \in \Gamma_\mathcal{T})\ m(\gamma) \leq m(\gamma')\}$$

Note that $m(\gamma) = m(\overline{\gamma})$. Thus the StableConfigs problem may be equivalently posed as finding the set of saturated partial configurations $\overline{\gamma}$ that minimize $m(\overline{\gamma})$.

We now describe how to handle infinite counts. A configuration is saturated if and only if none of its starred sites (elements of $S^*_\mathcal{T}$) are exposed. Thus we focus on the subset of monomers that contain starred sites: the *limiting monomers* $\mathcal{T}_L = \{\mathbf{m} \in \mathcal{T} : \mathbf{m} \cap S^*_\mathcal{T} \neq \emptyset\}$. Limiting monomers are required to have finite count, whereas nonlimiting monomers (those with all unstarred sites) are allowed to be finite or infinite count. Our IP representation of a configuration explicitly accounts for *all* the limiting monomers, but only the nonlimiting monomers (in $\mathcal{T} \setminus \mathcal{T}_L$) in a polymer with some limiting monomer; implicitly every other nonlimiting monomer is unbound (i.e., in its own singleton polymer). This allows us to describe infinite configurations where all but finitely many of the infinite count monomers are unbound, guaranteeing that the number of merges counted in Equation (1) is finite.

3.2 Casting StableConfigs as an IP

3.2.1 Finding a single stable configuration

We first describe how to find a single element from StableConfigs($\mathcal{T}$) by identifying its partial configuration in $\mathcal{T}$. We begin by fixing an upper bound B on the number of non-singleton polymers in any partial configuration. If no *a priori* bound for B is available, conservatively take $B = |\mathcal{T}_L|$, the total number of limiting monomers.

3.2.1.1 Nonnegative integer variables

Assume an arbitrary ordering of the m monomer types $\mathbf{m}_1, \mathbf{m}_2, \ldots$. Our IP formulation uses $B \cdot m + B$ nonnegative integer variables describing the solution via its partial configuration:

- $\mathsf{Count}(\mathbf{m}, j)$: count of monomer type $\mathbf{m} \in \mathcal{T}$ in polymer $\mathbf{P}_j$ where $j \in \{1, 2, \ldots, B\}$
- $\mathsf{Exists}(j)$: false (0) if polymer $\mathbf{P}_j$ is empty, possibly true (1) otherwise, $j \in \{1, 2, \ldots, B\}$

▶ Remark 1. The constraints described below allow $\mathsf{Exists}(j) = 0$ even if polymer j is nonempty, even though the variables ultimately aim to count exactly the number of nonempty polymers (as $\sum_{j=1}^{m} \mathsf{Exists}(j)$). A false negative undercounts the number of polymers, overcounting the number of merges in Equation (1). However, the number of merges is being *minimized* by the IP. For a given setting of Count variables, the minimum is achieved (subject to the constraints) by setting each $\mathsf{Exists}(j) = 1$ *if and only* if polymer j is nonempty.

▶ **Example 2.** Recall the TBN of Figure 1. Suppose the TBN has 1 each of $\mathbf{m}_1 = \{a^*, b^*\}, \mathbf{m}_2 = \{a, b\}, \mathbf{m}_3 = \{a\}, \mathbf{m}_4 = \{b\}$, with upper bound $B = 2$ on the number of non-singleton polymers. $\mathcal{T}_L = \{\mathbf{m}_1\}$ since $\mathbf{m}_1$ is the only monomer with starred sites. The linear constraints (see below for details) do not require all copies of $\mathbf{m}_2, \mathbf{m}_3, \mathbf{m}_4 \in \mathcal{T} \setminus \mathcal{T}_L$ to be included in a polymer. The stable configuration on the right of Figure 1 (partition $\{\mathbf{m}_1, \mathbf{m}_2\}, \{\mathbf{m}_3\}, \{\mathbf{m}_4\}$) is represented in the IP by setting $\mathsf{Count}(\mathbf{m}_1, 1) = \mathsf{Count}(\mathbf{m}_2, 1) = 1$, $\mathsf{Count}(\mathbf{m}_3, 1) = \mathsf{Count}(\mathbf{m}_4, 1) = 0$ (monomers 1 and 2 are in polymer 1, but monomers 3 and 4 are not), and setting $\mathsf{Count}(\mathbf{m}_i, 2) = 0$ for $i = 1, 2, 3, 4$ (no monomers are in polymer 2), $\mathsf{Exists}(1) = 1$ (polymer 1 is non-empty), and $\mathsf{Exists}(2) = 0$ (polymer 2 is empty).

▶ **Example 3.** Suppose a TBN with the same monomer types as Example 2 has 3 of $\mathbf{m}_1$ and infinitely many of the remaining monomers (allowed since they are not limiting), with $B = 4$. The partial configuration where two copies of $\mathbf{m}_1$ are each bound to a single $\mathbf{m}_2$ (forming two polymers with two monomers each, as in the stable configuration of Figure 1), and the third $\mathbf{m}_1$ is bound to an $\mathbf{m}_3$ and an $\mathbf{m}_4$ (forming one polymer with three monomers, as in the rightmost non-stable saturated configuration of Figure 1) is represented in the IP by setting $\mathsf{Exists}(1) = \mathsf{Exists}(2) = \mathsf{Exists}(3) = 1$ and $\mathsf{Exists}(4) = 0$, and

$$\begin{array}{llll}
\mathsf{Count}(\mathbf{m}_1, 1) = 1, & \mathsf{Count}(\mathbf{m}_2, 1) = 1, & \mathsf{Count}(\mathbf{m}_3, 1) = 0, & \mathsf{Count}(\mathbf{m}_4, 1) = 0, \\
\mathsf{Count}(\mathbf{m}_1, 2) = 1, & \mathsf{Count}(\mathbf{m}_2, 2) = 1, & \mathsf{Count}(\mathbf{m}_3, 2) = 0, & \mathsf{Count}(\mathbf{m}_4, 2) = 0, \\
\mathsf{Count}(\mathbf{m}_1, 3) = 1, & \mathsf{Count}(\mathbf{m}_2, 3) = 0, & \mathsf{Count}(\mathbf{m}_3, 3) = 1, & \mathsf{Count}(\mathbf{m}_4, 3) = 1, \\
\mathsf{Count}(\mathbf{m}_1, 4) = 0, & \mathsf{Count}(\mathbf{m}_2, 4) = 0, & \mathsf{Count}(\mathbf{m}_3, 4) = 0, & \mathsf{Count}(\mathbf{m}_4, 4) = 0.
\end{array}$$

3.2.1.2 Linear constraints

Let $\mathcal{T}(\mathbf{m})$ denote the number of monomers of type $\mathbf{m}$ in the TBN $\mathcal{T}$. Recall that $\mathbf{m}(s)$ is the net count of site type $s \in S_{\mathcal{T}}$ in monomer type $\mathbf{m}$ (negative if $\mathbf{m}$ has more s^* than s).

$$\mathsf{Exists}(j) \leq 1 \qquad \forall j \in \{1, 2, \ldots, B\} \tag{2}$$

$$\sum_{j=1}^{B} \mathsf{Count}(\mathbf{m}, j) = \mathcal{T}(\mathbf{m}) \qquad \forall \mathbf{m} \in \mathcal{T}_L \tag{3}$$

$$\sum_{j=1}^{B} \mathsf{Count}(\mathbf{m}, j) \leq \mathcal{T}(\mathbf{m}) \qquad \forall \mathbf{m} \in \mathcal{T} \setminus \mathcal{T}_L \tag{4}$$

$$\sum_{\mathbf{m} \in \mathcal{T}} \mathsf{Count}(\mathbf{m}, j) \cdot \mathbf{m}(s) \geq 0 \qquad \forall j \in \{1, 2, \ldots, B\}, \forall s \in S_{\mathcal{T}} \tag{5}$$

$$\sum_{\mathbf{m} \in \mathcal{T}_L} \mathsf{Count}(\mathbf{m}, j) \geq \mathsf{Exists}(j) \qquad \forall j \in \{1, 2, \ldots, B\} \tag{6}$$

Constraint (2) enforces that Exists variables are Boolean. Constraints (3) and (4) intuitively establish "monomer conservation" in the partial configuration. Constraint (3) enforces that we account for every limiting monomer in $\mathcal{T}$. Constraint (4) establishes that for *non*-limiting monomers, we cannot exceed their supply (trivially satisfied for any infinite-count monomer); any leftovers are assumed to be in singleton polymers in the full configuration, but are not explicitly described by Count variables. Constraint (5) ensures that all polymers are self-saturated. Specifically, the count of site $s \in S_{\mathcal{T}}$ within polymer j must meet or exceed that of s^*. Lastly, Constraint (6) enforces that nonempty polymers contain at least one limiting monomer. Ideally, this constraint should enforce that if a polymer contains no monomers at all, then it cannot be part of the nonempty polymer tally; however, if the constraint were modeled in this way, the formulation would admit invalid partial configurations that include explicit singleton polymers.

3.2.1.3 Linear objective function

Subject to the above constraints, we minimize the number of merges needed to go from a configuration where all monomers are separate to a saturated configuration. For finite count TBNs, this is the number of monomers minus the number of polymers in the partial configuration. Equivalently (and applying to infinite TBNs), this is the sum over all nonempty polymers of its number of monomers minus 1. Formally, the IP minimizes (7):

$$\sum_{j=1}^{B} \left[\left(\sum_{\mathbf{m} \in \mathcal{T}} \mathsf{Count}(\mathbf{m}, j) \right) - \mathsf{Exists}(j) \right] \tag{7}$$

If polymer j is empty ($\sum_{\mathbf{m} \in \mathcal{T}} \mathsf{Count}(\mathbf{m}, j) = 0$), then constraint (6) forces $\mathsf{Exists}(j) = 0$; otherwise $\mathsf{Exists}(j) = 1$ minimizes (7). Thus the outer sum is over the nonempty polymers.

3.2.2 Finding all stable configurations

While an IP formulation for finding a single stable configuration is well-defined above, without modification it is ill-suited as a formulation to find *all* stable configurations. In addition, tightening the available constraints (e.g., enforcing $\mathsf{Exists}(j) \iff$ polymer j is nonempty, described below) provides a more robust framework to which to add custom constraints (e.g. specifying a fixed number of polymers).

3.2.2.1 IP to find optimal objective value, CP to enumerate optimal solutions

One straightforward improvement is to solve for the optimal value of the objective function using a dedicated IP solver such as SCIP, whose primal-dual methods exploit the underlying real-valued geometry of the search space to find an objective value more efficiently than Constraint Programming (CP) solvers such as CP-SAT. Then, use this optimal value to bootstrap the CP formulation, which is better suited to enumerating all solutions with a given objective value. This works particularly well in our experiments: use SCIP to solve the optimization problem (but SCIP has no built-in ability to enumerate all feasible solutions), then use CP-SAT (which takes longer than SCIP to find the objective value) to locate all feasible solutions to the IP obtaining the objective value found by SCIP.

3.2.2.2 Enforcing that Exists variables exactly describe nonempty polymers

Constraint (5) enforces that $\mathsf{Exists}(j) = 0$ if polymer $\mathbf{P}_j$ is empty, but it does not enforce the converse. However, when using CP-SAT with a *fixed* objective value, we can no longer rely on the minimization of Equation (7) to enforce that $\mathsf{Exists}(j) = 1 \iff \mathbf{P}_j$ is nonempty.

We add a new constraint to handle this. Let

$$C = 1 + \sum_{s \in S_{\mathcal{T}}} \sum_{\mathbf{m} \in \mathcal{T}} \mathcal{T}(\mathbf{m}) \cdot \mathbf{m}(s^*). \tag{8}$$

C is an upper bound on the largest number of monomers in a polymer in any valid partial configuration of $\mathcal{T}$ minimizing Equation (7). This corresponds to the worst case in which a single polymer contains *every* limiting monomer, and each starred site is bound to its own unique monomer. The following constraint enforces that if $\mathsf{Exists}(j) = 0$, then polymer $\mathbf{P}_j$ contains no monomers:

$$\sum_{\mathbf{m} \in \mathcal{T}_L} \mathsf{Count}(\mathbf{m}, j) \leq C \cdot \mathsf{Exists}(j) \qquad \forall j \in \{1, 2, \ldots, B\} \tag{9}$$

3.2.2.3 Eliminating symmetries due to polymer ordering

In the formulation of Section 3.2, many isomorphic solutions exist in the feasible region. For instance, one could obtain a "new" solution by swapping the compositions of polymers $\mathbf{P}_1$ and $\mathbf{P}_2$. The number of isomorphic partial configurations grows factorially with the number of polymers. Before asking the solver to enumerate all solutions, we must add constraints that eliminate isomorphic solutions. We achieve this by using the (arbitrary) ordering of the monomer types to induce a lexicographical ordering on the polymers, then add constraints ensuring that any valid solution contains the polymers in sorted order.

Sorting non-binary vectors in an IP is generally a difficult task (for instance, see [21]). The primary reason for this difficulty is that encoding the sorting constraints involves logical implications ($p \implies q$), which, being a type of disjunction ($\neg p$ OR q), are difficult to encode into a convex formulation described as a conjunction (AND) of several constraints. However, we do have an upper bound C on the values that the Count variables can take, making certain "large-number" techniques possible.

Intuitively, when comparing two lists of scalars (i.e., vectors) to verify that they are correctly sorted, one must proceed down the list of entries until one of the entries is larger than its corresponding entry in the other list. For as long as the numbers are the same, they are considered "tied". When one entry exceeds the corresponding other, the tie is considered "broken", after which no further comparisons need be conducted between the two vectors.

We introduce $B \cdot m$ new Boolean (0/1-valued) variables ($\mathsf{Tied}(\mathbf{m}_i, j)$ for each $i = 1, \ldots, m$ and $j = 1, \ldots, B$), that reason about consecutive pairs of polymers $\mathbf{P}_{j-1}$, $\mathbf{P}_j$. We describe constraints enforcing that for each $h \leq i$, $\mathsf{Tied}(\mathbf{m}_i, j) = 1 \iff \mathsf{Count}(\mathbf{m}_h, j-1) = \mathsf{Count}(\mathbf{m}_h, j)$.

Let C be defined as in (8). For simplicity of notation below, define the constants $\mathsf{Tied}(\mathbf{m}_0, j) = 1$ for all $j = 1, \ldots, B$. The meaning of the sorting variables is then enforced by the following constraints, which we define for $i \in \{1, 2, \ldots, m\}$ and $j \in \{2, 3, \ldots, B\}$:

$$\mathsf{Tied}(\mathbf{m}_i, j) \leq \mathsf{Tied}(\mathbf{m}_{i-1}, j) \tag{10}$$

$$\mathsf{Count}(\mathbf{m}_i, j-1) - \mathsf{Count}(\mathbf{m}_i, j) \leq C \cdot (1 - \mathsf{Tied}(\mathbf{m}_i, j)) \tag{11}$$

$$\mathsf{Count}(\mathbf{m}_i, j-1) - \mathsf{Count}(\mathbf{m}_i, j) \geq -C \cdot (1 - \mathsf{Tied}(\mathbf{m}_i, j)) \tag{12}$$

$$\mathsf{Count}(\mathbf{m}_i, j-1) - \mathsf{Count}(\mathbf{m}_i, j) \geq 1 - C \cdot \big(1 + \mathsf{Tied}(\mathbf{m}_i, j) - \mathsf{Tied}(\mathbf{m}_{i-1}, j)\big) \tag{13}$$

Intuitively, (10) enforces $\mathsf{Tied}(\mathbf{m}_i, j) \implies \mathsf{Tied}(\mathbf{m}_{i-1}, j)$: a tie in the current entry is only relevant if the tie was not resolved before. (11) and (12) together enforce $\mathsf{Tied}(\mathbf{m}_i, j) \implies \big(\mathsf{Count}(\mathbf{m}_i, j-1) = \mathsf{Count}(\mathbf{m}_i, j)\big)$: ties can only continue for as long as the corresponding entries are equal.

(13) enforces $\mathsf{Tied}(\mathbf{m}_{i-1}, j) \wedge \neg\mathsf{Tied}(\mathbf{m}_i, j) \implies \big(\mathsf{Count}(\mathbf{m}_i, j-1) > \mathsf{Count}(\mathbf{m}_i, j)\big)$: ties can only be broken if the tie was not broken previously and the current entries are ordered correctly. Thus any solution satisfying these constraints must sort the polymers.

3.3 Empirical running time measurements

For our empirical tests we use as a benchmark the autocatalytic TBN described in [2, Section 4.2.2 and Fig. 6]. This TBN features two large monomers of size n^2 in which n is a parameter in the design, as well as a variable number of additional monomers ("fuels") intended to be present in large excess quantities.

In addition to the formulation we give in this paper, we also tested a number of formulation variants, including the StableGen algorithm originally posed in [3] for solving the StableConfigs problem, justifying some of our design choices. "No Heuristic" performs a thorough accounting of all monomers (not just those needed to achieve saturation against the limiting monomers). "Labelled Monomers" assumes that the monomers are provided as a set, rather than a multiset. "Network Model" is a modification of StableGen with an alternate saturation constraint which does not require the explicit invocation of site-level bonds.

Each data point represents the average of three runs, and the solver was allowed to run for up to 100 seconds before a timeout was forced. Figure 2 (left) shows the runtimes as they increase with the parameter n, holding the count of each fuel at 2. Figure 2 (right) fixes $n = 3$ and shows the runtimes as they increase with the multiplicity of the fuel monomers. Note that our formulation can solve the case when fuels are in unbounded excess, while the variant formulations require bounded counts of all monomers.

Figure 2 (bottom) shows the runtime of our formulation in the case of unbounded fuels as it grows with increasing size parameter n. The rote specification of the benchmark problem is quadratic in n, and the empirical runtime growth appears polynomial with exponent close to this lower bound (2.71). However, we note that this is just a single family of problem instances, and we expect the runtime to be exponential in the worst case, since the problem is NP-hard.

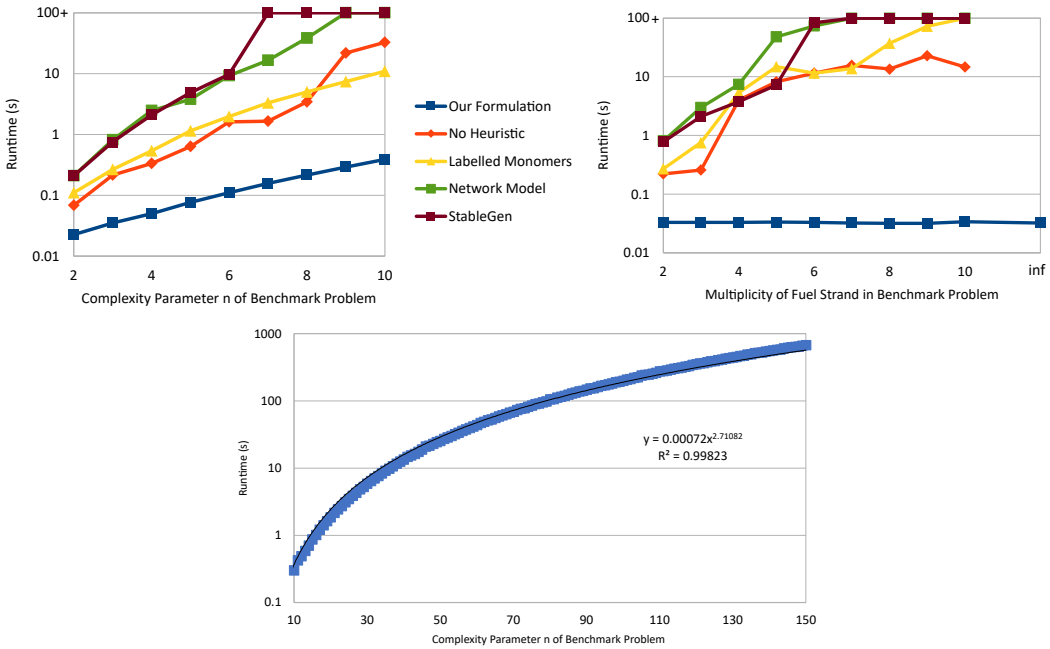

Figure 2 Empirical tests solving StableConfigs for our benchmark problem based upon its complexity parameter n (left and bottom), and the multiplicity of the unstarred "fuel" strands (right). Our formulation is tested against several variations on the approach (which are described in the text) and the StableGen algorithm from [3]. The TBN is parameterized by n and contains the monomers $G_n = \{x^*_{ij} : 1 \leq i, j \leq n\}$, $H_i = \{x_{ij} : 1 \leq j \leq n\}$ for all $1 \leq i \leq n$, and $V_j = \{x_{ij} : 1 \leq i \leq n\} \cup \{x_{ij} : j \leq i \leq n\}$ for all $1 \leq j \leq n$. See Fig. 6 from [2] for a detailed explanation of this TBN and its operation. The vertical axis is log scale. Points at the top of the scale timed out after 100 seconds. The alternate formulations cannot solve the instance in the case of infinite fuel strands.

4 Computing bases of locally stable configurations of TBNs

We now shift attention to *locally* stable configurations: those in which no polymer can be split without breaking a bond. Such a configuration may not be stable, but the only paths to create more polymers, without breaking any bonds, require first merging existing polymers (i.e., going uphill in energy before going down). The saturated configurations are precisely those obtained by merging polymers starting from some locally stable configuration. In this section we describe a technique for computing what we call the *polymer basis*: the (finite) set of polymers that can exist in locally stable configurations. In Section 4.1, we show that an algebraic concept called the *Hilbert basis* [7] characterizes the polymer basis. In Sections 4.2 and 4.3 we show how the polymer basis can be used to reason about TBN behavior.

4.1 Equivalence of polymer bases and Hilbert bases

We note that the connection between Hilbert bases and polymer bases is not particularly deep and does not require clever techniques to prove. Once the definitions are appropriately set up, the equivalence follows almost immediately. (Though we provide a self-contained proof.) The primary insight of this section is that casting TBNs in our IP formulation sets up the connection with Hilbert bases. Since highly optimized software exists for computing Hilbert bases [1], this software can be deployed to automate reasoning about TBNs.

Let M be a set of monomer types with $m = |M|$. Let $\mathcal{S}_M$ denote the *TBN schema of* M, the set of all TBNs containing only monomers from M, such that starred sites are limiting (i.e., such that saturated configurations have all starred sites bound). Let A_M be the *matrix representation* of the monomer types in $\mathcal{S}_M$, describing the contents of each monomer type: formally, the row-i, column-j entry of A_M is $\mathbf{m}_j(s_i)$, the net count of site type s_i in monomer type $\mathbf{m}_j$ (as an example, $\{a^*, b, a, a, a, c, c^*, c^*\}$ has net count 2 of a, 1 of b, and -1 of c). Formally, a TBN $\mathcal{T} \in \mathcal{S}_M$ if and only if $A_M\mathcal{T} \geq \mathbf{0}$.

Recall that $\Gamma_\mathcal{T}$ is the set of saturated configurations of the TBN $\mathcal{T}$, and that a polymer $\mathbf{P}$ is *self-saturated* if it has no exposed starred sites, i.e., $A_M\mathbf{P} \geq \mathbf{0}$. Define the *polymer basis* $\mathcal{B}_{\mathcal{S}_M}$ to be the set of all polymers $\mathbf{P}$ with the following properties:

- $(\exists \mathcal{T} \in \mathcal{S}_M)(\exists \alpha \in \Gamma_\mathcal{T})$ $\mathbf{P} \in \alpha$ (i.e., $\mathbf{P}$ appears in some saturated configuration of a TBN using only the monomer types from M.)
- There is no partition of $\mathbf{P}$ into two (or more) self-saturated polymers.

For example, consider the monomers $G = \{a^*, b^*, c^*, d^*\}$, $H_1 = \{a, b\}$, $H_2 = \{c, d\}$, $V_1 = \{a, c\}$, $V_2 = \{b, d\}$ and let $M = \{G, H_1, H_2, V_1, V_2\}$. The polymer basis $\mathcal{B}_{\mathcal{S}_M}$ is { $\{G, H_1, H_2\}$, $\{G, V_1, V_2\}$, $\{H_1\}$, $\{H_2\}$, $\{V_1\}$, $\{V_2\}$ }. All other self-saturated polymers are unions of these.

To show that polymer bases can be characterized by Hilbert bases, we must first define some additional terms. A *conical combination* of a set of vectors is a linear combination of the vectors using only nonnegative coefficients. An *integer conical combination* of a set of vectors is a conical combination of the vectors using only integer coefficients. A *(polyhedral) convex cone* $C = \{\lambda_1\mathbf{a}_1 + \cdots + \lambda_n\mathbf{a}_n : \lambda_1, \ldots, \lambda_n \geq 0\}$ is the space of all conical combinations of a finite set of vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ (and is said to be *generated* by $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$). C is *pointed* if $C \cap (-C) = \{\mathbf{0}\}$. A set of the form $\{\mathbf{x} \in \mathbb{R}^m : A\mathbf{x} \geq \mathbf{0} \text{ and } \mathbf{x} \geq \mathbf{0}\}$ is always a pointed convex cone [7].

A set is *inclusion-minimal* with respect to a property if it has no proper subset that satisfies the same property. The *Hilbert basis* of a pointed convex cone C is the unique inclusion-minimal set of integer vectors such that every integer vector in C is an integer conical combination of the vectors in the Hilbert basis. For example, the Hilbert basis of the convex cone generated (with nonnegative real coefficients) by $(1, 3)$ and $(2, 1)$ is $\{(1, 1), (1, 2), (1, 3), (2, 1)\}$; note that $\frac{2}{5} \cdot (1, 3) + \frac{4}{5} \cdot (2, 1) = (2, 2)$, which is not an integer combination of $(1, 3)$ and $(2, 1)$, but $2 \cdot (1, 1) = (2, 2)$.

Recall that the matrix-vector product $A_M\mathbf{P}$ gives the number of exposed sites of each type in the polymer, so that $A_M\mathbf{P} \geq \mathbf{0}$ iff the polymer is self-saturated (i.e. none of the starred sites are exposed).

We are then interested in vectors contained in $\{\mathbf{P} \in \mathbb{N}^m : A_M\mathbf{P} \geq \mathbf{0}\}$. Noting that $\mathbb{N}^m = \{\mathbf{P} \in \mathbb{R}^m : \mathbf{P} \geq \mathbf{0}\} \cap \mathbb{Z}^m$, we can equivalently state that we are interested in all integer vectors contained in the pointed convex cone $\{\mathbf{P} \in \mathbb{R}^m : A_M\mathbf{P} \geq \mathbf{0} \text{ and } \mathbf{P} \geq \mathbf{0}\}$.

▶ **Theorem 4.** *Let $\mathcal{S}_M$ be a TBN schema and let A_M be the matrix representation of its monomer types. Then the polymer basis $\mathcal{B}_{\mathcal{S}_M}$ of $\mathcal{S}_M$ is the Hilbert basis of $\{\boldsymbol{P} \in \mathbb{R}^m : A_M\boldsymbol{P} \geq \mathbf{0}$ and $\boldsymbol{P} \geq \mathbf{0}\}$.*

Proof. Note that the integer vectors in $\{\mathbf{P} \in \mathbb{R}^m : A_M\mathbf{P} \geq \mathbf{0} \text{ and } \mathbf{P} \geq \mathbf{0}\}$ are precisely the polymers that appear in saturated configurations of TBNs in $\mathcal{S}_M$, since $A_M\mathbf{P} \geq \mathbf{0} \iff$ polymer $\mathbf{P}$ is self-saturated, and $\mathcal{S}_M$ is defined to have starred sites limiting, so that a configuration is saturated if and only if each of its polymers is self-saturated.

We must show two properties to establish that $\mathcal{B}_{\mathcal{S}_M}$ is the Hilbert basis. First we must show that every polymer in saturated configurations of TBNs in $\mathcal{S}_M$ is a nonnegative integer combination of polymers in $\mathcal{B}_{\mathcal{S}_M}$. Next, to establish inclusion-minimality, we must show that no polymer can be removed from $\mathcal{B}_{\mathcal{S}_M}$ while satisfying the first property.

To see the first property, consider a polymer $\mathbf{P}$ in a saturated configuration of some TBN in $\mathcal{S}_M$. If it cannot be split into multiple self-saturated polymers, then we are done since it is in $\mathcal{B}_{\mathcal{S}_M}$ (it is the integer combination consisting of one copy of itself). Otherwise, we can iteratively split $\mathbf{P}$ into polymers $\mathbf{P}_1, \ldots, \mathbf{P}_k$ that themselves cannot be split into self-saturated polymers. Then $\mathbf{P} = \mathbf{P}_1 + \cdots + \mathbf{P}_k$.

To see the second property, suppose for the sake of contradiction that a polymer $\mathbf{P} \in \mathcal{B}_{\mathcal{S}_M}$ can be removed while maintaining the first property. By the definition of polymer basis, all polymers in $\mathcal{B}_{\mathcal{S}_M}$ are self-saturated, so $\mathbf{P}$ is self-saturated. If $\mathbf{P}$ can be removed from $\mathcal{B}_{\mathcal{S}_M}$ while maintaining the first property, then $\mathbf{P}$ is the nonnegative integer sum of some polymers remaining in $\mathcal{B}_{\mathcal{S}_M} \setminus \{\mathbf{P}\}$, and these polymers must also be self-saturated by virtue of being in $\mathcal{B}_{\mathcal{S}_M}$. This means that $\mathbf{P}$ can be partitioned into multiple self-saturated polymers, but then $\mathbf{P}$ does not satisfy the second constraint required to be in $\mathcal{B}_{\mathcal{S}_M}$ to begin with. ◀

4.2 Using the polymer basis to reason about TBN behavior

The complexity of computing the polymer basis in general can be very large; however, once it is calculated, reasoning about the stable configurations becomes a simpler task. For instance, in a previous example we had $\mathcal{B}_{\mathcal{S}_M} = \{\ \{G, H_1, H_2\}, \{G, V_1, V_2\}, \{H_1\}, \{H_2\}, \{V_1\}, \{V_2\}\ \}$. We can see from the above basis that in saturated configurations, G can only be present one of two unsplittable polymer types: $\{G, H_1, H_2\}$ or $\{G, V_1, V_2\}$, and we can optimize the number of polymers in a configuration by taking the other monomers as singletons (which is allowed, as these singletons are in the polymer basis). More generally, reasoning about stable configurations amounts to determining the number of each polymer type to use from the polymer basis so that the union of all polymers is the TBN, while using the maximum number of polymers possible. Our software can also solve for stable configurations in this way; specifically, for a TBN $\mathcal{T}$, it can calculate the polymer basis (abbreviated here as $\mathcal{B}$) and then solve for the stable configurations using the following IP:

$$\max_{\mathbf{c} \in \mathbb{N}^{|\mathcal{B}|}} \|\mathbf{c}\|_1 \text{ subject to } \sum_{i=1}^{|\mathcal{B}|} c_i \mathcal{B}_i = \mathcal{T}$$

Alternately, one can solve for the stable systems via an augmentation approach (see [7]).

If the goal is simply to solve the StableConfigs problem, we do not expect that solving for the stable configurations in this way will be more efficient than the previous formulation, as the time spent computing the Hilbert basis alone can require a great deal longer than solving via the formulation of the previous section. Instead, the true value of the basis is in its ability to describe *all* saturated configurations of a TBN.

For instance, in [2], the authors define an augmented TBN model in which a system can move between saturated configurations by two atomic operations: polymers can be pairwise merged (with an energetic penalty, i.e., higher energy) or they can be split into two so long as no bonds are broken (with an energetic benefit, i.e., lower energy; for instance $\{a, b\}, \{a^*, b^*\}, \{a\}, \{a^*\}$ can be split into $\{a, b\}, \{a^*, b^*\}$ and $\{a\}, \{a^*\}$, whereas $\{a\}, \{a^*\}$ cannot be split). Any saturated polymer not in the basis can split into its basis components without breaking any bonds. Thus the polymer basis contains all polymers that can form in a *local* minimum energy configuration, i.e., one where no polymer can split.

When designing a TBN, the designer will typically have a sense for which polymers are to be "allowed" in local energy minima. Proving that the system observes this behavior was not previously straightforward, but we can now observe that the TBN will behave ideally when its expected behavior matches its polymer basis.

4.3 A case example: Circular Translator Cascade

We now discuss an example of using the polymer basis to reason about a TBN's kinetic behavior, studying a TBN known as a *circular translator cascade*, first defined in [2]:

$$\{\{a,b,c\},\{b,c,d\},\{c,d,e\},\{d,e,f\},\{e,f,a\},\{f,a,b\},$$
$$\{a*,b*\},\{b*,c*\},\{c*,d*\},\{d*,e*\},\{e*,f*\},\{f*,a*\}\}$$

There are two stable configurations of this TBN, shown in Figure 3.

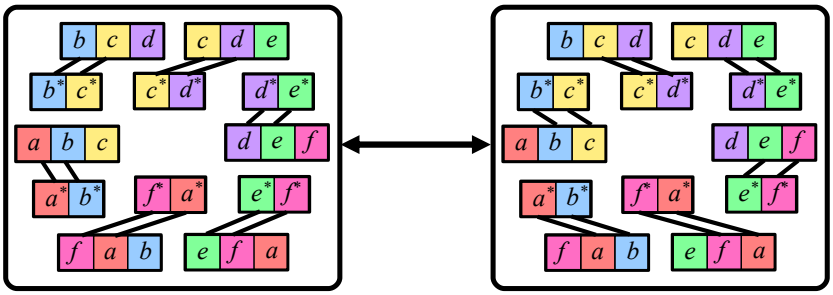

Figure 3 The two stable configurations of a variant of the circular translator cascade described in [2]. In the left configuration, the unstarred monomers are bound to their "left-side" companions (e.g. $\{a,b,c\}$ is bound to $\{a^*,b^*\}$), and in the right configuration, the unstarred monomers are bound to their "right-side" companions (e.g. $\{a,b,c\}$ is bound to $\{b^*,c^*\}$).

We consider now the "pathways" by which one of the stable configurations can "transition" to another. This process is described formally in [2]; here we give an intuitive description. Informally, we admit as atomic operations the ability for two polymers to merge or for one polymer to split into two polymers, so long as the resulting configuration remains saturated. In essence, these operations are modelling the physical phenomenon of solutes colocalizing in solution before reactions occur, specifically in dilute solutions in which enthalpic bond rearrangements occur on a timescale much faster than the timescale for entropic colocalization. If many polymers must be merged in some intermediate configuration to transition between stable configurations, then since each merge is individually unlikely, the successive merges required are exponentially unlikely i.e., a large *energy barrier* exists to transition between the configurations.

The design intention of this TBN is to have two stable configurations with a large energy barrier to transition between them. For the largest possible energy barrier, the transition should require the simultaneous merging of *all* of the polymers into a single polymer as an intermediate step. However, this is not the case for the TBN of Figure 3; the polymer basis gives insight into why. See [2, Section A.2] for an argument why more domain types and monomer types are required. We interpret this as a design error. We now explain how the error can be detected by reasoning about the polymer basis of the system, justifying that the automated computation of the polymer basis by our software enables one to automate some reasoning about the correct behavior of TBNs.

If it were true that the polymer basis contained only the 12 polymer types that are present in the two stable configurations of Figure 3, then that would be sufficient to prove the high energy barrier. To see why this is true, suppose there were a locally stable intermediate configuration that is part of a lower barrier transition. Since the configuration is locally stable, it is saturated, and none of its polymers can be partitioned into self-saturated polymers. By definition, the polymer basis should then contain all of the polymers present in this

intermediate configuration. However, all of the polymers in the basis have exactly two monomers, and so there must be 6 polymers in the intermediate configuration. The stable configurations also have 6 polymers, and so the intermediate configuration is also stable, but this contradicts that there are only two stable configurations.

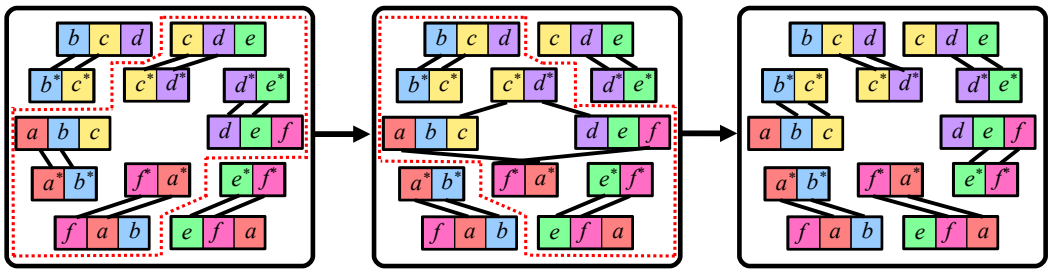

Figure 4 A counterexample to the claim that transitioning between stable configurations of this TBN requires the simultaneous merger of all monomers into a single polymer. Starting from the stable configuration on the left, by merging the four polymers within the red dotted outline, it is possible to re-arrange bonds and then split to the middle configuration. Then from the middle configuration, by merging the three polymers in the red dotted online, it is possible to re-arrange bonds and then split to the stable configuration on the right. Such intermediate configurations are evident by examining the elements of the polymer basis.

In fact, the polymer basis for this TBN has 57 entries (determined via our software), not 12, and we can use this basis to disprove the high energy barrier, i.e, to show that there is a sequence of merges and splits that transitions between the two stable configurations, without all monomers ever being merged into a single polymer. To discover a pathway that demonstrates the lower energy barrier, consider one unexpected entry in the polymer basis: $\mathbf{P} = \{\{a, b, c\}, \{d, e, f\}, \{c^*, d^*\}, \{f^*, a^*\}\}$. Its existence in the polymer basis tells us that there must be some saturated configuration that contains it. If we examine where these monomers were in one of the original stable configurations (Figure 3, left), we see that these were originally in polymers

$$\{\{a, b, c\}, \{a^*, b^*\}\}, \quad \{\{c, d, e\}, \{c^*, d^*\}\}, \quad \{\{d, e, f\}, \{d^*, e^*\}\}, \quad \{\{f, a, b\}, \{f^*, a^*\}\}.$$

From the starting configuration, if only these four polymers were merged, then they could then iteratively split into $\mathbf{P}$, $\{\{c, d, e\}, \{d^*, e^*\}\}$, and $\{\{f, a, b\}, \{a^*, b^*\}\}$. Since the latter two polymers are part of the target configuration, one could now greedily merge all polymers except for these latter two and then split into the target configuration. At no point in the interim were all polymers merged together into a single polymer. The resulting pathway is illustrated in Figure 4.

The difference between intended and actual barrier in this design becomes more pronounced if it is scaled up to include more site types and monomers. In [2] it is shown that by modifying the design, it is possible to achieve a linear energy barrier by using a quadratic number of site types.

5 Conclusion

In our investigation we observed that it was generally more efficient to solve Saturated-Configs by finding the optimal objective value using an IP solver as a first step, followed by using a CP solver on the same formulation with the objective value now constrained to the value found by the IP solver. Are further computational speedups possible by using IP as a callback during the CP search, instead of only in the beginning? How would one formulate the subproblems that would need to be solved in these callbacks?

In this paper we also note the value of polymer bases that are derived from the matrix containing the monomer descriptions. Such polymer bases can be used to describe all saturated configurations of a TBN, and so provide a valuable tool for analyzing potential behavior of a TBN when the model is augmented with rules that allow for dynamics. In practice, rather than discover unexpected behavior by calculating the polymer basis, a designer would instead like to begin with a set of behaviors and then create a TBN that respects them. Can we begin from verifiable polymer/Hilbert bases, encoding desired behavior, and transform them into TBN/DNA designs?

The full TBN model [2] can also be used to describe the regime in which there is a more modest tradeoff between the two energetic penalties of breaking bonds and merging complexes (i.e., saturation is not guaranteed). For example toehold-length binding sites in DNA strand displacement systems [4, 16, 19, 24, 25] are intended to dissociate over timescales comparable to association. Indeed, our software [13] includes an implementation of the StableConfigs formulation in which this relative weighting factor is included in the objective function. Under what conditions can a comparable polymer basis for such a system be found? Within the context of integer programming, it is known that by adding constraints to the design, one can reduce the complexity of finding/verifying Hilbert bases (and related *Graver bases*) [10], but it is not clear how to interpret these numeric constraints within the context of TBNs.

Satisfiability Modulo Theory (SMT) formulations have the ability to express TBN concepts (such as reachability along kinetic paths) without converting to a linear algebra framework, and existing solvers can solve SMT instances with surprising efficiency (for example, Z3 [8]). Can such solvers reason about TBNs directly within a reasonable time frame? Can they efficiently extract information beyond what is contained in the polymer basis?

References

1. 4ti2 team. 4ti2 – A software package for algebraic, geometric and combinatorial problems on linear spaces. `https://4ti2.github.io/hilbert.html`. URL: `https://4ti2.github.io`.
2. Keenan Breik, Cameron Chalk, David Haley, David Doty, and David Soloveichik. Programming substrate-independent kinetic barriers with thermodynamic binding networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(1):283–295, 2021. Special issue of invited papers from CMSB 2018.
3. Keenan Breik, Chris Thachuk, Marijn Heule, and David Soloveichik. Computing properties of stable configurations of thermodynamic binding networks. *Theoretical Computer Science*, 785:17–29, 2019.
4. Yuan-Jyue Chen, Neil Dalchau, Niranjan Srinivas, Andrew Phillips, Luca Cardelli, David Soloveichik, and Georg Seelig. Programmable chemical controllers made from DNA. *Nature Nanotechnology*, 8(10):755–762, 2013.
5. Yuan-Jyue Chen, Benjamin Groves, Richard A. Muscat, and Georg Seelig. DNA nanotechnology from the test tube to the cell. *Nature Nanotechnology*, 10:748–760, 2015.
6. Michele Conforti, Gérard Cornuéjols, Giacomo Zambelli, et al. *Integer programming*, volume 271. Springer, 2014.
7. Jesús A De Loera, Raymond Hemmecke, and Matthias Köppe. *Algebraic and geometric ideas in the theory of discrete optimization*. SIAM, 2012.
8. Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient smt solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340. Springer, 2008.
9. David Doty, Trent A Rogers, David Soloveichik, Chris Thachuk, and Damien Woods. Thermodynamic binding networks. In *DNA 2017: Proceedings of the 23rd International Meeting on DNA Computing and Molecular Programming*, pages 249–266. Springer, 2017.

10 Friedrich Eisenbrand, Christoph Hunkenschröder, Kim-Manuel Klein, Martin Koutecký, Asaf Levin, and Shmuel Onn. An algorithmic theory of integer programming. *arXiv preprint*, 2019. arXiv:1904.01361.

11 Gerald Gamrath, Daniel Anderson, Ksenia Bestuzheva, Wei-Kun Chen, Leon Eifler, Maxime Gasse, Patrick Gemander, Ambros Gleixner, Leona Gottwald, Katrin Halbig, Gregor Hendel, Christopher Hojny, Thorsten Koch, Pierre Le Bodic, Stephen J. Maher, Frederic Matter, Matthias Miltenberger, Erik Mühmer, Benjamin Müller, Marc E. Pfetsch, Franziska Schlösser, Felipe Serrano, Yuji Shinano, Christine Tawfik, Stefan Vigerske, Fabian Wegscheider, Dieter Weninger, and Jakob Witzig. The SCIP Optimization Suite 7.0. Technical report, Optimization Online, March 2020. URL: http://www.optimization-online.org/DB_HTML/2020/03/7705.html.

12 LLC Gurobi Optimization. Gurobi optimizer reference manual, 2020. URL: http://www.gurobi.com.

13 David Haley. Stable-TBN – a software package for computing the stable configurations of thermodynamic binding networks. URL: https://github.com/drhaley/stable_tbn.

14 Dionis Minev, Christopher M Wintersinger, Anastasia Ershova, and William M Shih. Robust nucleation control via crisscross polymerization of highly coordinated DNA slats. *Nature Communications*, 12(1):1–9, 2021.

15 Laurent Perron and Vincent Furnon. OR-tools. URL: https://developers.google.com/optimization.

16 Lulu Qian and Erik Winfree. Scaling up digital circuit computation with DNA strand displacement cascades. *Science*, 332(6034):1196–1201, 2011.

17 Paul W. K. Rothemund. Folding DNA to create nanoscale shapes and patterns. *Nature*, 440(7082):297–302, 2006.

18 Paul Shaw, Vincent Furnon, and Bruno De Backer. A constraint programming toolkit for local search. In *Optimization Software Class Libraries*, pages 219–261. Springer, 2003.

19 Niranjan Srinivas, James Parkin, Georg Seelig, Erik Winfree, and David Soloveichik. Enzyme-free nucleic acid dynamical systems. *Science*, 358(6369):eaal2052, 2017.

20 Chris Thachuk, Erik Winfree, and David Soloveichik. Leakless DNA strand displacement systems. In *DNA 2015: Proceedings of the 21st International Meeting on DNA Computing and Molecular Programming*, pages 133–153. Springer, 2015.

21 Andrew C Trapp and Oleg A Prokopyev. Solving the order-preserving submatrix problem via integer programming. *INFORMS Journal on Computing*, 22(3):387–400, 2010.

22 Boya Wang, Chris Thachuk, Andrew D Ellington, Erik Winfree, and David Soloveichik. Effective design principles for leakless strand displacement systems. *Proceedings of the National Academy of Sciences*, 115(52):E12182–E12191, 2018.

23 Damien Woods†, David Doty†, Cameron Myhrvold, Joy Hui, Felix Zhou, Peng Yin, and Erik Winfree. Diverse and robust molecular algorithms using reprogrammable DNA self-assembly. *Nature*, 567(7748):366–372, 2019. †joint first authors. doi:10.1038/s41586-019-1014-9.

24 David Yu Zhang and Georg Seelig. Dynamic DNA nanotechnology using strand-displacement reactions. *Nature chemistry*, 3(2):103–113, 2011.

25 David Yu Zhang and Erik Winfree. Control of DNA strand displacement kinetics using toehold exchange. *Journal of the American Chemical Society*, 131(47):17303–17314, 2009.