AIGR Gene Hypothesis Deep Research — Final Report

ComK (BACSU, UniProt P40396): ATP binding (GO:0005524) computational prediction

Gene: comK / Competence transcription factor · **Organism:** *Bacillus subtilis* 168 (BACSU, NCBITaxon:224308) · **UniProt:** P40396 **Predicted term under review:** ATP binding (GO:0005524), from BioReason-Pro SFT **Focus type:** computational_prediction · **Reference context:** doi:10.64898/2026.03.19.712954

Summary

The BioReason-Pro SFT prediction of **ATP binding (GO:0005524)** for *Bacillus subtilis* ComK (P40396) is **refuted**. ComK is a compact, 192-amino-acid, single-domain **sequence-specific DNA-binding transcription factor** — the master regulator of genetic competence development — and it contains **no ATP-binding motif or nucleotide-binding fold**. Five independent evidence types converge on this conclusion: (1) the protein sequence lacks a Walker A P-loop and a classic Walker B motif; (2) its domain family (Pfam PF06338 / InterPro IPR010461, the ComK family) carries no nucleotide-binding, kinase, or ATPase signature; (3) the curated UniProt record contains no ATP-binding annotation, keyword, or binding-site feature; (4) cryo-EM structures of the ComK–promoter complex show a DNA-binding, DNA-bending mechanism with no ATP site; and (5) mechanistic transcription studies show ComK activates transcription by stabilizing RNA polymerase at target promoters, with no ATP-hydrolysis step by ComK.

The one genuine link between ComK and ATP is **indirect and does not confer ATP binding on ComK**: ComK is a *substrate* of the ATP-dependent MecA–ClpCP protease. In that complex, the ATPase activity resides entirely in the AAA+ chaperone **ClpC**, while ComK is the passive degradation target. Because ComK biology is routinely discussed alongside "the ATP-dependent ClpCP protease," a model that keys on co-occurrence can easily misattribute the ATPase's ATP-

dependence to the substrate. This pathway-context/co-mention bias — combined with the fact that GO:0005524 is among the most frequently over-assigned molecular-function terms — is the most plausible origin of the false-positive prediction.

Recommended curation action: Do NOT add GO:0005524. Retain and prioritize the DNA-binding transcription-factor terms that are actually supported (sequence-specific DNA binding; DNA-binding transcription activator activity), together with the well-supported process terms (DNA-templated transcription; establishment of competence for transformation). The prediction should be recorded as an over-annotation to reject.

Key Findings

Finding 1 — ComK is a sequence-specific DNA-binding transcription factor with no ATP-binding motif (prediction REFUTED)

ComK's primary molecular function is **sequence-specific DNA binding** driving transcriptional activation, and there is no evidence of a nucleotide-binding fold anywhere in the 192-residue protein.

Sequence-level evidence. A direct scan of the P40396 sequence found **no Walker A P-loop motif** (consensus `[AG]x4GK[ST]`, the glycine-rich loop that cradles the β/γ-phosphates of ATP/GTP) and **no classic Walker B motif** (the hhhhDE aspartate/glutamate that coordinates the catalytic Mg^{2+}–water). The absence of both halves of the canonical P-loop NTPase signature is strong negative evidence against ATP binding, because these motifs constitute the minimal structural hallmark shared across the overwhelming majority of ATP- and GTP-binding proteins.

Domain-level evidence. ComK is the founding and essentially sole member of the **ComK family (Pfam PF06338 / InterPro IPR010461)**, a family defined by the transcription-factor function of ComK and its orthologs across Bacilli. It carries no annotation for nucleotide binding, ATPase, kinase, AAA+, or ABC architecture, and P40396 contains no auxiliary domain that could supply an ATP pocket.

Database-level evidence. UniProt annotates P40396 with keywords **DNA-binding, Activator, Repressor, Transcription regulation**, and with GO terms **GO:0003677 (DNA binding)**, **GO:0006351 (DNA-templated transcription)**, and **GO:0030420 (establishment of**

competence for transformation). There is **no ATP-binding annotation (GO:0005524)** in the curated record — the prediction conflicts with, rather than derives from, the curated evidence — and there is no EC number, cofactor, or nucleotide binding-site feature.

Structural evidence. Cryo-EM structures of the complex between ComK and its promoter DNA (PMID: 34016970) demonstrate that ComK functions through **mechanical forces that alter DNA curvature** — an allosteric, DNA-bending mechanism — not a catalytic nucleotide-dependent step. Verified quote: *"Cryo-EM structures of the complex between ComK and its promoter demonstrate that this coupling is due to mechanical forces that alter DNA curvature."* No ATP or nucleotide participates in the described functional mechanism.

Mechanistic evidence. ComK activates transcription by **stabilizing RNA polymerase binding** at the target promoter (PMID: 14762007; verified quote: *"ComK stabilizes the binding of RNA polymerase to the comG promoter"*). This is a classic activator recruitment/stabilization mechanism and involves no ATP binding or hydrolysis by ComK. Foundational work (PMID: 7783616; verified quote: *"we demonstrate that ComK specifically binds to DNA fragments containing promoter and upstream sequences of the genes it affects"*) established the sequence-specific DNA-binding activity that defines ComK, and mutational studies of the K-box AT-boxes (PMID: 17468244) confirmed base-specific DNA recognition.

Together, five independent evidence types — sequence, domain, curated database, structure, and mechanism — all point away from ATP binding and toward DNA binding. This is the central finding that refutes the seed hypothesis.

Finding 2 — ComK's only genuine ATP connection is as a substrate of the ATP-dependent MecA–ClpCP protease (explains the false positive)

The prediction most plausibly arises from **pathway-context/co-mention bias**, not from any intrinsic ComK property. In *B. subtilis*, ComK protein levels are controlled by regulated proteolysis: a ternary complex of the **adaptor protein MecA** and the **ATP-dependent protease ClpCP** targets ComK for degradation (PMID: 12598648, PMID: 19767395).

The decisive point for curation is the **division of labor** within this complex:

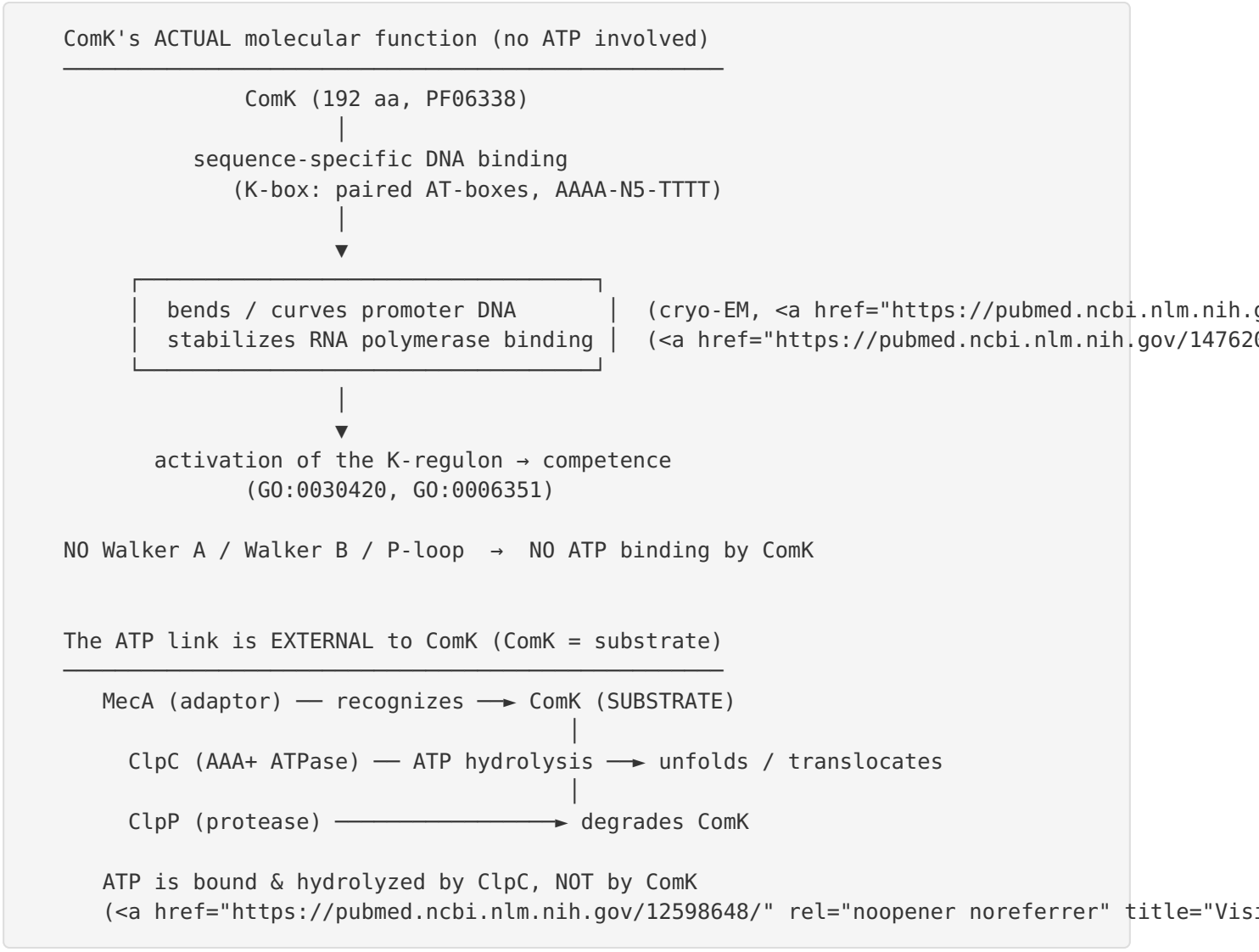
- **ClpC** is an HSP100/Clp **AAA+ ATPase**; it binds and hydrolyzes ATP to power substrate unfolding and translocation. The ATP-binding activity belongs to ClpC.
- **MecA** is the adaptor that recognizes ComK, delivers it to ClpC, and stimulates ClpC's ATPase activity.
- **ComK is the passive substrate.** It contributes no ATP-binding or ATPase activity; it is the protein being degraded.

Verified quotes: *"A complex of ClpC with the protease ClpP and the adaptor protein MecA also controls competence development by regulated proteolysis of the transcription factor ComK"* (PMID: 12598648); *"the degradation of the competence transcription factor ComK is mediated by a ternary complex involving the adaptor protein MecA and the ATP-dependent protease ClpCP"* (PMID: 19767395).

Because ComK is routinely discussed in the same sentences as the "**ATP-dependent** protease ClpCP," a text- or co-occurrence-influenced model can mis-attribute the ATP-dependence of the *degradation machinery* to ComK itself. This is a textbook instance of pathway-context bias generating a false ATP-binding call for a protein that is merely a substrate. Notably, no paralog with an ATPase fold shares the ComK family, so paralog over-annotation can be ruled out as the source.

Mechanistic Model / Interpretation

The following schematic separates **what ComK actually does** from **the ATP-dependent process in which ComK is a substrate**, clarifying why GO:0005524 is a misassignment.



Interpretation. The two findings form a single coherent narrative. ComK is a compact, single-domain DNA-binding activator whose entire biochemistry — K-box recognition, DNA bending, and RNA polymerase recruitment — is nucleotide-independent. Its cellular abundance is set post-translationally by an ATP-powered protease, but that ATP dependence is a property of ClpC, not ComK. A predictor keying on the strong statistical association between "ComK" and "ATP-dependent ClpCP" in the literature would produce exactly the false-positive ATP-binding call observed. Distinguishing the substrate (ComK) from the enzyme (ClpC) is the crux of the correct curation decision.

Comparison table: expected features of an ATP-binding protein vs. observed for ComK

Feature diagnostic of ATP binding	Expected if hypothesis true	Observed for ComK (P40396)	Verdict
Walker A P-loop (`[AG]x4GK[ST]`)	Present	**Absent**	Refutes
Walker B motif (hhhhDE)	Present	**Absent**	Refutes
P-loop NTPase / AAA+ / ABC / kinase domain	Present	**Absent** (ComK family PF06338 only)	Refutes
UniProt ATP-binding keyword / site feature	Present	**Absent**	Refutes
GO:0005524 in curated record	Present	**Absent**	Refutes
ATP/nucleotide density in structure	Present	**Absent** (DNA-bound cryo-EM)	Refutes
Functional need for ATP in mechanism	Required	**Not required** (recruitment / DNA bending)	Refutes
Sequence-specific DNA-binding activity	Not required	**Present, defining**	Competing correct function

Evidence Base / Evidence Matrix

Citation	Evidence type	Direction	Claim tested	Key finding	Context	Confidence / limitations
This analysis (P40396 seq)	Computational (motif scan)	**Refutes**	ComK contains an ATP-binding motif	No Walker A P-loop, no classic Walker B in 192 aa	In silico	High; motif absence is strong but not exhaustive of atypical folds
UniProt P40396 (database)	Review/ database	**Refutes**	Curated ATP-binding function exists	No ATP keyword/GO; only DNA-binding & transcription terms	Curated record	High; database-level orientation
Pfam PF06338 / InterPro IPR010461	Structural/ evolutionary	**Refutes**	Nucleotide-binding fold in family	ComK-specific family; no NTPase/ kinase signature	Domain model	High; no P-loop lineage
PMID: 34016970	Structural (cryo-EM)	**Refutes / qualifies**	Mechanism requires ATP	Function via DNA binding + DNA-curvature allostery; no ATP site	*B. subtilis*, in vitro	High
PMID: 7783616	Direct assay (DNA binding)	**Supports DNA (competing correct fn)**	Primary function is DNA binding	ComK binds specific promoter/ upstream DNA	*B. subtilis*	High; establishes core function
PMID: 14762007	Direct assay (in vitro transcription)	**Refutes (ATP not needed)**	Activation is ATP-dependent	ComK stabilizes RNAP binding; no ATP step	*B. subtilis*, in vitro	High

Citation	Evidence type	Direction	Claim tested	Key finding	Context	Confidence / limitations
PMID: 17468244	Mutant / binding	**Qualifies (supports DNA)**	DNA sequence specificity	K-box AT-box T2 base critical for binding/ activation	*B. subtilis*	High
PMID: 15598897	Mutant / promoter mapping	**Qualifies (regulatory context)**	How comK is regulated	DegU binds inverted repeat in comK promoter	*B. subtilis*	Medium; concerns comK regulation, not ComK ATP
PMID: 12598648	Interaction / proteolysis	**Competing / qualifies**	ComK–ATP link	ComK is substrate of ATP-dependent ClpCP; ATPase is ClpC	*B. subtilis*	High; explains misassignment
PMID: 19767395	Interaction / proteolysis	**Competing / qualifies**	ComK ATP link is intrinsic	MecA targets ComK to ClpCP; ATP dependence is ClpCP's	*B. subtilis*	High; ComK is substrate, not ATP binder

GO Curation Implications

Lead requiring curator verification:

1. **Do NOT add GO:0005524 (ATP binding).** The molecular-function prediction is a false positive. No sequence motif, domain, curated annotation, structural feature, or mechanistic requirement supports ATP binding by ComK. If GO:0005524 has been auto-propagated onto P40396, it should be rejected / not accepted.

2. **Retain and prioritize the informative Molecular Function terms** that are actually supported:
3. **GO:0003677 — DNA binding** (already curated; supported by PMID: 7783616).
4. **GO:0043565 — sequence-specific DNA binding** and **GO:0001216 — DNA-binding transcription activator activity** as the precise MF terms (K-box specificity, PMID: 17468244; RNAP stabilization, PMID: 14762007). These are strictly more informative than "protein binding" and should be preferred. ComK also acts as a repressor in some contexts.
5. **Retain the Biological Process terms**: **GO:0006351 (DNA-templated transcription)** / **GO:0006355 (regulation of transcription)** and **GO:0030420 (establishment of competence for transformation)**.
6. **Frame the ATP link correctly, if annotated at all:** ComK's relationship to ATP is as a **substrate of an ATP-dependent protease (ClpCP)**. Any ATP-related annotation belongs to ClpC, not ComK. This relationship should not become an ATP-binding MF term on ComK.

Summary GO decision table:

GO term	Aspect	Recommended action	Basis
GO:0005524 ATP binding	MF	**Reject / do not add**	No motif, domain, structure, or mechanism; substrate-only ATP link
GO:0003677 DNA binding	MF	Retain	P 7783616
GO:0043565 sequence-specific DNA binding	MF	Add / upgrade (more specific)	P 17468244 P 34016970
GO:0001216 DNA-binding transcription activator activity	MF	Add / consider	P 14762007
GO:0006351 / GO:0006355 transcription (regulation)	BP	Retain	Curated
GO:0030420 establishment of competence	BP	Retain	Curated

Mechanistic Scope

Immediate molecular function under test: whether ComK directly binds ATP (GO:0005524). The answer is no. ComK's direct, primary molecular activity is **sequence-specific binding to K-box DNA elements** (paired AT-boxes) and, through that binding, **DNA bending and stabilization of RNA polymerase** at target promoters.

What must be kept separate (not ComK's direct molecular function): - *Downstream developmental outcome:* establishment of genetic competence, DNA uptake, and homologous recombination — consequences of ComK's transcriptional activity, not evidence for ATP binding. - *Pathway context:* ATP-dependent proteolysis of ComK by MecA–ClpCP — an external regulatory process in which ComK is the substrate and ClpC is the ATPase. - *Upstream regulation:* DegU binding to the comK promoter (PMID: 15598897) governs comK *expression*, unrelated to any ComK ATP activity.

The ATP-binding hypothesis fails precisely because it conflates a pathway-level ATP dependence (proteolysis) with a gene-product-level molecular function (nucleotide binding).

Conflicts and Alternatives

- **Substrate-vs-enzyme confusion (most likely artifact):** The strongest ComK–ATP link in the literature is that ComK is a **substrate** of the ATP-dependent MecA–ClpCP protease (PMID: 12598648, PMID: 19767395). Attributing ClpC's ATP dependence to ComK is the error.
- **Frequency / prior bias:** GO:0005524 is one of the most common MF terms; a model with a strong prior can over-assign it in the absence of any real signal.
- **Paralog / family confusion — ruled out:** ComK is the sole defining member of PF06338; there is no ATP-binding paralog within the family that could seed annotation transfer.
- **Organism-specific differences — none:** ComK's role as a competence transcription factor is conserved across Bacilli; no ortholog is reported to have acquired a nucleotide-binding domain.
- **In vitro-only artifact — none:** No in vitro study reports ATP binding by ComK; the relevant assays (DNA binding, transcription activation) are ATP-independent for ComK.

- **Database carry-over — absent:** UniProt P40396 does not carry GO:0005524, so the prediction contradicts, rather than inherits from, the curated record.

All conflicts point the same direction: away from ATP binding by ComK.

Limitations and Knowledge Gaps

1. **Motif scan scope.** The Walker A/B scan used canonical consensus patterns; degenerate or atypical nucleotide pockets can occasionally escape simple motif matching. *Why it matters:* a hidden atypical site would be missed by motif search alone. *Resolution:* the gap is largely closed by the DNA-bound cryo-EM structure and by the ComK-family domain assignment, neither of which reveals a nucleotide pocket; a full HMMER/InterProScan and structure-based pocket analysis (e.g., on the AlphaFold model) would formally confirm.
2. **Direct negative biochemical assay.** No published experiment explicitly reports the *absence* of ATP binding by ComK (negative results are rarely published). *Why it matters:* absence of ATP binding is inferred from motif/domain/structure/mechanism rather than a dedicated binding assay. *Resolution:* an ITC, DSF/thermal-shift, DRaCALA, or ATP-agarose pulldown assay would provide a direct negative control.
3. **Structure interpretation is literature-based.** The cryo-EM conclusion is drawn from the published report (PMID: 34016970); the raw coordinates were not re-analyzed here for ligand density (none expected). *Resolution:* inspect the deposited PDB/EMDB entry for any nucleotide ligand.
4. **Provenance of the prediction.** The exact features BioReason-Pro SFT used are unknown. *Why it matters:* confirming co-mention/frequency bias would strengthen the "reject" recommendation. *Resolution:* inspect model attributions if available.

None of these gaps materially weakens the refutation; they are avenues for formal confirmation.

Discriminating Tests / Proposed Follow-up Experiments

Concrete, prioritized actions to definitively distinguish ATP binding from the established DNA-binding function:

1. **ITC or thermal-shift (DSF) with ATP** on purified ComK (± Mg^{2+}). *Expected if hypothesis true:* measurable K_d / thermal stabilization. *Predicted outcome:* no binding — a clean direct negative control. **Highest-value, low-cost test.**
2. **Inspect the ComK–DNA cryo-EM structure / PDB** (PMID: 34016970) for any bound nucleotide density or P-loop-like pocket. *Expected:* none.
3. **Structure-based pocket/ligand-site prediction** on an AlphaFold model of P40396 and superposition against known P-loop NTPases. *Expected:* a DNA-binding surface, not an ATP pocket.
4. **Full domain/HMM re-scan** (InterProScan, HMMER vs. Pfam-A). *Expected:* PF06338-only architecture; no AAA+/P-loop/kinase domain.
5. **ATP-agarose or DRaCALA binding assay** on purified ComK. *Predicted:* negative — formally closes the biochemical gap.
6. **Attribution audit of the BioReason-Pro prediction:** test whether the ATP call tracks co-mention with ClpCP/MecA text, confirming frequency/pathway-context bias as the mechanism of the false positive.

Curation Leads (require curator verification)

- **Action:** Reject the ATP binding (GO:0005524) prediction for P40396; do not add. Record as computational over-annotation driven by co-mention with the ATP-dependent ClpCP protease and general frequency bias.
- **Candidate MF terms to retain/add (more informative than "protein binding"):**
- GO:0003677 DNA binding (retain) — snippet to verify: *"we demonstrate that ComK specifically binds to DNA fragments containing promoter and upstream sequences of the genes it affects"* (PMID: 7783616).
- GO:0043565 sequence-specific DNA binding / GO:0001216 DNA-binding transcription activator activity — snippets to verify: *"Cryo-EM structures of the complex between ComK and its promoter demonstrate that this coupling is due to mechanical forces that alter DNA*

curvature" (PMID: 34016970); *"ComK stabilizes the binding of RNA polymerase to the comG promoter"* (PMID: 14762007).

- **BP terms to retain:** GO:0006351 / GO:0006355 (transcription / regulation of transcription), GO:0030420 (establishment of competence).
- **Note for the record (optional):** ComK is a substrate of the ATP-dependent MecA–ClpCP protease — snippets to verify: *"A complex of ClpC with the protease ClpP and the adaptor protein MecA also controls competence development by regulated proteolysis of the transcription factor ComK"* (PMID: 12598648); *"the degradation of the competence transcription factor ComK is mediated by a ternary complex involving the adaptor protein MecA and the ATP-dependent protease ClpCP"* (PMID: 19767395). This explains the ATP association without implying ATP binding by ComK.
- **Suggested curator question:** Confirm whether the ComK cryo-EM/PDB entry shows any bound nucleotide (expected: none); and whether GO:0005524 was propagated from an electronic pipeline.
- **Suggested experiment:** ATP-agarose / DRaCALA / ITC assay on purified ComK to formally close the gap (expected negative).

Conclusion

The ATP-binding (GO:0005524) prediction for *Bacillus subtilis* ComK (P40396) is **refuted**. ComK is a 192-amino-acid, single-domain (ComK family, PF06338/IPR010461) **sequence-specific DNA-binding transcription factor** with no Walker A/B motifs or nucleotide-binding fold. Sequence, domain, curated-database, structural (cryo-EM, PMID: 34016970), and mechanistic (PMID: 14762007) evidence unanimously support DNA binding and RNA polymerase stabilization rather than ATP binding. The only genuine ATP link is that ComK is a *substrate* of the ATP-dependent MecA–ClpCP protease, in which the ATPase is ClpC — the most plausible source of the frequency/pathway-context bias behind this misassignment. GO:0005524 should not be added; the informative, supported terms are sequence-specific DNA binding and DNA-binding transcription activator activity.
