THE GREEDY SUPERSTRING RATIO AT SMALL STRING LENGTHS: $\rho_8, \rho_9, \rho_{11} > 2$ AND $\rho_4, \rho_5 > 15/8$

KOREA SUPERINTELLIGENCE LABS

ABSTRACT. Given a finite set S of strings, GREEDY repeatedly replaces a pair of distinct elements of maximum overlap by their overlap merge; ties are broken arbitrarily, and ρ_k denotes the supremum of $|g|/\mathrm{OPT}(S)$ over all instances $S \subseteq \Sigma^k$ and all outputs g of a legal execution. The greedy superstring conjecture, that GREEDY never outputs more than twice the optimum, stood for nearly four decades and was refuted on 1 September 2026 by Shibata, who proved $\rho_k \geq (9k+2)/(4k+4)$ for every *even* $k \geq 10$, hence $23/11$ at $k = 10$ and $9/4$ in the limit. We reproduce his refutation as an explicit instance of 198 strings of length 10 with a proved optimum, and then extend it in two directions his construction does not reach. First, it refutes the conjecture already at string length 8: an instance of 252 strings of length 8 has optimum exactly 259 and admits a legal greedy execution of output length $519 > 2 \cdot 259$. Since $\rho_k \leq 2$ is a theorem for $k \leq 4$, string length 8 is now the smallest at which the conjecture is known to fail. Second, it fails at *odd* string lengths, which the source's construction — written for $k = 2s$ throughout — cannot address at all: at $k = 9$ an instance of 336 strings has optimum 344 and a legal greedy output of length 689, and at $k = 11$ an instance of 250 strings has optimum 260 and a legal greedy output of length 521. In each case we also exhibit the sharp boundary instance one block smaller, which lands exactly on twice the optimum, or one letter below it. The paper also retains, from its first version, the refutation of the Cazaux–Rivals conjecture $\rho_k = 2 - 1/k$ at the two string lengths $k = 4$ and $k = 5$ left open by Chukhin, Kulikov, Mihajlin and Smal: $\rho_4 \geq 437/233 > 15/8$ and $\rho_5 \geq 325/173 > 15/8$, in particular $\rho_4 > 9/5 = \rho_3$. In every instance the optimum is proved by a counting bound together with an exhibited witness, so no search over superstrings is involved; the executions are certified merge by merge. Every statement about an individual instance made in a theorem or proposition below has been checked by the Lean 4 kernel; the remarks that report simulations instead are marked as such, and the final section says exactly what the kernel does and does not cover. Section 1.2 corrects two sentences of the first version of this paper.

1. INTRODUCTION

1.1. **Superstrings and the greedy algorithm.** Let Σ be a finite alphabet. A string w is a *common superstring* of a finite set S of strings if every element of S occurs in w as a contiguous block, and $\mathrm{OPT}(S)$ denotes the length of a shortest common superstring. Computing OPT is NP-hard, and the algorithm that has dominated the subject for forty years is the simplest one available. For strings p, q write

$$\mathrm{ov}(p,q) \;=\; \max\{\, \ell \;:\; \text{the length-}\ell \text{ suffix of } p \text{ equals the length-}\ell \text{ prefix of } q \,\},$$

Date: September 3, 2026.

and let $p \odot q$ be the *overlap merge*, that is p followed by all of q except its first $\mathrm{ov}(p,q)$ letters. Greedy is:

> while more than one string is live, choose an ordered pair of distinct live strings (p,q) with $\mathrm{ov}(p,q)$ maximum, and replace p and q by $p \odot q$.

Ties are broken arbitrarily, so Greedy is a nondeterministic algorithm and $\textsc{Greedy}(S)$ is a *set* of outputs, each of them a common superstring of S. The *greedy conjecture* asserted that $|g| \le 2\,\mathrm{OPT}(S)$ for every S and every $g \in \textsc{Greedy}(S)$. It dates from the late 1980s. Blum, Jiang, Li, Tromp and Yannakakis [9] state it in the form used here — a common conjecture, in their words, that Greedy "produces a superstring of length at most two times optimal" — and attribute it there to Storer [10], to Tarhio and Ukkonen [11] and to Turner [12]; those same four sources are the ones [2] cites for it.

The conjecture is false. It was refuted on 1 September 2026 by Shibata [1], who exhibited, for every even $k \ge 10$, a family of instances all of whose strings have length k, together with legal greedy executions on them whose ratios increase to $(9k+2)/(4k+4)$; that value is $23/11 = 2.0909\ldots$ at $k = 10$ and tends to $9/4$ as $k \to \infty$. The present paper takes that refutation as its starting point and pushes it down to string length 8 and across to odd string lengths, neither of which [1] reaches.

For $k \ge 2$, *k-SCS* is the restriction of the problem to instances $S \subseteq \Sigma^k$, all input strings having the same length k, and

$$\rho_k \;=\; \sup\Big\{ \frac{|g|}{\mathrm{OPT}(S)} \;:\; S \subseteq \Sigma^k \text{ finite},\ g \in \textsc{Greedy}(S) \Big\}. \tag{1}$$

The restriction of the greedy conjecture to k-SCS is the assertion $\rho_k \le 2$; it is that restricted assertion, one string length at a time, that the theorems below decide. Two features of (1) matter throughout: the supremum is over *all* legal executions, so a lower bound on ρ_k is witnessed by a single instance together with a single adversarial tie-breaking; and inputs of a common length are automatically substring-free, so the variant of Greedy that first discards strings contained in others begins with the same instance — that no live string is ever contained in another along the executions exhibited below is recorded in Section 12.

1.2. **Corrections to the first version of this paper.** Two sentences of the first version of this paper, dated 30 August 2026, are corrected here. The first follows from [1]; the second from [3], which we reached through the bibliography of [1].

First, that version wrote of the greedy conjecture, in its Section 1.1, "The conjecture is open." It is not open: it is false, by [1], and this is the starting point of Sections 8–10 below. The refutation is reproduced here as Theorem 19 and extended in Theorems 21, 24 and 26. No theorem of the first version is affected. Every one of them is an assertion about an explicit instance of 4-SCS or 5-SCS, and every one of them stands unchanged as Theorems 12–15 below; what they refute is the Cazaux–Rivals conjecture $\rho_k = 2 - 1/k$ of (3) — a statement about the *exact value* of ρ_k, which at $k = 4$ and $k = 5$ is a separate question from the greedy conjecture. At those two string lengths nothing here bounds ρ_k from above: $\rho_4 \le 2$ is a theorem [6] but the exact value is unknown, and at $k = 5$ even $\rho_5 \le 2$ is open.

Second, the display (4) below, which the first version reproduced from the introduction of [2] as the state of the art, records $\rho_k \le 3.396$ for $k \ge 6$; that entry is

the bound of Englert, Matsakis and Veselý [4], which is what [2] cites for it. It is superseded: Chukhin, Kulikov, Mihajlin and Smal [3] prove that the greedy ratio is at most 3, so $\rho_k \le 3$ for every k. The entry was current when [2] printed it on 24 August 2026, and it was the same four authors who superseded it three days later: [3] was submitted to ECCC on 27 August 2026, and its record there gives 30 August 2026 as the date of publication — the date of the first version of this paper, whose searches did not find it. We keep (4) below because it is quoted, as what [2] printed, and record the corrected picture in Table 1.

1.3. **Two conjectures, and the state of the art.** Cazaux and Rivals [5] proved

$$2 - \frac{1}{k} \;\le\; \rho_k \;\le\; \min\Big\{\frac{k+1}{2},\; \frac{7}{2}\Big\}, \tag{2}$$

the lower bound by an explicit series of instances (their Proposition 2) and the upper bound as their Theorem 3, and immediately after it they wrote:

> "Theorem 3 suggests a more precise version of the greedy conjecture: the superstring approximation ratio of greedy on r-SSP is $2 - 1/r$."

We refer to the assertion

$$\rho_k \;=\; 2 - \frac{1}{k} \tag{3}$$

as the Cazaux–Rivals conjecture. It is a strengthening of the greedy conjecture at every k, and the two are refuted by different arguments at different string lengths, so they have to be kept apart.

Chukhin, Kulikov, Mihajlin and Smal [2] refuted (3) at both ends of the range. They prove $\rho_k \ge 2$ for every $k \ge 6$, so that the conjectured value is wrong there by a whole $1/k$, and they determine $\rho_3 = 9/5$ exactly, against the conjectured $5/3$. Their introduction records the resulting state of the art as

$$\begin{gathered} 1.66 \le \rho_3 \le 2, \qquad 1.75 \le \rho_4 \le 2.5, \qquad 1.8 \le \rho_5 \le 3, \\ 2 - \tfrac{1}{k} \le \rho_k \le 3.396 \quad (k \ge 6), \end{gathered} \tag{4}$$

the lower entries being the Cazaux–Rivals values $2 - 1/k$, with $5/3$ printed as 1.66; the upper entry for $k \ge 6$ is superseded, as explained in Section 1.2. Neither of their two theorems reaches $k = 4$ or $k = 5$, and the abstract of the current version of [2] claims only to be "completely characterizing the worst-case behavior of the greedy algorithm in case of strings of length 3". Those two string lengths are the subject of Sections 3–6.

Remark 1. The entry $\rho_4 \le 2.5$ of (4) is the generic bound $(k+1)/2$ of (2). In fact $\rho_4 \le 2$ has been known since Kulikov, Savinov and Sluzhaev [6], whose abstract reads, in full, "In this short note, we prove that the greedy conjecture for the shortest common superstring problem is true for strings of length 4"; Cazaux and Rivals report the same result as "in the case where all input words have length 4 (for 4-SSP) the greedy algorithm achieves a superstring ratio of at most 2" [5], and Nikolaev's survey lists it as "GC holds for strings of length exactly 4" [7, v1, §1]. So the sharpest published bracket at $k = 4$ is [1.75, 2] rather than [1.75, 2.5]. This is not an error in anything [2] proves; we note it because it is the bracket our results should be measured against, and because it makes the comparison $\rho_4 > \rho_3$ of Corollary 4 meaningful — the published bracket is consistent with $\rho_4 = \rho_3 = 9/5$.

We have not been able to consult the text of [6], which is closed access with no open copy, and say nothing here about how it is proved.

Table 1 collects what is known after this paper. The upper bound 3 in every row is [3]; at $k = 5$ it coincides with the generic $(k + 1)/2$ of (2).

k	lower bound on ρ_k	upper bound	is $\rho_k \leq 2$?
3	9/5, exact [2]	9/5 [2]	true
4	$> 15/8$ (Thm. 13)	2 [6]	true
5	$> 15/8$ (Thm. 15)	3	open
6	2 [2]	3	open
7	2 [2]	3	open
8	> 2 (Thm. 21)	3	**false**
9	> 2 (Thm. 24)	3	**false**
even $k \geq 10$	$(9k + 2)/(4k + 4)$ [1]	3	**false**
11	> 2 (Thm. 26)	3	**false**
odd $k \geq 13$	2 [2]	3	open
unrestricted	9/4 [1]	3 [3]	**false**

TABLE 1. The greedy superstring ratio at a fixed string length, after this paper. The last line is the unrestricted ratio, over instances with strings of arbitrary lengths.

Two readings of the table are worth stating. The conjecture $\rho_k \leq 2$ is a theorem for $k \leq 4$ and false at $k = 8$, so *string length* 8 *is the smallest at which the greedy conjecture is known to fail*; before Theorem 21 that string length was 10. And it is false at $k = 9$ and $k = 11$, which are odd, whereas [1] settles even string lengths only. The cells $k = 5$, $k = 6$, $k = 7$ and the odd $k \geq 13$ remain open.

1.4. **Results.** Everything below concerns explicit instances. The paper works upward through the string lengths. Sections 3–7 are the material of the first version: two block families, one for each of the string lengths 4 and 5, refuting (3) at the two cells [2] leaves open. Here $S_t^{(4)}$ consists of $10t$ strings of length 4 and $S_t^{(5)}$ of $13t$ strings of length 5.

Theorem 2 (String lengths four and five)**.** (i) $\rho_4 \geq 76/43 > 7/4$, *witnessed by* $S_4^{(4)}$ *(*40 *strings, optimum* 43*, a legal greedy output of length* 76*).*

(ii) $\rho_5 \geq 125/69 > 9/5$, *witnessed by* $S_5^{(5)}$ *(*65 *strings, optimum* 69*, a legal greedy output of length* 125*).*

(iii) $\rho_4 \geq 437/233 > 15/8$, *witnessed by* $S_{23}^{(4)}$ *(*230 *strings, optimum* 233*, a legal greedy output of length* 437*); already* $S_6^{(4)}$ *gives* $\rho_4 \geq 114/63 > 9/5$.

(iv) $\rho_5 \geq 325/173 > 15/8$, *witnessed by* $S_{13}^{(5)}$ *(*169 *strings, optimum* 173*, a legal greedy output of length* 325*); already* $S_7^{(5)}$ *gives* $\rho_5 \geq 175/95 > 11/6$.

Parts (i) and (ii) refute the Cazaux–Rivals conjecture at the two string lengths it survived; parts (iii) and (iv) say by how much. Three consequences are worth stating separately.

Corollary 3. *The Cazaux–Rivals conjecture $\rho_k = 2 - 1/k$ is false for every $k \geq 3$.*

Corollary 4. *$\rho_4 > 9/5 = \rho_3$: the greedy superstring ratio is strictly larger on strings of length 4 than on strings of length 3.*

Corollary 5. *$15/8 < \rho_4 \leq 2$ and $15/8 < \rho_5 \leq 3$.*

Sections 8–10 are new. They concern the greedy conjecture itself, at the string lengths 8, 9, 10 and 11.

Theorem 6 (String lengths eight, nine, ten and eleven)**.** (i) (Shibata [1], reproduced.) *There is an instance of 10-SCS on 198 strings whose optimum is exactly 207 and on which a legal greedy execution outputs a string of length $415 > 2 \cdot 207$; hence $\rho_{10} > 2$.*

(ii) *There is an instance of 8-SCS on 252 strings whose optimum is exactly 259 and on which a legal greedy execution outputs a string of length $519 > 2 \cdot 259$; hence $\rho_8 > 2$.*

(iii) *There is an instance of 9-SCS on 336 strings whose optimum is exactly 344 and on which a legal greedy execution outputs a string of length $689 > 2 \cdot 344$; hence $\rho_9 > 2$.*

(iv) *There is an instance of 11-SCS on 250 strings whose optimum is exactly 260 and on which a legal greedy execution outputs a string of length $521 > 2 \cdot 260$; hence $\rho_{11} > 2$.*

Part (i) is not new: the instance, the merge order and the two length formulas behind it are Shibata's, and what is added is that they are reproduced from his definitions and checked, including the optimum, which his Theorem 10 obtains by a route we do not follow (Section 8). Parts (ii)–(iv) are new. Part (ii) is his construction at $k = 8$, one even cell below the range $k \geq 10$ his Theorem 14 states; his proof of the border-overlap table it depends on does not cover that cell, and Section 9 supplies the two missing entries. Parts (iii) and (iv) use blocks that appear in no published source: they were found by a search over the *shape* of Shibata's block (Section 10), which, unlike his three explicit words, is not tied to $k = 2s$.

Corollary 7. *The greedy superstring conjecture fails at string lengths 8, 9 and 11. With the results recorded in [7] and [6] it is therefore true for $k \leq 4$, false for $k = 8$, $k = 9$, $k = 11$ and every even $k \geq 10$, and open for $k = 5$, $k = 6$, $k = 7$ and every odd $k \geq 13$. In particular 8 is the smallest string length at which it is known to fail.*

Each of the three new instances comes with a companion one block smaller which shows the block count cannot be reduced: Propositions 22, 25 and 27. Two of the three land exactly on $2\,\mathrm{OPT}$ and the third one letter below it, so they are the sharpest possible controls.

Theorem 2 is Theorems 12 and 13 at $k = 4$ and Theorems 14 and 15 at $k = 5$; Theorem 6 is Theorems 19, 21, 24 and 26. The corollaries are proved in Sections 6 and 10. Each of the five block families continues past the instances listed above; what they converge to is described in Remarks 11, 23 and 29, which are not part of the machine-checked development.

1.5. **Method.** Each statement of Theorems 2 and 6 needs two numbers about one explicit instance S: an upper bound on $\mathrm{OPT}(S)$ and a lower bound on the length of some legal greedy output.

The optimum is obtained without any search. Exhibiting a common superstring w of S bounds $\mathrm{OPT}(S) \le |w|$, and that inequality is all the ratio needs; but each of our instances comes with a witness of length exactly $n+k-1$, which by Lemma 9 — an elementary counting argument — is the least a common superstring of n distinct strings of length k can have. So $\mathrm{OPT}(S)$ is determined exactly, and the ratio reported is the instance's true ratio and not merely a lower bound for it. The one place where a mistake would flatter the result is the claim that the exhibited word is a common superstring at all: were it not, the true optimum could be larger and the ratio smaller. That claim is a decidable property of explicit data, it is checked, and Theorems 18(iv) and 28(iii) show the check is not vacuous.

The greedy output is a certificate: a list of merges, each recorded as an ordered pair of positions in the current list of live strings. Replaying it is a finite computation which checks, at every step, that the two chosen live strings are distinct and that their overlap equals the maximum overlap over *all* ordered pairs of distinct live strings — so the certificate is accepted only if every merge is one Greedy is permitted to make. The replay is the largest computation entering any theorem below: at each of the 335 steps of the longest certificate the checker maximises over up to 336^2 ordered pairs.

How the block words were found is a separate matter, described in Section 11. That section reports several searches, but no theorem in this paper depends on any of them: the blocks are simply data once written down.

2. Preliminaries

2.1. **Strings, overlap, merge.** Fix an alphabet Σ. Strings are finite sequences over Σ, $|w|$ is the length of w, and $s \sqsubseteq w$ means that s occurs in w as a contiguous block, that is $w = u\,s\,v$ for some strings u, v. For a finite set S of strings,

$$\mathrm{OPT}(S) \;=\; \min\{\, |w| \;:\; s \sqsubseteq w \text{ for all } s \in S \,\}.$$

For strings p, q the overlap $\mathrm{ov}(p,q)$ is the largest $\ell \le \min(|p|,|q|)$ such that the length-ℓ suffix of p equals the length-ℓ prefix of q; it always exists, $\ell = 0$ being admissible. The overlap merge is $p \odot q = p \cdot q[\mathrm{ov}(p,q):]$, of length $|p|+|q|-\mathrm{ov}(p,q)$.

For a cyclic word B of length b and any $k \ge 1$, the *cyclic k-mers* of B are its b windows of length k read with wraparound, one window per starting position, wrapping as many times as $k > b$ requires; they are b strings of length k, not necessarily distinct. All instances in this paper are unions of sets of cyclic k-mers.

2.2. **Legal executions.** A *state* is a finite list L of live strings. Write

$$\mathrm{maxov}(L) \;=\; \max\{\, \mathrm{ov}(p,q) \;:\; p, q \in L,\; p \ne q \,\}$$

(the maximum over ordered pairs of *distinct strings*, with $\mathrm{maxov}(L) = 0$ when no such pair exists). A step is a pair of positions (i, j) of L; it is *legal* if $L_i \ne L_j$ and $\mathrm{ov}(L_i, L_j) = \mathrm{maxov}(L)$, and it transforms L into the list obtained by deleting positions i and j and appending $L_i \odot L_j$.

Definition 8. A *greedy certificate* for an instance S (listed in some order) is a sequence of steps, each legal in the state it is applied to, that reduces S to a single string g. Then $g \in \textsc{Greedy}(S)$, and g is a common superstring of S.

Definition 8 is deliberately conservative in two ways. It compares strings, not positions: if a state ever held two equal strings, a positions-only reading would

allow a merge GREEDY does not. And it forbids merging two occurrences of the same string, which is what "a pair of distinct strings" means. Neither restriction is ever active in the executions below, but each makes the certified executions a subset, never a superset, of the executions GREEDY may perform.

2.3. **The counting bound.**

Lemma 9. *Let S be a nonempty finite set of pairwise distinct strings, each of length k, and let w be a common superstring of S. Then*

$$|w| \;\geq\; |S| + k - 1.$$

Proof. A string of length ℓ has exactly $\ell - k + 1$ starting positions for a block of length k (and none at all if $\ell < k$; since $S \neq \emptyset$, some $s \in S$ occurs in w, so $|w| \geq k$). Every $s \in S$ occurs at some starting position, distinct elements of S occupy distinct positions because a position determines its block, and hence $|S| \leq |w| - k + 1$. □

This is folklore; it is the elementary half of the lemma of Pevzner [8] quoted as such in [2], whose other half says that equality holds whenever the de Bruijn subgraph of S is balanced and weakly connected. Only the inequality is used here, and it is used in the form:

Lemma 10. *If in addition w is a common superstring of S with $|w| = |S| + k - 1$, then w is shortest, i.e.* $\mathrm{OPT}(S) = |S| + k - 1$.

Proof. Immediate from Lemma 9. □

Every optimum quoted in this paper is obtained this way, and the witness w is in every case a word spelling an Eulerian circuit of the instance's de Bruijn subgraph. We use nothing about Eulerian circuits: the word is data, and that it is a common superstring of the right length is checked directly.

2.4. **What a lower bound on ρ_k requires.** Every theorem in Sections 4–10 is an instance of the following statement, which we abbreviate $\rho_k > a/b$:

> there are a nonempty finite set S of pairwise distinct strings of length k, a greedy certificate for S with output g, and a common superstring w of S such that $|w| \leq |v|$ for every common superstring v of S, with $a\,|w| < b\,|g|$.

Given such data, $|g|/\mathrm{OPT}(S) = |g|/|w| > a/b$, and $\rho_k \geq |g|/\mathrm{OPT}(S)$ by (1), so indeed $\rho_k > a/b$. We emphasise that the formal development does not define ρ_k — it defines no supremum at all — and therefore asserts nothing beyond what one instance witnesses; the passage from the displayed statement to $\rho_k > a/b$ is the definition (1) and is carried out in prose, here and nowhere else.

3. The two block families at string lengths four and five

The instances of Theorems 12–15 are the members of two families described in this section. The description explains how the instances are built and why they behave as they do; it is not what the theorems rest on. Each theorem is proved on its own instance, whose optimum comes from Lemma 10 applied to an exhibited witness and whose execution is replayed merge by merge, so no property claimed below for a general t enters any proof.

3.1. **Blocks and spectra.** Let B be a cyclic word of length b over the alphabet $\{\mathtt{A}, \mathtt{B}, \mathtt{X}\}$, in which $\mathtt{A}$ and $\mathtt{B}$ are *shared* letters and $\mathtt{X}$ is *fresh*. For $t \geq 1$ let B_i denote B with $\mathtt{X}$ renamed to a letter $\mathtt{X}_i$ private to the index i, and let

$$S_t(B) \;=\; \bigcup_{i=1}^{t} \{\, \text{the } b \text{ cyclic } k\text{-mers of } B_i \,\}.$$

If the b cyclic k-mers of B are pairwise distinct and each contains $\mathtt{X}$, then $S_t(B)$ consists of exactly bt pairwise distinct strings of length k: distinct within a block by the first condition, and across blocks because a string of block i contains $\mathtt{X}_i$.

3.2. **The block at $k = 4$.** Take $k = 4$ and

$$B^{(4)} \;=\; \mathtt{AAXABAXBAX} \qquad (b = 10).$$

Its ten cyclic 4-mers, read from positions $0, \ldots, 9$, are

$$\mathtt{AAXA},\ \mathtt{AXAB},\ \mathtt{XABA},\ \mathtt{ABAX},\ \mathtt{BAXB},\ \mathtt{AXBA},\ \mathtt{XBAX},\ \mathtt{BAXA},\ \mathtt{AXAA},\ \mathtt{XAAX},$$

pairwise distinct, each containing $\mathtt{X}$. Write $S_t^{(4)} = S_t(B^{(4)})$, so $|S_t^{(4)}| = 10t$.

The optimum. Let $v^* = \mathtt{ABA}$ (the unique cyclic 3-mer of $B^{(4)}$ free of $\mathtt{X}$) and let B_i^{rot} be B_i rotated to begin at v^*, i.e. $\mathtt{ABAX}_i\mathtt{BAX}_i\mathtt{AAX}_i$. Then

$$w_t^{(4)} \;=\; B_1^{\mathrm{rot}} B_2^{\mathrm{rot}} \cdots B_t^{\mathrm{rot}}\, v^* \tag{5}$$

has length $10t + 3$ and is a common superstring of $S_t^{(4)}$ — it traverses each block's cyclic word once, returning each time to v^* — so by Lemma 10, $\mathrm{OPT}(S_t^{(4)}) = 10t{+}3$.

The execution. The certificate makes greedy assemble, inside each block, the three strings

$$\mathtt{X}_i\mathtt{AAX}_i\mathtt{AA}\ (6), \qquad \mathtt{X}_i\mathtt{ABAX}_i\mathtt{AB}\ (7), \qquad \mathtt{X}_i\mathtt{BAX}_i\mathtt{BA}\ (6), \tag{6}$$

of total length 19 per block, and then concatenates the resulting $3t$ strings with no further overlap, for an output of length $19t$. Two facts make every merge legal. First, each of the merges producing (6) joins two live strings at overlap $3 = k{-}1$, and by the endpoint-inheritance lemma of [9], recalled in Remark 30, no two distinct live strings can overlap in more than $k - 1$ letters. So these merges are legal whatever else is live. Second, once (6) is reached, no two of the $3t$ strings overlap at all. Indeed each of them begins and ends with one of the three words $\mathtt{X}_i\mathtt{AA}$, $\mathtt{X}_i\mathtt{AB}$, $\mathtt{X}_i\mathtt{BA}$, and any overlap of two of them is an overlap of length ≤ 3 between such words; for length 1 or 2 the suffix begins with a shared letter while the prefix begins with a fresh one, and for length 3 the two words would have to be equal, which forces the same block and the same string. Hence greedy's last $3t - 1$ merges are plain concatenations and

$$|g| \;=\; 19t, \qquad \mathrm{OPT}(S_t^{(4)}) \;=\; 10t + 3.$$

For the three values of t used in Section 4 both quantities are verified on the instance itself; for general t see Remark 11.

3.3. **The block at $k = 5$.** Take $k = 5$ and

$$B^{(5)} \;=\; \texttt{AAABXAABXBXBX} \qquad (b = 13),$$

whose thirteen cyclic 5-mers are pairwise distinct and each contain `X`; write $S_t^{(5)} = S_t(B^{(5)})$, so $|S_t^{(5)}| = 13t$. The `X`-free cyclic 4-mer is $v^* = \texttt{AAAB}$, the block already begins there, and

$$w_t^{(5)} \;=\; B_1 B_2 \cdots B_t \, v^* \tag{7}$$

is a common superstring of length $13t + 4 = |S_t^{(5)}| + k - 1$; by Lemma 10 it is shortest. The certificate assembles, per block,

$$\texttt{X}_i\texttt{AAABX}_i\texttt{AAA}\ (9), \qquad \texttt{X}_i\texttt{AABX}_i\texttt{BX}_i\texttt{AAB}\ (10), \qquad \texttt{X}_i\texttt{BX}_i\texttt{BX}_i\texttt{B}\ (6), \tag{8}$$

of total length 25, every merge again being at the maximum possible overlap $4 = k-1$, and then concatenates the $3t$ strings with no overlap, so that $|g| = 25t$ against $\mathrm{OPT}(S_t^{(5)}) = 13t + 4$. Here the pairwise-overlap-0 condition is not settled by the one-line argument of Section 3.2, because the third string of (8) begins with $\texttt{X}_i\texttt{BX}_i\texttt{B}$, whose fresh letter is not only initial; it is instead a finite check over all ordered pairs of the three initial 4-mers, same-block and cross-block, which two blocks already exhaust. Again, for the three values of t used in Section 5 the two quantities are verified on the instance itself.

3.4. **The instances used below.** All six instances of Theorem 2 are members of these two families:

k	instance	n	OPT	$\lvert g\rvert$	ratio
4	$S_4^{(4)}$	40	43	76	$76/43 = 1.7674\ldots > 7/4$
4	$S_6^{(4)}$	60	63	114	$114/63 = 1.8095\ldots > 9/5$
4	$S_{23}^{(4)}$	230	233	437	$437/233 = 1.8755\ldots > 15/8$
5	$S_5^{(5)}$	65	69	125	$125/69 = 1.8115\ldots > 9/5$
5	$S_7^{(5)}$	91	95	175	$175/95 = 1.8421\ldots > 11/6$
5	$S_{13}^{(5)}$	169	173	325	$325/173 = 1.8786\ldots > 15/8$

Each t is the least at which $19t/(10t + 3)$, respectively $25t/(13t + 4)$, exceeds the value in the last column, so the margins at the threshold are narrow: the last two lines are $15 \cdot 233 = 3495 < 3496 = 8 \cdot 437$ and $15 \cdot 173 = 2595 < 2600 = 8 \cdot 325$.

Remark 11 (not part of the machine-checked development). Both families continue. The five properties used in Sections 3.2 and 3.3 — that the b cyclic k-mers of the block are distinct and each contain `X`; that (5), (7) are common superstrings of length $bt + k - 1$; that each merge assembling (6), (8) is at overlap $k - 1$; that the resulting strings are pairwise overlap-free; and that the block's edge set decomposes into the three pieces used — are all finite checks on the block itself, independent of t, and each was carried out. Granting the induction over t, which is bookkeeping rather than mathematics, one obtains

$$\rho_4 \;\geq\; \lim_{t\to\infty} \frac{19t}{10t+3} \;=\; \frac{19}{10} \;=\; 1.9, \qquad \rho_5 \;\geq\; \lim_{t\to\infty} \frac{25t}{13t+4} \;=\; \frac{25}{13} \;=\; 1.9230\ldots,$$

narrowing the brackets of Corollary 5 to [1.9, 2] and [25/13, 3]. These two statements are *not* part of the machine-checked development, which instantiates the

families only at the six values of t tabulated above; they are recorded here as what the constructions give and are not used anywhere below.

4. String length four

Theorem 12. *Let $S_4^{(4)}$ be the instance of 40 strings of length 4 of Section 3.2. Then $\mathrm{OPT}(S_4^{(4)}) = 43$, and there is a greedy certificate for $S_4^{(4)}$ whose output has length 76. Consequently*

$$\rho_4 \;\geq\; \frac{76}{43} \;>\; \frac{7}{4} \;=\; 2 - \frac{1}{4},$$

so the Cazaux–Rivals conjecture fails at $k = 4$.

Proof sketch. The instance, the word $w_4^{(4)}$ of (5), the list of 39 merges and the output are explicit finite data. Three properties of them are decided by evaluation: that the 40 strings are pairwise distinct; that $w_4^{(4)}$ is a common superstring of the instance; and that the list of merges is a greedy certificate in the sense of Definition 8, that is, that at each of the 39 steps the two chosen live strings are distinct and their overlap equals the maximum over all ordered pairs of distinct live strings. The last is the largest of the three computations, and it is exhaustive only over the pairs available at each step. For the optimum, no search is performed: $|w_4^{(4)}| = 43 = 40 + 4 - 1$, so Lemma 10 makes it shortest. Finally $7 \cdot 43 = 301 < 304 = 4 \cdot 76$. □

Theorem 13. *With $S_6^{(4)}$ (60 strings, optimum 63, a legal greedy output of length 114) and $S_{23}^{(4)}$ (230 strings, optimum 233, a legal greedy output of length 437),*

$$\rho_4 \;\geq\; \frac{114}{63} \;>\; \frac{9}{5} \qquad \text{and} \qquad \rho_4 \;\geq\; \frac{437}{233} \;>\; \frac{15}{8}.$$

Proof sketch. Identical in shape to Theorem 12, with 59 and 229 merges respectively and with $|w_6^{(4)}| = 63 = 60+4-1$ and $|w_{23}^{(4)}| = 233 = 230+4-1$. The two arithmetic facts are $9 \cdot 63 = 567 < 570 = 5 \cdot 114$ and $15 \cdot 233 = 3495 < 3496 = 8 \cdot 437$. □

5. String length five

Theorem 14. *Let $S_5^{(5)}$ be the instance of 65 strings of length 5 of Section 3.3. Then $\mathrm{OPT}(S_5^{(5)}) = 69$, and there is a greedy certificate for $S_5^{(5)}$ whose output has length 125. Consequently*

$$\rho_5 \;\geq\; \frac{125}{69} \;>\; \frac{9}{5} \;=\; 2 - \frac{1}{5},$$

so the Cazaux–Rivals conjecture fails at $k = 5$.

Proof sketch. As for Theorem 12: distinctness of the 65 strings, the superstring property of $w_5^{(5)}$ and the legality of the 64 merges are decided by evaluation; $|w_5^{(5)}| = 69 = 65+5-1$ makes $w_5^{(5)}$ shortest by Lemma 10; and $9 \cdot 69 = 621 < 625 = 5 \cdot 125$. □

Theorem 15. *With $S_7^{(5)}$ (91 strings, optimum 95, a legal greedy output of length 175) and $S_{13}^{(5)}$ (169 strings, optimum 173, a legal greedy output of length 325),*

$$\rho_5 \;\geq\; \frac{175}{95} \;>\; \frac{11}{6} \;=\; 2 - \frac{1}{6} \qquad \text{and} \qquad \rho_5 \;\geq\; \frac{325}{173} \;>\; \frac{15}{8}.$$

In particular the conjectured value $2 - 1/5$ is wrong at $k = 5$ by more than one step of the sequence $2 - 1/k$ itself.

Proof sketch. As above, with 90 and 168 merges; $11 \cdot 95 = 1045 < 1050 = 6 \cdot 175$ and $15 \cdot 173 = 2595 < 2600 = 8 \cdot 325$. □

6. Consequences at string lengths four and five

Proof of Corollary 3. For $k = 3$ the value $\rho_3 = 9/5 > 5/3$ is [2]; alternatively, Theorem 17(iii) below exhibits an instance of 3-SCS with $\rho_3 > 5/3 = 2 - 1/3$, which needs their construction but not their proof. For $k = 4$ and $k = 5$, Theorem 2(i), (ii). For $k = 6$, Theorem 17(iii) again, or [2]. For $k \geq 7$, $\rho_k \geq 2 > 2 - 1/k$ by [2]. □

Proof of Corollary 4. $\rho_4 \geq 114/63 > 9/5$ by Theorem 13, and $\rho_3 = 9/5$ by [2]. □

Corollary 5 adds the upper bounds of [6] (see Remark 1) and of (2); no upper bound is proved in this paper. Corollary 4 is the qualitative statement we would single out. Before it, the sharpest published information at $k = 4$ was $1.75 \leq \rho_4 \leq 2$ (Remark 1), which is consistent with $\rho_4 = \rho_3$, and indeed consistent with ρ_k being constant on $3 \leq k \leq 5$; the two block families show instead that ρ_4 and ρ_5 both exceed $15/8$, hence both exceed $\rho_3 = 9/5$, and indeed both exceed the value $2 - 1/6$ that the Cazaux–Rivals conjecture assigns to a longer string length. We have no upper bound to offer. Whether $\rho_4 = 2$ — that is, whether the bound of [6] is attained — remains open, as does $\rho_5 \leq 2$, which by Table 1 is one of the four surviving cells of the greedy conjecture. The greedy conjecture is recorded in [7] as known for strings of length at most 3 and for strings of length *exactly* 4 — the mixed case of strings of length at most 4 is called "uncharted territory" there.

Remark 16. All lower bounds here, like those of [5], [2] and [1], are statements about adversarial tie-breaking: they exhibit one legal execution, not the behaviour of every one. That this is the right reading of ρ_k is the standard convention and is the framing of [2]; that it is not vacuous in the present instances is Theorem 18(v), which exhibits a second legal execution on the $k = 6$ instance of [2] attaining the optimum exactly.

7. The published constructions at $k = 3$ and $k = 6$, reproduced

The machinery of Section 2 was calibrated on the constructions of [2] before it was used for new ones. The following theorem is a reproduction and claims no novelty; every number in parts (i) and (ii) is printed in [2].

Theorem 17. (i) *(The $k = 6$ block of* [2]*.) The union of the cyclic* 6*-spectra of* `ABAABX` *and* `ABX` *consists of* 9 *distinct strings; the word* `ABAABXABXABAAB` *of length* $14 = 9 + 6 - 1$ *is a common superstring, hence shortest; there is a greedy certificate with output g of length* 18*; and the two words* $p_1 =$ `XABAABXABAA` *and* $p_2 =$ `BXABXABX`*, of lengths* 11 *and* 8*, satisfy* $\mathrm{ov}(p_2, p_1) = 1$ *and* $p_2 \odot p_1 = g$.

(ii) *(The $k = 3$ block of* [2]*.) The* 3*-spectrum of* `ABXAXAB` *consists of* 5 *distinct strings with optimum* 7*, and there is a greedy certificate with output of length* 9.

(iii) *Iterating those two blocks gives* $\rho_3 > 5/3$ *(an instance of* 30 *strings, optimum* 32*, greedy output* 54*) and* $\rho_6 > 11/6$ *(an instance of* 63 *strings, optimum* 68*, greedy output* 126*).*

Proof sketch. Each item is the same finite check as in Sections 4 and 5: distinctness, the superstring property of the exhibited word, the length identity $n+k-1$ feeding Lemma 10, and the replay of the certificate. For (i) the two piece lengths, their overlap and the identity $p_2 \odot p_1 = g$ are likewise evaluated. For (iii), $5 \cdot 32 = 160 < 162 = 3 \cdot 54$ and $11 \cdot 68 = 748 < 756 = 6 \cdot 126$. □

Part (iii) is what Corollary 3 uses at $k = 3$ and $k = 6$, so that only the range $k \geq 7$ of that corollary rests on a theorem quoted from elsewhere. Note that Theorem 17(iii) reproduces the *refutations* of [2] at those two string lengths, not their theorems: $\rho_3 = 9/5$ and $\rho_k \geq 2$ for all $k \geq 6$ are statements about infinite families and are not reproved here.

Theorem 18 (Controls)**.** (i) *Lemma 9 fails without distinctness: for the multiset* $(\mathtt{AA}, \mathtt{AA})$ *and* $w = \mathtt{AA}$ *one has* $|w| = 2 < 3 = 2 + 2 - 1$.

(ii) *A list of merges that leaves more than one string live is rejected.*

(iii) *One block alone does not suffice at* $k = 4$*:* $S_1^{(4)}$ *has* 10 *strings, optimum* 13 *and a legal greedy output of length* 19*, and* $19/13 < 7/4$*. The iteration over* t*, not the block, is what crosses the threshold.*

(iv) *The certificate checker is not permissive: replacing the first merge of the certificate of Theorem 12 by a pair of overlap* 2*, rather than the maximum* 3*, makes it reject the whole certificate. Nor is the superstring checker constant-true: deleting the first letter of* $w_4^{(4)}$ *destroys the property.*

(v) *On the* $k = 6$ *instance of Theorem 17(i) there is a second legal greedy execution whose output has length* 14*, the optimum.*

Items (i)–(iv) are negative controls on the definitions rather than results: (i) shows that the distinctness hypothesis of Lemma 9 is load-bearing, (ii) and (iv) that a certificate can fail to be accepted — so an accepted certificate carries information — and (iii) that the theorems are not artefacts of a single block. Item (v) is the content of Remark 16.

8. The refutation of the greedy conjecture, and string length ten

From here to the end of Section 10 the subject is the greedy conjecture itself.

8.1. **Shibata's construction.** We follow [1] and its notation, which is not the notation of Sections 3–7: here $\mathtt{x}$, $\mathtt{y}$ and $\mathtt{a}$ are the *shared* letters and $\mathtt{b}_i$ is the letter fresh to block i, whereas in Section 3 the fresh letter was written $\mathtt{X}$. Two collisions are worth naming, because in each the two letters differ only in case and their roles are opposite: the fresh letter of Section 3 is $\mathtt{X}$ while a shared letter here is $\mathtt{x}$, and $\mathtt{B}$ is shared in Section 3 while $\mathtt{b}_i$ is fresh here. No formula of Sections 3–7 may be read with the letters of this section, or conversely. Put

$$X_n \;=\; \begin{cases} (\mathtt{yx})^{n/2} & n \text{ even}, \\ \mathtt{x}(\mathtt{yx})^{(n-1)/2} & n \text{ odd}, \end{cases}$$

a string of length n; in particular X_0 is the empty string. For $k = 2s$ with $s \geq 3$ and for each block index i set

$$U_i = \mathtt{xa}X_{s-3}\,\mathtt{b}_i\,\mathtt{xa}X_{s-3}\,\mathtt{y}, \qquad V_i = \mathtt{xa}X_{s-3}\,\mathtt{b}_i, \qquad W_i = \mathtt{xa}X_{s-1}\,\mathtt{b}_i,$$

of lengths k, s and $s+2$, read as cyclic words, and let $S_{k,t}$ be the union over $i \le t$ of the cyclic k-mers of U_i, V_i and W_i. Then $|S_{k,t}| = t(4s+2)$, the strings are pairwise distinct (his Lemma 5),

$$\mathrm{OPT}(S_{k,t}) \;=\; t(4s+2) + k - 1, \tag{9}$$

and there is a legal greedy execution whose output has length $t(9s+1)+1$ (his Theorem 13). Its ratio increases to $(9s+1)/(4s+2) = (9k+2)/(4k+4)$ and never attains it, which is his Theorem 14 for even $k \ge 10$; letting $k \to \infty$ gives $\rho \ge 9/4$ for the unrestricted problem, his Corollary 15.

The execution proceeds in four phases: greedy first closes each of the $3t$ cyclic words by merges at the maximum possible overlap $k-1$, leaving $3t$ live strings $\widehat{U_i}, \widehat{V_i}, \widehat{W_i}$ whose pairwise overlaps are entries of a 3×3 table of *border overlaps* (his Table 1, in the two variants same block and different block); it then merges $\widehat{V_i} \odot \widehat{U_i}$ for each i, chains the results, and finally appends the $\widehat{W_i}$ at overlap 0. Three starting offsets — where each cyclic word is cut open — are fixed in [1] without comment; Remark 32 reports what they are worth.

Because the ratio only tends to $(9k+2)/(4k+4)$, no single instance witnesses $23/11$ at $k = 10$. What one instance can witness, and what refutes the conjecture, is a ratio strictly above 2; at $k = 10$ the first block count that achieves it is $t = 9$.

8.2. **String length ten.**

Theorem 19 (Shibata's refutation, reproduced)**.** *Let $S_{10,t}$ be as in Section 8.1 with $s = 5$.*

(i) *At $t = 1$: $|S_{10,1}| = 22$, $\mathrm{OPT}(S_{10,1}) = 31$, and there is a greedy certificate whose output has length 47 — the values $t(4s+2)+k-1$ and $t(9s+1)+1$ of* [1] *at $s = 5$, $t = 1$.*

(ii) *At $t = 9$: $|S_{10,9}| = 198$, $\mathrm{OPT}(S_{10,9}) = 207$, and there is a greedy certificate whose output has length 415. Since $415 > 414 = 2 \cdot 207$,*

$$\rho_{10} \;\ge\; \frac{415}{207} \;=\; 2.0048\ldots \;>\; 2,$$

so the greedy superstring conjecture is false.

Proof sketch. The instance, the witness word and the list of 197 merges are explicit finite data, emitted from an implementation of the definitions of Section 8.1. Three properties are decided by evaluation: that the 198 strings are pairwise distinct; that the exhibited word is a common superstring of the instance; and that the list of merges is a greedy certificate, that is, that at each of the 197 steps the two chosen live strings are distinct and their overlap is the maximum over all ordered pairs of distinct live strings — up to 198^2 pairs at the first step. For the optimum, no search is performed and (9) is not quoted: the exhibited word has length $207 = 198+10-1$, so Lemma 10 makes it shortest. Finally $415 > 2 \cdot 207$. □

Theorem 19 claims no novelty. The instance, the merge order and both length formulas are Shibata's; what is added is that they are reproduced from his definitions rather than transcribed, and that the two halves of the ratio are established independently of his proof — the optimum in particular, which (9) obtains from the Eulerian-circuit half of Pevzner's lemma, whereas here it comes from Lemma 10 and one exhibited word. All eighteen entries of his border-overlap table were also

recomputed, at $s = 4, 5, 6, 7$, and agree; that recomputation is what makes Section 9 possible.

Remark 20 (not part of the machine-checked development). Within Shibata's family, $k = 10$ with $t = 9$ is the smallest refuting instance in the number of strings. The condition $t(9s+1)+1 > 2\big(t(4s+2)+2s-1\big)$ is $t(s-3) > 4s-3$, so the least block count is $t_{\min}(s) = \lfloor (4s-3)/(s-3) \rfloor + 1$ and the instance has $t_{\min}(s)\,(4s+2)$ strings; this is $252, 198, 208, 210, 204$ at $s = 4, 5, 6, 7, 8$, minimised at $s = 5$. This is arithmetic on the two length formulas, cross-checked against simulation at every $s \le 8$, and is why Theorem 19 is stated at $k = 10$, $t = 9$.

9. String length eight

Theorem 14 of [1] is stated for even $k \ge 10$. The three words U_i, V_i, W_i are defined for every $s \ge 3$, since only X_{s-3} occurs in them; what needs $s \ge 5$ is the printed *proof* of the border-overlap table, because the borders themselves,

$$U_i' = X_{s-4}\,\mathtt{b}_i\,\mathtt{xa}X_{s-1}\,\mathtt{a}, \quad V_i' = \mathtt{a}X_{s-3}\,\mathtt{b}_i\,\mathtt{xa}X_{s-3}\,\mathtt{b}_i, \quad W_i' = X_{s-4}\,\mathtt{b}_i\,\mathtt{xa}X_{s-1}\,\mathtt{b}_i,$$

each of length $\ell = k-1 = 2s-1$, are written with X_{s-4}. Exactly two sentences of that proof fail at $s = 4$, where X_0 is empty: the entry $\mathrm{ov}(V_i', W_i') = s-3$ is obtained from "$V_i'[1] = \mathtt{a}$ and $W_i'[1] \in \{\mathtt{x}, \mathtt{y}\}$", but $W_i'[1] = \mathtt{b}_i$ when $s = 4$; and the entry $\mathrm{ov}(W_i', V_i') = 0$ is obtained from the relevant suffix of W_i' beginning "with the first character of X_{s-4}", which has no first character when $s = 4$.

Both entries are nonetheless true at $s = 4$, by the bookkeeping of positions the source uses elsewhere. With $s = 4$ and $\ell = 7$ the three borders are

$$U_i' = \mathtt{b}_i\mathtt{xaxyxa}, \qquad V_i' = \mathtt{axb}_i\mathtt{xaxb}_i, \qquad W_i' = \mathtt{b}_i\mathtt{xaxyxb}_i.$$

Entry $\mathrm{ov}(V_i', W_i')$. The last letter of V_i' is $\mathtt{b}_i$, so any positive overlap length is a position of $\mathtt{b}_i$ in W_i', that is 1 or ℓ. Length ℓ would force $V_i' = W_i'$, false since $V_i'[1] = \mathtt{a} \ne \mathtt{b}_i = W_i'[1]$. Length 1 does hold, as $V_i'[\ell] = \mathtt{b}_i = W_i'[1]$. So the overlap is $1 = s-3$, as tabulated. *Entry* $\mathrm{ov}(W_i', V_i')$. The last letter of W_i' is $\mathtt{b}_i$, so any positive overlap length is a position of $\mathtt{b}_i$ in V_i', that is $s-1 = 3$ or ℓ. Length ℓ would force $W_i' = V_i'$, false. Length 3 compares the suffix $\mathtt{yxb}_i$ of W_i' with the prefix $\mathtt{axb}_i$ of V_i', which differ in their first letter. So the overlap is 0, as tabulated.

Every other step of the argument of [1] is uniform in $s \ge 4$: his Lemma 5 needs $s+2 < k < 2(s+2)$, which at $s = 4$ reads $6 < 8 < 12$, and $s+2 \ne s$; his Lemma 7 needs $(k-1)-(s-2) = s+1$; and the merge $\widehat{V}_i \odot \widehat{U}_i$ at overlap $s-3$ is still legal at $s = 4$, where $s-3 = 1$ is still the maximum surviving overlap. All eighteen table entries were in addition recomputed at $s = 4, 5, 6, 7$.

None of this is needed for the theorem that follows, which is a statement about one instance and is checked on that instance; it is what explains why the instance exists, and it is what Remark 23 rests on.

Theorem 21. *Let $S_{8,t}$ be the construction of Section 8.1 at $s = 4$.*

(i) *At $t = 1$: $|S_{8,1}| = 18$, $\mathrm{OPT}(S_{8,1}) = 25$, and there is a greedy certificate whose output has length 38.*

(ii) *At $t = 14$: $|S_{8,14}| = 252$, $\mathrm{OPT}(S_{8,14}) = 259$, and there is a greedy certificate whose output has length 519. Since $519 > 518 = 2 \cdot 259$,*

$$\rho_8 \;\ge\; \frac{519}{259} \;=\; 2.0038\ldots \;>\; 2.$$

Hence the greedy superstring conjecture fails at string length 8*, one even string length below the range for which it is refuted in* [1].

Proof sketch. As for Theorem 19, on the instance itself: the 252 strings are checked pairwise distinct; the exhibited word of length $259 = 252 + 8 - 1$ is checked to be a common superstring, so Lemma 10 makes it shortest; and the list of 251 merges is replayed, each step maximising the overlap over all ordered pairs of distinct live strings. Finally $519 > 2 \cdot 259$. The proof of Theorem 21 does not use the border-overlap table, repaired or otherwise; the table is what predicted the instance. □

Proposition 22 (the boundary at $k = 8$)**.** *At* $t = 13$ *the same construction gives* $|S_{8,13}| = 234$, $\mathrm{OPT}(S_{8,13}) = 241$ *and a legal greedy output of length exactly* $482 = 2 \cdot 241$*. So fourteen blocks is the least number for which this construction refutes the conjecture at* $k = 8$*.*

Proof sketch. The same three evaluations on the $t = 13$ instance, with 233 merges and $241 = 234 + 8 - 1$; the final comparison is the equality $482 = 2 \cdot 241$ rather than a strict inequality. The minimality is arithmetic: the family's ratio is $(37t + 1)/(18t + 7)$ and this exceeds 2 exactly when $t \geq 14$. □

Remark 23 (not part of the machine-checked development). With the two repaired table entries above, the whole analysis of [1] applies at $s = 4$, and the family $S_{8,t}$ has $|S_{8,t}| = 18t$, $\mathrm{OPT} = 18t + 7$ and a legal greedy output of length $37t + 1$ for every t, whence

$$\rho_8 \;\geq\; \lim_{t\to\infty} \frac{37t+1}{18t+7} \;=\; \frac{37}{18} \;=\; 2.0555\ldots,$$

which is the value $(9k + 2)/(4k + 4)$ of [1] at $k = 8$. This moves the bracket at $k = 8$ from [2, 3] to [37/18, 3]. It is not part of the machine-checked development, which instantiates the family at $t = 1, 13, 14$ only; independently of the argument, the three lengths were confirmed by direct simulation — recomputing the true maximum overlap at every step rather than replaying a certificate — at $t = 1, \ldots, 17$ and at $t = 20, 25, 30$, every ratio matching $(37t + 1)/(18t + 7)$ exactly.

10. The odd string lengths nine and eleven

The construction of Section 8.1 is written for $k = 2s$ from its first definition onward: X_{s-3}, $|V_i| = s$, $|W_i| = s + 2$, the distinctness lemma and every entry of the border table are expressed in s. Nothing in it applies to an odd string length, where for $k \geq 7$ the best available lower bound was $\rho_k \geq 2$ [2].

Its *shape*, however, is not tied to $k = 2s$. Abstract a block as: one cyclic word U of length k carrying a single fresh letter $\mathtt{b}$, together with two cyclic words V, W obtained from periodic vertices of U. Here a vertex means a $(k-1)$-mer α of U; if α has period v then its length-v prefix is a cyclic word whose de Bruijn circuit passes through α, which is exactly how V_i and W_i arise in [1]. A block of this shape is balanced, is connected to the other blocks through the one $(k-1)$-mer of U that misses $\mathtt{b}$, and lets greedy close all three circuits at overlap $k - 1$; its ratio is then $1 + \big(3(k-1) - \lambda\big)/b$, where $b = |U| + |V| + |W|$ is the number of strings per block and λ is the overlap the continuation is forced to waste per block. The two blocks below were found by searching that shape; see Remark 33.

10.1. **String length nine.** The block is

$$U = \texttt{b}\,\texttt{xxyxaxyx}\ (9), \qquad V = \texttt{xyxbx}\ (5), \qquad W = \texttt{xbxxyxa}\ (7),$$

with $b = 21$ strings per block; the closing offsets are $(1, 2, 2)$. Numbering the letters of U from 1, its cyclic 8-mer starting at position 7 has period 5 and its length-5 prefix is V up to rotation; the one at position 9 has period 7 and gives W the same way; and `xxyxaxyx` is the unique 8-mer of U missing `b`, so it is the vertex through which the blocks are joined. All of this was verified rather than assumed. Writing $S_t^{(9)}$ for the union over $i \le t$ of the cyclic 9-mers of the three words with `b` renamed $\texttt{b}_i$, one has $|S_t^{(9)}| = 21t$.

Theorem 24. (i) *At $t = 1$: $|S_1^{(9)}| = 21$, $\mathrm{OPT}(S_1^{(9)}) = 29$, and there is a greedy certificate whose output has length 44.*

(ii) *At $t = 16$: $|S_{16}^{(9)}| = 336$, $\mathrm{OPT}(S_{16}^{(9)}) = 344$, and there is a greedy certificate whose output has length 689. Since $689 > 688 = 2 \cdot 344$,*

$$\rho_9 \ \ge\ \frac{689}{344} \ =\ 2.0029\ldots \ >\ 2.$$

Hence the greedy superstring conjecture fails at an odd string length.

Proof sketch. As before, on the instance: the 336 strings are checked pairwise distinct; the exhibited word of length $344 = 336 + 9 - 1$ is checked to be a common superstring of the instance, so Lemma 10 makes it shortest; and the certificate of 335 merges is replayed, each step maximising over all ordered pairs of distinct live strings — up to 336^2 pairs at the first step, the largest computation in this paper. Finally $689 > 2 \cdot 344$. How the block was found plays no part: it is data. □

Proposition 25 (the boundary at $k = 9$)**.** *At $t = 15$ the same block gives $|S_{15}^{(9)}| = 315$, $\mathrm{OPT}(S_{15}^{(9)}) = 323$ and a legal greedy output of length exactly $646 = 2 \cdot 323$. So sixteen blocks is the least number for which this block refutes the conjecture at $k = 9$.*

Proof sketch. The same three evaluations on the $t = 15$ instance, with 314 merges and $323 = 315 + 9 - 1$; the final comparison is the equality $646 = 2 \cdot 323$. The minimality is arithmetic: $(43t + 1)/(21t + 8) > 2$ exactly when $t \ge 16$. □

10.2. **String length eleven.** The block is

$$U = \texttt{b}\,\texttt{xxyxaxxyxx}\ (11), \qquad V = \texttt{xxyxxb}\ (6), \qquad W = \texttt{xxbxxyxa}\ (8),$$

again of the shape described above, with $b = 25$ strings per block and closing offsets $(11, 3, 2)$; write $S_t^{(11)}$ for the corresponding instance, so $|S_t^{(11)}| = 25t$.

Theorem 26. (i) *At $t = 1$: $|S_1^{(11)}| = 25$, $\mathrm{OPT}(S_1^{(11)}) = 35$, and there is a greedy certificate whose output has length 53.*

(ii) *At $t = 10$: $|S_{10}^{(11)}| = 250$, $\mathrm{OPT}(S_{10}^{(11)}) = 260$, and there is a greedy certificate whose output has length 521. Since $521 > 520 = 2 \cdot 260$,*

$$\rho_{11} \ \ge\ \frac{521}{260} \ =\ 2.0038\ldots \ >\ 2.$$

Hence the greedy superstring conjecture fails at a second odd string length. At 250 strings this is the smallest of the three instances of Theorem 6(ii)–(iv).

Proof sketch. As before: distinctness of the 250 strings, the superstring property of the exhibited word of length $260 = 250 + 11 - 1$ feeding Lemma 10, and the replay of 249 merges, each maximising over all ordered pairs of distinct live strings. Finally $521 > 2 \cdot 260$. □

Proposition 27 (the near-miss at $k = 11$)**.** *At $t = 9$ the same block gives $|S_9^{(11)}| = 225$, $\mathrm{OPT}(S_9^{(11)}) = 235$ and a legal greedy output of length exactly $469 = 2 \cdot 235 - 1$ — one letter short of a refutation. So ten blocks is the least number for which this block refutes the conjecture at $k = 11$.*

Proof sketch. The same three evaluations on the $t = 9$ instance, with 224 merges and $235 = 225 + 11 - 1$; the final comparison is $469 + 1 = 2 \cdot 235$. The minimality is arithmetic: $(52t + 1)/(25t + 10) > 2$ exactly when $t \geq 10$. □

Proof of Corollary 7. Falsity at $k = 8$, 9, 11 is Theorems 21, 24, 26; at even $k \geq 10$ it is [1], and at $k = 10$ also Theorem 19. Truth for $k \leq 3$ is the result recorded in [7] as "GC holds for strings of length no more than 3" — at $k = 3$ it also follows from $\rho_3 = 9/5$ [2] — and truth at $k = 4$ is [6] (see Remark 1). The remaining cells are those of Table 1. Smallest failing string length: the conjecture is a theorem at every $k \leq 4$, so no string length below 8 can fail, and 5, 6, 7 are open. □

10.3. **Controls.**

Theorem 28 (Controls for the new instances)**.**

(i) *One block is not enough. At $t = 1$ the ratios are*

$$\frac{47}{31}, \qquad \frac{38}{25}, \qquad \frac{44}{29}, \qquad \frac{53}{35}$$

at the string lengths 10, 8, 9 and 11 respectively, none of them above 1.53. With the three boundary instances above, this pins the block count from both sides.

(ii) *The certificate checker is not permissive: in each of the four refuting certificates, replacing the first merge by a pair whose overlap is one less than the maximum — 8 instead of 9 at $k = 10$, 6 instead of 7 at $k = 8$, 7 instead of 8 at $k = 9$, 9 instead of 10 at $k = 11$ — makes the checker reject the whole certificate, of 197, 251, 335 and 249 steps respectively.*

(iii) *The superstring checker is not constant-true: in each of the four cases, deleting the first letter of the exhibited optimal word destroys the common-superstring property.*

These are controls on the definitions, not results. Item (i) at $k = 8$ and $k = 9$ is sharp in the other direction as well, since the boundary instances land exactly on $2\,\mathrm{OPT}$; item (ii) shows that an accepted certificate carries information; item (iii) shows that the half of each theorem which could inflate a ratio is not vacuously true.

10.4. **What the two odd blocks give for all t.**

Remark 29 (not part of the machine-checked development). For both odd blocks the four phases of Section 8.1 apply verbatim, and the border overlaps surviving the closing phase are all 0 or 1. At $k = 9$ the arcs of overlap 1 are exactly $U_i' \to V_j'$ for all i, j, together with $V_i' \to U_i'$, $V_i' \to W_i'$ and $W_i' \to U_i'$; greedy therefore spends t merges on the $\widehat{V}_i \odot \widehat{U}_i$ and $t - 1$ on chaining the results, all at overlap 1, and the

last t merges are free, for a total loss of $2t-1$ against $3t$ closed strings of total length $t(21+3\cdot 8)=45t$. So

$$|S_t^{(9)}|=21t,\quad \mathrm{OPT}=21t+8,\quad |g|=43t+1,\qquad \rho_9\ \geq\ \frac{43}{21}=2.0476\ldots,$$

and the same argument with a per-block loss of 3 gives, at $k=11$,

$$|S_t^{(11)}|=25t,\quad \mathrm{OPT}=25t+10,\quad |g|=52t+1,\qquad \rho_{11}\ \geq\ \frac{52}{25}=2.08.$$

These move the brackets from [2, 3] to [43/21, 3] and [52/25, 3]. Neither is part of the machine-checked development, which instantiates the two families at $t=1,15,16$ and $t=1,9,10$ only; independently of the argument, all three lengths were confirmed by direct simulation, recomputing the true maximum overlap over all ordered pairs of distinct live strings at every step, for $t=1,\ldots,17$ at $k=9$ and $t=1,\ldots,12$ at $k=11$, with the optimal word checked to be a common superstring of length $n+k-1$ each time.

Both blocks have $b=2k+3$ and per-block loss $(k-5)/2$, which would give the ratio $(9k+5)/(4k+6)$ — exactly 43/21 at $k=9$ and 52/25 at $k=11$. Whether that pattern continues at any larger odd k is untested; at $k=13$ it would predict 61/29. We state it as an observation about two blocks, not as a claim about any third.

11. How the blocks were found

This section is expository: nothing in it is used by any theorem above, and none of it is part of the machine-checked development. It is included because the search space of k-SCS instances is astronomically large and none of the blocks above came out of it by luck.

Remark 30 (the design principle; not part of the machine-checked development). Let $S\subseteq\Sigma^k$ and let $G(S)$ be the de Bruijn subgraph whose vertices are $(k-1)$-mers and whose edges are the elements of S, the string $s_0\cdots s_{k-1}$ being an edge from $s_0\cdots s_{k-2}$ to $s_1\cdots s_{k-1}$. If $G(S)$ is balanced and weakly connected then $\mathrm{OPT}(S)=n+k-1$ with $n=|S|$ [8]. By the endpoint-inheritance lemma of [9] — for live strings $x\neq y$ one has $\mathrm{ov}(x,y)=\mathrm{ov}(\mathrm{right}(x),\mathrm{left}(y))\leq k-1$, where $\mathrm{right}(x)$ is the last *input* string merged into x and $\mathrm{left}(y)$ the first of y, the bound because those are two distinct inputs of length k — greedy's merges at overlap $k-1$ concatenate edge-disjoint trails of $G(S)$, and this phase can stop only when no live pair $x\neq y$ has the right endpoint vertex of x equal to the left endpoint vertex of y. On a balanced graph that forces every live trail to be *closed* with pairwise distinct base vertices: each open trail contributes an out-excess at its start and an in-excess at its end, balance makes the multisets of starts and ends coincide, and the stopping condition then matches each trail to itself. A closed trail of L edges spells a string of length $L+k-1$, so if the phase leaves c pieces and greedy then recovers total overlap O, the output has length $n+c(k-1)-O$ and

$$\frac{|g|}{\mathrm{OPT}(S)}\ =\ \frac{n+c(k-1)-O}{n+k-1}\ \longrightarrow\ 1+(k-1)\,\frac{c}{n}\qquad (n\to\infty,\ O/n\to 0).$$

Maximising greedy's ratio is minimising the average length of a closed trail. This observation is not new in substance — read in the opposite direction, bounding the number of trails in a cover bounds greedy from above, which is the engine of

the $\rho_3 \le 9/5$ proof of [2] — and the constraint that the average closed trail have length at least 2 is exactly the bound $\rho_k \le (k+1)/2$ of [5]. Its use here is as a design specification: over cyclic block words B of length b whose k-mers are distinct and `X`-carrying, with some $(k-1)$-mer `X`-free (which is what connects the blocks), decompose the block's closed walk into c closed trails with pairwise overlap-free bases, and maximise $1 + c(k-1)/b$. Shibata's blocks are of this kind with $c = 3$; so are the two odd blocks of Section 10, and the quantity λ there is the loss O per block.

11.1. **The two blocks at $k = 4$ and $k = 5$.** At $k = 4$ the block $B^{(4)}$ decomposes into the $c = 3$ closed trails carrying edges $\{9, 0, 8\}$, $\{2, 3, 7, 1\}$ and $\{6, 4, 5\}$, based at `XAA`, `XAB` and `XBA`, which spell the three strings (6); $1 + 3 \cdot 3/10 = 19/10$. At $k = 5$ the block $B^{(5)}$ decomposes into the trails carrying $\{0, 1, 2, 11, 12\}$, $\{3, 4, 5, 6, 9, 10\}$ and $\{7, 8\}$, based at `XAAA`, `XAAB` and `XBXB`, spelling (8); $1 + 3 \cdot 4/13 = 25/13$. The two-edge trail at $k = 5$ is the 2-cycle on the 4-mers `XBXB` and `BXBX`, and it is available only because k is odd: the two vertices of a 2-cycle have period 2, so for even k a $(k-1)$-mer on such a cycle has equal first and last letters and cannot both begin with a fresh letter and end with a shared one — which is what the overlap-0 argument of Section 3.2 asks of a base. That is one concrete reason the $k = 5$ optimum of the search below (25/13) exceeds the $k = 4$ one (19/10), and a hint about where a $k = 4$ upper bound might come from.

Remark 31 (a measurement; not part of the machine-checked development). We enumerated cyclic block words up to rotation and letter renaming, filtered by the three conditions of Remark 30, enumerated all simple cycles of the block graph, computed maximum-cardinality exact covers of the edge set by cycles, and tested every assignment of bases for the pairwise-overlap-0 condition; a second variant drops that condition and instead simulates greedy on the c pieces to measure the overlap O it recovers. Two calibrations came out exactly right: at $k = 3$, over blocks of length at most 12 on two shared and one fresh letter, the maximum is 9/5 and is attained only at $b = 5$, by four blocks up to rotation and renaming, one of which is `ABXAX`, the block of [2] — and 9/5 is the value they *prove* to be ρ_3; at $k = 6$, $b = 9$, the search returns the construction of [2] and its ratio $2 = 1 + (2 \cdot 5 - 1)/9$, including the one unit of overlap it loses. Eight sweeps were run in all, each over one alphabet and one range of block lengths:

k	alphabet	block lengths	threshold
3	2 shared, 1 fresh	$3 \le b \le 12$	ratio ≥ 1.8
4	2 shared, 1 fresh	$4 \le b \le 16$	ratio ≥ 1.9
4	3 shared, 1 fresh	$4 \le b \le 14$	ratio ≥ 1.9
4	2 shared, 2 fresh	$4 \le b \le 13$	ratio ≥ 1.9
4	2 shared, 1 fresh	$4 \le b \le 11$	ratio ≥ 1.9, overlap-0 dropped
5	2 shared, 1 fresh	$5 \le b \le 16$	ratio ≥ 1.8
5	2 shared, 1 fresh	$5 \le b \le 14$	ratio $\ge 25/13$, overlap-0 dropped
6	2 shared, 1 fresh	$6 \le b \le 14$	ratio > 2

No sweep used more than three shared or more than two fresh letters, and none combined three shared letters with two fresh ones. Nothing in these ranges beats 19/10 at $k = 4$, 25/13 at $k = 5$, or the value 2 of [2] at $k = 6$. This is a measurement of a search, not an upper-bound proof: it examines families built from

a single repeated block over small alphabets and short blocks. Widening it is not the productive direction — the number of block words grows exponentially in b — whereas an upper bound for $k = 4$ in the style of the trail-decomposition proof of $\rho_3 \le 9/5$ in [2] would settle the cell; Remark 30 says exactly what such an argument has to bound, namely the density of closed trails in a balanced de Bruijn subgraph whose base vertices are pairwise overlap-free.

11.2. **Shibata's offsets, and the sweep that produced the odd blocks.**

Remark 32 (a measurement; not part of the machine-checked development). The three starting offsets at which [1] cuts open his cyclic words are fixed there without comment, and they matter: after the closing phase the output length is $t(10s-1)$ minus the overlap greedy is then forced to spend, so a different triple could in principle waste less. We swept all $k \cdot s \cdot (s+2)$ triples at $s = 4, \dots, 8$, solving the continuation exactly — the minimum total overlap over all legal executions on the border graph of the closed strings — and reading the asymptotic per-block loss off as $\lambda = \mathrm{loss}(3) - \mathrm{loss}(2)$. The sweep is exhaustive: 158, 303, 514, 803, 1182 triples survive the validity filter at $s = 4, \dots, 8$, and none was dropped on the 10^6-node cap of the exact continuation. Its outcome is that the source's triple is the unique minimiser of λ at every one of those five values of s, with $\lambda = s - 2$ exactly, which is what produces the ratio $(9s+1)/(4s+2)$; nothing is even tied with it. Two by-products are worth recording. At $s = 7$ and $s = 8$ there are triples whose *maximum* border overlap is smaller than the source's and which are nevertheless worse, so the maximum is not the cost — the forced sequence of maxima is. And no triple at any $s \le 8$ makes the three closed strings pairwise overlap-free, which is the only way this shape could reach $1 + 3(k-1)/(4s+2) \to 5/2$; whether some other three-circuit block does is the concrete route to a bound above $9/4$.

Remark 33 (a measurement; not part of the machine-checked development). The blocks of Section 10 were found by enumerating blocks of the shape described there — one cyclic word U of length k over three shared letters carrying a single fresh letter, plus two cyclic words obtained from periodic vertices of U — together with every closing-offset triple of each, and solving the continuation exactly for each surviving pair. The counts, all with no candidate dropped on the node cap, are: at $k = 7$, 5 340 blocks and no candidate at all reaching ratio 2; at $k = 8$, 22 200 blocks, of which 24 offset choices beat 2 and all 24 give exactly $37/18$, on twelve blocks that are Shibata's up to rotation and relabelling — which is the sweep's calibration, since it rediscovers his $k = 8$ block; at $k = 9$, 87 162 blocks, of which 96 offset choices beat 2 and all 96 give exactly $43/21$; and at $k = 11$ a restricted sweep over 5 442 blocks with $|V| \in \{6, 7\}$ and $|W| \in \{8, 9\}$, of which 2 523 beat 2, the best at $52/25$.

The scope has to be stated exactly, because it is narrow. The sweep fixes the block shape and the alphabet at three shared letters plus one fresh — words over only two shared letters are included, since the enumeration runs over all of $\{\mathtt{x}, \mathtt{y}, \mathtt{a}\}^{k-1}$ — and it does not cover four or more shared letters, four or more circuits per block, two or more fresh letters per block, or blocks whose V and W do not come from a periodic vertex of U. At $k = 11$ it is restricted further, as above, so $52/25$ is the best *found* there and no optimality is claimed. The sweep also pre-filters candidates by the estimate $\lambda \approx \mathrm{loss}(2)/2$, which is a lower bound on λ exactly when the additive constant c in $\mathrm{loss}(t) = \lambda t + c$ is ≤ 0; it was $c = -1$ for

every survivor at $k = 8$, 9 and 11, but a candidate with $c > 0$ could in principle have been pruned. So this is a strong filter, not a proof of optimality, at any k; making it unconditional means computing $\mathrm{loss}(3) - \mathrm{loss}(2)$ for every pair passing the only unconditional test $1 + 3(k-1)/b > 2$, of which there are 722 112 at $k = 8$ and 4 834 404 at $k = 9$, and it was not run.

12. Verification

Lemmas 9 and 10, Theorems 12, 13, 14, 15, 17, 18, 19, 21, 24, 26, 28 and Propositions 22, 25, 27 have been formalised and checked by the Lean 4 [14] kernel against Mathlib [13], with the following division of labour. Overlap, merge, occurrence as a contiguous block, the common-superstring predicate, the legality test of Definition 8 and the replay of a certificate are defined as computable functions; that ov is the overlap and not merely a function of that name is itself proved, in the two halves "the length-$\mathrm{ov}(p, q)$ suffix of p is a prefix of q" and "no longer suffix is". Lemma 9 is proved from first principles and depends on no axiom beyond propositional extensionality and quotient soundness — in particular on no evaluation of compiled code — and Lemma 10 is an immediate consequence of it. This matters, because it is the half that could flatter the results. For each instance three finite facts — that the input strings are pairwise distinct, that the exhibited word is a common superstring, and that the list of merges replays as a legal execution — are discharged by compiled evaluation (Lean's `native_decide`), which trusts the compiler and the runtime for those three evaluations and records that trust as an axiom per evaluation; the largest of them replays 335 merges, each maximising the overlap over all ordered pairs of distinct live strings. No proof uses `sorry`; apart from the native-evaluation axioms just described, the only axioms occurring anywhere are Lean's standard `propext`, `Classical.choice` and `Quot.sound`; the final numerical comparisons of Theorem 28(i) and of Propositions 22–27 depend on no axiom at all; and there is no external certificate file. What is *not* formalised is: the passage from an instance to ρ_k (Section 2), since the development defines no supremum, and with it Theorems 2 and 6 and Corollaries 3–7 as stated; the results quoted from [2], [5], [6], [1] and [3] in those corollaries and in Table 1; the identification, in Sections 3, 8.1 and 10, of the instances as unions of cyclic spectra of block words, the instances entering the development as explicit lists of strings; the repair of the two border-overlap entries in Section 9; and everything in Remarks 11, 20, 23, 29, 30, 31, 32 and 33.

Independently of the formal development, implementations written against the definitions of [2] and of [1] were used first as calibrations and then as cross-checks. The first reproduces every number printed in [2] for the two constructions of Theorem 17, computes exact optima by dynamic programming over the overlap graph, simulates both block families of Section 3 for $t \le 5$, and at $t = 1$ enumerates *all* legal greedy executions, the maximum output length being exactly the certified 19 at $k = 4$ and 25 at $k = 5$ — so on one block the exhibited execution is not merely legal but worst possible. The second reproduces, from the definitions of Section 8.1 rather than from the printed numbers, the cardinality, the optimum, the length of the greedy output, the sequence of phase maxima and all eighteen border-overlap entries of [1] at $s = 4, 5, 6, 7$; it recomputes the true maximum overlap over all ordered pairs at every step rather than replaying a certificate, and it is what produced the simulations quoted in Remarks 23 and 29. Both implementations also confirmed

that along every certificate above no live string is ever a substring of another, so the executions are legal for the substring-eliminating formulation of Greedy as well. The generators that emitted the Lean data were re-run afterwards and every emitted block was checked to appear verbatim in the checked source. None of this is part of the machine-checked development. The repository is private: repository reference to be added at submission.

References

[1] H. Shibata, *Disproving the Greedy Superstring Conjecture*, arXiv:2609.01365 (v1, 1 September 2026), doi:10.48550/arXiv.2609.01365. Theorem 14 is the bound $\rho_k \geq (9k+2)/(4k+4)$ for every even $k \geq 10$ and Corollary 15 is $\rho \geq 9/4$ for the unrestricted problem; Theorem 10 is the optimum (9), Theorem 13 the length of the greedy output, Lemma 5 the distinctness of the strings of a block, and Table 1 the border-overlap table discussed in Section 9. All statements cited above are those of version v1, read from its LaTeX source. As of 3 September 2026 the arXiv record carries no journal reference and v1 is the current version.

[2] N. Chukhin, A. S. Kulikov, I. Mihajlin and A. Smal, *The Greedy Superstring Algorithm Achieves Ratio 2 for Strings of Length 6 Already*, arXiv:2608.20018 (v1, 20 August 2026; v2, 24 August 2026), doi:10.48550/arXiv.2608.20018. All statements cited above are those of version v2, read from its LaTeX source; the quoted display (4) is from its introduction, the theorem $\rho_3 = 9/5$ from its sections on the lower and upper bounds for ρ_3, and the theorem $\rho_k \geq 2$ for $k \geq 6$ from its section on the lower bound for ρ_k. As of 3 September 2026 the arXiv record carries no journal reference and v2 is the current version.

[3] N. Chukhin, A. S. Kulikov, I. Mihajlin and A. Smal, *A Tight Cycle-Cover Inequality for Shortest Common Superstring*, Electronic Colloquium on Computational Complexity, Report TR26-157 (2026), `https://eccc.weizmann.ac.il/report/2026/157/` (submitted 27 August 2026; the ECCC record gives 30 August 2026 as the date of publication). Cited above only for the bound 3 on the greedy ratio, which supersedes the entry 3.396 of (4); the report also gives a 7/3-approximation for the shortest common superstring problem. That bound carries no number of its own there: its abstract states it; its Theorem 3.1 is the cycle-cover inequality of its title, which is the case $\gamma = 1$ of the inequality its own Section 2.5 numbers (3); and that section reads "We prove the inequality (3) for $\gamma = 1$ (see Theorem 3.1). This implies that the greedy algorithm is 3-approximate", the implication being the one it attributes to [9, Section 5]. All numbering here is that of the report, not of this paper. It is the reference `ChukhinKMS26b` of the bibliography of [1].

[4] M. Englert, N. Matsakis and P. Veselý, *Approximation Guarantees for Shortest Superstrings: Simpler and Better*, in: 34th International Symposium on Algorithms and Computation (ISAAC 2023), Leibniz International Proceedings in Informatics **283**, Schloss Dagstuhl — Leibniz-Zentrum für Informatik, 2023, pp. 29:1–29:17, doi:10.4230/LIPIcs.ISAAC.2023.29. Cited only as the origin of the bound 3.396 quoted in (4), which its published abstract gives as an approximation guarantee of $(\sqrt{67}+2)/3 \approx 3.396$ for Greedy. We have not consulted the text.

[5] B. Cazaux and E. Rivals, *Relationship between superstring and compression measures: New insights on the greedy conjecture*, Discrete Applied Mathematics **245** (2018), 59–64, doi:10.1016/j.dam.2017.04.017. Proposition 2 is the lower bound $\rho_r \geq 2 - 1/r$, Theorem 3 the two-sided bound (2), and the conjecture (3) is the sentence quoted in Section 1.3, immediately following Theorem 3; the sentence about [6] quoted in Remark 1 is from Section 1.

[6] A. S. Kulikov, S. Savinov and E. Sluzhaev, *Greedy Conjecture for Strings of Length 4*, in: Combinatorial Pattern Matching (CPM 2015), Lecture Notes in Computer Science **9133**, Springer, 2015, pp. 307–315, doi:10.1007/978-3-319-19929-0_26. Cited here only through its published abstract, quoted in Remark 1; the full text is closed access.

[7] M. S. Nikolaev, *Greedy Conjecture for the Shortest Common Superstring Problem and Its Strengthenings*, in: String Processing and Information Retrieval (SPIRE 2024), Lecture Notes in Computer Science **14899**, Springer, 2024, pp. 233–248, doi:10.1007/978-3-031-72200-4_18; arXiv:2407.20422. The phrases quoted above are from the introduction of the arXiv version v1, which records the greedy conjecture as known for strings of length at most 3 and for strings of length exactly 4, and contains no statement of monotonicity of ρ_k in k.

[8] P. A. Pevzner, *l-Tuple DNA Sequencing: Computer Analysis*, Journal of Biomolecular Structure and Dynamics **7** (1989), no. 1, 63–73, doi:10.1080/07391102.1989.10507752. Only the elementary half of the lemma attributed to this paper is used, and it is proved here (Lemma 9).

[9] A. Blum, T. Jiang, M. Li, J. Tromp and M. Yannakakis, *Linear approximation of shortest superstrings*, in: Proceedings of the 23rd Annual ACM Symposium on Theory of Computing (STOC 1991), ACM, 1991, pp. 328–336, doi:10.1145/103418.103455; journal version, Journal of the ACM **41** (1994), no. 4, 630–647, doi:10.1145/179812.179818. Quotations and lemma references above are to the journal version, whose Section 1 carries the sentence quoted in Section 1.1 and whose Section 2 carries, after its Claim 1, the endpoint-inheritance statement that the maximum overlap of two live strings s, t equals $\mathrm{ov}(\mathrm{last}(s), \mathrm{first}(t))$.

[10] J. A. Storer, *Data Compression: Methods and Theory*, Computer Science Press, Rockville, Md., 1988.

[11] J. Tarhio and E. Ukkonen, *A greedy approximation algorithm for constructing shortest common superstrings*, Theoretical Computer Science **57** (1988), no. 1, 131–145, doi:10.1016/0304-3975(88)90167-3.

[12] J. S. Turner, *Approximation algorithms for the shortest common superstring problem*, Information and Computation **83** (1989), no. 1, 1–20, doi:10.1016/0890-5401(89)90044-8.

[13] The mathlib Community, *The Lean mathematical library*, in: Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020), ACM, 2020, pp. 367–381, doi:10.1145/3372885.3373824.

[14] L. de Moura and S. Ullrich, *The Lean 4 theorem prover and programming language*, in: Automated Deduction — CADE-28, Lecture Notes in Computer Science, Springer, 2021, pp. 625–635, doi:10.1007/978-3-030-79876-5_37.