

Scalable Enumeration of Pareto-optimal Polymers

Finding the molecular complexes that matter at equilibrium — fast.

Archit Patil *UT Austin*, Minki Hhan *KAIST*, David Soloveichik *UT Austin*
DNA32 Conference

Funding | Schmidt Sciences · DOE · NSF
Acknowledgements | Soloveichik Lab



DNA molecular systems are getting big

Engineered DNA systems now include large architectures with hundreds of interacting strands in one solution.

To verify such systems, we need **equilibrium concentrations**:

- **intended** products should dominate under the designed input condition
- unintended products correspond to **leakage**

A recent DNA neural network:

700

distinct complexes

200+

strands · one test tube

Cherry & Qian, Nature 2025; Song & Qian, Nature 2025

The gold-standard tool doesn't scale

NUPACK

Analyzes complexes at the individual base-pair level.

- > Accurate, physically detailed
- > Expensive for large systems
- > (many strands / large complexes)

Dirks et al. 2007; Fornace et al. 2025

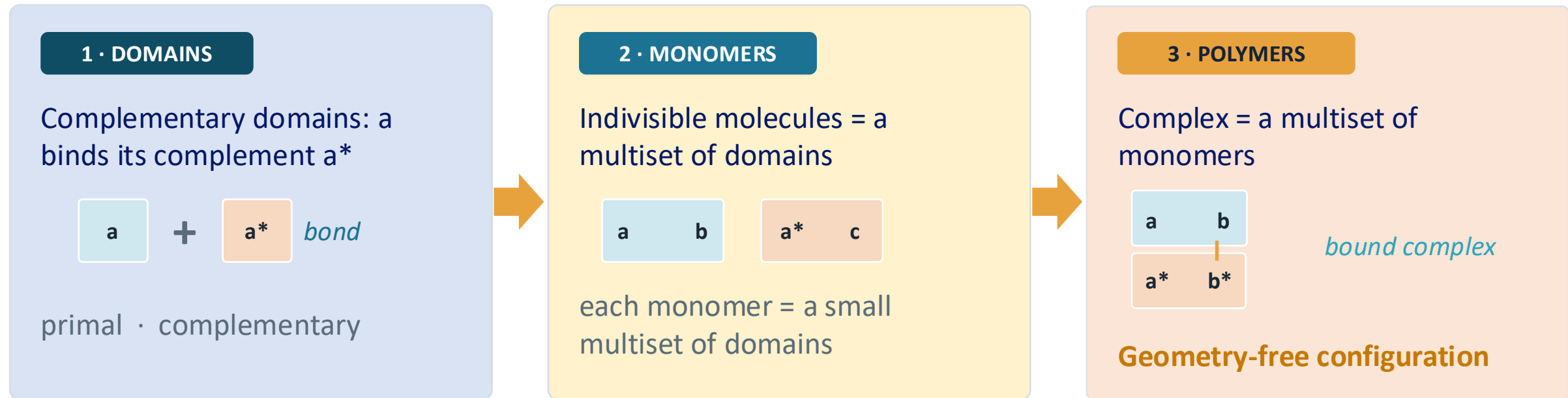


Zoom out

Work at a coarser level of abstraction —

- Treat specific sequences which can bind together as **complementary domains**.
- > Fewer, logical building blocks
- > Tractable for large designs

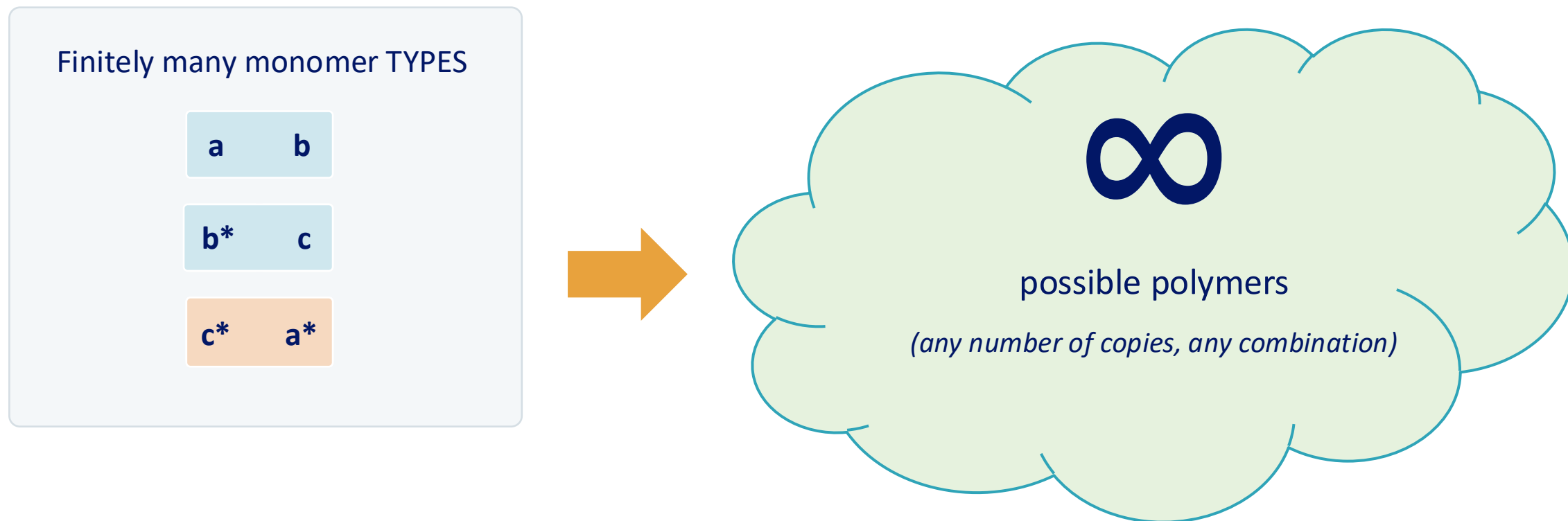
Domain-Monomer System



New here: we allow **unsaturated** binding. Not every possible bond must form, allowing enthalpy-entropy tradeoff.

Comparable to TBNs — Doty, Rogers, Soloveichik, Thachuk & Woods 2017; Breik et al. 2019.

A finite set of monomers -> infinitely many complexes



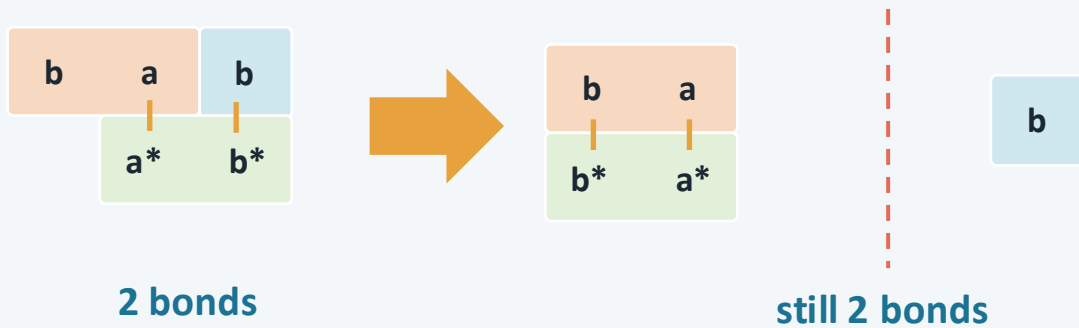
Yu et al. (COFFEE) 2025

Equilibrium solvers (NUPACK, COFFEE) need a finite candidate LIST. **But which candidates are likely to be present at equilibrium?**

The idea: Pareto-optimality

A polymer is **Pareto-suboptimal** if you can split it into two or more non-interacting parts.

Pareto-suboptimality:



Same enthalpy ($\Delta H \approx 0$)

+

More entropy ($\Delta S > 0$)



lower free energy ($\Delta G < 0$)

A **Pareto-optimal** polymer can't be split for free. There's an enthalpy-entropy tradeoff in every possible split.

Why is this justified?

DISCRETE MODEL

Minimum free-energy configurations

- No Pareto-suboptimal polymer ever appears in a minimum-free-energy configuration.

CONTINUOUS MODEL

Equilibrium concentrations

Total concentration of ALL suboptimal polymers is tiny:

$$C_{\text{sub}} \leq C_{\text{opt}}^2 / (1 - C_{\text{opt}})$$

In dilute solution, negligible vs. the optimal polymers.

- Restricting to Pareto-optimal polymers does not lose out any of the relevant polymers at equilibrium. (*Proofs in paper*)

How many are left? Finitely many and it's a Hilbert basis

What's a Hilbert basis?

The minimal set of **indecomposable building blocks** that generate every non-negative integer solution of a linear system.

Linear system: $S(A) = \{x \in \mathbb{N}^n : Ax=0\}$.

$H(A) = \{h \in S(A) \setminus \{0\} : \text{there do not exist nonzero } u, v \in S(A) \text{ such that } h = u + v\}$.

Proven to be finite for a pointed cone.

Our characterization

$$P^* = \pi(H) \setminus \{0\}$$

Pareto-optimal polymers = projection of a Hilbert basis

How: augment the system with unit monomers that neutralize dangling domains; indecomposable neutral solutions project down to exactly P^* .

Contribution: prior work tied Hilbert bases to the *saturated* TBN model; our neutralization extends it to the **unsaturated** setting.

Extends Haley & Doty, DNA 27 (2021). Hilbert-basis finiteness: Gordan's lemma, Gordan 1873

The catch: finite $\neq$ tractable

#P-hard

Just counting Hilbert-basis elements is #P-hard.

Hermann et al. 1999

Exponential

A tiny matrix can have an exponentially large basis.

see paper, App. B

>1000 s

7-module cascade: full computation takes over 1000 s.

this work, §7

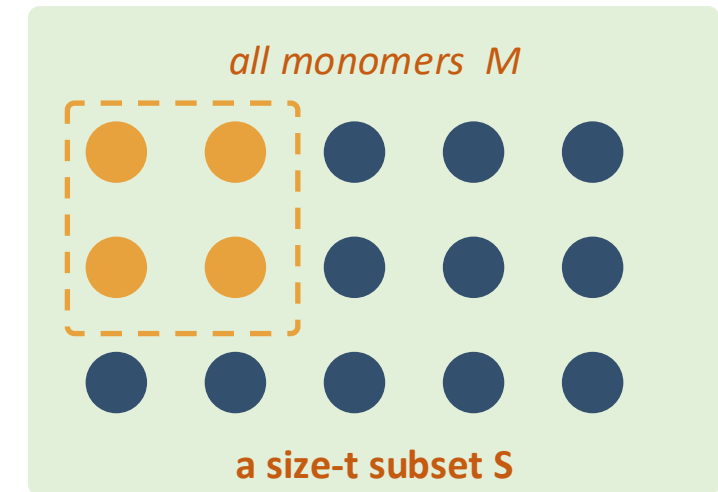
Fix 1: bound the number of monomer types in a polymer

Only enumerate polymers using at most t distinct monomer types.

$$P_t^* = \{p \in P^* : | \text{supp}(p) | \leq t\} \subseteq P^*$$

Larger t -> recover more, slower. **Smaller t** -> faster, less complete.

Compute P_t^* from small subsystems



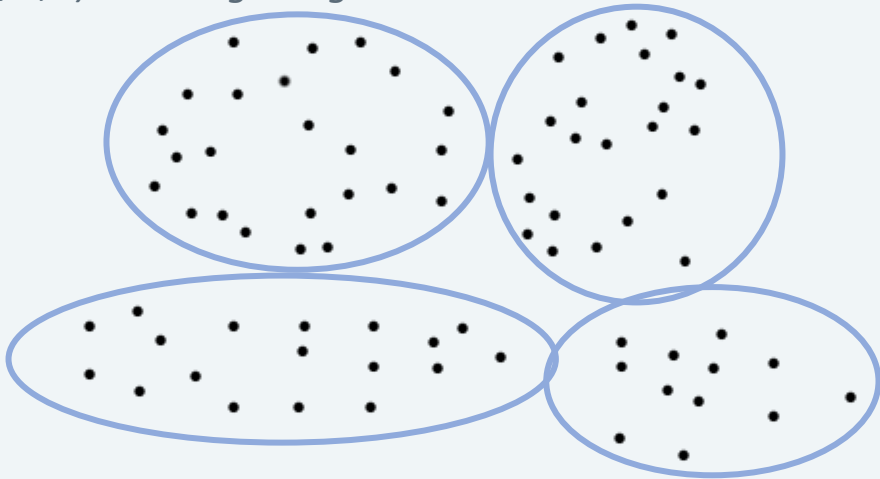
each subset -> one SMALL Hilbert basis

Fix 2: Covering design method

Naive: one Hilbert basis per t -subset $\rightarrow C(n, t)$ computations. Runtime increases combinatorically!

Key insight: computing on a bigger block of size k reveals **every** t -subset inside it.

(n, k, t) -covering design



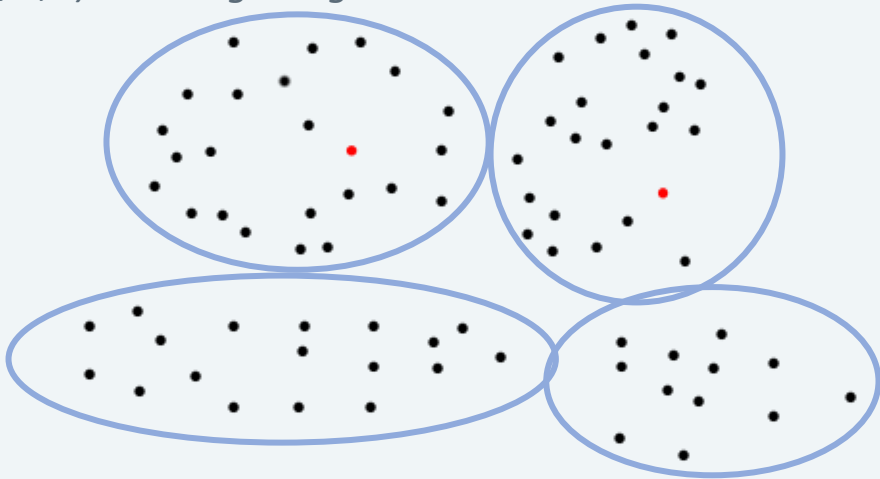
Covering designs: Schönheim 1964; Gordon, Patashnik & Kuperberg 1995; La Jolla Covering Repository (Gordon).

Fix 2: Covering design method

Naive: one Hilbert basis per t -subset $\rightarrow C(n, t)$ computations. Runtime increases combinatorically!

Key insight: computing on a bigger block of size k reveals **every** t -subset inside it.

(n, k, t) -covering design

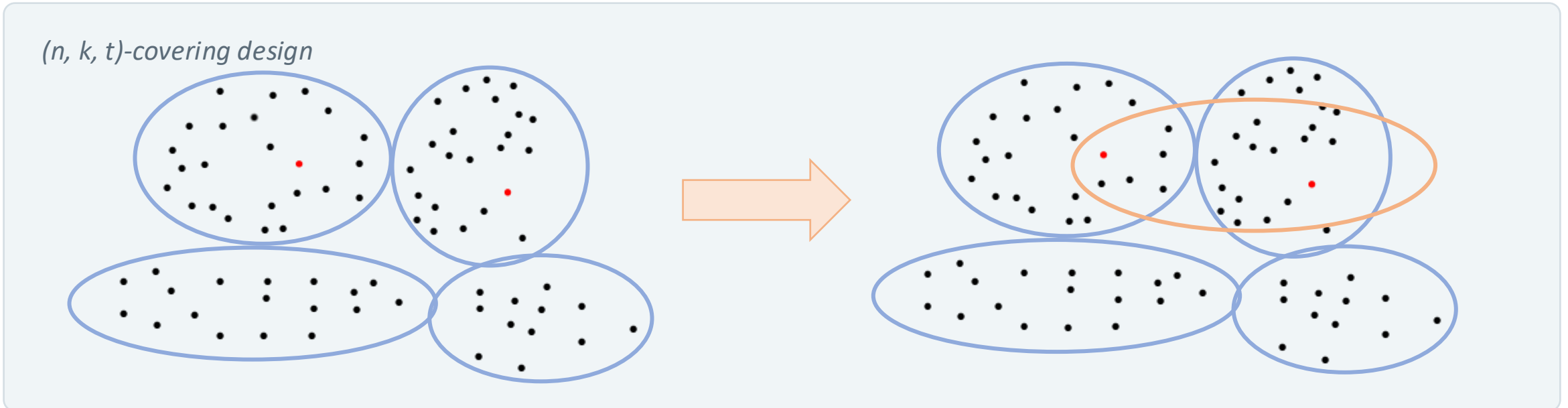


Covering designs: Schönheim 1964; Gordon, Patashnik & Kuperberg 1995; La Jolla Covering Repository (Gordon).

Fix 2: Covering design method

Naive: one Hilbert basis per t -subset $\rightarrow C(n, t)$ computations. Runtime increases combinatorically!

Key insight: computing on a bigger block of size k reveals **every** t -subset inside it.



Covering designs: Schönheim 1964; Gordon, Patashnik & Kuperberg 1995; La Jolla Covering Repository (Gordon).

Reducing the search space without losing the target

What the method returns:

$$P_t^* \text{ (t bounded pareto-optimal set)} \subseteq P_{k,t} \text{ (covering design algorithm output)} \subseteq P^* \text{ (set of pareto-optimal polymers)}$$

never lose a real t-bounded polymer

never invent a non-optimal one

Choosing the block size k

- Bigger k -> fewer blocks, but each Hilbert basis costs more.

We pick it automatically: a cheap 100-iteration probe estimates the runtime for each k

Does it work? Order-of-magnitude speedups

7-module AND cascade

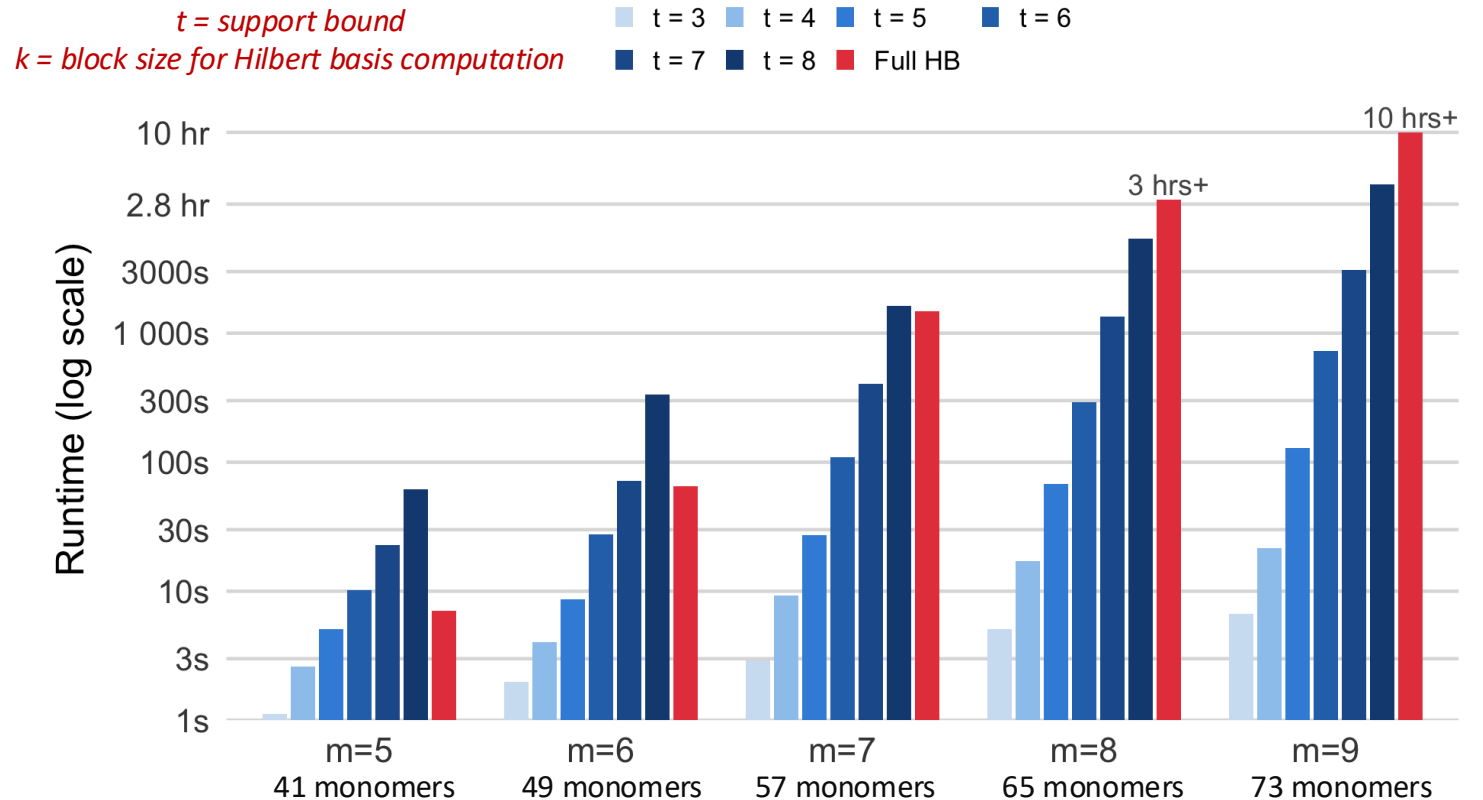
1105 s

full Hilbert basis

↓
24 s

covering design $(t,k)=(5,25)$

≈ 46× faster



AND gate linear cascade, Normaliz 3.11.1, COFFEE. Full-HB values for $m \geq 8$ are truncated/terminated runs. This work, §7.2; system from Wang et al., Sci. Adv. 2026.

Does it work? Order-of-magnitude speedups

7-module AND cascade

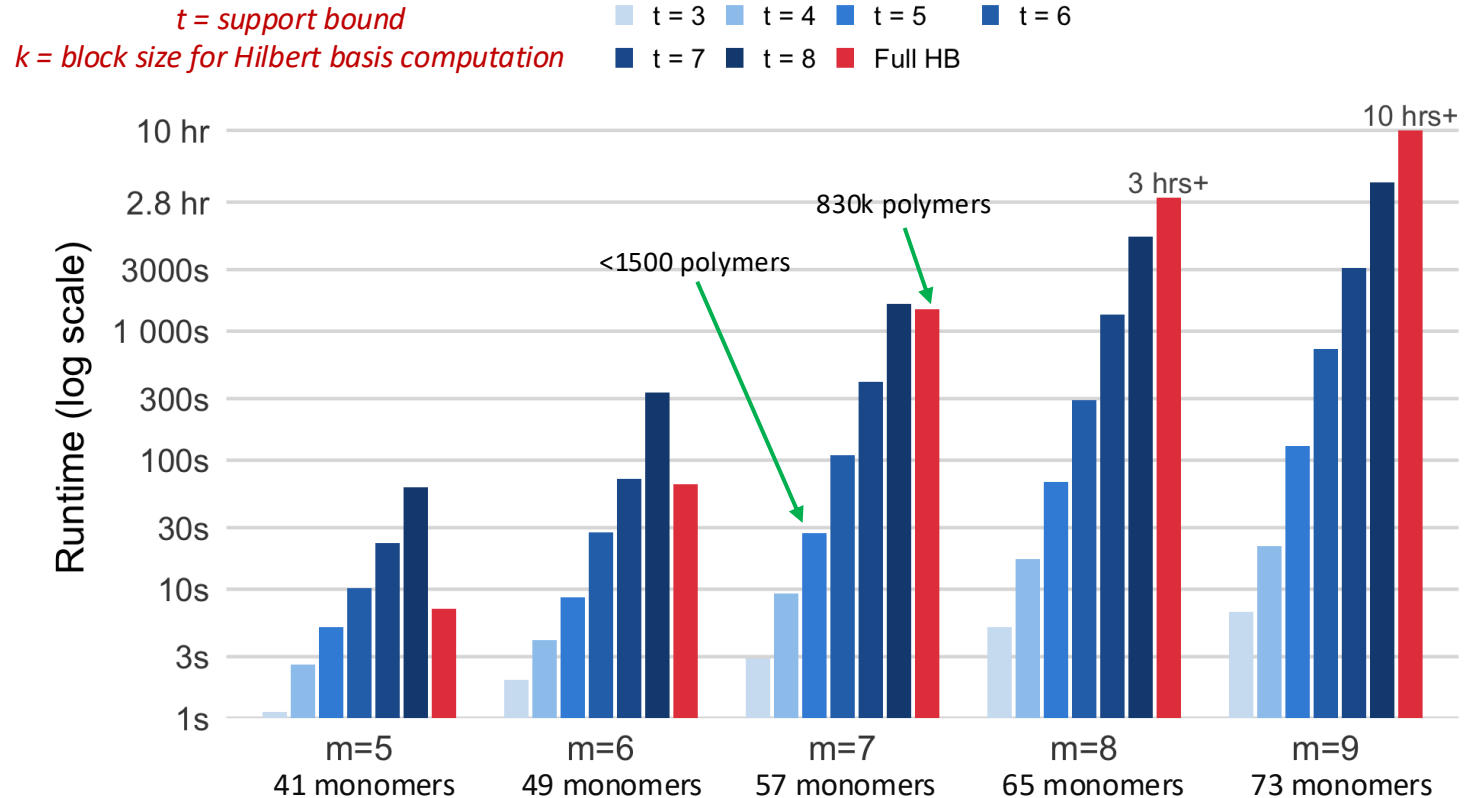
1105 s

full Hilbert basis

24 s

covering design $(t,k)=(5,25)$

$\approx 46\times$ faster

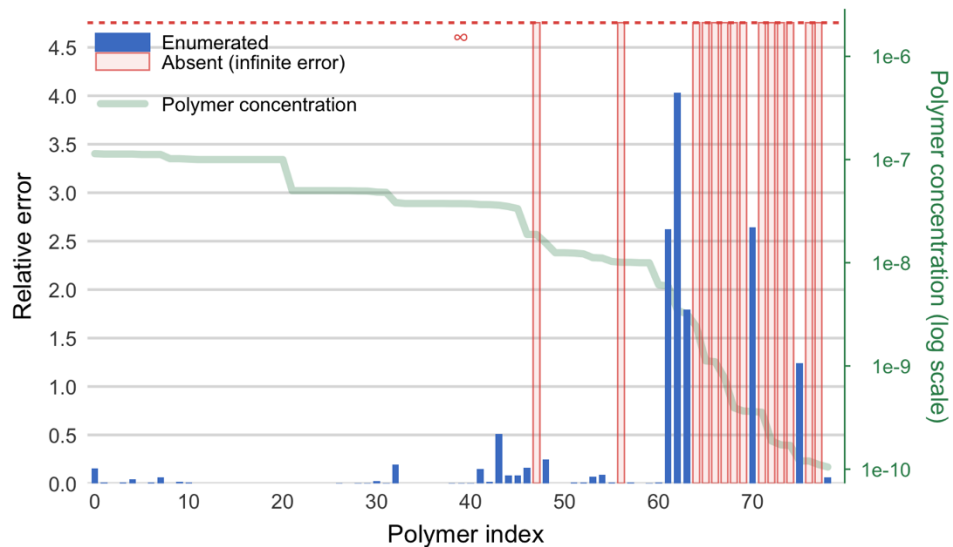


AND gate linear cascade, Normaliz 3.11.1, COFFEE. Full-HB values for $m \geq 8$ are truncated/terminated runs. This work, §7.2; system from Wang et al., Sci. Adv. 2026.

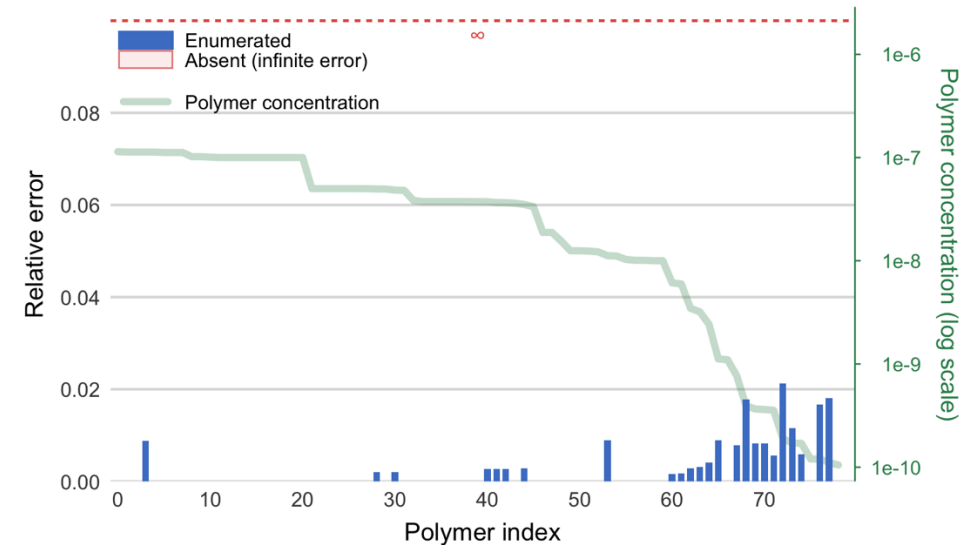
Fast! But did we drop what matters?

We sped up by *ignoring* candidates. Did we lose important polymers?

t (support bound) = 3



t = 5



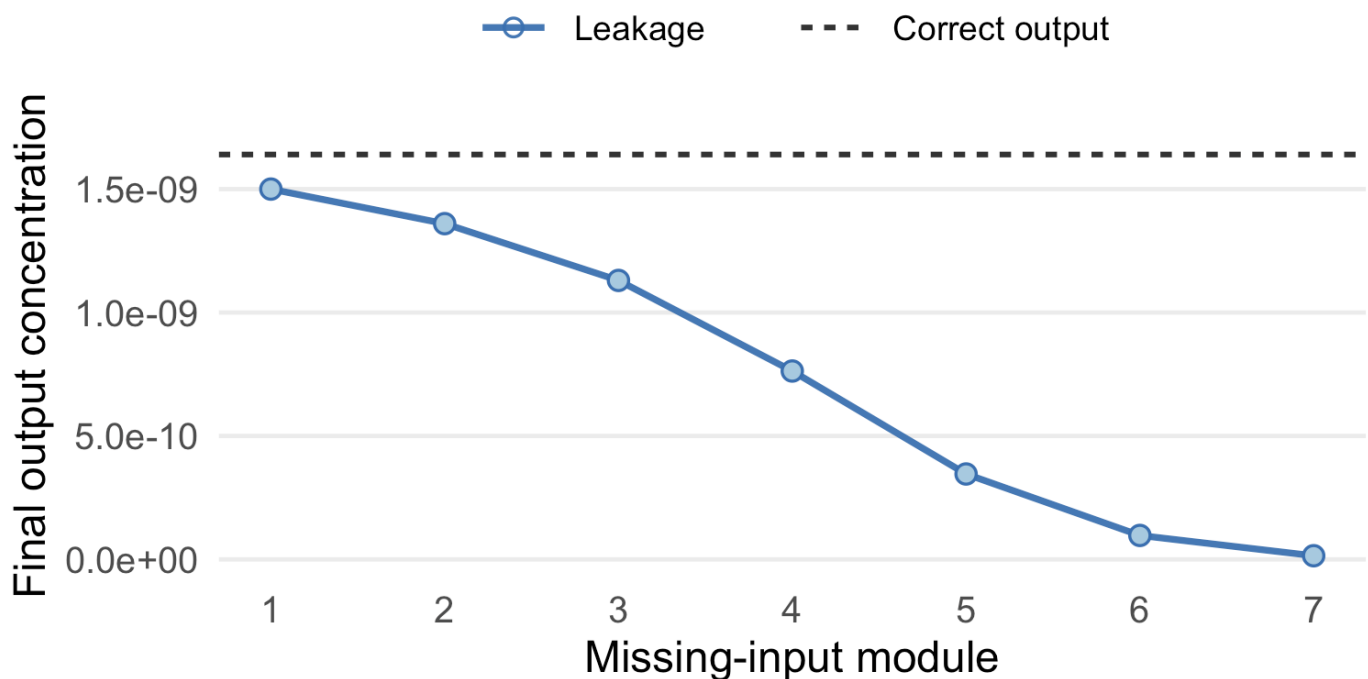
- ✗ Misses some equilibrium-relevant polymers
- ✗ Larger relative concentration errors
- ✓ Only low-concentration ones missed. Overall picture still fair

7-module cascade, Regime B leakage, $E = -20$ kcal/mol. Concentrations via COFFEE. This work, §7.3.

Applications: leakage analysis at scale

Leakage = unintended output formed through spurious pathways.

Remove one input, measure final-output leakage (7-module cascade):



Observation

- Removing an input from early module leaks the most.
- Early leakage propagates & amplifies down the cascade.

Regime B, $E = -10$ kcal/mol; curve is schematic of the reported trend. This work, §7.4; cascade from Wang et al. 2026.

Limitations

Large polymers

When designs expect very large complexes (e.g. the scaffolded-DNA computer), support-bounding by monomer types is less effective.

Stérin et al. 2025

Covering-design size

For very large $|M|$, the dynamic-programming construction isn't guaranteed minimal and may not scale optimally.

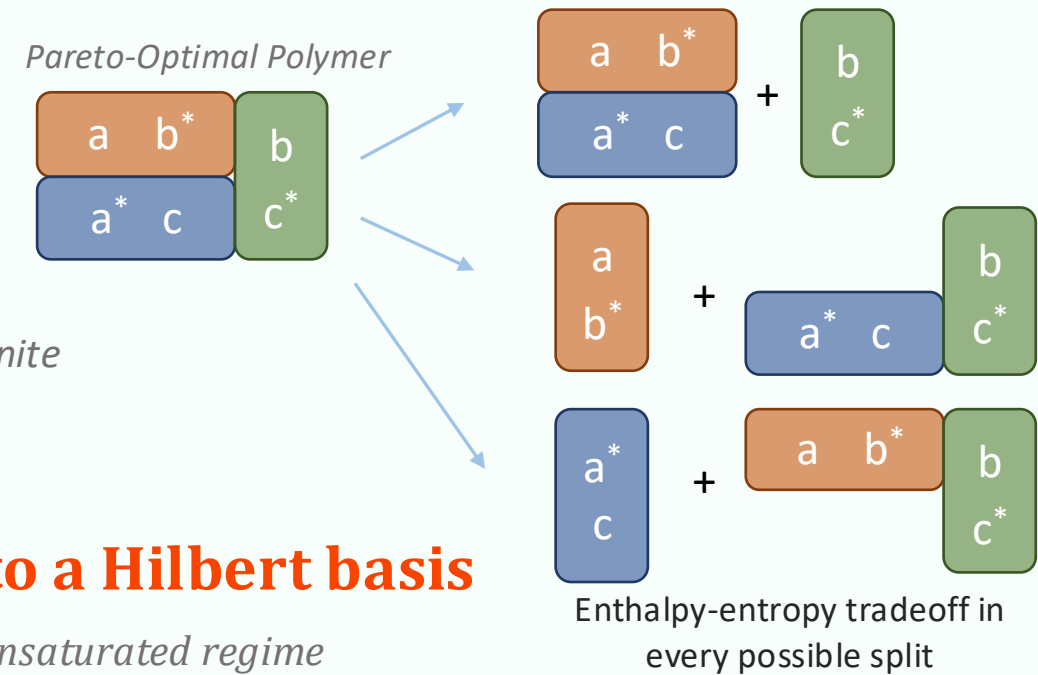
Gordon et al. 1995

Open question: beyond Pareto-optimality, what other structural properties could shrink the candidate set further while keeping every equilibrium-relevant polymer?

Takeaways

1. Pareto-optimality

Using thermodynamically justified idea to bring down infinite search space to finite



2. Set of Pareto-optimal polymers correspond to a Hilbert basis

Computing the set using an analogous mathematical object — extending to unsaturated regime

Augmented domain monomer system $A \rightarrow \text{Cone } C: Ax = 0 \rightarrow H(C) \rightarrow \pi(H(C)) \equiv \text{Pareto optimal Polymers}$

Hilbert Basis Projection back to the original system space

3. ...and we made it fast

Leveraging support based bound and covering design to bring down the finite (but large) search space drastically

